GENETICS OF HUMAN OBESITY IN THE POST-GENOME WIDE ASSOCIATION STUDY ERA

GENETICS OF HUMAN OBESITY IN THE POST-GENOME WIDE ASSOCIATION STUDY ERA

By AIHUA LI, MD, M.SC

A Thesis Submitted to the School of Graduate Studies in Partial Fulfilment of the Requirements for the Degree Doctor of Philosophy

McMaster University DOCTOR OF PHILOSOPHY (2016) Hamilton, Ontario

(Health Research Methodology)

TITLE: Genetics of human obesity in the post-genome wide association era

AUTHOR: Aihua Li, MD (Huanan University, Hunan, China), M.Sc. (Capital University of Medical Sciences, Beijing, China), M.Sc. (McMaster University, Ontario, Canada)

SUPERVISOR: Dr. David Meyre

NUMBER OF PAGES: xx 283

LAY ABSTRACT

Obesity is a chronic disorder triggered by multiple genes, environmental factors and their interactions. Currently most of the common genetic alterations, called single nucleotide polymorphisms (SNPs), associated with adult body mass index (BMI) were identified in populations of European ancestry. This thesis aims to: 1) investigate whether these BMI-associated SNPs are also associated with BMI in other ethnic groups; 2) explore the parental and child genetic contributions in children from birth to 5 years; 3) examine the maternal and child genetic contribution to maternal gestational weight gain (GWG) and postpartum weight retention. The major findings are: 1) BMI SNPs identified in Europeans are partially generalizable to other five ethnicities; 2) The collective SNPs contributing to adult BMI start to exert their effect at birth and in early childhood; and 3) There is a genetic link between pre-pregnancy BMI and offspring birth weight and maternal postpartum weight retention.

ABSTRACT

Obesity has more than doubled worldwide since 1980 and it has become the focus of public health due to a wide range of serious complications. It is believed to be a complex disorder triggered by multiple genes, environmental factors and their interactions. The total number of single nucleotide polymorphisms (SNPs) associated with adult body mass index (BMI) at genome-wide significance level ($P<5\times10^{-8}$) has recently increased to 136. However, these genome-wide association studies (GWAS) have been conducted primarily in populations of European ancestry. This thesis aims to: 1) investigate whether these BMI SNPs are also associated with BMI in other ethnicities (South Asian, East Asian, African, Latino American and Native American) using a multi-ethnic prospective EpiDREAM cohort study; 2) explore the parental and child genetic contributions to obesity-related traits in children from birth to 5 years in the FAMILY cohort; 3) examine the maternal and child genetic contribution of BMI SNPs to maternal gestational weight gain (GWG) and postpartum weight retention in the FAMILY cohort.

The major findings are: 1) most BMI susceptibility genes identified in Europeans are also associated with BMI in other five ethnicities. The effects of some SNPs and BMI genetic risk score (GRS) were modified by ethnicity; 2) SNPs contributing to adult BMI start to exert their effect at birth and in early childhood. Parent-of-origin effects may occur in a limited subset of obesity predisposing SNPs; and 3) there is no association between maternal and child GRS and GWG. But there is a genetic link between pre-pregnancy BMI variation and offspring birth weight and maternal postpartum weight retention. Taken together, these findings indicate that GWAS of specific ethnic group, children, birth weight and GWG are necessary to look for novel variants and alternative pathways influencing the development of obesity.

ACKNOWLEDGEMENTS

I would like to express my deepest gratitude to Dr. David Meyre for his comprehensive guidance, numerous discussions and constant support during the course of my training. I am grateful for his profound knowledge and insight that have brought me to the field of genetic epidemiology in obesity and to where I am today. This work would not have been possible without his tireless dedication.

I am deeply indebted to my supervisory committee members, Drs Sonia Anand, Parminder Raina and Hendrik Poinar. Dr. Anand's insights and expertise helped me immensely in understanding the concepts in child study and methodology. Dr. Raina's clear thoughts in statistics have both impressed and helped me a great deal. Dr. Poinar's lectures, TED talk and his comments always make science more interesting and fun. Without their contributions, this thesis could not be completed. A special thank-you also goes out to Dr. Guillaume Pare. His prompt assistance and efficient problem-solving helped finish my independent study smoothly and successfully. His knowledge and attitude toward science always encourage me and will benefit my future academic career.

I want to sincerely thank all members of the EpiDream and the FAMILY studies for their incredible feedback and genuine encouragement during the process of my manuscript preparation. I also want to extend my special thank-you to Jingyuan Chen, Karleen.Schulze, Nora.Abdalla, Senay Asma and Dipika Desai who helped me with the data extraction. I am particularly grateful to Dr. Sebastien Robiou du Pont, Dr. Arkan Abdai, Akram Alyass and other colleagues who helped me immensely. I am extremely grateful to Hudson Reddon for his invaluable comments and excellent editorial help, especially at the final stage of this thesis.

Also, I would like to deliver sincere thanks to my friends for their initial encouragement and continued support.

Finally, special thanks go to my husband, Tao, for his encouragement, love, and support over the years, and also to my lovely son, Andy, for his understanding and encouragement, even during the most difficult times. I am also grateful for the blessing and support from my parents, my brother, my sister and my nephews. I dedicate this thesis to them.

TABLE OF CONTENTS

Page

LIST OF FIGURES AND TABLES

FIGURES

CHAPTER II

CHPATER III

CHAPTER IV

CHAPTER VI:

CHAPTER VII:

TABLES

CHAPTER II

CHPATER III

CHAPTER IV

CHAPTER VI:

CHAPTER VII

LIST OF ABBREVIATIONS AND SYMBOLS

ACMG - American College of Medical Genetics and Genomics
AD - Alzheimer Disease
AUC - Under the Curve
ASCO - American Society of Clinical Oncology
BIA - Bioelectric Impedance Analysis
BMI - Body Mass Index
CCHS - Canadian Community Health Survey
CDC - Centers for Disease Control
CDCV - Common Disease-Common Variant
CDRV - Common Disease-Rare Variant
CNV - Copy Number Variant
CRT - Cyclic Reversible Termination
CT - Computerized Tomography
DEXA - Dual Energy X-ray Absorptiometry
DNA - DeoxyriboNucleic Acid
DTC - Direct-to-Consumer
EpiDREAM - epidemiological arm of the Diabetes Reduction Assessment with Ramipril and Rosiglitazone Medication (DREAM) study
ER - Estrogen-Receptor
FAMILY - the Family Atherosclerosis Monitoring In earLY life
FDA - Food and Drug Administration
FDR - False-Discovery Rate
FE - Fixed-Effects
GEWIS - Genome Wide Interaction Study
GIANT - Genetic Investigation of ANthropometric Traits
GWAS - Genome Wide Association Study
GWG - Gestational Weight Gain
HER2 - Human Epidermal Growth Factor Receptor 2
HWE - Hardy-Weinberg equilibrium
HR - Hazard Ratio
IADPSG - International Association of Diabetes and Pregnancy Study Groups
IBD - Identical By Descent
IOM - Institute of Medicine
LD - Linkage Disequilibrium
LOD - Logarithm of the Odds
MAF – Minor Allele Frequency
MODY - Maturity-Onset Diabetes of the Young
MRI - Magnetic Resonance Imaging
NGS - Next Generation Sequencing
NHANES - National Health and Nutrition Examination Survey
NIH - National Institute of Health
NRI - Net Reclassification Index
OGTT - Oral Glucose Tolerance Test

OMIM - Mendelian Inheritance in Man
OR - Odds Ratio
PCA - Principle Component Analysis
PR - Progesterone-Receptor
PRAMS - Pregnancy Risk Assessment Monitoring System
QC - Quality Control
RCT - Randomized Controlled Trial
RE - Random-Effects
RFLP - Restriction Fragment Length Polymorphism
RNA - Ribonucleic Acid
mRNA - messenger Ribonucleic Acid
tRNA - transfer RNA
ROC - Receiver Operating Characteristic
RR - Relative Risk
SCID - Severe Combined Immunodeficiency Disease
SNA - Single Nucleotide Addition
SNP - Single Nucleotide Polymorphism
STR - Short Tandem Repeats
T2D – Type 2 Diabetes
UTRs - Untranslated Regions
VNTR – variable number of tandem repeats
VP - Variance Prioritization
WC – Waist Circumference
WES - Whole-Exome Sequencing
WGS - Whole-Genome Sequencing
WHR – Waist-to-Hip Ratio
WTCCC - Wellcome Trust Case Control Consortium

DECLARATION OF ACADEMIC ACHIEVEMENT

I was the main contributor and the first author for all studies included in this thesis.

PREFACE AND OUTLINE

This thesis represents an investigation into the genetics of human obesity in the post-genome wide association study era. Currently, 136 single nucleotide polymorphisms (SNPs) have been identified to be associated with adult body mass index (BMI) or obesity at genome-wide significance level ($P<5\times10^{-8}$). The first objective of the thesis is to examine whether these SNPs most discovered in European ancestry are also associated with adult BMI in other ethnic populations. The second objective is to determine whether these SNPs contribute to the variation of obesity related traits at birth and early childhood. The third objective is to determine whether these SNPs account for the variation of maternal gestational weight gain and postpartum weight retention. These individual studies will follow a version of sandwich thesis and comprise individual chapters. All these studies are written into manuscripts for publication.

In addition, in order to expand my knowledge of genetic epidemiology at the beginning of my PhD study and under the supervision of Dr. Meyre, I completed an extensive review of literature and wrote a series of three articles that summarized the concepts of molecular genetics, genetic epidemiology and its applications which will be included in the supplementary as separate chapters.

Chapter I: A literature review consisting of the following aspects:

(1) Prevalence and health complications of obesity in children, adults and pregnant women.

(2) Environmental risk factors including individual-level lifestyle factors (diet, physical activity, behaviors and other lifestyle components such as sleep deprivation, socioeconomic status, smoking, depression, marital status, employment situation and

parity), perinatal, early life and intergenerational factors (such as maternal obesity/ excessive gestational weight gain/gestational diabetes mellitus and birth weight).

(3) Genetic risk factors including syndromic /non-syndromic obesity, oligogenic obesity and polygenic obesity.

(4) Gene and environment interactions and obesity

Chapter II will examine whether 23 SNPs (individually or collectively as a genetic risk score (GRS)) associated with adults BMI in Europeans are associated with adult BMI in other ethnicities using multi-ethnic prospective EpiDREAM cohort study (including South Asian, East Asian, African, Latino American and Native American) and test for interaction between SNPs/GRS and ethnicity.

Chapter III will investigate the parental and child contribution of 83 SNPs to obesity-related traits in children from birth to 5 years old in the FAMILY cohort. The parent-of-origin effects of each SNP are also explored. This manuscript has been submitted to *Obesity*.

Chapter IV will examine the associations between maternal prepregnancy BMI and gestational weight gain (GWG) and obesity-related traits in both mothers and offspring in the FAMILY birth cohort. The maternal and offspring genetic contributions of 83 BMI susceptibility variants to GWG and postpartum weight retention are further tested. This manuscript has been submitted to *Obesity*.

Chapter V will summarize the major contributions of this thesis to the current knowledge in the relevant fields. Future direction of research will be addressed.

Supplementary Chapter VI will focus on the fundamental concepts of molecular genetics. This manuscript has been published in *Current Psychiatry Reviews* 2014, 10(2):91-100

Supplementary Chapter VII will focus on the fundamental concepts and methods in genetic epidemiology including the classification of genetic disorders, study designs and their implementation, genetic marker selection, genotyping and sequencing technologies, gene identification strategies, data analyses and data interpretation. This manuscript has been published in *Current Psychiatry Reviews* 2014, 10(2):101-117

Supplementary Chapter VIII will discuss the evolution of personalized medicine and illustrate the most recent success in the fields of Mendelian and complex human diseases. This manuscript has been published in *Current Psychiatry Reviews* 2014, 10(2):118-132

CHAPTER I: FUNDAMENTAL EPIDEMIOLOGICAL METHODOLOGY AND LITERATURE REVIEW ON OBESITY

1. FUNDAMENTAL METHODOLOGY IN EPIDEMIOLOGICAL STUDIES

My thesis focuses on the genetics of human obesity in the post genome-wide association study era, but the fundamental methodology in classic epidemiology studies apply to genetic epidemiology as well. Therefore, I would like to review the relevant knowledge in the study designs, measurements of association, source of biases and confounding factors.

1.1. Study designs

To study genetic association, there are two types of distinctive data: data from unrelated individuals and family data. If unrelated individuals are recruited for the study, it resembles classical epidemiological studies. The choice of a study design depends on the particular research question, time and cost. If ethical consideration permits, in clinical epidemiology studies, the randomized controlled trial (RCT) is considered the gold standard to test the efficacy or effectiveness of an intervention as it minimizes known and unknown confounding via randomization. Compared to that, cross-sectional and case-control study designs are more commonly used in genetic association studies as genes are blindly allocated which avoid the selection bias of the predictor. In a cross-sectional study, outcome and predictive variables (including genotype data) are all measured on a single occasion or within a short period of time. In a case-control study, the outcomes (cases and controls) are chosen first and then the predictors (including genotype data) are collected from these two samples by recall. The strengths of these two approaches are cost and time efficiency. Another advantage of the case-control study is its efficiency for rare diseases. The major weakness of a cross-sectional study is the difficulty of establishing causal relationships. The biggest limitation of a case-control study is its

susceptibility to biases because the cases and controls come from two different samples and the retrospective measurements of predictor variables creates the potential for recall bias. In a prospective cohort study, the predictive variables are measured at the beginning and the outcomes are measured as they develop over follow-up. This is a powerful strategy to assess the causal relationship between the risk factors and the incidence of a disease or a quantitative trait. However, it is more expensive and less efficient for collecting a sufficient number of cases for a rare disease. As a solution, a nested case-control study in a prospective cohort is theoretically ideal for a genetic association study in which disease cases collected during the follow-up are matched to non-disease controls selected from a portion of the entire cohort subjects. This design minimizes the recall bias, selection bias and inadequate/unreliable records of the environmental exposure from a retrospective case-control study, particularly when gene × environment hypotheses are being tested. However, it sacrifices the statistical power due to loss of substantial sample size. Quantitative trait studies (e.g. BMI in GIANT) in large-scale population-based samples have proved to be an efficient approach to identify novel susceptibility loci.[1] However, it requires a larger sample size than case-control studies to reach the same statistical power, and this limitation greatly impacts the cost if expensive technologies are used (e.g. genome-wide DNA arrays).[2] Thus, a case-control study is considered the most powerful and cost-efficient method to perform genetic association studies, if the two study groups are recriuted properly. An enrichment sampling strategy can increase the power of a case-control study.[3] Enrichment strategies may select patients who are more homogenous at baseline, who have a greater likelihood of having a disease-related endpoint event or who are more likely to response to the drug treatment. For example, obesity cases selected from familial forms of childhood and adult extreme obesity, or having an early age of onset have narrow range of BMI, thus decreasing

inter-patient variability. The decreased variability will increase study power.[4] These individuals with more severe phenotype are likely to be enriched for genetic susceptibility in comparison with obese subjects randomly selected in the general population. Bezinous *et al.* use this strategy to discover novel genetic variants rs6232 and rs6234/rs6235 in *PCSK1* associated with obesity.[5] Though the enrichment sampling strategy is advantageous to improve power in genetic association studies, it artificially inflates the relative risk (i.e. winner's curse) and population attributable risk of the associated gene variants. Therefore, population-based follow-up cohort studies will be needed to obtain a reliable estimation of these parameters. The genetic variants rs6232 and rs6234/rs6235 in *PCSK1* were later shown to be associated with obesity and BMI in the general population.[6] This example illustrates that an enrichment sampling strategy is a cost-effective and efficient approach to identify loci associated with a trait of interest.

The family-based design is optimal in specific situations, such as the identification of disease-associated variants subjected to parental imprinting, or in haplotype studies (the reconstruction of the haplotype phase is improved by availability of parental genotypes). It is robust to population stratification bias,[7] but this design requests 50% more participants compared to a case-control study assuming the same power and risk allele frequency.[8]

1.2. Measurements of disease frequency and association

1.2.1. Disease frequency

It is essential to quantify the occurrence of a disease for an epidemiologic investigation and to allow public health decision-makers to allocate the health care resources in a particular community. Prevalence and incidence are the two categories most frequently used for measures of disease frequency. Prevalence quantifies the proportion of individuals in a population who

have the disease or condition at a specific point in time and provides an estimate of the probability (risk) that an individual will be ill at that time point. The formula for calculating the prevalence is

$$P = \frac{\text{number of existing cases of a disease}}{\text{total population}} \text{ at a given point in time}$$

In contrast to prevalence, incidence quantifies the number of new events or cases of diseases that develop in a population of individuals at risk during a specified time interval and provides an estimate of the probability (risk) that an individual will develop a disease during a specified period of time. The formula for calculating the incidence is

$$I = \frac{\text{number of new cases of a disease during a given period of time}}{\text{total population at risk}}$$

1.2.2. Measurements of disease association

When the outcome is binary (dichotomous) in which every participant is one of the two possibilities, for example, with or without disease, the data are often presented in the form of a two-by-two table, also called contingency table. The most commonly used effect measures of association in clinical trials with dichotomous data include the relative risk (RR) (also called the relative risk, the odds ratio (OR), the risk difference (RD) (also called the absolute risk reduction) and the number needed to treat (NNT).

		Disease		
		YES	NO	Total
Exposure	YES	a	b	a+b
	NO	c	d	c+d
	Total	a+c	b+d	a+b+c+d

Data in a contingency table from a case-control or cohort study

In a cohort study, relative risk estimates the magnitude of an association between the exposure and disease and is the ratio of likelihood of developing the disease in the exposed group relative to that in the control group, also called relative ratio. The formula for calculating RR is:

$$RR = \frac{a/(a+b)}{c/(c+d)}$$

RR is easy to compute and interpret, but the disadvantage is that the same value of RR may represent very different clinical scenarios. For example, a RR of 0.75 could be a clinically important reduction in hypertension from 60% to 45% or a modest and less clinically important reduction in hypertension from 4% to 3%. Therefore, it is suggested to report RR for summarizing the evidence and absolute measures for an actual clinical or public health situation.[9]

In a case-control study in which the participants are selected on the basis of disease status, it is not possible to calculate the risk of developing the disease given the presence or absence of exposure. Therefore, the estimate of RR in a cohort cannot be applied to the rationale in a case-control study. However, odds ratio estimates the ratio of the odds of exposure among the cases to that among the controls. It also represents the ratio of the odds of an outcome will occur given a particular exposure to the odds of the outcome occurring in the absence of that exposure. The formula for calculating OR is:

$$OR = \frac{a/c}{b/d} = \frac{a/b}{c/d} = \frac{ad}{bc}$$

If a disease is rare, the OR provides a good estimate of RR.[10, 11] The OR is not only commonly reported in case-control studies, it also reported in cohort studies, cross-sectional studies, or clinical trials.[12] The OR is the only measure of association estimated from a logistic model, without special assumption and requirement of study design.[12]

As discussed above, the rationale for RR and OR is different. When events are common in studies, the RR and OR differ. Because OR is more difficult to understand and interpret, it is unfortunately common in the literature to misinterpret an OR as a RR. If the OR is misinterpreted as a RR, it may be misleading because OR always overestimates the effect compared to a RR.[13, 14] When the OR is less than 1, it is smaller than the RR; and when it is greater than 1, it is greater than the RR.

Absolute risk reduction (ARR) estimates the absolute effect of the exposure or the excess risk of disease in those exposed compared to those not exposed. The formula for calculating ARR is:

$$ARR = \frac{a}{a+b} - \frac{c}{c+d}$$

The advantage of using ARR is that it is easy to compute and interpret. But it is worth noting that the difference in risk of fixed size may have greater importance when the values are close to 0 or 1 than those near the middle of the range. For example, the difference between 0.01 and 0.05 should get more attention than the difference between 0.45 and 0.49, when severe side effects are considered.[12] The clinical importance of a ARR may depend on the underlying risk of events and the consequences of the events when interpreting.[15]

Number needed to treat (NTT) is defined mathematically as the reciprocal of the ARR. The interpretation of NTT is dependent on whether the exposure-disease relationship is causal. If the causal-effect relationship exists, the NNT means the number of patients needed to be treated to prevent one case. Because it is an important measurement addressing both statistical and clinical significance and easy to interpret, NTT is often used to summarize the results of clinical trials.[16] The value of NTT is a function of the disease, the exposure and the outcome, and therefore, it is only appropriate to compare NNTs directly when the same disease with the same severity and the same outcome are compared.[17]

Another related measure, based on the ARR, is the population attributable risk (PAR). The formula for calculating PAR is:

$$PAR = (ARR)(P_e)$$

P_e is the proportion of exposure individuals in the population

It estimates the proportion of cases that are attributable to the exposure. In other words, it indicates the number (or proportion) of cases that would not occur in a population if the exposure were removed.

Overall, the choice of measurement depends on the types of study design. In retrospective and cross-sectional studies, in which the aim is to look at the association rather than differences, the OR is suggested. OR is also used in case-control studies in which the RR cannot be estimated. RR is recommended in cohort studies because it is clinically meaningful and easy to interpret.

1.2.3. Measurements of association with quantitative traits

Since the qualitative characteristics of disorders (with or without a disorder) are appeal to clinicians and patients, many genome-wide association studies (GWAS) are cases-control studies

that focus on binary traits and typically compare allele frequencies for cases and controls.[1] But GWAS indicate that many genes with small effects affect these disorders, indicating that common diseases are quantitative traits and their genetic liability is distributed quantitatively rather than qualitatively.[18] Identifying the genetic associations with quantitative traits that are related to a disease is important to understand the quantitative mechanisms underlying the disease. For example, a wave of GWAS have focused on quantitative traits related to obesity, including body mass index (BMI),[1] waist circumference,[19] waist-to-hip ratio (WHR) [20] and fat mass.[21]

Linear regression models are generally used to test the association between the predictors (including genotype data) and quantitative traits. The derived coefficients indicate the increase (positive association) or decrease (negative association) on the level of quantitative traits for every unit increase of the predictor or each additional risk allele.

1.3. Biases in case-control studies

Biases can occur at different stages of a research study: specification and selection of the study sample, execution of the protocol, data analysis, interpretation of the results and publication.[22] Since case-control studies are commonly used in observational and genetic association studies and the biggest weakness of case-control studies is their susceptibility to a variety of biases, we will address the sources of biases and how to control them below in epidemiological and genetic association studies.

1.3.1. Selection bias

This error is introduced when the study population does not represent the target population. It may arise with inclusion and exclusion criteria of the eligible population at design stage and non-random sample recruitment process. In practice, the cases are usually selected by

the investigators from the accessible sources of subjects. The sample of cases may not represent the target population. The general goal to control such a sampling bias is to select controls from a population having similar risk for the disease to the selected cases.[23] One approach is to sample the cases and controls from the same population, for example from hospital- or clinic-based population. The second is to match the controls with the selected cases for potential risk factors such as age, sex, and other additional variables (e.g. level of physical activity for obesity study). Because ethnicity is an obvious risk factor for a spurious genetic association, ideally subjects in two groups should be matched for ethnic and even geographical origin. "Super control" subjects can also be selected (normal-weight subjects with no familial background of obesity or extremely lean phenotypes). The third strategy is to use two or more control groups to corroborate a real association.[24] For example, if the cases are recruited from an emergency room, and different controls are used (emergency room controls the same as the cases, inpatient controls in the same hospital, and community control from the same city), the consistently strong associations derived from different controls support a true association in the population. The fourth way is to choose population-based cases and controls to manage the sample selection bias as disease registries are becoming increasingly available.

Population stratification refers to one type of sampling biases in genetic association studies. The risk allele frequency of a genetic variant may vary among different ethnic backgrounds or even the geographical location. When cases and controls come from multiple ethnic or geographic groups, the risk allele frequency may be associated with the disease, thus leading to false-positive associations.[25] For example, some obesity predisposing variants show highly varying levels of inter-ethnic or inter-geographic allele frequency variation, possibly due to positive diversifying selection (e.g. *ENPP1* rs1044498, *FTO* rs9939609 or *LCT* rs4988235).[26-

[28] In addition, the prevalence of obesity varies across ethnic backgrounds in a country.[29] Genome-wide association studies (GWAS) for obesity-related traits have reported modest but real evidence of population stratification.[30, 31]

1.3.2. Differential measurement bias

Differential measurement bias in case-control studies is due to measurement error caused by the retrospective approach to measure the dependent and predictor variables. It also refers to information bias and misclassification.[32] When the measurement tool used to detect the outcome or to measure the exposures is not perfect, the exposed or diseased individuals can be misclassified as non-exposed/non-diseased and vice versa.[33] In addition to measurement tool, observer/interviewer bias, recall bias and reporting bias (e.g. social desirability) commonly produce misclassification.[32] There are two types of misclassification: non-differential misclassification and differential misclassification.

Differential misclassification bias: The different error rates or probability of being classified in case and control groups will lead to differential misclassification. For example, Individuals with obesity are more likely to report lower weight and individuals with underweight are more likely to report higher weight. The estimates of such misclassification may be biased in either direction, underestimate or overestimate the true value.[33]

Non-differential misclassification bias: It is when all the variables (exposure, outcome, or covariate) have the same error rate or probability of being misclassified in case and control groups. For binary outcome, the estimation is biased to underestimate the true value.[33]

Some strategies have been used to control biased measurements including standardizing the measurement methods, training the observer, refining the instrument, automating the instrument, calibrating the instrument. In addition to these, blinding is a good strategy to reduce

observer/interviewer bias, recall bias and reporting bias. In observational study, it is hard to blind the participants and the observer the intervention they receive or give, but they are more frequently blind to the main hypothesis.

Genotype misclassification refers to a particular measurement error in genetic association studies. It is frequent in genetic association studies, leading to non-differential genotype misclassifications (same probability of being misclassified for all study subjects) or differential genotype misclassifications (varying probability of being misclassified according to the study groups).[34] A 1% increase in genotyping errors will require a sample size increase of 2-8% to keep the same type I and II errors.[35] Genotyping errors from batch to batch, laboratory to laboratory or preferential rejection of particular genotypes (usually heterozygotes) can result in differential genotype misclassifications and significant differences between case and control groups, leading to false-positive associations. A recent large-scale family-based study using TaqMan technology excluded a role of VNTR *INS* polymorphism in childhood obesity despite previous positive association using PCR-based restriction fragment length polymorphism (RFLP).[36, 37] As the reproducibility of RFLP genotyping data has been questioned, this method being highly subjective, the authors suggested that the lack of replication may be a result of previous genotype misclassification from the RFLP method.

In summary, a nested case-control study can increase the accuracy of the measurements from the study design stage to execution of the protocols and to the data collection, thus leading to a more reliable association.

1.4 Confounding

Confounding occurs in a situation in which a measure of association between exposure and outcome is distorted by the presence of another variable. This extraneous variable is a

confounding factor and can results in an inaccurate association. A confounding factor meets three criteria: 1) it is associated with the disease, independent of the exposure; 2) it is associated with the exposure; 3) It is not in the causal pathway between exposure and disease. Because it can positively and negatively influence an association between exposure and disease, confounding factors need to be taken into account and adjusted in the analytic model. As mentioned above, any inaccurate measurement of confounding factors will also lead to misclassification, thus any strategies used to reduce biases in exposure and outcome variables apply to measure confounding factors as well.

Unlike classic epidemiology studies, genotype-phenotype association studies are less likely to be confounded by other covariates because these factors usually do not alter genotype. If a covariate is associated with the phenotype independently of the genetic variant, adjustment of this covariate may increase the precision of the association.[38] For example, GWAS of height adjusted BMI identified a functional variate in *ADCY3* in children.[39] If a covariate is associated with both the genetic variant and the phenotype, adjustment for this covariate will remove confounding resulting from this covariate.[38] For example, rs9939609 in *FTO* was associated with type 2 diabetes. This association was abolished by adjustment for BMI, which indicated that the association was mediated through BMI.[40]

2. PREVALENCE AND COMPLICATIONS OF OBESITY

Obesity is a chronic disease that is defined as the condition of excess body fat and is associated with impaired health according to the World Health Organization (WHO).[41] Several methods are routinely used to measure body fat, from basic body measurements to high-tech instrument scan, including body mass index (BMI), waist circumference (WC), waist-to-hip ratio (WHR), skinfold thickness, bioelectric impedance analysis (BIA), dual energy X-ray

absorptiometry (DEXA) and computerized tomography (CT) and magnetic resonance imaging (MRI).[42] For practical purpose, BMI (weight in kilograms divided by height in meters squared) is commonly used to classify overweight and obesity in adults and children because it is closely correlated with body fat and obesity-related consequences.[41, 43] The WHO currently uses BMI cutoffs of 25 and 30 kg/m^2 to classify overweight and obesity in adults, respectively.[41] Given that the risk of adverse health effects increases with higher levels of BMI, obesity is further divided into three categories (class I-moderately obese: BMI of 30.0–34.9 kg/m^2; class II-severely obese: BMI of 35.0–39.9 kg/m^2; and class III-morbidly obese: BMI of 40.0 kg/m^2 or greater). BMI in childhood changes substantially with age and the criteria of overweight and obesity in children and adolescents differ across epidemiological studies.[44, 45] WHO defines obesity as BMI greater than 3 standard deviations above the WHO child growth standard median and overweight as BMI greater than 2 standard deviations above the median in children from birth to age 5.[46] In children from age 5 to 19, obesity is defined as BMI greater than 2 standard deviations above WHO reference 2007 growth standard median and overweight is defined as BMI greater than 1 standard deviation.[47] The Centers for Disease Control and Prevention recommends the age- and gender-specific BMI 85th and 95th percentiles as cut-offs for overweight and obesity for children aged 2-19 years, using Centers for Disease Control and Prevention Growth charts in 2000 as the reference.[48, 49] In children from birth to age 2, the CDC uses a modified version of the WHO criteria.[50] The 97th percentile of BMI is used for the definition of childhood obesity published in 1995 by the European Childhood Obesity Group (ECOG).[45] Obesity has more than doubled worldwide since 1980 and it has become the focus of public health due to a wide range of serious complications.[29]

2.1. Obesity in children

Obesity has been affecting vulnerable populations of children and adolescents at an alarming rate, faster than adult obesity.[51] The prevalence of overweight and obesity for children aged 5 to 17 years in Canada was 31% in 2012-2013 (19% for overweight and 12% for obesity).[52] The situation in the US was worse in which 31.8% children aged 2 to 19 years were overweight and 16.9% were obese in 2011-2012.[53] Recent studies suggest that the prevalence of childhood obesity is stabilizing in some geographic and ethnic groups;[54-57] however, the children with severe obesity have seen a continuing increase in the last decade.[53] Differentiated prevalence of obesity in different ethnic groups was also observed in children and adolescents in the US, with highest prevalence observed in Hispanic (22.4%) and lowest in Non-Hispanic Asian (8.6%).[57]

Childhood obesity is linked to early puberty, type 1 and type 2 diabetes (T2D), hypertension,[58, 59] obstructive sleep apnea, asthma,[60, 61] poor mental and physical health during childhood,[62-66] as well as adult obesity and the associated comorbidities.[67-69] A population-based survey among grade 5 children in the Canadian province of Nova Scotia with a 3-year follow-up period showed that the total health costs were 21% higher in obese children than normal weight counterparts.[70] It is estimated that the direct incremental lifetime medical cost of an obese child at the age of 10 is $19,000 compared to a child of a normal weight, which corresponds to a total direct medical cost of approximately $14 billion for this age alone in the US.[71]

2.2 Obesity in adults

In 2014, 20.2% of Canadians aged 18 and older were obese (self-reported data from the Canadian Community Health Survey (CCHS), http://www.statcan.gc.ca/eng/help/bb-/info/obesity). The prevalence of adult obesity reached 34.9% in the US in 2011-2012 according to the National Health and Nutrition Examination Survey (NHANES).[29] The differences by sex and age were significant.[29] The trend seems to have leveled off over the past decade, but there

was an increase in prevalence in women aged 60 years and older.[29] The prevalence of extreme obesity (BMI>35) is still on the rise in both Canada and the US.[29, 72, 73] Furthermore, obesity, long held to be a health problem that is restricted to the developed countries, has expanded into low- and middle-income countries with their adaption of Westernized diet and continued decrease in physical activity.[74-76] In 2014, more than 1.9 billion adults were overweight and of these over 600 million were obese worldwide (http://www.who.int/mediacentre-/factsheets/fs311/en/). Although obesity is pandemic, its prevalence varies across countries, ethnicities and even regions within the same country.[57, 77, 78] Some ethnic groups are more prone to obesity than others. In the US, the age-adjusted prevalence of obesity in adults were 47.8% in non-Hispanic Black, 42.0% in Hispanic, 33.4% non-Hispanic White and 10.9% in non-Hispanic Asian in 2011-2012. Non-Hispanic Black had the highest proportion of severe obesity (BMI≥35) of 23.4%.[29]

Excessive weight carries detrimental risks to a wide range of conditions including T2D, cardiovascular diseases, stroke, hypertension, nonalcoholic fatty liver disease, osteoarthritis, gallstones, sleep apnea, asthma, infertility and certain types of cancers (e.g. leukemia, breast, prostate, colon).[79, 80] Obesity is also associated with social and emotional consequences, including loss of self-esteem, depression, psychological disorders, and lower quality of life.[81] The life years lost associated with obesity is 8-13 in whites aged 20-30 years with severe form of obesity (BMI>45)[82] and the life years lost associated with obesity-related diseases is 4.7-5.23 years in non-smoking whites aged 40-49 years when BMI is greater than 40.[83] It is estimated that the direct healthcare costs associated with obesity in Canada are approximately 2-12% of total health expenditures while indirect costs associated with obesity-related co-morbidities account for 37-55%.[84] It was estimated that the medical cost of obesity was $190 billion in the US in

2012.[85] A systematic review showed that 0.7-2.8% of a country's total health care expenditure was spent on obesity-related medical costs worldwide and the medical costs for people who are obese were 30% higher than those of normal weight.[86]

2.3. Obesity in pregnancy

Obesity in pregnancy is usually assessed by a recent prepregnancy BMI (self-reported or measured by research staff) using the same definition as that used in the general population.[87] More than one-third of women at reproductive age (aged 20-44 years) in the US are obese, more than one half of them are overweight or obese, and 8% are extremely obese.[29, 88] Although limited documents about the prevalence of obesity in pregnant women are available, cohort studies in the UK have shown that the rise in obesity among pregnant women has occurred in parallel with the upward trend of obesity in the general population.[89, 90] Data from the Pregnancy Risk Assessment Monitoring System (PRAMS) in nine states in the US during 1993-2003 showed an increase of 70% in obesity at the beginning of pregnancy.[91]

The epidemic obesity in adults seems to be preceded by the increased prevalence of obesity in children, indicating the importance of the fetal period in the development of obesity in offspring.[92] A growing body of evidence supports the fetal origins of obesity.[93, 94] Our study (Chapter III) also adds evidence to this field showing that the obesity susceptibility genetic variants collectively start to act in the fetus with an increase of birth weight. Obesity in pregnancy, in part reflecting maternal and intrauterine nutrition conditions during fetal development, has been linked to maternal and offspring health complications. Women who are obese during pregnancy may experience a variety of short- and long-term health impairments including gestational hypertension, gestational diabetes, preeclampsia, postpartum hemorrhage, cesarean delivery, post-partum weight retention and subsequent obesity.[95, 96] Higher BMI during

pregnancy is also associated with macrosomia, stillbirth, congenital abnormality, childhood and adulthood obesity, as well as related metabolic diseases in offspring.[95-99] Although a specific cost of these conditions has not been estimated, the adverse maternal and offspring health outcomes, and economic implications of obesity during pregnancy, have been recognized as an important burden in healthcare settings.[100, 101]

Gestational weight gain (GWG) is another important characteristic during pregnancy. Excessive gestation weight gain can confer the same health risks in both mothers and children as obesity in pregnancy.[96-98, 102-108] Generally, the associations of GWG with maternal and offspring outcomes are weaker than those of maternal pre-pregnancy BMI.[96, 109] In contrast, inadequate GWG is also a risk factor for infant mortality, but the risk weakens with increasing pre-pregnancy BMI.[110] Based on this well-established clinical epidemiologic evidence and the dramatic changes in the population of women having babies between 1999-2009, the Institute of Medicine (IOM) reexamined the guidelines for weight gain during pregnancy and set new optimal ranges of GWG in 2009 for underweight women of 12.5-18 kg (28-40 lbs), normal weight women of 11.5-16 kg (25-35 lbs), overweight women of 7-11.5 kg (15-25 lbs) and obese women of 5-9 kg (11-20 lbs).[111] In 2002-2003, the Pregnancy Risk Assessment Monitoring System (PRAMS) data showed that underweight and normal women had higher mean GWG than overweight and obese women had. Sixty-three percent of overweight women and 46.3% of obese women had weight gains greater than the recommended ranges. In contrast, 19.5% of underweight women and 38.4% of normal weight women gained excessive weight.[111] Overall, among women in all BMI categories, less than 50% of women gained weight within the recommended ranges.[111]

3. ENVIRONMENTAL RISK FACTORS

The escalation of obesity among people at all the ages and the limited success of prevention and treatment strategies call for an urgent need to understand the determinants of excess body weight in order to manage the burden of childhood obesity and prevent adult obesity more efficiently. It is believed that obesity is a multi-factorial disorder caused by environmental and genetic factors and the interplay between these determinants.

3.1.Childhood obesity risk factors

Studies of overweight and obesity in childhood are of particular importance because overweight and obese children are more likely to become obese adults.[112-115] The odds ratios of childhood obesity associated with obesity in adulthood increase with age, ranging from 1.3 at 1 or 2 years of age to 17.5 at 15-17 years of age after adjustment of parental obesity.[115] Children whose weights are at the upper end of normal weight ranges are also at increased risk of being overweight adults.[116] Strategies to prevent adult obesity emphasize the significance of understanding the early life determinants of childhood obesity.

The regulation of body weight is under extraordinarily precise control in the physiological conditions during child growth. Any factors that tip the energy balance in favor of weight gain will contribute to obesity development in the long-term. For example, an excess of positive energy intake of only 120 kcal (about one serving of a sugar-sweetened soft drink per day) theoretically would produce a 50-kg increase in body mass in 10 years.[117] In addition to sugar-sweetened beverages, other risk factors include dietary fat, diet energy-density, physical inactivity, sedentary behaviors and short sleep durations.[51] Among those risk factors, sugar-sweetened beverages have been consistently shown to be robustly associated with obesity in school-aged children and adolescents.[118, 119] Some new risk factors such as chronic inflammation, anxiety, depression and behavioral problems warrant further investigation in this age group.[119]

Determinants of diet and physical activity will be discussed in the next session. In addition to the individual-level lifestyle factors, other prenatal, perinatal, early life and intergenerational factors are linked to childhood obesity and will be addressed below.

3.1.1. Maternal obesity and birth weight

Birth weight is positively associated with obesity in childhood and adulthood.[120-123] Furthermore, obese pregnant women are more likely to have newborns with increased birth weight,[124, 125] especially those either greater than the 90th centile for gestational age (large for gestational age) or macrosomia (birth weight >4 kg).[95, 96, 126] Maternal obesity is also associated with offspring obesity.[126] Such intergenerational associations may be accounted by genetic transmission, intrauterine environment, shared postnatal environment and lifestyle or epigenetics. It is very difficult to disentangle these factors in human studies.[94] Experimental and animal studies have provided some support for the "fetal overnutrition hypothesis,"[94, 127] which suggests that the intrauterine environment plays an important role in the development of obesity in the offspring.[128, 129] According to this hypothesis, hypernutrition (glucose, free fatty acid and amino acids) in the plasma of obese pregnant women enters the fetus through the placenta, and results in permanent changes in the fetus including increased insulin secretion, appetite control, neuroendocrine functioning and energy metabolism, thus leading to an increased risk of obesity in later life. Some epidemiological evidence has also supported this hypothesis. Paternal BMI has generally not been associated with birth weight, and the paternal-offspring BMI associations tend to emerge later.[124, 130, 131] Several large cohort studies demonstrated that the magnitudes of the maternal-offspring BMI associations were stronger compared to those of the paternal-offspring BMI associations, especially in the early postnatal period.[124, 131, 132] Indirect evidence comes from studies of pre-pregnancy bariatric surgery.[133-135] Kral *et al.* compared the prevalence

of obesity in 172 children who were aged 2 to 18 years and born to 113 obese mother (mean BMI=31±9 kg/m^2) after bariatric surgery with 45 same-age siblings who were born before bariatric surgery (mean BMI=48±8 kg/m^2). They found that after the surgery, the prevalence of obesity in children decreased by 52% and severe obesity by 45.1%, with no risk of underweight.[134]

3.1.2. Maternal excessive gestational weight gain and birth weight

Excessive weight gain during pregnancy has been associated with increased birth weight and the offspring's risk of obesity later in life.[96, 97, 103, 105, 136] These studies were unable to distinguish between the effects of genetic risk factors and shared environments. Ludwig *et al.* examined the association between GWG and birth weight by comparing several pregnancies in the same mother using a large within-family data (513,501 mothers and their 1,164,750 offspring).[104] They found that higher GWG increased birth weight after minimizing the confounding factors, supporting the "fetal overnutrition hypothesis".

3.1.3. Gestational diabetes mellitus and birth weight

Gestational diabetes, which is a characteristic of new onset hyperglycemia during pregnancy, will alter the intrauterine environment through excessive fetal insulin production. Thus, the offspring of mothers with gestational diabetes mellitus have a higher birth weight and a higher percentage of macrosomia.[137] A large cohort study showed that children born to mothers with gestational diabetes had increased birth weight, and also had increased risk of being obese during adolescence. However, this association was not significant after controlling for birth weight, indicating this association in adolescents was mediated by birth weight.[138]

3.1.4. Low birth weight and rapid weight gain

Barker *et al.* have proposed the "thrifty phenotype hypothesis" that intrauterine malnutrition, marked by low birth weight, predisposes individuals to T2D, hypertension, dyslipidemia, and cardiovascular diseases in adult life.[139] A new dimension added to the Barker hypothesis is that low birth weight and rapid weight gain in infancy and early childhood play a more important role in increasing the risk of cardiovascular diseases and their risk factors than low birth weight alone.[140] Maternal malnutrition at important stages of fetal development can disturb central endocrine regulatory systems established in gestation that result in obesity, a result supported by an analysis of Dutch famine cohort.[141] In historical cohorts, several studies showed that low birth weight and rapid weight gain in childhood were consistently associated with an increased risk of childhood and adult obesity and cardiovascular diseases.[120, 121, 142-146] These observations are explained by the mismatch between the intrauterine and subsequent postnatal environments, and epigenetic reprogramming.[147-149] For example, the individuals who experienced periconceptional exposure to the Dutch famine had a lower methylation of the imprinted insulin like growth factor 2 (*IGF2*) sixty years later compared to their unexposed, same-sex siblings, reinforcing that very early development period is important for establishing and maintaining epigenetic marks.[149]

3.1.5. Maternal smoking and low birth weight

Infants born to smoking mothers are more likely to be smaller for gestational age or to have lower birth weight compared to those born to non-smoking mothers.[150-152] A prospective birth cohort in Australia with 3,253 population-based children demonstrated a positive association between maternal smoking during pregnancy and childhood obesity at 14 years of age after adjustment of potential confounding factors (odds ratio=1.4 (95% CI: 1.01-1.94).[153] Rapid postnatal weight gain and epigenetic modification in the gene expression in utero might explain the observed association.[154]

3.2. Adult obesity risk factors

3.2.1. Diet

Overweight and obesity result from a long-term energy imbalance. An energy imbalance occurs when the calories intake is greater than the energy requirements of the body. In the past few decades, the size of food portions has increased dramatically in commercial settings such as restaurants and in homes. Not surprisingly, these portion increases have been accompanied by increased consumption.[155] In 2002, Yang and Nestle reported that the increases in the portion sizes of restaurant foods, grocery products and recipes in cookbooks were paralleled with the rise in the prevalence of obesity in the US.[155] Several studies demonstrated that larger portion of food increases energy intake during an isolated meal. Rolls *et al.* reported that adults were offered four portion sizes of a macaroni and cheese entrees, and energy intake of the largest portion of 1000 g was 30% more than of the smallest portion of 500 g.[156] Interestingly, when unware of the amount of food served in the restaurant setting, the customers who were served 50% more of a pasta dish ate 43% more than those served a standard portion.[157] Further evidence shows that accumulative energy intake is associated with large portions of food over a number of days. An 11-day study showed that increasing portions of all food and beverages by 50% resulted in a mean increase in daily energy intake by 423 calories.[158] Randomized controlled trials (RCTs) further demonstrated that portion-controlled entrees or prepared meal plan could efficiently control weight loss.[159-161] These findings explain the epidemic of obesity in an obesogenic setting where large energy dense food portions are readily available.

Because of the high energy density of fat (9 kcal/g) and the more favorable palatability of high-fat food, increased intake of dietary fat is believed to contribute to obesity. Cross-sectional

studies generally show a positive association between dietary fat and body fatness.[162] However, prospective studies of fat intake in relation to weight gain have produced controversial results.[162] A recent meta-analysis of 33 RCTs suggested that diets lower in total fat are associated with lower body weight.[163] But, this study was criticized based on an important methodological flaw such as inappropriate exclusion criteria. A meta-analysis of 28 mainly short-term trials demonstrated that a 10% reduction in total energy from fat could reduce body weight.[164] Longer-term trials, however, have not corroborated these findings.[165, 166]

Severely restricting carbohydrates has been shown to be an alternative strategy for weight loss. A meta-analysis of five RCTs showed that low-carbohydrate diets were more effective in weight loss than low-fat diets after 6 months. However, the difference disappeared after 12 months.[167] Similar results were observed in the most recent meta-analysis of 53 RCTs.[165] In addition to fat intake and carbohydrates, many other dietary factors including protein, whole grains, fiber, fruits and vegetables, nuts, caffeine, sugar-sweetened beverages, and alcohol have been studied for their associations with obesity.[42] Individual dietary factors may exert moderate effect in weight control, and their effects may accumulate to be significant over time. Current studies have shown that the relative influence of single nutrition factor on body fatness is unclear. Therefore, studies on dietary pattern emerged as a complementary approach to single nutrient analysis. Schulze *et al.* compared weight changes in women aged 26 to 46 years between a Western pattern (high intakes of red and processed meats, refined grains, sweets and desserts, and potatoes) and a Prudent pattern (high intakes of fruits, vegetables, whole grains, fish, poultry, and salad dressing) with a follow-up of 8 years.[168] They found that the prudent pattern effectively prevented weight gain. However, the results from many other studies are inconclusive.[42]

Inconsistent findings from epidemiologic studies are influenced by measurement errors in dietary assessment as well residual confounding. Most clinical trials also have serious limitations, such as short duration, small sample size, and sub-optimal adherence to dietary interventions. In summary, current evidence from meta-analysis provides no convincing recommendation that one nutrition factor or one dietary pattern prevents obesity in the general population.

3.2.2. Physical activity

An energy imbalance also occurs when the calorie intake is not expended by physical activity. In the past several decades, numerous studies have been performed to explore the relationships between physical activity and weight loss or weight maintenance in overweight and obese persons, and the relationship between physical activity and the prevention of weight gain in the general population. In cross-sectional studies, negative associations between physical activity and obesity are consistently reported. In general, high-intensity physical activity was more negatively associated with obesity than moderate- and low-intensity activity.[169-171] For example, Bernstein *et al.* demonstrated a clear dose-response relationship between high-intensity activities and lower risk of being obese among population-based adults, but not in moderate-intensity activities.[169] These cross-sectional studies could not determine the causal relationship between physical activity and obesity.[42, 172] The nature of this relationship may also be bi-directional; that is, being less physically active may be the consequence of carrying too much weight and less physically fit. The effect of physical activity on obesity is usually substantially stronger in cross-sectional studies than prospective studies due to lack of controlling for confounding factors (such as diet). Wareham *et al.* conducted a systematic review of 14 prospective cohort studies and found that more physically active persons had less weight gain in

follow-up periods that ranged from 3-10 years.[172] This association was also seen in different race, sex and age subgroups.[173-176] Many RCTs have also been used to examine the effect of physical activity on the weight loss in overweight and obese persons. Wing conducted a narrative review of RCTs and found that six out ten studies showed significantly greater weight loss in exercise-alone group compared to control. But the effect was moderate which averaged 1-2 kg.[177] Jakicic *et al.* completed an 18-month RCT study to compare the effects of different amount of exercise on long-term weight loss and maintenance in overweight adult women.[178] A dose-response relationship between amount of exercise and long-term weight loss was significant and the average weight loss in individuals who exercised more than 200 min/week, 150-200 min/week, and less than 150 min/week was 13.1±8.0, 8.5±5.8, 3.5±6.5 kg, respectively. Recently, a meta-analysis of RCTs showed that physical activity alone exerted a significant post-partum weight loss at 12 months of delivery.[179]

Epidemiological studies of physical activity and obesity also face the same methodological challenges as the investigations of diet. People who are physically active usually tend to have a healthier lifestyle; therefore the effect of physical activity on the weight loss may be exaggerated without control for diet patterns and additional lifestyle factors.

In addition to diet and physical activity, other lifestyle components influence the development of obesity, including sleep deprivation, excessive sleep, socioeconomic status, smoking, depression, marital status, employment situation, social network and parity.[42, 97, 180, 181]

3.2.3. Combined effect of diet, physical activity and lifestyle

Currently, although there is no consensus in the literatures about the primary cause of the adult obesity, combined physical activity and diet have been accepted as the most likely culprits.[182] Two seminal RCTs have demonstrated that lifestyle intervention (reduced energy

intake and increased physical activity and behavioral modification) can lead to effective long-term weight loss.[183-185] In the Diabetes Prevention Program (DPP), 3,234 prediabetic adults with a mean BMI of 34 kg/m^2 were randomly assigned to a medication of metformin, a lifestyle-modification program (including diet modification (1,200-1,800 kcal/day) and physical activity (>175 min/week) or placebo.[183] The lifestyle intervention aimed to loss ≥7% of their initial body weight and to maintain that weight loss. After 24 weeks, fifty percent of intervention participants met the 7% weight reduction goal. After 2.8 years, participants in the lifestyle intervention group had shown a greater increase in physical activity and greater weight loss (−5.6 kg) than those on metformin (−2.1 kg) or placebo (−0.1 kg).[183] In the second ongoing study Look AHEAD (Action for Health in Diabetes), 5,145 adults with type 2 diabetes who were overweight or obese were randomly assigned to either usual care or one-to-one intensive lifestyle intervention, which was adapted from the DPP and also targeted a weight loss of ≥7% of initial weight.[184] The intervention group, compared to control group, had lost a greater percentage of their initial weight by year 1 (-8.6% *vs* -0.7%), by year 4 (-6.15% *vs* -0.88%) and by year 8 (-4.7% *vs* -2.1%).[185, 186] This intervention did not yield significant reduction in cardiovascular morbidity and mortality but did improve some cardiovascular disease risk factors over a mean follow-up of 9.6 years.[185, 186] These studies indicate that the potential determinants of adult obesity function in concert to create the "obesogenic" environment that is expanding across the world.

4. GENETIC RISK FACTORS

Not everyone exposed to the "obesogenic" environment experiences the same risk of developing obesity. This difference could be due to genetic risk factors that influence individual differences in obesity development and response to weight loss interventions.

4.1. Heritability of obesity

A child with two obese parents is 10 times more likely to be obese compared to a child whose parents are of normal weight.[120] This familial aggregation may be the result of shared environment and genetic components. Heritability measures the proportion of total phenotypic variability explained by genetic variance in a particular population at a specific time. Twin and adoption studies are ideal experimental designs to estimate heritability because of their natural separation of genetic and environmental components.[187] Numerous studies involving twins, adoption and families have reported heritability estimates ranging from 25 to 90% for BMI.[188-191] The significant variability in the heritability of BMI can be explained by difference in population and settings (ethnicity, age, specific environment), study designs (twin, adoption, family) and methodology (self-reported *vs* measured BMI, self-reported *vs* DNA-based determination of zygosity, different analytic models).[192] The genetic contribution to BMI varies with age. It increases with age before young adulthood and then decreases with age in adult life.[192, 193] Cross-sectional studies reported that heritability was low for birthweight (40%), moderate at age 4 (60%), high at age 10 (77%) and highest in adolescents (70-90%).[194-197] A longitudinal study with repeated measurements of BMI in twins corroborated this trend.[193] High heritability estimates have also been reported in other obesity-related traits with 65% to 75% for fat mass and percentage of body fat, 46% to 90% for waist circumference, and 48% to 69% for skinfolds.[198, 199] The heritability estimates from previous family-based studies may be inflated due to shared environmental exposures.[192] More recently, Yang *et al.* developed a method (GREML-LDMS) to estimate heritability for BMI in general population using whole-genome sequencing data.[200] They reported that heritability is likely to be 30-40% for BMI. Taken together, obesity and related traits are highly heritable, thus driving intensive efforts to identify obesity susceptibility genes in the past two decades.

4.2. Genetics of monogenic obesity

Monogenic obesity (also named Mendelian obesity) refers to severe forms of obesity caused by a single gene mutation and follows a Mendelian pattern of inheritance, typically beginning in childhood.[190] There are syndromic and non-syndromic varieties of monogenic obesity.

4.2.1. Syndromic monogenic obesity

Syndromic obesity is rare and co-occurs in the clinical context with other characteristics of mental retardation, dysmorphic features and organ-specific abnormalities.[201] Over 30 syndromic forms of obesity have been identified.[202] The genetic basis underlying some of these syndromes have been elucidated and provided insights into the pathogenesis of the derangements of energy homeostasis. Severe or morbid obesity developed in Prader Willi syndrome (PWS) is the most common syndromic form of obesity. PWS is the consequence of loss of expression of paternal genes on the imprinted region of 15q11-13 (65-75% of the cases), or maternal uniparental disomy 15 (20-30%), or an imprinting defect (1-3).[203, 204] Bardet-Biedl syndrome (BBS) is the first reported obesity syndrome but is rare.[205] To date, 19 BBS genes that cause BBS have been identified. The severity and age of onset of disease depend on the number of mutant alleles. Although genetic mutations in many of the syndromic obesity are waiting to be discovered, the known forms of syndromic obesity seem to have specific effects on food intake.

4.2.2. Non-syndromic monogenic obesity

Leptin deficiency was the first cause of monogenic obesity to be demonstrated in a human patient.[206] In 1997, two severely, intractably obese Pakistani children of a highly consanguineous family were identified to have a frame-shift mutation in the gene *LEP* encoding

leptin by candidate gene approach. Four parents were heterozygotes and, of the four siblings, one was a heterozygote and three were wild-type homozygotes. The phenotype and genotype from this family indicated that this disorder complied with an autosomal recessive inheritance pattern. Currently, eight mutations in leptin gene have been confirmed to cause extreme obesity in infancy.[207] Homozygous / heterozygous compound loss of function mutations in other four genes (leptin receptor (*LEPR*), proopiomelanocortin (*POMC*), prohormone convertase 1 (*PCSK1*), melanocortin 4 receptor (*MC4R*)) have been reported.[208-211] Complete mutations of these genes are fully penetrant and very rare, and follow recessive inheritance pattern in their families. The patients with complete mutations of these genes present early onset morbid obesity, hyperphagia and reduced energy expenditure, and also have additional clinical features specific to their mutation. The studies of these recessive forms of monogenic obesity in human and in animals have demonstrated that the leptin-melanocortin pathway plays a critical role in the regulation of body weight.[212]

In several other severely obese children, heterozygous carriers of mutations in brain-derived neurotrophic factor (*BDNF*), or its receptor neurotrophic tyrosine kinase receptor type 2 (*NTRK2*) (encoding TrkB), or *SIM1* also present early-onset hyperphagic obesity.[213-215] Because complete deficiency in BDNF or SIM1 is lethal and complete deficiency in TrkB dramatically reduces life span based on studies in mice, there are no human cases reported with complete deficiency in these proteins.[213-215] BDNF and TrkB participate in proliferation, survival, and differentiation of neurons during fetus development and post-natal synaptic plasticity in the central nervous system, especially in hypothalamic neurons. SIM plays a major role in neuronal differentiation in the hypothalamus, along the leptin-melanocortin pathway to control food intake.

In a large cohort study of 300 patients with severe early-onset obesity and 500 controls, five probands were identified with either a frameshift or a missense mutation in Src homology 2 B adapter protein 1 (*SH2B1*).[216] Loss-of-function mutations in *SH2B1* were associated with early-onset obesity, insulin resistance, reduced final height and a range of behavioral abnormalities. SH2B1 modulates signaling by a variety of ligands that bind to receptor tyrosine kinases or JAK-associated cytokine receptors, including leptin, insulin, growth hormone (GH), and nerve growth factor (NGF).[216] All of these mutations were associated with loss of function in the GH/NGF-mediated signaling. Intriguingly, only the frameshift mutation impaired leptin signaling. These findings provided insights into the alternative pathways that influence the weight control.

Recently, whole-exome sequencing technology has helped to identify novel mutations associated with monogenic obesity. A homozygous frameshift mutation in Tubby bipartite transcription factor (*TUB)* occurred in an eleven years old boy from a consanguineous UK family and resulted in retinal dystrophy and early-onset obesity.[217] Recessively inherited mutations in *TUB* in mice were demonstrated to cause retinal degeneration, obesity and insulin resistance.[218] This study provided evidence that TUB was important in energy homeostasis in humans. Another study identified several rare mutations in kinase suppressor of Ras 2 (*KSR2*) in 2,101 severe early-onset obesity and 1,536 controls.[219] KSR2 acts as a positive regulator of the Ras-Raf-MEK signaling pathway by acting as scaffolding proteins and plays a role in energy homeostasis.[220] Loss-of-function mutations in *KSR2* lead to hyperphagia in childhood, low heart rate, reduced basal metabolic rate and severe insulin resistance.[219] This study suggests that Ras-Raf-MEK signaling pathway may be a novel therapeutic target for obesity and type 2 diabetes.

The advent of a new generation of genotyping technology extends the single nucleotide polymorphisms (SNPs) to copy number variants (CNVs) or structural variants in the genotype platforms. Rare deletions in the region 16p11.2 have been reported in about 0.5-0.7% of individuals with severe obesity in two independent studies.[221, 222] The examination of rare CNVs offers a new avenue to explore the susceptibility variants of obesity.

4.3. Genetics of oligogenic obesity

Some individuals carrying heterozygous deleterious coding mutations present less extreme or incompletely penetrant forms of obesity. This phenomenon explains a substantial number of severe obesity cases in human.[212] Heterozygous loss-of-function mutations have been observed in some monogenic genes such as *MC4R*, *POMC*, *LEP*, *LEPR* and *PCSK1*.[211, 223-226] Heterozygous carriers of *MC4R* loss-of-function mutations are reported to consume three-times more food than their unaffected siblings and prefer high-fat food.[211, 227] Partial deficiency of MC4R or POMC is associated with an incomplete penetrant form of Mendelian obesity, whereas partial LEP or LEPR deficiency is associated with higher body fat mass. However, heterozygous carriers of p.Arg80* mutations show a dominant form of Mendelian obesity in a large French pedigree.[228]

4.4. Genetics of polygenic obesity

Linkage analyses and candidate gene studies are hypothesis-driven approaches and have been widely used in genetic association studies before the advent of genome-wide association study (GWAS). Linkage analysis aims to map the location of a disease-causing locus[229] and candidate gene approach tests selected genes based on preceding knowledge of their potential role on the trait of interest from *in vivo*, *in vitro* or *in silico* studies in animals or humans.[230-232]

Using linkage analysis, K121Q in *ENPP1* was identified to be associated with childhood and adult obesity in European populations,[233, 234] and R125W in *TBC1D1* was associated with severe obesity in females in US and French populations.[235, 236] After the region was found to be linked to obesity through linkage analysis, *PCSK1* was later confirmed to be associated with severe obesity and was also associated with BMI and obesity in general population.[5, 6, 237] Candidate gene approach has been successful in the discoveries of the Val66Met polymorphism in *BDNF* and a non-synonymous variant p.R270H in *GPR120,* which were later replicated and confirmed to be associated with BMI.[238, 239] The biological function of *GPR120* suggests that lactose consumption is significantly associated with obesity.[240]

Hypothesis-free genome wide association studies (GWAS) exhaustively test the genotype-phenotype associations across up to several million genetic markers and currently represent the most efficient way to identify common variants (minor allele frequency (MAF)> 1%) associated with complex diseases.[241] Along with the increase in the number of novel genetic variants validated by the 1000 Genomes Project and advanced high-throughput technology, rare variants and CNVs in addition to more common variants are available in a single genotyping array.[242] This ensures an enhanced capability to detect a novel genotype-phenotype association. Until recently, a total of 136 SNPs are associated with BMI or binary obesity at a genome-wide significance level ($P<5\times10^{-8}$). The first BMI locus rs9939609 in the fat mass and obesity-associated gene (*FTO*) was discovered in 2007.[40] Several other common variants that have the strongest associations with BMI or obesity lie in introns 1 and 2 of *FTO,* within a highly linkage disequilibrium block with rs9939609 in Europeans.[243-245] *FTO* has now unequivocally been replicated in other populations and ethnicities, and in both adults and children.[246] Following the discovery of *FTO, MC4R* was identified in 2008, loci in or near *TMEM18*, *SH2B1*, *KCTD15*,

MTCH2, *NEGR1*, *BDNF*, *SEC16B*, *GNPDA2* and *ETV5* in 2009 and additional 18 loci in 2010.[1] The most recent meta-analysis of GWAS and Metabochip studies in 339,226 individuals (majority Europeans) has detected 56 novel genetic loci associated with adult BMI.[247] Furthermore, the genetic variants in almost all the genes associated with monogenic obesity have been identified by GWAS with contributions to common forms of obesity as well.

Several genes are expressed in the hypothalamus and genetically driven hyperphagic behavior plays an important role in appetite regulation and energy homeostasis, as demonstrated in monogenic obesity.[1, 247] Locke *et al.*'s study applied novel bioinformatics tools to explore the potential biological pathways involved in BMI regulations. Their findings provided strong support for the role of the central nervous system in obesity susceptibility and also indicated new pathways including synaptic function, glutamate signaling, insulin secretion, energy metabolism, lipid biology and adipogenesis.[247] In addition to the brain, other organs, including the liver, pancreas, adipose tissue, skeletal muscle, and intestine, take part in the pathogenesis of obesity. Evidence of genetic associations, in combination of animal and human studies, has proved that the liver plays a crucial role in the development of obesity. For example, several variants are associated with hepatic lipid and cholesterol metabolism (*NPC1*, *HMGCR*, *NR1H3*, *CYP27A1*),[248-251] lipoprotein transport (*APOC1, APOE*),[252] or glycogen storage (*GBE1*).[253] Other genetic variants have been reported to be associated with dysfunction in β-cells and insulin secretion (*GIPR*, *CDKAL1*, *TCF7L2*)[1, 254], fat distribution (*FTO*),[255] and skeletal muscle (*TBC1D1*).[256] However, the exact molecular mechanisms responsible for these potential pathways remain largely unknown since many of the associated loci are not causal and are not located in protein coding regions, which indicate that numerous gene regulatory studies are needed. For example, although some progress has been achieved in the identification of a

mechanistic basis for the association between the *FTO* locus and obesity in humans, the relevant cell types and target genes remain unresolved, and the causal variant remains uncharacterized.[247, 255, 257]

From GWAS to multiple-stage GWAS to GWAS meta-analysis, all the loci currently identified account for 2.7% of the variation in BMI.[247] Excitingly, Yang *et al.*'s recent study suggests that as much as 27% of BMI variation in 44,126 unrelated individuals can be explained by common genetic variation using ~17 million imputed variants.[200] GWAS use a stringent threshold of $P<5\times10^{-8}$ to accept a genotype-phenotype association, which may exclude many potential obesity susceptibility variants. To fill in the gap between higher heritability of BMI and small portion of BMI variation explained by identified SNPs, strategies to search for this missing heritability include rare variant association, structure variants, CNVs, ethnic-specific variants, gene-environment and gene-gene interactions, and methodological innovations.[258]

4.5. Generalizability of identified adult SNPs across different ethnicities

GWAS have been conducted primarily in populations of Northern European ancestry. More recently, GWAS for BMI have been conducted in East Asian and African populations.[259-262] The BMI-associated loci identified by GWAS display a large overlap in European, East Asian or African populations. However, these studies also identified several novel loci (*GALNT10* in African ancestry; *KLF9, CDKAL1* and *GP2* in East Asian ancestry) missed previously by GWAS of European ancestry.[259-262] Whereas SNPs in / near *GALNT10*, *CDKAL1* and *GP2* genes showed nominal evidence of associations with BMI in European populations from the GIANT consortium, the *KLF9* signal was not replicated.[260-262] In addition, independent SNPs at the *PCSK1* locus are associated with obesity-related traits in European and East Asian populations.[5, 260] These data suggest some ethnic heterogeneity in the genetic variants of obesity.[263]

Numerous studies have reported the effect of European-derived GWAS obesity signals in other ethnic backgrounds and the results always show at least partial overlap of association with obesity.[263] Only one study to date assessed the generalizability of European-derived obesity signals in a multi-ethnic population. Fesinmeyer *et al.* studied 8-13 BMI SNPs derived from European populations in 69,775 individuals from 6 ethnic groups (East Asians, African Americans, Latino Americans, Pacific Islanders, Native Americans, N between 604 and 15,415), suggesting a limited overlap in genetic variants for obesity across diverse ethnic groups.[264] However, this pioneer study displayed two major limitations. First, the authors analyzed the separate contribution of 8-13 SNPs to obesity, but their collective effect on BMI variation as measured by a genetic risk score was not assessed. In addition, the South Asian ethnic group, which represents one-quarter of the worldwide population, was not included in this study.

My first thesis project (Chapter II) aimed to investigate the generalizability of 23 obesity SNPs, analyzed separately or collectively as a genetic risk score, across six ethnic groups (European, South Asian, East Asian, African, Latino American and Native American; N_{total} =17,423) using EpiDREAM cohort study.

4.6. Generalizability of identified adult SNPs across different ages

GWAS have identified 39 and 20 SNPs associated with adult and child binary obesity, respectively, and 12 of them overlap.[31, 244, 265, 266] Current known BMI susceptibility variants were originally derived from adult population, and the total number of SNPs associated with adult BMI at genome-wide significance level ($P<5\times10^{-8}$) has recently increased to 116.[247] A meta-analysis of GWAS in European adolescents and young adults (aged 16-25 years) found seven loci associated with BMI level[267]. All seven loci were previously identified in European adults, but four of them displayed larger effects and one locus had smaller effect compared with

older adults. A large-scale genome-wide interaction study identified 15 loci of which the effect on BMI was different between younger and older adults, 11 of the 15 age-dependent BMI loci having stronger effects in the younger.[268] Recently, two meta-analyses of GWAS of childhood BMI identified 12 loci previously associated with adult BMI or childhood obesity (*ADCY3*, *GNPDA2*, *TMEM18*, *SEC16B*, *FAIM2*, *FTO*, *TFAP2B*, *TNNI3K*, *MC4R*, *GPR61*, *LMX1B*, *OLFM4*) and four novel loci (*FAM120AOS, ELP3, RAB27B, ADAM23*).[269, 270] Evidence from post-GWAS indicates that some of the adult BMI loci also affect BMI level at specific ages in children.[1, 30, 40, 271, 272] De Hoed *et al.* compared the effect sizes of 17 BMI SNPs in adults and children and reported that *TMEM18*, *SEC16B*, and *KCTD15* had stronger effects in children and adolescents compared to adults, while *BDNF*, *SH2B1* and *MTCH2* had weaker effects.[273] Because children grow rapidly and obesity typically develops over a period of time, longitudinal analyses of repeated measures of weight and BMI are considered to be an optimal approach to look for specific developmental windows during which the genetic variants are associated with BMI. In the prospective 1946 British Birth Cohort, variants in *FTO* and *MC4R* demonstrated significant associations with growth up to adulthood. The associations peaked at 20 years old and then weakened with age in adults.[274] Belsky *et al.* reported that the genetic risk score by summing the number of risk alleles of 32 BMI loci predicted faster growth in childhood (ages 3 through 13 years).[275] These findings indicate that most adult BMI variants start to influence BMI in early childhood. However, in combination, the BMI variants have little influence on birth weight but promote accelerated weight gain.[272]

Birthweight is heritable with 40% of its variation accounted by genetic components.[194] Previous association studies of the effects of genetic risk scores and birth weight and obesity-related traits in childhood are limited by the small group of SNPs analyzed (at most 32 SNPs

which were identified until 2010), and many adult BMI risk variants recently identified have not been investigated. My second thesis project (Chapter III) aimed to investigate the effects of genetic risk score summing the risk alleles of 83 robustly associated adult BMI SNPs on birth weight, weight gain and growth trajectory from birth to 5 years of age in the children using the FAMILY birth cohort. We further examined whether parental risk alleles in specific genes contributed to child's weight and BMI variation in early life.

5. GENE AND ENVIRONMENT INTERACTIONS AND OBESITY

The prevalence of obesity has doubled in adults and tripled in children in the last three decades which indicates the changes in affluent food supply and sedentary lifestyle have driven this epidemic. Some individuals are more likely to be obese when exposed to an obesogenic environment because of different genetic background. For example, when people from developing countries with a low prevalence of obesity migrate to the western countries, the risk of obesity increases substantially in some of them compared to those stay in their country. In those people who become obese may have genetic susceptibility to common forms of obesity and its modest effects are amplified in the presence of environmental factors such as Westernized food and lifestyle. As another example, some individuals fail to lose weight and even gain weight in response to lifestyle intervention, even though they actively involved in the program. One can only assume that the combination of genetic and environmental determinants is responsible for gaining weight or losing weight more readily for some people. Thus, a better understanding of the etiology of obesity requires a careful investigation of gene-environment interactions.

5.1. Obesity susceptibility variants interact with diet

While the *FTO* variant rs99309609 confers a predisposition to obesity in children, it is associated with food intake and food choice independent of body weight, but not energy expenditure.[276] Many studies have demonstrated that *FTO* variants mediated the impact of diet patterns on obesity.[277-280] Moleres *et al.*'s cross-sectional study showed that children and adolescents carrying risk allele(s) of rs9939609 in *FTO* with higher consume in total energy and high saturated fat had an increased obesity risk compared to those with no risk allele.[281] The Preventing Overweight Using Novel Dietary Strategies (POUNDS LOST) trial randomized participants into one of the four weight-loss diets of varying macronutrient composition (the same total energy with different percentages of fat, protein and carbohydrate) and followed for 2 years.[278] The results demonstrated that carriers of the risk allele of *FTO* rs1558902 more successfully reduced weight, body composition and fat distribution in response to high-protein diet than a low-protein diet. Using a nested case-control study, Qi *et al.* found an association between the genetic risk score based on 32 BMI variants and adiposity appeared to be more pronounced in people who took more sugar-sweetened beverages or consumed more fried food.[277, 280] However, a recent meta-analysis from 177,330 adults did not detect an interaction between the *FTO* genetic variants and total energy or macronutrients on the risk of obesity.[279] As pinpointed by the authors, different study designs and inevitable measurement errors in self-reported data on BMI and dietary intakes may substantially influence the pooled estimations.

Peroxisome proliferator-activated receptor-γ 2 (*PPARG2*) Pro12Ala (P12A) is another obesity genetic variant widely studied for its modification effect for the diets on the risk of obesity. Several studies have reported significant interactions between dietary fat/total energy/carbohydrate/ratio of polyunsaturated fat to saturated fat and the *PPARG2* gene on adiposity/BMI.[282-285] For example, Robitaille *et al.* examined the interaction between dietary fat

and the *PPARG2* P12A polymorphism on the level of BMI and waist circumference in a cohort of 720 adults participating in the Quebec Family Study.[283] Intake of total fat and saturated fat were significantly associated with BMI and waist circumference in P homozygotes, but not among carriers of the A allele, indicating significant interactions between total fat and saturated fat intake and *PPARG2* genotype.

5.2. Obesity susceptibility variants interact with physical activity and sedentary lifestyle

Numerous studies have shown that physical activity attenuates the risk of some obesity susceptibility variants on obesity. These genetic variants locate in/near genes *UCP3*, *ADRB3*, *ADRB2*, *PPARGC1A* and *FTO.*[42] Specifically, the interaction between *FTO* and physical activity, particularly moderate to vigorous physical activity, is well documented. The effects of interactions between the *FTO* variants and physical activity on BMI level/obesity/fat mass/fat distribution/waist circumference have been reported in children, adolescents and adults and different ethnic groups.[286-293] In Andreasen *et al.*'s study with 17,162 middle-aged Danes, a significant difference in BMI between the risk allele homozygotes and non-risk allele homozygotes was observed only among physically inactive subjects, but not among those who were physically active. These results suggested that higher physical activity might attenuate the adverse effects of the *FTO* variant on obesity. In addition to numerous independent studies showing significant associations between the *FTO* variants and physical activity on BMI variation in children, adolescents and adults, a recent meta-analysis of 218,166 adults has corroborated that the association of rs9939609 in *FTO* with BMI and obesity was attenuated by physical activity.[294] However, no such interaction was found in 19,268 children and adolescents.[294] In this meta-analysis, physical activity was measured differently in each individual study but was categorized into a dichotomous variable in the pooled analysis, which

may decrease the statistical power. In addition, BMI Z-score for children and adolescent may be calculated using different criteria. As highlighted by the authors, a greater consistency and statistical power could ultimately be reached only through the establishment of large single or multicenter studies using standardized methods and precise measurement of physical activity and obesity-related traits.

In addition to single genetic variants, accumulated effect of multiple genetic variants (as measured by genetic risk score) on the BMI variation has also been reported to be accentuated in the people with less physical activity.[295-297] In a large prospective cohort study, the association of genetic risk score with BMI was reduced with increased level of physical activity and strengthened with increased hours of TV watching.[297] These findings suggested that sedentary lifestyle might boost the effect of genetic association and increased physical activity might lessen the susceptibility to obesity of the genetic variants.

5.3 Obesity susceptibility variants interact with pregnancy and in utero factors

We discussed previously that maternal pre-pregnancy BMI and GWG, which in part reflect maternal and intrauterine nutrition conditions during fetal development, have been linked to subsequent overweight and obesity in mothers and children. As expected, BMI susceptibility variants are associated with maternal pre-pregnancy BMI (Chapter IV). Whether obesity susceptibility variants or genetic variants from other pathway contribute to GWG is still unknown. Recently, Andersson *et al.* estimated that genetic factors explained 43% of the variation in GWG in the first pregnancy and 26% in the second pregnancy using twin mother-pairs.[298] Given that a considerable fraction of GWG is attributable to the fat in both fetus and mother,[111] fetal and maternal genetic variants associated with BMI are anticipated to contribute

to GWG. Stube *et al* first examined 8 obesity-associated genetic loci and none of them were associated with GWG.[299] Using repeated measures of GWG during pregnancy (a median of 8 weight measurements), Lawlor *et al* reported that 4 BMI loci (*FTO*, *MC4R*, *TMEM18* and *GNPDA2*), individually or collectively as a genetic risk score from fetus or mother, were not associated with GWG either.[300]

However, the contribution of many recent identified BMI risk variants to the variation of GWG has not been examined. Furthermore, no such a study has been done in Canada where the environment is believed to be more obesogenic than Europe and Asian. The third project of my thesis aimed to investigate the associations between the genetic risk score (GRS) summing the risk alleles of 83 BMI SNPs of both mother and fetus with GWG (Chapter IV).

5.4 Challenges in gene-environment interaction studies

Theoretically, more gene-environment interactions with larger effect sizes are expected to be found to explain the variation in obesity-related traits. The major obstacles include the inadequate sample size and lack of accurate measurement of tested variables. The sample size needed to test departure from a multiplicative gene-environment interaction is at least four times that required to evaluate the main genetic or environmental associations.[301] To compensate for the measurement errors in environment factors, even larger sample sizes are needed.[302] The identified gene-environment associations above are based on candidate genes. If the whole-genome hypothesis-free approach or Genome-Wide Interaction Study (GEWIS) is applied, it amplifies the problems of multiple-testing and sample size.[303] Currently no study has reported the GEWIS for obesity. To overcome these challenges, some novel approaches have been proposed, including gene- or pathway-based approaches, a module-based cocktail approach, a

joint test of marginal associations and gene-environment interaction, variance prioritization approach, and a set-based gene-environment interaction test.[304]

6. SUMMARY

Considerable progress in understanding of both environmental and genetic risk factors associated with the development of obesity has been achieved. Research findings indicate that increases in physical activity and healthy lifestyle interventions can not only substantially reduce the risk of obesity but also can attenuate the risk of genetic variants on the development of obesity. This has essential implications to clinical and public health. However, there are still many challenges. The future progress will depend on the development of methodology in study designs, analytic methods, accurate measurements of environment variables, and accruing large sample sizes.

REFERENCES

1. Speliotes EK, Willer CJ, Berndt SI, Monda KL, Thorleifsson G, Jackson AU *et al.* Association analyses of 249,796 individuals reveal 18 new loci associated with body mass index. *Nature genetics* 2010; **42**(11)**:** 937-48.
2. Yang J, Wray NR, Visscher PM. Comparing apples and oranges: equating the power of case-control and quantitative trait association studies. *Genetic epidemiology* 2010; **34**(3)**:** 254-7.
3. Zondervan KT, Cardon LR. Designing candidate gene and genome-wide case-control association studies. *Nature protocols* 2007; **2**(10)**:** 2492-501.
4. Leber PD, Davis CS. Threats to the validity of clinical trials employing enrichment strategies for sample selection. *Controlled clinical trials* 1998; **19**(2)**:** 178-87.
5. Benzinou M, Creemers JW, Choquet H, Lobbens S, Dina C, Durand E *et al.* Common nonsynonymous variants in PCSK1 confer risk of obesity. *Nature genetics* 2008; **40**(8)**:** 943-5.
6. Nead KT, Li A, Wehner MR, Neupane B, Gustafsson S, Butterworth A *et al.* Contribution of common non-synonymous variants in PCSK1 to body mass index variation and risk of obesity: a systematic review and meta-analysis with evidence from up to 331 175 individuals. *Human molecular genetics* 2015; **24**(12)**:** 3582-94.
7. Ott J, Kamatani Y, Lathrop M. Family-based designs for genome-wide association studies. *Nature reviews. Genetics* 2011; **12**(7)**:** 465-74.
8. Risch N, Merikangas K. The future of genetic studies of complex human diseases. *Science* 1996; **273**(5281)**:** 1516-7.
9. Egger M, Smith GD, Phillips AN. Meta-analysis: principles and procedures. *BMJ* 1997; **315**(7121)**:** 1533-7.
10. Sinclair JC, Bracken MB. Clinically useful measures of effect in binary analyses of randomized trials. *Journal of clinical epidemiology* 1994; **47**(8)**:** 881-9.
11. Sargent JD, Beach ML, Adachi-Mejia AM, Gibson JJ, Titus-Ernstoff LT, Carusi CP *et al.* Exposure to movie smoking: its relation to smoking initiation among US adolescents. *Pediatrics* 2005; **116**(5)**:** 1183-91.
12. Schechtman E. Odds ratio, relative risk, absolute risk reduction, and the number needed to treat--which of these should we use? *Value in health : the journal of the International Society for Pharmacoeconomics and Outcomes Research* 2002; **5**(5)**:** 431-6.
13. Altman DG, Deeks JJ, Sackett DL. Odds ratios should be avoided when events are common. *BMJ* 1998; **317**(7168)**:** 1318.
14. Deeks J. When can odds ratios mislead? Odds ratios should be used only in case-control studies and logistic regression analyses. *BMJ* 1998; **317**(7166)**:** 1155-6; author reply 1156-7.
15. Laupacis A, Sackett DL, Roberts RS. An assessment of clinically useful measures of the consequences of treatment. *The New England journal of medicine* 1988; **318**(26)**:** 1728-33.
16. Elferink JA, Van Zwieten-Boot BJ. New antiepileptic drugs. Analysis based on number needed to treat shows differences between drugs studied. *BMJ* 1997; **314**(7080)**:** 603.
17. McQuay HJ, Moore RA. Using numerical results from systematic reviews in clinical practice. *Annals of internal medicine* 1997; **126**(9)**:** 712-20.
18. Plomin R, Haworth CM, Davis OS. Common disorders are quantitative traits. *Nature reviews. Genetics* 2009; **10**(12)**:** 872-8.

19. Lindgren CM, Heid IM, Randall JC, Lamina C, Steinthorsdottir V, Qi L *et al.* Genome-wide association scan meta-analysis identifies three Loci influencing adiposity and fat distribution. *PLoS genetics* 2009; **5**(6)**:** e1000508.
20. Heid IM, Jackson AU, Randall JC, Winkler TW, Qi L, Steinthorsdottir V *et al.* Meta-analysis identifies 13 new loci associated with waist-hip ratio and reveals sexual dimorphism in the genetic basis of fat distribution. *Nature genetics* 2010; **42**(11)**:** 949-60.
21. Pei YF, Zhang L, Liu Y, Li J, Shen H, Liu YZ *et al.* Meta-analysis of genome-wide association data identifies novel susceptibility loci for obesity. *Human molecular genetics* 2014; **23**(3)**:** 820-30.
22. Sackett DL. Bias in analytic research. *Journal of chronic diseases* 1979; **32**(1-2)**:** 51-63.
23. Hulley SB, Cummings SR, Browner WS, Grady DG, Newman TB. *Designing clinical research*, Third edn Lippincott Williams &Wilkins: Philadelphia, PA 19106 USA, 2007.
24. Hurwitz ES, Barrett MJ, Bregman D, Gunn WJ, Pinsky P, Schonberger LB *et al.* Public Health Service study of Reye's syndrome and medications. Report of the main study. *JAMA : the journal of the American Medical Association* 1987; **257**(14)**:** 1905-11.
25. Hirschhorn JN, Lohmueller K, Byrne E, Hirschhorn K. A comprehensive review of genetic association studies. *Genetics in medicine : official journal of the American College of Medical Genetics* 2002; **4**(2)**:** 45-61.
26. Chandalia M, Grundy SM, Adams-Huet B, Abate N. Ethnic differences in the frequency of ENPP1/PC1 121Q genetic variant in the Dallas Heart Study cohort. *Journal of diabetes and its complications* 2007; **21**(3)**:** 143-8.
27. Klimentidis YC, Abrams M, Wang J, Fernandez JR, Allison DB. Natural selection at genomic regions associated with obesity and type-2 diabetes: East Asians and sub-Saharan Africans exhibit high levels of differentiation at type-2 diabetes regions. *Human genetics* 2011; **129**(4)**:** 407-18.
28. Kettunen J, Silander K, Saarela O, Amin N, Muller M, Timpson N *et al.* European lactase persistence genotype shows evidence of association with increase in body mass index. *Human molecular genetics* 2010; **19**(6)**:** 1129-36.
29. Ogden CL, Carroll MD, Kit BK, Flegal KM. Prevalence of childhood and adult obesity in the United States, 2011-2012. *JAMA : the journal of the American Medical Association* 2014; **311**(8)**:** 806-14.
30. Willer CJ, Speliotes EK, Loos RJ, Li S, Lindgren CM, Heid IM *et al.* Six new loci associated with body mass index highlight a neuronal influence on body weight regulation. *Nature genetics* 2009; **41**(1)**:** 25-34.
31. Meyre D, Delplanque J, Chevre JC, Lecoeur C, Lobbens S, Gallina S *et al.* Genome-wide association study for early-onset and morbid adult obesity identifies three new risk loci in European populations. *Nature genetics* 2009; **41**(2)**:** 157-9.
32. Delgado-Rodriguez M, Llorca J. Bias. *Journal of epidemiology and community health* 2004; **58**(8)**:** 635-41.
33. Copeland KT, Checkoway H, McMichael AJ, Holbrook RH. Bias due to misclassification in the estimation of relative risk. *American journal of epidemiology* 1977; **105**(5)**:** 488-95.
34. Li A, Meyre D. Challenges in reproducibility of genetic association studies: lessons learned from the obesity field. *Int J Obes (Lond)* 2013; **37**(4)**:** 559-67.

35. Gordon D, Finch SJ, Nothnagel M, Ott J. Power and sample size calculations for case-control genetic association tests when errors are present: application to single nucleotide polymorphisms. *Human heredity* 2002; **54**(1)**:** 22-33.
36. Bouatia-Naji N, De Graeve F, Bronner G, Lecoeur C, Vatin V, Durand E *et al.* INS VNTR is not associated with childhood obesity in 1,023 families: a family-based study. *Obesity (Silver Spring)* 2008; **16**(6)**:** 1471-5.
37. Le Stunff C, Fallin D, Schork NJ, Bougneres P. The insulin gene VNTR is associated with fasting insulin levels and development of juvenile obesity. *Nature genetics* 2000; **26**(4)**:** 444-6.
38. Lunetta KL. Genetic association studies. *Circulation* 2008; **118**(1)**:** 96-101.
39. Stergiakouli E, Gaillard R, Tavare JM, Balthasar N, Loos RJ, Taal HR *et al.* Genome-wide association study of height-adjusted BMI in childhood identifies functional variant in ADCY3. *Obesity (Silver Spring)* 2014; **22**(10)**:** 2252-9.
40. Frayling TM, Timpson NJ, Weedon MN, Zeggini E, Freathy RM, Lindgren CM *et al.* A common variant in the FTO gene is associated with body mass index and predisposes to childhood and adult obesity. *Science* 2007; **316**(5826)**:** 889-94.
41. Obesity: preventing and managing the global epidemic. Report of a WHO consultation. *World Health Organization technical report series* 2000; **894:** i-xii, 1-253.
42. Hu F. Obesity Epidemiology. In: Hu F, (ed). New York City: Oxford University Press, 2008.
43. Wang Y. Epidemiology of childhood obesity--methodological aspects and guidelines: what is new? *International journal of obesity and related metabolic disorders : journal of the International Association for the Study of Obesity* 2004; **28 Suppl 3:** S21-8.
44. Cole TJ, Freeman JV, Preece MA. Body mass index reference curves for the UK, 1990. *Archives of disease in childhood* 1995; **73**(1)**:** 25-9.
45. Rolland-Cachera MF. Childhood obesity: current definitions and recommendations for their use. *International journal of pediatric obesity : IJPO : an official journal of the International Association for the Study of Obesity* 2011; **6**(5-6)**:** 325-31.
46. de Onis M, Onyango AW, Borghi E, Siyam A, Nishida C, Siekmann J. Development of a WHO growth reference for school-aged children and adolescents. *Bulletin of the World Health Organization* 2007; **85**(9)**:** 660-7.
47. Organizatoin WH. World Health Organization Child Growth Standards. In, 2006.
48. Kuczmarski RJ, Flegal KM. Criteria for definition of overweight in transition: background and recommendations for the United States. *The American journal of clinical nutrition* 2000; **72**(5)**:** 1074-81.
49. Kuczmarski RJ, Ogden CL, Grummer-Strawn LM, Flegal KM, Guo SS, Wei R *et al.* CDC growth charts: United States. *Advance data* 2000; (314)**:** 1-27.
50. Grummer-Strawn LM, Reinold C, Krebs NF. Use of World Health Organization and CDC growth charts for children aged 0-59 months in the United States. *MMWR. Recommendations and reports : Morbidity and mortality weekly report. Recommendations and reports / Centers for Disease Control* 2010; **59**(RR-9)**:** 1-15.
51. Lakshman R, Elks CE, Ong KK. Childhood obesity. *Circulation* 2012; **126**(14)**:** 1770-9.
52. Vogel L. Active play key to curbing child obesity. *CMAJ : Canadian Medical Association journal = journal de l'Association medicale canadienne* 2015; **187**(9)**:** E269-70.

53. Skinner AC, Skelton JA. Prevalence and trends in obesity and severe obesity among children in the United States, 1999-2012. *JAMA pediatrics* 2014; **168**(6)**:** 561-6.
54. Pan L, Blanck HM, Sherry B, Dalenius K, Grummer-Strawn LM. Trends in the prevalence of extreme obesity among US preschool-aged children living in low-income families, 1998-2010. *JAMA : the journal of the American Medical Association* 2012; **308**(24)**:** 2563-5.
55. Wen X, Gillman MW, Rifas-Shiman SL, Sherry B, Kleinman K, Taveras EM. Decreasing prevalence of obesity among young children in Massachusetts from 2004 to 2008. *Pediatrics* 2012; **129**(5)**:** 823-31.
56. Robbins JM, Mallya G, Polansky M, Schwarz DF. Prevalence, disparities, and trends in obesity and severe obesity among students in the Philadelphia, Pennsylvania, school district, 2006-2010. *Preventing chronic disease* 2012; **9:** E145.
57. Ogden CL, Carroll MD, Kit BK, Flegal KM. Prevalence of obesity and trends in body mass index among US children and adolescents, 1999-2010. *JAMA : the journal of the American Medical Association* 2012; **307**(5)**:** 483-90.
58. Verbeeten KC, Elks CE, Daneman D, Ong KK. Association between childhood obesity and subsequent Type 1 diabetes: a systematic review and meta-analysis. *Diabetic medicine : a journal of the British Diabetic Association* 2011; **28**(1)**:** 10-8.
59. Reilly JJ, Methven E, McDowell ZC, Hacking B, Alexander D, Stewart L *et al.* Health consequences of obesity. *Archives of disease in childhood* 2003; **88**(9)**:** 748-52.
60. Kohler MJ, Thormaehlen S, Kennedy JD, Pamula Y, van den Heuvel CJ, Lushington K *et al.* Differences in the association between obesity and obstructive sleep apnea among children and adolescents. *Journal of clinical sleep medicine : JCSM : official publication of the American Academy of Sleep Medicine* 2009; **5**(6)**:** 506-11.
61. Caprio S, Daniels SR, Drewnowski A, Kaufman FR, Palinkas LA, Rosenbloom AL *et al.* Influence of race, ethnicity, and culture on childhood obesity: implications for prevention and treatment. *Obesity (Silver Spring)* 2008; **16**(12)**:** 2566-77.
62. Schwimmer JB, Burwinkle TM, Varni JW. Health-related quality of life of severely obese children and adolescents. *JAMA : the journal of the American Medical Association* 2003; **289**(14)**:** 1813-9.
63. Lawlor DA, Mamun AA, O'Callaghan MJ, Bor W, Williams GM, Najman JM. Is being overweight associated with behavioural problems in childhood and adolescence? Findings from the Mater-University study of pregnancy and its outcomes. *Archives of disease in childhood* 2005; **90**(7)**:** 692-7.
64. Sawyer MG, Miller-Lewis L, Guy S, Wake M, Canterford L, Carlin JB. Is there a relationship between overweight and obesity and mental health problems in 4- to 5-year-old Australian children? *Ambulatory pediatrics : the official journal of the Ambulatory Pediatric Association* 2006; **6**(6)**:** 306-11.
65. Srinivasan SR, Myers L, Berenson GS. Changes in metabolic syndrome variables since childhood in prehypertensive and hypertensive subjects: the Bogalusa Heart Study. *Hypertension* 2006; **48**(1)**:** 33-9.
66. Bradford NF. Overweight and obesity in children and adolescents. *Primary care* 2009; **36**(2)**:** 319-39.
67. Kindblom JM, Lorentzon M, Hellqvist A, Lonn L, Brandberg J, Nilsson S *et al.* BMI changes during childhood and adolescence as predictors of amount of adult subcutaneous and visceral adipose tissue in men: the GOOD Study. *Diabetes* 2009; **58**(4)**:** 867-74.

68. Franks PW, Hanson RL, Knowler WC, Sievers ML, Bennett PH, Looker HC. Childhood obesity, other cardiovascular risk factors, and premature death. *The New England journal of medicine* 2010; **362**(6)**:** 485-93.
69. Baker JL, Olsen LW, Sorensen TI. Childhood body-mass index and the risk of coronary heart disease in adulthood. *The New England journal of medicine* 2007; **357**(23)**:** 2329-37.
70. Kuhle S, Kirk S, Ohinmaa A, Yasui Y, Allen AC, Veugelers PJ. Use and cost of health services among overweight and obese Canadian children. *International journal of pediatric obesity : IJPO : an official journal of the International Association for the Study of Obesity* 2011; **6**(2)**:** 142-8.
71. Finkelstein EA, Graham WC, Malhotra R. Lifetime direct medical costs of childhood obesity. *Pediatrics* 2014; **133**(5)**:** 854-62.
72. Twells LK, Gregory DM, Reddigan J, Midodzi WK. Current and predicted prevalence of obesity in Canada: a trend analysis. *CMAJ open* 2014; **2**(1)**:** E18-26.
73. Shields M, Carroll MD, Ogden CL. Adult obesity prevalence in Canada and the United States. *NCHS data brief* 2011; (56)**:** 1-8.
74. Hossain P, Kawar B, El Nahas M. Obesity and diabetes in the developing world--a growing challenge. *The New England journal of medicine* 2007; **356**(3)**:** 213-5.
75. Popkin BM, Adair LS, Ng SW. Global nutrition transition and the pandemic of obesity in developing countries. *Nutrition reviews* 2012; **70**(1)**:** 3-21.
76. Ng M, Fleming T, Robinson M, Thomson B, Graetz N, Margono C *et al.* Global, regional, and national prevalence of overweight and obesity in children and adults during 1980-2013: a systematic analysis for the Global Burden of Disease Study 2013. *Lancet* 2014; **384**(9945)**:** 766-81.
77. Flegal KM, Carroll MD, Kit BK, Ogden CL. Prevalence of obesity and trends in the distribution of body mass index among US adults, 1999-2010. *JAMA : the journal of the American Medical Association* 2012; **307**(5)**:** 491-7.
78. Ogden CL, Yanovski SZ, Carroll MD, Flegal KM. The epidemiology of obesity. *Gastroenterology* 2007; **132**(6)**:** 2087-102.
79. Calle EE, Rodriguez C, Walker-Thurmond K, Thun MJ. Overweight, obesity, and mortality from cancer in a prospectively studied cohort of U.S. adults. *The New England journal of medicine* 2003; **348**(17)**:** 1625-38.
80. Finucane MM, Stevens GA, Cowan MJ, Danaei G, Lin JK, Paciorek CJ *et al.* National, regional, and global trends in body-mass index since 1980: systematic analysis of health examination surveys and epidemiological studies with 960 country-years and 9.1 million participants. *Lancet* 2011; **377**(9765)**:** 557-67.
81. Avila C, Holloway AC, Hahn MK, Morrison KM, Restivo M, Anglin R *et al.* An Overview of Links Between Obesity and Mental Health. *Current obesity reports* 2015; **4**(3)**:** 303-10.
82. Fontaine KR, Redden DT, Wang C, Westfall AO, Allison DB. Years of life lost due to obesity. *JAMA : the journal of the American Medical Association* 2003; **289**(2)**:** 187-93.
83. Chang SH, Pollack LM, Colditz GA. Life Years Lost Associated with Obesity-Related Diseases for U.S. Non-Smoking Adults. *PloS one* 2013; **8**(6)**:** e66550.

84. Tran BX, Nair AV, Kuhle S, Ohinmaa A, Veugelers PJ. Cost analyses of obesity in Canada: scope, quality, and implications. *Cost effectiveness and resource allocation : C/E* 2013; **11**(1)**:** 3.
85. Cawley J, Meyerhoefer C. The medical care costs of obesity: an instrumental variables approach. *Journal of health economics* 2012; **31**(1)**:** 219-30.
86. Withrow D, Alter DA. The economic burden of obesity worldwide: a systematic review of the direct costs of obesity. *Obesity reviews : an official journal of the International Association for the Study of Obesity* 2011; **12**(2)**:** 131-41.
87. Lawlor DA, Relton C, Sattar N, Nelson SM. Maternal adiposity--a determinant of perinatal and offspring outcomes? *Nature reviews. Endocrinology* 2012; **8**(11)**:** 679-88.
88. Drake AJ, Reynolds RM. Impact of maternal obesity on offspring obesity and cardiometabolic disease risk. *Reproduction* 2010; **140**(3)**:** 387-98.
89. Kanagalingam MG, Forouhi NG, Greer IA, Sattar N. Changes in booking body mass index over a decade: retrospective analysis from a Glasgow Maternity Hospital. *BJOG : an international journal of obstetrics and gynaecology* 2005; **112**(10)**:** 1431-3.
90. Heslehurst N, Ells LJ, Simpson H, Batterham A, Wilkinson J, Summerbell CD. Trends in maternal obesity incidence rates, demographic predictors, and health inequalities in 36,821 women over a 15-year period. *BJOG : an international journal of obstetrics and gynaecology* 2007; **114**(2)**:** 187-94.
91. Kim SY, Dietz PM, England L, Morrow B, Callaghan WM. Trends in pre-pregnancy obesity in nine states, 1993-2003. *Obesity (Silver Spring)* 2007; **15**(4)**:** 986-93.
92. Olsen LW, Baker JL, Holst C, Sorensen TI. Birth cohort effect on the obesity epidemic in Denmark. *Epidemiology* 2006; **17**(3)**:** 292-5.
93. Barker DJ. Obesity and early life. *Obesity reviews : an official journal of the International Association for the Study of Obesity* 2007; **8 Suppl 1:** 45-9.
94. Whitaker RC, Dietz WH. Role of the prenatal environment in the development of obesity. *The Journal of pediatrics* 1998; **132**(5)**:** 768-76.
95. Sebire NJ, Jolly M, Harris JP, Wadsworth J, Joffe M, Beard RW *et al.* Maternal obesity and pregnancy outcome: a study of 287,213 pregnancies in London. *International journal of obesity and related metabolic disorders : journal of the International Association for the Study of Obesity* 2001; **25**(8)**:** 1175-82.
96. Gaillard R, Durmus B, Hofman A, Mackenbach JP, Steegers EA, Jaddoe VW. Risk factors and outcomes of maternal obesity and excessive weight gain during pregnancy. *Obesity (Silver Spring)* 2013; **21**(5)**:** 1046-55.
97. Reynolds RM, Osmond C, Phillips DI, Godfrey KM. Maternal BMI, parity, and pregnancy weight gain: influences on offspring adiposity in young adulthood. *The Journal of clinical endocrinology and metabolism* 2010; **95**(12)**:** 5365-9.
98. Hochner H, Friedlander Y, Calderon-Margalit R, Meiner V, Sagy Y, Avgil-Tsadok M *et al.* Associations of maternal prepregnancy body mass index and gestational weight gain with adult offspring cardiometabolic risk factors: the Jerusalem Perinatal Family Follow-up Study. *Circulation* 2012; **125**(11)**:** 1381-9.
99. Stothard KJ, Tennant PW, Bell R, Rankin J. Maternal overweight and obesity and the risk of congenital anomalies: a systematic review and meta-analysis. *JAMA : the journal of the American Medical Association* 2009; **301**(6)**:** 636-50.

100. Chu SY, Bachman DJ, Callaghan WM, Whitlock EP, Dietz PM, Berg CJ *et al.* Association between obesity during pregnancy and increased use of health care. *The New England journal of medicine* 2008; **358**(14)**:** 1444-53.
101. Galtier-Dereure F, Boegner C, Bringer J. Obesity and pregnancy: complications and cost. *The American journal of clinical nutrition* 2000; **71**(5 Suppl)**:** 1242S-8S.
102. Fraser A, Tilling K, Macdonald-Wallis C, Hughes R, Sattar N, Nelson SM *et al.* Associations of gestational weight gain with maternal body mass index, waist circumference, and blood pressure measured 16 y after pregnancy: the Avon Longitudinal Study of Parents and Children (ALSPAC). *The American journal of clinical nutrition* 2011; **93**(6)**:** 1285-92.
103. Fraser A, Tilling K, Macdonald-Wallis C, Sattar N, Brion MJ, Benfield L *et al.* Association of maternal weight gain in pregnancy with offspring obesity and metabolic and vascular traits in childhood. *Circulation* 2010; **121**(23)**:** 2557-64.
104. Ludwig DS, Currie J. The association between pregnancy weight gain and birthweight: a within-family comparison. *Lancet* 2010; **376**(9745)**:** 984-90.
105. Viswanathan M, Siega-Riz AM, Moos MK, Deierlein A, Mumford S, Knaack J *et al.* Outcomes of maternal weight gain. *Evidence report/technology assessment* 2008; (168)**:** 1-223.
106. Ludwig DS, Rouse HL, Currie J. Pregnancy weight gain and childhood body weight: a within-family comparison. *PLoS medicine* 2013; **10**(10)**:** e1001521.
107. Mamun AA, O'Callaghan M, Callaway L, Williams G, Najman J, Lawlor DA. Associations of gestational weight gain with offspring body mass index and blood pressure at 21 years of age: evidence from a birth cohort study. *Circulation* 2009; **119**(13)**:** 1720-7.
108. Mannan M, Doi SA, Mamun AA. Association between weight gain during pregnancy and postpartum weight retention and obesity: a bias-adjusted meta-analysis. *Nutrition reviews* 2013; **71**(6)**:** 343-52.
109. Schack-Nielsen L, Michaelsen KF, Gamborg M, Mortensen EL, Sorensen TI. Gestational weight gain in relation to offspring body mass index and obesity from infancy through adulthood. *Int J Obes (Lond)* 2010; **34**(1)**:** 67-74.
110. Davis RR, Hofferth SL, Shenassa ED. Gestational weight gain and risk of infant death in the United States. *American journal of public health* 2014; **104 Suppl 1:** S90-5.
111. Rasmussen KM, Yaktine AL. Weight gain during pregnancy:reexamining the guidelines. In: guidelines CtrIpw, (ed). Washington, DC: National Academies Press, 2009.
112. Singh AS, Mulder C, Twisk JW, van Mechelen W, Chinapaw MJ. Tracking of childhood overweight into adulthood: a systematic review of the literature. *Obesity reviews : an official journal of the International Association for the Study of Obesity* 2008; **9**(5)**:** 474-88.
113. Magarey AM, Daniels LA, Boulton TJ, Cockington RA. Predicting obesity in early adulthood from childhood and parental obesity. *International journal of obesity and related metabolic disorders : journal of the International Association for the Study of Obesity* 2003; **27**(4)**:** 505-13.
114. Serdula MK, Ivery D, Coates RJ, Freedman DS, Williamson DF, Byers T. Do obese children become obese adults? A review of the literature. *Preventive medicine* 1993; **22**(2)**:** 167-77.

115. Whitaker RC, Wright JA, Pepe MS, Seidel KD, Dietz WH. Predicting obesity in young adulthood from childhood and parental obesity. *The New England journal of medicine* 1997; **337**(13)**:** 869-73.
116. Field AE, Cook NR, Gillman MW. Weight status in childhood as a predictor of becoming overweight or hypertensive in early adulthood. *Obesity research* 2005; **13**(1)**:** 163-9.
117. Speakman JR, Levitsky DA, Allison DB, Bray MS, de Castro JM, Clegg DJ *et al.* Set points, settling points and some alternative models: theoretical options to understand how genes and environments combine to regulate body adiposity. *Disease models & mechanisms* 2011; **4**(6)**:** 733-45.
118. Malik VS, Schulze MB, Hu FB. Intake of sugar-sweetened beverages and weight gain: a systematic review. *The American journal of clinical nutrition* 2006; **84**(2)**:** 274-88.
119. Must A, Barish EE, Bandini LG. Modifiable risk factors in relation to changes in BMI and fatness: what have we learned from prospective studies of school-aged children? *Int J Obes (Lond)* 2009; **33**(7)**:** 705-15.
120. Reilly JJ, Armstrong J, Dorosty AR, Emmett PM, Ness A, Rogers I *et al.* Early life risk factors for obesity in childhood: cohort study. *BMJ* 2005; **330**(7504)**:** 1357.
121. Parsons TJ, Power C, Manor O. Fetal and early life growth and body mass index from birth to early adulthood in 1958 British cohort: longitudinal study. *BMJ* 2001; **323**(7325)**:** 1331-5.
122. Rogers I. The influence of birthweight and intrauterine environment on adiposity and fat distribution in later life. *International journal of obesity and related metabolic disorders : journal of the International Association for the Study of Obesity* 2003; **27**(7)**:** 755-77.
123. Rogers IS, Ness AR, Steer CD, Wells JC, Emmett PM, Reilly JR *et al.* Associations of size at birth and dual-energy X-ray absorptiometry measures of lean and fat mass at 9 to 10 y of age. *The American journal of clinical nutrition* 2006; **84**(4)**:** 739-47.
124. Regnault N, Botton J, Forhan A, Hankard R, Thiebaugeorges O, Hillier TA *et al.* Determinants of early ponderal and statural growth in full-term infants in the EDEN mother-child cohort study. *The American journal of clinical nutrition* 2010; **92**(3)**:** 594-602.
125. Botton J, Heude B, Maccario J, Borys JM, Lommez A, Ducimetiere P *et al.* Parental body size and early weight and height growth velocities in their offspring. *Early human development* 2010; **86**(7)**:** 445-50.
126. Catalano PM, Ehrenberg HM. The short- and long-term implications of maternal obesity on the mother and her offspring. *BJOG : an international journal of obstetrics and gynaecology* 2006; **113**(10)**:** 1126-33.
127. Pedersen J. Weight and length at birth of infants of diabetic mothers. *Acta endocrinologica* 1954; **16**(4)**:** 330-42.
128. Freinkel N. Banting Lecture 1980. Of pregnancy and progeny. *Diabetes* 1980; **29**(12)**:** 1023-35.
129. Levin BE, Govek E. Gestational obesity accentuates obesity in obesity-prone progeny. *The American journal of physiology* 1998; **275**(4 Pt 2)**:** R1374-9.
130. Knight B, Shields BM, Hill A, Powell RJ, Wright D, Hattersley AT. The impact of maternal glycemia and obesity on early postnatal growth in a nondiabetic Caucasian population. *Diabetes care* 2007; **30**(4)**:** 777-83.
131. Lawlor DA, Smith GD, O'Callaghan M, Alati R, Mamun AA, Williams GM *et al.* Epidemiologic evidence for the fetal overnutrition hypothesis: findings from the mater-

university study of pregnancy and its outcomes. *American journal of epidemiology* 2007; **165**(4)**:** 418-24.
132. Linabery AM, Nahhas RW, Johnson W, Choh AC, Towne B, Odegaard AO *et al.* Stronger influence of maternal than paternal obesity on infant and early childhood body mass index: the Fels Longitudinal Study. *Pediatric obesity* 2013; **8**(3)**:** 159-69.
133. Marceau P, Kaufman D, Biron S, Hould FS, Lebel S, Marceau S *et al.* Outcome of pregnancies after biliopancreatic diversion. *Obesity surgery* 2004; **14**(3)**:** 318-24.
134. Kral JG, Biron S, Simard S, Hould FS, Lebel S, Marceau S *et al.* Large maternal weight loss from obesity surgery prevents transmission of obesity to children who were followed for 2 to 18 years. *Pediatrics* 2006; **118**(6)**:** e1644-9.
135. Smith J, Cianflone K, Biron S, Hould FS, Lebel S, Marceau S *et al.* Effects of maternal surgical weight loss in mothers on intergenerational transmission of obesity. *The Journal of clinical endocrinology and metabolism* 2009; **94**(11)**:** 4275-83.
136. Tie HT, Xia YY, Zeng YS, Zhang Y, Dai CL, Guo JJ *et al.* Risk of childhood overweight or obesity associated with excessive weight gain during pregnancy: a meta-analysis. *Archives of gynecology and obstetrics* 2014; **289**(2)**:** 247-57.
137. Silverman BL, Rizzo T, Green OC, Cho NH, Winter RJ, Ogata ES *et al.* Long-term prospective evaluation of offspring of diabetic mothers. *Diabetes* 1991; **40 Suppl 2:** 121-5.
138. Gillman MW, Rifas-Shiman S, Berkey CS, Field AE, Colditz GA. Maternal gestational diabetes, birth weight, and adolescent obesity. *Pediatrics* 2003; **111**(3)**:** e221-6.
139. Hales CN, Barker DJ. Type 2 (non-insulin-dependent) diabetes mellitus: the thrifty phenotype hypothesis. *Diabetologia* 1992; **35**(7)**:** 595-601.
140. Kelishadi R, Haghdoost AA, Jamshidi F, Aliramezany M, Moosazadeh M. Low birthweight or rapid catch-up growth: which is more associated with cardiovascular disease and its risk factors in later life? A systematic review and cryptanalysis. *Paediatrics and international child health* 2015; **35**(2)**:** 110-23.
141. Ravelli AC, van Der Meulen JH, Osmond C, Barker DJ, Bleker OP. Obesity at the age of 50 y in men and women exposed to famine prenatally. *The American journal of clinical nutrition* 1999; **70**(5)**:** 811-6.
142. Ong KK, Ahmed ML, Emmett PM, Preece MA, Dunger DB. Association between postnatal catch-up growth and obesity in childhood: prospective cohort study. *BMJ* 2000; **320**(7240)**:** 967-71.
143. Eriksson JG, Forsen T, Tuomilehto J, Winter PD, Osmond C, Barker DJ. Catch-up growth in childhood and death from coronary heart disease: longitudinal study. *BMJ* 1999; **318**(7181)**:** 427-31.
144. Barker DJ, Osmond C, Forsen TJ, Kajantie E, Eriksson JG. Trajectories of growth among children who have coronary events as adults. *The New England journal of medicine* 2005; **353**(17)**:** 1802-9.
145. Baird J, Fisher D, Lucas P, Kleijnen J, Roberts H, Law C. Being big or growing fast: systematic review of size and growth in infancy and later obesity. *BMJ* 2005; **331**(7522)**:** 929.
146. Ong KK, Loos RJ. Rapid infancy weight gain and subsequent obesity: systematic reviews and hopeful suggestions. *Acta Paediatr* 2006; **95**(8)**:** 904-8.
147. Yajnik CS, Fall CH, Coyaji KJ, Hirve SS, Rao S, Barker DJ *et al.* Neonatal anthropometry: the thin-fat Indian baby. The Pune Maternal Nutrition Study.

International journal of obesity and related metabolic disorders : journal of the International Association for the Study of Obesity 2003; **27**(2)**:** 173-80.
148. Yajnik C. Interactions of perturbations in intrauterine growth and growth during childhood on the risk of adult-onset disease. *The Proceedings of the Nutrition Society* 2000; **59**(2)**:** 257-65.
149. Heijmans BT, Tobi EW, Stein AD, Putter H, Blauw GJ, Susser ES *et al.* Persistent epigenetic differences associated with prenatal exposure to famine in humans. *Proceedings of the National Academy of Sciences of the United States of America* 2008; **105**(44)**:** 17046-9.
150. Aagaard-Tillery KM, Porter TF, Lane RH, Varner MW, Lacoursiere DY. In utero tobacco exposure is associated with modified effects of maternal factors on fetal growth. *American journal of obstetrics and gynecology* 2008; **198**(1)**:** 66 e1-6.
151. Secker-Walker RH, Vacek PM. Relationships between cigarette smoking during pregnancy, gestational age, maternal weight gain, and infant birthweight. *Addictive behaviors* 2003; **28**(1)**:** 55-66.
152. Wang X, Zuckerman B, Pearson C, Kaufman G, Chen C, Wang G *et al.* Maternal cigarette smoking, metabolic gene polymorphism, and infant birth weight. *JAMA : the journal of the American Medical Association* 2002; **287**(2)**:** 195-202.
153. Al Mamun A, Lawlor DA, Alati R, O'Callaghan MJ, Williams GM, Najman JM. Does maternal smoking during pregnancy have a direct effect on future offspring obesity? Evidence from a prospective birth cohort study. *American journal of epidemiology* 2006; **164**(4)**:** 317-25.
154. Bakker R, Timmermans S, Steegers EA, Hofman A, Jaddoe VW. Folic acid supplements modify the adverse effects of maternal smoking on fetal growth and neonatal complications. *The Journal of nutrition* 2011; **141**(12)**:** 2172-9.
155. Young LR, Nestle M. The contribution of expanding portion sizes to the US obesity epidemic. *American journal of public health* 2002; **92**(2)**:** 246-9.
156. Rolls BJ, Morris EL, Roe LS. Portion size of food affects energy intake in normal-weight and overweight men and women. *The American journal of clinical nutrition* 2002; **76**(6)**:** 1207-13.
157. Diliberti N, Bordi PL, Conklin MT, Roe LS, Rolls BJ. Increased portion size leads to increased energy intake in a restaurant meal. *Obesity research* 2004; **12**(3)**:** 562-8.
158. Rolls BJ, Roe LS, Beach AM, Kris-Etherton PM. Provision of foods differing in energy density affects long-term weight loss. *Obesity research* 2005; **13**(6)**:** 1052-60.
159. Metz JA, Stern JS, Kris-Etherton P, Reusser ME, Morris CD, Hatton DC *et al.* A randomized trial of improved weight loss with a prepared meal plan in overweight and obese patients: impact on cardiovascular risk reduction. *Archives of internal medicine* 2000; **160**(14)**:** 2150-8.
160. Wing RR, Jeffery RW. Food provision as a strategy to promote weight loss. *Obesity research* 2001; **9 Suppl 4:** 271S-275S.
161. Wing RR, Jeffery RW, Burton LR, Thorson C, Nissinoff KS, Baxter JE. Food provision vs structured meal plans in the behavioral treatment of obesity. *International journal of obesity and related metabolic disorders : journal of the International Association for the Study of Obesity* 1996; **20**(1)**:** 56-62.
162. Lissner L, Heitmann BL. Dietary fat and obesity: evidence from epidemiology. *European journal of clinical nutrition* 1995; **49**(2)**:** 79-90.

163. Hooper L, Abdelhamid A, Moore HJ, Douthwaite W, Skeaff CM, Summerbell CD. Effect of reducing total fat intake on body weight: systematic review and meta-analysis of randomised controlled trials and cohort studies. *BMJ* 2012; **345:** e7666.
164. Bray GA, Popkin BM. Dietary fat intake does affect obesity! *The American journal of clinical nutrition* 1998; **68**(6)**:** 1157-73.
165. Tobias DK, Chen M, Manson JE, Ludwig DS, Willett W, Hu FB. Effect of low-fat diet interventions versus other diet interventions on long-term weight change in adults: a systematic review and meta-analysis. *The lancet. Diabetes & endocrinology* 2015.
166. Seidell JC. Dietary fat and obesity: an epidemiologic perspective. *The American journal of clinical nutrition* 1998; **67**(3 Suppl)**:** 546S-550S.
167. Nordmann AJ, Nordmann A, Briel M, Keller U, Yancy WS, Jr., Brehm BJ *et al.* Effects of low-carbohydrate vs low-fat diets on weight loss and cardiovascular risk factors: a meta-analysis of randomized controlled trials. *Archives of internal medicine* 2006; **166**(3)**:** 285-93.
168. Schulze MB, Fung TT, Manson JE, Willett WC, Hu FB. Dietary patterns and changes in body weight in women. *Obesity (Silver Spring)* 2006; **14**(8)**:** 1444-53.
169. Bernstein MS, Costanza MC, Morabia A. Association of physical activity intensity levels with overweight and obesity in a population-based sample of adults. *Preventive medicine* 2004; **38**(1)**:** 94-104.
170. Visser M, Launer LJ, Deurenberg P, Deeg DJ. Total and sports activity in older men and women: relation with body fat distribution. *American journal of epidemiology* 1997; **145**(8)**:** 752-61.
171. Chan CB, Spangler E, Valcour J, Tudor-Locke C. Cross-sectional relationship of pedometer-determined ambulatory activity to indicators of health. *Obesity research* 2003; **11**(12)**:** 1563-70.
172. Wareham NJ, van Sluijs EM, Ekelund U. Physical activity and obesity prevention: a review of the current evidence. *The Proceedings of the Nutrition Society* 2005; **64**(2)**:** 229-47.
173. Wenche DB, Holmen J, Kruger O, Midthjell K. Leisure time physical activity and change in body mass index: an 11-year follow-up study of 9357 normal weight health women 20-49 years old. *J Womens Health (Larchmt)* 2004; **13**(1)**:** 55-62.
174. Sternfeld B, Wang H, Quesenberry CP, Jr., Abrams B, Everson-Rose SA, Greendale GA *et al.* Physical activity and changes in weight and waist circumference in midlife women: findings from the Study of Women's Health Across the Nation. *American journal of epidemiology* 2004; **160**(9)**:** 912-22.
175. DiPietro L, Kohl HW, 3rd, Barlow CE, Blair SN. Improvements in cardiorespiratory fitness attenuate age-related weight gain in healthy men and women: the Aerobics Center Longitudinal Study. *International journal of obesity and related metabolic disorders : journal of the International Association for the Study of Obesity* 1998; **22**(1)**:** 55-62.
176. Schmitz KH, Jacobs DR, Jr., Leon AS, Schreiner PJ, Sternfeld B. Physical activity and body weight: associations over ten years in the CARDIA study. Coronary Artery Risk Development in Young Adults. *International journal of obesity and related metabolic disorders : journal of the International Association for the Study of Obesity* 2000; **24**(11)**:** 1475-87.

177. Wing RR. Physical activity in the treatment of the adulthood overweight and obesity: current evidence and research issues. *Medicine and science in sports and exercise* 1999; **31**(11 Suppl)**:** S547-52.
178. Jakicic JM, Winters C, Lang W, Wing RR. Effects of intermittent exercise and use of home exercise equipment on adherence, weight loss, and fitness in overweight women: a randomized trial. *JAMA : the journal of the American Medical Association* 1999; **282**(16)**:** 1554-60.
179. Lim S, O'Reilly S, Behrens H, Skinner T, Ellis I, Dunbar JA. Effective strategies for weight loss in post-partum women: a systematic review and meta-analysis. *Obesity reviews : an official journal of the International Association for the Study of Obesity* 2015; **16**(11)**:** 972-87.
180. Theorell-Haglow J, Berglund L, Berne C, Lindberg E. Both habitual short sleepers and long sleepers are at greater risk of obesity: a population-based 10-year follow-up in women. *Sleep medicine* 2014; **15**(10)**:** 1204-11.
181. Christakis NA, Fowler JH. The spread of obesity in a large social network over 32 years. *The New England journal of medicine* 2007; **357**(4)**:** 370-9.
182. Ross SE, Flynn JI, Pate RR. What is really causing the obesity epidemic? A review of reviews in children and adults. *Journal of sports sciences* 2015**:** 1-6.
183. Knowler WC, Barrett-Connor E, Fowler SE, Hamman RF, Lachin JM, Walker EA *et al.* Reduction in the incidence of type 2 diabetes with lifestyle intervention or metformin. *The New England journal of medicine* 2002; **346**(6)**:** 393-403.
184. Ryan DH, Espeland MA, Foster GD, Haffner SM, Hubbard VS, Johnson KC *et al.* Look AHEAD (Action for Health in Diabetes): design and methods for a clinical trial of weight loss for the prevention of cardiovascular disease in type 2 diabetes. *Controlled clinical trials* 2003; **24**(5)**:** 610-28.
185. Eight-year weight losses with an intensive lifestyle intervention: the look AHEAD study. *Obesity (Silver Spring)* 2014; **22**(1)**:** 5-13.
186. Pi-Sunyer X, Blackburn G, Brancati FL, Bray GA, Bright R, Clark JM *et al.* Reduction in weight and cardiovascular disease risk factors in individuals with type 2 diabetes: one-year results of the look AHEAD trial. *Diabetes care* 2007; **30**(6)**:** 1374-83.
187. Visscher PM, Hill WG, Wray NR. Heritability in the genomics era--concepts and misconceptions. *Nature reviews. Genetics* 2008; **9**(4)**:** 255-66.
188. Loos RJ. Recent progress in the genetics of common obesity. *British journal of clinical pharmacology* 2009; **68**(6)**:** 811-29.
189. Must A, Spadano J, Coakley EH, Field AE, Colditz G, Dietz WH. The disease burden associated with overweight and obesity. *JAMA : the journal of the American Medical Association* 1999; **282**(16)**:** 1523-9.
190. Walley AJ, Asher JE, Froguel P. The genetic contribution to non-syndromic human obesity. *Nature reviews. Genetics* 2009; **10**(7)**:** 431-42.
191. Maes HH, Neale MC, Eaves LJ. Genetic and environmental factors in relative body weight and human adiposity. *Behavior genetics* 1997; **27**(4)**:** 325-51.
192. Elks CE, den Hoed M, Zhao JH, Sharp SJ, Wareham NJ, Loos RJ *et al.* Variability in the heritability of body mass index: a systematic review and meta-regression. *Frontiers in endocrinology* 2012; **3:** 29.

193. Haworth CM, Carnell S, Meaburn EL, Davis OS, Plomin R, Wardle J. Increasing heritability of BMI and stronger associations with the FTO gene over childhood. *Obesity (Silver Spring)* 2008; **16**(12)**:** 2663-8.
194. Vlietinck R, Derom R, Neale MC, Maes H, van Loon H, Derom C *et al.* Genetic and environmental variation in the birth weight of twins. *Behavior genetics* 1989; **19**(1)**:** 151-61.
195. Koeppen-Schomerus G, Wardle J, Plomin R. A genetic analysis of weight and overweight in 4-year-old twin pairs. *International journal of obesity and related metabolic disorders : journal of the International Association for the Study of Obesity* 2001; **25**(6)**:** 838-44.
196. Wardle J, Carnell S, Haworth CM, Plomin R. Evidence for a strong genetic influence on childhood adiposity despite the force of the obesogenic environment. *The American journal of clinical nutrition* 2008; **87**(2)**:** 398-404.
197. Pietilainen KH, Kaprio J, Rissanen A, Winter T, Rimpela A, Viken RJ *et al.* Distribution and heritability of BMI in Finnish adolescents aged 16y and 17y: a study of 4884 twins and 2509 singletons. *International journal of obesity and related metabolic disorders : journal of the International Association for the Study of Obesity* 1999; **23**(2)**:** 107-15.
198. Faith MS, Pietrobelli A, Nunez C, Heo M, Heymsfield SB, Allison DB. Evidence for independent genetic influences on fat mass and body mass index in a pediatric twin sample. *Pediatrics* 1999; **104**(1 Pt 1)**:** 61-7.
199. Schousboe K, Visscher PM, Erbas B, Kyvik KO, Hopper JL, Henriksen JE *et al.* Twin study of genetic and environmental influences on adult body size, shape, and composition. *International journal of obesity and related metabolic disorders : journal of the International Association for the Study of Obesity* 2004; **28**(1)**:** 39-48.
200. Yang J, Bakshi A, Zhu Z, Hemani G, Vinkhuyzen AA, Lee SH *et al.* Genetic variance estimation with imputed variants finds negligible missing heritability for human height and body mass index. *Nature genetics* 2015; **47**(10)**:** 1114-20.
201. Waalen J. The genetics of human obesity. *Translational research : the journal of laboratory and clinical medicine* 2014; **164**(4)**:** 293-301.
202. Garver WS, Newman SB, Gonzales-Pacheco DM, Castillo JJ, Jelinek D, Heidenreich RA *et al.* The genetics of childhood obesity and interaction with dietary macronutrients. *Genes & nutrition* 2013; **8**(3)**:** 271-87.
203. Cassidy SB, Driscoll DJ. Prader-Willi syndrome. *European journal of human genetics : EJHG* 2009; **17**(1)**:** 3-13.
204. Angulo MA, Butler MG, Cataletto ME. Prader-Willi syndrome: a review of clinical, genetic, and endocrine findings. *Journal of endocrinological investigation* 2015; **38**(12)**:** 1249-63.
205. Forsythe E, Beales PL. Bardet-Biedl syndrome. *European journal of human genetics : EJHG* 2013; **21**(1)**:** 8-13.
206. Montague CT, Farooqi IS, Whitehead JP, Soos MA, Rau H, Wareham NJ *et al.* Congenital leptin deficiency is associated with severe early-onset obesity in humans. *Nature* 1997; **387**(6636)**:** 903-8.
207. Wabitsch M, Funcke JB, Lennerz B, Kuhnle-Krahl U, Lahr G, Debatin KM *et al.* Biologically inactive leptin and early-onset extreme obesity. *The New England journal of medicine* 2015; **372**(1)**:** 48-54.

208. Clement K, Vaisse C, Lahlou N, Cabrol S, Pelloux V, Cassuto D *et al.* A mutation in the human leptin receptor gene causes obesity and pituitary dysfunction. *Nature* 1998; **392**(6674)**:** 398-401.
209. Krude H, Biebermann H, Luck W, Horn R, Brabant G, Gruters A. Severe early-onset obesity, adrenal insufficiency and red hair pigmentation caused by POMC mutations in humans. *Nature genetics* 1998; **19**(2)**:** 155-7.
210. Jackson RS, Creemers JW, Ohagi S, Raffin-Sanson ML, Sanders L, Montague CT *et al.* Obesity and impaired prohormone processing associated with mutations in the human prohormone convertase 1 gene. *Nature genetics* 1997; **16**(3)**:** 303-6.
211. Farooqi IS, Keogh JM, Yeo GS, Lank EJ, Cheetham T, O'Rahilly S. Clinical spectrum of obesity and mutations in the melanocortin 4 receptor gene. *The New England journal of medicine* 2003; **348**(12)**:** 1085-95.
212. Choquet H, Meyre D. Molecular basis of obesity: current status and future prospects. *Current genomics* 2011; **12**(3)**:** 154-68.
213. Kernie SG, Liebl DJ, Parada LF. BDNF regulates eating behavior and locomotor activity in mice. *The EMBO journal* 2000; **19**(6)**:** 1290-300.
214. Michaud JL, Boucher F, Melnyk A, Gauthier F, Goshu E, Levy E *et al.* Sim1 haploinsufficiency causes hyperphagia, obesity and reduction of the paraventricular nucleus of the hypothalamus. *Human molecular genetics* 2001; **10**(14)**:** 1465-73.
215. Silos-Santiago I, Fagan AM, Garber M, Fritzsch B, Barbacid M. Severe sensory deficits but normal CNS development in newborn mice lacking TrkB and TrkC tyrosine protein kinase receptors. *The European journal of neuroscience* 1997; **9**(10)**:** 2045-56.
216. Doche ME, Bochukova EG, Su HW, Pearce LR, Keogh JM, Henning E *et al.* Human SH2B1 mutations are associated with maladaptive behaviors and obesity. *The Journal of clinical investigation* 2012; **122**(12)**:** 4732-6.
217. Borman AD, Pearce LR, Mackay DS, Nagel-Wolfrum K, Davidson AE, Henderson R *et al.* A homozygous mutation in the TUB gene associated with retinal dystrophy and obesity. *Human mutation* 2014; **35**(3)**:** 289-93.
218. Noben-Trauth K, Naggert JK, North MA, Nishina PM. A candidate gene for the mouse mutation tubby. *Nature* 1996; **380**(6574)**:** 534-8.
219. Pearce LR, Atanassova N, Banton MC, Bottomley B, van der Klaauw AA, Revelli JP *et al.* KSR2 mutations are associated with obesity, insulin resistance, and impaired cellular fuel oxidation. *Cell* 2013; **155**(4)**:** 765-77.
220. Nguyen A, Burack WR, Stock JL, Kortum R, Chaika OV, Afkarian M *et al.* Kinase suppressor of Ras (KSR) is a scaffold which facilitates mitogen-activated protein kinase activation in vivo. *Molecular and cellular biology* 2002; **22**(9)**:** 3035-45.
221. Bochukova EG, Huang N, Keogh J, Henning E, Purmann C, Blaszczyk K *et al.* Large, rare chromosomal deletions associated with severe early-onset obesity. *Nature* 2010; **463**(7281)**:** 666-70.
222. Walters RG, Jacquemont S, Valsesia A, de Smith AJ, Martinet D, Andersson J *et al.* A new highly penetrant form of obesity due to deletions on chromosome 16p11.2. *Nature* 2010; **463**(7281)**:** 671-5.
223. Farooqi IS, Drop S, Clements A, Keogh JM, Biernacka J, Lowenbein S *et al.* Heterozygosity for a POMC-null mutation and increased obesity risk in humans. *Diabetes* 2006; **55**(9)**:** 2549-53.

224. Farooqi IS, Keogh JM, Kamath S, Jones S, Gibson WT, Trussell R *et al.* Partial leptin deficiency and human adiposity. *Nature* 2001; **414**(6859)**:** 34-5.
225. Farooqi IS, Wangensteen T, Collins S, Kimber W, Matarese G, Keogh JM *et al.* Clinical and molecular genetic spectrum of congenital deficiency of the leptin receptor. *The New England journal of medicine* 2007; **356**(3)**:** 237-47.
226. Creemers JW, Choquet H, Stijnen P, Vatin V, Pigeyre M, Beckers S *et al.* Heterozygous mutations causing partial prohormone convertase 1 deficiency contribute to human obesity. *Diabetes* 2012; **61**(2)**:** 383-90.
227. van der Klaauw AA, Farooqi IS. The hunger genes: pathways to obesity. *Cell* 2015; **161**(1)**:** 119-32.
228. Philippe J, Stijnen P, Meyre D, De Graeve F, Thuillier D, Delplanque J *et al.* A nonsense loss-of-function mutation in PCSK1 contributes to dominantly inherited human obesity. *Int J Obes (Lond)* 2015; **39**(2)**:** 295-302.
229. Dawn Teare M, Barrett JH. Genetic linkage studies. *Lancet* 2005; **366**(9490)**:** 1036-44.
230. Zhu M, Zhao S. Candidate gene identification approach: progress and challenges. *International journal of biological sciences* 2007; **3**(7)**:** 420-7.
231. Tranchevent LC, Capdevila FB, Nitsch D, De Moor B, De Causmaecker P, Moreau Y. A guide to web tools to prioritize candidate genes. *Briefings in bioinformatics* 2011; **12**(1)**:** 22-32.
232. Li A, Meyre D. Jumping on the Train of Personalized Medicine: A Primer for Non-Geneticist Clinicians: Part 2. Fundamental Concepts in Genetic Epidemiology. *Current psychiatry reviews* 2014; **10**(2)**:** 101-117.
233. Meyre D, Bouatia-Naji N, Tounian A, Samson C, Lecoeur C, Vatin V *et al.* Variants of ENPP1 are associated with childhood and adult obesity and increase the risk of glucose intolerance and type 2 diabetes. *Nature genetics* 2005; **37**(8)**:** 863-7.
234. Meyre D, Lecoeur C, Delplanque J, Francke S, Vatin V, Durand E *et al.* A genome-wide scan for childhood obesity-associated traits in French families shows significant linkage on chromosome 6q22.31-q23.2. *Diabetes* 2004; **53**(3)**:** 803-11.
235. Stone S, Abkevich V, Russell DL, Riley R, Timms K, Tran T *et al.* TBC1D1 is a candidate for a severe obesity gene and evidence for a gene/gene interaction in obesity predisposition. *Human molecular genetics* 2006; **15**(18)**:** 2709-20.
236. Meyre D, Farge M, Lecoeur C, Proenca C, Durand E, Allegaert F *et al.* R125W coding variant in TBC1D1 confers risk for familial obesity and contributes to linkage on chromosome 4p14 in the French population. *Human molecular genetics* 2008; **17**(12)**:** 1798-802.
237. Bell CG, Benzinou M, Siddiq A, Lecoeur C, Dina C, Lemainque A *et al.* Genome-wide linkage analysis for severe obesity in french caucasians finds significant susceptibility locus on chromosome 19q. *Diabetes* 2004; **53**(7)**:** 1857-65.
238. Gunstad J, Schofield P, Paul RH, Spitznagel MB, Cohen RA, Williams LM *et al.* BDNF Val66Met polymorphism is associated with body mass index in healthy adults. *Neuropsychobiology* 2006; **53**(3)**:** 153-6.
239. Shugart YY, Chen L, Day IN, Lewis SJ, Timpson NJ, Yuan W *et al.* Two British women studies replicated the association between the Val66Met polymorphism in the brain-derived neurotrophic factor (BDNF) and BMI. *European journal of human genetics : EJHG* 2009; **17**(8)**:** 1050-5.

240. Malek AJ, Klimentidis YC, Kell KP, Fernandez JR. Associations of the lactase persistence allele and lactose intake with body composition among multiethnic children. *Genes & nutrition* 2013; **8**(5)**:** 487-94.
241. Visscher PM, Brown MA, McCarthy MI, Yang J. Five years of GWAS discovery. *American journal of human genetics* 2012; **90**(1)**:** 7-24.
242. Auton A, Brooks LD, Durbin RM, Garrison EP, Kang HM, Korbel JO *et al.* A global reference for human genetic variation. *Nature* 2015; **526**(7571)**:** 68-74.
243. Scuteri A, Sanna S, Chen WM, Uda M, Albai G, Strait J *et al.* Genome-wide association scan shows genetic variants in the FTO gene are associated with obesity-related traits. *PLoS genetics* 2007; **3**(7)**:** e115.
244. Hinney A, Nguyen TT, Scherag A, Friedel S, Bronner G, Muller TD *et al.* Genome wide association (GWA) study for early onset extreme obesity supports the role of fat mass and obesity associated gene (FTO) variants. *PloS one* 2007; **2**(12)**:** e1361.
245. Dina C, Meyre D, Gallina S, Durand E, Korner A, Jacobson P *et al.* Variation in FTO contributes to childhood obesity and severe adult obesity. *Nature genetics* 2007; **39**(6)**:** 724-6.
246. Loos RJ, Bouchard C. FTO: the first gene contributing to common forms of human obesity. *Obesity reviews : an official journal of the International Association for the Study of Obesity* 2008; **9**(3)**:** 246-50.
247. Locke AE, Kahali B, Berndt SI, Justice AE, Pers TH, Day FR *et al.* Genetic studies of body mass index yield new insights for obesity biology. *Nature* 2015; **518**(7538)**:** 197-206.
248. Maetzel D, Sarkar S, Wang H, Abi-Mosleh L, Xu P, Cheng AW *et al.* Genetic and chemical correction of cholesterol accumulation and impaired autophagy in hepatic and neural cells derived from Niemann-Pick Type C patient-specific iPS cells. *Stem cell reports* 2014; **2**(6)**:** 866-80.
249. Sharpe LJ, Cook EC, Zelcer N, Brown AJ. The UPS and downs of cholesterol homeostasis. *Trends in biochemical sciences* 2014; **39**(11)**:** 527-35.
250. Krasowski MD, Ni A, Hagey LR, Ekins S. Evolution of promiscuous nuclear hormone receptors: LXR, FXR, VDR, PXR, and CAR. *Molecular and cellular endocrinology* 2011; **334**(1-2)**:** 39-48.
251. Heo GY, Liao WL, Turko IV, Pikuleva IA. Features of the retinal environment which affect the activities and product profile of cholesterol-metabolizing cytochromes P450 CYP27A1 and CYP11A1. *Archives of biochemistry and biophysics* 2012; **518**(2)**:** 119-26.
252. Kersten S. Physiological regulation of lipoprotein lipase. *Biochimica et biophysica acta* 2014; **1841**(7)**:** 919-33.
253. Ward TL, Valberg SJ, Adelson DL, Abbey CA, Binns MM, Mickelson JR. Glycogen branching enzyme (GBE1) mutation causing equine glycogen storage disease IV. *Mammalian genome : official journal of the International Mammalian Genome Society* 2004; **15**(7)**:** 570-7.
254. Zhao J, Bradfield JP, Zhang H, Annaiah K, Wang K, Kim CE *et al.* Examination of all type 2 diabetes GWAS loci reveals HHEX-IDE as a locus influencing pediatric BMI. *Diabetes* 2010; **59**(3)**:** 751-5.

255. Claussnitzer M, Dankel SN, Kim KH, Quon G, Meuleman W, Haugen C *et al.* FTO Obesity Variant Circuitry and Adipocyte Browning in Humans. *The New England journal of medicine* 2015; **373**(10)**:** 895-907.

256. Hatakeyama H, Kanzaki M. Regulatory mode shift of Tbc1d1 is required for acquisition of insulin-responsive GLUT4-trafficking activity. *Molecular biology of the cell* 2013; **24**(6)**:** 809-17.
257. Zhao X, Yang Y, Sun BF, Zhao YL, Yang YG. FTO and obesity: mechanisms of association. *Current diabetes reports* 2014; **14**(5)**:** 486.
258. Zuk O, Schaffner SF, Samocha K, Do R, Hechter E, Kathiresan S *et al.* Searching for missing heritability: designing rare variant association studies. *Proceedings of the National Academy of Sciences of the United States of America* 2014; **111**(4)**:** E455-64.
259. Guo Y, Lanktree MB, Taylor KC, Hakonarson H, Lange LA, Keating BJ. Gene-centric meta-analyses of 108 912 individuals confirm known body mass index loci and reveal three novel signals. *Human molecular genetics* 2013; **22**(1)**:** 184-201.
260. Wen W, Cho YS, Zheng W, Dorajoo R, Kato N, Qi L *et al.* Meta-analysis identifies common variants associated with body mass index in east Asians. *Nature genetics* 2012; **44**(3)**:** 307-11.
261. Okada Y, Kubo M, Ohmiya H, Takahashi A, Kumasaka N, Hosono N *et al.* Common variants at CDKAL1 and KLF9 are associated with body mass index in east Asian populations. *Nature genetics* 2012; **44**(3)**:** 302-6.
262. Monda KL, Chen GK, Taylor KC, Palmer C, Edwards TL, Lange LA *et al.* A meta-analysis identifies new loci associated with body mass index in individuals of African ancestry. *Nature genetics* 2013; **45**(6)**:** 690-6.
263. Lu Y, Loos RJ. Obesity genomics: assessing the transferability of susceptibility loci across diverse populations. *Genome medicine* 2013; **5**(6)**:** 55.
264. Fesinmeyer MD, North KE, Ritchie MD, Lim U, Franceschini N, Wilkens LR *et al.* Genetic risk factors for BMI and obesity in an ethnically diverse population: results from the population architecture using genomics and epidemiology (PAGE) study. *Obesity (Silver Spring)* 2013; **21**(4)**:** 835-46.
265. Bradfield JP, Taal HR, Timpson NJ, Scherag A, Lecoeur C, Warrington NM *et al.* A genome-wide association meta-analysis identifies new childhood obesity loci. *Nature genetics* 2012; **44**(5)**:** 526-31.
266. Scherag A, Dina C, Hinney A, Vatin V, Scherag S, Vogel CI *et al.* Two new Loci for body-weight regulation identified in a joint analysis of genome-wide association studies for early-onset extreme obesity in French and german study groups. *PLoS genetics* 2010; **6**(4)**:** e1000916.
267. Graff M, Ngwa JS, Workalemahu T, Homuth G, Schipf S, Teumer A *et al.* Genome-wide analysis of BMI in adolescents and young adults reveals additional insight into the effects of genetic loci over the life course. *Human molecular genetics* 2013; **22**(17)**:** 3597-607.
268. Winkler TW, Justice AE, Graff M, Barata L, Feitosa MF, Chu S *et al.* The Influence of Age and Sex on Genetic Associations with Adult Body Size and Shape: A Large-Scale Genome-Wide Interaction Study. *PLoS genetics* 2015; **11**(10)**:** e1005378.
269. Warrington NM, Howe LD, Paternoster L, Kaakinen M, Herrala S, Huikari V *et al.* A genome-wide association study of body mass index across early life and childhood. *International journal of epidemiology* 2015; **44**(2)**:** 700-12.

270. Felix JF, Bradfield JP, Monnereau C, van der Valk RJ, Stergiakouli E, Chesi A *et al.* Genome-wide association analysis identifies three new susceptibility loci for childhood body mass index. *Human molecular genetics* 2015.

271. Loos RJ, Lindgren CM, Li S, Wheeler E, Zhao JH, Prokopenko I *et al.* Common variants near MC4R are associated with fat mass, weight and risk of obesity. *Nature genetics* 2008; **40**(6)**:** 768-75.
272. Elks CE, Loos RJ, Sharp SJ, Langenberg C, Ring SM, Timpson NJ *et al.* Genetic markers of adult obesity risk are associated with greater early infancy weight gain and growth. *PLoS medicine* 2010; **7**(5)**:** e1000284.
273. den Hoed M, Ekelund U, Brage S, Grontved A, Zhao JH, Sharp SJ *et al.* Genetic susceptibility to obesity and related traits in childhood and adolescence: influence of loci identified by genome-wide association studies. *Diabetes* 2010; **59**(11)**:** 2980-8.
274. Hardy R, Wills AK, Wong A, Elks CE, Wareham NJ, Loos RJ *et al.* Life course variations in the associations between FTO and MC4R gene variants and body size. *Human molecular genetics* 2010; **19**(3)**:** 545-52.
275. Belsky DW, Moffitt TE, Houts R, Bennett GG, Biddle AK, Blumenthal JA *et al.* Polygenic risk, rapid childhood growth, and the development of obesity: evidence from a 4-decade longitudinal study. *Archives of pediatrics & adolescent medicine* 2012; **166**(6)**:** 515-21.
276. Cecil JE, Tavendale R, Watt P, Hetherington MM, Palmer CN. An obesity-associated FTO gene variant and increased energy intake in children. *The New England journal of medicine* 2008; **359**(24)**:** 2558-66.
277. Qi Q, Chu AY, Kang JH, Jensen MK, Curhan GC, Pasquale LR *et al.* Sugar-sweetened beverages and genetic risk of obesity. *The New England journal of medicine* 2012; **367**(15)**:** 1387-96.
278. Zhang X, Qi Q, Zhang C, Smith SR, Hu FB, Sacks FM *et al.* FTO genotype and 2-year change in body composition and fat distribution in response to weight-loss diets: the POUNDS LOST Trial. *Diabetes* 2012; **61**(11)**:** 3005-11.
279. Qi Q, Kilpelainen TO, Downer MK, Tanaka T, Smith CE, Sluijs I *et al.* FTO genetic variants, dietary intake and body mass index: insights from 177,330 individuals. *Human molecular genetics* 2014; **23**(25)**:** 6961-72.
280. Qi Q, Chu AY, Kang JH, Huang J, Rose LM, Jensen MK *et al.* Fried food consumption, genetic risk, and body mass index: gene-diet interaction analysis in three US cohort studies. *BMJ* 2014; **348:** g1610.
281. Moleres A, Ochoa MC, Rendo-Urteaga T, Martinez-Gonzalez MA, Azcona San Julian MC, Martinez JA *et al.* Dietary fatty acid distribution modifies obesity risk linked to the rs9939609 polymorphism of the fat mass and obesity-associated gene in a Spanish case-control study of children. *The British journal of nutrition* 2012; **107**(4)**:** 533-8.
282. Luan J, Browne PO, Harding AH, Halsall DJ, O'Rahilly S, Chatterjee VK *et al.* Evidence for gene-nutrient interaction at the PPARgamma locus. *Diabetes* 2001; **50**(3)**:** 686-9.
283. Robitaille J, Despres JP, Perusse L, Vohl MC. The PPAR-gamma P12A polymorphism modulates the relationship between dietary fat intake and components of the metabolic syndrome: results from the Quebec Family Study. *Clinical genetics* 2003; **63**(2)**:** 109-16.

284. Vaccaro O, Lapice E, Monticelli A, Giacchetti M, Castaldo I, Galasso R *et al.* Pro12Ala polymorphism of the PPARgamma2 locus modulates the relationship between energy intake and body weight in type 2 diabetic patients. *Diabetes care* 2007; **30**(5)**:** 1156-61.
285. Marti A, Corbalan MS, Martinez-Gonzalez MA, Forga L, Martinez JA. CHO intake alters obesity risk associated with Pro12Ala polymorphism of PPARgamma gene. *Journal of physiology and biochemistry* 2002; **58**(4)**:** 219-20.
286. Andreasen CH, Stender-Petersen KL, Mogensen MS, Torekov SS, Wegner L, Andersen G *et al.* Low physical activity accentuates the effect of the FTO rs9939609 polymorphism on body fat accumulation. *Diabetes* 2008; **57**(1)**:** 95-101.
287. Richardson AS, North KE, Graff M, Young KM, Mohlke KL, Lange LA *et al.* Moderate to vigorous physical activity interactions with genetic variants and body mass index in a large US ethnically diverse cohort. *Pediatric obesity* 2014; **9**(2)**:** e35-46.
288. Vimaleswaran KS, Li S, Zhao JH, Luan J, Bingham SA, Khaw KT *et al.* Physical activity attenuates the body mass index-increasing influence of genetic variation in the FTO gene. *The American journal of clinical nutrition* 2009; **90**(2)**:** 425-8.
289. Demerath EW, Lutsey PL, Monda KL, Linda Kao WH, Bressler J, Pankow JS *et al.* Interaction of FTO and physical activity level on adiposity in African-American and European-American adults: the ARIC study. *Obesity (Silver Spring)* 2011; **19**(9)**:** 1866-72.
290. Ruiz JR, Labayen I, Ortega FB, Legry V, Moreno LA, Dallongeville J *et al.* Attenuation of the effect of the FTO rs9939609 polymorphism on total and central body fat by physical activity in adolescents: the HELENA study. *Archives of pediatrics & adolescent medicine* 2010; **164**(4)**:** 328-33.
291. Ahmad T, Lee IM, Pare G, Chasman DI, Rose L, Ridker PM *et al.* Lifestyle interaction with fat mass and obesity-associated (FTO) genotype and risk of obesity in apparently healthy U.S. women. *Diabetes care* 2011; **34**(3)**:** 675-80.
292. Mitchell JA, Church TS, Rankinen T, Earnest CP, Sui X, Blair SN. FTO genotype and the weight loss benefits of moderate intensity exercise. *Obesity (Silver Spring)* 2010; **18**(3)**:** 641-3.
293. Xi B, Shen Y, Zhang M, Liu X, Zhao X, Wu L *et al.* The common rs9939609 variant of the fat mass and obesity-associated gene is associated with obesity risk in children and adolescents of Beijing, China. *BMC medical genetics* 2010; **11:** 107.
294. Kilpelainen TO, Qi L, Brage S, Sharp SJ, Sonestedt E, Demerath E *et al.* Physical activity attenuates the influence of FTO variants on obesity risk: a meta-analysis of 218,166 adults and 19,268 children. *PLoS medicine* 2011; **8**(11)**:** e1001116.
295. Li S, Zhao JH, Luan J, Ekelund U, Luben RN, Khaw KT *et al.* Physical activity attenuates the genetic predisposition to obesity in 20,000 men and women from EPIC-Norfolk prospective population study. *PLoS medicine* 2010; **7**(8).
296. Zhu J, Loos RJ, Lu L, Zong G, Gan W, Ye X *et al.* Associations of genetic risk score with obesity and related traits and the modifying effect of physical activity in a Chinese Han population. *PloS one* 2014; **9**(3)**:** e91442.
297. Qi Q, Li Y, Chomistek AK, Kang JH, Curhan GC, Pasquale LR *et al.* Television watching, leisure time physical activity, and the genetic predisposition in relation to body mass index in women and men. *Circulation* 2012; **126**(15)**:** 1821-7.
298. Andersson ES, Silventoinen K, Tynelius P, Nohr EA, Sorensen TI, Rasmussen F. Heritability of Gestational Weight Gain-A Swedish Register-Based Twin Study. *Twin*

research and human genetics : the official journal of the International Society for Twin Studies 2015**:** 1-9.
299. Stuebe AM, Lyon H, Herring AH, Ghosh J, Wise A, North KE *et al.* Obesity and diabetes genetic variants associated with gestational weight gain. *American journal of obstetrics and gynecology* 2010; **203**(3)**:** 283 e1-17.
300. Lawlor DA, Fraser A, Macdonald-Wallis C, Nelson SM, Palmer TM, Davey Smith G *et al.* Maternal and offspring adiposity-related genetic variants and gestational weight gain. *The American journal of clinical nutrition* 2011; **94**(1)**:** 149-55.
301. Hunter DJ. Gene-environment interactions in human diseases. *Nature reviews. Genetics* 2005; **6**(4)**:** 287-98.
302. Wong MY, Day NE, Luan JA, Chan KP, Wareham NJ. The detection of gene-environment interaction for continuous traits: should we deal with measurement error by bigger studies or better measurement? *International journal of epidemiology* 2003; **32**(1)**:** 51-7.
303. Cordell HJ. Detecting gene-gene interactions that underlie human diseases. *Nature reviews. Genetics* 2009; **10**(6)**:** 392-404.
304. Huang T, Hu FB. Gene-environment interactions and obesity: recent developments and future directions. *BMC medical genomics* 2015; **8 Suppl 1:** S2.

CHAPTER II: GENERALIZABILITY OF OBESITY SUSCEPTIBILITY LOCI ACROSS MULTIPLE ETHNIC GROUPS

Aihua Li[1], Sebastien Robiou du Pont[1], Hertzel Gerstein[1,2,3], James C. Engert[4,5], Viswanathan Mohan[6], Rafael Diaz[7], Salim Yusuf[1,2,3], Sonia S. Anand[1,2,3], David Meyre[1,2*]

[1]Department of Clinical Epidemiology and Biostatistics, McMaster University, Hamilton, Ontario, Canada
[2]Population Health Research Institute, McMaster University and Hamilton Health Sciences, Hamilton General Hospital, Hamilton, Ontario, Canada
[3]Department of Medicine, McMaster University, Hamilton, Ontario, Canada
[4]Department of Medicine, McGill University, Montreal, Quebec, Canada
[5]Department of Human Genetics, McGill University, Montreal, Quebec, Canada
[6]Madras Diabetes Research Foundation, Gopalapuram, Chennai, India
[7]Estudios Cardiológicos Latino américa, Rosario, Argentina

ABSTRACT

Genome-wide association studies have identified 136 loci associated with body mass index (BMI) or obesity and these studies have been conducted primarily in the populations of Northern European ancestry. We hypothesize that genome-wide associated single nucleotide polymorphisms (SNPs) may have differentiated effect sizes across different ethnic groups. Twenty-three obesity-susceptibility SNPs were genotyped in six ethnic groups (Europeans, East Asians, South Asians, Africans, Latinos and Native Americans) in the EpiDREAM cohort study (N=17,423). Linear mixed models were used to examine the genetic associations between SNP and genetic risk score (GRS) with BMI in each ethnicity and overall. The results showed that 19 out of 22 selected SNPs and the GRS had associations directionally consistent with previous reports and six of them reached at least nominal statistical significance. Two SNPs (rs1805081 in *NPC1* and rs611203 in *USP37*) and the GRS had significantly different effect sizes across six ethnicities (I^2=57%, 72%, 60%, and P_{het}=0.04, 0.003, 0.03, respectively). Therefore their generalizability across different ethnic groups is partial.

INTRODUCTION

The prevalence of obesity in adults (defined as a body mass index (BMI) greater than 30) has risen tremendously since 1980s, and reached 34.9% in the United States in 2012.[1] Obesity is a risk factor for type 2 diabetes (T2D), cardiovascular diseases, stroke, hypertension, nonalcoholic fatty liver diseases and certain specific cancers. It ultimately leads to an 8-13 years shorter life expectancy in its more severe forms.[2-4] Economic burden is also an important consequence of this epidemic that was estimated to be $147 billion in the United States in 2008, corresponding to 9.1% of the total annual health care expenditures.[5]

Although obesity is pandemic worldwide, its prevalence varies across and within countries.[1, 3, 6, 7] Some ethnic groups are more prone to be obese than others. In 2011-2012, the age-adjusted prevalence of obesity in adults from the United States was 47.8%, 42.5%, 32.6%, and 10.8% in non-Hispanic Blacks, Hispanics, non-Hispanic White Americans and non-Hispanic Asians, respectively.[1] These disparities may be due to differences in lifestyle, diets, socioeconomic status, or access to health care across the different ethnic groups. However, these differences may also reflect differences in genetic susceptibility to obesity.[8] Admixture studies for obesity also argue for genuine genetic differences across various ethnic groups.[9]

Obesity is a multi-factorial disorder in which genetic and environmental factors act in concert.[10] Environmental risk exposures for obesity include sedentary lifestyle, excessive energy intake and sleep debt, among many others.[11] However, the fact that only a subset of people exposed to an obesity-prone environment develop obesity is mainly explained by inherited components.[12] Heritability estimations (which measures the fraction of phenotype variance in a population attributed by genetic components) of 47-90% for BMI from twin studies have been reported in literature.[13] Rare mutations or structural variations have been implicated in

monogenic forms of obesity with or without syndromic features.[14, 15] Genome-wide association studies (GWAS) and Metabochip meta-analyses have identified and validated 136 loci associated with BMI or obesity.[14, 16] These GWAS have been conducted primarily in the populations of Northern European ancestry. More recently, GWAS for BMI have been conducted in East Asian and African populations.[17-19] The BMI-associated loci identified by GWAS display a large overlap in European, East Asian or African populations. However, these studies also identified several novel loci (*GALNT10* in African ancestry; *KLF9, CDKAL1* and *GP2* in East Asian ancestry) missed in GWAS of European ancestry.[17-19] In addition, independent SNPs at the *PCSK1* locus have been associated with obesity-related traits in European and East Asian populations.[17, 20] These data suggest some ethnic heterogeneity in the genetic architecture of obesity and question the generalizability of GWAS signals across different ethnic groups.[21]

Numerous studies have reported the effect of European-derived GWAS obesity signals in other ethnic backgrounds and the results always showed at least partial overlap of association with obesity.[21] Surprisingly, only one study to date assessed the generalizability of European-derived obesity signals in a multi-ethnic population.[22] However, this pioneer study displayed two major limitations. First, the authors analyzed the separate contribution of 8-13 SNPs to obesity, but their cumulative effect on BMI variation as measured by a genetic risk score was not assessed. In addition, the South Asian ethnic group, which represents one-quarter of the worldwide population, was not included in this study.

This study aimed to investigate the generalizability of 23 obesity SNPs, analyzed separately or collectively as a genetic risk score, across six ethnic groups (European, South Asian, East Asian, African, Latino American and Native American; N_{total} =17,423) using a multi-ethnic prospective EpiDREAM cohort study.

PARTICIPANTS AND METHODS

Study population

The individuals were at risk for dysglycemia and originally collected through a prospective cohort, as described in detail elsewhere.[23, 24] Briefly, EpiDREAM enrolled a total of 24,872 individuals across 191 centers in 21 countries that were screened to enter the DREAM clinical trial.[23] Individuals who were considered as at risk for dysglycemia based on family history, ethnicity, abdominal adiposity and gestational diabetes which was screened using a 75-gram oral glucose tolerance test (OGTT) after an overnight fasting. All participants were between the age of 18 and 85 years and were recruited between July 2001 and August 2003. Our study used cross-sectional data from the baseline screening visit for the EpiDREAM and DREAM studies. Self-reported ethnicity has been confirmed by a principal component analysis using the EIGENSOFT software.[25] Samples that failed to cluster with individuals of the same self-reported ethnicity were removed. Individuals who lacked relevant clinical data, were sex discordant with genotyping information, or did not pass the genotyping quality control (genotyping call rates <97%) were removed. Consequently, 17,423 subjects who were finally analyzed were from six ethnic groups mentioned above and had both required phenotypic measurements at baseline and genotyping data (Supplementary Figure 1). The EpiDREAM and DREAM studies have been approved by local research ethics board and informed consent was obtained from each subject in accordance with the Declaration of Helsinki.

Genotyping

DNA extracted from buffy coats was genotyped using the Illumina cardiovascular gene-centric bead chip microarray ITMAT Broad Care which captures genetic variants in great depth of coverage in the genes for cardiovascular, metabolic and inflammatory diseases.[26] Genotyping

was conducted at McGill University and Genome Quebec Innovation Centre using the Illumina Bead Studio genotyping module version 3.2. Within each ethnicity, SNPs were further examined for minor allele frequency (MAF), risk allele frequency and deviation from Hardy-Weinberg equilibrium (HWE). SNPs were excluded if MAF<1% or P values $<1\times10^{-6}$ in HWE tests. A list of 23 SNPs (lead SNP or its proxy, detailed information in Supplementary Table 1) was selected based on two criteria of: 1) being associated with BMI and/or binary obesity status at a genome-wide significance level ($P<5\times10^{-8}$), and 2) being available in the cardiovascular gene-centric 50K SNP array. We used the following criteria to select proxy SNPs: 1) SNPs were included in the Illumina cardiovascular gene-centric array; 2) r^2 >0.90 in a population of similar ancestry in which the lead SNP was identified from the 1000 Genomes Project; 3) Pairwise linkage disequilibrium in the European population of all the SNPs was also examined and $r^2<0.1$ was ensured to avoid any overlap in the final SNP list. We used the Broad Institute website tool SNAP (SNP Annotation and Proxy Search) and identified seven proxy SNPs (Supplementary Table 1). The 23 SNPs passed the genotype quality control and the call rates for each of the 23 SNPs were between 99.72 and 100 % (Supplementary Table 1).

Phenotypes

At baseline visit, participants completed a questionnaire collecting demographic data, medical history, physical activity behaviors and diet patterns. Anthropometric measurements including height and weight were performed using standardized protocols.[23] Weight in kilograms (kg) and height in meters (m) were measured by trained medical staff. Standing height was measured to the nearest 0.1 cm and weight was measured to the nearest 0.1 kg in light clothing. BMI was calculated as weight divided by height squared.

Statistical analysis

The comparisons of risk allele frequencies and BMI at baseline among ethnic groups were conducted using Chi-square and ANOVA tests. The obesity risk alleles for each of the 23 SNPs were chosen based on what were originally identified in literature (Supplementary Table 1, 2). An additive mode of inheritance was applied in all relevant analyses and genotypes were coded as 0, 1 and 2 designating the number of copies of the risk allele. A genetic risk score (GRS) was calculated by adding up the risk alleles of 23 SNPs and therefore the theoretical scores ranged from 0 to 46. We used an unweighted GRS as recommended by Janssens *et al.*[27] The scores for the missing genotypes were imputed with arithmetic average of the coded genotypes from individuals who were successfully genotyped within each ethnicity.

In each ethnic population, the associations between each SNP or GRS and BMI were analyzed using linear mixed effect model in which age, sex and SNP/GRS were the variables for the fixed effects and identified relatedness as a random effect. In pooled analyses, the same linear mixed effect model was used with an additional variable of ethnicity for the fixed effect. Interaction term of SNP/GRS by ethnicity was also added to the model to examine the modification effect of SNP/GRS on BMI by ethnicity. The overall β coefficients across 6 ethnicities of each SNP/GRS were furthermore calculated using the fixed-effects meta-analysis. When heterogeneity existed, the random-effects meta-analysis was applied. P-values <0.00026 (0.05/(24×8)) were considered as statistically significant after Bonferroni correction and P-values between 0.00026 and 0.05 were regarded as nominally significant.

Within each ethnic group of South Asians, East Asians, Africans, Latinos and Native Americans, the risk allele frequencies of each SNP was compared with those in Europeans and the counts of increases and decreased were recorded. Whether the increased or decreased counts

were significant was examined using a binomial test (two-sided). Similarly, whether the associations of SNPs or GRS with BMI were directionally consistent with previous reports was also examined by binomial tests (one-sided). All the binomial tests used a null expected ratio of 0.5. All the statistical analyses were performed using PLINK (version 1.07) and R (version 2.15.2.).[28, 29]

RESULTS

Baseline characteristics of participants

The descriptive baseline characteristics of participants in each ethnic group and overall population are presented in Table 1. Their average age was 52.7 years and 61% were women. BMI was lowest in East Asians (mean±SD, 26.1±4.3 kg/m^2) and South Asians (26.4±4.4 kg/m^2), highest in Africans (32.4±7.0 kg/m^2) and Native Americans (32.5±6.4 kg/m^2) and intermediate in Europeans (30.6±6.1 kg/m^2) and Latinos (30.1±6.2 kg/m^2). The distribution pattern of the BMI in 6 ethnicities was Native Americans>Africans>Latinos>Europeans>South Asians>East Asians. The difference of BMI was statistically significant among ethnicities (P=0).

Frequencies of risk alleles across 6 ethnicities

The risk allele frequencies of the 23 tested SNPs in each ethnic group were shown in Table 2. The allele frequencies of the 19 out of 23 SNPs were significantly different across ethnicities. To further understand the extent of difference, we counted the number of risk alleles which had higher or lower frequencies in South Asian, East Asian, African, Latino and Native American compared to those in Europeans. None of them achieved statistical significance at binominal tests in any ethnicity, meaning that there was no any evidence for a specific enrichment or reduction of obesity risk alleles among the ethnic groups (Supplementary Table 3).

Meanwhile, the GRSs were significantly different across the 6 ethnic groups ($P=3.51 \times 10^{-270}$). The order of the GRS values among 6 populations was Africans (27.1±2.9)>Europeans (25.8±3.1)>Latinos (24.7±2.9)>Native Americans (24.5±2.9)>South Asians (24.5±2.9)>East Asian (22.3±2.8). Although both BMI and GRS were statistically different among the 6 ethnicities, their patterns were not the same.

Effects of SNPs and GRS on BMI levels across 6 ethnicities

The minor allele A of the SNP rs761 in *ALDH2* was very rare (< 0.7%) in all ethnic groups and the β coefficients calculated for this SNP had wide-range confidence intervals which we considered as unreliable. Therefore we decided not to report the associations between rs671 and BMI in 6 ethnicities, but the number of risk alleles G of rs671 still contributed to GRS. We first examined whether the direction of effects of the 22 SNP risk alleles on BMI was consistent with previous reports (β coefficients> 0) in each ethnic group. We observed that 19 out of 22 SNPs showed directional consistency in Europeans (binomial probability, P=0.004), 17/22 in both East Asians and Latinos (P=0.01), 15/22 in South Asians (P=0.07), 14/22 in Native Americans (P=0.14), and 11/22 in Africans (P=0.58) (Table 3). The number of associations that reached nominal significance (P<0.05) were 0 in East Asians and Native Americans, 1 in Africans, 2 in South Asians, 3 in Latinos, and 6 in Europeans (Table 3).

We next examined the overall effects of each SNP using both linear mixed regressions and meta-analyses across the 6 ethnic groups. Nineteen of twenty-two SNPs showed associations directionally consistent with previous reports and six of them reached at least a nominal significance: rs1514176 in *TNNI3K*, rs2206734 in *CDKAL1*, rs9939609 in *FTO*, rs11671664 in *GIPR*, rs2984618 in *TAL1*, and rs7903146 in *TCF7L2* (Table 3). No significant interactions between SNP and ethnicity on BMI levels were detected except the interactions between

rs2206734 (*CDKAL1*), rs9939609 (*FTO*), rs1805081 (*NPC1*), rs749767 (*KAT8*) and African ancestry (P=0.04, 0.04, 0.002, and 0.007, respectively), and the interactions between rs1211166 (*NTRK2*) rs2075650 (*TOMM40-APOE-APOC1*) and Latino ancestry (P=0.01 and 0.02, respectively), which did not depart from random associations after Bonferroni correction for this specific experiment (Table 4). The meta-analyses of the effect sizes of each SNP across 6 ethnicities showed that similar β coefficients to those obtained from linear mixed regression analyses (adjusted for age, sex and ethnicity) (Table 3). Two out of the 22 risk variants demonstrated evidence of heterogeneity in the effect size across ethnic groups (rs1805081 in *NPC1*, I^2=57%, P_{het}=0.04; rs611203 in *USP37*, I^2=72%, P_{het}=0.003) (Table 3). The heterogeneity of rs611203 in *USP37* was statistically significant after Bonferroni correction, but not rs1805081 in *NPC1* (P<0.004 (0.1/24)).

Regarding the GRS, each additional risk allele resulted in an average increase of 0.15 units (95% CI: 0.11-0.194) in BMI in Europeans. The same directional effects were observed in all the other ethnic groups except in Africans (Table 3). The associations between GRS and BMI were statistically significant in Europeans ($P=2.86\times10^{-12}$) and nominally significant in both South Asians (P=0.04) and Latinos (P=0.002), but not significant in the other ethnic groups (P > 0.14). The effect sizes of the GRS on BMI in six ethnic groups varied from -0.02 to 0.16. Europeans and East Asians had the largest and similar increase in BMI per risk allele, whereas Africans and South Asians had least increase in BMI per risk allele, and Latinos and Native Americans were intermediate (Table 3). The analysis of the GRS × ethnicity interaction showed that the effect sizes of GRS significantly decreased in South Asians and Africans compared to that in Europeans (Table 4). We further combined β coefficients of GRS on BMI derived from linear mixed regression across 6 ethnic populations using meta-analysis. The result showed that

the overall effect size was 0.10 units increase in BMI per risk allele and this finding was not consistent across 6 ethnicities (I^2=60%, P_{het}=0.03), summarizing what we observed for the SNPs analyzed separately. Overall, these data suggest that the SNPs may display to some extent differences in effects of genetic variants on BMI across six ethnic groups.

DISSCUSSION

In this large multi-ethnic study of 17,423 participants at high risk of obesity / dysglycemia, we observed that the mean of BMI was significantly different across ethnicities. The pattern of BMI was: Native American>African>Latino>European>South Asian> East Asian (Table 1), which is consistent with what has been described in general populations.[1, 30] Even though the studied individuals were selected to be more obese, an ethnic-dependent adiposity pattern was still observed. This supports the observation that certain ethnic groups are not only more likely to be obese but display higher severe / morbid obesity rates.[1]

Despite the BMI was significantly different across the six populations in our study, no ethnic group harbored significant enrichment or deprivation of obesity risk alleles when compared to participants of European ancestry (Supplementary Table 3). This result is in line with a previous report by Chen *et al.*[31] They did not observe any consistent pattern of risk allele frequencies for 12 obesity SNPs across 11 HapMap populations.[31] Even though most SNPs have been originally discovered in Europeans, only 4 out of 23 risk alleles displayed highest frequencies in this specific ethnic group in EpiDREAM. Thirteen out of 23 obesity risk alleles had lowest frequencies in East Asians. The low GRS value in East Asians (African>European>Latino>Native American>South Asian>East Asian) is consistent with a low propensity to obesity. On the contrary, Native Americans had the intermediate level of GRS despite the highest risk of obesity. This may suggest that additional susceptibility loci missed by

European GWAS should be more specific to highly differentiated Native American populations.[32] Therefore, the ethnic differences in risk allele frequencies call for GWAS in diverse ancestries to uncover the genetic architecture underlying obesity, as recently illustrated in East Asian and African populations.[17-19]

Overall, associations between BMI and 19 SNPs and GRS were in a consistent direction of effect with previous literature. Only did six individual SNPs and the GRS reach at least a nominal evidence of association ($P < 0.05$). Our study illustrates the difficulty to replicate association signals with small effect size issued from large GWAS meta-analyses (e.g. GIANT).[33] The results of our study support a partial generalizability of SNPs identified in European populations to other ethnic groups, in line with previous observations.[21, 22, 34, 35] For instance, Lu and Loos recently reported that 50% of BMI-associated SNPs were shared by European and East Asian ancestries.[21] Fesinmeyer *et al.* studied 8-13 BMI SNPs derived from European populations in 69,775 individuals from 6 ethnic groups.[22] Whereas 10 out of 13 SNPs were associated with BMI in European Americans, only 0-3 out of 8-13 SNPs showed nominal associations with BMI in the other ethnic groups (East Asians, African Americans, Latino Americans, Pacific Islanders, Native Americans, N between 604 and 15,415), suggesting a limited overlap in genetic architecture for obesity across diverse ethnic groups.[22] Heterogeneity analysis found that the effect sizes varied significantly in two SNPs (rs1805081 in *NPC1* and rs611203 in *USP37*) and GRS across ethnicities. For example, the effect sizes of GRS varied from a 0.016 BMI unit (kg/m^2) decrease per additional risk allele in Africans to 0.158 units increase in East Asians (Table 3). This suggests that SNPs tagging the associations with obesity traits in Europeans might not be the best proxies of causal variants in other ethnic groups (e.g. Africans and Native Americans). In addition to different risk allele frequencies and ethnic-

specific associations, allelic heterogeneity, different linkage disequilibrium patterns, or gene × gene, gene × environment interactions may explain the incomplete generalizability of all known associations across ethnicities.[36] The heterogeneity across ethnicities of *NPC1* may deserve further investigation. We studied only one polymorphism (rs1805081) in EpiDREAM but three non-synonymous SNPs (rs1805081, rs1805082, rs1788799) are in the same LD block in Europeans, their minor allele frequencies differing dramatically in different ethnic groups. It is tempting to speculate that these three coding SNPs may have accumulatively detrimental effect on *NPC1* function, meaning that haplotype rather than single SNP analyses may better capture the association.[37, 38] The generalizability of the 23 studied SNPs is partial and our data argue for the completion of large-scale GWAS meta-analyses with dense SNP arrays in multi-ethnic designs to capture the universal proxies for associations and eventually identify the causal variants.[39-41]

Some studies showed that BMI, waist circumference (WC) and waist-to-hip ratio (WHR) were similarly strong predictors of T2D,[42, 43] however, other studies demonstrated that WC and WHR were more associated with increased risk of T2D and cardiovascular diseases than BMI.[44, 45] Recent meta-analyses have identified 33 new loci associated with WHR and 19 additional loci associated with waist and hip circumferences in European ancestry individuals.[46] These loci are mostly expressed in adipose tissue, which are different from those loci associated with BMI.[16, 46] Furthermore, the prevalence of T2D and cardiovascular diseases is significantly different across different ethnic populations.[47, 48] This evidence indicates that it is important to investigate whether the WC or WHR susceptibility variants identified in Europeans are also associated with WC or WHR in other populations in future studies.

Our study has several strengths. First, it is a large prospective cohort consisting of multiple ethnicities. Second, it enriches with obese participants and has reasonable sample sizes in each ethnic population except East Asia (N=225) and Native America (N=500). However, our study also presents some limitations including an incomplete list of BMI genetic variants used to examine the generalizability of associations from Europeans to other ethnic populations and limited statistical power to replicate a GWAS association. Compared to original discovery studies, the statistical power of our study is modest. Except rs9939609 in *FTO* had a power of greater than 80% to detect a significant association, the rest had limited power as shown in Supplementary Figure 2.

Taken together, our results have shown that 19 out of the 22 selected SNPs and GRS are overall directionally associated with BMI. The effect sizes of rs1805081 in *NPC1*, rs611203 in *USP37* and GRS are not consistent across six ethnic populations. Therefore their generalizability across different ethnic groups is partial.

Conflict of interest

The authors confirm that this article content has no conflict of interest.

Acknowledgements

We acknowledge the support from EpiDREAM and DREAM investigators. We would also like to thank all the study participants and clinical collaborators for their cooperation. We also thank the reviewers for their helpful comments. David Meyre is supported by a Tier 2 Canada Research Chair. Aihua Li is supported by an Ontario Graduate Scholarship. Sonia S. Anand holds a Canada Research Chair in Ethnic Diversity and Cardiovascular Disease, and is the Michael G.

DeGroote and Heart and Stroke Foundation of Ontario Chair in Population Health, McMaster University. Hertzel C. Gerstein holds the Population Health Research Institute Chair in Diabetes Research sponsored by Aventis. Salim Yusuf holds the Heart and Stroke Chair in Cardiovascular Research.

Table 1. Baseline characteristics by ethnic group in EpiDREAM study.

	European	**South Asian**	**East Asian**	**African**	**Latino**	**Native American**	**Overall**	**P value**
N	9395	2762	225	1249	3292	500	17423	
Age, years (SD)	55.0 (10.8)	44.9 (9.4)	53.1 (11.2)	54.1 (11.0)	52.6 (11.7)	48.9 (10.9)	52.7 (11.4)	0
Female (%)	5707 (60.8)	1335 (48.4)	129 (57.3)	897 (71.8)	2197 (66.8)	348 (69.6)	10613 (60.9)	7.17×10^{-65}
BMI Mean (SD)	30.1 (6.1)	26.4 (4.4)	26.1 (4.3)	32.4 (7.0)	31.1 (6.2)	32.5 (6.4)	30.2 (6.2)	0

SD: standard deviation

Table 2. Risk allele frequencies by ethnic group and comparison across ethnicities.

SNP	Gene	Risk Allele	Major Allele	Minor Allele	European	South Asian	East Asian	African	Latino	Native American	p value*
rs1514176	*TNNI3K*	G	A	G	0.416	0.545	0.678	0.669	0.520	0.555	1.75E-188
rs6235	*PCSK1*	G	C	G	0.266	0.295	0.296	0.155	0.207	0.234	1.85E-53
rs6232	*PCSK1*	G	A	G	0.049	0.061	0.002	0.008	0.029	0.029	1.36E-39
rs2206734	*CDKAL1*	C	C	T	0.801	0.767	0.664	0.759	0.800	0.774	3.22E-20
rs2272903	*TFAP2B*	G	G	A	0.893	0.781	0.789	0.709	0.857	0.892	1.94E-195
rs1211166	*NTRK2*	A	A	G	0.807	0.722	0.808	0.671	0.800	0.866	6.28E-89
rs6265	*BDNF*	G	G	A	0.814	0.775	0.526	0.966	0.841	0.849	7.75E-154
rs1401635	*BDNF*	C	G	C	0.290	0.382	0.081	0.253	0.216	0.224	3.41E-111
rs997295	*MAP2K5*	T	T	G	0.589	0.458	0.197	0.545	0.453	0.440	2.10E-167
rs7203521	*FTO*	A	A	G	0.610	0.429	0.261	0.633	0.472	0.390	1.97E-234
rs9939609	*FTO*	A	T	A	0.417	0.329	0.178	0.489	0.336	0.236	4.42E-110
rs1805081	*NPC1*	A	A	G	0.611	0.767	0.756	0.930	0.690	0.682	1.98E-286
rs2075650	*TOMM40-APOE-APOC1*	A	A	G	0.860	0.869	0.872	0.884	0.886	0.900	3.40E-07
rs11671664	*GIPR*	G	G	A	0.893	0.891	0.626	0.884	0.909	0.889	7.16E-78
rs2984618	*TAL1*	T	G	T	0.477	0.487	0.412	0.462	0.476	0.489	3.79E-13
rs1011527	*LEPR*	A	G	A	0.096	0.136	0.121	0.108	0.127	0.141	1.32E-19
rs7605927	*POMC*	G	C	G	0.298	0.370	0.301	0.316	0.372	0.326	4.36E-34
rs611203	*USP37*	G	A	G	0.397	0.390	0.372	0.393	0.380	0.382	0.28
rs2535633	*ITIH4*	G	C	G	0.446	0.443	0.438	0.426	0.449	0.442	0.58
rs3824755	*NT5C2*	C	G	C	0.128	0.179	0.126	0.137	0.168	0.140	2.28E-23
rs7903146	*TCF7L2*	C	C	T	0.697	0.705	0.711	0.708	0.706	0.707	0.68
rs671	*ALDH2*	G	G	A	0.998	0.997	0.993	0.997	0.997	0.999	0.09
rs749767	*KAT8*	A	A	G	0.632	0.653	0.611	0.619	0.634	0.649	0.04

P-values for Chi-squared tests of risk allele frequencies across 6 ethnicities

Table 3. Associations between 23 SNPs / GRS and BMI overall and by ethnicity

SNP	Gene	β ± SE (P value)							meta-anlysis (fixed-effect)	
		European *	South Asian*	East Asian*	African*	Latinos*	Native American*	Overall**	β ± SE (P value)	heterogeneity I^2 (P value)
rs1514176	*TNNI3K*	0.21 ± 0.09 (**0.02**)	0.10 ± 0.12 (0.39)	0.42 ± 0.39 (0.30)	0.43 ± 0.29 (0.15)	0.16 ± 0.15 (0.28)	0.09 ± 0.41 (0.83)	0.18 ± 0.06 (**5×10^{-3}**)	0.18 ± 0.06 (**3.0×10^{-3}**)	0% (0.88)
rs6235	*PCSK1*	0.05 ± 0.10 (0.64)	0.11 ± 0.13 (0.40)	-0.74 ± 0.44 (0.13)	-0.44 ± 0.38 (0.26)	-0.05 ± 0.19 (0.775)	-0.13 ± 0.48 (0.79)	0.002 ± 0.07 (0.97)	0.03 ± 0.07 (0.62)	12% (0.34)
rs6232	*PCSK1*	0.17± 0.21 (0.43)	0.19 ± 0.24 (0.42)	0.18 ± 3.06 (0.95)	-0.64 ± 1.62 (0.70)	-0.26 ± 0.45 (0.57)	-0.33 ± 1.25 (0.79)	0.09 ± 0.16 (0.55)	0.12 ± 0.15 (0.42)	0% (0.95)
rs2206734	*CDKAL1*	0.36 ± 0.11 (**1.0×10^{-3}**)	0.18 ± 0.14 (0.19)	-0.27 ±0.41 (0.52)	-0.34 ± 0.32 (0.30)	0.18 ± 0.19 (0.35)	0.30 ± 0.51 (0.56)	0.24 ± 0.08 (**2.0×10^{-3}**)	0.22 ± 0.08 (**3.0×10^{-3}**)	20% (0.28)
rs2272903	*TFAP2B*	0.11 ± 0.14 (0.43)	0.10 ± 0.14 (0.49)	0.25 ± 0.51 (0.63)	0.05 ± 0.29 (0.87)	0.32 ± 0.22 (0.14)	1.07 ± 0.66 (0.11)	0.15 ± 0.09 (0.09)	0.15 ± 0.08 (0.07)	0% (0.71)
rs1211166	*NTRK2*	-0.12 ± 0.11 (0.27)	0.14 ± 0.13 (0.28)	0.54 ± 0.49 (0.30)	0.06 ± 0.29 (0.83)	0.40 ± 0.19 (**0.03**)	-0.46 ± 0.60 (0.44)	0.04 ± 0.08 (0.59)	0.06 ± 0.07 (0.39)	39% (0.15)
rs6265	*BDNF*	0.16 ± 0.11 (0.15)	0.04 ± 0.14 (0.76)	0.53 ± 0.42 (0.24)	-0.29 ± 0.74 (0.70)	0.35 ± 0.21 (0.09)	-0.65 ± 0.55 (0.24)	0.14 ± 0.08 (0.09)	0.15 ± 0.08 (**0.05**)	0% (0.48)
rs1401635	*BDNF*	0.21 ± 0.10 (**0.03**)	0.02 ± 0.12 (0.86)	0.18 ± 0.71 (0.81)	0.11 ± 0.32 (0.73)	-0.08 ± 0.19 (0.67)	-0.21 ± 0.48 (0.67)	0.10 ± 0.07 (0.59)	0.10 ± 0.07 (0.14)	0% (0.69)
rs997295	*MAP2K5*	0.06 ± 0.09 (0.52)	0.06 ± 0.12 (0.83)	0.42 ± 0.49 (0.42)	-0.15 ± 0.27 (0.58)	0.08 ± 0.15 (0.24)	-0.15 ± 0.40 (0.72)	0.04 ± 0.06 (0.49)	0.04 ± 0.06 (0.50)	0% (0.92)
rs7203521	*FTO*	0.12 ± 0.09 (0.20)	-0.10 ± 0.12 (0.39)	0.95 ± 0.45 (0.07)	0.03 ± 0.28 (0.92)	-0.18 ± 0.15 (0.24)	0.29 ± 0.40 (0.47)	0.04 ± 0.06 (0.57)	0.02 ± 0.06 (0.71)	42% (0.12)
rs9939609	*FTO*	0.63 ± 0.09 (**5.30×10^{-12}**)	0.37 ± 0.12 (**2.0×10^{-3}**)	-0.23 ± 0.55 (0.69)	0.11 ± 0.28 (0.70)	0.31 ± 0.16 (0.06)	0.15 ±0.48 (0.75)	0.48 ± 0.07 (**2.63×10^{-13}**)	0.46 ± 0.06 (**7.34×10^{-14}**)	43% (0.12)
rs1805081	*NPC1*	0.16 ± 0.09 (0.08)	-0.07 ± 0.14 (0.60)	0.74 ± 0.46 (0.14)	-1.36 ± 0.54 (**0.02**)	0.01 ± 0.16 (0.94)	-0.28 ± 0.45 (0.54)	0.07 ± 0.07 (0.35)	-0.01 ± 0.13 (0.97)	57% (**0.04**)
rs2075650	*TOMM40-APOE-APOC1*	-0.05 ± 0.13 (0.69)	-0.23 ± 0.17 (0.18)	0.003 ± 0.60 (0.99)	0.43 ± 0.43 (0.32)	0.51 ± 0.24 (**0.04**)	0.67 ± 0.66 (0.31)	0.06 ± 0.09 (0.52)	0.02 ± 0.09 (0.87)	41% (0.13)
rs11671664	*GIPR*	0.24 ± 0.14 (0.10)	0.23 ± 0.19 (0.23)	0.16 ± 0.38 (0.68)	0.85 ± 0.43 (0.06)	0.05 ± 0.27 (0.86)	0.90 ± 0.66 (0.18)	0.26 ± 0.10 (**0.01**)	0.25 ± 0.10 (**0.01**)	0% (0.62)

rs2984618	*TAL1*	0.29 ± 0.90 (**1.0×10^{-3}**)	0.08 ± 0.12 (0.49)	1.24 ± 0.81 (0.13)	0.33 ± 0.31 (0.30)	0.41 ± 0.15 (**6.0×10^{-3}**)	0.51 ± 0.40 (0.21)	0.30 ± 0.06 (**4.06×10^{-6}**)	0.24 ± 0.09 (**2.0×10^{-3}**)	4% (0.39)
rs1011527	*LEPR*	0.04 ± 0.17 (0.81)	-0.003 ± 0.13 (0.98)	0.07 ± 0.72 (0.93)	-0.42 ± 0.40 (0.30)	0.39 ± 0.30 (0.20)	1.39 ± 0.93 (0.14)	-0.06 ± 0.10 (0.56)	0.04 ± 0.09 (0.67)	0% (0.42)
rs7605927	*POMC*	0.06 ± 0.10 (0.56)	-0.12 ± 0.12 (0.32)	-0.39 ± 0.41 (0.37)	-0.12 ± 0.27 (0.66)	0.02 ± 0.16 (0.88)	0.08 ± 0.41 (0.84)	-0.01 ± 0.07 (0.85)	-0.02 ± 0.07 (0.75)	0% (0.82)
rs611203	*USP37*	0.08 ±0.09 (0.37)	-0.11 ± 0.12 (0.37)	-0.51 ± 0.48 (0.32)	-0.11 ± 0.29 (0.71)	-0.10 ± 0.16 (0.52)	0.64 ± 0.42 (0.13)	0.02 ± 0.07 (0.81)	0.06 ± 0.13 (0.67)	72% (**0.003**)
rs2535633	*ITIH4*	-0.08 ± 0.09 (0.38)	-0.01 ± 0.12 (0.93)	0.37 ± 0.43 (0.42)	-0.21 ± 0.29 (0.48)	0.17 ± 0.15 (0.27)	-0.19 ± 0.40 (0.64)	-0.03 ± 0.06 (0.61)	-0.02 ± 0.06 (0.76)	0% (0.64)
rs3824755	*NT5C2*	0.19 ± 0.15 (0.19)	0.05 ± 0.13 (0.72)	0.04 ± 0.46 (0.93)	0.20 ± 0.35 (0.57)	0.35 ± 0.20 (0.08)	0.19 ± 0.60 (0.76)	0.17 ± 0.09 (0.06)	0.16 ± 0.08 (0.06)	0% (0.89)
rs7903146	*TCF7L2*	0.38 ± 0.10 (**6.23×10^{-5}**)	0.35 ± 0.13 (**6.0×10^{-3}**)	0.59 ± 0.81 (0.49)	0.31 ± 0.30 (0.300)	0.12 ± 0.17 (0.49)	0.665 ± 0.57 (0.25)	0.32 ± 0.07 (**3.14×10^{-6}**)	0.33 ± 0.06 (**2.87×10^{-7}**)	0% (0.79)
rs671	*ALDH2*	NA	NA	NA	NA	NA	NA	NA	NA	NA
rs749767	*KAT8*	0.15 ± 0.09 (0.10)	0.13 ± 0.17 (0.47)	0.21 ± 0.50 (0.69)	-0.59 ± 0.30 (0.05)	0.07 ± 0.15 (0.65)	0.28 ± 0.39 (0.47)	0.09 ± 0.07 (0.19)	0.10 ± 0.066 (0.15)	17% (0.30)
	GRS	0.15 ± 0.02 (**2.86×10^{-12}**)	0.05 ± 0.03 (**0.04**)	0.16 ± 0.10 (0.14)	-0.02 ± 0.07 (0.82)	0.12 ± 0.04 (**2.0×10^{-3}**)	0.15 ± 0.10 (0.14)	0.12 ± 0.02 (**3.93×10^{-14}**)	0.10 ± 0.03 (**1.72×10^{-4}**)	60% (**0.03**)

SE: standard error

NA: not analyzed

* adjustment for age, sex

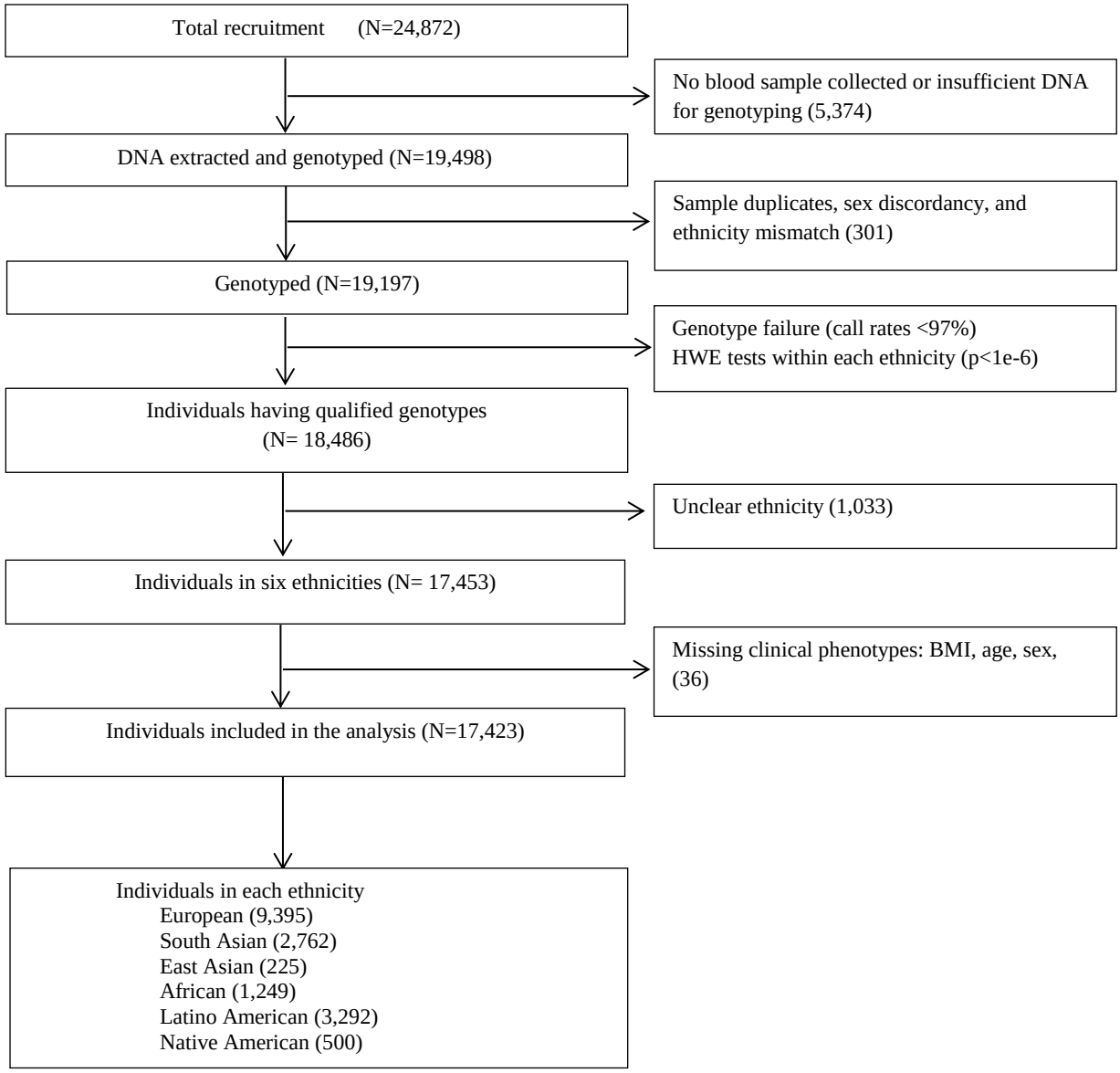
**adjustment for age, sex and ethnicity

Table 4. Interactions between SNP/GRS and ethnicity

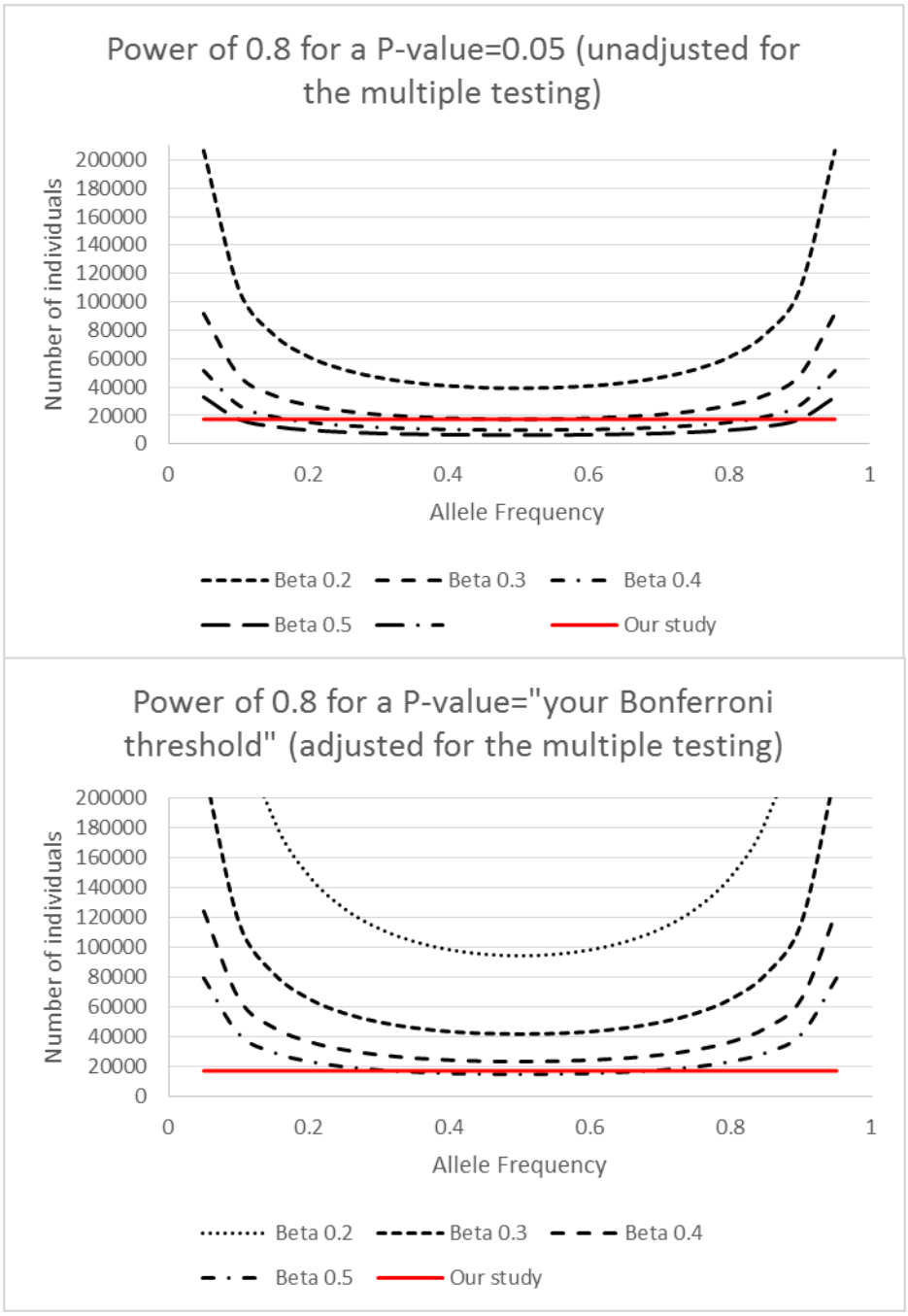
SNP	Main effect of SNP/GRS			β (P value) of interactions between SNP/GRS and ethnicity*				
	β coefficient	SE	P value	South Asian	East Asian	African	Latino	Native American
rs1514176	0.21	0.09	**0.02**	-0.13 (0.45)	0.13 (0.81)	0.18 (0.49)	-0.05 (0.75)	-0.19 (0.62)
rs6235	0.05	0.10	0.64	0.05 (0.74)	-0.83 (0.16)	-0.43 (0.21)	-0.10 (0.630)	-0.20 (0.66)
rs6232	0.17	0.20	040	0.02 (0.97)	-0.39 (0.93)	-0.99 (0.48)	-0.46 (0.34)	-0.43 (0.71)
rs2206734	0.36	0.11	**1.0×10^{-3}**	-0.14 (0.51)	-0.63 (0.28)	-0.61 **(0.04)**	-0.19 (0.38)	0.14 (0.77)
rs2272903	0.11	0.14	0.44	-0.03 (0.92)	0.11 (0.88)	-0.04 (0.89)	0.20 (0.42)	0.83 (0.18)
rs1211166	-0.12	0.11	0.28	0.28 (0.18)	0.68 (0.33)	0.08 (0.78)	0.54 **(0.01)**	-0.38 (0.50)
rs6265	0.16	0.11	0.15	-0.08(0.71)	0.30 (0.62)	-0.66 (0.31)	0.16 (0.47)	-0.83 (0.11)
rs1401635	0.21	0.09	**0.026**	-0.18 (0.33)	-0.01 (0.99)	-0.14 (0.62)	-0.30 (0.13)	-0.53 (0.23)
rs997295	0.06	0.09	0.53	0.001 (0.99)	0.33 (0.64)	-0.18 (0.48)	0.01 (0.94)	-0.13 (0.73)
rs7203521	0.12	0.09	0.17	-0.25 (0.18)	092 (0.15)	-0.13 (0.61)	-0.27 (0.11)	0.27 (0.47)
rs9939609	0.63	0.09	**3.94×10^{-13}**	-0.27 (0.15)	-0.93 (0.22)	-0.52 **(0.04)**	-0.33 (0.06)	-0.39 (0.38)
rs1805081	0.17	0.09	0.06	-0.25 (0.22)	0.73 (0.26)	-1.51 **(2.0×10^{-3})**	-0.16 (0.38)	-0.55 (0.19)
rs2075650	-0.05	0.13	0.70	-0.21 (0.44)	0.05 (0.96)	0.44 (0.26)	0.59 **(0.03)**	0.57 (0.36)
rs11671664	0.22	0.14	0.12	0.03 (0.93)	0.03 (0.95)	0.54 (0.18)	0.19 (0.51)	0.89 (0.15)
rs2984618	0.30	0.09	**1.0×10^{-3}**	-0.19 (0.29)	0.69 (0.54)	0.05 (0.86)	0.09 (0.58)	0.16 (0.67)
rs1011527	0.04	0.16	0.83	-0.05 (0.83)	0.17 (0.86)	-0.46 (0.22)	-0.45 (0.18)	1.42 (0.10)
rs7605927	0.05	0.10	0.59	-0.21 (0.27)	-0.45 (0.44)	-0.17 (0.51)	-0.05(0.78)	0.09 (0.81)
rs611203	0.07	0.09	0.41	-0.17 (0.37)	-0.50 (0.47)	-0.17 (0.52)	-0.17 (0.33)	0.60 (0.13)
rs2535633	-0.08	0.09	0.34	0.07 (0.69)	0.38 (0.53)	-0.09 (0.75)	0.23 (0.18)	-0.14 (0.70)
rs3824755	0.20	0.14	0.16	-0.18 (0.46)	-0.20 (0.76)	-0.05 (0.88)	0.12 (0.61)	-0.20 (0.73)
rs7903146	0.38	0.09	**5.07E-05**	0.02 (0.91)	0.10 (0.93)	-0.10 (0.73)	0.27 (0.14)	0.07 (0.89)
rs749767	0.15	0.09	0.09	-0.03 (0.90)	0.13 (0.85)	-0.74 **(7.0×10^{-3})**	-0.07 (0.71)	0.06 (0.87)
GRS	0.15	0.02	**3.37E-13**	**-0.10(0.02)**	0.01 (0.95)	-0.18 **(4.0×10^{-3})**	-0.04 (0.32)	-0.02 (0.86)

*European was set as reference

Supplementary Figure 1. Flowchart for participant selection and quality control



Supplementary Figure 2. Power to detect a main effect of a SNP on BMI in EpiDREAM.



The power was calculated using QUANTO software

Supplementary Table 1. Genotypes distributions of 23 BMI/obesity SNPs in each ethnicity in EpiDREAM study

Ethnicity	Genotype Counts			Risk Allele Frequency	Genotype Call Rate (%)	HWE P-Value
rs1514176 in *TNNI3K*	**GG**	**GA**	**AA**	**G**		
European	1490	4149	2929	0.416	100	0.76
South Asian	772	1288	538	0.545	100	1.00
East Asian	110	66	35	0.678	100	6.30E-05
African	504	498	124	0.669	100	0.95
Latinos	839	1420	720	0.520	100	0.01
Native American	136	220	87	0.555	100	1.00
Total	3851	7641	4433	0.482	100	-
rs6235 in *PCSK1*	**CC**	**CG**	**GG**	**C**		
European	633	3283	4651	0.266	99.99	0.11
South Asian	250	1034	1313	0.295	99.96	0.03
East Asian	17	91	103	0.296	100	0.74
African	26	297	803	0.155	100	0.91
Latinos	134	963	1882	0.207	100	0.43
Native American	22	163	258	0.234	100	0.69
Total	1082	5831	9010	0.251	99.99	-
rs6232 in *PCSK1*	**GG**	**GA**	**AA**	**G**		
European	19	799	7750	0.049	100	0.82
South Asian	15	287	2296	0.061	100	0.08
East Asian	0	1	210	0.002	100	1.00
African	0	18	1108	0.008	100	1.00
Latinos	2	168	2809	0.029	100	1.00
Native American	0	26	417	0.029	100	1.00
Total	36	1299	14590	0.043	100	-
rs2206734 in *CDKAL1*	**TT**	**TC**	**CC**	**C**		
European	318	2779	5470	0.801	99.99	0.14
South Asian	150	912	1536	0.767	100	0.35
East Asian	28	86	97	0.664	100	0.22
African	66	411	649	0.759	100	0.94
Latinos	124	946	1909	0.800	100	0.61
Native American	17	166	260	0.774	100	0.17
Total	703	5300	9921	0.789	99.99	-
rs2272903 in *TFAP2B*	**AA**	**AG**	**GG**	**G**		
European	109	1609	6850	0.893	100	0.19
South Asian	125	887	1586	0.781	100	0.95
East Asian	8	73	130	0.789	100	0.68
African	109	438	579	0.709	100	0.05
Latinos	67	716	2196	0.857	100	0.33
Native American	4	88	351	0.892	100	0.81

Ethnicity	**Genotype Counts**			**Risk Allele Frequency**	**Genotype Call Rate (%)**	**HWE P-Value**
Total	422	3811	11692	0.854	100	-
rs1211166 in *NTRK2*	**GG**	**GA**	**AA**	**A**		
European	333	2634	5599	0.807	99.98	0.30
South Asian	199	1045	1354	0.722	100	0.92
East Asian	11	59	141	0.808	100	0.18
African	122	496	507	0.671	99.91	1.00
Latinos	135	921	1923	0.800	100	0.07
Native American	10	99	334	0.866	100	0.41
Total	810	5254	9858	0.784	99.98	-
rs6265 in *BDNF*	**AA**	**AG**	**GG**	**G**		
European	322	2551	5695	0.814	100	0.09
South Asian	136	895	1567	0.775	100	0.57
East Asian	45	110	56	0.526	100	0.58
African	4	68	1054	0.966	100	0.03
Latinos	88	774	2117	0.841	100	0.10
Native American	11	112	320	0.849	100	0.71
Total	606	4510	10809	0.820	100	-
rs1401635 in *BDNF*	**CC**	**CG**	**GG**	**C**		
European	747	3482	4338	0.290	99.99	0.20
South Asian	400	1186	1011	0.382	99.96	0.10
East Asian	4	26	181	0.081	100	0.03
African	69	431	626	0.253	100	0.69
Latinos	139	1008	1832	0.216	100	1.00
Native American	23	152	268	0.224	100	0.79
Total	1382	6285	8256	0.284	99.99	-
rs997295 in *MAP2K5*	**GG**	**GT**	**TT**	**T**		
European	1457	4128	2983	0.589	100	0.66
South Asian	779	1260	559	0.458	100	0.25
East Asian	139	61	11	0.197	100	0.20
African	235	554	337	0.545	100	0.81
Latinos	919	1421	638	0.453	99.97	0.05
Native American	151	194	98	0.440	100	0.02
Total	3680	7618	4626	0.530	99.99	-
rs7203521 in *FTO*	**GG**	**GA**	**AA**	**A**		
European	1328	4022	3218	0.610	100	0.23
South Asian	855	1257	485	0.429	99.96	0.55
East Asian	115	82	14	0.261	100	1.00
African	158	511	457	0.633	100	0.44
Latinos	874	1400	704	0.472	99.97	2.0×10^{-3}
Native American	179	181	82	0.390	99.77	4.0×10^{-3}
Total	3509	7453	4960	0.546	99.98	-
rs9939609 in *FTO*	**AA**	**AT**	**TT**	**A**		
European	1539	4074	2955	0.417	100	0.04

Ethnicity	**Genotype Counts**			**Risk Allele Frequency**	**Genotype Call Rate (%)**	**HWE P-Value**
South Asian	303	1104	1191	0.329	100	0.06
East Asian	5	65	141	0.178	100	0.64
African	262	576	288	0.489	100	0.44
Latinos	360	1281	1338	0.336	100	0.05
Native American	30	149	264	0.236	100	0.19
Total	2499	7249	6177	0.3845	100	-
rs1805081 in *NPC1*	**GG**	**GA**	**AA**	**A**		
European	1317	4028	3223	0.611	100	0.32
South Asian	142	927	1529	0.767	100	0.91
East Asian	14	75	122	0.756	100	0.58
African	6	146	974	0.930	100	0.82
Latinos	305	1236	1438	0.690	100	0.10
Native American	41	200	202	0.682	100	0.44
Total	1825	6612	7488	0.678	100	-
rs2075650 in *TOMM40-APOE-APOC1*	**GG**	**GA**	**AA**	**A**		
European	168	2060	6340	0.860	100	0.96
South Asian	48	587	1963	0.869	100	0.61
East Asian	5	44	162	0.872	100	0.35
African	11	239	876	0.884	100	0.31
Latinos	43	591	2345	0.886	100	0.41
Native American	3	83	357	0.900	100	0.60
Total	278	3604	12043	0.869	100	-
rs11671664 in *GIPR*	**AA**	**AG**	**GG**	**G**		
European	113	1607	6845	0.893	99.97	0.10
South Asian	30	505	2063	0.891	100	1.00
East Asian	36	86	89	0.626	100	0.06
African	14	234	878	0.884	100	0.89
Latinos	25	492	2461	0.909	99.97	0.91
Native American	3	92	348	0.889	100	0.34
Total	221	3016	12684	0.891	99.97	-
rs2984618 in *TAL1*	**TT**	**TG**	**GG**	**T**		
European	1385	4062	3121	0.477	99.99	0.30
South Asian	530	1235	833	0.487	100	0.07
East Asian	2	18	191	0.412	100	0.10
African	78	426	622	0.462	100	0.70
Latinos	698	1369	912	0.476	100	3.16E-5
Native American	109	211	123	0.489	100	0.34
Total	2802	7321	5802	0.490	99.99	-
rs1011527 in *LEPR*	**AA**	**AG**	**GG**	**A**		
European	52	1160	7354	0.096	99.98	0.38
South Asian	234	1080	1284	0.136	99.97	0.74
East Asian	1	38	172	0.121	100	0.70
African	24	270	832	0.108	100	0.71

Ethnicity	**Genotype Counts**			**Risk Allele Frequency**	**Genotype Call Rate (%)**	**HWE P-Value**
Latinos	11	391	2576	0.127	100	0.40
Native American	2	44	397	0.141	100	0.37
Total	324	2983	12615	0.111	99.98	-
rs7605927 in *POMC*	**GG**	**GC**	**CC**	**G**		
European	525	3082	4959	0.298	99.94	0.12
South Asian	547	1296	744	0.370	99.78	0.72
East Asian	36	108	67	0.301	100	0.57
African	242	523	361	0.316	99.91	0.05
Latinos	461	1359	1153	0.372	99.72	0.07
Native American	67	190	186	0.326	100	0.12
Total	1878	6558	7470	0.326	99.87	-
rs611203 in *USP37*	**GG**	**GA**	**AA**	**G**		
European	1468	4109	2991	0.397	99.98	0.38
South Asian	279	1052	1266	0.390	100	0.01
East Asian	8	63	140	0.372	100	0.82
African	160	556	410	0.393	100	0.21
Latinos	511	1477	991	0.380	100	0.35
Native American	53	198	192	0.382	100	0.83
Total	2479	7455	5990	0.391	99.87	-
rs2535633 in *ITIH4*	**GG**	**GC**	**CC**	**G**		
European	1422	4127	3016	0.446	99.97	0.88
South Asian	428	1210	960	0.443	100	0.16
East Asian	26	100	85	0.438	100	0.77
African	112	483	531	0.426	100	0.89
Latinos	697	1469	812	0.449	100	0.51
Native American	113	215	115	0.442	100	0.57
Total	2798	7604	5519	0.444	99.98	-
rs3824755 in *NT5C2*	**CC**	**CG**	**GG**	**C**		
European	94	1526	6948	0.128	100	0.31
South Asian	164	901	1533	0.179	100	0.04
East Asian	13	92	106	0.126	100	0.31
African	42	354	730	0.137	100	1.00
Latinos	89	825	2065	0.168	100	0.55
Native American	8	108	327	0.140	100	1.00
Total	410	3806	11709	0.145	100	-
rs7903146 in *TCF7L2*	**TT**	**TC**	**CC**	**C**		
European	886	3574	4108	0.697	100	0.01
South Asian	258	1116	1224	0.705	100	0.89
East Asian	1	21	189	0.711	100	0.47
African	100	476	550	0.708	100	0.89
Latinos	245	1187	1547	0.706	100	0.42
Native American	10	131	302	0.707	100	0.40
Total	1500	6505	7920	0.701	100	-
rs671 in *ALDH2*	**AA**	**AG**	**GG**	**G**		

Ethnicity	**Genotype Counts**			**Risk Allele Frequency**	**Genotype Call Rate (%)**	**HWE P-Value**
European	0	1	8567	0.998	100	1.00
South Asian	0	1	2597	0.997	100	1.00
East Asian	12	47	152	0.830	100	0.01
African	0	3	1123	0.997	100	1.00
Latinos	0	1	2978	0.997	100	1.00
Native American	0	0	443	0.999	100	1.00
Total	12	53	15860	0.998	100	-
rs749767 in *KAT8*	**GG**	**GA**	**AA**	**A**		
European	1321	4102	3145	0.632	99.99	0.80
South Asian	48	587	1961	0.653	99.96	0.61
East Asian	10	56	145	0.611	100	0.16
African	104	453	569	0.619	100	0.32
Latinos	642	1412	924	0.634	99.97	0.02
Native American	95	213	135	0.649	100	0.57
Total	2220	6823	6879	0.635	99.98	-

Supplementary Table 2. Literature resources of the 23 SNPs selected in EpiDREAM

SNP	Gene	Proxy	Chromosome position (GRCh37/hg19)	Risk Allele	Traits	References
rs1514175	*TNNI3K*	rs1514176	chr1:74991596	G	BMI/obesity	Speliotes, Nat Genet,2010
rs6235	*PCSK1*		chr5:95729398	G	obesity	Benzinou, Nat Genet, 2008
rs6232	*PCSK1*		chr5:95752285	G	obesity	Benzinou, Nat Genet, 2008
rs2206734	*CDKAL1*		chr6:20695384	C	BMI	Wen, Nat Genet,2012; Okada, Nat Genet,2012
rs2272903	*TFAP2K5*		chr6:50787071	G	BMI	Guo, Hum Mol Genet,2013
rs1211166	*NTRK2*		chr9:87286492	A	BMI	Guo, Hum Mol Genet,2013
rs6265	*BDNF*		chr11:27680416	G	BMI	Thorleifsson, Nat Genet, 2009
rs1401635	*BDNF*		chr11:27694491	C	BMI/obesity	Thorleifsson, Nat Genet, 2009; Jiao, BMC Med Genet,2011
rs997295	*MAP2K5*		chr15:68016843	T	BMI	Guo, Hum Mol Genet,2013
rs7203521	*FTO*		chr16:53769793	A	BMI	Thorleifsson, Nat Genet, 2009
rs9939609	*FTO*		chr16:53821027	A	BMI/obesity	Fraying, Science,2007; Dina, Nat Genet, 2007
rs1805081	*NPC1*		chr18:21140932	A	obesity	Meyre, Nat Genet, 2009
rs2075650	*TOMM40-APOE-APOC1*		chr19:45396119	A	BMI	Guo, Hum Mol Genet,2013
rs11671664	*GIPR*		chr19:46172278	G	BMI	Okada, Nat Genet,2012
rs977747	*TAL1*	rs2984618	chr1:47690438	T	BMI	Locke, Nat Genet, 2015
rs11208659	*LEPR*	rs1011527	chr1:65988093	A	obesity	Wheeler, Nat Genet, 2013
rs1561288	*POMC*	rs7605927	chr2:25375905	G	BMI	Graff, Hum Mol Genet, 2013
rs492400	*USP37*	rs611203	chr2:219472325	G	BMI	Locke, Nat Genet, 2015
rs2535633	*ITIH4*		chr3:52859630	G	BMI	Wen, Hum Mol Genet, 2014
rs11191560	*NT5C2*	rs3824755	chr10:104595849	C	BMI	Wen, Hum Mol Genet, 2014
rs7903146	*TCF7L2*		chr10:114758349	C	BMI	Locke, Nat Genet, 2015
rs671	*ALDH*		chr12:112241766	G	BMI	Wen, Hum Mol Genet, 2014
rs9925964	*KAT8*	rs749767	chr16:31124407	A	BMI	Locke, Nat Genet, 2015

Supplementary Table 3. Assessment of frequencies of risk alleles in other ethnic groups compared to those in European

	South Asian	East Asian	African	Latinos	Native American
Increased	12	8	10	11	12
Decreased	11	15	13	12	11
P value*	1	0.21	0.678	1	1

* P values of two-sided binominal tests. Null hypothesis of binominal tests was that 50% of risk allele would have increased/decreased frequencies compared to those in European.

References

1. Ogden CL, Carroll MD, Kit BK, Flegal KM. Prevalence of childhood and adult obesity in the United States, 2011-2012. *JAMA : the journal of the American Medical Association* 2014; **311**(8)**:** 806-14.
2. Calle EE, Rodriguez C, Walker-Thurmond K, Thun MJ. Overweight, obesity, and mortality from cancer in a prospectively studied cohort of U.S. adults. *The New England journal of medicine* 2003; **348**(17)**:** 1625-38.
3. Finucane MM, Stevens GA, Cowan MJ, Danaei G, Lin JK, Paciorek CJ *et al.* National, regional, and global trends in body-mass index since 1980: systematic analysis of health examination surveys and epidemiological studies with 960 country-years and 9.1 million participants. *Lancet* 2011; **377**(9765)**:** 557-67.
4. Fontaine KR, Redden DT, Wang C, Westfall AO, Allison DB. Years of life lost due to obesity. *Jama* 2003; **289**(2)**:** 187-93.
5. Finkelstein EA, Trogdon JG, Cohen JW, Dietz W. Annual medical spending attributable to obesity: payer-and service-specific estimates. *Health Aff (Millwood)* 2009; **28**(5)**:** w822-31.
6. Flegal KM, Carroll MD, Kit BK, Ogden CL. Prevalence of obesity and trends in the distribution of body mass index among US adults, 1999-2010. *Jama* 2012; **307**(5)**:** 491-7.
7. Ogden CL, Carroll MD, Kit BK, Flegal KM. Prevalence of obesity and trends in body mass index among US children and adolescents, 1999-2010. *Jama* 2012; **307**(5)**:** 483-90.
8. Sankar P, Cho MK, Condit CM, Hunt LM, Koenig B, Marshall P *et al.* Genetic research and health disparities. *Jama* 2004; **291**(24)**:** 2985-9.
9. Fernandez JR, Pearson KE, Kell KP, Bohan Brown MM. Genetic admixture and obesity: recent perspectives and future applications. *Human heredity* 2013; **75**(2-4)**:** 98-105.
10. Choquet H, Meyre D. Genetics of Obesity: What have we Learned? *Current genomics* 2011; **12**(3)**:** 169-79.
11. McAllister EJ, Dhurandhar NV, Keith SW, Aronne LJ, Barger J, Baskin M *et al.* Ten putative contributors to the obesity epidemic. *Critical reviews in food science and nutrition* 2009; **49**(10)**:** 868-913.
12. Wardle J, Carnell S, Haworth CM, Plomin R. Evidence for a strong genetic influence on childhood adiposity despite the force of the obesogenic environment. *The American journal of clinical nutrition* 2008; **87**(2)**:** 398-404.
13. Elks CE, den Hoed M, Zhao JH, Sharp SJ, Wareham NJ, Loos RJ *et al.* Variability in the heritability of body mass index: a systematic review and meta-regression. *Front Endocrinol (Lausanne)* 2012; **3:** 29.
14. Choquet H, Meyre D. Molecular basis of obesity: current status and future prospects. *Current genomics* 2011; **12**(3)**:** 154-68.
15. Doche ME, Bochukova EG, Su HW, Pearce LR, Keogh JM, Henning E *et al.* Human SH2B1 mutations are associated with maladaptive behaviors and obesity. *The Journal of clinical investigation* 2012; **122**(12)**:** 4732-6.
16. Locke AE, Kahali B, Berndt SI, Justice AE, Pers TH, Day FR *et al.* Genetic studies of body mass index yield new insights for obesity biology. *Nature* 2015; **518**(7538)**:** 197-206.
17. Wen W, Cho YS, Zheng W, Dorajoo R, Kato N, Qi L *et al.* Meta-analysis identifies common variants associated with body mass index in east Asians. *Nature genetics* 2012; **44**(3)**:** 307-11.

18. Okada Y, Kubo M, Ohmiya H, Takahashi A, Kumasaka N, Hosono N *et al.* Common variants at CDKAL1 and KLF9 are associated with body mass index in east Asian populations. *Nature genetics* 2012; **44**(3)**:** 302-6.
19. Monda KL, Chen GK, Taylor KC, Palmer C, Edwards TL, Lange LA *et al.* A meta-analysis identifies new loci associated with body mass index in individuals of African ancestry. *Nat Genet* 2013; **45**(6)**:** 690-6.
20. Benzinou M, Creemers JW, Choquet H, Lobbens S, Dina C, Durand E *et al.* Common nonsynonymous variants in PCSK1 confer risk of obesity. *Nature genetics* 2008; **40**(8)**:** 943-5.
21. Lu Y, Loos RJ. Obesity genomics: assessing the transferability of susceptibility loci across diverse populations. *Genome medicine* 2013; **5**(6)**:** 55.
22. Fesinmeyer MD, North KE, Ritchie MD, Lim U, Franceschini N, Wilkens LR *et al.* Genetic risk factors for BMI and obesity in an ethnically diverse population: results from the population architecture using genomics and epidemiology (PAGE) study. *Obesity (Silver Spring)* 2013; **21**(4)**:** 835-46.
23. Gerstein HC, Yusuf S, Holman R, Bosch J, Pogue J. Rationale, design and recruitment characteristics of a large, simple international trial of diabetes prevention: the DREAM trial. *Diabetologia* 2004; **47**(9)**:** 1519-27.
24. Anand SS, Dagenais GR, Mohan V, Diaz R, Probstfield J, Freeman R *et al.* Glucose levels are associated with cardiovascular disease and death in an international cohort of normal glycaemic and dysglycaemic men and women: the EpiDREAM cohort study. *European journal of preventive cardiology* 2012; **19**(4)**:** 755-64.
25. Price AL, Patterson NJ, Plenge RM, Weinblatt ME, Shadick NA, Reich D. Principal components analysis corrects for stratification in genome-wide association studies. *Nature genetics* 2006; **38**(8)**:** 904-9.
26. Keating BJ, Tischfield S, Murray SS, Bhangale T, Price TS, Glessner JT *et al.* Concept, design and implementation of a cardiovascular gene-centric 50 k SNP array for large-scale genomic association studies. *PloS one* 2008; **3**(10)**:** e3583.
27. Janssens AC, Moonesinghe R, Yang Q, Steyerberg EW, van Duijn CM, Khoury MJ. The impact of genotype frequencies on the clinical validity of genomic profiling for predicting common chronic diseases. *Genetics in medicine : official journal of the American College of Medical Genetics* 2007; **9**(8)**:** 528-35.
28. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MA, Bender D *et al.* PLINK: a tool set for whole-genome association and population-based linkage analyses. *American journal of human genetics* 2007; **81**(3)**:** 559-75.
29. Ihaka R, Genteman R. R: a language for data analysis and graphics. *Journal of Computational and Graphical Statistics* 1996; **5:** 299-314.
30. Wang Y, Beydoun MA. The obesity epidemic in the United States--gender, age, socioeconomic, racial/ethnic, and geographic characteristics: a systematic review and meta-regression analysis. *Epidemiologic reviews* 2007; **29:** 6-28.
31. Chen R, Corona E, Sikora M, Dudley JT, Morgan AA, Moreno-Estrada A *et al.* Type 2 diabetes risk alleles demonstrate extreme directional differentiation among human populations, compared to other diseases. *PLoS genetics* 2012; **8**(4)**:** e1002621.
32. Reich D, Patterson N, Campbell D, Tandon A, Mazieres S, Ray N *et al.* Reconstructing Native American population history. *Nature* 2012; **488**(7411)**:** 370-4.

33. Li A, Meyre D. Challenges in reproducibility of genetic association studies: lessons learned from the obesity field. *Int J Obes (Lond)* 2013; **37**(4)**:** 559-67.
34. Rouskas K, Kouvatsi A, Paletas K, Papazoglou D, Tsapas A, Lobbens S *et al.* Common variants in FTO, MC4R, TMEM18, PRL, AIF1, and PCSK1 show evidence of association with adult obesity in the Greek population. *Obesity (Silver Spring)* 2012; **20**(2)**:** 389-95.
35. Graff M, North KE, Mohlke KL, Lange LA, Luo J, Harris KM *et al.* Estimation of genetic effects on BMI during adolescence in an ethnically diverse cohort: The National Longitudinal Study of Adolescent Health. *Nutrition & diabetes* 2012; **2:** e47.
36. Gabriel SB, Schaffner SF, Nguyen H, Moore JM, Roy J, Blumenstiel B *et al.* The structure of haplotype blocks in the human genome. *Science* 2002; **296**(5576)**:** 2225-9.
37. Al-Daghri NM, Cagliani R, Forni D, Alokail MS, Pozzoli U, Alkharfy KM *et al.* Mammalian NPC1 genes may undergo positive selection and human polymorphisms associate with type 2 diabetes. *BMC medicine* 2012; **10:** 140.
38. Tregouet DA, Konig IR, Erdmann J, Munteanu A, Braund PS, Hall AS *et al.* Genome-wide haplotype association study identifies the SLC22A3-LPAL2-LPA gene cluster as a risk locus for coronary artery disease. *Nature genetics* 2009; **41**(3)**:** 283-5.
39. Cooper RS, Tayo B, Zhu X. Genome-wide association studies: implications for multiethnic samples. *Human molecular genetics* 2008; **17**(R2)**:** R151-5.
40. Pulit SL, Voight BF, de Bakker PI. Multiethnic genetic association studies improve power for locus discovery. *PloS one* 2010; **5**(9)**:** e12600.
41. Li YR, Keating BJ. Trans-ethnic genome-wide association studies: advantages and challenges of mapping in diverse populations. *Genome medicine* 2014; **6**(10)**:** 91.
42. Vazquez G, Duval S, Jacobs DR, Jr., Silventoinen K. Comparison of body mass index, waist circumference, and waist/hip ratio in predicting incident diabetes: a meta-analysis. *Epidemiologic reviews* 2007; **29:** 115-28.
43. Qiao Q, Nyamdorj R. Is the association of type II diabetes with waist circumference or waist-to-hip ratio stronger than that with body mass index? *European journal of clinical nutrition* 2010; **64**(1)**:** 30-4.
44. Yusuf S, Hawken S, Ounpuu S, Bautista L, Franzosi MG, Commerford P *et al.* Obesity and the risk of myocardial infarction in 27,000 participants from 52 countries: a case-control study. *Lancet* 2005; **366**(9497)**:** 1640-9.
45. Wang Y, Rimm EB, Stampfer MJ, Willett WC, Hu FB. Comparison of abdominal adiposity and overall obesity in predicting risk of type 2 diabetes among men. *The American journal of clinical nutrition* 2005; **81**(3)**:** 555-63.
46. Shungin D, Winkler TW, Croteau-Chonka DC, Ferreira T, Locke AE, Magi R *et al.* New genetic loci link adipose and insulin biology to body fat distribution. *Nature* 2015; **518**(7538)**:** 187-96.
47. Spanakis EK, Golden SH. Race/ethnic difference in diabetes and diabetic complications. *Current diabetes reports* 2013; **13**(6)**:** 814-23.
48. Yusuf S, Reddy S, Ounpuu S, Anand S. Global burden of cardiovascular diseases: Part II: variations in cardiovascular disease by specific ethnic groups and geographic regions and prevention strategies. *Circulation* 2001; **104**(23)**:** 2855-64.

CHAPTER III: PARENTAL AND CHILD GENETIC CONTRIBUTION TO OBESITY TRAITS IN EARLY LIFE BASED ON 83 LOCI VALIDATED IN ADULTS: THE FAMILY STUDY

Aihua Li[1], Sébastien Robiou-du-Pont[1], Sonia S. Anand[1,2], Katherine M. Morrison[2,3], Sarah D. McDonald[1,4], Stephanie A. Atkinson[3], Koon K. Teo[1,2], and David Meyre[1,5] *

[1]Department of Clinical Epidemiology and Biostatistics, McMaster University, Hamilton, ON, Canada

[2]Department of Medicine, McMaster University, Hamilton, ON, Canada

[3]Department of Pediatrics, Hamilton Health Sciences and McMaster University, Hamilton, ON, Canada

[4]Department of Obstetrics and Gynecology, McMaster University, Hamilton, ON, Canada

[5]Department of Pathology and Molecular Medicine, McMaster University, Hamilton, ON, Canada.

ABSTRACT

The effect of recently identified single nucleotide polymorphisms (SNPs) associated with adult body mass index (BMI) on excess body weight in early life is unclear. This study aimed to investigate the parental and child contribution of 83 SNPs to obesity-related traits in children from birth to 5 years old in the FAMILY cohort. A total of 1,402 individuals (541 children, 541 mothers, and 320 fathers) were genotyped for 83 SNPs associated with adult BMI. An unweighted genetic risk score (GRS) was generated by the sum of BMI-increasing alleles. Repeated weight and length/height were measured at birth, 1, 2, 3, and 5 years of age, and age-, sex-specific weight and BMI Z-scores were computed. Multiple linear and mixed-effects regression models were used in genetic association tests, adjusting for relevant covariates (maternal pre-pregnancy BMI, gestational weight gain, parity, gestational diabetes, smoking status and ethnicity). Cross-sectional analyses showed that the GRS was significantly associated

with birth weight Z-score (β±SE, 0.019±0.009/allele; P=0.026). It was also associated with weight and BMI gain Z-score between birth and 2 years old and between birth and 5 years old. In longitudinal analyses, the GRS was associated with weight and BMI Z-score from birth to 5 years (β±SE, 0.019±0.007/allele; P=0.01 and 0.017±0.007/allele; P=0.011, respectively). The maternal effect of rs3736485 in *DMXL2* on offspring weight and BMI Z-score from birth to 5 years was significantly greater compared to paternal effect (Z-test $P=1.20\times10^{-6}$ and $P=1.55\times10^{-5}$, respectively). Therefore, SNPs contributing to adult BMI start to exert their effect at birth and in early childhood. Parent of origin effects may occur in a limited subset of obesity predisposing SNPs.

INTRODUCTION

The prevalence of overweight and obesity among children aged 5 to 17 years in Canada was 31% in 2012-2013.[1] Childhood obesity has been linked to early puberty, type 1 and type 2 diabetes, hypertension, poor mental and physical health during childhood, as well as adult obesity and its associated comorbidities.[2] The total health costs were 21% higher in children with obesity compared to their normal weight counterparts in Canada.[3] Therefore, it is urgent to understand the determinants of obesity in early life in order to manage the burden of childhood obesity and prevent adult obesity more efficiently.

The risk factors for childhood obesity are multifactorial, including environmental, behavioral and genetic components and their interactions.[2] Temporal effects also influence susceptibility to the development of obesity, the critical windows being defined as gestation, early infancy, adiposity rebound (5-7 years old) and adolescence.[4] Parental BMI, socioeconomic status, mother's gestational weight gain, gestational smoking status, and child's birth weight, weight gain during infancy and infant nutrition are all significant predictors of childhood/adolescent obesity.[2, 5] Rare deleterious mutations and chromosome deletions lead to monogenic and early-onset obesity in childhood.[6] Meanwhile, the total number of common genetic variants, single nucleotide polymorphisms (SNPs), associated with adult BMI at genome-wide significance level ($P<5\times10^{-8}$) has recently increased to 116.[7] Evidence from cross-sectional studies and longitudinal birth cohorts indicates that some of the adult BMI loci affect BMI in childhood and adolescent.[8-13] Recent meta-analyses of genome-wide association studies (GWAS) not only corroborated the effects of some genetic loci underlying adult BMI in European children, adolescents and young adults, but also identified four novel loci (*FAM120AOS*, *ELP3*, *RAB27B* and *ADAM23*) for childhood BMI.[14-16] Furthermore, the effects of BMI SNPs/genetic risk score on the variation of birth weight and weight gain during early childhood were not

certain.[8, 9, 17] Taken together, these studies greatly improved our knowledge of genetic contribution to childhood obesity; there are limitations to this evidence. First, many recently identified adult BMI risk variants have not been investigated in children (previous studies examined at most 32 SNPs identified until 2010). Second, several important maternal variables during pregnancy such as maternal prepregnancy BMI, gestational age, gestational weight gain, gestational diabetes, parity, and smoking status were not accounted for when examining the association between genetic risk score and birth weight. Third, the effects of maternal and paternal genotypes on childhood adiposity have not been thoroughly investigated.[18, 19]

In response, we aimed to investigate the effects of genetic risk score (GRS) combing 83 SNPs robustly associated adult BMI on birth weight, weight/BMI gain and growth trajectory from birth to 5 years of age in children using the longitudinal FAMILY birth cohort. We also hypothesized that parental risk alleles in specific genes contributed to child's weight and BMI in early life. Whether the extent of association of parental risk alleles on child weight and BMI variation depends on the transmission from the mother or the father was explored by comparing the effects of each SNP between the parents.

PARTICIPANTS AND METHODS

Participants

The Family Atherosclerosis Monitoring In earLY life (FAMILY) study was designed to longitudinally examine the fetal and early childhood determinants for the development of adiposity, cardiovascular diseases and atherosclerosis in childhood and has been described in detail previously.[20] Briefly, 857 families including 901 newborns, 857 mothers and 530 fathers were enrolled from three hospitals in the great Hamilton region (Ontario, Canada) from 2004 to

2009 and were followed for up to 5 years, with a planned follow-up for 10 years or more. We selected only singletons into our study because multiple births (n=85) have a strong impact on birth weight and postnatal growth velocity. Informed consent was obtained from all the adult participants, and the parents provided consent for their children. All experiments were performed in accordance with relevant guidelines and regulations. The research ethics boards at Hamilton Health Sciences and St Joseph's Health Center in Hamilton, and Joseph Brant Memorial Hospital in Burlington, Ontario, Canada approved the FAMILY study.

Genotyping

Genomic DNA of all the participants was extracted from buffy coats and genotyping was conducted using the Illumina Cardio-Metabochip (San Diego, CA, USA). Eighty-three selected SNPs met two criteria of having genome-wide significant associations with adult BMI ($P<5\times10^{-8}$) and being available on the Cardio-Metabochip array (lead SNP or proxy) (Table S1). Five proxy SNPs were identified using the 1000 Genomes data of European population via the Broad Institute website tool SNAP (SNP Annotation and Proxy Search) (Table S1). To ensure all 83 SNPs were independent from each other, pairwise linkage disequilibrium in European population were examined using the 1000 Genomes Project data and all the pairwise r^2 were less than 0.1. Standard procedures were conducted to assess the quality of genotyping. All 83 SNPs had call rates ranging from 95.7-100% and met with Hardy Weinberg Equilibrium ($P>0.001$). Genotyping data also found that 26 individuals had SNP missingness rates >10%, 16 individuals were from 5 families having non-biological fathers, and 11 individuals had sex-discordance between reported and SNP-identified sex; all of these were excluded from the analysis. Six pairs of individuals had cryptic relatedness (2^{nd} degree relatives) and one of each pair was randomly

selected for analysis. The ethnicity reported by participants was verified by principal component analysis (PCA) using EIGENSTRAT and the first 10 components were used to adjust for population stratification. A majority of the FAMILY participants were white Caucasians (92.8% mothers, 89.3% fathers and 91.1% children).

Phenotypes

Child weight and length/height were measured at birth, 1, 2, 3 and 5 years of age by trained staff with standard measure scales and the age at measurement were recorded. Maternal age, height, self-reported prepregnancy weight, gestational weight gain, parity, gestational diabetes and gestational age at birth were obtained either from the mothers at the baseline visit or from the medical chart review. BMI was calculated as weight in kilograms divided by length/height in meters squared. Individual weight, length/height and BMI in children were converted to age- and sex-specific Z-scores using World Health Organization growth reference (2006) (http://www.who.int/childgrowth/standards/en/). Child weight/BMI gain in Z-scores at 1, 2, 3 and 5 years were assessed as the differences between weight/BMI Z-score at 1, 2, 3 and 5 years of age and weight/BMI Z-score at birth. Maternal gestational weight gain was assessed as the difference between the last measured weight prior to delivery and self-reported weight before pregnancy. Gestational diabetes mellitus, defined as "any degree of glucose intolerance with onset or first recognition during pregnancy", was diagnosed according to the International Association of Diabetes and Pregnancy Study Groups (IADPSG) criteria using a 75 g oral glucose tolerance test (OGTT).[21] Children whose birth weight was less than 1 kg and greater than 7 kg were excluded. In total, 1,402 individuals were included in data analyses (541 unrelated children, 541 mothers and 320 fathers) (See flow chart in Figure S1). The numbers of missing

values for each of the exposures and outcomes are different and the numbers for the available data on each exposure and outcome for this study are presented in Table 1. For each association test, maximal data available were included and the numbers included in each analysis are reported in the tables.

Statistical analysis

We chose the risk alleles for each of the 83 SNPs as previously reported in the literature (Table S1). An additive mode of inheritance was applied to each SNP to code three genotypes as 0, 1 and 2 designating the number of the BMI increasing allele. A GRS was calculated by adding up the risk alleles of 83 SNPs and therefore the theoretical scores ranged from 0 to 166. We used an unweighted GRS as recommended by Dudbridge.[22] The values for the missing genotypes were imputed with arithmetic average of the coded genotypes from individuals who were successfully genotyped.

The effects of GRS on cross-sectional associations with weight or BMI Z-score at birth, 1, 2, 3 and 5 years of age were tested using multiple linear regression models. Each model was adjusted for relevant covariates including maternal prepregnancy BMI and the first 10 PCA components for ethnicity, and the associations between the GRS and birth weight/BMI Z-score were adjusted for additional covariates to eliminate the intrauterine effect including gestational age, gestational weight gain, parity, gestational diabetes and maternal smoking status. The GRS was also tested for associations with weight gain and BMI gain Z-score during the first five years. Linear mixed-effects models were used to investigate the effect of the GRS on the overall change of weight and BMI Z-score during 0-5 years. This approach was selected since it takes the correlations between repeated measures on the same individual into account and allows for

missing measurement data and measurement at different time points, assuming the missing events are random.[23] The weight and BMI Z-score at birth of each individual (intercept) and correlation among measurements on the same subject (slope of age) were modeled as random effects, and time (in years), GRS, maternal prepregnancy BMI and ethnicity were modeled as fixed effects. Interaction term, GRS×time, which tests if the effects of the GRS on weight and BMI Z-score change over time were also initially added to the linear mixed-effects regression model. Due to its non-significance, this term was removed from the final analysis to produce the most parsimonious model.

If the effects of genetic variants on phenotypic variation depend upon the parent from whom the variant is inherited, parent-of-origin effects occur.[24] Imprinting is considered as the underlying mechanism of parent-of-origin effects and it has been confirmed in less than 1% of human genes.[25] As a result, the GRS as an overall measurement of all BMI-associated SNPs was not expected to have a parent-of-origin effect and we therefore performed an exploratory analysis of the parent-of-origin effect on each SNP in addition to the GRS. To assess the effects of maternal/paternal SNPs on the child weight and BMI Z-score patterns, linear mixed-effects regression was also performed by using the maternal/paternal SNP and child's SNP and ethnicity as fixed effects and the weight/BMI Z-score at birth of each individual (intercept) and correlation among measurements on the same subject (slope of age) as random effects. Existence of a parent-origin-effect on a specific SNP was tested by a Z-test which compared the effect sizes between maternal and paternal genetic variants. Considering the low statistical power of this exploratory test, we did a sensitive analysis in which only families with complete data from mother, father and child were included. This approach would substantially reduce the bias.

Bonferroni corrected P-values are routinely applied to exploratory genetic association studies. However, they are overly conservative given the high prior likelihood of association in post-GWAS experiments. Previous studies reported the associations between GWAS BMI genetic variants and weight and BMI during childhood without applying a Bonferroni correction.[8, 9, 12, 17] Therefore, a two-tailed α-level of 0.05 was considered significant for the analyses of associations between GRS and child obesity-related traits. A P-value less than the threshold corrected for Bonferroni procedures was considered statistically significant ($P<1.5\times10^{-4}=0.05/(4*84)$) when exploratory analyses of parent-origin effects were applied. All the statistical analyses were performed using PLINK (version 1.07) and R (version 2.15.2).[26, 27]

RESULTS

Participant characteristics

The characteristics of children included in the analysis are shown in Table 1. Among the 816 singletons, genotyping data were available in 541 children. The average of birth weight in the analyzed individuals was 3.4 kg (SD=0.5). The mean of BMI at birth was 13.5±1.3 kg/m^2. The average gestational age at birth in included children was 39.4±1.5 weeks.

Associations of the child GRS with weight from birth to 5 years of age

The GRS based on 83 adult BMI-associated SNPs ranged from 59 to 94 (Mean±SD, 78.66±5.74) and it demonstrated a normal distribution. Cross-sectional analyses indicated that the GRS was associated with higher birth weight Z-score (β±SE: 0.019±0.009/allele; P=0.026), but not associated with weight at 1 year. The GRS was also associated with higher weight Z-score at 2, 3 and 5 years of age ($0.019\leq\beta\leq0.026$, $0.005\leq P\leq0.022$) (Figure 1A). The GRS was not

associated with weight gain Z-score between birth and 1 year and between birth and 3 years. In contrast, it was associated with weight gain Z-score between birth and 2 years (β±SE: 0.016±0.007/allele; P=0.035) and between birth and 5 years (β±SE: 0.021±0.009/allele; P=0.020) (Table 2). Using longitudinal analysis there was an association of the GRS with weight Z-score from birth to 5 years of age (β±SE: 0.019±0.007/allele; P=0.010) (Table 3).

Associations of the child GRS with BMI from birth to 5 years of age

Cross-sectional analyses showed that the GRS was not associated with birth BMI Z-score (β±SE: 0.015±0.010/allele; P=0.123). Similar to weight Z-score, there was no association between the GRS and BMI at 1 year, but associations of the GRS with BMI at 2, 3 and 5 years were observed ($0.018 \leq \beta \leq 0.028$, $0.004 \leq P \leq 0.040$) (Figure 1B). The GRS was not associated with BMI gain Z-score between birth and 1 year and between birth and 3 years. It was associated with BMI gain Z-score between birth and 2 years (β±SE: 0.017±0.008/allele; P=0.03) and birth and 5 years old (β±SE: 0.025±0.009/allele; P=0.007) (Table 2). An association between the GRS and BMI Z-score from birth to 5 years of age (β±SE: 0.017±0.007/allele; P=0.011) was identified in the longitudinal analysis (Table 3).

Associations of the parental SNPs/GRS with weight variation from birth to 5 years of age

After controlling for the contribution of child corresponding SNPs, the longitudinal analyses showed that only maternal risk allele for rs3736485 in *DMXL2* was significantly positively associated with weight Z-score (β±SE: 0.252±0.063/allele, $P=6.16\times10^{-5}$), whereas paternal risk alleles for rs3736485 was nominally negatively associated with weight Z-score (β±SE: -0.285±0.095/allele, $P=2.66\times10^{-3}$) from birth to 5 years of age. The difference in the effects of rs3736485 between mothers and fathers were statistically significant ($P=1.20\times10^{-6}$)

after Bonferroni correction, indicating a parent-of-origin effect (Table 4). The GRS had no parent-of-origin effect on childhood weight from birth to 5 years of age (P=0.107). The sensitivity analysis with complete family data showed that the parent-of-origin effect was stronger ($P=9.01\times10^{-8}$) (Table 4).

Associations of the parental SNPs/GRS with BMI variation from birth to 5 years of age

After controlling for the contribution of child corresponding SNPs, the longitudinal analyses showed that maternal risk allele for rs3736485 was nominally positively associated with BMI Z-score (β±SE: 0.212±0.059/allele, $P=3.07\times10^{-4}$), whereas paternal risk alleles for rs3736485 was nominally negatively associated with BMI Z-score (β±SE: -0.221±0.086/allele, $P=9.97\times10^{-3}$) from birth to 5 years of age. The maternal association of rs3736485 with increased child BMI Z-score was the only one to be statistically significant after Bonferroni correction compared to the paternal effect ($P=1.55\times10^{-5}$), indicating a parent-of-origin effect of rs3736485 (*DMXL2*) in the development of childhood obesity (Table 4). The GRS had no parent-of-origin effect on childhood BMI Z-score from birth to 5 years of age (P=0.216). The sensitivity analysis showed that the parent-of-origin effect became stronger ($P=8.66\times10^{-7}$) with complete family data (Table 4).

DISCUSSION

In the present study, we demonstrate that the GRS based on 83 adult BMI GWAS SNPs is associated with birth weight. This finding suggests that SNPs previously associated with adult BMI start to have an effect on body composition during fetal growth. These data may support, to some extent, the link between high birth weight and subsequent increased risk of obesity.[28] It is

important to note that this relationship was observed after adjustment for newborn sex, age at measurement, gestational age at birth, ethnicity, mother's prepregnancy BMI, gestational weight gain, parity, gestational diabetes, and smoking status. This result aligns with observations by Elks *et al.* of a borderline association (P=0.05) between a genetic risk score of 11 BMI variants and birth weight using the 1946 British birth cohort (N=2,537).[12] Whereas alterations in the intrauterine environment play important roles in birth weight, 10-40% of the variation in birth weight may be explained by inherited factors.[29, 30] Other studies that did not detect an association between BMI genetic variants and birth weight may be due to a small subset of adult BMI SNPs being examined and not accounting for some critical confounding factors.[8, 9, 17, 31] GWAS for birth weight have demonstrated that other genetic variants, other than BMI loci, influence fetal growth.[32, 33] The GRS examined in our study covered more of the currently identified BMI adult variants (N=83 SNPs) compared to previous studies (N≤32 SNPs).

The findings of the associations between the GRS and weight and BMI at 2, 3, and 5 years in cross-sectional analyses are consistent with the results of several other studies.[8-10, 12, 13, 17] The increasing trend of the effects on BMI (Figure 1) also supports the increasing heritability of BMI over childhood.[34] Epidemiologic studies have shown that rapid early weight gain predicts later obesity and metabolic diseases.[35, 36] Rapid weight gain may occur at any stage, but the greatest variation is commonly seen in the first 1-2 years of life.[35] This may be related to influences from the intrauterine environment and then children return to their genetic trajectory around two years old.[35] In our study, the GRS was associated with both weight and BMI gain between birth and 2 and 5 years, which has been shown in other studies even in infancy as early as 6 weeks and 3 years after birth.[8, 9]

Obesity typically develops over a long period of time. Using longitudinal analyses of repeated measurements of weight and height may identify some specific developmental windows during which the GRS shows stronger associations with child growth. Such questions cannot be answered from cross-sectional analyses. In our study, longitudinal analysis of five measurements from birth to 5 years showed that children at higher genetic risk had higher overall BMI. This is in line with three independent longitudinal studies. Elks *et al.* found that the GRS of 8 or 11 adult BMI loci was positively associated with weight and BMI from birth to 11years in two cohorts.[8, 12] Belsky *et al.* reported that the GRS derived from 32 BMI loci predicted higher BMI across childhood (ages 3 through 13 years) and adulthood (ages 13 through 38 years).[9]

GWAS assumes that the maternal and paternal alleles have equal effect on the phenotype variation. But in some circumstances the phenotypic variation caused by genetic variants may depend on parental origin. It has been observed that the parental-of-origin effects are involved in fetal and placental growth and function and have also been reported to be associated with the development of obesity.[18, 37] Recently Hoggart *et al.* developed a novel method and detected two paternal risk alleles in *SLC2A10* and *KCNK9* that increased BMI compared to the respective maternal alleles using genome-wide genotype data of unrelated individual.[18] In our study, the parent-of-origin effect of rs3736485 in *DMXL2* on BMI remained significant after Bonferroni correction. Maternal rs3736485 risk allele increased BMI whereas the same allele was protective when inherited from father, as previously observed for SNPs having strong parent-of-origin effects with type 2 diabetes.[38] Dmx-like 2 or Rabconnectin-3, encoded by *DMXL2*, is involved in regulation of the Notch signaling pathway.[39] Recently, an *in vivo* study reported that the Notch pathway inhibited brown adipocytes in white adipose tissue and therefore decreased energy expenditure and deteriorated obesity.[40] There is no evidence showing that the variant in *DMXL2*

activates the Notch signaling pathway and promotes obesity. *DMXL2* is located on chromosome 15q21.2, not close to any known imprinted genes (http://www.geneimprint.com/). There has been no report of *DMXL2* on imprinting and epigenetic modification, therefore, this parent-of-origin effect warrants further replication in an independent study.

Our study had several strengths. First, our longitudinal birth cohort, with weight and BMI at birth, 1, 2, 3, and 5 years of age and critical confounding factors influencing intrauterine environment, enabled us to more accurately investigate the effect of GRS on the weight and BMI at birth and in early childhood.[23] Second, the GRS derived from an updated list of 83 SNPs provide the most current state of evidence (until February 2015). Third, a family-based design with genotypes from both parents and child which are not available in most birth cohort studies permitted us to investigate the potential parent-of-origin effects of genetic variants on weight and BMI in early childhood. This study also had some limitations. First, although the longitudinal nature of our data increased the statistical power, our study had a relatively modest sample size, which increased the chance of random sampling error and false positive associations. Thus, the parent-of-origin effect of rs3736485 in *DMXL2* needs to be replicated in a large family-based birth cohort. Second, we do not have sufficient power to study the impact of individual SNPs on BMI early in life as done by others.[11] Third, we had fewer fathers than mothers having genotype data, which may bias the analysis of parent-of-origin effects.

In summary, we provide evidence that the GRS derived from 83 adult BMI SNPs is associated with weight at birth. The GRS also predicts the overall weight and BMI from birth and 5 years of age, suggesting adult BMI SNPs start to exert effect in early childhood. Parent-of-origin effects may occur in a limited subset of obesity predisposing SNPs, providing new insights of the mechanisms underlying the childhood obesity.

Achnowledgements

This study was funded by the Heart and Stroke Foundation of Ontario (grant # NA 7293 "Early genetic origins of cardiovascular risk factors"). AL is funded by the Ontario Graduate Scholarship. SSA holds the Heart and Stroke Foundation of Ontario, Michael G. DeGroote endowed Chair in Population Health and a Canada Research Chair in Ethnicity and Cardiovascular Disease. DM holds a Canada Research Chair in Genetics of Obesity, and SDM is supported by a Canada Research Chair in Maternal and Child Obesity Prevention and Intervention. We would also like to thank Hudson Reddon for his comments and help in proofreading of this manuscript.

Conflict of interest

The authors confirm that this article content has no conflict of interest.

Table 1. Characteristics of participants in FAMILY study.

Measurements	Boys (n=273)			Girls (n=268)			Total (n=541)		
	n	Mean	SD	n	Mean	SD	n	Mean	SD
Weight (kg)									
Birth	247	3.5	0.5	251	3.3	0.5	498	3.4	0.5
1 y	250	10.5	1.3	238	9.8	1.2	488	10.2	1.3
2 y	239	13.1	1.5	234	12.5	1.5	473	12.8	1.5
3 y	220	15.2	1.7	222	14.7	1.8	442	15.0	1.8
5 y	179	19.6	2.6	184	19.5	3.0	363	19.6	2.8
Length/Height (cm)									
Birth	266	50.7	2.3	259	49.8	2.2	525	50.3	2.3
1 y	246	77.0	3.3	237	74.9	3.0	483	76.0	3.3
2 y	235	88.8	3.5	233	87.5	3.4	468	88.2	3.5
3 y	216	96.6	4.0	221	95.3	3.7	437	95.9	3.9
5 y	179	111.1	5.3	184	110.1	4.5	363	110.6	4.9
BMI (kg/m^2)									
Birth	244	13.5	1.3	245	13.4	1.4	489	13.5	1.3
1 y	246	17.7	1.4	237	17.4	1.4	483	17.5	1.4
2 y	235	16.5	1.2	231	16.3	1.3	466	16.4	1.3
3 y	216	16.3	1.2	221	16.2	1.3	437	16.2	1.2
5 y	179	15.8	1.3	184	16.1	1.9	363	15.9	1.6
Gestational age at birth (week)									
	273	39.4	1.5	268	39.4	1.5	541	39.4	1.5
Maternal gestational weight gain (kg)									
	195	15.1	5.0	207	15.3	5.7	402	15.2	5.3
Mother prepregnancy BMI (kg/m^2)									
	196	26.4	6.6	216	26.9	6.8	412	26.7	6.7
Father BMI (kg/m^2)									
	114	28.0	5.1	126	28.9	4.8	240	28.5	4.9

Table 2. Cross-sectional associations between the GRS and weight and BMI gain Z-score at different ages.

Trait	0-1 year			0-2 year			0-3 year			0-5 year		
	N	β (SE)	P	N	β (SE)	P	N	β (SE)	P	N	β (SE)	P
Weight	445	0.014 (0.007)	0.059	431	0.016 (0.007)	**0.035**	402	0.013 (0.007)	0.068	332	0.021 (0.009)	**0.020**
BMI	434	0.006 (0.007)	0.411	417	0.017 (0.008)	**0.03**	391	0.013 (0.007)	0.071	326	0.025 (0.009)	**0.007**

The associations between the GRS and weight/BMI Z-score change were adjusted for age interval between measurements, birth weight/birth BMI Z-score and ethnicity.

Table 3. Longitudinal linear mixed modeling of the associations between the GRS and overall changes in weight and BMI Z-score from birth to 5 years of age.

Trait	N	β±SE	P
Weight	411 (1757*)	0.019 ± 0.007	**0.010**
BMI	410 (1733*)	0.017 ± 0.007	**0.011**

* The numbers indicated the total number of measurements in the longitudinal data.
In the linear mixed-effects models, the weight and BMI Z-scores at birth of each individual (intercept) and correlation among measurements on the same subject (slope of age) were modeled as random effects, and GRS, ethnicity and mother's prepregnancy BMI were modeled as fixed effects.

Table 4. Parent-of-origin effects of rs3736485 in *DMXL2* on childhood obesity traits (overall effects from birth to 5 years of age).

Obesity traits	Maternal			Paternal			Z-test P value
	N	β±SE	P value	N	β±SE	P value	
Weight	411 (1752*)	0.252±0.063	**6.16×10^{-5}**	216 (947*)	-0.285±0.095	**2.66×10^{-3}**	**1.20×10^{-6}**
BMI	410 (1733*)	0.212±0.059	**3.07×10^{-4}**	216 (937*)	-0.221±0.086	**9.97×10^{-3}**	**1.55×10^{-5}**
Sensitive analysis							
Weight	216 (947*)	0.393±0.089	**9.51×10^{-6}**	216 (947*)	-0.285±0.095	**2.66×10^{-3}**	**9.01×10^{-8}**
BMI	216 (937*)	0.340±0.080	**2.13×10^{-5}**	216 (937*)	-0.221±0.086	**9.97×10^{-3}**	**8.86×10^{-7}**

* The total number of measurements in the longitudinal analyses.
Linear mixed-effects regression was performed to assess the overall effects of maternal/paternal SNPs on the child's weight and BMI Z-score from birth to 5 years old, by using the maternal/paternal genetic variants and child's genotypes and ethnicity as fixed effects and the weight and BMI Z-score at birth of each individual (intercept) and correlation among measurements on the same subject (slope of age) as random effects. The comparison of the effect sizes between maternal and paternal genetic variants was tested using Z-test.

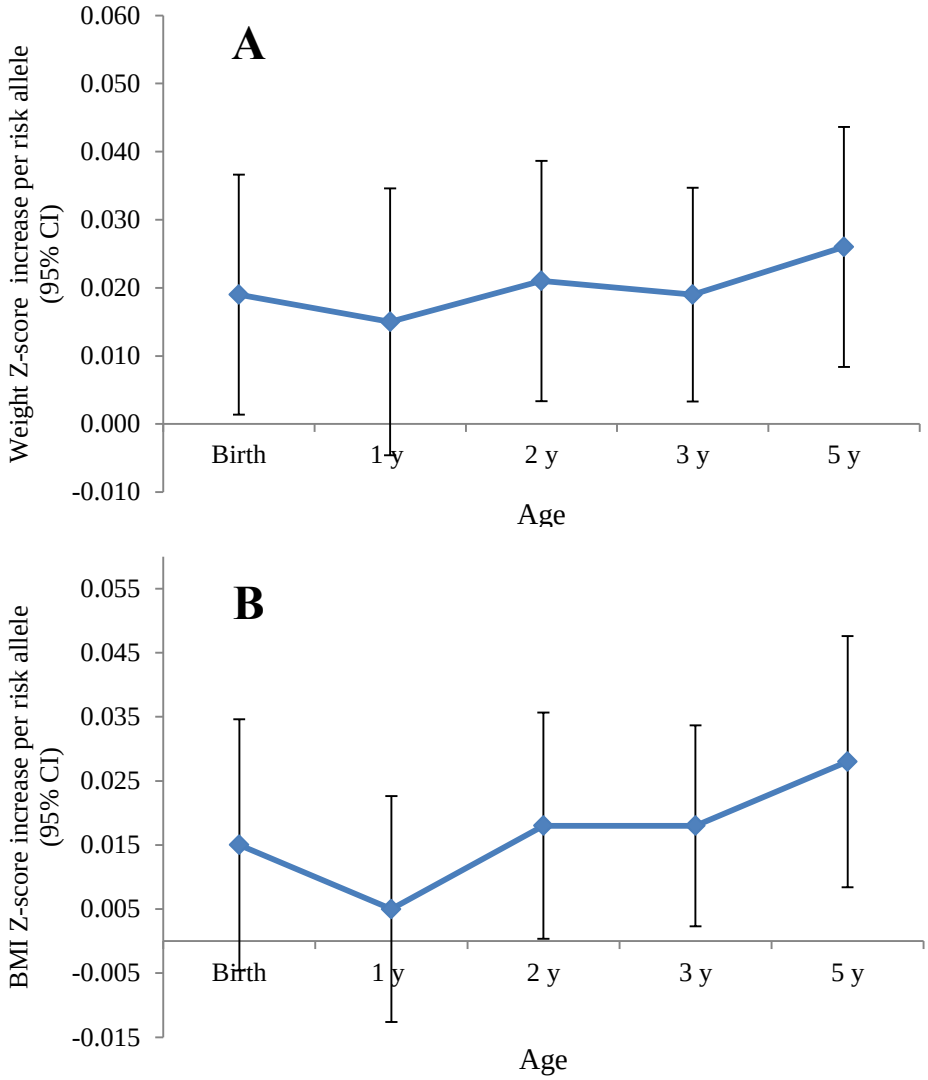


Figure 1. Longitudinal associations between the genetic risk score (GRS) and (A) weight Z-score and (B) BMI Z-score from birth to 5 years old. Regression coefficients ±95% CI are shown from multiple linear regression models. Birth weight Z-score and birth BMI Z-score were adjusted for child ethnicity, gestational age at birth, maternal prepregnancy BMI, gestational weight gain, parity, gestational diabetes, and smoking status. Weight and BMI Z-score at other ages were adjusted for child ethnicity and maternal prepregnancy BMI.

Supplementary Table 1. Characteristics of the 83 SNPs associated with adult BMI variation.

Nearest Gene	SNP	Proxy	Chr	Effect/Other Allele	Risk Allele Frequency*	Call Rate	Genotype		HWE test
TAL1	rs977747		1	T/G	0.409	100%	259/727/532	TT/TG/GG	0.948
AGBL4	rs657452		1	A/G	0.421	99.67%	270/723/520	AA/AG/GG	0.896
ELAVL4	rs11583200		1	C/T	0.405	100%	251/729/538	CC/CT/TT	0.947
NEGR1	rs2815752		1	A/G	0.627	100%	220/687/611	GG/GA/AA	0.089
TNNI3K	rs1514175		1	A/G	0.432	100%	283/749/486	AA/AG/GG	0.846
PTBP2	rs1555543	rs11165643	1	T/C	0.596	99.93%	268/722/524	CC/CT/TT	0.291
SEC16B	rs543874		1	G/A	0.199	100%	69/485/964	GG/GA/AA	0.689
NAV1	rs2820292		1	C/A	0.557	100%	309/749/459	AA/AC/CC	1.000
TMEM18	rs6548238		2	C/T	0.830	100%	40/451/1027	TT/TC/CC	0.499
POMC	rs713586	rs10182181	2	G/A	0.469	100%	337/786/395	GG/GA/AA	0.655
KCNK3	rs11126666		2	A/G	0.274	100%	138/568/812	AA/AG/GG	0.055
FANCL	rs887912		2	T/C	0.308	100%	142/630/746	TT/TC/CC	0.412
EHBP1	rs11688816		2	G/A	0.531	100%	324/790/404	AA/AG/GG	0.444
FIGN	rs1460676		2	C/T	0.162	100%	48/415/1055	CC/CT/TT	0.101
UBE2E3	rs1528435		2	T/C	0.616	100%	222/724/572	CC/CT/TT	0.502
CREB1	rs17203016		2	G/A	0.194	99.93%	58/450/1009	GG/GA/AA	0.308
ERBB4	rs7599312		2	G/A	0.741	100%	106/574/838	AA/AG/GG	0.869

USP37	rs492400		2	C/T	0.427	99.74%	287/733/494	CC/CT/TT	0.044
RARB	rs6804842		3	G/A	0.580	99.93%	255/781/481	AA/AG/GG	0.005
FHIT	rs2365389		3	C/T	0.593	99.93%	253/766/498	TT/TC/CC	0.947
CADM2	rs13078807		3	G/A	0.199	99.93%	53/486/978	GG/GA/AA	0.369
RASA2	rs16851483	rs2035935	3	G/A	0.074	100.00%	16/201/1301	GG/GA/AA	0.057
ETV5	rs7647305		3	C/T	0.800	100%	64/487/967	TT/TC/CC	0.111
GNPDA2	rs10938397		4	G/A	0.425	99.93%	259/761/497	GG/GA/AA	0.948
SCARB2	rs17001654	rs17001561	4	A/G	0.157	100%	32/394/1092	AA/AG/GG	0.277
SLC39A8	rs13107325		4	T/C	0.071	100%	13/187/1318	TT/TC/CC	0.328
HHIP	rs11727676		4	T/C	0.908	100%	14/247/1257	CC/CT/TT	0.707
FLJ35779	rs2112347		5	T/G	0.619	100%	241/704/573	GG/GT/TT	0.736
HMGA1	rs206936		6	G/A	0.197	100%	63/503/952	GG/GA/AA	0.920
TDRG1	rs2033529		6	G/A	0.281	100%	116/609/793	GG/GA/AA	0.270
TFAP2B	rs987237		6	G/A	0.181	100%	53/450/1015	GG/GA/AA	0.914
FOXO3	rs9400239		6	C/T	0.696	100%	142/646/730	TT/TC/CC	0.764
LOC285762	rs9374842		6	T/C	0.768	100%	67/548/903	CC/CT/TT	0.245
IFNGR1	rs13201877		6	G/A	0.144	100%	22/370/1126	GG/GA/AA	0.439
PARK2	rs13191362		6	A/G	0.890	100%	20/296/1202	GG/GA/AA	0.745
HIP1	rs1167827		7	G/A	0.591	100%	260/754/504	AA/AG/GG	0.323
ASB4	rs6465468		7	T/G	0.332	94.73%	140/657/641	TT/TG/GG	0.462

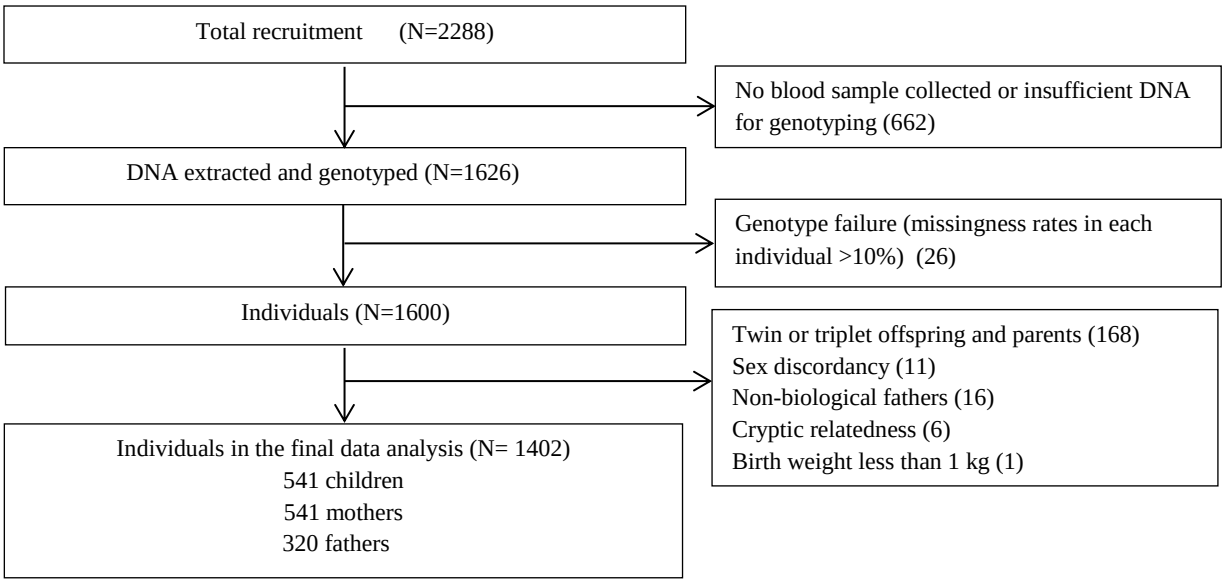
ZBTB10	rs16907751	8	C/T	0.899	100%	21/286/1211	TT/TC/CC	0.727
RALYL	rs2033732	8	C/T	0.745	100%	108/582/828	TT/TC/CC	0.676
C9orf93	rs4740619	9	T/C	0.554	100%	314/726/476	CC/CT/TT	0.440
LRRN6C	rs10968576	9	G/A	0.316	100%	139/653/726	GG/GA/AA	0.556
EPB41L4B	rs6477694	9	C/T	0.348	100%	165/736/617	CC/CT/TT	0.049
TLR4	rs1928295	9	T/C	0.564	100%	297/725/496	CC/CT/TT	0.897
LMX1B	rs10733682	9	A/G	0.478	100%	380/747/391	AA/AG/GG	0.484
GRID1	rs7899106	10	G/A	0.048	100%	4/147/1367	GG/GA/AA	1.000
HIF1AN	rs17094222	10	C/T	0.218	100%	65/521/932	CC/CT/TT	0.514
NT5C2	rs11191560	10	C/T	0.091	100%	11/259/1248	CC/CT/TT	1.000
TCF7L2	rs7903146	10	C/T	0.700	100%	131/646/741	TT/TC/CC	0.405
TUB	rs4929949	11	C/T	0.521	100%	350/775/392	TT/TC/CC	0.308
BDNF	rs925946	11	T/G	0.303	100%	125/655/738	TT/TG/GG	0.201
BDNF	rs6265	11	G/A	0.805	100%	56/479/983	AA/AG/GG	0.104
HSD17B12	rs2176598	11	T/C	0.249	100%	109/530/879	TT/TC/CC	0.050
MTCH2	rs10838738	11	G/A	0.364	99.80%	204/697/614	GG/GA/AA	0.192
CADM1	rs12286929	11	G/A	0.530	99.93%	348/720/449	AA/AG/GG	0.055
FAIM2	rs7138803	12	A/G	0.346	100%	203/639/676	AA/AG/GG	0.003
CLIP1	rs11057405	12	G/A	0.895	99.47%	16/286/1208	AA/AG/GG	0.733
MIR548X2	rs9540493	13	A/G	0.435	100%	279/767/471	AA/AG/GG	0.651

MIR548A2	rs1441264		13	A/G	0.602	100%	242/735/541	GG/GA/AA	0.947
STXBP6	rs10132280		14	CA	0.684	100%	145/655/718	AA/AC/CC	0.606
PRKD1	rs11847697		14	T/C	0.044	100%	4/132/1382	TT/TC/CC	0.426
NRXN3	rs10150332	rs17109256	14	A/G	0.218	100%	64/534/920	AA/AG/GG	0.225
DMXL2	rs3736485		15	A/G	0.466	100%	353/748/417	AA/AG/GG	0.523
MAP2K5	rs2241423		15	G/A	0.770	100%	95/531/892	AA/AG/GG	0.720
LOC100287559	rs7164727		15	T/C	0.666	100%	170/675/673	CC/CT/TT	0.253
NLRC3	rs758747		16	T/C	0.278	100%	127/629/762	TT/TC/CC	0.579
GPRC5B	rs12444979		16	C/T	0.853	100%	35/366/1117	TT/TC/CC	0.899
SBK1	rs2650492		16	A/G	0.291	100%	110/640/767	AA/AG/GG	0.877
SH2B1	rs7498665		16	G/A	0.383	99.93%	216/722/579	GG/GA/AA	0.946
INO80E	rs4787491		16	G/A	0.553	100%	321/741/456	AA/AG/GG	0.898
KAT8	rs9925964		16	A/G	0.652	100%	197/693/628	GG/GA/AA	0.888
CBLN1	rs2080454		16	C/A	0.391	100%	216/736/566	CC/CA/AA	0.229
FTO	rs9939609		16	A/T	0.391	100%	234/700/584	AA/AT/TT	0.463
SMG6	rs9914578		17	G/C	0.212	100%	65/526/926	GG/GC/CC	0.390
RABEP1	rs1000940		17	G/A	0.306	100%	161/614/743	GG/GA/AA	0.410
LOC284260	rs7239883		18	G/A	0.409	100%	231/753/534	GG/GA/AA	0.645
GRP	rs7243357		18	T/G	0.833	100%	47/401/1070	GG/GT/TT	0.568
MC4R	rs571312		18	A/C	0.247	100%	86/572/860	AA/AC/CC	0.669

PGPEP1	rs17724992	19	A/G	0.745	100%	108/583/826	GG/GA/AA	0.801
KCTD15	rs11084753	19	G/A	0.649	99.08%	184/690/630	AA/AG/GG	1.000
TOMM40-APOE-APOC1	rs2075650	19	A/G	0.855	100%	38/372/1108	GG/GA/AA	0.898
GIPR	rs2287019	19	C/T	0.813	100%	50/463/1005	TT/TC/CC	1.000
TMEM160	rs3810291	19	A/G	0.673	98.09%	147/690/652	GG/GA/AA	0.024
ETS2	rs2836754	21	C/T	0.628	100%	225/693/600	TT/TC/CC	0.376

*Risk allele frequencies were calculated from independent mothers and fathers

Supplementary Figure 1. Flowchart for quality control.



References:

1. Body mass index of children and youth, 2012 to 2013: Statistics Canada; 2014. Available from: http://www.statcan.gc.ca/pub/82-625-x/2014001/article/14105-eng.htm.
2. Lakshman R, Elks CE, Ong KK. Childhood obesity. *Circulation* 2012;**126:** 1770-1779.
3. Kuhle S, Kirk S, Ohinmaa A, Yasui Y, Allen AC, Veugelers PJ. Use and cost of health services among overweight and obese Canadian children. *International journal of pediatric obesity : IJPO : an official journal of the International Association for the Study of Obesity* 2011;**6:** 142-148.
4. Dietz WH. Critical periods in childhood for the development of obesity. *The American journal of clinical nutrition* 1994;**59:** 955-959.
5. Morandi A, Meyre D, Lobbens S, Kleinman K, Kaakinen M, Rifas-Shiman SL, *et al.* Estimation of newborn risk for child or adolescent obesity: lessons from longitudinal birth cohorts. *PloS one* 2012;**7:** e49919.
6. Choquet H, Meyre D. Molecular basis of obesity: current status and future prospects. *Current genomics* 2011;**12:** 154-168.
7. Locke AE, Kahali B, Berndt SI, Justice AE, Pers TH, Day FR, *et al.* Genetic studies of body mass index yield new insights for obesity biology. *Nature* 2015;**518:** 197-206.
8. Elks CE, Loos RJ, Sharp SJ, Langenberg C, Ring SM, Timpson NJ, *et al.* Genetic markers of adult obesity risk are associated with greater early infancy weight gain and growth. *PLoS medicine* 2010;**7:** e1000284.
9. Belsky DW, Moffitt TE, Houts R, Bennett GG, Biddle AK, Blumenthal JA, *et al.* Polygenic risk, rapid childhood growth, and the development of obesity: evidence from a 4-decade longitudinal study. *Archives of pediatrics & adolescent medicine* 2012;**166:** 515-521.
10. Sovio U, Mook-Kanamori DO, Warrington NM, Lawrence R, Briollais L, Palmer CN, *et al.* Association between common variation at the FTO locus and changes in body mass index from infancy to late childhood: the complex nature of genetic association through growth and development. *PLoS genetics* 2011;**7:** e1001307.
11. Hardy R, Wills AK, Wong A, Elks CE, Wareham NJ, Loos RJ, *et al.* Life course variations in the associations between FTO and MC4R gene variants and body size. *Human molecular genetics* 2010;**19:** 545-552.
12. Elks CE, Loos RJ, Hardy R, Wills AK, Wong A, Wareham NJ, *et al.* Adult obesity susceptibility variants are associated with greater childhood weight gain and a faster tempo of growth: the 1946 British Birth Cohort Study. *The American journal of clinical nutrition* 2012;**95:** 1150-1156.
13. Warrington NM, Howe LD, Wu YY, Timpson NJ, Tilling K, Pennell CE, *et al.* Association of a body mass index genetic risk score with growth throughout childhood and adolescence. *PloS one* 2013;**8:** e79547.
14. Warrington NM, Howe LD, Paternoster L, Kaakinen M, Herrala S, Huikari V, *et al.* A genome-wide association study of body mass index across early life and childhood. *International journal of epidemiology* 2015;**44:** 700-712.
15. Felix JF, Bradfield JP, Monnereau C, van der Valk RJ, Stergiakouli E, Chesi A, *et al.* Genome-wide association analysis identifies three new susceptibility loci for childhood body mass index. *Human molecular genetics* 2015.

16. Graff M, Ngwa JS, Workalemahu T, Homuth G, Schipf S, Teumer A, *et al.* Genome-wide analysis of BMI in adolescents and young adults reveals additional insight into the effects of genetic loci over the life course. *Human molecular genetics* 2013;**22:** 3597-3607.
17. Elks CE, Heude B, de Zegher F, Barton SJ, Clement K, Inskip HM, *et al.* Associations between genetic obesity susceptibility and early postnatal fat and lean mass: an individual participant meta-analysis. *JAMA pediatrics* 2014;**168:** 1122-1130.
18. Hoggart CJ, Venturini G, Mangino M, Gomez F, Ascari G, Zhao JH, *et al.* Novel approach identifies SNPs in SLC2A10 and KCNK9 with evidence for parent-of-origin effect on body mass index. *PLoS genetics* 2014;**10:** e1004508.
19. Liu X, Hinney A, Scholz M, Scherag A, Tonjes A, Stumvoll M, *et al.* Indications for potential parent-of-origin effects within the FTO gene. *PloS one* 2015;**10:** e0119206.
20. Morrison KM, Atkinson SA, Yusuf S, Bourgeois J, McDonald S, McQueen MJ, *et al.* The Family Atherosclerosis Monitoring In earLY life (FAMILY) study: rationale, design, and baseline data of a study examining the early determinants of atherosclerosis. *American heart journal* 2009;**158:** 533-539.
21. Metzger BE, Gabbe SG, Persson B, Buchanan TA, Catalano PA, Damm P, *et al.* International association of diabetes and pregnancy study groups recommendations on the diagnosis and classification of hyperglycemia in pregnancy. *Diabetes care* 2010;**33:** 676-682.
22. Dudbridge F. Power and predictive accuracy of polygenic risk scores. *PLoS genetics* 2013;**9:** e1003348.
23. Thompson WK, Hallmayer J, O'Hara R. Design considerations for characterizing psychiatric trajectories across the lifespan: application to effects of APOE-epsilon4 on cerebral cortical thickness in Alzheimer's disease. *The American journal of psychiatry* 2011;**168:** 894-903.
24. Lawson HA, Cheverud JM, Wolf JB. Genomic imprinting and parent-of-origin effects on complex traits. *Nature reviews Genetics* 2013;**14:** 609-617.
25. Morison IM, Ramsay JP, Spencer HG. A census of mammalian imprinting. *Trends in genetics : TIG* 2005;**21:** 457-465.
26. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MA, Bender D, *et al.* PLINK: a tool set for whole-genome association and population-based linkage analyses. *American journal of human genetics* 2007;**81:** 559-575.
27. Ihaka R, Genteman R. R: a language for data analysis and graphics. *Journal of Computational and Graphical Statistics* 1996;**5:** 299-314.
28. Zhao Y, Wang SF, Mu M, Sheng J. Birth weight and overweight/obesity in adults: a meta-analysis. *European journal of pediatrics* 2012;**171:** 1737-1746.
29. Lunde A, Melve KK, Gjessing HK, Skjaerven R, Irgens LM. Genetic and environmental influences on birth weight, birth length, head circumference, and gestational age by use of population-based parent-offspring data. *American journal of epidemiology* 2007;**165:** 734-741.
30. Ehrenberg HM, Mercer BM, Catalano PM. The influence of obesity and diabetes on the prevalence of macrosomia. *American journal of obstetrics and gynecology* 2004;**191:** 964-968.
31. Kilpelainen TO, den Hoed M, Ong KK, Grontved A, Brage S, Jameson K, *et al.* Obesity-susceptibility loci have a limited influence on birth weight: a meta-analysis of up to 28,219 individuals. *The American journal of clinical nutrition* 2011;**93:** 851-860.
32. Freathy RM, Mook-Kanamori DO, Sovio U, Prokopenko I, Timpson NJ, Berry DJ, *et al.* Variants in ADCY5 and near CCNL1 are associated with fetal growth and birth weight. *Nature genetics* 2010;**42:** 430-435.

33. Horikoshi M, Yaghootkar H, Mook-Kanamori DO, Sovio U, Taal HR, Hennig BJ, *et al.* New loci associated with birth weight identify genetic links between intrauterine growth and adult height and metabolism. *Nature genetics* 2013;**45:** 76-82.
34. Haworth CM, Carnell S, Meaburn EL, Davis OS, Plomin R, Wardle J. Increasing heritability of BMI and stronger associations with the FTO gene over childhood. *Obesity (Silver Spring)* 2008;**16:** 2663-2668.
35. Ong KK, Ahmed ML, Emmett PM, Preece MA, Dunger DB. Association between postnatal catch-up growth and obesity in childhood: prospective cohort study. *BMJ* 2000;**320:** 967-971.
36. Barker DJ, Osmond C, Forsen TJ, Kajantie E, Eriksson JG. Trajectories of growth among children who have coronary events as adults. *The New England journal of medicine* 2005;**353:** 1802-1809.
37. Frost JM, Moore GE. The importance of imprinting in the human placenta. *PLoS genetics* 2010;**6:** e1001015.
38. Hanson RL, Guo T, Muller YL, Fleming J, Knowler WC, Kobes S, *et al.* Strong parent-of-origin effects in the association of KCNQ1 variants with type 2 diabetes in American Indians. *Diabetes* 2013;**62:** 2984-2991.
39. Sethi N, Yan Y, Quek D, Schupbach T, Kang Y. Rabconnectin-3 is a functional regulator of mammalian Notch signaling. *The Journal of biological chemistry* 2010;**285:** 34757-34764.
40. Bi P, Shan T, Liu W, Yue F, Yang X, Liang XR, *et al.* Inhibition of Notch signaling promotes browning of white adipose tissue and ameliorates obesity. *Nature medicine* 2014;**20:** 911-918.

CHAPTER IV: EVIDENCE OF A GENETIC LINK BETWEEN PREPREGNANCY BMI VARIATION AND POSTPARTUM WEIGHT RETENTION: THE FAMILY STUDY

Aihua Li[1], Koon K. Teo[1,2], Katherine M. Morrison[3], Sarah D. McDonald[1,4], Stephanie A. Atkinson[3], Sonia S. Anand[1,2], David Meyre[1,5]

[1]Department of Clinical Epidemiology and Biostatistics, McMaster University, Hamilton, ON, Canada

[2]Department of Medicine, McMaster University, Hamilton, ON, Canada

[3]Department of Pediatrics, Hamilton Health Sciences and McMaster University, Hamilton, ON, Canada

[4]Department of Obstetrics and Gynecology, McMaster University, Hamilton, ON, Canada

[5]Department of Pathology and Molecular Medicine, McMaster University, Hamilton, ON, Canada.

Abstract

Objectives: We examined the associations of prepregnancy BMI and gestational weight gain (GWG) with maternal postpartum weight retention and offspring obesity-related traits. We also investigated the maternal and offspring genetic contribution of BMI-associated SNPs to GWG and postpartum weight retention in the FAMILY birth cohort.

Methods: Blood samples from mothers (n=608) and offspring (n=541) were genotyped for 83 SNPs associated with BMI. Maternal prepregnancy BMI, GWG, 1-year and 5-year postpartum weight retention were obtained. Repeated weight and length/height in offspring were measured from birth to 5 years old. Multiple linear regression and mixed-effects regression models were performed.

Results: Prepregnancy BMI was positively associated with offspring weight and BMI Z-score at birth and longitudinally from birth to 5 years. GWG was positively associated with maternal postpartum weight retention at 1 and 5 years and with offspring weight and BMI Z-score at birth.

The maternal BMI susceptibility genetic risk score (GRS) was associated with prepregnancy BMI. It was also associated with offspring weight Z-score at birth and postpartum weight retention at 5 years, but not associated with GWG.

Conclusions: Adult BMI-associated SNPs may contribute to the genetic link between maternal prepregnancy BMI variation and long-term postpartum weight retention and offspring birth weight.

Introduction

The fetal origins hypothesis (also named prenatal programming or thrifty phenotype hypothesis) proposed by David Barker in 1995 states that fetal undernutrition in middle to late gestation leads to chronic diseases in adulthood.[1] These findings have been confirmed in several independent studies.[2, 3] A growing body of evidence also supports the fetal origins of obesity.[4, 5] High prepregnancy body mass index (BMI) and excess gestational weight gain (GWG) have been associated with adverse maternal and offspring outcomes. Pregnant women with obesity or whose GWG exceeds the recommended ranges may experience a variety of adverse short- and long-term health outcomes including gestational hypertension, gestational diabetes mellitus, cesarean delivery, postpartum weight retention and obesity.[6, 7] High prepregnancy BMI or excessive GWG is also associated with higher birthweight, childhood and adulthood obesity, as well as related metabolic diseases in offspring.[7-11] Based on this well-established clinical epidemiologic evidence, the 2009 Institute of Medicine (IOM) GWG guidelines recommend optimal ranges according to the mother's prepregnancy BMI: 12.5-18 kg for underweight women (<18.5 kg/m^2), 11.5-16 kg for normal weight women (18.5-24.9 kg/m^2), 7-11.5 kg for overweight women (25-29.9 kg/m^2) and 5-9 kg for women with obesity (>30 kg/m^2).[12]

GWG consists of the accretion of proteins, fat, water and minerals which are deposited in the fetus, placenta, amniotic fluid, uterus, mammary gland, blood and adipose tissue [12]. Using a four-composition model to measure fat content, studies have shown that maternal fat mass gain during pregnancy parallels GWG and approximately 30% of GWG is fat mass in those pregnant women who gain weight within IOM recommendation.[12, 13] The fetus accounts for 20-25% of overall GWG and fat mass consists of 8-20% birth weight.[12, 14] The difference in fat mass is more variable than fat-free mass in both mothers and fetuses during gestation.[15] In addition to

risk factors such as unhealthy lifestyles (e.g. diet, smoking, and physical inactivity) or medical conditions (e.g. maternal gestational diabetes mellitus), genetic factors have recently been shown to explain 43% of the variation in GWG in the first pregnancy and 26% in the second pregnancy in twin studies using structural equation modeling.[16] Given that a considerable fraction of GWG is attributable to the fat in both fetus and mother, it is plausible that fetal and maternal genetic variants associated with BMI might also contribute to GWG.[12, 17] In addition, our previous study using FAMILY cohort has shown that fetal BMI-associated genetic risk score (GRS) is associated with birth weight. Furthermore, it has been hypothesized that the fetal genotype likely to influence maternal metabolism through placenta hormone and proteins.[17] Therefore we examined whether the fetal GRS was linked to maternal GWG and postpartum weight retention. Two previous studies in European or North-American populations did not identify any associations between 4 to 9 maternal or offspring BMI susceptibility single nucleotide polymorphisms (SNPs) and GWG.[18, 19] However, the total number of SNPs associated with adult BMI at genome-wide significance level ($P<5\times10^{-8}$) has recently increased to 113.[20] Therefore, the contribution of an exhaustive list of BMI susceptibility variants to the variation of GWG warrants further investigation. Furthermore, no study has been done in Canada where the environment is believed to be more obesogenic than in Europe, but less obesogenic than in the United States.

In the Canadian prospective FAMILY birth cohort, we investigated: 1) the associations between prepregnancy BMI and GWG and postpartum weight retention at 1 year and 5 years; 2) the effects of maternal prepregnancy BMI and GWG on offspring weight and BMI from birth to 5 years of age; 3) the contribution of the maternal BMI-associated GRS combining 83 SNPs robustly associated with adult BMI to maternal prepregnancy BMI, GWG, postpartum weight

retention at 1 year and 5 years, and offspring weight and BMI from birth to 5 years of age; and 4) the contribution of the offspring BMI-associated GRS to maternal GWG and postpartum weight retention at 1 year and 5 years.

Methods

Participants

The Family Atherosclerosis Monitoring In earLY life (FAMILY) study is a longitudinal birth cohort examining the fetal and early childhood determinants for the development of adiposity, cardiovascular diseases and atherosclerosis in childhood. The study has been described in detail previously.[21] Briefly, 857 families including 901 newborns, 857 mothers and 530 fathers were enrolled from three hospitals in the Hamilton and Burlington regions (Ontario, Canada) between 2004 and 2009, and were followed for up to 5 years, with a planned follow-up for 10 years or more. Only were singletons selected (N=816 families) in our analyses because multiple births have a strong impact on offspring birthweight and maternal gestational weight gain. Informed consent was obtained from all the adult participants, and the parents provided consent for their children. All procedures were performed in accordance with relevant guidelines and regulations. The research ethics boards at Hamilton Health Sciences and St Joseph's Health Center in Hamilton, and Joseph Brant Memorial Hospital in Burlington approved the protocol.

Genotyping

Buffy coats were used for genomic DNA extraction and genotyping was conducted using the Illumina Cardio-Metabochip (San Diego, CA, USA). Eighty-three SNPs were selected based on two criteria of: being associated with BMI at genome-wide significance level ($P<5\times10^{-8}$) in

adults and being available on the Cardio-Metabochip array (lead SNP or proxy) (See Table S1). The SNP selection procedure was completed in March 2015. Five out of 83 were proxy SNPs ($r^2 \geq 0.95$ with the lead SNP) which were identified using the 1000 Genomes data of European population via the website tool SNAP (SNP Annotation and Proxy Search).[22] We double checked the availability of the SNPs and proxies using SNAP and their chromosomal position in the Illumina data file.[22] To ensure all 83 SNPs were independent, pairwise linkage disequilibrium was examined. All the pairwise r^2 were less than 0.1 in European population tested using the 1000 Genomes Project data. The call rates of 83 SNPs ranged from 95.7 to 100% and the genotypes satisfied Hardy Weinberg Equilibrium test (P>0.001). Genotyping data indicated that 26 individuals had SNP missingness rates >10%, 16 individuals were from 5 families having non-biological fathers, and 11 individuals had sex-discordance between reported and SNP-identified sex, all of them were excluded from the analysis. Ten pairs of individuals had cryptic relatedness (2^{nd} degree relatives) and one individual of each pair was randomly selected for exclusion (See Figure S1). The ethnicity reported by participants was verified by principal component analysis (PCA) using EIGENSTRAT.[23] The first 10 components of PCAs were used for adjustment of ethnicity. The majority of FAMILY participants were white Caucasians (92.8% mothers and 91.1% offspring).

Phenotypes

Newborn characteristics of sex and gestational age at birth were recorded from medical charts. Offspring weight and length/height were measured by trained research staff with standard scales at birth, 1, 2, 3 and 5 years of age. The BMI was calculated as weight in kilograms divided by length/height in meters squared. Individual weight, length/height and BMI in offspring were

converted to age- and sex-specific Z-scores using World Health Organization growth reference (2006) (http://www.who-.int/childgrowth/standards/en/).

Mothers' weight and height during pregnancy visits and at 1 and 5 years after delivery were measured by trained staff. Maternal age, self-reported prepregnancy weight, parity and smoking were obtained at baseline visit. Gestational weight gain, gestational diabetes, and all subsequent live births within 5 years of the index birth were obtained from the questionnaires at subsequent visits or from the medical chart review. GWG was assessed as the difference between the last measured weight prior to delivery and self-reported weight before pregnancy. Analyses including GWG were adjusted by adding the variable of gestational age at the time of the last maternal weight measured into the regression models. Maternal postpartum weight retention at 1- and 5-years was evaluated from the difference in weights between 1- or 5-years after delivery and before pregnancy. Gestational diabetes mellitus, defined as "any degree of glucose intolerance with onset or first recognition during pregnancy", was diagnosed according to the International Association of Diabetes and Pregnancy Study Groups criteria using a 75 g oral glucose tolerance test. Overall, 1149 individuals were included in data analyses in which 608 mothers were analyzed for maternal epidemiological associations and 541 mother-child pairs were analyzed for associations between mothers and offspring (See flow chart in Figure S1). The numbers of missing values for each of the exposures and outcomes are different and the numbers for the available data on each exposure and outcome for this study are presented in Table 1. For each association test, maximal data available were included and the numbers included in each analysis are reported in the tables.

Statistical analysis

We selected the risk alleles for each of 83 SNPs as previously reported in literature (Table S1). The genotypes of each locus were coded as 0, 1 and 2 designating the number of risk alleles following an additive mode of inheritance. The coding for the missing genotypes (<0.1% of the total genotype) was imputed with arithmetic average of the coded genotypes from individuals who were successfully genotyped. A GRS was obtained by summing the risk alleles of 83 SNPs and an unweighted GRS as recommended by Dudbridge was used.[24]

All the associations between prepregnancy BMI/GWG and postpartum weight retention, between prepregnancy BMI/GWG and offspring weight and BMI Z-score at birth, between the maternal GRS and prepregnancy BMI, or GWG, or offspring weight and BMI Z-score at birth, and between the offspring GRS and maternal GWG and postpartum weight retention were tested using multiple linear regression models. Each regression model was adjusted for relevant covariates which were presented in the footnotes of each table, including offspring gestational age at birth, maternal age, prepregnancy BMI, gestational weight gain, parity, gestational diabetes, smoking, all subsequent live births within 5 years of the index birth and ethnicity. There was no collinearity between covariates.

Linear mixed-effects models were used to investigate the associations of the maternal prepregnancy BMI, GWG, and GRS on offspring weight and BMI Z-score from 0 to 5 years old with repeated measurements. This approach was selected since it takes into account the correlations between repeated measurements on the same individual and allows for missing measurement data and measurement at different time, assuming the missing events are random.[25] The offspring weight and BMI Z-score at birth of each individual (intercept) and correlation among measurements on the same subject (slope of offspring's age) were modeled as random effects, and time, ethnicity and prepregnancy BMI or GWG were modeled as fixed effects. An

interaction term of prepregnancy BMI×time (in year) or GWG×time (in year) or GRS×time (in year), which tests whether the effect of the prepregnancy BMI/GWG/GRS changes over time, was also initially added to the linear mixed-effects regression model. Due to its non-significance, this term was removed from the final analysis to produce the most parsimonious model.

The analyses of the associations between the maternal GRS and offspring weight and BMI Z-score at birth and longitudinally from birth to 5 years old were repeated including the fetal GRS in the models to correct for potential confounding caused by direct effect of the fetal genotype.[26]

Multiple outcomes were tested in this study, including offspring obesity-related traits, maternal prepregnancy BMI and GWG, and postpartum weight retention. Applying a Bonferroni corrected P-value across the outcomes will reduce the chance of making type I errors, but it will increase the chance of making type II errors. Therefore, we applied Bonferroni correction to outcomes under each question.[27] A P-value of less than 0.0125 was considered statistically significant when analyses of the associations between maternal prepregnancy BMI/GWG/GRS and offspring weight and BMI Z-score at birth and longitudinally from 0-5 years old were performed (a total of 4 analyses). A P-value of less than 0.025 was considered statistically significant when analyses of associations between prepregnancy BMI/GWG and maternal postpartum weight retention were performed (a total of 2 analyses). A P-value of less than 0.025 was considered statistically significant when analyses of GRS on maternal postpartum weight retention were applied (a total of 2 analyses). All the statistical analyses were performed using PLINK (version 1.07) and R (version 2.15.2).[28, 29]

Results

Participant characteristics

The characteristics of mothers and offspring included in the analyses are summarized in Table 1. Mother's prepregnancy BMI was 26.5 kg/m^2 (SD=6.4). The percentages of overweight and obesity in pregnancy in FAMILY were 26.2% and 21.3%, respectively. The mean of GWG was 15.9 kg (SD=5.4) (Table 1). According to the IOM definition, 57.8% of mothers had excessive weight gain, 13.8% had insufficient weight gain and 28.4% gained weight within the optimal ranges (Supplementary Table 2).

Association between prepregnancy BMI and GWG

After adjusting for maternal age, parity, gestational diabetes, smoking and ethnicity, prepregnancy BMI was significantly negatively associated with GWG. Each one unit increase in prepregnancy BMI was associated with a 0.25 kg decrease in GWG (SE=0.04; P=8.91×10^{-11}).

Associations of prepregnancy BMI/GWG and postpartum weight retention at 1 year and 5 years

Maternal prepregnancy BMI was not associated with either postpartum weight retention at 1 year or at 5 years (β±SE, 0.05±0.04 kg/BMI unit; P=0.28; and -0.09±0.07 kg/BMI unit; P=0.16; respectively) with adjustments of maternal age, time interval, GWG, parity, subsequent live births within 5 years of the index birth, and ethnicity. In contrast, GWG was significantly and positively associated with postpartum weight retention at both 1 year and 5 years (β±SE, 0.46±0.05 kg/kg GWG; P=1.59×10^{-17} and 0.24±0.08 kg/kg GWG; P=4.42×10^{-3}, respectively) with adjustments of maternal age, time interval, prepregnancy BMI, subsequent live births within 5 years of the index birth, and ethnicity.

Effects of prepregnancy BMI/GWG on offspring weight and BMI from birth to 5 years old

Maternal prepregnancy BMI was significantly associated with offspring birth weight Z-score (β±SE, 0.027±0.007/BMI unit; $P=4.24\times10^{-4}$) and BMI Z-score at birth (β±SE, 0.03±0.008/BMI unit; $P=4.82\times10^{-4}$) (Table 2). In the longitudinal analyses, significant associations of maternal prepregnancy BMI with offspring weight and BMI Z-score from birth to 5 years of age (β±SE, 0.021±0.006/BMI unit; $P=3.59\times10^{-4}$; and 0.025±0.005/BMI unit; $P=6.30\times10^{-6}$, respectively) were identified (Table 2).

GWG was significantly associated with offspring birth weight Z-score (β±SE, 0.033±0.008/kg GWG; $P=8.53\times10^{-5}$) and BMI Z-score at birth (β±SE, 0.031±0.009/kg GWG; $P=1.19\times10^{-3}$). In the longitudinal analyses, the overall effect of GWG on offspring weight Z-score from birth to 5 years of age was statistically significant (β±SE, 0.019±0.007/kg GWG; $P=8.39\times10^{-3}$), but not on offspring BMI Z-score.

Associations of the maternal BMI-associated GRS with offspring weight and BMI from birth to 5 years

The maternal GRS was associated with offspring birth weight Z-score (β±SE, 0.021±0.008/risk allele; P=0.01), but not BMI Z-score at birth. In the longitudinal analyses, the maternal GRS was also associated with offspring weight Z-score from birth to 5 years of age was identified (β±SE, 0.016±0.007/risk allele; P=0.01), but not BMI Z-score. After adjusting for the fetal GRS, the estimated effects changed to be not significant (Table 3).

Associations of the maternal and offspring BMI GRS with maternal prepregnancy BMI, GWG and postpartum weight retention

The maternal GRS was positively associated with prepregnancy BMI (β±SE, 0.11±0.04 BMI units/risk allele; P=0.01) and explained 1.07% variation of preprgnancy BMI, but it was not associated with GWG. The maternal GRS was not associated with 1 year postpartum weight retention. On the contrary, it was associated with weight retention at 5 years (β±SE, 0.15±0.06 kg/risk allele; P=0.02). The offspring GRS was not associated with maternal GWG or weight retention at 1 year and 5 years (Table 4).

Discussion

The prevalence of obesity in pregnancy in FAMILY is 21%, higher than 14% reported in Ontario, Canada from the Maternity Experience Survey 2006-2007 (http://www.phac-aspc.gc.ca/rhs-ssg/survey-eng.php). Overall, 58% of the mothers in our study gained excessive weight during pregnancy according to the IOM recommended ranges, similar to that in a Canadian cohort of 52%.[30] Furthermore, 73% of women with overweight and 72% of women with obesity had GWG greater than the recommended ranges. In contrast, 27% of underweight women and 46% of normal weight women gained excessive weight.

Our results are consistent with the findings of previous studies that maternal prepregnancy BMI was associated with offspring weight and BMI at different ages.[8, 31, 32] This supports the hypothesis that intergenerational factors contribute to obesity in offspring at early age. However, these associations may be confounded by genetic or environmental influence shared by mother and offspring. Using siblings born from the same mother and fixed-effects models to minimize the shared familial effects, Branum *et al.* found that the genetic effects

contributed to the association between maternal prepregnancy BMI and GWG and child BMI.[33] In a most recent mendelian randomization study using the maternal GRS as an instrumental variable, Tyrrell *et al.* found that genetically elevated maternal BMI is casually associated with higher offspring birth weight.[26] Our study is the first to examine the effect of the maternal GRS on maternal prepregnancy BMI, GWG, postpartum weight retention and offspring weight and BMI based on an exhaustive list of SNPs (N=83 SNPs). The results showed that the maternal GRS was associated with both maternal prepregnancy BMI and offspring birth weight. In our study, there was no evidence of associations between the maternal GRS and any confounding variables. Therefore we replicated the causal association between maternal BMI and offspring birth weight in the Family cohort. The maternal GRS was also associated with maternal weight retention at 5 years after delivery, indicating that the genes contributing to BMI variation in general adults may also be involved in the resistance to the long-term postpartum weight loss.

Neither maternal nor offspring BMI GRS were found to be associated with GWG in our study. This is consistent with two previous studies.[18, 19] Stube *et al.* first reported that none of 9 obesity-associated genetic loci was associated with GWG.[18] Later, Lawlor *et al.* demonstrated that 4 BMI loci (*FTO*, *MC4R*, *TMEM18* and *GNPDA2*), individually or combined as a genetic risk score from fetus or mother, were not associated with GWG.[19] Compared to these two studies, we increased the number of obesity susceptible SNPs to 83 which was assumed to increase the statistical power to detect an association. Several reasons may explain the null finding. First, GWG is a complex trait and the fat accretion in both mother and fetus caused by genetic variants may not be sufficient to significantly impact the total GWG. Second, genetic variants involved in different biological pathways rather than obesity susceptible variants may account for the variation in GWG. Two studies have demonstrated that genetic variants in

KCNQ1 and *TPH1*, which are involved in the regulation of glucose and β-cell proliferation during pregnancy, respectively, are associated with GWG.[18, 34] Third, maternal GWG measures weight change during pregnancy, and family and twin studies have shown that genetic factors have an important role in response to weight gain or loss [35]. Although all 83 SNPs are robustly associated with BMI, whether they influence the weight change, especially during pregnancy, is still unclear.[36] Fourth, genetic variants may be associated with trimester-specific GWG.[37] Therefore, GWAS of GWG, total and trimester-specific GWG, may discover novel genes and alternative pathways influencing GWG.

In addition to the findings of our study, a methodological challenge is worth addressing. Assessment of both prepregnancy BMI and GWG requires rigorous methods of data collection. For example, prepregnancy weight should be measured at a preconceptional visit, though high concordance between self-reported and measured weight has been reported.[11] GWG should be the last measured weight right before delivery from clinical records subtracting prepregnancy weight. Unfortunately, most of the data in the literature were not collected with a high level of rigor, and most studies relied on self-reported weight values which likely introduce recall error and/or bias, such as socially desirable reporting.[11, 33, 38]

The strengths of our study include repeated measurements of weight and BMI from birth to 5 years of age in offspring, appropriate adjustment for critical variables during pregnancy and an updated list of 83 obesity-susceptible SNPs. Limitations of this study include a relatively modest sample size and self-reported prepregnancy weight.

Conclusions

In summary, we have made several important discoveries in the longitudinal FAMILY birth cohort representative of the South Canadian population (Ontario). Maternal prepregnancy BMI is negatively associated with GWG. It is positively associated with offspring weight and BMI at birth and longitudinally birth to 5 years. GWG is positively associated with maternal weight retention at 1 and 5 years and it is also associated with offspring weight and BMI at birth, and weight longitudinally from birth to 5 years. The maternal BMI GRS is associated with prepregnancy BMI and weight retention at 5 years and offspring birth weight, but is not associated with GWG. The offspring BMI GRS is not associated with GWG or weight retention at 1 and 5 years. Our findings suggest that though adult BMI susceptibility genetic variants have no discernable effect to GWG, they may contribute to the genetic link between maternal prepregnancy BMI variation and long-term postpartum weight retention and offspring birth weight.

Acknowledgements

This study was funded by the Heart and Stroke Foundation of Ontario (grant # NA 7293 "Early genetic origins of cardiovascular risk factors"). The FAMILY study is funded by grants from CIHR, Hear and Stroke Foundation of Ontario and the Population Research Institute (PHRI), Hamilton Health Science (HHS) and McMaster University. AL is funded by the Ontario Graduate Scholarship. SSA holds the Heart and Stroke Foundation of Ontario, Michael G. DeGroote endowed Chair in Population Health and a Canada Research Chair in Ethnicity and Cardiovascular Disease. DM holds a Canada Research Chair in Genetics of Obesity, and SDM is supported by a Canada Research Chair in Maternal and Child Obesity Prevention and Intervention. The contribution of the mothers and their offspring to this study is gratefully

acknowledged. We would like to thank Sebastien Robiou-du-Pont and Akram Alyass for their technical assistance. We would like to thank Hudson Reddon for his help in proofreading the manuscript.

Table 1. Characteristics of mothers and offspring.

Characteristics	N	Mean ± SD
Mother's age at pregnancy	608	32.3±4.7
Mother's prepregnancy weight (kg)	593	72.2±18.1
Mother's prepregnancy BMI (kg/m^2)	591	26.5±6.4
Gestational weight gain (kg)	576	15.9±5.4
Smoking in pregnancy (%)		
Never smoker	368	61.2%
Former smoker	203	33.8%
Current smoker	30	5.0%
Parity (%)		
0	258	42.4%
1	254	41.8%
2	69	11.3%
≥3	27	4.5%
GDM (%)		
Yes	92	15.6%
No	498	84.4%
Maternal weight at 1 year after delivery (kg)	553	73.4±18.6
Maternal weight at 5 years after delivery (kg)	362	74.4±19.3
Maternal weight retention at 1 year after delivery (kg)	541	1.5±6.2
Maternal weight retention at 5 years after delivery (kg)	354	2.1±7.4
Gestational age (week)	541	39.4±1.5
Male offspring (%)	541	50.4%
Offspring weight (kg)		
Birth	498	3.4±0.5
1 y	488	10.2±1.3
2 y	473	12.8±1.5
3 y	442	15.0±1.8
5 y	363	19.6±2.8
Offspring BMI (kg/m^2)		
Birth	489	13.5±1.3
1 y	483	17.5±1.4
2 y	466	16.4±1.3
3 y	437	16.2±1.2
5 y	363	15.9±1.6

Table 2. Effects of maternal prepregnancy BMI or GWG on offspring weight and BMI Z-score at birth and from birth to 5 years old.

Traits	Prepregnancy BMI						GWG					
	Birth			0-5 years (longitudinal analyses)			Birth			0-5 years (longitudinal analyses)		
	N	β (SE)	P	N	β (SE)	P	N	β (SE)	P	N	β (SE)	P
Weight	444	0.027 (0.007)	**4.24×10^{-4}**	497	0.021 (0.006)	**3.59×10^{-4}**	444	0.033 (0.008)	**8.53×10^{-5}**	497	0.019 (0.007)	**8.39×10^{-3}**
BMI	436	0.030 (0.008)	**4.82×10^{-4}**	496	0.025 (0.005)	**6.30×10^{-6}**	436	0.031 (0.009)	**1.19×10^{-3}**	496	0.013 (0.006)	4.62×10^{-2}

The effects of maternal prepregnancy BMI or GWG on offspring weight and BMI Z-score at birth were analyzed using multiple linear regression models. Each model was adjusted for offspring gestational age at birth, maternal GWG/prepregnancy BMI, parity, gestational diabetes mellitus, smoking and ethnicity.

The overall effects of maternal prepregnancy BMI or GWG on offspring weight and BMI Z-score from birth to 5 years old were analyzed using linear mixed models. In each model, the weight or BMI Z-score in offspring at birth of each individual (intercept) and correlation among measurements on the same subject (slope of age) were set as random effects, and time, offspring ethnicity, and maternal GWG/prepregnancy BMI were set as fixed effects.

Table 3. Effects of the maternal BMI GRS on offspring weight and BMI Z-score from birth to 5 years old.

Trait	Birth					0-5 years (longitudinal analysis)				
	no offspring GRS			offspring GRS*		no offspring GRS			offspring GRS*	
	N	β (SE)	P	β (SE)	P	N	β (SE)	P	β (SE)	P
Weight	338	0.021 (0.008)	**0.01**	0.015 (0.010)	0.12	411	0.016 (0.007)	**0.01**	0.011 (0.008)	0.17
BMI	333	0.014 (0.009)	0.14	0.008 (0.011)	0.47	410	0.013 (0.006)	0.03	0.008 (0.007)	0.29

The effects of the maternal prepregnancy BMI-associated GRS on offspring weight and BMI Z-score at birth were analyzed using multiple linear regression models. Each model was adjusted for offspring gestational age at birth, sex, maternal GWG, parity, gestational diabetes mellitus, smoking and ethnicity.

The overall effects of the maternal BMI-associated GRS on offspring weight and BMI Z-score from birth to 5 years old were analyzed using linear mixed models. In each model, the weight or BMI Z-score in offspring at birth of each individual (intercept) and correlation among measurements on the same subject (slope of age) were set as random effects, and time, offspring sex and ethnicity were set as fixed effects.

*Adjusted additionally for the fetal GRS

Table 4. Effects of the offspring and maternal BMI GRS on maternal prepregnancy BMI, GWG and postpartum weight retention.

Maternal trait	Offspring GRS			Maternal GRS§		
	N	β (SE)	P	N	β (SE)	P
Prepregnancy BMI (kg/m^2)[a]		-	-	568	0.11 (0.04)	**0.01**
GWG (kg) [b]	383	0.01 (0.05)	0.90	538	0.03 (0.04)	0.50
Weight retention at 1 year (kg)[c]	342	0.05 (0.06)	0.36	475	0.03 (0.04)	0.48
Weight retention at 5 years (kg)[d]	250	-0.01 (0.08)	0.93	317	0.15 (0.06)	**0.02**

§: using maternal phenotypes only and the sample size was larger than those using phenotypes in both maternal and offspring
a: The associations were adjusted for maternal age at pregnancy, parity, smoking and ethnicity
b: The associations were adjusted for maternal age at pregnancy, parity, smoking and ethnicity
c: The associations were adjusted for maternal age at pregnancy, the time interval between measurements of weight at 1 year and prepregnancy weight, GWG and ethnicity
d: The associations were adjusted for maternal age at pregnancy, the time interval between measurements of weight at 5 years and prepregnancy weight, GWG, subsequent live births within 5 years of the index birth and ethnicity

Supplementary Table 1. Characteristics of the 83 SNPs associated with BMI variation.

Nearest Gene	SNP	Proxy	Chr	Effect/Other Allele	Risk Allele Frequency*	Call Rate	Genotype		HWE test
TAL1	rs977747		1	T/G	0.409	100%	259/727/532	TT/TG/GG	0.948
AGBL4	rs657452		1	A/G	0.421	99.67%	270/723/520	AA/AG/GG	0.896
ELAVL4	rs11583200		1	C/T	0.405	100%	251/729/538	CC/CT/TT	0.947
NEGR1	rs2815752		1	A/G	0.627	100%	220/687/611	GG/GA/AA	0.089
TNNI3K	rs1514175		1	A/G	0.432	100%	283/749/486	AA/AG/GG	0.846
PTBP2	rs1555543	rs11165643 (r^2=0.98)	1	T/C	0.596	99.93%	268/722/524	CC/CT/TT	0.291
SEC16B	rs543874		1	G/A	0.199	100%	69/485/964	GG/GA/AA	0.689
NAV1	rs2820292		1	C/A	0.557	100%	309/749/459	AA/AC/CC	1.000
TMEM18	rs6548238		2	C/T	0.830	100%	40/451/1027	TT/TC/CC	0.499
POMC	rs713586	rs10182181 (r^2=0.97)	2	G/A	0.469	100%	337/786/395	GG/GA/AA	0.655
KCNK3	rs11126666		2	A/G	0.274	100%	138/568/812	AA/AG/GG	0.055
FANCL	rs887912		2	T/C	0.308	100%	142/630/746	TT/TC/CC	0.412
EHBP1	rs11688816		2	G/A	0.531	100%	324/790/404	AA/AG/GG	0.444
FIGN	rs1460676		2	C/T	0.162	100%	48/415/1055	CC/CT/TT	0.101
UBE2E3	rs1528435		2	T/C	0.616	100%	222/724/572	CC/CT/TT	0.502
CREB1	rs17203016		2	G/A	0.194	99.93%	58/450/1009	GG/GA/AA	0.308
ERBB4	rs7599312		2	G/A	0.741	100%	106/574/838	AA/AG/GG	0.869

USP37	rs492400		2	C/T	0.427	99.74%	287/733/494	CC/CT/TT	0.044
RARB	rs6804842		3	G/A	0.580	99.93%	255/781/481	AA/AG/GG	0.005
FHIT	rs2365389		3	C/T	0.593	99.93%	253/766/498	TT/TC/CC	0.947
CADM2	rs13078807		3	G/A	0.199	99.93%	53/486/978	GG/GA/AA	0.369
RASA2	rs16851483	rs2035935 (r^2=0.95)	3	G/A	0.074	100.00%	16/201/1301	GG/GA/AA	0.057
ETV5	rs7647305		3	C/T	0.800	100%	64/487/967	TT/TC/CC	0.111
GNPDA2	rs10938397		4	G/A	0.425	99.93%	259/761/497	GG/GA/AA	0.948
SCARB2	rs17001654	rs17001561 (r^2=0.95)	4	A/G	0.157	100%	32/394/1092	AA/AG/GG	0.277
SLC39A8	rs13107325		4	T/C	0.071	100%	13/187/1318	TT/TC/CC	0.328
HHIP	rs11727676		4	T/C	0.908	100%	14/247/1257	CC/CT/TT	0.707
FLJ35779	rs2112347		5	T/G	0.619	100%	241/704/573	GG/GT/TT	0.736
HMGA1	rs206936		6	G/A	0.197	100%	63/503/952	GG/GA/AA	0.920
TDRG1	rs2033529		6	G/A	0.281	100%	116/609/793	GG/GA/AA	0.270
TFAP2B	rs987237		6	G/A	0.181	100%	53/450/1015	GG/GA/AA	0.914
FOXO3	rs9400239		6	C/T	0.696	100%	142/646/730	TT/TC/CC	0.764
LOC285762	rs9374842		6	T/C	0.768	100%	67/548/903	CC/CT/TT	0.245
IFNGR1	rs13201877		6	G/A	0.144	100%	22/370/1126	GG/GA/AA	0.439
PARK2	rs13191362		6	A/G	0.890	100%	20/296/1202	GG/GA/AA	0.745
HIP1	rs1167827		7	G/A	0.591	100%	260/754/504	AA/AG/GG	0.323

ASB4	rs6465468	7	T/G	0.332	95.73%	140/657/641	TT/TG/GG	0.462
ZBTB10	rs16907751	8	C/T	0.899	100%	21/286/1211	TT/TC/CC	0.727
RALYL	rs2033732	8	C/T	0.745	100%	108/582/828	TT/TC/CC	0.676
C9orf93	rs4740619	9	T/C	0.554	100%	314/726/476	CC/CT/TT	0.440
LRRN6C	rs10968576	9	G/A	0.316	100%	139/653/726	GG/GA/AA	0.556
EPB41L4B	rs6477694	9	C/T	0.348	100%	165/736/617	CC/CT/TT	0.049
TLR4	rs1928295	9	T/C	0.564	100%	297/725/496	CC/CT/TT	0.897
LMX1B	rs10733682	9	A/G	0.478	100%	380/747/391	AA/AG/GG	0.484
GRID1	rs7899106	10	G/A	0.048	100%	4/147/1367	GG/GA/AA	1.000
HIF1AN	rs17094222	10	C/T	0.218	100%	65/521/932	CC/CT/TT	0.514
NT5C2	rs11191560	10	C/T	0.091	100%	11/259/1248	CC/CT/TT	1.000
TCF7L2	rs7903146	10	C/T	0.700	100%	131/646/741	TT/TC/CC	0.405
TUB	rs4929949	11	C/T	0.521	100%	350/775/392	TT/TC/CC	0.308
BDNF	rs925946	11	T/G	0.303	100%	125/655/738	TT/TG/GG	0.201
BDNF	rs6265	11	G/A	0.805	100%	56/479/983	AA/AG/GG	0.104
HSD17B12	rs2176598	11	T/C	0.249	100%	109/530/879	TT/TC/CC	0.050
MTCH2	rs10838738	11	G/A	0.364	99.80%	204/697/614	GG/GA/AA	0.192
CADM1	rs12286929	11	G/A	0.530	99.93%	348/720/449	AA/AG/GG	0.055
FAIM2	rs7138803	12	A/G	0.346	100%	203/639/676	AA/AG/GG	0.003
CLIP1	rs11057405	12	G/A	0.895	99.47%	16/286/1208	AA/AG/GG	0.733

MIR548X2	rs9540493		13	A/G	0.435	100%	279/767/471	AA/AG/GG	0.651
MIR548A2	rs1441264		13	A/G	0.602	100%	242/735/541	GG/GA/AA	0.947
STXBP6	rs10132280		14	CA	0.684	100%	145/655/718	AA/AC/CC	0.606
PRKD1	rs11847697		14	T/C	0.044	100%	4/132/1382	TT/TC/CC	0.426
NRXN3	rs10150332	rs17109256 (r^2=0.99)	14	A/G	0.218	100%	64/534/920	AA/AG/GG	0.225
DMXL2	rs3736485		15	A/G	0.466	100%	353/748/417	AA/AG/GG	0.523
MAP2K5	rs2241423		15	G/A	0.770	100%	95/531/892	AA/AG/GG	0.720
LOC100287559	rs7164727		15	T/C	0.666	100%	170/675/673	CC/CT/TT	0.253
NLRC3	rs758747		16	T/C	0.278	100%	127/629/762	TT/TC/CC	0.579
GPRC5B	rs12444979		16	C/T	0.853	100%	35/366/1117	TT/TC/CC	0.899
SBK1	rs2650492		16	A/G	0.291	100%	110/640/767	AA/AG/GG	0.877
SH2B1	rs7498665		16	G/A	0.383	99.93%	216/722/579	GG/GA/AA	0.946
INO80E	rs4787491		16	G/A	0.553	100%	321/741/456	AA/AG/GG	0.898
KAT8	rs9925964		16	A/G	0.652	100%	197/693/628	GG/GA/AA	0.888
CBLN1	rs2080454		16	C/A	0.391	100%	216/736/566	CC/CA/AA	0.229
FTO	rs9939609		16	A/T	0.391	100%	234/700/584	AA/AT/TT	0.463
SMG6	rs9914578		17	G/C	0.212	100%	65/526/926	GG/GC/CC	0.390
RABEP1	rs1000940		17	G/A	0.306	100%	161/614/743	GG/GA/AA	0.410
LOC284260	rs7239883		18	G/A	0.409	100%	231/753/534	GG/GA/AA	0.645
GRP	rs7243357		18	T/G	0.833	100%	47/401/1070	GG/GT/TT	0.568

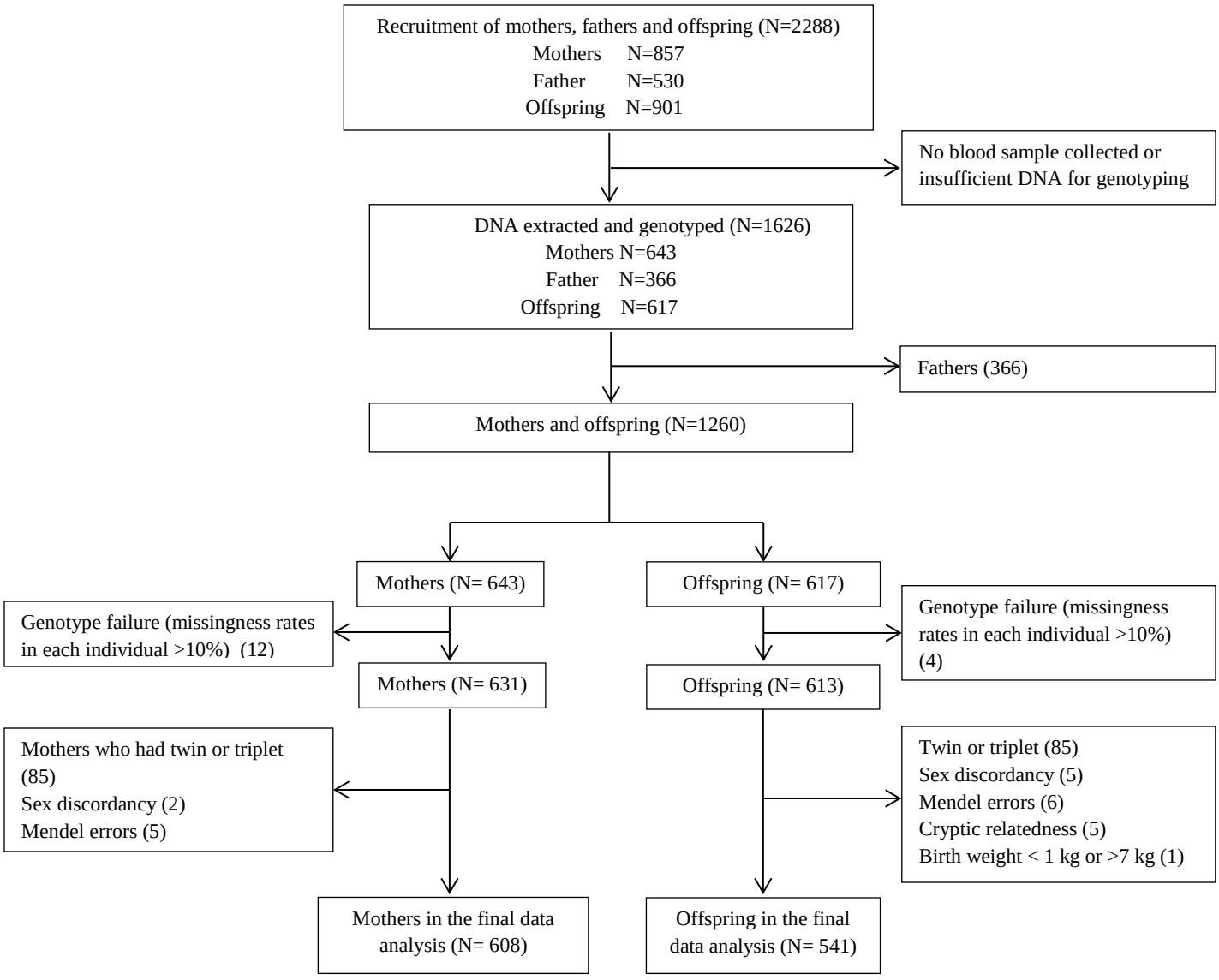
MC4R	rs571312	18	A/C	0.247	100%	86/572/860	AA/AC/CC	0.669
PGPEP1	rs17724992	19	A/G	0.745	100%	108/583/826	GG/GA/AA	0.801
KCTD15	rs11084753	19	G/A	0.649	99.08%	184/690/630	AA/AG/GG	1.000
TOMM40-APOE-APOC1	rs2075650	19	A/G	0.855	100%	38/372/1108	GG/GA/AA	0.898
GIPR	rs2287019	19	C/T	0.813	100%	50/463/1005	TT/TC/CC	1.000
TMEM160	rs3810291	19	A/G	0.673	98.09%	147/690/652	GG/GA/AA	0.024
ETS2	rs2836754	21	C/T	0.628	100%	225/693/600	TT/TC/CC	0.376

Supplementary Table 2. Maternal prepregnancy BMI and GWG categories in the FAMILY cohort.

BMI[a]	IOM GWG ranges [b,c]					
	Lower		Normal		Higher	
Category N (%)	N	%	N	%	N	%
Underweight: 11 (1.9)	2	18.2	6	54.5	3	27.3
Normal weight: 299 (50.6)	57	19.4	103	35.0	134	45.6
Overweight: 155 (26.2)	5	3.2	37	24.0	112	72.7
Obesity: 126 (21.3)	15	13.0	17	14.8	83	72.2
Total: 591	79	13.8	163	28.4	332	57.8

a. BMI was categorized using World Health Organization (WHO) definition: a BMI less than 18.5 kg/m^2 is underweight, a BMI between 18.5 and 24.99 kg/m^2 is normal weight, a BMI between 25 and 29.99 kg/m^2 is overweight, and a BMI greater than 30 kg/m^2 is obesity.
b. GWG was categorized according to the 2009 Institute of Medicine (IOM) definition: the recommend optimal ranges according to the mother's prepregnancy BMI: 12.5-18 kg for underweight women (<18.5 kg/m^2), 11.5-16 kg for normal weight women (18.5-24.9 kg/m^2), 7-11.5 kg for overweight women (25-29.9 kg/m^2) and 5-9 kg for obese women (>30 kg/m^2).
c. The mothers having both measurements of BMI and GWG were included for GWG categories.

Supplementary Figure 1. Flowchart for quality control.



References:

1. Barker DJ. Fetal origins of coronary heart disease. *BMJ* 1995; **311**(6998)**:** 171-4.
2. Eriksson JG. The fetal origins hypothesis--10 years on. *BMJ* 2005; **330**(7500)**:** 1096-7.
3. Rich-Edwards JW, Stampfer MJ, Manson JE, Rosner B, Hankinson SE, Colditz GA *et al.* Birth weight and risk of cardiovascular disease in a cohort of women followed up since 1976. *BMJ* 1997; **315**(7105)**:** 396-400.
4. Barker DJ. Obesity and early life. *Obesity reviews : an official journal of the International Association for the Study of Obesity* 2007; **8 Suppl 1:** 45-9.
5. Whitaker RC, Dietz WH. Role of the prenatal environment in the development of obesity. *The Journal of pediatrics* 1998; **132**(5)**:** 768-76.
6. Mannan M, Doi SA, Mamun AA. Association between weight gain during pregnancy and postpartum weight retention and obesity: a bias-adjusted meta-analysis. *Nutrition reviews* 2013; **71**(6)**:** 343-52.
7. Gaillard R, Durmus B, Hofman A, Mackenbach JP, Steegers EA, Jaddoe VW. Risk factors and outcomes of maternal obesity and excessive weight gain during pregnancy. *Obesity (Silver Spring)* 2013; **21**(5)**:** 1046-55.
8. Reynolds RM, Osmond C, Phillips DI, Godfrey KM. Maternal BMI, parity, and pregnancy weight gain: influences on offspring adiposity in young adulthood. *The Journal of clinical endocrinology and metabolism* 2010; **95**(12)**:** 5365-9.
9. Ludwig DS, Currie J. The association between pregnancy weight gain and birthweight: a within-family comparison. *Lancet* 2010; **376**(9745)**:** 984-90.
10. Fraser A, Tilling K, Macdonald-Wallis C, Sattar N, Brion MJ, Benfield L *et al.* Association of maternal weight gain in pregnancy with offspring obesity and metabolic and vascular traits in childhood. *Circulation* 2010; **121**(23)**:** 2557-64.
11. Ludwig DS, Rouse HL, Currie J. Pregnancy weight gain and childhood body weight: a within-family comparison. *PLoS medicine* 2013; **10**(10)**:** e1001521.
12. Rasmussen KM, Yaktine AL. Weight gain during pregnancy:reexamining the guidelines. In: guidelines CtrIpw, (ed). Washington, DC: National Academies Press, 2009.
13. Butte NF, Ellis KJ, Wong WW, Hopkinson JM, Smith EO. Composition of gestational weight gain impacts maternal fat retention and infant birth weight. *American journal of obstetrics and gynecology* 2003; **189**(5)**:** 1423-32.
14. Koo WW, Walters JC, Hockman EM. Body composition in human infants at birth and postnatally. *The Journal of nutrition* 2000; **130**(9)**:** 2188-94.
15. Sparks JW. Human intrauterine growth and nutrient accretion. *Seminars in perinatology* 1984; **8**(2)**:** 74-93.
16. Andersson ES, Silventoinen K, Tynelius P, Nohr EA, Sorensen TI, Rasmussen F. Heritability of Gestational Weight Gain-A Swedish Register-Based Twin Study. *Twin research and human genetics : the official journal of the International Society for Twin Studies* 2015**:** 1-9.
17. Petry CJ, Ong KK, Dunger DB. Does the fetal genotype affect maternal physiology during pregnancy? *Trends in molecular medicine* 2007; **13**(10)**:** 414-21.
18. Stuebe AM, Lyon H, Herring AH, Ghosh J, Wise A, North KE *et al.* Obesity and diabetes genetic variants associated with gestational weight gain. *American journal of obstetrics and gynecology* 2010; **203**(3)**:** 283 e1-17.

19. Lawlor DA, Fraser A, Macdonald-Wallis C, Nelson SM, Palmer TM, Davey Smith G *et al.* Maternal and offspring adiposity-related genetic variants and gestational weight gain. *The American journal of clinical nutrition* 2011; **94**(1)**:** 149-55.
20. Locke AE, Kahali B, Berndt SI, Justice AE, Pers TH, Day FR *et al.* Genetic studies of body mass index yield new insights for obesity biology. *Nature* 2015; **518**(7538)**:** 197-206.
21. Morrison KM, Atkinson SA, Yusuf S, Bourgeois J, McDonald S, McQueen MJ *et al.* The Family Atherosclerosis Monitoring In earLY life (FAMILY) study: rationale, design, and baseline data of a study examining the early determinants of atherosclerosis. *American heart journal* 2009; **158**(4)**:** 533-9.
22. Robiou-du-Pont S, Li A, Christie S, Sohani ZN, Meyre D. Should we have blind faith in bioinformatics software? Illustrations from the SNAP web-based tool. *PloS one* 2015; **10**(3)**:** e0118925.
23. Price AL, Patterson NJ, Plenge RM, Weinblatt ME, Shadick NA, Reich D. Principal components analysis corrects for stratification in genome-wide association studies. *Nature genetics* 2006; **38**(8)**:** 904-9.
24. Dudbridge F. Power and predictive accuracy of polygenic risk scores. *PLoS genetics* 2013; **9**(3)**:** e1003348.
25. Lindstrom ML, Bates DM. Nonlinear mixed effects models for repeated measures data. *Biometrics* 1990; **46**(3)**:** 673-87.
26. Tyrrell J, Richmond RC, Palmer TM, Feenstra B, Rangarajan J, Metrustry S *et al.* Genetic Evidence for Causal Relationships Between Maternal Obesity-Related Traits and Birth Weight. *JAMA : the journal of the American Medical Association* 2016; **315**(11)**:** 1129-40.
27. Feise RJ. Do multiple outcome measures require p-value adjustment? *BMC medical research methodology* 2002; **2:** 8.
28. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MA, Bender D *et al.* PLINK: a tool set for whole-genome association and population-based linkage analyses. *American journal of human genetics* 2007; **81**(3)**:** 559-75.
29. Ihaka R, Genteman R. R: a language for data analysis and graphics. *Journal of Computational and Graphical Statistics* 1996; **5:** 299-314.
30. Davies GA, Maxwell C, McLeod L, Gagnon R, Basso M, Bos H *et al.* SOGC Clinical Practice Guidelines: Obesity in pregnancy. No. 239, February 2010. *International journal of gynaecology and obstetrics: the official organ of the International Federation of Gynaecology and Obstetrics* 2010; **110**(2)**:** 167-73.
31. Schack-Nielsen L, Michaelsen KF, Gamborg M, Mortensen EL, Sorensen TI. Gestational weight gain in relation to offspring body mass index and obesity from infancy through adulthood. *Int J Obes (Lond)* 2010; **34**(1)**:** 67-74.
32. Lawrence GM, Shulman S, Friedlander Y, Sitlani CM, Burger A, Savitsky B *et al.* Associations of maternal pre-pregnancy and gestational body size with offspring longitudinal change in BMI. *Obesity (Silver Spring)* 2014; **22**(4)**:** 1165-71.
33. Branum AM, Parker JD, Keim SA, Schempf AH. Prepregnancy body mass index and gestational weight gain in relation to child body mass index among siblings. *American journal of epidemiology* 2011; **174**(10)**:** 1159-65.

34. Kwak SH, Park BL, Kim H, German MS, Go MJ, Jung HS *et al.* Association of variations in TPH1 and HTR2B with gestational weight gain and measures of obesity. *Obesity (Silver Spring)* 2012; **20**(1)**:** 233-8.
35. Austin MA, Friedlander Y, Newman B, Edwards K, Mayer-Davis EJ, King MC. Genetic influences on changes in body mass index: a longitudinal analysis of women twins. *Obesity research* 1997; **5**(4)**:** 326-31.
36. Bray MS, Loos RJ, McCaffery JM, Ling C, Franks PW, Weinstock GM *et al.* NIH working group report-using genomic information to guide weight management: From universal to precision treatment. *Obesity (Silver Spring)* 2016; **24**(1)**:** 14-22.
37. Karachaliou M, Georgiou V, Roumeliotaki T, Chalkiadaki G, Daraki V, Koinaki S *et al.* Association of trimester-specific gestational weight gain with fetal growth, offspring obesity, and cardiometabolic traits in early childhood. *American journal of obstetrics and gynecology* 2015; **212**(4)**:** 502 e1-14.
38. Mamun AA, O'Callaghan M, Callaway L, Williams G, Najman J, Lawlor DA. Associations of gestational weight gain with offspring body mass index and blood pressure at 21 years of age: evidence from a birth cohort study. *Circulation* 2009; **119**(13)**:** 1720-7.

CHAPTER V: SUMMARY OF NOVEL CONTRIBUTIONS AND FUTURE DIRECTIONS

This thesis addresses some novel questions surrounding the genetics of obesity in the post-GWAS era. This work provides several original contributions and I will highlight them by chapter and discuss future research directions.

Chapter II: Transferability of obesity susceptibility loci across multiple ethnic groups

EpiDREAM is a large multi-ethnic cohort study and well represents the variety of worldwide ethnic populations. BMI is significantly different across six ethnic groups in our study, in line with the evidence that the prevalence of obesity varies across countries and ethnicities although this was not a random sample.[1-4] The major findings of this study are that the risk allele frequencies of most of the tested SNPs are significantly different across ethnic groups and the obesity susceptibility genes are partially generalizable across 6 ethnicities. Most of the SNPs and the GRS displayed associations with BMI that were directionally consistent with previous reports, yet the effects of a few SNPs and GRS on the level of BMI may be influenced by the ethnicity. However, the DNA samples in this study were genotyped using a cardiovascular gene-centric 50K SNP array [5] and only 23 out of 136 currently identified BMI or obesity SNPs were available to be tested.[6] Definitely, such a conclusion drawn from GWAS array data would be more convincing and biologically meaningful. Our results argue for the completion of large-scale GWAS meta-analyses with dense SNP arrays in multi-ethnic designs to capture the universal proxies for associations and eventually identify the causal variants.[7-9] Meanwhile, it would be interesting to explore the identified heterogeneity in some genes and GRS across ethnicities. *NPC1* provides an interesting example. Only was one polymorphism (rs1805081) studied in EpiDREAM and three non-synonymous SNPs (rs1805081, rs1805082, rs1788799) are

in the same LD block in Europeans, but their minor allele frequencies varied dramatically in the different ethnic groups. It is tempting to speculate that these three coding SNPs may have accumulative detrimental effect on *NPC1* function, meaning that haplotype rather than single SNP analyses may better capture the association.[10,11] The effect sizes of GRS were not consistent across ethnic groups. The ethnicity of South Asian and African significantly influence the effects of GRS on the level of BMI. This indicates GWAS of BMI or obesity in these two populations may of great importance.

In addition to different risk allele frequencies, ethnic-specific associations and gene × environment interaction, allelic heterogeneity, different linkage disequilibrium patterns or gene × gene interactions may explain the incomplete generalizability of all known associations across ethnicities.[12] Furthermore, the "thrifty genotype" hypothesis proposed by the geneticist JV Neel in 1962 implies that genetic variants that favor highly efficient fat metabolism and storage may have undergone positive selection during historical periods of erratic food supply.[13] If the hypothesis that an ethnic group experiences positive selection at obesity associated alleles holds, some of the characteristics of a risk variant may develop: high frequency of the derived allele, longer haplotype, and highly differentiated risk allele frequencies across different populations.[14] This may explain in part why some SNPs are associated with BMI in one population but not in another. Thus, a relevant follow-up project would be to examine the signatures of natural selection (including Tajima's D, Fay and Wu's H, EHH, XP-EHH, iHS and Fst) for all known BMI and obesity susceptibility SNPs in different ethnic populations using the 1000 Genomes Project sequencing data.[15,16] If most of the 136 SNPs underwent natural selection in a specific population when exposed to a specific environmental change, this will explain the differences in

the frequencies of risk alleles or genetic structure and can infer that obesity is the consequence of natural selection, supporting the "thrifty genotype" hypothesis.

Chapter III: Parental and child genetic contributions to obesity traits in early life based on 83 loci validated in adults: the FAMILY study

One of the major findings of this study is that SNPs contributing to adult BMI start to exert their effect at birth and in early childhood. The GRS derived from 83 adult BMI SNPs was associated with birth weight after adjustment for important maternal confounding factors including pre-pregnancy BMI, GWG, parity, GDM and smoking status. It is generally accepted that the *in utero* environment plays a critical role in regulating offspring birth weight.[17,18] The results of the association between SNP/GRS and birth weight in previous studies are conflicting, but they did not account for critical confounding variables.[19-21] Furthermore, GWAS for birth weight have demonstrated that other genetic variants, other than BMI loci, may influence fetal growth.[22,23] All these findings indicate that we need to replicate this association in a larger study with the availability of additional maternal confounding variables.

Another major finding is that a parent-of-origin effect of rs3736485 in *DMXL2* is significantly associated with BMI variation from birth to 5 years old in children. Imprinting is one specific type of parent-of-origin effect[24] and its occurrence relies on the stage of development and the tissue in question. It has been observed that the majority of imprinted genes are involved in fetal and placental growth and function.[25,26] Parent-of-origin effects have also been reported to be associated with the development of obesity.[27,28] Imprinted genes are especially sensitive to environmental signals. Because imprinted genes have only a single active copy and no back-up, any epigenetic changes will have a greater impact on gene expression. Therefore, a relevant follow-up project would use candidate gene approach to examine the

methylation status of gene *DMXL2* and other genes showing nominal significant parent-of-origin effects in our exploratory analysis, including *IFNGR1*, *HHIP*, *FTO*, *SEC16B*, *TMEM18* and *C9orf93*. This may provide novel insights into the mechanisms underlying childhood obesity.

Chapter IV: Evidence of a genetic link between pre-pregnancy BMI variation and postpartum weight retention

We found no association between maternal and child GRS and GWG. Although genetic factors have recently been shown to explain 43% of the variation in GWG in the first pregnancy and 26% in the second pregnancy,[29] our results, along with two other studies, show that BMI SNPs are not associated with GWG.[30,31] Two directions may be worth to pursue to identify the genetic variants associated with GWG. One way is to produce high-quality data with accurate measurements of maternal pre-pregnancy BMI and GWG, because most studies in the literature rely on recalled weight values which likely introduce recall error and/or bias.[32-35] Then candidate gene approach can be applied to investigate whether BMI SNPs are associated with GWG. The second approach is to conduct GWAS of GWG to search alternative pathways involving GWG.

We also found a genetic link between pre-pregnancy BMI variation and postpartum weight retention. Maternal GRS is associated with both pre-pregnancy BMI and weight retention at 5 years after delivery. This association needs a replication in a larger study with more accurate measurement of GWG.

Summary:

Taken together, these findings indicate that GWAS of specific ethnic group, children, birth weight and GWG are necessary and essential to look for novel variants and alternative pathways influencing the development of obesity.

Epilogue and personal reflections:

This thesis witnesses my growth from having a limited understanding in the field of genetic epidemiology to being able to write comprehensive genetic epidemiology reviews and to perform sophisticated genetic association studies. I experienced both the challenges and cheerfulness during this journey. I have been fortunate to study in such a stimulating scientific area, participate in wonderful courses and work with students in a stimulating and collaborative learning environment. I am also lucky to work in a harmonized and cooperative environment in Dr. Meyre's lab. In addition, I have gained an appreciation for the rapid expansion of knowledge in genetic epidemiology, from the HapMap Project to Phase I, Phase II and Phase III of the 1000 Genomes Project, from single studies to consortia, from meta-analysis to network meta-analysis and from 32 BMI SNPs to 136, just to name a few. This also inspires me to push forward my research investigations in genetic epidemiology. What I have learned from this experience is that I do not know what the truth is, but I will strive to approach it.

References:

1 Flegal, K. M., Carroll, M. D., Kit, B. K. & Ogden, C. L. Prevalence of obesity and trends in the distribution of body mass index among US adults, 1999-2010. *Jama* **307**, 491-497 (2012).

2 Ogden, C. L., Carroll, M. D., Kit, B. K. & Flegal, K. M. Prevalence of obesity and trends in body mass index among US children and adolescents, 1999-2010. *Jama* **307**, 483-490 (2012).

3 Finucane, M. M. *et al.* National, regional, and global trends in body-mass index since 1980: systematic analysis of health examination surveys and epidemiological studies with 960 country-years and 9.1 million participants. *Lancet* **377**, 557-567, doi:10.1016/S0140-6736(10)62037-5 (2011).

4 Ogden, C. L., Carroll, M. D., Kit, B. K. & Flegal, K. M. Prevalence of childhood and adult obesity in the United States, 2011-2012. *JAMA : the journal of the American Medical Association* **311**, 806-814, doi:10.1001/jama.2014.732 (2014).

5 Keating, B. J. *et al.* Concept, design and implementation of a cardiovascular gene-centric 50 k SNP array for large-scale genomic association studies. *PloS one* **3**, e3583, doi:10.1371/journal.pone.0003583 (2008).

6 Locke, A. E. *et al.* Genetic studies of body mass index yield new insights for obesity biology. *Nature* **518**, 197-206, doi:10.1038/nature14177 (2015).

7 Cooper, R. S., Tayo, B. & Zhu, X. Genome-wide association studies: implications for multiethnic samples. *Human molecular genetics* **17**, R151-155, doi:10.1093/hmg/ddn263 (2008).

8 Pulit, S. L., Voight, B. F. & de Bakker, P. I. Multiethnic genetic association studies improve power for locus discovery. *PloS one* **5**, e12600, doi:10.1371/journal.pone.0012600 (2010).

9 Li, Y. R. & Keating, B. J. Trans-ethnic genome-wide association studies: advantages and challenges of mapping in diverse populations. *Genome medicine* **6**, 91, doi:10.1186/s13073-014-0091-5 (2014).

10 Al-Daghri, N. M. *et al.* Mammalian NPC1 genes may undergo positive selection and human polymorphisms associate with type 2 diabetes. *BMC medicine* **10**, 140, doi:10.1186/1741-7015-10-140 (2012).

11 Tregouet, D. A. *et al.* Genome-wide haplotype association study identifies the SLC22A3-LPAL2-LPA gene cluster as a risk locus for coronary artery disease. *Nature genetics* **41**, 283-285, doi:10.1038/ng.314 (2009).

12 Gabriel, S. B. *et al.* The structure of haplotype blocks in the human genome. *Science* **296**, 2225-2229, doi:10.1126/science.1069424 (2002).

13 Neel, J. V. Diabetes mellitus: a "thrifty" genotype rendered detrimental by "progress"? *American journal of human genetics* **14**, 353-362 (1962).

14 Grossman, S. R. *et al.* A composite of multiple signals distinguishes causal variants in regions of positive selection. *Science* **327**, 883-886, doi:10.1126/science.1183863 (2010).

15 Pybus, M. *et al.* 1000 Genomes Selection Browser 1.0: a genome browser dedicated to signatures of natural selection in modern humans. *Nucleic acids research* **42**, D903-909, doi:10.1093/nar/gkt1188 (2014).

16 Sudmant, P. H. *et al.* An integrated map of structural variation in 2,504 human genomes. *Nature* **526**, 75-81, doi:10.1038/nature15394 (2015).

17 Ehrenberg, H. M., Mercer, B. M. & Catalano, P. M. The influence of obesity and diabetes on the prevalence of macrosomia. *American journal of obstetrics and gynecology* **191**, 964-968, doi:10.1016/j.ajog.2004.05.052 (2004).
18 Bouret, S. G. Early life origins of obesity: role of hypothalamic programming. *Journal of pediatric gastroenterology and nutrition* **48 Suppl 1**, S31-38, doi:10.1097/MPG.0b013e3181977375 (2009).
19 Elks, C. E. *et al.* Genetic markers of adult obesity risk are associated with greater early infancy weight gain and growth. *PLoS medicine* **7**, e1000284, doi:10.1371/journal.pmed.1000284 (2010).
20 Belsky, D. W. *et al.* Polygenic risk, rapid childhood growth, and the development of obesity: evidence from a 4-decade longitudinal study. *Archives of pediatrics & adolescent medicine* **166**, 515-521, doi:10.1001/archpediatrics.2012.131 (2012).
21 Elks, C. E. *et al.* Associations between genetic obesity susceptibility and early postnatal fat and lean mass: an individual participant meta-analysis. *JAMA pediatrics* **168**, 1122-1130, doi:10.1001/jamapediatrics.2014.1619 (2014).
22 Freathy, R. M. *et al.* Variants in ADCY5 and near CCNL1 are associated with fetal growth and birth weight. *Nature genetics* **42**, 430-435, doi:10.1038/ng.567 (2010).
23 Horikoshi, M. *et al.* New loci associated with birth weight identify genetic links between intrauterine growth and adult height and metabolism. *Nature genetics* **45**, 76-82, doi:10.1038/ng.2477 (2013).
24 Lawson, H. A., Cheverud, J. M. & Wolf, J. B. Genomic imprinting and parent-of-origin effects on complex traits. *Nature reviews. Genetics* **14**, 609-617, doi:10.1038/nrg3543 (2013).
25 Abu-Amero, S., Monk, D., Apostolidou, S., Stanier, P. & Moore, G. Imprinted genes and their role in human fetal growth. *Cytogenetic and genome research* **113**, 262-270, doi:10.1159/000090841 (2006).
26 Frost, J. M. & Moore, G. E. The importance of imprinting in the human placenta. *PLoS genetics* **6**, e1001015, doi:10.1371/journal.pgen.1001015 (2010).
27 Lindsay, R. S., Kobes, S., Knowler, W. C., Bennett, P. H. & Hanson, R. L. Genome-wide linkage analysis assessing parent-of-origin effects in the inheritance of type 2 diabetes and BMI in Pima Indians. *Diabetes* **50**, 2850-2857 (2001).
28 Hoggart, C. J. *et al.* Novel approach identifies SNPs in SLC2A10 and KCNK9 with evidence for parent-of-origin effect on body mass index. *PLoS genetics* **10**, e1004508, doi:10.1371/journal.pgen.1004508 (2014).
29 Andersson, E. S. *et al.* Heritability of Gestational Weight Gain-A Swedish Register-Based Twin Study. *Twin research and human genetics : the official journal of the International Society for Twin Studies*, 1-9, doi:10.1017/thg.2015.38 (2015).
30 Stuebe, A. M. *et al.* Obesity and diabetes genetic variants associated with gestational weight gain. *American journal of obstetrics and gynecology* **203**, 283 e281-217, doi:10.1016/j.ajog.2010.06.069 (2010).
31 Lawlor, D. A. *et al.* Maternal and offspring adiposity-related genetic variants and gestational weight gain. *The American journal of clinical nutrition* **94**, 149-155, doi:10.3945/ajcn.110.010751 (2011).
32 Mamun, A. A. *et al.* Associations of gestational weight gain with offspring body mass index and blood pressure at 21 years of age: evidence from a birth cohort study. *Circulation* **119**, 1720-1727, doi:10.1161/CIRCULATIONAHA.108.813436 (2009).

33 Ludwig, D. S., Rouse, H. L. & Currie, J. Pregnancy weight gain and childhood body weight: a within-family comparison. *PLoS medicine* **10**, e1001521, doi:10.1371/journal.pmed.1001521 (2013).

34 Branum, A. M., Parker, J. D., Keim, S. A. & Schempf, A. H. Prepregnancy body mass index and gestational weight gain in relation to child body mass index among siblings. *American journal of epidemiology* **174**, 1159-1165, doi:10.1093/aje/kwr250 (2011).

35 Fraser, A. *et al.* Associations of gestational weight gain with maternal body mass index, waist circumference, and blood pressure measured 16 y after pregnancy: the Avon Longitudinal Study of Parents and Children (ALSPAC). *The American journal of clinical nutrition* **93**, 1285-1292, doi:10.3945/ajcn.110.008326 (2011).

SUPPLEMENTARY CHAPTER VI: JUMP ON THE TRAIN OF PERSONALIZED MEDICINE: A PRIMER FOR NON-GENETICIST CLINICIANS

PART1. FUNDAMENTAL CONCEPTS IN MOLECULAR GENETICS

Aihua Li, David Meyre

ABSTRACT

With the decrease in sequencing cost and the rise of companies providing sequencing services, it is likely that personalized whole-genome sequencing will eventually become an instrument of common medical practice. We write this series of three reviews to help non-geneticist clinicians get ready for the major breakthroughs that are likely to occur in the coming years in the fast-moving field of personalized medicine. This first paper focuses on the fundamental concepts of molecular genetics. We review how recombination occurs during meiosis, how *de novo* genetic variations including single nucleotide polymorphisms (SNPs), insertions and deletions are generated and how they are inherited from one generation to the next. We detail how genetic variants can impact protein expression and function, and summarize the main characteristics of the human genome. We also explain how the achievements of the Human Genome Project, the HapMap Project, and more recently, the 1000 Genome Project, have boosted the identification of genetic variants contributing to common diseases in human populations. The second and third papers will focus on genetic epidemiology and clinical applications in personalized medicine.

Introduction

Most human diseases have a genetic component. Non-genetic clinicians are familiar with single-gene disorders for the simple reason that the medical training, including human genetic courses, mainly refers to Mendelian diseases. An example of single-gene disorder is Huntington disease which is caused by a single mutation in the *HD* gene and that follows the easily recognized pattern of autosomal dominant inheritance across generations.[1] Some clinicians are, however, less comfortable with the principles of genetic contributions to complex disorders, despite the fact that a majority of human diseases (e.g. diabetes, cardiovascular diseases, cancers and psychiatric disorders) fall into this category. Complex diseases are triggered by multiple genetic variants in multiple genes acting in combination with environmental factors, and they typically do not follow any Mendelian patterns of inheritance. This limited knowledge in the medical community is understandable as genetic determinants for complex diseases were uncovered in the last 15 years and new discoveries are ongoing. Two important breakthroughs have revolutionized the search for genetic variants contributing to complex diseases and have boosted the elucidation of complex traits in the last five years. First, the commercialization of high throughput genotyping microarrays has led to the emergence of genome-wide association studies (GWAS) and to an unparalleled harvest of disease-associated loci.[2,3] Since the first report of GWAS in 2005, more than 2000 loci have been conclusively associated with one or more complex traits.[4,5] However, most genetic variants from GWAS can only be correlated with a disease and the underlying mechanism may not be known. Over the past three years, the advent of high-throughput next generation sequencing platforms has led to the availability of whole-exome sequencing experiments which specifically sequence the subset of the human genome that code proteins, and to the tremendous progress in the elucidation of Mendelian and complex disorders.[6,7] With the decrease in sequencing cost and growing patient willingness to participate,

[8] personalized whole-genome sequencing may eventually become an instrument of common medical practice.[9,10] These new perspectives challenge the clinicians to jump into the fast-moving field of personalized medicine, an emerging practice that uses an individual's genetic profile to guide decision-making in regard to the prevention, diagnosis, and treatment of diseases.[11] Despite all the recent 'buzz' around personalized medicine, the potential benefits of genetics in clinical practice are regarded with a certain degree of skepticism by the majority of clinicians.[12,13] Obvious reasons include ethical concerns about privacy and discrimination or negative consequences of genetic testing for the patients (worry and anxiety).[14] A less acknowledged but important reason is that genetics is regarded as a hermetic scientific field. The fact that geneticists use a highly technical language with terms like genome-wide association study (GWAS), single nucleotide polymorphism (SNP), and haplotype, certainly does not help. Ignorance begets fear and clinicians lacking the scientific background in genetic epidemiology may be more prone to mistrust or to scorn the promises and potential applications of genetic discoveries in their fields. Taking that into account, now is the time for clinicians to become more familiar with the key concepts of genetic epidemiology in order to become active participants of the personalized medicine revolution. We intend to write this series of three reviews to help non-geneticist clinicians prepare for the major genetic breakthroughs that are likely to occur in the coming years and to welcome genomic medicine into their spheres of practice with the hope of achieving better prevention and care of human genetic disorders. In this Part I of the series of three reviews, we will detail the basic concepts of molecular genetics in user-friendly language. In the next two reviews we will then discuss the study designs and statistical procedures classically used in genetic epidemiology (in Part II) and the realistic promises and challenges in application of recent genetic discoveries in medicine (in Part III).

DNA, RNA and proteins

It has been known since time immemorial that offspring inherit, to a certain extent, their appearance, characteristics, and personality from their parents. As early as 1920s, a chemical substance called DeoxyriboNucleic Acid (DNA) was identified to carry the genetic information and transmit these characteristics from one generation to another.[15] In 1953, James Watson and Francis Crick confirmed the double helix model of DNA structure which is the fundamental discovery for the central dogma of molecular biology.[16] Nucleotides are the basic units of the complex DNA molecule. They contain three parts of a five-carbon sugar, a phosphate molecule and a nitrogen-containing base, which is either adenine (A), thymine (T), cytosine (C) or guanine (G). Nucleotides polymerize into long chains by phosphodiester bonds. Because phosphodiester bonds link the 3' carbon atom of one sugar to the 5' carbon atom of the next sugar, the 5' end has a terminal sugar residue in which the 5' carbon atom is free and the 3' end has a terminal sugar residue in which the 3' carbon atom is free. Adequate evidence has corroborated that cellular DNA forms a double stranded helix. The two coiled long polynucleotide chains are in an antiparallel formation in which the sugar-phosphate backbones are on the outside of the double helix, and the nitrogenous bases are on the inside and perpendicular to the backbones. Therefore, one strand runs in the direction of 5' to 3', whereas the other runs from 3' to 5'. The hydrogen bonds between pairs of bases join the two strands following a specific and base-pairing rule: A to T only and C to G only. Although most of a cell's DNA in humans is contained in the nucleus, mitochondria have their own independent genome that bears a strong resemblance to bacterial genomes.[17] Every cell in the human body has a complete set of DNA called a genome with the exception of mature red blood cells (erythrocytes), which lack a nucleus and most organelles. A gene is a segment of DNA along the genome encompassing specific regulatory elements (5'-untranslated regions (5'-UTRs) and 3'-untranslated regions (3'-UTRs)), non-coding regions

(introns) and coding-regions (exons) which give instructions to messenger ribonucleic acid (mRNA) in the form of three base-pair sets called codons that assemble amino acids to a functional protein (Figure 1). Therefore, a gene is considered to be the basic unit of heredity. During cell growth and division, DNA replication initiates when two DNA strands unwind at a specific origin and serve as their own templates and synthesize the second copy of each DNA strand with the assistance of DNA polymerase and other enzymes. DNA self-replication is conducted in an extremely accurate manner (less than 1 mismatched nucleotide in 10^7).[18,19] Once an error occurs, a repair system including DNA polymerases, exonuclease and other enzymes will proofread DNA sequence and excise the incorrect base pair, ensuring the stability and high fidelity of DNA within an individual and across generations.[20,21] On the other hand, if the repair system fails, a mismatch will lead to a *de novo* mutation. The DNA composition of the different types of cells in human is basically identical. However, the extent to which a given gene is "converted into" a functional protein may vary greatly in different cell types or even in the same type of cells at different states. Generally speaking, DNA is the instruction book, RNA is a photocopy of a specific page of the book and this page tells the cell how to make the protein. The RNA step ensures that energy is not wasted because the entire book is contained in every cell in the body, but certain cells only need to read certain pages (e.g. a nerve cell and a muscle cell use different sets of genes). When a specific protein is required, a process of transcription, the first step of gene expression, is initiated in which DNA is copied into an intermediate molecule named ribonucleic acid (RNA). One of the DNA strand serves as a template (called template or antisense strand), and RNA synthesis is also oriented in a 5' to 3' direction (corresponding to the N-terminus to C-terminus of the sequence of a polypeptide). The RNA transcript is complementary to the template, just like during DNA replication, except that a nucleobase uracil

(U) pairs with A; therefore it has the same sequence as the non-template strand of DNA (which is called sense strand) except that U replaces T. A proofreading mechanism is also involved in transcription, but it is not as accurate as that of DNA replication.[19] Corresponding introns (non-coding sequencing in the RNA transcript) in a newly synthesized RNA molecule are subsequently removed by RNA splicing and a final mature messenger RNA (mRNA) is produced, which contains only exons (sequences that directly code for amino acids). The mRNA needs to be exported into organelle called ribosomes in the cytoplasm where the proteins are assembled. The sequence of an mRNA molecule is the template used to synthesize the corresponding a protein. The process by which mRNA is converted into a linear sequence of amino acids is called translation, the second step of gene expression. Specific nucleotide triplets, called codons, on the mRNA determine the start, the stop or the addition of an amino acid, leading to the creation of a polypeptide chain. Then a transfer RNA (tRNA), carrying the anticodon sequence (complementary to the codon on the mRNA) and a corresponding amino acid, binds to the codon on the mRNA and delivers the new amino acid to extend the polypeptide being synthesized. The maximum number of combinations of three bases out of four is theoretically 4^3=64 (Table 1). Except for one specific codon (AUG) that initiates the translation of mRNA into protein and 3 codons (UGA, UAG, UAA) that stop translation, 61 out of the 64 triplets encode 20 different amino acids. Most amino acids are represented by more than one codon (e.g., six codons of UUA, UUG, CUU, CUC, CUA, and CUG for leucine). A specific codon always encodes a specific amino acid except in the case of the mitochondrial genome, which has four codons used differently from the nuclear DNA. This determines two important characteristics of the genetic code: specificity and degeneracy (also termed as redundancy). The degeneracy makes the protein more tolerant to some point mutations in coding regions and

accounts for synonymous coding mutations. This means that a substitution of one nucleotide by another nucleotide does not necessarily result in an amino acid change (synonymous mutation) but others do change the coding sequence (non-synonymous mutations). All kinds of biological functions need the participation of proteins. However, the physiological roles of a protein depend on its amino acid sequence, configuration, and modulations from other relevant factors such as regulator proteins, ligands/receptors or substrates. Mutations outside of the coding regions of the gene of interest may rather influence its mRNA expression or stability.

Chromosomes, mitosis and meiosis

Most normal human somatic cells are diploid, and in their nucleus there are 46 continuous DNA molecules and each of them is named a chromosome.[22] The 46 chromosomes make up two sets, and therefore two copies of each chromosome have the same length, same centromere and identical genes and are designated homologs. One homolog is maternally inherited and the other is paternally inherited. Each set has 23 single chromosomes-22 autosomes and an X or Y sex chromosome. A male has an X and Y chromosome pair and female has a pair of X chromosomes (Figure 2). The twenty-two autosomes have been ordered from chromosome 1 to 22 according to the length of DNA base pairs (from the longest to the shortest). The X chromosome is much larger than the Y chromosome. The DNA sequences of two homologous chromosomes are usually not completely identical. DNA in a chromosome is packed in many complex units called nucleosomes consisting of two copies of core histones H2A, H2B, H3 and H4 around which is wound by a fragment of DNA, like many beads (histones) on a string (DNA). A fifth histone H1 is located in the spacer region between any two nucleosomes. Histone H3 and H4 can be modified by post-translational regulation mechanisms such as methylation, acetylation, ubiquitination and phosphorylation.[23,24] These proteins are involved in epigenetic

mechanisms, which determine in part the stable gene expression pattern from cell to cell or from generation to generation in the absence of change to DNA sequences.[25,26] Along each chromosome, a constriction point called the centromere divides the chromosome into two arms: the shorter arm or "p arm" and the longer arm or "q arm". In addition to identifying genetic diseases based on the patterns of G-banding (stained by Giemsa's solution at the metaphase),[27] chromosome arms are useful to describe the location of a specific gene mutations.

There are two types of cell divisions, mitosis and meiosis.[28] Mitosis occurs in the context of body growth, cell differentiation, self-renewal and regeneration in somatic cells. DNA replication and partitioning go along with mitosis, ensuring the maintenance of a diploid chromosome stock (2×23 chromosomes) in daughter cells. Meiosis is a specialized reductive cell division which occurs exclusively in germ cells and gives rise to sperm and egg cells. In a single diploid spermatocyte or oocyte, DNA duplication generates two identical sister chromatids, followed by two DNA segregations and cell divisions known as meiosis I and II. During meiosis I, the homologous chromosomes, which are paired together to form a bivalent may possibly exchange a fragment of DNA between maternal and paternal strands. This process of exchange of genetic material is called recombination (crossover) and is one of the key mechanisms by which genetic diversity between daughter cells is generated. Subsequently, a complete set of 2×23 chromosomes are pulled to either pole and separated to form two haploid cells, each with one of the homologs. Which homolog in a bivalent pair ends up to in which daughter cell is independent and this is called the independent assortment. Independent assortment is the second major mechanism of genetic diversity. Therefore in humans, the total number of possible combinations of chromosomes in one gamete is 2^{23}. Meiosis II is similar to mitosis except that final daughter cells have 23 chromosomes instead of 46. As a result, meiosis eventually produces

four haploid gametes. All eggs have a 23,X chromosome constitution representing 22 autosomes plus a single X chromosome, and 50% of the sperms have a 23,X chromosome constitution and the other half are 23,Y (Figure 3). When a sperm fuses to an egg, a zygote is formed and the diploid chromosomal status is re-established. Taken together, the daughter cells from mitosis are genetically identical, whereas the daughter cells from meiosis are genetically different as a consequence of independent assortment and recombination. As discussed earlier, *de novo* DNA mutation may be caused by a failure in the repair system during DNA replication. In addition to this, an error in combination process may also generate structure abnormalities in DNA. The average number of crossovers per cell is about 55 in males and is approximately 50% more in females, which means crossovers are not rare events. Crossovers are essential in maintaining the genetic variability that is transferred from parent to offspring. The exception again is mitochondrial DNA, which is inherited as a single linked molecule through the female line. It does not undergo recombination. Just as in DNA replication, errors during recombination do occur at a very low frequency, giving rise to translocations, inversions, duplications, or deletions.

Abnormalities of chromosome structure are reported to contribute to a small portion of cases in psychiatric disorders. Balanced translocation is an exchange of chromosome segments between two non-homologous chromosomes. A balanced translocation between chromosome 1 and 11 disturbed *DISC1* gene is associated with increased risk of schizophrenia.[29,30] Chromosome inversion occurs when there are two breaks in one chromosome and the same segment is re-constituted with the orientation inverted. A pericentric inversion on chromosome 9 was found to be associated with schizophrenia.[31]

Characteristics of the human genome

The completion of the Human Genome Project in April 2003 and the 1000 Genomes Project in 2012 has revealed several important characteristics of the Human genome:[32,33] 1) there are about 3 billion of base pairs in the human genome; 2) 99% of nucleotide bases are the same in all humans; 3) an estimated 30,000 genes exist in humans, with an average length of 3000 base pairs; 4) genes represent less than 2% of the human genome; 5) more than 50% of genomic DNA consist of non-repetitive DNA sequences and most of the genes display unique DNA sequences; 6) about 45% of genomic DNA consists of repetitive sequences which are thought to contribute to maintaining chromosome structure; 7) there are 38 million validated SNPs in which a single nucleotide differs at a particular position among 1,092 human genomes.

Genetic variations

The current entire database of human genomic variation was recently derived from a panel of whole-genome sequence data in 1,092 individuals from 14 populations in the context of the 1000 Genomes Project.[33] The next targeted milestone of the 1000 Genomes Project is sequencing the genome of 2,500 individuals from 27 populations across the world.[34] Although 99% of the genomic DNA sequences are identical, 1% still signifies 38 million genetic variants between unrelated individuals, indicating there is one allele variant in every 80 base pairs on average. During the assembling of consensus sequences, differences between (among) the nucleotide sequences of different individuals were noticed. The genetic variants of SNPs represent more than 90% of all human variation.[35] If the frequency of a SNP is greater than 5%, it is considered a common variant or polymorphism. If the frequency is between 1-5%, it is a low-frequency SNP. If the frequency is less than 1% in population, it is defined as a mutation. In addition to SNPs, other genetic variants have been observed in the human genome, including microsatellites, variable number of tandem repeats (VNTR), and copy number variants (CNV).

The 1000 Genomes Project also identified 1.4 million bi-allelic short insertions and deletions, and more than 14,000 large deletions.[33] Because these genetic variants were discovered during the sequence assembly, their locations are inherently known, providing a key resource in mapping genes that predispose to common diseases.

Genetic variants may occur in any region of the genome. A SNP that is located in the coding region without changing the corresponding amino acid is called synonymous, while coding SNPs that lead to changes of the amino acid, shifting of the reading frame or to an earlier stop code are called non-synonymous, frameshift or non-sense, respectively. SNPs found in a non-protein coding area in a gene may influence the protein expression by changing regulatory elements such as transcription factor, binding sites or configuration. In humans, there are usually only two alleles at a SNP location, but three alleles are sometimes reported, such as e2, e3, e4 alleles at the *APOE* gene locus.[36] The most common nomenclature of a SNP uses a unique reference SNP (rs) number. An example is the rs10994336 SNP in the *ANK3* gene that has been associated with bipolar disorder.[37] SNP data are available from publicly accessible resources and are constantly updated, such as dbSNP polymorphism repository, Human Genome Variation Database, the International HapMap Project, SNP consortium or 1000 Genomes Project database.

Microsatellites or short tandem repeats (STR) refer to repeated sequences of less than 10 bp of DNA. When the repeat units have 10-100 nucleotides and the copy number reaches hundreds to thousands, this repeat cluster is referred as a minisatellite or a variable number tandem repeat (VNTR).[22] The number of alleles in microsatellites and minisatellites is usually 5 or more. Though both microsatellites and minisatellites are highly unstable, the majority of the variations have no detrimental clinical consequences. The mutation mechanisms in

microsatellites and minisatellites are different. In minisatellites, mutations occur during homologous recombination at meiosis, but the rate is approximately 10 times greater than that of other DNA sequence. Microsatellites undergo slip-strand mispairing during replication and subsequently the genes in the repair systems are inactivated, leading to expansion of the repeats [22]. This mutation rate is also several of orders of magnitude higher than the mutation processes that lead to SNPs. Some of them result in increased risks of diseases. For example, whereas healthy individuals carry less than 36 repeats of CAG in the *HD* gene, the number of repeats increases to more than 40 in individuals who will develop Huntington disease.[1]

Another type of polymorphism is called copy number variant (CNV). CNV refers to the duplication or reduction of a DNA segment (200 bp to 1.5Mb) and they usually have 2 alleles. CNVs (deletion or duplication) can have important functional consequences and have been convincingly associated with psychiatric disorders such as schizophrenia.[38]

Alleles and genotypes

The location of a DNA sequence or a gene on a chromosome is called a locus. If there is more than one type of nucleotide at a specific locus in a population, each nucleotide is called an allele. Most polymorphic sites have only two alleles, while a few have more than two alleles. Individuals are called homozygotes when the two alleles of homologous chromosomes at a specific locus are identical. When the two alleles are different, individuals are classified as heterozygotes. At bi-allelic SNPs, the allele with higher frequency in a given population is called the major, and the less common one is called minor allele. The three (or more) possible combinations of alleles at a specific locus (e.g. major allele / major allele, major allele/ minor allele, minor allele / minor allele) are called genotypes. Sometimes, a genotype refers to the overall genetic constitution of an individual.

For instance, the SNP rs1024582 in the *CACNA1C* gene is associated with bipolar disorder and schizophrenia.[39] There are two alleles A and G, A being the minor allele with frequency of 33.7% and G being the major allele. The three genotypes of an individual at this locus can be AA, AG or GG. The minor allele A increases the risk of bipolar disorder and schizophrenia.[39]

Haplotypes and linkage disequilibrium

Alleles of different loci are sometimes not independently transmitted from one generation to another. They may be physically linked on the chromosome and the crossovers across generations do not break them apart. Such a cluster of alleles is called a haplotype (Figure 4). The US National Institute of Health initiated the International HapMap Project in 2002 to develop a human haplotype map [40]. In phase I more than 1 million common SNPs were genotyped in 2005 in 270 individuals from four geographically distinct populations, Japanese, Han Chinese, Yoruba of Nigeria and Americans of North Western European ancestry [41]. These data were used to explore the patterns of association among SNPs in the genome, and how these patterns vary across populations. In Phase II HapMap, over 3.1 million SNPs were genotyped to create a second generation human haplotype map.[42]

Linkage disequilibrium (LD) measures the non-random association of alleles at two or more loci that may or may not be on the same chromosome. For instance, we may consider two loci with alleles A1/A2 and B1/B2, A1 and B1 alleles being on one chromosome and A2, B2 alleles being on the other homologous chromosome. The frequency of A1 is 60% and B1 is 30% in a population. If the recombination of the two loci is independent, the expected frequency of the four possible haplotypes A1B1, A1B2, A2B1 and A2B2 would be 18%, 42%, 12% and 21%, respectively. If the distribution of these four haplotypes is consistent with the theoretical

frequency, the alleles are in linkage equilibrium. If the distribution significantly departs from the theoretical frequency, the alleles are in linkage disequilibrium, indicating the two loci are not independent. If one of the alleles is a disease causing allele, the haplotype including this allele is considered as a disease-containing haplotype. In most circumstances, a genetic variant that is found to be associated with a disease is not the functional diseasing causing allele; rather it is a proxy SNP. This indicates that this proxy SNP is in the same LD block with the potential causal SNP which is not genotyped in the array. LD may change over time and the patterns of LD may vary depending on the population. The sizes of LD blocks, which reflect the frequency of recombination, have been reported to be smaller in African than in Asian and European populations (Figure 5).[43] Thus, knowing the LD pattern in a specific ethnic group (e.g. from the HapMap Project) is useful to refine the association signal and ultimately lead to the discovery of the causal variant.[44,45] Many other factors may influence LD patterns, including random genetic drift, population growth, admixture, inbreeding, natural selection, and *de novo* mutation.[46]

Using the example of haplotype given above, one statistical test to measure LD is $D'=D/Dmax=(P_{A1B1}-P_{A1}P_{B1})/Dmax$, where Dmax is the maximum difference between P_{A1B1} and $P_{A1}P_{B1}$. D' ranges from -1 to 1. One or -1 denotes there is no recombination between two loci A and B, and 0 indicates that A and B are in linkage equilibrium. If the allele frequencies of A1 and B1 are similar, a high D' value indicates A is a good surrogate for B. However, if the sample size is small or one allele is rare, D' will be inflated. There is a second measurement of LD, using the squared coefficient of determination r^2 (ranging from 0 to 1). r^2 takes into account the sample size and allele frequency. Therefore, D' is extensively used by population geneticist to assess recombination patterns such as defining haplotype patterns, whereas r^2 is a more appropriate measure of linkage disequilibrium in association studies.[47] For example, two SNPS can display a

D' value of 0.85 and a r^2 value of 0.18. In an association study, these two SNPs cannot be tagged or substituted for each other because of low r^2. Pairwise measurement of LD for neighboring SNPs are used to group more than 2 loci into a haplotype termed LD block if the values of D' between any two SNPs within the group are above a certain threshold (e.g. D' > 0.8). This knowledge is essential to guide the design of whole-genome SNP genotyping arrays because carefully selecting a single or a few SNPs representing a haplotype block due to their strong associations can be used to identify an important haplotype, rather than genotyping all the SNPs in this haplotype (Figure 5). The common SNPs in commercial genotyping arrays captures untyped common variation with an average maximum r^2 (a correlation coefficient between genotyped and untyped SNPs) from 0.9 to 0.96 depending on the population. Therefore, the advances from Phase II HapMap, in combination with increased density of high-throughput technology and capability of imputation of untyped SNPs, greatly improved the power of association studies. HapMap 3 was completed in 2009 and it genotyped 1.6 million common and rare variants including CNVs in 1,184 reference individuals from 11 global populations.[48] The integrated map of genetic variation from the complete HapMap data and the 1000 Genomes Project [33] enables analysis of common and rare variants and CNVs in populations of different ethnic background. For instance, Sung and colleagues recently derived the genotypic distribution of 6.7 million SNPs from the information of 324,607 SNPs genotyped in their sample, using the 1000 Genome reference panel.[49]

Conclusions

Having a stronger background in molecular genetics, we are ready to discuss the subtle concepts of genetic epidemiology including study design implementation, gene identification strategies, genetic marker selection, genotyping strategies, data analysis, data interpretation and

their potential applications in the context of personalized medicine. The two next articles in this series will review these topics.

Conflict of interest

The authors confirm that this article content has no conflict of interest.

Acknowledgements

We thank Jackie Hudson and Arkan Al Abadi for editing of the manuscript, and the reviewers for their helpful comments. David Meyre is supported by a Tier 2 Canada Research Chair. Aihua Li is supported by a Queen Elizabeth II Graduate Scholarship in Science and Technology.

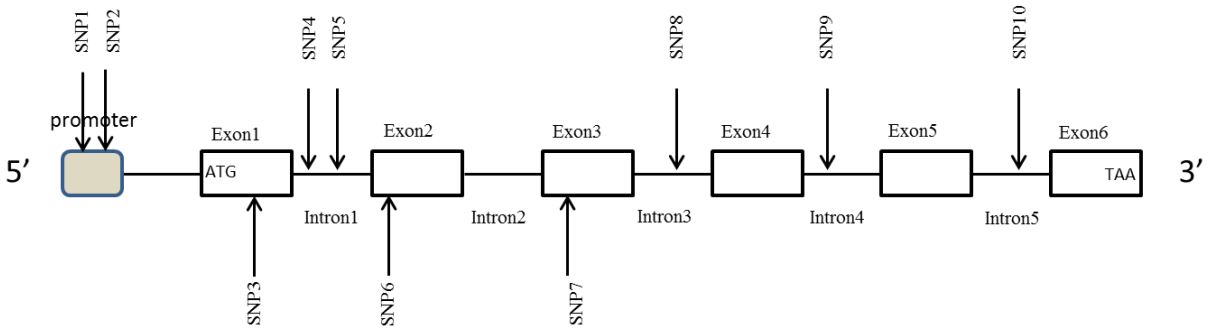


Figure 1. Schematic gene structure. This gene has 5 introns and 6 exons. It is assumed there are 10 SNPs along the gene which may locate at promoter, introns or exons.

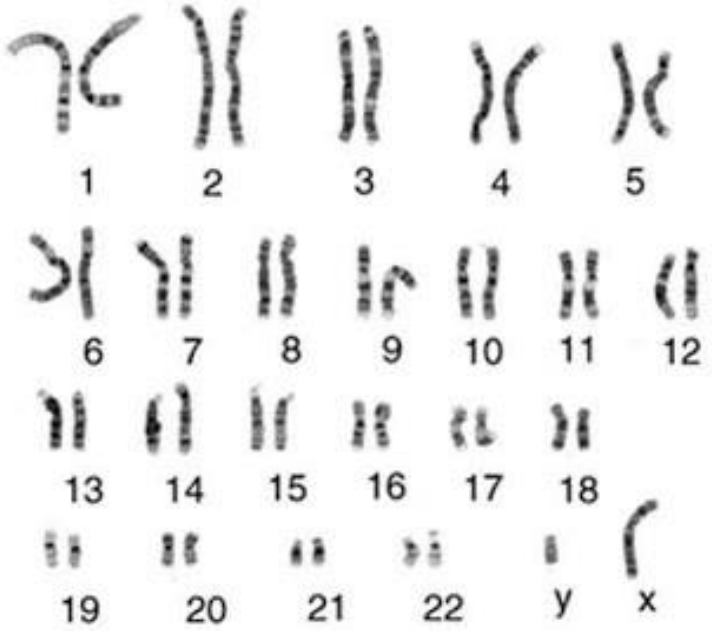


Figure 2. A human male karyotype with Giemsa banding. The autosomes are arranged from 1 to 22 according to their length. Sexual X and Y chromosomes are displayed separately.

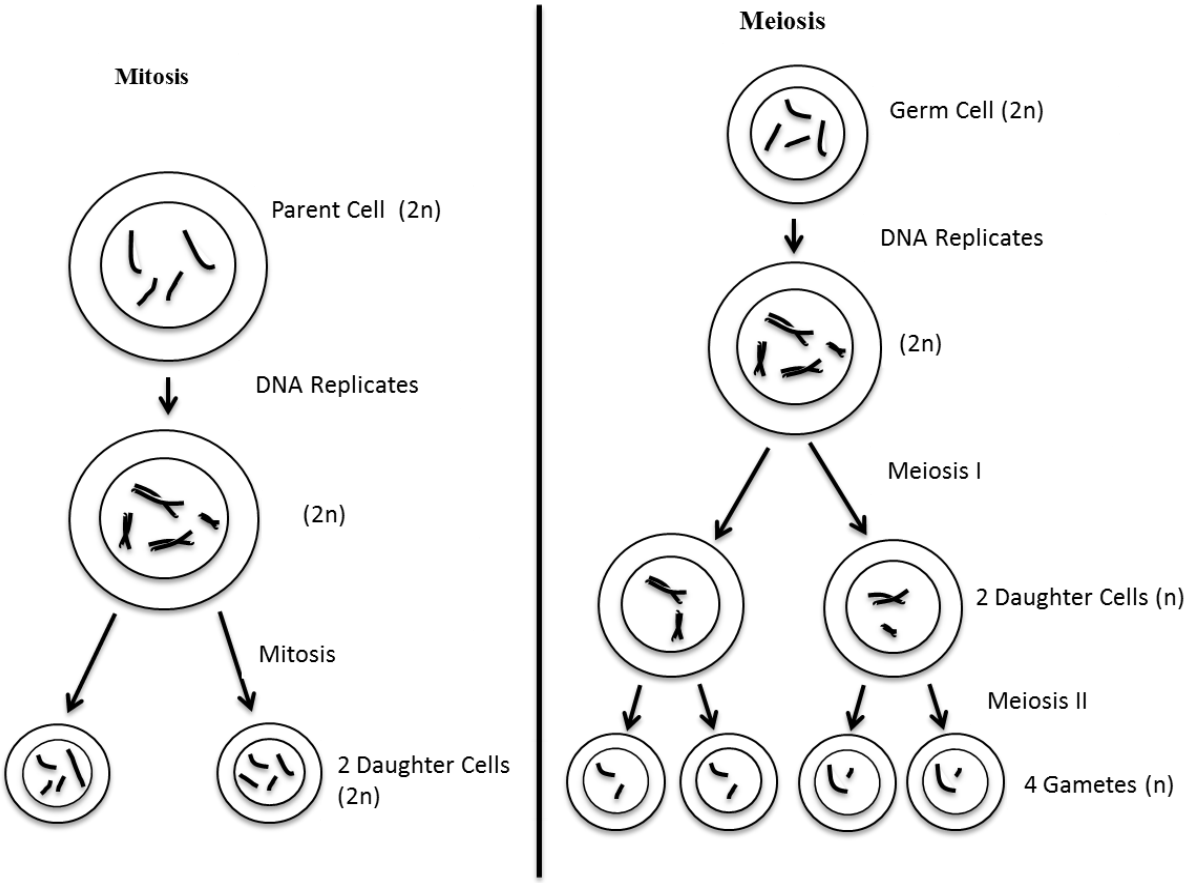


Figure 3. Mitosis and Meiosis. In mitosis, one cell produces two identical daughter cells through DNA duplication, and division. In meiosis, one diploid germ cell gives rise to four haploid gametes through DNA duplication, and two cell divisions (meiosis I and meiosis II). Four chromosome pairs are shown as demonstrations.

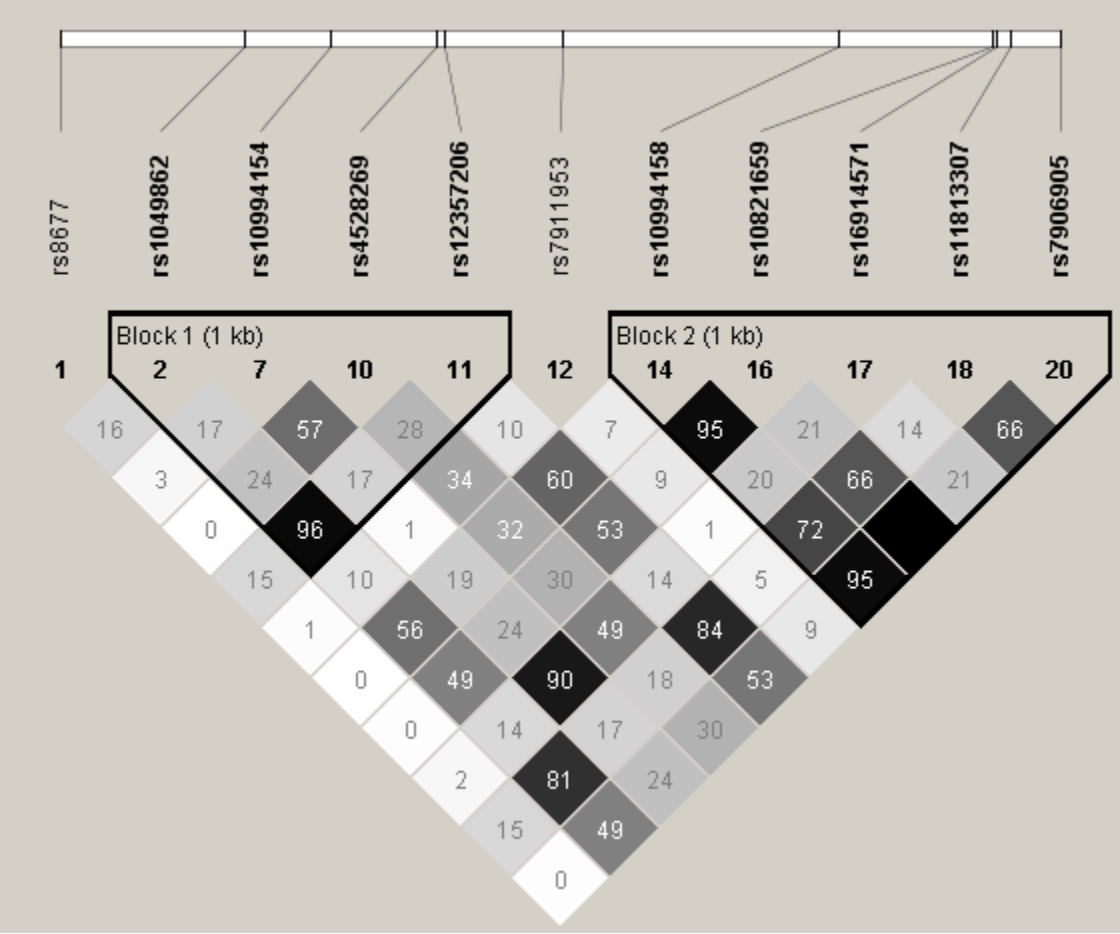


Figure 4. Schematics of linkage disequilibrium (LD) plot. LD blocks among the 11 SNPs in *ANK3* gene are shown. The LD between the SNPs is measured as r^2 and shown (× 100) in the diamond at the intersection of the diagonals from each SNP. $r^2 = 0$ is shown as white, $0 < r^2 < 1$ is shown in gray and $r^2 = 1$ is shown in black. The top shows the relative physical positions of the SNPs on the chromosome 10. Two haplotype blocks (outlined in bold black line) indicate markers that are in high LD.

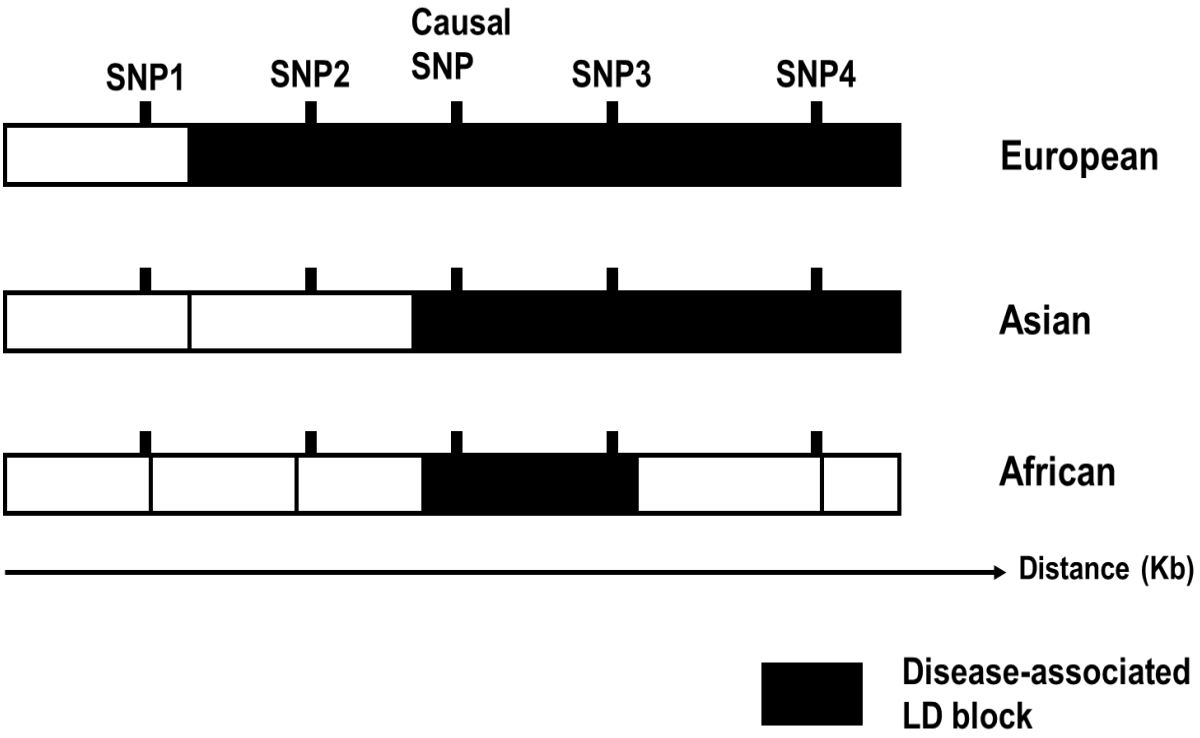


Figure 5. Linkage disequilibrium patterns in different ethnic/racial groups. The size of the LD block where locates the causal SNP is smaller in African than in Asian and European populations. SNP2, SNP3 and SNP4 are proxy SNPs.

FIRST LETTER	SECOND LETTER U	SECOND LETTER C	SECOND LETTER A	SECOND LETTER G	THIRD LETTER
U	UUU Phe UUC Phe UUA Leu UUG Leu	UCU Ser UCC Ser UCA Ser UCG Ser	UAU Tyr UAC Tyr UAA Stop UAG Stop	UGU Cys UGC Cys UGA Stop UGG Trp	U C A G
C	CUU Leu CUC Leu CUA Leu CUG Leu	CCU Pro CCC Pro CCA Pro CCG Pro	CAU His CAC His CAA Gln CAU Gln	CGU Arg CGC Arg CGA Arg CGG Arg	U C A G
A	AUU Ile AUC Ile AUA Ile AUG Met	ACU Thr ACC Thr ACA Thr ACG Thr	AAU Asn AAC Asn AAA Lys AAG Lys	AGU Ser AGC Ser AGA Arg AGG Arg	U C A G
G	GUU Val GUC Val GUA Val GUG Val	GCU Ala GCC Ala GCA Ala GCG Ala	GAU Asp GAC Asp GAA Glu GAG Glu	GGU Gly GGC Gly GGA Gly GGG Gly	U C A G

Table 1. The genetic codes. Sixty-four different combinations of triplet codons are derived from 4 unique bases. Except ATG for the start codon and TAG, GTA, TAA for stop codons, each codon codes for one of the 20 amino acids.

GLOSSARY

Mendelian diseases: Phenotypes that are caused by a single gene mutation and display a clear pattern of inheritance

Complex diseases: Phenotypes that are caused by multiple genetic variants, environmental risk factors and interplays between them. They do not exhibit classic patterns of Mendelian inheritance

Genome-wide association studies: A study evaluating simultaneously associations between a dense subset of genetic variants theoretically covering the whole genome genetic diversity and a phenotype of interest

Single nucleotide polymorphism (SNP): A DNA variant in which a single base pair changes at a particular position compared with a "wild-type" allele

Gene: A segment of DNA embedding specific regulatory elements, non-coding regions and coding-regions which give instruction how amino acids assemble to a protein

Genotype: The genetic constitution at a specific locus or sometimes the overall genetic constitution of an individual

Allele: Each type of nucleotide at a given locus in a DNA fragment if there are two or more than two different types of nucleotides

Locus: The unique location on a chromosome at which a SNP or a gene is located

Homozygote: Individuals in whom the two alleles on the homologous chromosomes at a specific locus are identical

Heterozygote: Individuals in whom the two alleles on the homologous chromosomes at a specific locus are different

Linkage disequilibrium: A measure of non-random association between alleles at different loci

References

1 A novel gene containing a trinucleotide repeat that is expanded and unstable on Huntington's disease chromosomes. The Huntington's Disease Collaborative Research Group. *Cell* **72**, 971-983 (1993).

2 Ewis, A. A. *et al.* A history of microarrays in biomedicine. *Expert Rev Mol Diagn* **5**, 315-328, doi:10.1586/14737159.5.3.315 (2005).

3 The human genome at ten. *Nature* **464**, 649-650, doi:10.1038/464649a (2010).

4 Klein, R. J. *et al.* Complement factor H polymorphism in age-related macular degeneration. *Science* **308**, 385-389, doi:10.1126/science.1109557 (2005).

5 Visscher, P. M., Brown, M. A., McCarthy, M. I. & Yang, J. Five years of GWAS discovery. *American journal of human genetics* **90**, 7-24, doi:10.1016/j.ajhg.2011.11.029 (2012).

6 Bamshad, M. J. *et al.* Exome sequencing as a tool for Mendelian disease gene discovery. *Nat Rev Genet* **12**, 745-755 (2011).

7 Kiezun, A. *et al.* Exome sequencing and the genetic basis of complex traits. *Nat. Genet.* **44**, 623-630, doi:10.1038/ng.2303 (2012).

8 Hood, L. & Friend, S. H. Predictive, personalized, preventive, participatory (P4) cancer medicine. *Nat Rev Clin Oncol* **8**, 184-187, doi:10.1038/nrclinonc.2010.227 (2011).

9 Pettersson, E., Lundeberg, J. & Ahmadian, A. Generations of sequencing technologies. *Genomics* **93**, 105-111, doi:10.1016/j.ygeno.2008.10.003 (2009).

10 Patterson, K. 1000 genomes: a world of variation. *Circ. Res.* **108**, 534-536, doi:10.1161/RES.0b013e31821470fe (2011).

11 Gonzaga-Jauregui, C., Lupski, J. R. & Gibbs, R. A. Human genome sequencing in health and disease. *Annu Rev Med* **63**, 35-61 (2012).

12 Hindorff, L. A. *et al.* Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *Proc. Natl. Acad. Sci. U. S. A.* **106**, 9362-9367, doi:10.1073/pnas.0903103106 (2009).

13 Wacholder, S. *et al.* Performance of common genetic variants in breast-cancer risk models. *N. Engl. J. Med.* **362**, 986-993, doi:10.1056/NEJMoa0907727 (2010).

14 Scheuner, M. T., Sieverding, P. & Shekelle, P. G. Delivery of genomic medicine for common chronic adult diseases: a systematic review. *JAMA : the journal of the American Medical Association* **299**, 1320-1334, doi:10.1001/jama.299.11.1320 (2008).

15 Avery, O. T., Macleod, C. M. & McCarty, M. Studies on the Chemical Nature of the Substance Inducing Transformation of Pneumococcal Types : Induction of Transformation by a Desoxyribonucleic Acid Fraction Isolated from Pneumococcus Type Iii. *J. Exp. Med.* **79**, 137-158 (1944).

16 Watson, J. D. & Crick, F. H. Molecular structure of nucleic acids; a structure for deoxyribose nucleic acid. *Nature* **171**, 737-738 (1953).

17 Andersson, S. G., Karlberg, O., Canback, B. & Kurland, C. G. On the origin of mitochondria: a genomics perspective. *Philos. Trans. R. Soc. Lond. B. Biol. Sci.* **358**, 165-177; discussion 177-169, doi:10.1098/rstb.2002.1193 (2003).

18 McCulloch, S. D. & Kunkel, T. A. The fidelity of DNA synthesis by eukaryotic replicative and translesion synthesis polymerases. *Cell Res.* **18**, 148-161, doi:10.1038/cr.2008.4 (2008).

19 Berg JM, T. J., Stryer L. *Biochemistry*. 7th edn, (W.H. Freeman and Company, 2012).

20 Sancar, A., Lindsey-Boltz, L. A., Unsal-Kacmaz, K. & Linn, S. Molecular mechanisms of mammalian DNA repair and the DNA damage checkpoints. *Annu. Rev. Biochem.* **73**, 39-85, doi:10.1146/annurev.biochem.73.011303.073723 (2004).

21 Wood, R. D., Mitchell, M., Sgouros, J. & Lindahl, T. Human DNA repair genes. *Science* **291**, 1284-1289, doi:10.1126/science.1056154 (2001).

22 Krebs J. E., G. E. S., Kilpatrick S.T. *Lewin's Gene X*. 10th edn, (Jones and Bartlet Publishers, 2011).

23 Luger, K., Mader, A. W., Richmond, R. K., Sargent, D. F. & Richmond, T. J. Crystal structure of the nucleosome core particle at 2.8 A resolution. *Nature* **389**, 251-260, doi:10.1038/38444 (1997).

24 Thompson, L. L., Guppy, B. J., Sawchuk, L., Davie, J. R. & McManus, K. J. Regulation of chromatin structure via histone post-translational modification and the link to carcinogenesis. *Cancer Metastasis Rev.*, doi:10.1007/s10555-013-9434-8 (2013).

25 Bonasio, R., Tu, S. & Reinberg, D. Molecular signals of epigenetic states. *Science* **330**, 612-616, doi:10.1126/science.1191078 (2010).

26 Martin, C. & Zhang, Y. Mechanisms of epigenetic inheritance. *Curr. Opin. Cell Biol.* **19**, 266-272, doi:10.1016/j.ceb.2007.04.002 (2007).

27 Speicher, M. R. & Carter, N. P. The new cytogenetics: blurring the boundaries with molecular biology. *Nat Rev Genet* **6**, 782-792, doi:10.1038/nrg1692 (2005).

28 Nussbaum RL, M. R., Willard HF, Hamosh A. *Thompson & Thompson Genetics in Medicine*. 7th edn, (2007).

29 St Clair, D. *et al.* Association within a family of a balanced autosomal translocation with major mental illness. *Lancet* **336**, 13-16 (1990).

30 Millar, J. K. *et al.* Disruption of two novel genes by a translocation co-segregating with schizophrenia. *Hum. Mol. Genet.* **9**, 1415-1423 (2000).

31 Kunugi, H., Lee, K. B. & Nanko, S. Cytogenetic findings in 250 schizophrenics: evidence confirming an excess of the X chromosome aneuploidies and pericentric inversion of chromosome 9. *Schizophr. Res.* **40**, 43-47 (1999).

32 Yang, Y. W., Wen, G. D., Wu, C. J., Ren, Q. L. & Wu, P. D. Preparation of natural alpha-tocopherol from non-alpha-tocopherols. *Journal of Zhejiang University. Science* **5**, 1524-1527, doi:10.1631/jzus.2004.1524 (2004).

33 Abecasis, G. R. *et al.* An integrated map of genetic variation from 1,092 human genomes. *Nature* **491**, 56-65, doi:10.1038/nature11632 (2012).

34 Pennisi, E. Genomics. 1000 Genomes Project gives new map of genetic diversity. *Science* **330**, 574-575, doi:10.1126/science.330.6004.574 (2010).

35 Reich, D. E., Gabriel, S. B. & Altshuler, D. Quality and completeness of SNP databases. *Nat. Genet.* **33**, 457-458, doi:10.1038/ng1133 (2003).

36 Corbo, R. M. & Scacchi, R. Apolipoprotein E (APOE) allele distribution in the world. Is APOE*4 a 'thrifty' allele? *Ann. Hum. Genet.* **63**, 301-310 (1999).

37 Ferreira, M. A. *et al.* Collaborative genome-wide association analysis supports a role for ANK3 and CACNA1C in bipolar disorder. *Nat. Genet.* **40**, 1056-1058, doi:10.1038/ng.209 (2008).

38 Guha, S. *et al.* Implication of a rare deletion at distal 16p11.2 in schizophrenia. *JAMA Psychiatry* **70**, 253-260, doi:10.1001/2013.jamapsychiatry.71 (2013).

39 Ripke, S. *et al.* Genome-wide association study identifies five new schizophrenia loci. *Nature genetics* **43**, 969-976, doi:10.1038/ng.940 (2011).

40 The International HapMap Project. *Nature* **426**, 789-796, doi:10.1038/nature02168 (2003).
41 A haplotype map of the human genome. *Nature* **437**, 1299-1320, doi:10.1038/nature04226 (2005).
42 Frazer, K. A. *et al.* A second generation human haplotype map of over 3.1 million SNPs. *Nature* **449**, 851-861, doi:10.1038/nature06258 (2007).
43 Gabriel, S. B. *et al.* The structure of haplotype blocks in the human genome. *Science* **296**, 2225-2229, doi:10.1126/science.1069424 (2002).
44 Helgason, A. *et al.* Refining the impact of TCF7L2 gene variants on type 2 diabetes and adaptive evolution. *Nat. Genet.* **39**, 218-225, doi:10.1038/ng1960 (2007).
45 Hassanein, M. T. *et al.* Fine mapping of the association with obesity at the FTO locus in African-derived populations. *Hum. Mol. Genet.* **19**, 2907-2916, doi:10.1093/hmg/ddq178 (2010).
46 Ardlie, K. G., Kruglyak, L. & Seielstad, M. Patterns of linkage disequilibrium in the human genome. *Nat Rev Genet* **3**, 299-309, doi:10.1038/nrg777 (2002).
47 Mueller, J. C. Linkage disequilibrium for different scales and applications. *Brief Bioinform* **5**, 355-364 (2004).
48 Altshuler, D. M. *et al.* Integrating common and rare genetic variation in diverse human populations. *Nature* **467**, 52-58, doi:10.1038/nature09298 (2010).
49 Sung, Y. J., Wang, L., Rankinen, T., Bouchard, C. & Rao, D. C. Performance of genotype imputations using data from the 1000 Genomes Project. *Hum. Hered.* **73**, 18-25, doi:10.1159/000334084 (2012).

SUPPLEMENTARY CHAPTERVII: JUMP ON THE TRAIN OF PERSONALIZED MEDICINE: A PRIMER FOR NON-GENETICIST CLINICIANS

PART2. FUNDAMENTAL CONCEPTS IN GENETIC EPIDEMIOLOGY

Aihua Li, David Meyre

Abstract: With the decrease in sequencing costs, personalized genome sequencing will eventually become common in medical practice. We therefore write this series of three reviews to help non-geneticist clinicians to jump into the fast-moving field of personalized medicine. In the first article of this series, we reviewed the fundamental concepts in molecular genetics. In this second article, we cover the key concepts and methods in genetic epidemiology including the classification of genetic disorders, study designs and their implementation, genetic marker selection, genotyping and sequencing technologies, gene identification strategies, data analyses and data interpretation. This review will help the reader critically appraise a genetic association study. In the next article, we will discuss the clinical applications of genetic epidemiology in the personalized medicine area.

What is genetic epidemiology?

Genetic epidemiology emerged in the 1960s at the crossroads of multiple disciplines such as molecular genetics, epidemiology and biostatistics. Genetic epidemiology studies the role of genetic factors in determining health and disease in families and in populations, as well as the interplay of genetic determinants with specific environmental exposures. Morton elegantly defined genetic epidemiology as "a science which deals with the etiology, distribution, and control of disease in groups of relatives and with inherited causes of disease in populations".[1] In this article, we aim to illustrate how to identify genetic variants associated with a disease including the relevant concepts, study designs and statistical analyses classically used in genetic epidemiology. Due to the complexity of the steps needed to explore genetic variation in common diseases, we provide a diagram which outlines how this paper is structured (Figure 1). The questions illustrate the step by step procedures to conduct genetic epidemiology research; the methods show the parameters which are measured, and the third column lists the study designs most commonly used in genetic epidemiology.

Phenotype

A phenotype represents the observable physical or biochemical characteristics of an individual or a group of organisms, as determined by both genetic make-up and environmental influences. In human genetics, phenotypes refer to traits as diverse as diseases, biochemical measurements or the levels of expression of a gene transcript. A phenotype can be binary (e.g. presence or absence of schizophrenia), categorical (e.g. personality disorders) or quantitative (e.g. hippocampal volume).[2] The ideal phenotype should be clinically and biologically relevant, not too rare, and inexpensive, thus allowing large-scale discovery and replication studies

feasible. It should be well defined so that measurement errors, misclassification and heterogeneity can be minimized. [3]

Modes of inheritance

There are five basic patterns of Mendelian inheritances (Figure 1). Punnett squares which are used to predict the chance of genetic disease in children for parents with an increased risk are presented in Figure 2. First, autosomal dominant inheritance explains more than 50% of Mendelian diseases. One deleterious copy of the gene is sufficient to confer the disease. Both males and females have 50% risk of being affected and the disease occurs in every generation. Huntington's disease follows an autosomal dominant mode of inheritance.[4] If each copy of the gene contributes to the trait and the heterozygote generates an intermediate phenotype, this is called co-dominant (e.g. ABO blood type) or additive inheritance (e.g. genetic effects from most risk alleles). Generally speaking, the concept of co-dominant includes additive models. If the trait is quantitative, when the heterozygotes have a mean level which is the average of two types of homozygotes means it is an additive model. An autosomal recessive disease only occurs when an individual harbors two deleterious copies at the locus. In most cases, both parents of the affected person are healthy heterozygous carriers of risk allele.[5] In accordance with Mendel's Laws, every offspring has a 25% probability of developing the disease. Offspring of consanguineous marriages are more likely to develop autosomal recessive disorders because consanguinity increases the risk to inherit two identical mutations.[5] Sometimes, individuals develop autosomal recessive disorders in non-consanguineous pedigrees because they carry two mutant alleles for the same gene, but with those two alleles being different from each other (for example, two mutant alleles are at different loci). This phenomenon is called compound heterozygosity. Compound heterozygotes usually get ill later in life with less severe symptoms.

Phenylketonuria, an inherited disorder that is characterized by seizures, delayed development, behavioral problems and psychiatric disorders, follows an autosomal recessive pattern of inheritance.[6] The fourth mode is X-linked recessive inheritance. A mutation in a gene located on the X chromosome causes a disease in males who are also called hemizygous (the gene mutation only occurs on the X chromosome) and in females who carry the mutant on each of the X chromosome. Thus, X-linked recessive diseases, such as X-linked mental retardation,[7] affect more males than females. On the other hand, if only the father is affected, none of his sons will develop the disease, whereas all his daughters will carry the mutant allele. Fifth, X-linked dominant disorders are less common compared with X-linked recessive type. All the offspring of affected females have a 50% chance that they will inherit from such a disease whereas all the daughters of an affected male will develop it. Usually, males are affected more severely than females as observed in Fragile X syndrome.[5] However, more female patients with X-linked dominant disorders are sometimes observed. In the Rett syndrome for instance, 50% of the males with the mutant allele miscarry before birth.[8]

Departure from classical Mendelian patterns of inheritance often occurs and can be explained by different mechanisms that include incomplete penetrance, variable expressivity, genomic imprinting effects, mosaicism, mitochondrial inheritance, *de novo* mutations, overdominance or digenic inheritance. Incomplete penetrance refers to a situation in which the occurrence of the disease in individuals who harbour the same disease-causing allele is less than 100%. Although the mutant allele does not inevitably cause the disease, it is still passed to the offspring. On the other hand, individuals who inherit the same mutant allele may experience a different level of severity of the disease. This phenomenon is called variable expressivity. Incomplete penetrance and variable expressivity are commonly observed in autosomal dominant

and X-linked recessive disorders and can be explained by the effect of modifying genes or by differential regulation of gene expression.[9] For instance, microdeletion of 15q13.3 shows incomplete penetrance of autism and a wide spectrum of mental retardation.[10,11] Genomic imprinting is a phenomenon by which imprinted alleles are silenced such that the genes are expressed in a parent-of-origin-specific and mono-allelic manner.[12] In other words, the genes are expressed only from the non-imprinted allele inherited from the mother (maternal imprinting) or from the father (paternal imprinting). Imprinting is an epigenetic process that involves DNA methylation or histone methylation mechanisms with no alteration of the genetic sequence [12]. These epigenetic marks are established in the germline cells and are maintained throughout all somatic cells of an organism. Genomic imprinting has an important role in fetal and placental growth and development.[13,14] Angelman or Prader–Willi syndromes are classical examples of genetic defects in genes submitted to parental imprinting.[15] When the paternal copy is imprinted and silenced, a deletion of 15q12 inherited from the mother causes Angelman syndrome. On the other contrary, if the maternal copy is imprinted and silenced, the deletion inherited from the father leads to Prader-Willi syndrome. Genomic DNA in every single cell of an individual is the same. But, if a mutation occurs during mitotic cell divisions of the developing fetus, it can give rise to mosaicism of at least two populations of cells (somatic or germline) that are genetically different. Mosaicism may explain a substantial fraction of unusual clinical observations, for example, mosaic structural variations are two-fold more frequent in schizophrenic cases than in controls.[16] A very small but functionally important portion of genomic DNA resides in the cytoplasm of mitochondria. Mitochondrial DNA can only be inherited from the mother, because mitochondria present in sperm are eliminated from the embryo. Another unique feature of mitochondrial DNA is that it is randomly distributed into daughter cells during mitosis and

meiosis, leading to remarkably variable expressivity in mitochondrial diseases. Schizophrenia and bipolar disease have been reported to present excessive maternal inheritance, and mutations in mitochondrial DNA are also related to these disorders.[17-19] There is a probability of 10^{-6} to have a *de novo* mutation in any types of inheritance modes. The *de novo* mutations in autosomal recessive diseases are more frequent than autosomal dominant and X-linked disorders. The over-dominant mode of inheritance is rarely observed in humans.[20] In that model, the mean of the heterozygotes is higher than the mean of two types of homozygotes. Sometimes, a disease occurs only if two mutations in two different genes are present in the same individual which belongs to a digenic mode of inheritance.[21] Digenic inheritance has been reported in severe familial forms of insulin resistance.[22] Most of the time, non-Mendelian modes of inheritance observed in human diseases result from polygenic genetic architectures (see the section below).

Familial aggregation, heritability and segregation analyses

Clinicians are used to collecting family history information related to a particular disease in order to assess whether a person is at risk of developing similar problems. A more frequent recurrence of a disease in a pedigree may be because of their shared environmental exposure (e.g. toxin), however, most of the time it indicates that the disease has a hereditary component. Familial aggregation analysis answers the question of whether the relatives of the affected person (proband) are more likely to suffer the same disease compared with the general population at a specific point of time. If the phenotype is qualitative, familial aggregation is measured by recurrence risk ratio in relatives λ_R (Table 2).[23] A greater λ is expected in first degree than in second degree relatives of the affected person if genetic factors play a role in the occurrence of the disease.[23] A λ_R of 2 and above is a good indication that the causes of the underlying familial aggregation warrant further study.[24] Very high relative risk ratios λ_S for siblings have been

observed for autism (λ_S=75), schizophrenia (λ_S=10) and bipolar disorder (λ_S=15) [25] in which shared genes greatly contribute to the familial recurrence of the diseases. If the phenotype is quantitative, familial aggregation is measured by intra-family correlation coefficients (ICC) which is the proportion of the total variance in the phenotype attributed to differences between families. The larger λ_R or ICC, the greater the familial component of the trait will be.[23] Neither λ_R nor ICC distinguishes genetic from environmental components, because family relatives share not only genes but also similar environment. For example, familial aggregation for depression could be due to either shared genes or similar environmental factors, such as socioeconomic status of the family.

Heritability reflects the proportion of total phenotypic variability explained by genetic variance in a particular population at a specific time. When only additive genetic effects are accounted for in the genetic variance, heritability is named narrow-sense heritability or just heritability (h^2); when all genetic variance from additive, dominant and epistatic (gene × gene interaction) effects is accounted for, heritability is defined as broad-sense heritability (H^2).[26] Twin and adoption studies are ideal experimental designs to estimate heritability because of their natural separation of genetic and environmental components.[26] In twin studies, monozygotic (MZ) twins share 100% of their genome whereas dizygotic (DZ) twins share 50%. If genetic factors play a role in the phenotype, the correlation coefficient of the phenotype between MZs should be significantly higher than in DZs. The calculation of the heritability is listed in Table 2. These calculations are based on the assumption that MZ pairs and DZ pairs grow up in an identical environment.[27] There is a methodological concern that twins are not representatives of the general population.[28] In practice, the assumption of identical environment in twin studies may be difficult to hold. Twins may display difference in delivery process, special life events,

and interactions with teachers or friends. In an alternative adoption study, a biological parent and an adopted-away offspring, or a full sibling and an adopted-away full sibling share 50% of genes that attribute to their resemblance in the trait. The heritability in this situation assumes they have different environmental exposures (Table 2). When the traits are binary, a liability scale model in which a disease arises when the determined probability exceeds a certain threshold, or the statistical models developed for quantitative traits may be applied.[29,30] Although the assumptions underlying the twin and adoption studies are not always met in practice, many important findings have been discovered from such designs.[31] More recently, structural equation models have been used to estimate heritability with consideration of shared and non-shared environment effects by collecting diverse environmental variables.[32] Recently, Yang *et al.* has developed a GCTA model, a tool that estimates heritability using GWAS data and unrelated individuals for both quantitative and binary traits.[33,34] The phenotypic variance explained by this model is from all the SNPs (including imputed SNPs) rather than individual SNPs associated with this phenotype. It has been applied to estimate the heritability in intelligence and schizophrenia.[35,36] Heritability is an important concept in genetics but is often misunderstood.[26] Heritability does not influence a trait in itself, but it can play a role in the variation of a trait. Therefore, heritability estimate cannot be used as an indicator of the individual risk. Heritability may vary in different populations and change over time. It is important to select a phenotype in a population with a substantial heritability to identify the genetic determinants underlying the trait.[24] Studies have shown that schizophrenia, bipolar disorder and autism are highly heritable traits with heritability greater than 80%, whereas drug dependence shows moderate heritability of 50-60%.[37] We do not encourage gene identification programs if traits show heritability estimates lower than 30%, as these programs may become a 'geneticist's nightmare.[3,38]

Once twin studies, adoption studies, family studies or population based studies of unrelated individuals have provided evidence that a trait has a genetic component, a segregation analysis with family data will answer the question of what is the best inheritance mode this trait follows.[39] It determines whether the transmission pattern of a trait in families is consistent with the expectation of one of the Mendelian inheritance modes we discussed above. Likelihood ratio test or chi-squared test is usually applied to examine whether a segregation ratio deviates from the expected under Mendelian laws, with no need for genetic marker information. For example, a dominant disease has a theoretical segregation ratio of 0.5. If the hypothesized Mendelian segregation ratio is true, it indicates the disease is determined by a single gene. Otherwise, the deviation may be an indication that the disease is determined by multiple genes, or caused by interplay between genetic and environmental factors, or the disease has an incomplete penetrance. Under these complicated circumstances, maximum likelihood tests are used to compare different inheritance models.[40] Therefore, segregation analysis seems appealing to typical Mendelian modes of inheritance. To a few notable exceptions (e.g. type 1 diabetes) [41] segregation analyses for psychiatric diseases did not succeed in revealing the presence of a major gene and a clear pattern of inheritance.[42]

Single gene disorders versus complex diseases

A single-gene disorder (also called a Mendelian or monogenic disorder) is caused by a single mutation in a single gene. It exhibits a familial pattern consistent with one of the Mendelian inheritance modes. According to the statistics of Mendelian Inheritance in Man (OMIM) (www.ncbi.nlm.nih.gov/omim), more than 5200 diseases follow a Mendelian inheritance pattern, and the underlying molecular basis of 66% of them has been elucidated. Sometimes, mutations in only one gene elucidate 100% of disease cases (e.g. Huntington's

disease). Sometimes, mutations in different genes lead to similar disease presentation. For instance, mutations in 15 different genes lead to the Bardet-Biedl syndrome.[43,44] In that situation, the disease is referred as a heterogeneous monogenic disorder. The identification of genes responsible for single-gene diseases has made tremendous progress in the past 15 years and has greatly facilitated the understanding of disease-related molecular mechanisms. However, Mendelian segregation law which predicts discrete traits (like yellow/green, wrinkled/smooth peas in the original experiments) cannot explain many anthropometric features such as height and weight that show continuous variation. These quantitative traits do display familial clustering (e.g. relatives of the taller individuals tend to be taller than the general population), however, their transmission across generations does not follow clear Mendelian patterns of inheritance. In 1918, Ronald A Fisher, together with Sewall Wright and JBS Haldane, solved the dilemma by developing a polygenic inheritance theory using analysis of variance.[45] Multiple genes contribute to the continuous variation of a trait, each with allelic variation. Meanwhile, each allele follows Mendel's segregation law and makes a small change in the total variance.[45,46] Many common diseases (eg. cancers, diabetes, cardiovascular diseases, Alzheimer's disease and schizophrenia) follow a polygenic model.[47,48] Though the etiology of them is not completely understood, it is believed that they are caused by multiple genes and environmental factors and their interplay. The term complex disease is exchangeable with common disease and polygenic disease in the literature. It is important to pinpoint that monogenic genes exist in polygenic diseases, often initially identified in extreme end of the distribution of a trait. For example, more than sixty loci modestly contribute to the risk of obesity.[49] In addition, rare mutations or deletions at nine loci lead to monogenic forms of early-onset severe obesity and may explain 5-10% of obesity cases. [49,50]

Different models have been proposed to explain the genetic architecture of complex diseases. First, the common disease-common variant hypothesis (CDCV) states that risk variants are at relatively high frequency (>1%) in populations and modestly contribute to the risk of disease.[51,52] The advent of genome-wide association studies (GWAS) has identified more than 2000 common loci modestly associated with complex traits and has given some credit to the CDCV hypothesis. However, the fact that common variants identified through large-scale GWAS consortium initiatives only explain a small proportion of heritability for most complex diseases excludes the possibility that CDCV hypothesis is the only relevant model.[53,54] The second hypothesis, common disease-rare variant (CDRV), states that most of the common phenotypic variance are caused by rare variants (allele frequency <1%) with large effect sizes.[55] Recently, rare variants have been identified to play a role in several multifactorial disorders such as prostate cancer,[56] inflammatory bowel disease [57] or type 2 diabetes.[58] Third, Dickson *et al.* recently proposed the synthetic association model in which the association of a common non-functional SNP with a disease may be the result of several disease-causing rare variants that have stronger effects and are tagged by the common SNPs.[59] Although the synthetic association hypothesis has been validated for specific SNPs associated with hearing loss, sickle cell anemia or Crohn's disease,[59,60] it is unlikely to explain most of the associations between common variants and complex traits identified through GWAS.[60,61] In fact, CDCV and CDRV models are complementary, and there is a growing consensus that multifactorial diseases may result from a combination of rare and common risk variants.[62,63]

Identification of disease predisposing genetic variants: study designs

Different study designs can be used to identify disease-associated genetic variants in different contexts. Case-control and prospective cohort studies commonly used in classical

epidemiology are also applied to genetic epidemiology. A case-control study recruits two groups of individuals who are diagnosed with (cases) or without (controls) a disease and determines the risk of being affected depending on different genotypes. This enables researchers to identify genes responsible for a disease (especially a less-common disease) in a time- and cost-efficient way, because adequate sample size is required to reach sufficient power to detect modest genetic effects. The major weaknesses of a case-control design are biases brought up by the retrospective recalls of exposures and misclassification of cases and controls.[3,64] However, such biases are not a significant concern in a genetic association study because the genotypes (exposures) of individuals does not change with time.[65] However, when confounding factors of some exposures or gene-environment interactions are assessed, considerations to such biases are still relevant. Since genetic associations are sensitive to population stratification between cases and controls, individuals in both groups should come from the same population.[66] In some case-control studies, an enrichment sampling strategy may be applied to increase power to detect a novel genetic variant.[67] Such a strategy increases power but usually overestimates the relative risk. Therefore, it is necessary to replicate in a population-based sample or make a conclusion based on a specific group of people.

In a prospective cohort study, individuals without the disease at baseline are followed for a period of time and then the associations between genotypes and the incident disease status are assessed at the end of the study. Because the disease has not yet presented during sampling, it allows the researchers to control the potential selection bias and minimize the misclassification errors as well. This is why cohort studies are considered the gold standard for both classical and genetic epidemiology studies, but this is with the sacrifice of time and cost. For this reason, case-control studies are more popular in genetic epidemiology. An alternative study design, the nested

case-control study, collects cases in a defined cohort and selects a specific number of controls among those who have not developed the disease yet at the time of assessment.[68] Such an approach shows its unique value in gene-environment interaction association studies because it increases the measurement accuracy of environmental exposures which is essential to increase statistical power to detect interactions.[64,69]

Population-based designs are desirable in genetic epidemiology but they require larger sample sizes than case-control designs to reach the same statistical power, the latter being enriched in a greater proportion of cases.[70] This limitation can become critical if expensive technologies are used (e.g. genome-wide DNA arrays, whole genome sequencing).

Family-based designs are also widely used in genetic epidemiology, which are ideal to assess parental imprinting effects or in haplotype studies (the reconstruction of the haplotype phase is improved by the availability of parental genotypes).[71] A case-parent triad design which consists of one affected offspring and the two parents in each family is commonly used. Given the same power, type I error threshold and risk allele frequency, the number of trios in family-based study is the same as the pairs in a case-control study, signifying 50% more individuals and 50% increased genotyping or sequencing costs are needed. For example, if the power is 90%, using two-sided P-value of 0.001 and an allele frequency of 20% in the control group, 3731 trios will be requested to detect an odds ratio of 1.20 in family-based design and 3731 pairs of case and control in a case-control study, representing 50% more participants. Case-parent triad design is also used to confirm an association from a case-control study because it is robust to population stratification. However, it is not well-adapted to late-onset diseases due to the difficulty or unavailability of DNA collection in parents.[72] There are also other family-based matching designs and corresponding statistical methods.[73] The main limitations are the lack of power,

especially if the effect sizes are small, difficulties in recruiting required number of samples and the generalization of the discoveries from family-based studies to general populations.[74,75]

As mentioned above, the choice of an appropriate control is critical to conduct a valid case-control study. The case-only study is one of the designs which have no controls involved. As well explained by Khoury *et al.*, this design is especially efficient in the context of gene-environment interaction studies when the assumption that the tested genotype and environmental exposure are independent in a given population is met [76,77]. Case-only studies can only examines the departure from a multiplicative interaction model rather than an additive interaction model, which is also less accepted by the scientific community. Although the case-only study design provides better estimation and needs a smaller sample size than traditional case-control design, it also may increase type I error if the assumption is not true.[77,78] In addition to gene-environment interactions, it has also been used in gene-gene interaction and pharmacogenetic studies.[79,80] Pharmacogenetic interaction is a special type of gene-environment interaction and is designed to identify genetic variants which predict response to treatment. When case-only study design is applied, the assumption that there is no correlation between genetic variants and treatment assignment must been examined. Thus, a case-only design nested in a randomized controlled trial (RCT) provides an ideal model for pharmacogenetic studies in which treatment assignment is random and unrelated to genotypes.[80]

How do we get the genetic information?

DNA extraction

Adequate quantity and quality of DNA from a large number of individuals are prerequisites for a successful genetic epidemiology study, both of which depend on the samples

collected and DNA extraction. The samples stored in the Biobank of study centres may be buffy coat (mainly blood leukocytes), saliva (mainly buccal cells) or tissue biopsies. The buffy coat is most commonly used, but saliva is getting more and more popular because of its non-invasive nature and stability at room temperature. Modern DNA extraction methods are fast, non-toxic and reach high yields. A general DNA extraction procedure consists of cell lysis by alkaline, protein removal by salt precipitation and DNA recovery by ethanol precipitation.[81] Extracted DNA is dissolved in appropriate buffer and stored in small aliquots at -70°C for long-term storage, but repeated freezing and thawing should be avoided.

Genotyping

Single nucleotide polymorphisms (SNPs) represent more than 90% of the entire genomic variants. SNPs have been initially detected by direct sequencing and genotyping of 270 individuals in the context of the Human Genome Project and HapMap Project and more recently through the 1000 Genomes Project. There are over 38 million validated human SNPs in the dbSNP database (dbSNP Build 137) (https://www.ncbi.nlm.nih.gov/SNP/). In the past two decades, many genotyping principles have been developed, with most of them assuming a bi-allelic feature of most SNPs in human. The commonly used approaches include restriction fragment length polymorphism (RFLP), differential hybridization (TaqMan), allele-specific primer extension (SNaPshot, SNPstream, pyrosequencing), allele-specific oligonucleotide ligation (Applied Biosystems SNPlex), allele-specific extension (Illumina Omni Whole-Genome Arrays) and single-base extension (Affymetrix 6.0) which can be detected by mass spectrometry (Sequenom MassArray), fluorescent light (TaqMan, Applied Biosystems), bioluminescent light, electrophoresis or high-resolution melting curves (Roche Applied Sciences LightTyper).[82,83] Generally speaking, all these methods are performed in two different formats: homogeneous

reactions (in solution) and heterogeneous reactions (in solution and a solid phase such as a microtiter well plate, latex beads, a glass slide, or a silicon chip). The former has limited capability of multiplexing which is to examine more than one SNP at a time; while the latter one is flexible in multiplexing ranging from a few to a hundred to several millions SNPs. Because of their intrinsic characteristics, each genotyping method has unique application and multiplexing capability. For examples, TaqMan SNP Genotyping Assays (Applied Biosystems) identify the genotypes of single SNP at a time with great precision and is widely used in candidate-gene association and replication studies even with large sample size.[84] The Sequenom MassArray uses a single-base primer extension genotyping method followed by distinguishing DNA base by molecular weight. It has high resolution but moderate multiplexing, and it is appropriate for small number of SNPs.[85] The more recent genome-wide genotyping arrays can accommodate up to 4.8 million genetic markers, including single nucleotide polymorphisms (SNPs) and probes for the detection of copy number variations (CNV). Therefore, some platforms work better for single SNPs or a few targeted SNPs in many individuals, some are suitable for small to moderate number (hundreds to thousands) of SNPs on a few subjects at one time, and others are the best choice for several millions of SNPs on one subject at one time, depending on the aim and design of a particular study. Customized design may also be applied to genotyping on a single SNP or moderate number of SNPs. More than 4.5 million predesigned probes are available to customized uses with TaqMan genotyping.[83] Table 1 gives a simple guideline on how to choose an appropriate genotyping platform, and the updated capacity of each platform is always available on the commercial websites.

Sequencing

Sequencing is a method to determine the exact sequence of nucleotides from a fragment of DNA or the whole genome. It not only examines the presence of the bi-allelic variants reported in databases, but also provides information on all possible polymorphisms (including those with 3 or 4 alleles). Sequencing is the ideal method to characterize the sequence of a new genome or to identify rare genetic variants not reported in SNP databases. Due to its current cost, sequencing has not yet been an efficient and economical way to genotype SNPs. Sanger sequencing (the first generation sequencing method), which was described in 1977,[86] experienced many technical revolutions and eventually developed into today's automated Sanger sequencing.[87,88] The completion of the Human Genome Project led to tremendous improvements in the Sanger sequencing method, including the development of whole-genome shotgun sequencing and a parallel sequencing initiative of the human genome by the company Celera Genomics.[89] However, Sanger sequencing is still expensive and laborious, and faster and more affordable methods to sequence DNA were in great demand from broad research interests such as variant association studies, comparative genomics, population evolution and clinical diagnostics. High-throughput next generation sequencing (NGS), first launched in 2005, involves "massively parallel" sequencing and offers to sequence up to hundreds of millions of DNA fragments in a single platform. It cost $2.7 billion and 12 years to complete the Human Genome Project with Sanger sequencing, but it is now possible to obtain a personal whole-genome sequence at a cost of $1,000.[90]

Currently the DNA polymerase-dependent sequencing strategies are widespread on the market and can be classified as single nucleotide addition (SNA), cyclic reversible termination (CRT) and real-time sequencing. Here we will introduce three major platforms which are commercially available, in combination with their unique sequencing principles (Table 4).

Roche/454 was first developed NGS, using "pyrosequencing" technique of DNA.[91,92] The current Roche/454 GS FLX+ Sequencer is able to produce 700 Mb of sequence with 99.997% accuracy for single reads of 1,000 bases in length (http://454.com/products/gs-flx-system/index.asp).

The second NGS approach is the Illumina/Solexa Genome Analyzer which currently dominates the market. The capacity of the newest model generates up to 600 Gb of bases per run with a read length of about 100 bases (http://www.illumina.com/technology/solexatechnology.ilmn). This is less than Roche/454 due to less efficient incorporation of modified nucleotides.

Another NGS system is Applied Biosystems Supported Oligonucleotide Ligation and Detection (SOLiD) sequencer based on sequencing by ligation [93]. The complicated process is well illustrated in Metzker's paper.[92] SOLiD systems have two independent flow cells and allow two completely different experiments to be run at the same time. The updated SOLiD system can yield 320 Gb of sequence per run with a 99.99% accuracy and a read length of 50-75 bases (http://www.invitrogen.com/site/us/en/home/Products-and-Services/Applications/Sequencing-/NextGenerationSequencing/).

Recently, the novel sequencing technology ION Torrent arose on the market. It does not need any modified nucleotides. Its chemistry rationale is very simple. During the process of DNA synthesis, the incorporation of each dNTP causes the release of a hydrogen ion. The hydrogen ion changes pH in the solution, which can be detected by an ion-sensitive field-effect transistor (ISFET) detector.[94] This method enables a fast, accurate, inexpensive, and simple massively parallel sequencing. Ion Personal Genome Machine (PGM) and Ion Proton sequencers load amplified DNA fragments into micro wells of a high-density Ion chip to perform sequencing. The changed pH can be detected by an ion sensitive layer beneath the wells and

converted into voltage changes. The change in voltage is proportional to the type and number of nucleotides incorporated and recorded. These smaller and cheaper sequencers can produce up to 2 Gb output per run with a read length of 200-400 bases.

In addition to the strategies discussed above, many other technologies are under development and all the methods will continue to compete and improve.[88] Currently, it is not easy to predict which approach will be the winner of the future sequencing market. NGS is certainly another ground-breaking revolution in biology and medicine after the completion of the Human Genome Project, making personal whole-genome studies more than just a dream. The 1000 Genomes Project has used Illumina/Solexa and Roche/454 platforms to sequence whole genomes and has validated up to 38 million SNPs, 1.4 million short insertions and deletions, and more than 14,000 larger deletions.[95] Whole-genome sequencing plays a unique role in facilitating a deeper and broader understanding of the spectrum of genetic variants and their pathogenesis in complex diseases, clinical diagnosis and personalized health decision-making. It will eventually come into daily practice in the near future; however, current cost and analytical challenges limit its applicability.[90,96] An alternative solution to this may be to apply NGS to target specific sequences of interest, for example, whole-exome sequencing which sequences the entire protein-coding genes. In spite of constituting approximately 1% of the human genome, protein-coding regions include 85% of mutations associated with Mendelian diseases.[97] Meanwhile, non-synonymous variants predict with a high likelihood a functional change.[98] As such, the whole-exome is a relevant subset of the genome to search for genetic variants with large effect sizes and has been used to dissect the genetic architectures of Mendelian and complex disorders.[99,100] Exome sequencing by NGS, in conjunction with developed strategies in study design and analytic methods, has had a great success in identifying causal alleles for several dozen

Mendelian disorders [99]. Although it is more challenging, whole-exome sequencing has also been an effective strategy in identifying coding variants associated with complex diseases such as autism spectrum disorders and schizophrenia.[101-103] Compared to whole-genome sequencing, whole-exome sequencing is currently a more widely accepted strategy to search for rare variants because of its cost-effectiveness, the simpler data analysis and interpretation.

Gene identification strategies

The identification of genes responsible for Mendelian and complex diseases may enable a better understanding of their pathology, provide efficient molecular targets for innovative therapeutic drugs, and help to better predict disease risk in populations for targeted prevention. In the past decade, a remarkable progress has been made in the journey of discovering disease-causing genes. However, more than 30% of the underlying genes leading to Mendelian disorders are still unknown, and the identified genetic variants to complex diseases account for only a small portion of heritability. In order to pursue gene identification efforts, traditional and novel gene identification strategies are introduced below.

Genetic linkage studies

Linkage analysis aims to map the location of a disease-causing loci by looking for genetic markers that co-segregate with the disease within pedigrees, though the disease causing allele has not to be directly genotyped.[75] Linkage is based on the facts that recombination occurs between homologous chromosomes during meiosis and recombination likelihood increases with the distance between two loci, a random probability from zero to 0.5. When a marker allele is inherited along with the disease in pedigrees, it strongly suggests that the disease-causing locus is located in the vicinity of the genetic marker on the chromosome. A set of 400 highly-

informative microsatellite markers (repeated sequences of DNA fragments less than 10 bp [104]) equally distributed across the genome is generally selected in a whole-genome linkage analysis. More recently, a set of 6,000-10,000 markers have been proposed by different companies to perform linkage analysis.

Different linkage approaches are chosen depending on the type of disease (monogenic or polygenic) or trait (dichotomous or quantitative). Parametric or model-based linkage analysis is used if the disease follows one of the typical Mendelian inheritance modes. Results of linkage analysis are often reported as logarithm of the odds (LOD) score which is a function of the parameter θ. θ is the probability of a recombination event (recombination fraction) between a genetic marker and the disease locus.[75] LOD score analysis is equivalent to likelihood ratio test, assessing the null hypothesis H_0 of $\theta=0.5$ (absence of linkage) versus alternative hypothesis H_1 of $\theta<0.5$ (presence of linkage). In the simplest scenario with a known inheritance model, complete penetrance, no *de novo* mutations and no phenocopies (different environmental exposures and genetic variants lead to the same disease), θ is estimated by the maximum likelihood method, thus giving rise to a maximum LOD score (Table 2). The higher the LOD score is, the stronger the evidence of linkage will be. Historically, a rule of thumb states that a LOD score above 3 is sufficient to claim a significant linkage, based on the critical value from Morton.[104] An even higher LOD score of 3.3 is required to ensure the genome-wide type I error of 0.05. Other complicated model-based cases with incomplete penetrance, phenocopies and mutations, and more relaxed LOD score thresholds are discussed in detail by Ziegler and Konig. [30] Linkage analysis has successfully mapped genes responsible for Mendelian disorders such as the Wolfram syndrome on the short arm of chromosome 4.[105,106]

Little is known about loci predisposing to complex diseases, and attributing a clear Mendelian pattern of inheritance within families for such a locus is impossible. As a result, model-based linkage analyses do not apply to complex trait linkage analyses and model-free linkage analyses have been developed. The fundamental rationale underlying model-free linkage analysis is that the genetic resemblance in the affected sibling pairs is more similar in certain regions of the genome if the disease is heritable. Therefore, the statistical tests assess whether the observed degree of genotypic similarity exceeds the expected value. Instead of measuring recombinant fraction of θ, genotypic similarity is measured by the identical by descent (IBD) value which refers to the number of alleles inherited from the same common ancestor in a pair of relatives. The IBD values can be 0, 1, or 2. If the distribution of IBD values is determined, model-free linkage analysis examines whether allele sharing in affected siblings is different from the expected distribution. More generally, it tests whether the mean number of IBD shared alleles departs from the expected value of 1 in sibling pairs.[107] Excess of IBD sharing can also be tested by other methods such as the maximum non-parametric LOD score test and Wald test [30] which successfully identified the HLA region associated with type I diabetes.[108]

Linkage studies also apply to quantitative traits such as cholesterol or glucose level. The approaches for model-free linkage analysis of quantitative traits include the Haseman-Elston, variance component methods among others.[30] A region between markers D9S925 and D9S741 on chromosome 9p associated with high-density lipoprotein-cholesterol concentration in Mexican Americans was initially identified with variance component analysis.[109] However, true linkage has been hard to find in complex trait studies, likely due to the modest effect sizes of genetic variants, allelic heterogeneity, or gene by environment interactions in complex diseases. [25,110]

Homozygosity mapping

Homozygosity mapping is a powerful tool to map genes responsible for recessive Mendelian disorders in consanguineous pedigrees.[111] With this approach less than a dozen of affected individuals are needed and more importantly no additional family members are required to identity a disease-causing locus. These advantages render it possible to map disease loci of many rare recessive disorders when it is impossible to collect adequate number of families as linkage analysis usually requires. The principle underlying this approach is that if the offspring of a consanguineous marriage (for example sibling, first-cousin, and second-cousin) is affected with a recessive inherited disease, a large region spanning the disease locus is homozygous by decent [111]. For instance, a child of a consanguineous couple has a coefficient of inbreeding F of 1/4, 1/16, 1/64 for sibling, first-cousin, and second cousin, respectively. Assuming the frequency of the disease allele in this population is q, the probability of homozygosity by decent at the disease locus is $\alpha = F*q/[F*q+(1-F)*q^2]$. If q is far smaller than F, α is close to 1, indicating the greatest chance to be homozygous. The comparison of homozygous regions in several affected family members, along with traditional linkage analysis and a sufficiently dense genetic map, can narrow down the location of a gene underlying a recessive disease. Low-density restriction fragment length polymorphism (RFLP), microsatellite linkage maps, and more recently high-density SNP arrays have been used in homozygosity mapping gene identification. For instance, the use of a high-density GeneChip containing 57,244 SNPs identified the linked region for autosomal recessive Bardet-Biedl syndrome which was initially missed by linkage studies with 400 highly informative microsatellites in a small Israeli Bedouin consanguineous pedigree.[112]

Candidate gene studies

This approach is hypothesis-driven and has been widely used in genetic association studies before the advent of GWAS. Candidate genes are selected based on prior knowledge of their potential role on the trait of interest from *in vivo*, *in vitro* or *in silico* studies in animals or humans.[113,114] One important advantage of the candidate gene approach is to restrict the number of hypotheses tested and to relax the multiple testing correction thresholds in comparison with genome-wide approaches. One limitation of the candidate gene approach is its dependence on the level of current knowledge of a specific gene. The success rate of candidate gene studies has been low, in part due to the limited understanding of the molecular and genetic mechanisms in complex diseases.[66] Selecting strong candidate genes on the basis of converging arguments from different research disciplines has been more successful, as illustrated by the identification of SNPs in *APOE4* associated with Alzheimer disease (AD).[115]. *APOE4* gene was indeed located on the proximal long arm of chromosome 19, in a region of linkage for late-onset AD.[116] In addition, apolipoprotein E (ApoE) was a key protein related to AD.[115]

Genome-wide association studies

Hypothesis-free GWAS exhaustively test the genotype-phenotype associations across up to 4.8 million genetic markers and represent to date the most efficient way to identify common variants (MAF> 1%) associated with complex diseases.[117] Along with the advanced high-throughput technology, more and more SNPs and copy number variants (CNVs) are validated by the 1000 Genomes Project, which enable the current genotyping arrays to include rare variants and CNVs in addition to common variants., GWAS have identified several risk variants associated with bipolar disorder [118] or schizophrenia.[119] However, there are two major limitations

of GWAS. First, a very stringent level of significance is required to adjust for multiple testing. Second, most of the statistically significant associations lack a biological support.[120-122]

Whole-genome/whole-exome sequencing

Whole-genome/whole-exome sequencing strategies are currently efficiently applied to identify rare variants associated with Mendelian or complex traits. Whole-genome/whole-exome sequencing is not just an alternative way for genotyping as it also detects novel mutations not catalogued in SNP databases and additional alleles beyond bi-alleles. The biggest challenge in whole-genome/whole-exome sequencing experiments is how to analyze a huge sequencing dataset to identify the novel causal genes for either Mendelian or complex diseases.[123] Usually, 20,000 to 30,000 variants are found through each whole-exome run. Unreliable variants are first removed by data quality control procedures (e.g. read coverage less than five, inconsistency among the reads). If the investigators focus their attention on potentially deleterious rare coding variants, variants located outside the coding regions and synonymous coding variants are filtered out. Then the most important step with substantial reduction of the number of variants is to exclude known polymorphisms in human population based on appropriate databases.[124] At this step, approximately 150-500 non-synonymous or splicing variants remain to be potentially causal variants. Additional filtering methods may include *in silico* functional evaluation of mutations, candidate gene, linkage, homozygosity mapping, *de novo* and overlap strategies.[123] Becker *et al*. have successfully used homozygosity mapping, in combination with an exome sequencing strategy, to elucidate the genetic basis of osteogenesis imperfecta.[125] They found 318 non-synonymous variants after several filtering strategies. Among them, 17 were autosomal homozygous, but only three were in the regions with the larger stretch of homozygous loci. In combination with overlap strategy and functional testing, truncating mutations in gene

SERPINF1 were identified as causal loci leading to autosomal-recessive osteogenesis imperfecta.[125]

How to interpret genetic associations in complex disease?

Power of a study

In genetic epidemiology, most genetic variants confer small to modest effect sizes with an odds ratio (OR) lower than 1.5, indicating that a large sample size is needed in a population-based association study. For example, if the risk allele frequency in controls is 20%, 1763 cases and 1763 controls are needed to detect an OR of 1.3 at a type I error level of 0.001 (two-sided) and power of 90%.[3] The requirement for such large sample sizes can be difficult to achieve by single teams and as a result researchers have to pool samples in large-scale international consortium initiatives to reach an adequate power. These power estimations also imply that many previously published case-control studies were underpowered. This may explain why many promising associations were never replicated.[126] Replication of an association study in an independent sample is recommended. The sample size for the replication study should take into account of the risk of overestimation of the true effect in the initial sample (a phenomenon called the Winner's curse effect).[127,128]. Statistical power may be even more a concern in genetic association studies involving rare variants, and the desired number of individuals may not be feasible in practice.[129] To deal with these issues, researchers select designs where additional copies of the variant of interest can be sampled (perhaps in large pedigrees or in a founder population). They also pool together variants likely to have an impact on the function of a specific gene and compare the global distribution of these variants in case control designs.[130]

Data quality control (QC)

Genotyping errors cause genotype misclassification and have the potential risk of decreased power, leading to false associations.[131] The procedures to remove the uncertain individuals and DNA markers are critical steps before statistical analysis of associations. It is recommended to conduct QC on the individuals before QC on the DNA markers.[132] Individuals with discordant sex information, inaccurate phenotypic data, or a conflicting ethnicity between self-reported and genetically determined should be identified and removed. Individuals with low DNA quality (e.g. displaying >10% missing genotypes in a genotyping array) should also be taken out. At the genetic marker level, the genotyping method should be reliable and the laboratory protocols should be standard. The concordance rate of duplicated samples must be higher than 99% (usually > 10% of the entire sample are re-genotyped with the same or a different genotyping method). SNPs with a genotyping call rate (percentage of successfully genotyped individuals) <95%, a significant deviation from a Hardy-Weinberg equilibrium (HWE) test [133] ($P_{HWE} < 0.005$ in the control group), a significant difference in the missing genotype rates between cases and controls, or a very low allele frequency should be filtered out. In a family-based study, an additional check of Mendelian inconsistencies should be conducted .[30]

According to the workflow of NGS, standard protocols for QC should be developed and implemented at each step including DNA extraction, targeted gene enrichment, library preparation and sequencing. Current NGS technologies have higher raw per-base error rates than Sanger sequencing.[134] However, this shortcoming can be compensated to some extent by increasing the coverage depth of sequencing, checking the presence of a mutation in related individuals or validating the findings by Sanger sequencing.[135,136] False-positive association may also result from a difference of coverage depth between cases and control groups.[137]

Statistical analysis

A genetic model (i.e., dominant, additive, recessive) needs to be defined prior to any genotype-phenotype association study. If the underlying genetic model is unknown, an additive model is frequently assumed, but testing the three models is more informative. Given two alleles A and B (B is risk allele) and three genotypes AA, AB and BB at a locus, AA is coded as 0, AB as 1 and BB as 2, and a 2×3 contingency table is created under an additive model as illustrated in the table 3. In the simplest scenario in which cases and controls are matched for confounding factors (e.g. age, sex), the Cochran-Armitage test is used to test the association between the allele B and a trait, which is similar to Peason's $\chi 2$ test but taking into account the order of risk of the three genotypes (AA<AB<BB) [138]. Meanwhile, the ORs are often calculated to provide a measure of the strength of the associations. If individuals have one risk allele B, the risk of having the disease is OR1=(b/a)/(e/d)=bd/ae times higher than those who has no risk allele B; and if individuals have two copies of B, the risk of being affected is OR2=(c/a)/(f/d)=cd/af times higher than those who has no B. If the outcome is binary (presence or absence of the disease), a simple logistic regression can also be applied. The exponential of the regression coefficient equals to the increased OR with per additional B. If the outcome is a continuous (or quantitative) variable, a linear regression model will be used. The beta coefficient from a linear regression analysis means how much increase in the outcome for each additional risk allele B. Compared to Cochran-Armitage and Peason's $\chi 2$ tests, the advantage of using a linear or logistic regression is that they allow for the adjustment for the confounding factors such as age, sex and including of gene × gene and gene × environment interaction terms into the model.[139] When the outcome is a count/rate or a time-to-events, a Poisson regression model or a Cox proportional hazard model will be chosen, respectively. As a result, relative risk (RR) or hazard ratio (HR) will be

estimated.[140] Sophisticated methods such as the kernel association test have been recently developed to assess the association of groups of rare variants with a disease or a quantitative trait.[141,142]

From the perspective of statistics, GWAS analysis is just an extension of the single-SNP analysis and covariates can also be adjusted in linear or logistic regression models. One issue is that most of the significant associations at the nominal level ($P < 0.05$) are likely to be spurious in the context of the many tests performed in GWAS [143]. There is no universal standard to obtain a critical value for adjustment; nevertheless, the Bonferroni correction, Bayesian procedures and false-discovery rate (FDR) are widely used to define an appropriate threshold of significance level accounting for multiple testing. The Bonferroni correction considers a simple setting in which the type I error α level is 0.05 and n independent SNPs are tested, the adjusted significance level α' should meet $\alpha=1-(1-\alpha')^n$ and then $\alpha' \approx \alpha/n$. If 1 million SNPs are independently tested whether they are associated with a trait in a GWAS context, the Bonferroni-adjusted threshold will be $0.05 / 1{,}000{,}000 = 5 \times 10^{-8}$, which is a genome-wide significance level frequently reported in the GWAS literature.[144] The Bonferroni correction is overly conservative because many SNPs being tested are in linkage disequilibrium and tightly correlated each other. The Bayesian approach is based on the prior probability of true positive association from previous evidence.[145] As a result, the P-values are far less stringent and the thresholds are different from study to study and from researcher to researcher. The FDR method measures the false rate of the rejected null hypotheses (detected associations) rather than focusing on the presence of at least one error, resulting in an increase in power.[146,147] A FDR of 0.05 is usually adapted and indicates that 5% of the detected associations are random results. However, in

GWAS, because the majority of the null hypotheses are true, FDR does not provide a substantial advantage in comparison with the Bonferroni correction.

Multiple testing presents new challenges in whole-exome and whole-genome sequencing experiments due to the massive amount of genetic data generated by these methods. Because there are many rare variants which are expected to have larger effect sizes and more severe functional impacts, it is not practical to use the same threshold across all the variants. Several recommendations are proposed and different analytic packages are in implementation.[130,148] Some authors suggest gene-based or pathway-based tests,[130] while others recommend different thresholds would be generated according to cut-offs derived from different allele frequencies. Probably, a permutation-based approach is more accurate to handle multiple testing by naturally taking into account allele frequency and correlated alleles.[100]

Most genetic association studies focus on the main effects of variants contributing to the development of a disease. However, predisposing SNPs identified to date only explain a small portion of the heritability of many complex diseases. Gene by gene (G×G) and gene by environment (G×E) interactions are critical components of the architecture of complex traits and have been proposed to explain at least a fraction of the "missing heritability". [47,54] May a variant missed by a classical GWAS have an increased effect in presence of another genetic variant or in a specific environment? This hypothesis can be tested by incorporating interaction terms into a SNP-based linear or logistic regression model [65]. When a systematic search for G×G epistatic interactions is undertaken, the power dramatically decreases due to the numerous combinations of any two SNP tests. If two SNPs interactions are systematically investigated in a first generation GWAS (e.g. 300,000 SNPs), 100 billion epistasis tests will be performed, resulting in an exceptionally stringent Bonferroni-corrected significance threshold of 5×10^{-13}.[149] As a result,

the few epistasis studies using GWAS data published up to date failed to identify G×G interactions significant after multiple testing correction.[150,151] Compared to an epistatic study, a G×E interaction study is more feasible in the context of GWAS, although an empirical rule states that the samples needed are four times larger than those needed for studying the main effect.[152] Currently, three classical methods are used to identify G×E interactions.[153] The first tests G×E interactions using biologic candidate genes and/or GWAS validated loci. This is currently the more commonly used approach in literature. The second approach is the hypothesis-free Genome-Environment Wide Interaction Study (GEWIS), which systematically tests G×E interactions across the genome. Multiple testing decreases the statistical power in GEWIS. The third method of variance prioritization (VP) prioritizes SNPs on the basis of heterogeneity in the variance of a quantitative trait among three genotypes of a bi-allelic SNP.[154] It selects a subset of SNPs for G×E interaction tests, thus increasing the chance to detect potential associations missed by GEWIS.

All the commonly used statistical software (such as SAS, SPSS or STAT *etc*) can be used to analyze genetic data. PLINK [155] is a free and very efficient tool to deal with genetic quality control and data analysis, especially for GWAS data. R software is more and more used in genetic epidemiology as many packages with specific genetic functions are programmed and it is free online.

Meta-analysis

An individual linkage or association study is rarely conclusive in genetic epidemiology; therefore replication studies are always required. Following the same rules as in traditional clinical epidemiology, meta-analysis is also applied to genetic epidemiology. Meta-analysis combines relevant but independent studies and increases the power of the analysis and the

precision of the effect size by increasing sample size, thus providing more precise evidence of association.[156] Usually, more weight is assigned in the meta-analysis to studies displaying a larger sample size or a greater event rate. Both the sample size and event rate can be reflected in the variance estimate. Therefore, a usual way to assign a weight to individual studies in a meta-analysis is to use inverse variance, even though alternative methods exist (e.g. Mantel-Haenszel test). The estimation of the degree of between-study heterogeneity is important in the interpretation of meta-analyses.[157] Between-study heterogeneity is measured by I^2 which is a modified Cochran's Q statistic.[158] Because this test has a low power, a p value of less than 0.1 is considered as significant heterogeneity. Usually, I^2 values of 25%, 50% and 75% represent low, moderate and high levels of between-study heterogeneity, respectively. If heterogeneity exists, subgroup or sensitivity analysis may further be performed to assess the causes of such heterogeneity (e.g. study ascertainment). New global fixed-effect (FE) and random-effects (RE) meta-analytic methods have been recently proposed to deal with heterogeneity between studies.[159] The recent emergence of international consortiums and the conduct of large-scale meta-analyses of genetic association studies have revolutionized the field and have led to an important yield of novel disease-predisposing loci. For instance, a recent meta-analysis of the 5, 10-methylenetetrahydrofolate reductase (*MTHFR*) gene variant C677T in 29,502 subjects has confirmed its associations with schizophrenia, bipolar disorder and unipolar depressive disorder and suggests a shared genetic susceptibility among distinct psychiatric disorders.[160] Numbers matter but do not always lead to success. Recently, the psychiatric GWAS consortium conducted a mega-analysis for major depressive disorder in $18,759$ subjects followed by a replication in 57,478 samples. They did not find genome-wide significant association signal and concluded that

the sample was still underpowered to identify common variants associated with major depression.[161]

Conclusions

Genetic epidemiology is a relatively recent but fascinating research field in which expertise from different disciplines converge to elucidate genetic factors responsible for Mendelian and complex diseases. We comprehensively reviewed the key concepts and methods in genetic epidemiology including single gene disorders and complex diseases, study design implementation, genotyping and sequencing strategies, gene identification strategies, data analysis and data interpretation. We hope this review will help non-geneticist clinicians critically appraise a genetic association study and understand what makes a good genetic association study. With the decrease in sequencing costs, personalized genome sequencing will eventually become an instrument of common medical practice. In the next paper, we will review the past, current and coming applications of genetic knowledge in medical practice, and we will appreciate how far we are from the personalized medicine revolution.

Conflict of interest

The authors confirm that this article content has no conflict of interest.

Acknowledgements

We thank Jackie Hudson, Alexandra Mayhew and Hudson Reddon for editing of the manuscript, and the reviewers for their helpful comments. David Meyre is supported by a Tier 2 Canada

Research Chair. Aihua Li is supported by a Queen Elizabeth II Graduate Scholarship in Science and Technology.

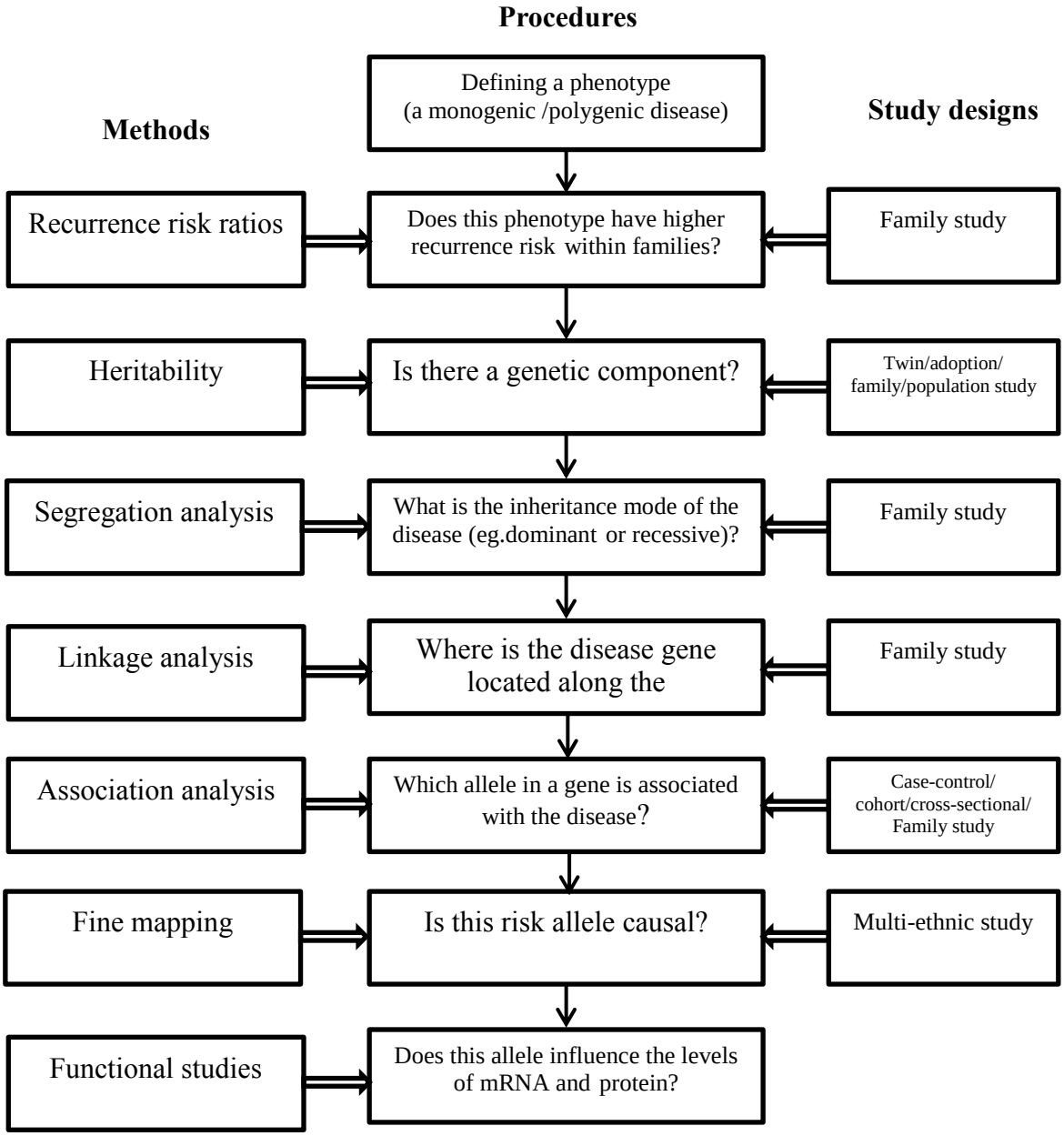


Figure 1. Framework outlining the procedures, methods and study designs to identify the genetic determinants of common diseases

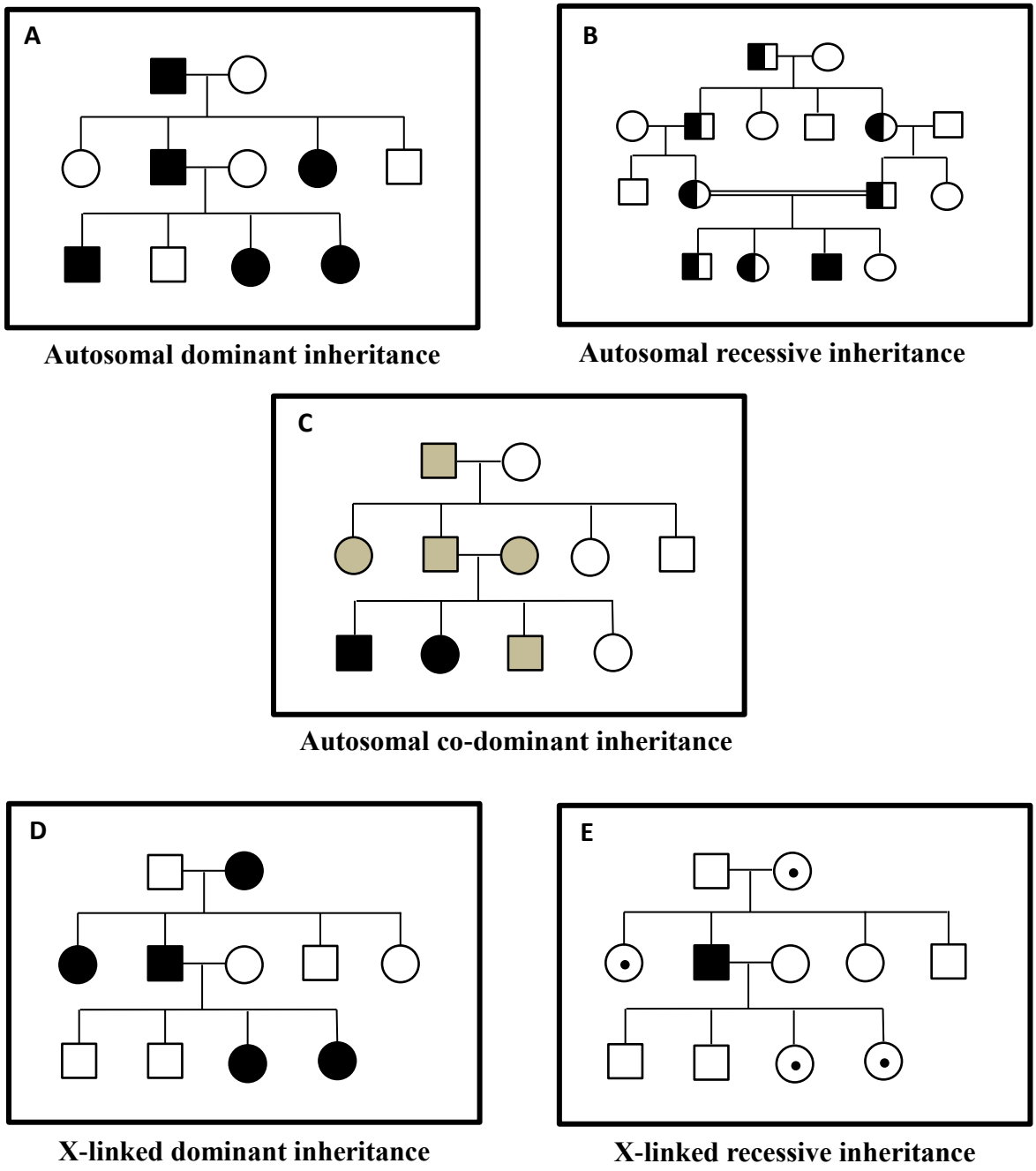


Figure 2. Modes of inheritance. Pedigrees with autosomal dominant inheritance (A), autosomal recessive inheritance (B), autosomal co-dominant inheritance (C), X-linked dominant inheritance (D), X-linked recessive inheritance (E).

= unaffected male/ female

= male/ female carrier

= affected male/ female

= affected male/ female with less severe symptoms

= X-linked female carrier

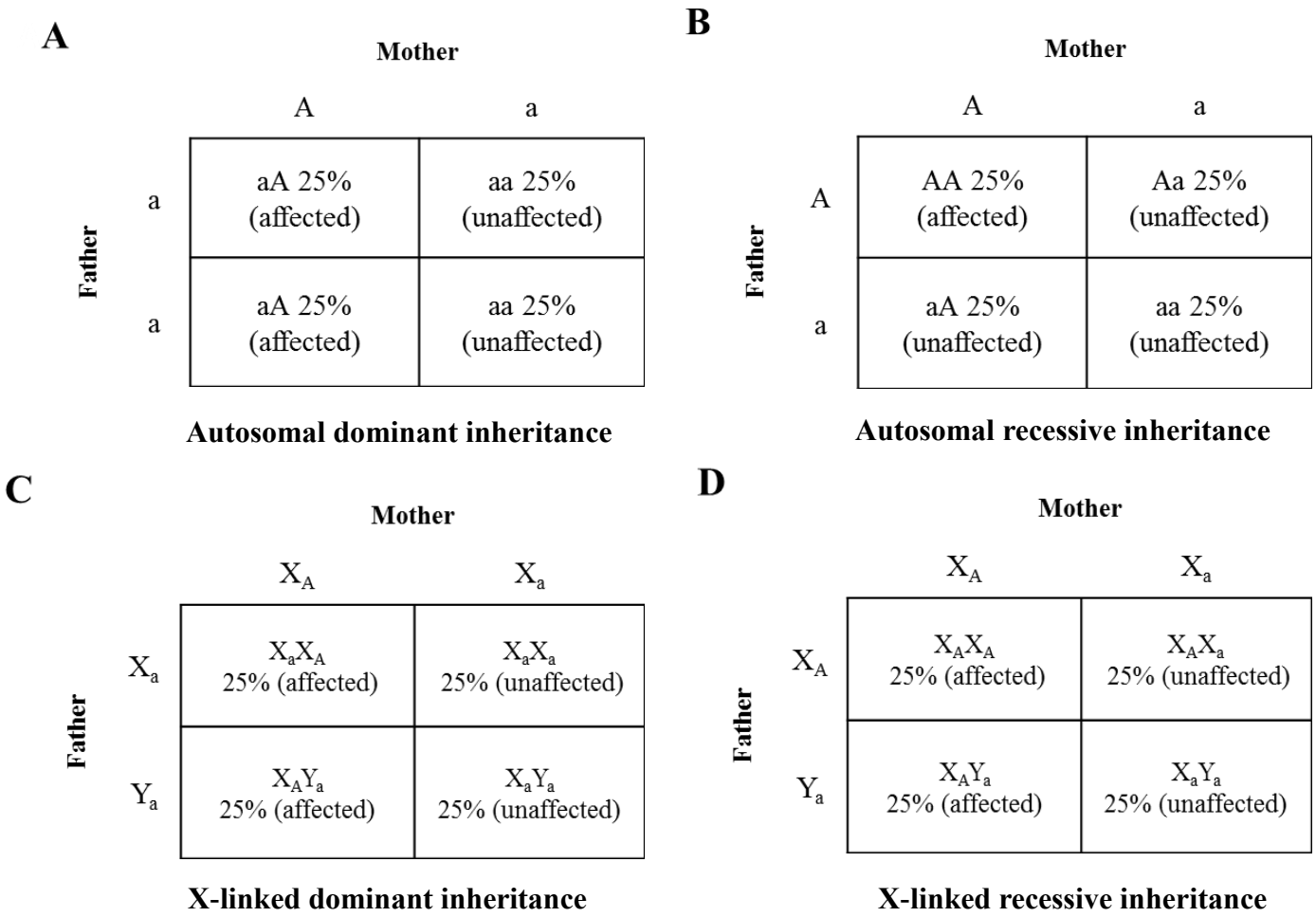


Figure 3. Punnett squares of inherited traits. Punnett squares are used to predict the chance of genetic disease in children for parents with an increased risk. The disease-causing mutation is denoted by A and the normal gene is denoted by a**.** A) Autosomal

dominant inheritance: A mother with an autosomal dominant mutation has children with a father who is normal. They have 50% chance with each pregnancy of having a child (boy or girl) affected by the disease and a 50% chance having a child (boy or girl) unaffected. B) Autosomal recessive inheritance: A mother with an autosomal mutation has children with a father who also has the same autosomal mutation. They have 25% chance with each pregnancy of having a child (boy or girl) affected, a 50% of chance having a child unaffected but with the same mutation (carriers), and 25% chance having a child unaffected with normal genotypes. C) X-linked dominant inheritance: A mother with an X-linked mutation has children with a father who is normal. They have 25% chance with each pregnancy of having a girl affected by the disease and a 25% chance having a boy affected. The rest of the children are unaffected with normal genotypes. D) X-linked recessive inheritance: A mother with an X-linked mutation has children with a father who also has a copy of X-linked mutation. They have 25% chance with each pregnancy of having a girl affected by the disease and a 25% chance having a boy affected. The other half of the girls are unaffected but are the mutant carriers and the other half of the boys are unaffected with normal genotypes.

Table 1. Genotyping methods and study designs

Number of SNPs to be genotyped	Study designs	Genotyping methods
1-10	Candidate gene studies Replication studies	TaqMan LightTyper Pyrosequencing
1-500	Replication studies Linkage studies Fine-mapping studies	SNaPshot SNPlex Sequenom MassARRAY Illumina Golden Gate with BeadXpress readout
384-3,072	Linkage studies Fine-mapping studies Disease-specific SNPs Pathway-specific SNPs	Illumina Golden Gate with iScan readout
6,000-70,000	Linkage studies Fine-mapping studies Disease-specific SNPs Pathway-specific SNPs	Illumina Infinium iSelec Custom Beadchip
>500,000 Up to 4.8 million	GWAS (SNPs, CNVs)	Illumina Omni Whole-Genome Array Affymetrix 6.0 Array

Table 2. Measurements of familial aggregation, heritability and linkage analysis

	Measurements	**Formula**	**Thresholds**
Familial Aggregation	recurrence risk ratio in relatives λ_R [23]	λ_R =prevalence of the disease in the relatives of the affected individual / prevalence of the disease in the general population [24]	2 [24]
Heritability	the proportion of total phenotypic variability explained by genetic variance in a particular population [26]	Twin study: $h^2=2(rMZ-rDZ)$ Adoption study : $h^2=2rPO$ [27] Population-based: (narrow-sense) h^2= variance of additive genetic effects/total variance of the observed phenotype [26]	There is no consensus on the minimum threshold of heritability needed to follow-up with gene identification program. A heritability estimate of 30% maybe considered as the minimum.[3]
Linkage study	LOD: logarithm of the odds score [75]	$LOD(\theta)=\log_{10}[Likelihood(\hat{\theta})/Likelihood(\theta=0.5)]$ [75]	3.3 [75]

rMZ: correlation coefficient of the trait between monozygotic twins

rDZ: correlation coefficient of the trait between dizygotic twins

rPO: correlation coefficient of the trait between a biological parent and an adopted-away child

θ is the probability of a recombination event (recombination fraction) between a genetic marker and the disease locus. Observed $\hat{\theta}$ can be obtained by counting recombinants and non-recombinants when the genotypes of individuals within a family are available.

Table 3. A 2×3 contingency table in an additive model

	AA	AB	BB
Case	a	b	c
Control	d	e	f

a, b, c are the counts of individuals with genotypes of AA, AB, BB respectively in cases, and d, e, f are the counts of individuals with genotypes of AA, AB, BB respectively in controls.

Table 4. Characteristics of sequencing platforms

Platform	Sequencing technology	Sequencing reaction	Capacity	Efficiency (bp/read)
Roche/454	Single nucleotide addition (pyrosequencing)	Synthesis	700 MB	1,000
Illumina/Solexa Genome Analyzer	Cyclic reversible termination	Synthesis	600Gb	100
Applied Biosystems/ (SOLiD)	Real-time sequencing	Ligation	320Gb	50-75
Applied Biosystems/ION Torrent	**Semiconductor**	Synthesis	2Gb	200-400

References

1 Morton, N. E. The future of genetic epidemiology. *Ann. Med.* **24**, 557-562 (1992).

2 Erk, S. *et al.* Functional impact of a recently identified quantitative trait locus for hippocampal volume with genome-wide support. *Transl Psychiatry* **3**, e287, doi:10.1038/tp.2013.57 (2013).

3 Li, A. & Meyre, D. Challenges in reproducibility of genetic association studies: lessons learned from the obesity field. *Int J Obes (Lond)* **37**, 559-567, doi:10.1038/ijo.2012.82 (2013).

4 A novel gene containing a trinucleotide repeat that is expanded and unstable on Huntington's disease chromosomes. The Huntington's Disease Collaborative Research Group. *Cell* **72**, 971-983 (1993).

5 Sudbery, P. & Sudbery, I. *Human molecular genetics*. Third edition edn, p3 (Pearson Education Limited, 2009).

6 Scriver, C. R. The PAH gene, phenylketonuria, and a paradigm shift. *Hum. Mutat.* **28**, 831-845, doi:10.1002/humu.20526 (2007).

7 Ropers, H. H. & Hamel, B. C. X-linked mental retardation. *Nature reviews. Genetics* **6**, 46-57, doi:10.1038/nrg1501 (2005).

8 Strachan, T. & Read, A. *Human molecular genetics* 4th edition edn, 76 (Garland Science, 2011).

9 Raj, A. & van Oudenaarden, A. Nature, nurture, or chance: stochastic gene expression and its consequences. *Cell* **135**, 216-226, doi:10.1016/j.cell.2008.09.050 (2008).

10 Ben-Shachar, S. *et al.* Microdeletion 15q13.3: a locus with incomplete penetrance for autism, mental retardation, and psychiatric disorders. *Journal of medical genetics* **46**, 382-388, doi:10.1136/jmg.2008.064378 (2009).

11 van Bon, B. W. *et al.* Further delineation of the 15q13 microdeletion and duplication syndromes: a clinical spectrum varying from non-pathogenic to a severe outcome. *Journal of medical genetics* **46**, 511-523, doi:10.1136/jmg.2008.063412 (2009).

12 Reik, W. & Walter, J. Genomic imprinting: parental influence on the genome. *Nat Rev Genet* **2**, 21-32, doi:10.1038/35047554 (2001).

13 Miozzo, M. & Simoni, G. The role of imprinted genes in fetal growth. *Biol. Neonate* **81**, 217-228, doi:56752 (2002).

14 Frost, J. M. & Moore, G. E. The importance of imprinting in the human placenta. *PLoS genetics* **6**, e1001015, doi:10.1371/journal.pgen.1001015 (2010).

15 Nicholls, R. D. The impact of genomic imprinting for neurobehavioral and developmental disorders. *The Journal of clinical investigation* **105**, 413-418, doi:10.1172/JCI9460 (2000).

16 Ruderfer, D. M. *et al.* Mosaic copy number variation in schizophrenia. *Eur. J. Hum. Genet.*, doi:10.1038/ejhg.2012.287 (2013).

17 McMahon, F. J., Stine, O. C., Meyers, D. A., Simpson, S. G. & DePaulo, J. R. Patterns of maternal transmission in bipolar affective disorder. *American journal of human genetics* **56**, 1277-1286 (1995).

18 Goldstein, J. M., Faraone, S. V., Chen, W. J. & Tsuang, M. T. Gender and the familial risk for schizophrenia. Disentangling confounding factors. *Schizophrenia research* **7**, 135-140 (1992).

19 Rollins, B. *et al.* Mitochondrial variants in schizophrenia, bipolar disorder, and major depressive disorder. *PloS one* **4**, e4913, doi:10.1371/journal.pone.0004913 (2009).
20 Wermter, A. K. *et al.* Preferential reciprocal transfer of paternal/maternal DLK1 alleles to obese children: first evidence of polar overdominance in humans. *Eur. J. Hum. Genet.* **16**, 1126-1134, doi:10.1038/ejhg.2008.64 (2008).
21 Burghes, A. H., Vaessin, H. E. & de La Chapelle, A. Genetics. The land between Mendelian and multifactorial inheritance. *Science* **293**, 2213-2214, doi:10.1126/science.1065930 (2001).
22 Savage, D. B. *et al.* Digenic inheritance of severe insulin resistance in a human pedigree. *Nature genetics* **31**, 379-384, doi:10.1038/ng926 (2002).
23 Risch, N. Linkage strategies for genetically complex traits. I. Multilocus models. *American journal of human genetics* **46**, 222-228 (1990).
24 Burton, P. R., Tobin, M. D. & Hopper, J. L. Key concepts in genetic epidemiology. *Lancet* **366**, 941-951, doi:10.1016/S0140-6736(05)67322-9 (2005).
25 Altmuller, J., Palmer, L. J., Fischer, G., Scherb, H. & Wjst, M. Genomewide scans of complex human diseases: true linkage is hard to find. *American journal of human genetics* **69**, 936-950, doi:10.1086/324069 (2001).
26 Visscher, P. M., Hill, W. G. & Wray, N. R. Heritability in the genomics era--concepts and misconceptions. *Nature reviews. Genetics* **9**, 255-266, doi:10.1038/nrg2322 (2008).
27 Urbanoski, K. A. & Kelly, J. F. Understanding genetic risk for substance use and addiction: a guide for non-geneticists. *Clinical psychology review* **32**, 60-70, doi:10.1016/j.cpr.2011.11.002 (2012).
28 Kamin, L. J. & Goldberger, A. S. Twin studies in behavioral research: a skeptical view. *Theor. Popul. Biol.* **61**, 83-95, doi:10.1006/tpbi.2001.1555 (2002).
29 Falconer, D. S. The inheritance of liability to diseases with variable age of onset, with particular reference to diabetes mellitus. *Ann. Hum. Genet.* **31**, 1-20 (1967).
30 Ziegler, A. K., I.R. *A statistical approach to genetic epidemiology*. 2nd edn, (Wiley-VCH Verlag GmbH & Co., 2010).
31 Boomsma, D., Busjahn, A. & Peltonen, L. Classical twin studies and beyond. *Nat Rev Genet* **3**, 872-882, doi:10.1038/nrg932 (2002).
32 Carlsson, S., Ahlbom, A., Lichtenstein, P. & Andersson, T. Shared genetic influence of BMI, physical activity and type 2 diabetes: a twin study. *Diabetologia* **56**, 1031-1035, doi:10.1007/s00125-013-2859-3 (2013).
33 Yang, J., Lee, S. H., Goddard, M. E. & Visscher, P. M. GCTA: a tool for genome-wide complex trait analysis. *Am J Hum Genet* **88**, 76-82, doi:10.1016/j.ajhg.2010.11.011 (2011).
34 Visscher, P. M., Yang, J. & Goddard, M. E. A commentary on 'common SNPs explain a large proportion of the heritability for human height' by Yang et al. (2010). *Twin Res Hum Genet* **13**, 517-524, doi:10.1375/twin.13.6.517 (2010).
35 Davies, G. *et al.* Genome-wide association studies establish that human intelligence is highly heritable and polygenic. *Mol. Psychiatry* **16**, 996-1005, doi:10.1038/mp.2011.85 (2011).
36 Purcell, S. M. *et al.* Common polygenic variation contributes to risk of schizophrenia and bipolar disorder. *Nature* **460**, 748-752, doi:10.1038/nature08185 (2009).
37 Dick, D. M., Riley, B. & Kendler, K. S. Nature and nurture in neuropsychiatric genetics: where do we stand? *Dialogues in clinical neuroscience* **12**, 7-23 (2010).

38 Neel, J. V. Diabetes mellitus: a "thrifty" genotype rendered detrimental by "progress"? *American journal of human genetics* **14**, 353-362 (1962).

39 Elston, R. C. Segregation analysis. *Advances in human genetics* **11**, 63-120, 372-123 (1981).

40 Moll, P. P., Burns, T. L. & Lauer, R. M. The genetic and environmental sources of body mass index variability: the Muscatine Ponderosity Family Study. *Am J Hum Genet* **49**, 1243-1255 (1991).

41 Barrai, I. & Cann, H. M. Segregation Analysis of Juvenile Diabetes Mellitus. *J. Med. Genet.* **2**, 8-11 (1965).

42 Martinez, M. [Genetic markers and risk factors in diseases with complex etiology: psychiatric diseases]. *Rev. Epidemiol. Sante Publique* **41**, 306-314 (1993).

43 Leitch, C. C. *et al.* Hypomorphic mutations in syndromic encephalocele genes are associated with Bardet-Biedl syndrome. *Nat. Genet.* **40**, 443-448, doi:10.1038/ng.97 (2008).

44 Stoetzel, C. *et al.* Identification of a novel BBS gene (BBS12) highlights the major role of a vertebrate-specific branch of chaperonin-related proteins in Bardet-Biedl syndrome. *Am J Hum Genet* **80**, 1-11, doi:10.1086/510256 (2007).

45 Risch, N. J. Searching for genetic determinants in the new millennium. *Nature* **405**, 847-856, doi:10.1038/35015718 (2000).

46 Farrall, M. Quantitative genetic variation: a post-modern view. *Hum. Mol. Genet.* **13 Spec No 1**, R1-7, doi:10.1093/hmg/ddh084 (2004).

47 Eichler, E. E. *et al.* Missing heritability and strategies for finding the underlying causes of complex disease. *Nat Rev Genet* **11**, 446-450, doi:10.1038/nrg2809 (2010).

48 Schork, N. J., Greenwood, T. A. & Braff, D. L. Statistical genetics concepts and approaches in schizophrenia and related neuropsychiatric research. *Schizophr. Bull.* **33**, 95-104, doi:10.1093/schbul/sbl045 (2007).

49 Choquet, H. & Meyre, D. Molecular basis of obesity: current status and future prospects. *Curr Genomics* **12**, 154-168, doi:10.2174/138920211795677921 (2011).

50 Choquet, H. & Meyre, D. Genetics of Obesity: What have we Learned? *Current genomics* **12**, 169-179, doi:10.2174/138920211795677895 (2011).

51 Lander, E. S. The new genomics: global views of biology. *Science* **274**, 536-539 (1996).

52 Reich, D. E. & Lander, E. S. On the allelic spectrum of human disease. *Trends in genetics : TIG* **17**, 502-510 (2001).

53 Maher, B. Personal genomes: The case of the missing heritability. *Nature* **456**, 18-21, doi:10.1038/456018a (2008).

54 Manolio, T. A. *et al.* Finding the missing heritability of complex diseases. *Nature* **461**, 747-753, doi:10.1038/nature08494 (2009).

55 Cirulli, E. T. & Goldstein, D. B. Uncovering the roles of rare variants in common disease through whole-genome sequencing. *Nat Rev Genet* **11**, 415-425, doi:10.1038/nrg2779 (2010).

56 Gudmundsson, J. *et al.* A study based on whole-genome sequencing yields a rare variant at 8q24 associated with prostate cancer. *Nat. Genet.* **44**, 1326-1329, doi:10.1038/ng.2437 (2012).

57 Rivas, M. A. *et al.* Deep resequencing of GWAS loci identifies independent rare variants associated with inflammatory bowel disease. *Nat. Genet.* **43**, 1066-1073, doi:10.1038/ng.952 (2011).

58 Bonnefond, A. *et al.* Rare MTNR1B variants impairing melatonin receptor 1B function contribute to type 2 diabetes. *Nat. Genet.* **44**, 297-301, doi:10.1038/ng.1053 (2012).

59 Dickson, S. P., Wang, K., Krantz, I., Hakonarson, H. & Goldstein, D. B. Rare variants create synthetic genome-wide associations. *PLoS biology* **8**, e1000294, doi:10.1371/journal.pbio.1000294 (2010).

60 Anderson, C. A., Soranzo, N., Zeggini, E. & Barrett, J. C. Synthetic associations are unlikely to account for many common disease genome-wide association signals. *PLoS Biol* **9**, e1000580, doi:10.1371/journal.pbio.1000580 (2011).

61 Hunt, K. A. *et al.* Negligible impact of rare autoimmune-locus coding-region variants on missing heritability. *Nature* **498**, 232-235, doi:10.1038/nature12170 (2013).

62 Goldstein, D. B. & Chikhi, L. Human migrations and population structure: what we know and why it matters. *Annu Rev Genomics Hum Genet* **3**, 129-152, doi:10.1146/annurev.genom.3.022502.103200 (2002).

63 Gibson, G. Rare and common variants: twenty arguments. *Nat Rev Genet* **13**, 135-145, doi:10.1038/nrg3118 (2011).

64 Manolio, T. A., Bailey-Wilson, J. E. & Collins, F. S. Genes, environment and the value of prospective cohort studies. *Nat Rev Genet* **7**, 812-820, doi:10.1038/nrg1919 (2006).

65 Suarez, E., Sariol, C. A., Burguete, A. & McLachlan, G. A tutorial in genetic epidemiology and some considerations in statistical modeling. *Puerto Rico health sciences journal* **26**, 401-421 (2007).

66 Hirschhorn, J. N., Lohmueller, K., Byrne, E. & Hirschhorn, K. A comprehensive review of genetic association studies. *Genetics in medicine : official journal of the American College of Medical Genetics* **4**, 45-61 (2002).

67 Meyre, D. *et al.* Genome-wide association study for early-onset and morbid adult obesity identifies three new risk loci in European populations. *Nature genetics* **41**, 157-159, doi:10.1038/ng.301 (2009).

68 Thomas, D. Gene--environment-wide association studies: emerging approaches. *Nat Rev Genet* **11**, 259-272, doi:10.1038/nrg2764 (2010).

69 Moffitt, T. E., Caspi, A. & Rutter, M. Strategy for investigating interactions between measured genes and measured environments. *Arch. Gen. Psychiatry* **62**, 473-481, doi:10.1001/archpsyc.62.5.473 (2005).

70 Yang, J., Wray, N. R. & Visscher, P. M. Comparing apples and oranges: equating the power of case-control and quantitative trait association studies. *Genetic epidemiology* **34**, 254-257, doi:10.1002/gepi.20456 (2010).

71 Cordell, H. J., Barratt, B. J. & Clayton, D. G. Case/pseudocontrol analysis in genetic association studies: A unified framework for detection of genotype and haplotype associations, gene-gene and gene-environment interactions, and parent-of-origin effects. *Genetic epidemiology* **26**, 167-185, doi:10.1002/gepi.10307 (2004).

72 Cardon, L. R. & Palmer, L. J. Population stratification and spurious allelic association. *Lancet* **361**, 598-604, doi:10.1016/S0140-6736(03)12520-2 (2003).

73 Cordell, H. J. & Clayton, D. G. Genetic association studies. *Lancet* **366**, 1121-1131, doi:10.1016/S0140-6736(05)67424-7 (2005).

74 Risch, N. & Merikangas, K. The future of genetic studies of complex human diseases. *Science* **273**, 1516-1517 (1996).

75 Dawn Teare, M. & Barrett, J. H. Genetic linkage studies. *Lancet* **366**, 1036-1044, doi:10.1016/S0140-6736(05)67382-5 (2005).

76 Khoury, M. J. & Flanders, W. D. Nontraditional epidemiologic approaches in the analysis of gene-environment interaction: case-control studies with no controls! *Am. J. Epidemiol.* **144**, 207-213 (1996).

77 Piegorsch, W. W., Weinberg, C. R. & Taylor, J. A. Non-hierarchical logistic models and case-only designs for assessing susceptibility in population-based case-control studies. *Stat. Med.* **13**, 153-162 (1994).

78 Kazma, R., Dizier, M. H., Guilloud-Bataille, M., Bonaiti-Pellie, C. & Genin, E. Power comparison of different methods to detect genetic effects and gene-environment interactions. *BMC Proc* **1 Suppl 1**, S74 (2007).

79 Yang, Q., Khoury, M. J., Sun, F. & Flanders, W. D. Case-only design to measure gene-gene interaction. *Epidemiology* **10**, 167-170 (1999).

80 Pierce, B. L. & Ahsan, H. Case-only genome-wide interaction study of disease risk, prognosis and treatment. *Genet. Epidemiol.* **34**, 7-15, doi:10.1002/gepi.20427 (2010).

81 Visvikis, S., Schlenck, A. & Maurice, M. DNA extraction and stability for epidemiological studies. *Clin. Chem. Lab. Med.* **36**, 551-555, doi:10.1515/CCLM.1998.094 (1998).

82 Kwok, P. Y. Methods for genotyping single nucleotide polymorphisms. *Annu Rev Genomics Hum Genet* **2**, 235-258, doi:10.1146/annurev.genom.2.1.235 (2001).

83 Edenberg, H. J. & Liu, Y. Laboratory methods for high-throughput genotyping. *Cold Spring Harb Protoc* **2009**, pdb top62, doi:10.1101/pdb.top62 (2009).

84 Song, Y. *et al.* FTO polymorphisms are associated with obesity but not diabetes risk in postmenopausal women. *Obesity (Silver Spring)* **16**, 2472-2480, doi:10.1038/oby.2008.408 (2008).

85 Jurinke, C., van den Boom, D., Cantor, C. R. & Koster, H. Automated genotyping using the DNA MassArray technology. *Methods Mol. Biol.* **187**, 179-192, doi:10.1385/1-59259-273-2:179 (2002).

86 Sanger, F., Nicklen, S. & Coulson, A. R. DNA sequencing with chain-terminating inhibitors. *Proc. Natl. Acad. Sci. U. S. A.* **74**, 5463-5467 (1977).

87 Hutchison, C. A., 3rd. DNA sequencing: bench to bedside and beyond. *Nucleic Acids Res* **35**, 6227-6237, doi:10.1093/nar/gkm688 (2007).

88 Metzker, M. L. Emerging technologies in DNA sequencing. *Genome research* **15**, 1767-1776, doi:10.1101/gr.3770505 (2005).

89 Venter, J. C. Genome-sequencing anniversary. The human genome at 10: successes and challenges. *Science* **331**, 546-547, doi:10.1126/science.1202812 (2011).

90 Mardis, E. R. The $1,000 genome, the $100,000 analysis? *Genome medicine* **2**, 84, doi:10.1186/gm205 (2010).

91 Ronaghi, M., Karamohamed, S., Pettersson, B., Uhlen, M. & Nyren, P. Real-time DNA sequencing using detection of pyrophosphate release. *Anal. Biochem.* **242**, 84-89, doi:10.1006/abio.1996.0432 (1996).

92 Metzker, M. L. Sequencing technologies - the next generation. *Nature reviews. Genetics* **11**, 31-46, doi:10.1038/nrg2626 (2010).

93 Shendure, J. *et al.* Accurate multiplex polony sequencing of an evolved bacterial genome. *Science* **309**, 1728-1732, doi:10.1126/science.1117389 (2005).

94 Rothberg, J. M. *et al.* An integrated semiconductor device enabling non-optical genome sequencing. *Nature* **475**, 348-352, doi:10.1038/nature10242 (2011).

95 Abecasis, G. R. *et al.* An integrated map of genetic variation from 1,092 human genomes. *Nature* **491**, 56-65, doi:10.1038/nature11632 (2012).
96 Gonzaga-Jauregui, C., Lupski, J. R. & Gibbs, R. A. Human genome sequencing in health and disease. *Annual review of medicine* **63**, 35-61, doi:10.1146/annurev-med-051010-162644 (2012).
97 Botstein, D. & Risch, N. Discovering genotypes underlying human phenotypes: past successes for mendelian disease, future approaches for complex disease. *Nat. Genet.* **33 Suppl**, 228-237, doi:10.1038/ng1090 (2003).
98 Kryukov, G. V., Pennacchio, L. A. & Sunyaev, S. R. Most rare missense alleles are deleterious in humans: implications for complex disease and association studies. *Am J Hum Genet* **80**, 727-739, doi:10.1086/513473 (2007).
99 Bamshad, M. J. *et al.* Exome sequencing as a tool for Mendelian disease gene discovery. *Nature reviews. Genetics* **12**, 745-755, doi:10.1038/nrg3031 (2011).
100 Kiezun, A. *et al.* Exome sequencing and the genetic basis of complex traits. *Nat. Genet.* **44**, 623-630, doi:10.1038/ng.2303 (2012).
101 O'Roak, B. J. *et al.* Sporadic autism exomes reveal a highly interconnected protein network of de novo mutations. *Nature*, doi:10.1038/nature10989 (2012).
102 Sanders, S. J. *et al.* De novo mutations revealed by whole-exome sequencing are strongly associated with autism. *Nature*, doi:10.1038/nature10945 (2012).
103 Girard, S. L. *et al.* Increased exonic de novo mutation rate in individuals with schizophrenia. *Nature genetics* **43**, 860-863, doi:10.1038/ng.886 (2011).
104 Morton, N. E. Sequential tests for the detection of linkage. *Am J Hum Genet* **7**, 277-318 (1955).
105 Polymeropoulos, M. H., Swift, R. G. & Swift, M. Linkage of the gene for Wolfram syndrome to markers on the short arm of chromosome 4. *Nat. Genet.* **8**, 95-97, doi:10.1038/ng0994-95 (1994).
106 Inoue, H. *et al.* A gene encoding a transmembrane protein is mutated in patients with diabetes mellitus and optic atrophy (Wolfram syndrome). *Nat. Genet.* **20**, 143-148, doi:10.1038/2441 (1998).
107 Blackwelder, W. C. & Elston, R. C. A comparison of sib-pair linkage tests for disease susceptibility loci. *Genetic epidemiology* **2**, 85-97, doi:10.1002/gepi.1370020109 (1985).
108 Davies, J. L. *et al.* A genome-wide search for human type 1 diabetes susceptibility genes. *Nature* **371**, 130-136, doi:10.1038/371130a0 (1994).
109 Arya, R. *et al.* Linkage of high-density lipoprotein-cholesterol concentrations to a locus on chromosome 9p in Mexican Americans. *Nat. Genet.* **30**, 102-105, doi:10.1038/ng810 (2002).
110 Saunders, C. L. *et al.* Meta-analysis of genome-wide linkage studies in BMI and obesity. *Obesity (Silver Spring)* **15**, 2263-2275, doi:10.1038/oby.2007.269 (2007).
111 Lander, E. S. & Botstein, D. Homozygosity mapping: a way to map human recessive traits with the DNA of inbred children. *Science* **236**, 1567-1570 (1987).
112 Chiang, A. P. *et al.* Homozygosity mapping with SNP arrays identifies TRIM32, an E3 ubiquitin ligase, as a Bardet-Biedl syndrome gene (BBS11). *Proc. Natl. Acad. Sci. U. S. A.* **103**, 6287-6292, doi:10.1073/pnas.0600158103 (2006).
113 Zhu, M. & Zhao, S. Candidate gene identification approach: progress and challenges. *International journal of biological sciences* **3**, 420-427 (2007).

114 Tranchevent, L. C. *et al.* A guide to web tools to prioritize candidate genes. *Briefings in bioinformatics* **12**, 22-32, doi:10.1093/bib/bbq007 (2011).

115 Saunders, A. M. *et al.* Association of apolipoprotein E allele epsilon 4 with late-onset familial and sporadic Alzheimer's disease. *Neurology* **43**, 1467-1472 (1993).

116 Pericak-Vance, M. A. *et al.* Linkage studies in familial Alzheimer disease: evidence for chromosome 19 linkage. *American journal of human genetics* **48**, 1034-1050 (1991).

117 Visscher, P. M., Brown, M. A., McCarthy, M. I. & Yang, J. Five years of GWAS discovery. *American journal of human genetics* **90**, 7-24, doi:10.1016/j.ajhg.2011.11.029 (2012).

118 Craddock, N. & Sklar, P. Genetics of bipolar disorder. *Lancet* **381**, 1654-1662, doi:10.1016/S0140-6736(13)60855-7 (2013).

119 Ripke, S. *et al.* Genome-wide association study identifies five new schizophrenia loci. *Nature genetics* **43**, 969-976, doi:10.1038/ng.940 (2011).

120 Feingold, S. B., Smith, J., Houtz, J., Popovsky, E. & Brown, R. S. Prevalence and functional significance of thyrotropin receptor blocking antibodies in children and adolescents with chronic lymphocytic thyroiditis. *The Journal of clinical endocrinology and metabolism* **94**, 4742-4748, doi:10.1210/jc.2009-1243 (2009).

121 Pearson, T. A. & Manolio, T. A. How to interpret a genome-wide association study. *JAMA* **299**, 1335-1344, doi:10.1001/jama.299.11.1335 (2008).

122 Manolio, T. A. Genomewide association studies and assessment of the risk of disease. *N. Engl. J. Med.* **363**, 166-176, doi:10.1056/NEJMra0905980 (2010).

123 Gilissen, C., Hoischen, A., Brunner, H. G. & Veltman, J. A. Disease gene identification strategies for exome sequencing. *European journal of human genetics : EJHG* **20**, 490-497, doi:10.1038/ejhg.2011.258 (2012).

124 Abecasis, G. R. *et al.* A map of human genome variation from population-scale sequencing. *Nature* **467**, 1061-1073, doi:10.1038/nature09534 (2010).

125 Becker, J. *et al.* Exome sequencing identifies truncating mutations in human SERPINF1 in autosomal-recessive osteogenesis imperfecta. *Am J Hum Genet* **88**, 362-371, doi:10.1016/j.ajhg.2011.01.015 (2011).

126 Ioannidis, J. P. Why most published research findings are false. *PLoS Med* **2**, e124, doi:10.1371/journal.pmed.0020124 (2005).

127 Zollner, S. & Pritchard, J. K. Overcoming the winner's curse: estimating penetrance parameters from case-control data. *American journal of human genetics* **80**, 605-615, doi:10.1086/512821 (2007).

128 Zhong, H. & Prentice, R. L. Correcting "winner's curse" in odds ratios from genomewide association findings for major complex human diseases. *Genetic epidemiology* **34**, 78-91, doi:10.1002/gepi.20437 (2010).

129 Fawcett, K. A. *et al.* Detailed investigation of the role of common and low-frequency WFS1 variants in type 2 diabetes risk. *Diabetes* **59**, 741-746, doi:10.2337/db09-0920 (2010).

130 Do, R., Kathiresan, S. & Abecasis, G. R. Exome sequencing and complex disease: practical aspects of rare variant association studies. *Hum. Mol. Genet.* **21**, R1-9, doi:10.1093/hmg/dds387 (2012).

131 Pompanon, F., Bonin, A., Bellemain, E. & Taberlet, P. Genotyping errors: causes, consequences and solutions. *Nat Rev Genet* **6**, 847-859, doi:10.1038/nrg1707 (2005).

132 Anderson, C. A. *et al.* Data quality control in genetic case-control association studies. *Nat Protoc* **5**, 1564-1573, doi:10.1038/nprot.2010.116 (2010).

133 Wittke-Thompson, J. K., Pluzhnikov, A. & Cox, N. J. Rational inferences about departures from Hardy-Weinberg equilibrium. *Am J Hum Genet* **76**, 967-986, doi:10.1086/430507 (2005).

134 Chan, E. Y. Next-generation sequencing methods: impact of sequencing accuracy on SNP discovery. *Methods Mol. Biol.* **578**, 95-111, doi:10.1007/978-1-60327-411-1_5 (2009).

135 Stitziel, N. O., Kiezun, A. & Sunyaev, S. Computational and statistical approaches to analyzing variants identified by exome sequencing. *Genome Biol* **12**, 227, doi:10.1186/gb-2011-12-9-227 (2011).

136 Ku, C. S. *et al.* Exome sequencing: dual role as a discovery and diagnostic tool. *Annals of neurology* **71**, 5-14, doi:10.1002/ana.22647 (2012).

137 Garner, C. Confounded by sequencing depth in association studies of rare alleles. *Genet. Epidemiol.*, doi:10.1002/gepi.20574 (2011).

138 Slager, S. L. & Schaid, D. J. Case-control studies of genetic markers: power and sample size approximations for Armitage's test for trend. *Human heredity* **52**, 149-153 (2001).

139 Saito, Y. A., Talley, N. J., de Andrade, M. & Petersen, G. M. Case-control genetic association studies in gastrointestinal disease: review and recommendations. *The American journal of gastroenterology* **101**, 1379-1389, doi:10.1111/j.1572-0241.2006.00587.x (2006).

140 Burton, P. R., Scurrah, K. J., Tobin, M. D. & Palmer, L. J. Covariance components models for longitudinal family data. *International journal of epidemiology* **34**, 1063-1077; discussion 1077-1069, doi:10.1093/ije/dyi069 (2005).

141 Li, B. & Leal, S. M. Methods for detecting associations with rare variants for common diseases: application to analysis of sequence data. *Am J Hum Genet* **83**, 311-321, doi:10.1016/j.ajhg.2008.06.024 (2008).

142 Wu, M. C. *et al.* Rare-variant association testing for sequencing data with the sequence kernel association test. *Am J Hum Genet* **89**, 82-93, doi:10.1016/j.ajhg.2011.05.029 (2011).

143 Rice, T. K., Schork, N. J. & Rao, D. C. Methods for handling multiple testing. *Adv. Genet.* **60**, 293-308, doi:10.1016/S0065-2660(07)00412-9 (2008).

144 Dudbridge, F. & Gusnanto, A. Estimation of significance thresholds for genomewide association scans. *Genet. Epidemiol.* **32**, 227-234, doi:10.1002/gepi.20297 (2008).

145 Colhoun, H. M., McKeigue, P. M. & Davey Smith, G. Problems of reporting genetic associations with complex outcomes. *Lancet* **361**, 865-872 (2003).

146 Bretz, F., Landgrebe, J. & Brunner, E. Multiplicity issues in microarray experiments. *Methods Inf. Med.* **44**, 431-437, doi:10.1267/METH05030431 (2005).

147 Storey, J. D. & Tibshirani, R. Statistical significance for genomewide studies. *Proceedings of the National Academy of Sciences of the United States of America* **100**, 9440-9445, doi:10.1073/pnas.1530509100 (2003).

148 Goldstein, D. B. *et al.* Sequencing studies in human genetics: design and interpretation. *Nat Rev Genet* **14**, 460-470, doi:10.1038/nrg3455 (2013).

149 Balding, D. J. A tutorial on statistical methods for population association studies. *Nature reviews. Genetics* **7**, 781-791, doi:10.1038/nrg1916 (2006).

150 Tao, S. *et al.* Genome-wide two-locus epistasis scans in prostate cancer using two European populations. *Hum. Genet.* **131**, 1225-1234, doi:10.1007/s00439-012-1148-4 (2012).
151 Greliche, N. *et al.* A genome-wide search for common SNP x SNP interactions on the risk of venous thrombosis. *BMC Med Genet* **14**, 36, doi:10.1186/1471-2350-14-36 (2013).
152 Hunter, D. J. Gene-environment interactions in human diseases. *Nature reviews. Genetics* **6**, 287-298, doi:10.1038/nrg1578 (2005).
153 Franks, P. W. Gene x environment interactions in type 2 diabetes. *Curr Diab Rep* **11**, 552-561, doi:10.1007/s11892-011-0224-9 (2011).
154 Pare, G., Cook, N. R., Ridker, P. M. & Chasman, D. I. On the use of variance per genotype as a tool to identify quantitative trait interaction effects: a report from the Women's Genome Health Study. *PLoS Genet* **6**, e1000981, doi:10.1371/journal.pgen.1000981 (2010).
155 Purcell, S. *et al.* PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am J Hum Genet* **81**, 559-575, doi:10.1086/519795 (2007).
156 Normand, S. L. Meta-analysis: formulating, evaluating, combining, and reporting. *Stat. Med.* **18**, 321-359 (1999).
157 Minelli, C., Thompson, J. R., Abrams, K. R., Thakkinstian, A. & Attia, J. The quality of meta-analyses of genetic association studies: a review with recommendations. *American journal of epidemiology* **170**, 1333-1343, doi:10.1093/aje/kwp350 (2009).
158 Higgins, J. P., Thompson, S. G., Deeks, J. J. & Altman, D. G. Measuring inconsistency in meta-analyses. *BMJ* **327**, 557-560, doi:10.1136/bmj.327.7414.557 (2003).
159 Neupane, B., Loeb, M., Anand, S. S. & Beyene, J. Meta-analysis of genetic association studies under heterogeneity. *Eur. J. Hum. Genet.* **20**, 1174-1181, doi:10.1038/ejhg.2012.75 (2012).
160 Peerbooms, O. L. *et al.* Meta-analysis of MTHFR gene variants in schizophrenia, bipolar disorder and unipolar depressive disorder: evidence for a common genetic vulnerability? *Brain, behavior, and immunity* **25**, 1530-1543, doi:10.1016/j.bbi.2010.12.006 (2011).
161 Ripke, S. *et al.* A mega-analysis of genome-wide association studies for major depressive disorder. *Mol. Psychiatry* **18**, 497-511, doi:10.1038/mp.2012.21 (2013).

SUPPLEMENTARY CHAPTER VIII: JUMP ON THE TRAIN OF PERSONALIZED MEDICINE: A PRIMER FOR NON-GENETICIST CLINICIANS

PART3. CLINICAL APPLICATIONS IN THE PERSONALIZED MEDICINE AREA

Aihua Li, David Meyre

Abstract

The rapid decline of sequencing costs brings hope that personal genome sequencing will become a common feature of medical practice. This series of three reviews aim to help non-geneticist clinicians to jump into the fast-moving field of personalized genetic medicine. In the first two articles, we covered the fundamental concepts of molecular genetics and the methodologies used in genetic epidemiology. In this third article, we discuss the evolution of personalized medicine and illustrate the most recent success in the fields of Mendelian and complex human diseases. We also address the challenges that currently limit the use of personalized medicine to its full potential.

Introduction

The observation of a familial clustering for human diseases was first reported by the Greek physician Hippocrates at the time of the 5th century BC.[1] He believed that hereditary material in all parts of the body affected health of next generation.[1] In 1865, Gregor Mendel published his seminal work on the laws of Mendelian inheritance from his experiments in peas. [2] In 1902, Archibald Garrod postulated that inborn errors of metabolism in humans might follow Mendel's laws and described how alkaptonuria, a rare human disorder, followed a pattern of recessive inheritance. This was the first report linking Mendel's laws and a human disease.[3] Garrod can be considered as the founder of human genetics, a field that has long been considered by most physicians as an esoteric academic specialty.[4] Times have changed with the development of clinical genetics and more recently with the emergence of the concept of personalized medicine.

Personalized medicine, also known as genomic medicine or precision medicine, originated with the idea of using an individual's unique genetic make-up to assess the risk of developing disease, predict the course and prognosis of disease, and tailor therapeutic interventions accordingly.[4,5] It was this blueprint that inspired the United States (US) National Research Council in 1990s to initiate the Human Genome Project.[6,7] Completion of the Human Genome Project, the HapMap project and more recently the 1000 Genomes Project has resulted in an explosion of genetic discoveries related to human disorders.[8-10] Since then, there has been marked improvement in high-throughput technologies for both genotyping and sequencing; which along with advances in computational biotechnology, has fostered great promise in the potential of personalized medicine to revolutionize how we understand, diagnose, prevent and treat diseases.

Genetic screening is an important tool to use advances in genetics and genomics to improve public health.[11] However, in the first half of the 20th century, many scientifically unsound and socially harmful policies and laws based on "perceived genetic risks", had been adapted and implemented in many countries in the name of eugenics. Eugenics was coined by Sir Francis Galton in 1883 and he claimed that "a highly gifted race of men" could be generated by the process of selective breeding.[12] Among the most famous proponents of the eugenic idea, the United States was the first country to take some actions. On one side, the US advocated "positive eugenics" to encourage reproduction among those who were presumed to hold superior gifted genes. On the other side, as many as 33 American states passed "negative eugenics" laws to promote compulsory sterilization surgeries to disabled individuals who were mentally disabled or ill, morally undesirable (like the prisoners), or who belonged to socially disadvantaged groups living on the margins of society.[13] These laws were upheld by the US Supreme Court in 1927, but the "negative eugenics" movement led to more than 60,000 sterilizations across the US.[13,14] German politicians and scientists endorsed the Nazi "racial hygiene" eugenic movement during 1933-1945. As a consequence of such motivation and actions, approximately 400,000 feeble patients were sterilized without consent and 275,000 of them were murdered by the Nazi "euthanasia" programs.[15-18] Some other countries also adapted such sterilisation programs, for example in Sweden, Canada and Japan.[19-21] In reaction to Nazi abuses, eugenics became almost universally reviled in many of the nations where it had once been popular. Scientists recognized the difficulty of predicting characteristics of offspring from their parents and demonstrated the inadequacy of simplistic theories of eugenics. The Universal Declaration of Human Rights was adopted by the United Nations in 1948 and affirmed, "Men and women of full age, without any limitation due to race, nationality or religion, have the right to marry and to found a family".

The modern concept of personalized medicine aims to use personal genetic information to predict or diagnose a disease (through prenatal diagnosis, neonatal screening, diagnosis of genetic disease in children, screening prospective parents for the carrier status of specific disorders, prediction for a serious late-onset disease), to minimize the exposure to environmental risks or to assess the differentiated response to a therapeutic drug.[11,22] In this review, we will first discuss how to estimate the clinical utility of genetic testing; second, illustrate the current status of personalized medicine with examples; third, highlight the challenges on the way towards personalized medicine; and last, envision the future of personalized medicine.

How to assess the clinical utility of a genetic marker

Whereas some genetic variants have an obvious clinical utility in disease diagnosis (e.g. mutation F508del in the *CFTR* gene and cystic fibrosis [23,24]), others genetic variants despite being strongly associated with diseases do not necessarily imply a predictive value in clinical practice.[25] The measurements of genetic variant's effect sizes (odds ratio, relative risk, hazard ratio) commonly used in traditional epidemiology are not adequate to determine the potential value of a genetic marker for predicting individual risk. The efficiency of a new test is typically evaluated by discrimination using a receiver operating characteristic (ROC) curve[26] or an alternative *c* statistic in survival data.[27] The ROC curve is a plot of sensitivity or the true positive (the probability of a positive test among those with the disease) verse 1-specificity or the false positive (the possibility of a positive test among those without the disease). Each point on the ROC curve represents the decision criterion at a given threshold. With a specific threshold, the predictor values above this are classified as positive (diseased category) and those lower than this are classified as negative (non-diseased category). The ROC curve also shows the trade-off

between sensitivity and specificity. In other words, any increase in sensitivity will be accompanied by a decrease in specificity. The area under the curve (AUC) from the ROC analysis is used to assess how well the model can distinguish people who do have the disease from those who do not. By definition, an AUC of 0.5 indicates classification of cases and controls by chance and 1 designates a perfect classification. AUCs of 0.50-0.70 are considered as low, 0.70-0.90 are considered as moderate, and > 0.9 are considered as high.[28] For example, in a study of prediction of depression in dementia in mild to moderate Alzheimer patients which is measured by the Cornell Scale based on signs and symptoms, an AUC of 0.91 meant that the probability was 91% that a randomly selected case had a higher Cornell Scale than a randomly selected non-case.[29,30] This approach has been widely used to examine the clinical utility of common and rare genetic variants in predicting the risk of having common diseases.[31-33] These results for the most part have showed that the addition of genetic variants only slightly or modestly improve the performance of risk prediction compared with the models with standard clinical risk factors. This phenomenon may be explained by the small individual effect size (odds ratios<1.5) of genetic variants analyzed separately and by an insufficient knowledge of disease predisposing genetic variants. Notably, Pepe *et al*., have suggested that an odds ratio of 3.0 or smaller may be of clinical importance in characterizing population variations in risk but may have little impact on the ROC curve or c statistics.[34] In other words, a strong association between an outcome and a predictor does not imply that the ROC curve analysis or *c* statistic will give rise to a good estimate of discrimination. Additionally, the ROC curve and c-statistic are insensitive to assessing the impact of adding new markers to an existing predictive model, especially when there is a correlation between them.[30]

In the end, when it comes to risk factors, patients and physicians alike are interested in the likelihood of disease development and options for a better medical management afterwards, rather than the true positive rate and true negative rate if the patient has been diagnosed. This can be measured by calibration or reclassification, another measurement of clinical utility. If a model with novel predictive markers can more accurately classify individuals into higher or lower risk categories, it is better calibrated and will lead to a better clinical outcome. For instance, three independent studies performed reclassification analysis using genetic variants to predict the risks of cardiovascular diseases, type 2 diabetes and breast cancer.[35-37] These studies showed various risk reclassification improvements from 4 to 53%. [35-37] For example, in Wacholder's study, after the addition of 10 common genetic variants associated with breast cancer into the traditional risk model, the AUC increased from 58% to 61.8% which was modest; but 32.5% of patients was reclassified into a higher quintile, 20.4% into a lower quintile, and 47.2% remained in the same quintile.[37] Thus, different therapeutic options would be applied to different subgroups and improved outcomes would be expected. Furthermore, whether the reclassification is correct can be tested using Hosmer-Lemeshow test.[38] Based on the reclassification table, a single measure named net reclassification index (NRI) was proposed by Pencina *et al*.[39] It examines the proportions moving up or down categories among cases and controls separately and NRI = [Pr(up|case)-Pr(down|case)]-[Pr(up|control)-Pr(down|control)]. The most advantageous feature of NRI over ROC curve analysis and reclassification is that the categories of up and down can be defined according to clinically important risk estimates. As a result NRI can detect the prediction of clinically significant improvement due to genetic markers. Strictly speaking, NRI is a measure of discrimination rather than calibration. Therefore, when clinical utility of genetic variants and

other molecular signatures are investigated, careful selection of relevant statistical metrics, such as risk reclassification and NRI, is essential.

Current personalized medicine applications

In the post-genomic era, the elucidation of genetic basis of human disorders is progressing with unprecedented rapidity. Genome-wide association studies (GWAS) have identified several thousand common and low-frequency single-nucleotide polymorphisms (SNPs) associated with human diseases. Whole-exome sequencing (WES) and whole-genome sequencing (WGS) have more recently led to the discovery of disease-causing rare variants. WES selectively sequences the coding regions and is useful to discover rare coding variants which usually have more severe functional consequences. WES has been successfully used to identify genetic determinants of both common and rare diseases.[40-42] WES is currently cheaper and more commonly used than WGS.[43] The applications of this new body of knowledge to state-of-the-art personalized medicine are described below.

Mendelian diseases

Until the advent of high-throughput technology, positional cloning and candidate gene approach were the primary methodologies by which approximately 2,000 genes causing Mendelian diseases were identified.[44,45] These genes represent the foundation on which the routine genetic tests that are widely used in clinical laboratories provide early diagnosis or early prediction. The relevance of mutations or structural variants responsible for Mendelian disorders is obvious in genetic tests as they have very clear effects on phenotype. The diagnosis of Mendelian disorders is more beneficial if efficient treatments are available. For example, permanent neonatal diabetes is caused by mutations in the *KCNJ11* and *ABCC8* genes.[46,47] The

two genes encode Kir6.2 and sulfonylurea receptor 1 (SUR1), the two subunits of the ATP-sensitive potassium (K_{ATP}) channel, and trans-activating mutations in these genes result in a failure of the beta-cell K_{ATP} channel to close in response to increased intracellular ATP and impaired insulin secretion.[48] Ninety percent of patients carrying a mutation in *KCNJ11* or *ABCC8* genes reverse diabetes when they are shifted from insulin to oral sulfonylurea medication.[47,49] However, the clinical diagnosis of permanent neonatal diabetes is based on Sanger sequencing of the PCR fragments from the *KCNJ11* and *ABCC8* genes. This molecular diagnosis is restricted to a limited number of the known mutations and other possible genetic loci elsewhere in the genome are not assessed. Recently, Bonnefond *et al*. performed WES for a permanent neonatal diabetes patient and identified a novel non-synonymous mutation (c.1455G>C/p.Q485H) in *ABCC8* gene which was missed by classical Sanger sequencing.[50] Using WES in the maturity-onset diabetes of the young (MODY) patients, the same research group found one mutation (p.Glu227Lys) in *KCNJ11*, indicating that such MODY patients can be ideally treated with oral sulfonylureas.[51] Although Sanger sequencing is the gold-standard DNA sequencing method, next generation sequencing (NGS) has its unique advantage at finding a novel disease-causing mutation in larger areas of the genome when the exact site of mutation is unknown.

When WES is performed, 20,000-30,000 genetic variants are typically identified in patients comparing to reference genomic sequences. A series of filtering strategies are then required to isolate the disease-causing variant(s).[52] Since the first report of the targeted capture and massively parallel sequencing of the exomes of 12 humans in 2009,[43] WES has identified many novel disease mutations that contribute to both Mendelian and common diseases.[52] In 2010, Sarah Ng and colleagues used WES to sequence four patients who were affected with

Miller syndrome (MIM#263750), an autosomal recessive inherited disorder. By simple filtering procedures using dbSNP and the HapMap databases to prioritize the candidate variants, they found Miller syndrome was caused by mutations in *DHODH* gene.[53] This was the first WES study that identified a causal gene for a Mendelian disorder. Targeted re-sequencing in another four affected individuals using Sanger approach found that all of them were compound heterozygotes for missense mutations in *DHODH*. Furthermore, each parent of the affected individual was a heterozygous carrier, none of the mutations appeared to be *de novo*, and none of the unaffected siblings were compound heterozygotes. All of these features supported the hypothesis that *DHODH* was the causal gene of Miller disorder.[53]

More recently, WES has not only led to the identification of a novel Mendelian mutation and the elucidation of a novel mechanism underlying inflammatory bowel disease (IBD), but also provided key information for the clinicians to find an effective treatment.[54] A boy started to present Crohn's disease-like symptoms when he was 15 months old. Comprehensive clinical evaluation and laboratory examinations (including genetic tests of defined forms of IBD) could not reach a conclusive diagnosis, thus his illness could not be controlled and was getting worse and life-threatening. When the patient was at age of 5 years and 8 month, a WES was conducted and a mutation in the X-linked inhibitor of apoptosis gene *XIAP* was identified. The affected boy was a hemizygote for a cysteine to tyrosine amino acid substitution, leading to a previously undefined form of IBD. This variant was confirmed and his mother was heterozygous carrier for the same mutation. XIAP protein has a central role in the pro-inflammatory response and bacterial sensing through the NOD signaling pathway.[55,56] In *in vitro* tests with the patient's cells, the mutated protein had an increased susceptibility to activation-induced cell death and defective response to NOD2 ligands and its function was tested. After receiving an allogeneic

hematopoietic progenitor cell transplant, the boy was able to eat and drink normally and there was no recurrence of gastrointestinal symptoms.[54]

These studies clearly demonstrate that disease-causing variants for Mendelian disorders can be directly identified by WES in several unrelated individuals or in a single family. In addition to filtering variants based on a variety of reference databases, another strategy used to remove benign variants is bioinformatics-based prediction of the putative impact of point mutations on the structure and function of human proteins like software PolyPhen-2 (Polymorphism Phenotyping v2) [57] which has been used in Bonnefond *et al.*'s study.[50] It should be known that such computational algorithms have at least 20% of false prediction.[52] In combination with other challenges encountered by WES during filtering and interpretation, current success rate of identifying causal mutations with WES is approximately 50%.[52] Theoretically, WES is expected to be more efficient when applied to recessive disorders because the likelihood to find homozygous or compound heterozygous carriers for rare non-synonymous variants is low.

Common diseases

Unlike Mendelian diseases, the predictive value of common genetic variants with modest effects identified by GWAS is limited in the context of common diseases. Some common loci with unusual large effect sizes have been used for disease prediction in clinical settings, for example, *HLA* variants in autoimmune disease like type 1 diabetes and rheumatoid arthritis, *APOE* in Alzheimer's disease, and *BRCA1* and *BRCA2* in breast and ovarian cancers.[58] It is important to mention that these variants were identified by linkage studies or candidate gene approach before the GWAS advent. Among thousands of genetic variants identified by GWAS, except for a handful of variants having odds ratios greater than 3, most of them so far have small

effects with a median odds ratio of 1.33.[59] When the associated variants thus far are considered together they generally account for a small proportion of the heritability of a specific disease.[60]

Is it too early to implement genomic information in the prediction of the risk of having a common disease? ROC analysis using genetic information from common variants identified by GWAS did not provide clinically relevant improvement in the prediction of type 2 diabetes or cardiovascular disease, even using more than 20 SNPs together.[32,61,62] Such failures are not surprising, as the variants selected in these studies are usually associated with the disease exceeding a stringent level of statistical significance ($P < 5 \times 10^{-8}$). Beyond these 'top hits', many genetic variants with true modest effects on the trait do not reach such a level of association because of statistical power issues. These variants are consequently excluded from the prediction analyses. Genome-wide association consortium initiatives studies with very large samples and the use of new algorithms may enable a better prediction of the risk of common diseases.

Height is a polygenic trait with an estimated heritability of 80%. To date, a large-scale GWAS meta-analysis in close to 200,000 subjects identified hundreds of genetic variants in 180 loci conclusively associated with height that together explain 20% of the genetic variation of height.[63] Yang *et al*, chose a method of restricted maximum likelihood that simultaneously accounted for all the SNPs (N=294,831) genotyped in a DNA array and explained 45% of the genetic variation of height.[64] Stahl *et al*. developed a novel method based on Bayesian inference and evidenced that thousands of common SNPs were able to explain approximate of 50% of the heritability for both cardiovascular diseases and Type 2 diabetes.[65] This suggests that many more SNPs contributing to the trait remain to be discovered and that GWAS from even larger studies and with better imputation methods (e.g. using the 1000 Genomes Project reference panel) will

continue to be highly productive for the discovery of additional susceptibility loci for common diseases. In another study, Wei *et al*. used a sophisticated Support Vector Machine (SVM) algorithm to assess the risk of type 1 diabetes using whole-genome genotyping array data.[66] They demonstrated that SVM could accurately assess the risk of type 1 diabetes with an AUC of approximate 0.84 in two independent datasets. This study also reported that the higher the heritability is, the more accurate prediction SVM provides. These studies suggest that the current lack of clinical relevance of prediction models for common diseases may be related to incomplete knowledge of the disease-associated SNPs and to the use of suboptimal methodologies. The integration of common genetic variation information into efficient prediction models is definitely relevant in personalized medicine.

Psychiatric diseases are currently diagnosed by symptoms and psychopathological tests with criteria from the Diagnostic and Statistical Manual of Mental Disorders (DSM, 5th edition).[67] These criteria are more categorical than quantitative, sometimes making the diagnosis ambiguous. Furthermore, it is common that different psychiatric disorders share biologic background and environmental exposures. Recently, Bragazzi proposed to apply OMICS science and personalized medicine to the field of psychiatry to refine the disease classification and diagnosis and tailor the therapeutic regimen.[68] Recently, Professor Bernard Lerer, the director of the Biological Psychiatry Laboratory at Hadassah-Hebrew University Medical Center, Israel, won the Werner Kalow Responsible Innovation Prize in Global Omics and Personalized Medicine because of his achievements in the development of methodology and novel discoveries in the field of psychiatric pharmacogenetics.[69] This shows a strong international peer-recognition for the success and potentials of personalized medicine in psychiatric disorders.

Along with common variants, low-frequency SNPs and rare variants are also important in the elucidation of missing heritability and in prediction of the risk for common diseases.[70,71] Many studies have provided clear evidence that rare variants contribute to chronic diseases.[72-75] By resequencing the exons and regulatory regions of 10 candidate genes, Nejentsev *et al*. identified that four rare variants in the exons and introns of *IFIH1* (encoding interferon induced with helicase C domain 1) gene were associated with type 1 diabetes, none of which was dependent on a known common SNP in the same gene, suggesting *IFIH1* gene is casual.[72] Large-scale exon re-sequencing of *MTNR1B* gene (encoding melatonin receptor 1B), which was initially found to be associated with type 2 diabetes by GWAS, revealed that 36 very rare variants with minor allele frequency less than 0.1% were associated with type 2 diabetes, and a pool of 13 of them having partial- or total-loss-of-function strongly increased the risk (odds ratio=5.67, 95% confidential interval: 2.17-14.82, $P=4.09 \times 10^{-4}$).[73] Subsequent biological evaluation of these rare variants further confirmed the functional link between *MTNR1B* and type 2 diabetes. An extended haplotype association study in an enrichment population of Ashkenazi Jewish, in which the prevalence of Crohn's disease is several-fold higher compared with non-Jewish European ancestry, has found an ethnic-specific missense rare mutation R642S in *HEATR3* to be associated with Crohn's disease.[74] An integrated simulation framework to mimic the empirical genetic data of common diseases suggested that rare variants played a significant causal role in explaining missing heritability, but it also excluded such an extreme hypothesis that rare variants are entirely responsible for disease.[76] Therefore, the combined effect of both common and rare genetic variants may significantly improve disease prediction. [77]

In addition to prediction based on GWAS data, the potential applications of WGS are being explored to predict the risk of common diseases. A report by Roberts *et al*.[78] constructed a

mathematical model and used the information of incidence of 27 common diseases from large monozygotic twin studies to assess the capacity of WGS data in predicting who were at risk of these diseases. They concluded that the predictive value of this approach was small. This study raised much debate.[79-81]. Begg and Golan criticized the analytic caveats in this study and proved that WGS could theoretically offer more optimistic risk prediction compared with what presented by Roberts *et al.*[79,80] As pointed out by Topol, the predictive capacity of WGS is unlikely to be sufficiently powerful until the sequences of many individuals with the same well-defined trait and advanced analytic approaches are available.[81] He stated with optimism that his lab would sequence 14 million people at the end of 2014. Another study sequenced whole-genome for eight individuals, four at upper and four at lower deciles of risk for metabolic, cardiovascular, skeletal and mental health.[82] Approximately two-thirds of the genetic predictions were concordant with longitudinal clinical measurements.

Combining genomic information with regular monitoring of clinical status which measures other "omics" profiling with different high-throughput platforms will theoretically improve personalized medicine. Recently, Chen *et al.* first used "integrative personal omics profiling" (iPOP), which included genomics, transcriptomic, proteomic, metabolomics and autoantibody profiles, to evaluate healthy and diseased status.[83] They collected blood samples from a 54-year-old male volunteer at 20 time points during a 24-month study and captured snapshots of several molecular metrics at different conditions of health (i.e. healthy, during viral infection, recovery). The subject coincidentally developed type 2 diabetes during the monitored time frame. The results captured extensive and dynamic changes in diverse molecular features and biological pathways that occurred as the subject transitioned from healthy to diseased conditions. Using poly-omics dataset, Heather *et al.* recently developed a method called

OmicKriging and showed substantially better performance in prediction of seven diseases than any single OMIC dataset in the study from the Wellcome Trust Case Control Consortium (WTCCC).[84] With this strategy, collective databases with "omics" profiles from more individuals with different diseases may be valuable in the diagnosis and monitoring diseases, even if this approach may not be realistic in a clinical setting.

Pharmacogenetics

Traditionally, clinical trials classify patients into different groups on a basis of symptoms (e.g. mild/severe depression) or histological patterns (e.g. breast cancer stage I/II/III), assuming that the patients within the same subgroup will have similar responses to treatment. This current symptom-based treatment regimen leads to more than 2 million adverse drug reactions annually in US alone with a cost of $76 billion for drug-related morbidity and mortality.[85] Generally speaking, with a given medication, 25-60% of the patients gain therapeutic benefits and the rest either do not respond or suffer from drug toxicity.[85] Administrating a drug to non-responders also induces colossal loss of money for the public health system. For example, 30-40% of the psychiatric patients with major depression do not respond to treatment with fluoxetine.[86] These numbers highlight the fact that individuals vary greatly in their response to treatment, and part of this response may be inherited. If the patients are stratified using genetic markers (or genomic markers such as gene expression signatures in the broader context of pharmacogenomics), subgroups are expected to become more homogenous and display a more similar response to the same treatment.

Pharmacogenetics refers to genetic variations that affect individual responses to drugs, in terms of both clinical efficacy and adverse effects, thus predicting efficacy and toxicity and indicating dosage adjustments.[87] The genes harboring these genetic markers usually encode

enzymes which are involved in the course of the pharmacokinetics and pharmacodynamics of the drug.

Cardiovascular medicine offers a good illustration of the impact of pharmacogenetics in clinical practice. Warfarin has been the most widely used oral anticoagulant for 60 years and it achieves therapeutic anticoagulation without excess risk of bleeding or thromboembolic events only within a narrow range of concentrations in the blood. The response to warfarin varies greatly from patient to patient and 10-20 fold differences in warfarin dosage have been reported to achieve the therapeutic effect.[88] As a result, warfarin use is associated with multiple dose adjustments, long periods of over- or under-anticoagulation for the patients, and inappropriate dosage of this drug is the leading cause of emergency department visits and hospitalizations due to an adverse drug reaction.[88] Finding new strategies for an effective and safe use of warfarin is therefore an ongoing and vital concern. Sequence variants in genes that encode cytochrome P450 2C9 (*CYP2C9*), a major enzyme that metabolizes warfarin, and vitamin K epoxide reductase (*VKORC1*), the molecular target of warfarin, have proved to contribute to more than 50% of dose variation among the patients.[89,90] In 2009, the International Warfarin Pharmacogenetics Consortium established a dose algorithm based on these genetic variants and clinical relevant indicators.[91] The results showed that this algorithm was superior to predominant strategy, using clinical variables only, at directing the initial dosage to achieve desirable and stable therapeutic concentrations. It identified 49.4% of the patients that needed lower doses and 24.8% that required higher doses, in comparison to 33.3% and 7.2% from clinical algorithm, thus providing a better dose adjustment and improved treatment. This algorithm has been followed by evidence-based studies to evaluate its effectiveness. Initial warfarin dosage adjusted from the patient's genotype data could reduce the risk of hospitalization in outpatients by 31% [92] and globally

improve the clinical outcomes including significantly lower rate of serious hemorrhage.[93] Based on this evidence, Food and Drug Administration (FDA) modified the warfarin label, stating that *CYP2C9* and *VKORC1* genotypes may be useful in determining the optimal initial dose of warfarin [94,95]. Most recently, two large randomized controlled trials tested the effect of the genotype-guided algorithm for warfarin dosing.[96,97] The study by Kimmel *et al* recruited patients from different ethnic groups in US and showed that the percentage of time reaching the therapeutic range was almost identical in both genotype-guided and clinically guided groups (45.2% vs. 45.4%) and the rates of side effects did not differ either.[96] A significant interaction between dosing and race was observed. Controversially, Pirmohamed *et al.* reported significant improvement in the percentage of time reaching the therapeutic range (67.4% vs. 60.3%) and significant decrease in the rate of side effects in the genotype-guided *versus* clinically-guided groups of Europeans.[97] However, these two studies were underpowered to assess the more important end-point, the rate of bleeding and thrombotic complications, which was reported as the secondary outcome in both trials.[98] Therefore, meta-analysis of these outcomes or randomized controlled trials based on ethnic-specific algorithms may be necessary, indicating that the promise of genotype-based algorithm is proving to be more difficult than first predicted.

Another example of pharmacogenetics at work is statin, a cholesterol-lowering drug that effectively reduces the incidence of heart attack and stroke.[99] However, high doses of statin (e.g. 80mg/day) may induce myopathy.[100] A GWAS that selected 175 matched cases and controls from a 12,000-participant trial identified a non-coding SNP rs4149056 strongly associated with statin-induced myopathy.[101] This variant is located in the gene *SLCO1B1*, a well-known regulator of the hepatic uptake of statin. The homozygotes of the risk allele (CC) have 16.9 times higher risk of myopathy than non-risk allele homozygotes (TT). The screening of this genetic

variant may help avoid serious side-effect of statin. However, the very low incidence of myopathy lowers the positive predictive value of this variant and reduces its cost-effectiveness, therefore, this pharmacogenetic indication has not been pursued by FDA.

Cytochrome P450s (CYPs) consist of a large family of metabolizing enzymes which are active in the metabolism of clinically used drugs like warfarin discussed above. P450 genes are polymorphic and variations in *CYP2D6* and *CYP2C19,* alone or together, have also been shown to cause the ultra-rapid or delayed clearance of many psychiatric medications.[102-104] For example, citalopram is one of the widely prescribed antidepressant medications, but more than 50% of the patients do not have a complete remission of their symptoms.[105] Citalopram is a highly selective serotonin reuptake inhibitor metabolized by CYP2C19, CYP3A4 and CYP2D6 enzymes.[106,107] Individual who are homozygous for *CYP2C19**17/*17 genotype (ultra-rapid metabolizer) have 42% lower of serum concentration of citalopram compared with those with normal function alleles and increase the probability of therapeutic failure.[108] Therefore, increasing the starting dose is recommended. On the other hand, individuals with the *CYP2C19**2/*2, *2/*3, *3/*3 (poor metabolizer) genotypes have higher serum concentration and increased risk of side effects, thus using 61% of the standard dose has been suggested.[109] Although minimal downward dose adjustment has been suggested for poor CYP2D6 metabolizers, a potential interaction between *CYP2C19* and *CYP2D6* effect has been reported and labeled by FDA.[104,110]

The number of pharmacogenetic associations is increasing steadily [111] and the FDA has appended pharmacogenetic information to approximate 140 drug labels across a variety of diseases and 23 of them are psychiatric medications (http://www.fda.gov/Drugs/ScienceResearch/ResearchAreas/Pharmacogenetics/).[112] Black box warnings on some drugs denote serious or life-threatening risk of adverse effects to patients related with specific genetic variants.

Importantly, such pharmacogenetics-based genotype tests should be considered before initiating drug treatment to maximize the patients' benefits and minimize the drug side effects. When someday a clinical genetic program which integrates drug-gene interactions will be applied into patient electronic medical record system, a patient's tested genetic information will help the physicians to choose the optimal drug and its appropriate initial dosage.[113] In fact, patient electronic medical records are gradually being introduced into clinical practice and will keep updated with evidence from pharmacogenetic research.[113]

Cancers

Cancer is a common disease that is standing on the frontier of personalized medicine. The importance of inherited cancer risk has long been realized and the American Society of Clinical Oncology (ASCO) released its first statement on genetic testing for cancer susceptibility in 1996. [114] This statement has since been updated repeatedly to keep up with the rapid pace of new discoveries in genetics.[115] Some of the genetic variants identified from germline genetic testing are highly penetrant and confer substantial increases in cancer risk. *BRCA1* and *BRCA2* are such examples, where breast-cancer risk by the age of 80 years in carriers of the *BRCA1* and *BRCA2* pathogenic mutations are 90% and 40%, respectively, though their frequencies in the population are low.[116] Therefore, if the mutations in *BRCA1* and *BRCA2* are detected in a woman with multiple affected family members, clinical decisions of intensive screening with mammography or magnetic resonance image, and even preventive surgery would be prudent. [115] Most genetic variants identified from GWAS are low-penetrant and have limited clinical relevance in the context of the currently applied methodologies. Thus, they are not currently used as part of standard cancer diagnostics.[115] The challenge is how to parse the flood of data into simple and usable information. Recently, Massachusetts-based Foundation Medicine has developed software

to interpret sequenced genomic data in tumor tissues and are now capable of sequencing up to 300 cancer related genes and extracting potentially actionable information for clinicians and studies are ongoing to link the results to care recommendations.[117]

Beyond genetic information, gene expression markers which measure the levels of messenger RNA (mRNA) are extremely useful in all aspects of cancer management, from disease classification, response to chemotherapy, development of new therapeutics, and prognosis.[118] In some tumors, like breast cancers and glioblastmas,[119] molecular markers have been implemented as disease classification criteria. Breast cancer has been distinguished into four molecular categories on the basis of histological patterns and gene-expression markers [120,121]: basal-like cancers (estrogen-receptor (ER)-negative, progesterone-receptor (PR)-negative, and human epidermal growth factor receptor 2 (HER2)-negative), luminal-A cancers (ER-positive and histological low-grade), luminal-B cancers (ER-positive and histological high-grade), and HER2-positive cancers. This classification is still evolving as more data from microarray profiling, which measures thousands of mRNA transcripts simultaneously, increase the number of categories and classifications under each type of cancer, providing more precise targeted and efficient therapy. Gene-expression signatures also provide a unique approach to identify certain primary tissue of a metastatic tissue, because expression pattern of the origin tissue are often retain in the cancer.[118]

Another two categories of biomarkers, epigenetic changes and microRNA, are increasingly thought to drive the development of cancers.[122-125] Epigenetic changes are heritable and cause the changes of gene expression without alteration of DNA sequence.[126] DNA methylation is the currently most studied epigenetic mechanism which has been linked to both normal development and human diseases.[126] In cancer, epigenetic mechanisms act in term of

silencing tumor suppressor genes and DNA repair genes and activating oncogenes.[122] For examples, methylation of tumor suppressor gene *BRCA1* is associated with breast cancer, activated DNA repair gene *MGMT* is associated with glioblastomas.[127,128] Recently, the genome-wide methylation technologies enable the comparison of DNA methylation patterns in normal and cancer cells.[129] Distinct patterns of DNA methylation have been reported to be associated with several cancers and their progression.[130] MicroRNAs are endogenous small (about 18-24 nucleotides) non-coding RNA molecules and are thought to play a key role in the regulation of translation and degradation of mRNA in the physiological and pathological process, including cancer.[131,132] MicroRNA expression profiling using microarrays has been linked to a wide range of human cancers such as prostate and colorectal cancers.[133] Importantly, abnormal DNA methylation and microRNA expression levels in the plasma or serum are non-invasive and are consistent with the methylation and microRNA status in the primary tumor. Because both epigenetic changes and microRNA expression are involved at every step of cancer development, and are potentially reversible by methylation inhibitors or antisense microRNAs, they hold promise in diagnosis, prognosis and specific tailored cancer therapies. But the clinical benefits are uncertain and lack scientific rigor at this early stage of evidence.[125,134]

Targeted therapy in cancer may also be directed by gene-expression based classification. Among breast cancer patients, 25-30% of them overexpress *HER2* which encodes a trans-membrane glycoprotein receptor and stimulates cell proliferation.[135] Meanwhile, the overexpressed *HER2* is highly associated with relapse within a short time and low survival rate. Trastuzumab, a recombinant monoclonal antibody, specifically targets *HER2*-postive breast cancer and improves the survival.[136,137] Similarly, Gefitinib targets the tyrosine kinase domain of the epidermal growth factor receptor, which is overexpressed in 40-80% of non-small-cell lung

cancers and other epithelial cancers. However, only 10% of non-small-cell lung patients harbor specific somatic mutations in the tyrosine domain and response quickly and well.[138] In the patients with the mutations, the response rate is 71% compared with 1% for those without.[139]

Gene expression signatures including several dozens of genes have been applied to predict clinical outcomes, thus avoiding the hazards of unnecessary or ineffective chemotherapy and expensive costs. Before the prognostic gene signature for breast cancer, the clinical guidelines based on histologic and clinical characteristics recommended chemotherapy for 85-90% of lymph-node-negative patients, even though about 60-70% of them would survive without it. A 70-gene signature (MammaPrint) derived from primary tumors has been used to predict distal metastasis and select patient for adjuvant systemic treatment.[140] The results showed that 52% of patients with "poor prognosis" needed chemotherapy, rather than 82% and 92% suggested by St Gallen and the National Institute of Health (NIH) guidelines, respectively. This predictive signature was later attested in an evidence-based study and approved by FDA.[141,142] This signature provides a powerful tool to allow clinician to avoid adjuvant systematic therapy to a specific group of patients with low metastatic scores. Another 76-gene-expression profiling from an independent study was reported to present similar results.[143] In parallel, many other gene expression profiles have been developed to optimize the use of therapeutics, identify the novel targets for drugs, and design clinical trials.[118,144]

In spite of unprecedented development of genomic application in cancers, and their promising potentials in personalized medicine, most of them do not have sufficient evidence to move to clinical application yet. Currently, there are only a few diseases and molecular subgroups in which the prognostic and therapeutic strategies are proved or recommended by

FDA, ASCO or the Evaluation of Genomic Applications in Practice and Prevention Initiative (EGAPP) working group.

Challenges and concerns

Technology and computational analysis development

Massively paralleled technology has made the cost of DNA sequencing plummet. Nevertheless, WGS remains too expensive to study most common diseases as well-powered studies typically require several thousand individuals. WES is a cost-effective alternative to WGS, but it does not include copy number variants and non-coding variants which may also be critical to the development of diseases.[145] Because NGS technology which is currently used in WGS and WES can only reads short lengths per run, identifying the copy number variants from WGS can be an arduous task. However, many NGS companies have been making significant improvement in read length and algorithms are being developed to capture these variants with WGS data.[146,147]

Another challenge is how to store and interpret the massive amount data of WGS from a group of participants. Even in the context of affordable WES/WGS strategies, other costs including storage of the data, analysis, validation and implementation may be still too expensive to extend their application in common diseases.[148] There is also an urgent need to develop software to figure out the "actionable" components which can be used in a more straightforward way to make a diagnosis, guide the change of the patients' lifestyle, or provide specific targets for pharmaceutics.[117]

Accuracy of prediction

GWAS have identified numerous genetic variants associated with common diseases, pharmacogenetic studies have discovered many variants associated with the efficacy or hazards

of a drug in a specific group of individuals, and plenty of gene-expression signatures have been reported to predict the outcomes of treatment; however, only a small portion of them have been approved for clinical use. There are three reasons for this. First, a genome-wide or an array-wide test may lead to many abnormal genomic findings which are unrelated to the primary reason, which is a phenomenon called "incidentalome". [149,150] As the number of tests (SNPs or gene expression) increases, the chance of a false-positive association increases as well. Second, researchers who discover novel genetic tests usually do not have the resources to conduct the evidence-based studies to examine their clinical utility. Third, there is insufficient clinical validation.[151] Three clinical trials testing the prediction of gene signatures on the outcomes of chemotherapy in non-small-cell lung cancer and breast cancer were suspended in 2011 because of the faults in the original data processing and analysis, and non-reproducibility.[152]

Recently, some genomic companies (23andMe, deCODEme, GeneticHealth and Navigenic) have started to provide genetic and genomic test on demand.[153] The relevance of this direct-to-consumer (DTC) medical service on disease risk estimation is controversial. The advocates may consider that DTC will improve the screening practices and motivate the buyers to switch to a healthier lifestyle; the opponents may ponder its safety, privacy and effectiveness.[154] The DTC results are not consistent when the same individual is assessed using different platforms offered by different companies, which may leave consumers confused or cause unnecessary anxiety from an unreliable diagnosis.[155,156] The risk predictions, especially for some serious diseases, are somewhat contradictory. Ng *et al.* ordered DTC tests for five individuals from two firms and they found that less than 50% of the risk estimations were consistent across them for seven diseases.[155] These discrepancies may be the consequences of different genetic markers used in different platforms. The genetic markers included in each

platform are chosen from GWAS, but different companies may have their own criteria and more than 40% of the genomic variants used in commercial tests have not been replicated in meta-analyses.[157] The algorithms they use to calculate the risk only include genes that explain small portion of heritability and rely on preliminary clinical relevance.[158] Moreover, some companies may update the markers with the ongoing discoveries in research, and some may not. This exemplifies the lack of validation and oversight and the insufficient medical input in the DTC business.

Training physicians and medical students

Today's physicians are facing the challenge of a transition from traditional to genomic medicine. Considering the growing number of approved genetic tests, a survey of American Medical Association members reported that only 10% respondents were confident enough to apply them in their practice.[159] Although the usefulness of epidermal growth factor receptor genetic testing in directing chemotherapy in lung cancer patients has been incorporated into the guidelines, one third of all physicians have yet to adopt it.[160] The emergence of DTC genomic service raises another challenge for traditional physicians. DTC has broken the established physician-patient relationship in which the clinical tests are ordered by physicians. Now thousands of people order their own genomic tests through DTC and bring the genomic profiles to their physicians. Many doctors are not familiar with the concepts of genomics and genomic medicine and are hard pressed to explain the estimated risks from such data.[161,162] Some physicians may take the uncertainty of the genetic test results as an excuse to reject them. On one hand, many patients believe that the doctors have an obligation to help them interpret and use the genetic results;[163] on the other hand, 83% of Americans do not believe their doctors are sufficiently trained in this capacity.[161] These facts highlight the urgent need to integrate the

education about the principles of genomic, targeted therapy, biomarker development, and biomarker-based clinical trials into the training curriculum and teaching program in the medical schools. Johns Hopkins University is leading this evolution by changing the teaching plans and opening new programs in the school of medicine.[159] The impetus came from the belief that every case is unique. A study introduced the 21-gene recurrence score assay to oncologists over standard tools to quantify the risk of distant recurrence and predict the extent of chemotherapy benefit in tamoxifen-treated patients with lymph node-negative, ER-positive breast cancer.[164] Before and after obtaining the score assay, the recommendation from the oncologists changed in 28 out of 89 cases. Among them, chemotherapy was removed from the treatment regimen in 20 cases. Meanwhile, the oncologists were more confident in their decision-making with the evidence from the score assay. Though this was a small study, it reflected the impact of genomic knowledge on the doctors' decision-making.[159]

Cost-effectiveness of genomic tests

Cost-effectiveness, which assesses whether a new diagnostic tool or a new drug is worth of its investment, is a critical concern for a health agency in allocation of limited health resources. Therefore, beyond clinical validity, cost-effectiveness presents another barrier to implement personalized genomic tests. In fact, genome-based diagnosis and therapies possess great potential to improve cost-effectiveness. Pharmacogenetic applications in cardiovascular diseases will improve effectiveness and decrease adverse effects; and predictive magnitude of chemotherapy in cancers will prevent prescription of expensive drugs in the non-responders and avoid toxicity as well. The examples from rare diseases may even better demonstrate this. Without a definite diagnosis, the patient will seek a variety of examinations and treatments which are actually useless. A baby suffering from a cascade of infections caused by severe

combined immunodeficiency disease (SCID) spent more than two months looking for many physicians before he got a conclusive diagnosis. At the end, he missed the treatment and died at 6 months and 15 days with a medical cost of $500,000. His younger sister who had the same disease was conclusively diagnosed by genotyping tests, received bone marrow transplantation at 16 days after birth, and survived with a lower bill than what her brother cost.[165]

Currently, most of the research grants are invested in basic discovery research, diagnostic and therapeutic clinical trials. There is only a small portion of research evaluating candidate applications and developing evidence-based recommendations, even fewer studies investigating cost-effectiveness in genomic research. The genomic research is still being ever-improving, with test accuracy keeping improved over time and costs dropping even faster. Re-evaluation of the cost-effectiveness might be necessary. Someday when everyone has his own genome sequence and the technologies are mature, cost-effectiveness may eventually not be a worry any more.

Gene patenting and prediction

A gene patent gives the owner of the gene exclusive rights for its application in research, diagnosis and therapeutics for 17 to 20 years and excludes anyone else from making, using or selling it. Currently, approximately 20% of the human genes had been patented and more than 40,000 DNA-related patents have been generated since 1982, when gene patents were first allowed.[166] Although gene patents are incentive to innovation, they also impede other institutes and companies to contribute to important genetic discoveries and limit patient access to health services. Whether genes should be patentable was a hot topic in the last couple of years because of the lawsuit in 2009 involving Myriad Genetics, a biotechnology company, which had owned the patents of *BRCA1* and *BRCA2*. Since Myriad won these patents in 1998, all laboratories across US that were doing such tests stopped their practice, whereas Myriad started to monopoly

the market with high price.[167] When a WES or a specific panel is able to sequence all exons and cancer-related genes in a single experiment, definitely including *BRCA1* and *BRCA2* and many other patent genes, doctors had to order them separately from other companies with authority or reported the results without the information of these genes if they did not buy licences. Furthermore, expensive cost for the patent genes adds another layer of complexity to cost-effectiveness analysis of genomic testing. In polygenic diseases, gene patents do stand in the middle to prevent scientist from doing better jobs towards personalized medicine. Fortunately and reasonably, on June 13, 2013, the US Supreme Court rejected Myriad's arguments and overturned the gene patents by saying that "genes are a product of nature and therefore are not patentable by law and myriad did not create anything". As hoped by many scientists and doctors, including Francis Collins, the director of the National Institute of Health, *BRCA1*, *BRCA2* and many other patent genes are set free.[167]

Ethical and legal issues

Many ethical and legal issues should be considered in the course of implementation of genetic and genomic testing.[85] People may reject genetic or genomic testing because they are afraid of genetic discrimination from insurance companies by denying coverage or from employers in employment decision. In 2008, the US Senate passed Genetic Information Nondiscrimination Act (GINA) to protect an individual's genetic information from insurance and employer discrimination.[168] This Act is also important to encourage Americans to make good use of genetic testing to prevent and prepare for potential diseases. Who else, except the patient, can the results be released to, and how to protect genetic privacy from the third party in the system of electronic medical record? There are still no answers for these questions. It is a challenging decision whether to inform children, adolescents or young adults when they have a diagnosis of a

cancer due to the special age window. It is however admitted that their awareness of their disease should offer a psyco-social support, thus leading to better compliance and adherence to the treatment and better clinical outcomes.[169] There is always a consensus to conduct newborn screening for a panel of early-onset but treatable diseases; however, newborn screening for late-onset or no cure diseases is controversial.[170] Some may consider screening for late-onset or no cure diseases adds extra anxiety for the individuals and their families if there is no preventive and early treatment options or no immediate intervention needed or the complications of a newborn screening are not clear;[171] others may think the testing can inform the individuals for their reproductive decision-making and the family for financial and psychological preparation. Some interesting concerns come with the advent of DTC. What are the proper procedures to obtain informed consents from DTC customers? Should only the results with sufficient clinical validity be reported to the patient or all of them? How to avoid the misleading or the uncertain results from DTC? Currently, there is no sufficient regulation on genetic and genomic testing. Some agencies like American Society of Clinical Oncology are calling for oversight from FDA and Center for Medicare and Medicaid Services to ensure highest standards for quality, accuracy, and reliability, but, on the flipping side, not hinder the scientific development or delivery of best available treatment and preventive care.[115] Fortunately, the FDA and other organizations have been active in addressing regulatory issues on personalized medicine. Very recently, the FDA has granted authorization for the first high-throughput genome sequencer, Illumina's MiSeqDx, for its clinical laboratory use because of its best performance in precision and reproducibility.[172] In February 2014, the FDA also withdrew the personal genome service from 23andMe due to its potential risks of inaccurate results [158]. We believe that this decision is a step in the right direction, as the accuracy of genetic testing must be controlled by authorized agencies in the best

interest of the patient. Some authorized organizations are making recommendations when personalized medicine is practiced.[173,174] For example, the American College of Medical Genetics and Genomics (ACMG) published a policy statement on clinical sequencing that a minimal list of genes and variants (currently in 24 diseases) should be routinely evaluated and reported as the incidental or secondary findings to the clinician who orders the test.[173]

The future of personalized medicine

Although many challenges and hurdles remain, for personalized medicine the future is bright. Recently, the term P4 medicine was coined by Leroy Hood.[175] It includes Predictive, Preventive, Personalized and Participatory aspects.[175-177] It is a system approach beyond genomics and uses each person's system biology, in combination with bioinformatics, to generate "actionable" regimen and convert billions of data points into an intelligible synopsis that is accessible to physicians and care providers. System biology consists of unique genomic sequence data that is combined with dynamic molecular and cellular information, as well as elastic environmental and phenotypic measurements that are fundamental health determinants. Compared with genomic medicine using one-dimensional data, P4 medicine utilizes biological information in totality to detect the disease-disturbed components, providing deep insights into disease mechanisms and new targets for diagnosis and therapeutic drugs. By identifying the actionable information from a vast composite of information, P4 medicine is quasi-holistic in its aim not only to demystify diseases but also to improve wellness, which meets with the latest definition of health edited by the World Health Organization as a state of complete physical, mental and social well-being. This P4 model expands personalized medicine beyond genomic medicine. Furthermore, P5 medicine with additional fifth P of Population science is proposed by Khoury, which is to be incorporated into each aspect of P4.[178] Population science covers almost

every aspects of health and uses ecologic model systems and mixed methods to input intelligence from multiple disciplines. It assesses the validity of evidence from P4 and is useful in guiding policy making.[179] From a population perspective, biological signatures from P4 models of uncertain clinical utility require strong evidence from randomized controlled trials before clinical use is recommended.[178] Among hundreds of reported predictive gene signatures of different cancers, only a handful of tests passed the FDA approval.[152,180] Without sufficient clinical validation, the newly developed personalized medicine strategies from P4 medicine may be misleading and consequently may be a waste of resources and do more harm than good to the patients. Meanwhile, a different P5 model with the different fifth P of Psyco-cognitive aspect was proposed by Gorini and Pravettoni.[181,182] Such a P5 medicine will not only inform the patients of their health status, but also empower them to be involved in their decision-making with doctors by their specific needs, values, behaviours, hopes and fears. Following this, a sixth P of Public was introduced by Bragazzi who was inspired by a Salvatore Iaconesi's clinical story.[183] Salvatore Iaconesi is a skilled computer scientist and one day was diagnosed with a brain tumor. He posted his medical records on his website and desired to seek help from various sources and shared his experience with anyone who needed it.[183] In other words, P6 approach brings up the additional notion of e-health into personalized medicine. The sixth P is an interesting concept but it may lead to important ethical considerations such as confidentiality, discrimination and implications to family members, and therefore its applications are limited.

Hood and Flores also portray a stunning picture of future P4/P5 medicine and predict that it would likely become true within the next decade.[184] They assume that accurate assessments from genomic sequence to proteomics and their function, to conventional medical data, to enormous amounts of clinical diagnostic imaging and environmental measurements would be

available, affordable and accessible for individuals. The leading edge biology and medicine in every field of "omics" will drive the development of new high-throughput technology and analytic tools to explore the multi-dimensional data from individuals, families, and across the population. P4/P5 medicine considers each person as unique, thus each has his own genome which would need to be sequenced only once, while measurements of other dynamic parameters, would require more regular assessments (e.g annually or biannually). By analyzing these data any transition from health to disease will be marked.[185] Genomic variants and protein profiles will also be used to assess drug toxicities, avoiding unnecessary adverse effects. P4/P5 medicine model is characterized by stratifying health and disease based on different markers and extracting actionable variants. Assuming that targeted drugs that are effective at different stages of disease progression are available in the future, tailored interventions will be engaged to correct a disease-perturbed network to restore an individual to wellness. All these information is linked to the individual's electronic medical records and the doctors will receive health messages in time such as health status change, drug choice and dosage, or progress/prognosis of a disease, achieving personalized prevention and treatment. More importantly, the P4/P5 medicine model postulates that individuals are active and networked rather than simple passive recipients of doctors' advice. Their participation will contribute to the advancement of medical and health knowledge and will eventually maximize their own wellness. They will be the most powerful drivers of the emergence of P4/P5 medicine. P4/P5 medicine also has the potential to drop the ever increasing costs of health care by active prevention, early diagnosis and specific treatments.

Does this sound like a scientific fiction story? Are they castles in the sky? Because we have witnessed the unprecedented success of human genomic, this ambitious vision should not be rejected. However, in the first decade after deciphering the human genome, only a handful of

genetic discoveries have been applied into routine medical practice and the clinical benefits are still far from enhancing the wellness and treating diseases for most individuals.[25,186] In addition to genomics, integration of other types of personal "omics" profiles including transcriptomics, proteomics, metabolomics, epigenomics, metagenomics will theoretically enable to understand the onset, progression and prognosis of common diseases, thus broadening the capability of personalized medicine.[187] The laboratory experiments have shown that the levels of these "omics" vary greatly across time, within individuals, and between individuals, and this massive variation has made clear interpretations difficult. Meanwhile, many of these analyses are currently prohibitively expensive. Importantly, P4/P5 medicine is built on stringent assumptions that all these "omics" are accurately measured. Therefore, it is too optimistic to build up such a system with integration of huge data that are not yet fully-understood.

Will this become reality in 10 years? P4/P5 medicine will use multi-level data within individuals and across a population to generate lots of information which can be used to improve health. Obviously, this complicated system in P4/P5 medicine model cannot be mimicked in the experiment settings. Therefore, one critical prerequisite to practice P4/P5 medicine is that all the elements in system biology should be clinically valid before they are used for final outcome syntheses. Over the past a few years, many evidence-based studies were undertaken to assess the clinical validity and utility of emerging genetic testing. The Evaluation of Genomic Applications in Practice and Prevention Initiative (EGAPP) Working Group, established in 2005, reviews evidence reports from randomized controlled trials and/or observational studies and assesses the analytic validity and clinical validity, providing recommendations on the appropriate use of genetic tests in specific clinical scenarios. Currently, EGAPP have released 11 recommendations, in which only 3 have sufficient evidence. The lack of information on the clinical validity for

most genetic and molecular tests is a major practical barrier to the implementation of P4/P5 medicine.[188] Another concern is that it takes average of 17 years to translate a new scientific discovery to clinical practice, with a success rate of less than 15%.[160] Furthermore, this P4/P5 medicine revolution will not happen without a new generation of experts who are able to create algorithms to integrate and interpret all the diverse sources of information from genetics, molecular biology, clinical knowledge, statistics and bioinformatics, and eventually synthesize the actionable messages for the clinicians and patients. We believe that P4/P5 medicine can progress with exponential acceleration as genomic science does, but it will be a long journey to reach the full potential of personalized medicine.

Conclusions

Because an individual's DNA sequence is static unless exceptional circumstances occur (eg. tumor, exposure to mutagenesis compounds), it is considered to be an easier and more reliable tool to predict long-term risk.[189] This review illustrates some of the successes of using personal genomic data in Mendelian and polygenic diseases. Personalized medicine is in its infancy and is also moving steadily forward, but many challenges remain. We describe the hopes and hypes of personalized P4/P5 medicine which is driven by advances in technologies such as OMICS platforms, computation, information integration, and analyses. We hope this review will encourage clinicians to be active contributors in this medical revolution.

Conflict of interest

The authors confirm that this article content has no conflict of interest.

Acknowledgements

We thank Arkan Al Abadi for his suggestive comments at the early stage of this manuscript and his editing of the manuscript. We also thank the reviewers for their helpful comments. David Meyre is supported by a Tier 2 Canada Research Chair. Aihua Li is supported by a Queen Elizabeth II Graduate Scholarship in Science and Technology.

References

1 Adams, F. L. *The genuine works of Hippocrates.*, Vol. 2 (William Wood, 1886).
2 Strachan, T. & Read, A. *Human molecular genetics*. 7th edn, (Garland Science, 2011).
3 Harris, H. (Oxford University, London, 1963).
4 Guttmacher, A. E. & Collins, F. S. Genomic medicine--a primer. *N. Engl. J. Med.* **347**, 1512-1520, doi:10.1056/NEJMra012240 (2002).
5 Goodman, D. M., Lynm, C. & Livingston, E. H. JAMA patient page. Genomic medicine. *JAMA* **309**, 1544, doi:10.1001/jama.2013.1927 (2013).
6 Lander, E. S. *et al.* Initial sequencing and analysis of the human genome. *Nature* **409**, 860-921, doi:10.1038/35057062 (2001).
7 Venter, J. C. *et al.* The sequence of the human genome. *Science* **291**, 1304-1351, doi:10.1126/science.1058040 (2001).
8 Finishing the euchromatic sequence of the human genome. *Nature* **431**, 931-945, doi:10.1038/nature03001 (2004).
9 Altshuler, D. M. *et al.* Integrating common and rare genetic variation in diverse human populations. *Nature* **467**, 52-58, doi:10.1038/nature09298 (2010).
10 Abecasis, G. R. *et al.* An integrated map of genetic variation from 1,092 human genomes. *Nature* **491**, 56-65, doi:10.1038/nature11632 (2012).
11 Khoury, M. J., McCabe, L. L. & McCabe, E. R. Population screening in the age of genomic medicine. *N. Engl. J. Med.* **348**, 50-58, doi:10.1056/NEJMra013182 (2003).
12 Gillham, N. W. Sir Francis Galton and the birth of eugenics. *Annu. Rev. Genet.* **35**, 83-101, doi:10.1146/annurev.genet.35.102401.090055 (2001).
13 Kevles, D. in

(New York: Knopf, 1985).
14 Carey, A. Gender and compulsory sterilization programs in America: 1907–1950. *J Historical Sociol* **11**, 74-105 (1998).
15 Cohen, M. M., Jr. Genetic drift. Overview of German, Nazi, and Holocaust medicine. *Am J Med Genet A* **152A**, 687-707, doi:10.1002/ajmg.a.32807 (2010).
16 Hughes, J. T. Neuropathology in Germany during World War II: Julius Hallervorden (1882-1965) and the Nazi programme of 'euthanasia'. *J Med Biogr* **15**, 116-122 (2007).
17 Zeidman, L. A. Neuroscience in Nazi Europe part I: eugenics, human experimentation, and mass murder. *Can. J. Neurol. Sci.* **38**, 696-703 (2011).
18 Zeidman, L. A. Neuroscience in Nazi Europe part II: resistance against the third reich. *Can. J. Neurol. Sci.* **38**, 826-838 (2011).
19 Adams, M. *The wellborn science: Eugenics in Germany, France, Brazil , and Russia*. (Oxford University Press, 1990).
20 Caulfield, T. & Robertson, J. Genetic policies in Alberta: from the systematic to the systemic. . *Alberta Law Review* **35**, 59-80 (1996).
21 Tsuchiya, T. Eugenic sterilizations in Japan and recent demands for an apology: a report. *Ethics Intellect Disabil* **3**, 1-4 (1997).
22 Andermann, A. & Blancquaert, I. Genetic screening: A primer for primary care. *Can. Fam. Physician* **56**, 333-339 (2010).
23 McIntosh, I. & Cutting, G. R. Cystic fibrosis transmembrane conductance regulator and the etiology and pathogenesis of cystic fibrosis. *FASEB J.* **6**, 2775-2782 (1992).

24 Castellani, C. *et al.* Consensus on the use and interpretation of cystic fibrosis mutation analysis in clinical practice. *J Cyst Fibros* **7**, 179-196, doi:10.1016/j.jcf.2008.03.009 (2008).

25 Collins, F. Has the revolution arrived? *Nature* **464**, 674-675, doi:10.1038/464674a (2010).

26 Hanley, J. A. & McNeil, B. J. The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology* **143**, 29-36 (1982).

27 Harrell, F. E., Jr., Califf, R. M., Pryor, D. B., Lee, K. L. & Rosati, R. A. Evaluating the yield of medical tests. *JAMA* **247**, 2543-2546 (1982).

28 Fischer, J. E., Bachmann, L. M. & Jaeschke, R. A readers' guide to the interpretation of diagnostic test properties: clinical example of sepsis. *Intensive Care Med.* **29**, 1043-1051, doi:10.1007/s00134-003-1761-8 (2003).

29 Vida, S., Des Rosiers, P., Carrier, L. & Gauthier, S. Depression in Alzheimer's disease: receiver operating characteristic analysis of the Cornell Scale for Depression in Dementia and the Hamilton Depression Scale. *J. Geriatr. Psychiatry Neurol.* **7**, 159-162 (1994).

30 Cook, N. R. Use and misuse of the receiver operating characteristic curve in risk prediction. *Circulation* **115**, 928-935, doi:10.1161/CIRCULATIONAHA.106.672402 (2007).

31 Morandi, A. *et al.* Estimation of newborn risk for child or adolescent obesity: lessons from longitudinal birth cohorts. *PloS one* **7**, e49919, doi:10.1371/journal.pone.0049919 (2012).

32 Anand, S. S. *et al.* Genetic information and the prediction of incident type 2 diabetes in a high-risk multiethnic population: the EpiDREAM genetic study. *Diabetes Care* **36**, 2836-2842, doi:10.2337/dc12-2553 (2013).

33 Simard, J. *et al.* Evaluation of BRCA1 and BRCA2 mutation prevalence, risk prediction models and a multistep testing approach in French-Canadian families with high risk of breast and ovarian cancer. *J. Med. Genet.* **44**, 107-121, doi:10.1136/jmg.2006.044388 (2007).

34 Pepe, M. S., Janes, H., Longton, G., Leisenring, W. & Newcomb, P. Limitations of the odds ratio in gauging the performance of a diagnostic, prognostic, or screening marker. *Am. J. Epidemiol.* **159**, 882-890 (2004).

35 Kathiresan, S. *et al.* Polymorphisms associated with cholesterol and risk of cardiovascular events. *N. Engl. J. Med.* **358**, 1240-1249, doi:10.1056/NEJMoa0706728 (2008).

36 Meigs, J. B. *et al.* Genotype score in addition to common risk factors for prediction of type 2 diabetes. *N. Engl. J. Med.* **359**, 2208-2219, doi:10.1056/NEJMoa0804742 (2008).

37 Wacholder, S. *et al.* Performance of common genetic variants in breast-cancer risk models. *N. Engl. J. Med.* **362**, 986-993, doi:10.1056/NEJMoa0907727 (2010).

38 Cook, N. R. Comments on 'Evaluating the added predictive ability of a new marker: From area under the ROC curve to reclassification and beyond' by M. J. Pencina et al., Statistics in Medicine (DOI: 10.1002/sim.2929). *Stat. Med.* **27**, 191-195, doi:10.1002/sim.2987 (2008).

39 Pencina, M. J., D'Agostino, R. B., Sr., D'Agostino, R. B., Jr. & Vasan, R. S. Evaluating the added predictive ability of a new marker: from area under the ROC curve to reclassification and beyond. *Stat. Med.* **27**, 157-172; discussion 207-112, doi:10.1002/sim.2929 (2008).

40 Kiezun, A. *et al.* Exome sequencing and the genetic basis of complex traits. *Nat. Genet.* **44**, 623-630, doi:10.1038/ng.2303 (2012).
41 Sanders, S. J. *et al.* De novo mutations revealed by whole-exome sequencing are strongly associated with autism. *Nature* **485**, 237-241, doi:10.1038/nature10945 (2012).
42 Neale, B. M. *et al.* Patterns and rates of exonic de novo mutations in autism spectrum disorders. *Nature* **485**, 242-245, doi:10.1038/nature11011 (2012).
43 Ng, S. B. *et al.* Targeted capture and massively parallel sequencing of 12 human exomes. *Nature* **461**, 272-276, doi:10.1038/nature08250 (2009).
44 Botstein, D. & Risch, N. Discovering genotypes underlying human phenotypes: past successes for mendelian disease, future approaches for complex disease. *Nat. Genet.* **33 Suppl**, 228-237, doi:10.1038/ng1090 (2003).
45 Hamosh, A., Scott, A. F., Amberger, J. S., Bocchini, C. A. & McKusick, V. A. Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders. *Nucleic Acids Res* **33**, D514-517, doi:10.1093/nar/gki033 (2005).
46 Gloyn, A. L. *et al.* Activating mutations in the gene encoding the ATP-sensitive potassium-channel subunit Kir6.2 and permanent neonatal diabetes. *N. Engl. J. Med.* **350**, 1838-1849, doi:10.1056/NEJMoa032922 (2004).
47 Babenko, A. P. *et al.* Activating mutations in the ABCC8 gene in neonatal diabetes mellitus. *N. Engl. J. Med.* **355**, 456-466, doi:10.1056/NEJMoa055068 (2006).
48 Ashcroft, F. M. ATP-sensitive potassium channelopathies: focus on insulin secretion. *J. Clin. Invest.* **115**, 2047-2058, doi:10.1172/JCI25495 (2005).
49 Pearson, E. R. *et al.* Switching from insulin to oral sulfonylureas in patients with diabetes due to Kir6.2 mutations. *N. Engl. J. Med.* **355**, 467-477, doi:10.1056/NEJMoa061759 (2006).
50 Bonnefond, A. *et al.* Molecular diagnosis of neonatal diabetes mellitus using next-generation sequencing of the whole exome. *PLoS One* **5**, e13630, doi:10.1371/journal.pone.0013630 (2010).
51 Bonnefond, A. *et al.* Whole-exome sequencing and high throughput genotyping identified KCNJ11 as the thirteenth MODY gene. *PLoS One* **7**, e37423, doi:10.1371/journal.pone.0037423 (2012).
52 Robinson, P. N., Krawitz, P. & Mundlos, S. Strategies for exome and genome sequence data analysis in disease-gene discovery projects. *Clin. Genet.* **80**, 127-132, doi:10.1111/j.1399-0004.2011.01713.x (2011).
53 Ng, S. B. *et al.* Exome sequencing identifies the cause of a mendelian disorder. *Nat. Genet.* **42**, 30-35, doi:10.1038/ng.499 (2010).
54 Worthey, E. A. *et al.* Making a definitive diagnosis: successful clinical application of whole exome sequencing in a child with intractable inflammatory bowel disease. *Genet Med* **13**, 255-262, doi:10.1097/GIM.0b013e3182088158 (2011).
55 Huang, Y. *et al.* Structural basis of caspase inhibition by XIAP: differential roles of the linker versus the BIR domain. *Cell* **104**, 781-790 (2001).
56 Krieg, A. *et al.* XIAP mediates NOD signaling via interaction with RIP2. *Proc. Natl. Acad. Sci. U. S. A.* **106**, 14524-14529, doi:10.1073/pnas.0907131106 (2009).
57 Adzhubei, I. A. *et al.* A method and server for predicting damaging missense mutations. *Nat Methods* **7**, 248-249, doi:10.1038/nmeth0410-248 (2010).
58 Jostins, L. & Barrett, J. C. Genetic risk prediction in complex disease. *Hum. Mol. Genet.* **20**, R182-188, doi:10.1093/hmg/ddr378 (2011).

59 Hindorff, L. A. *et al.* Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *Proc. Natl. Acad. Sci. U. S. A.* **106**, 9362-9367, doi:10.1073/pnas.0903103106 (2009).
60 Goldstein, D. B. Common genetic variation and human traits. *N. Engl. J. Med.* **360**, 1696-1698, doi:10.1056/NEJMp0806284 (2009).
61 Paynter, N. P. *et al.* Cardiovascular disease risk prediction with and without knowledge of genetic variation at chromosome 9p21.3. *Ann. Intern. Med.* **150**, 65-72 (2009).
62 Lango, H. *et al.* Assessing the combined impact of 18 common genetic variants of modest effect sizes on type 2 diabetes risk. *Diabetes* **57**, 3129-3135, doi:10.2337/db08-0504 (2008).
63 Lango Allen, H. *et al.* Hundreds of variants clustered in genomic loci and biological pathways affect human height. *Nature* **467**, 832-838, doi:10.1038/nature09410 (2010).
64 Yang, J. *et al.* Common SNPs explain a large proportion of the heritability for human height. *Nat. Genet.* **42**, 565-569, doi:10.1038/ng.608 (2010).
65 Stahl, E. A. *et al.* Bayesian inference analyses of the polygenic architecture of rheumatoid arthritis. *Nat. Genet.* **44**, 483-489, doi:10.1038/ng.2232 (2012).
66 Wei, Z. *et al.* From disease association to risk assessment: an optimistic view from genome-wide association studies on type 1 diabetes. *PLoS Genet* **5**, e1000678, doi:10.1371/journal.pgen.1000678 (2009).
67 (American Psychiatric Association, Arlington, VA, 2013).
68 Bragazzi, N. L. Rethinking psychiatry with OMICS science in the age of personalized P5 medicine: ready for psychiatome? *Philos Ethics Humanit Med* **8**, 4, doi:10.1186/1747-5341-8-4 (2013).
69 Ozdemir, V. *et al.* Bernard Lerer: recipient of the 2014 inaugural Werner Kalow Responsible Innovation Prize in Global Omics and Personalized Medicine (Pacific Rim Association for Clinical Pharmacogenetics). *OMICS* **18**, 211-221, doi:10.1089/omi.2014.0029 (2014).
70 Manolio, T. A. *et al.* Finding the missing heritability of complex diseases. *Nature* **461**, 747-753, doi:10.1038/nature08494 (2009).
71 Gibson, G. Rare and common variants: twenty arguments. *Nat Rev Genet* **13**, 135-145, doi:10.1038/nrg3118 (2011).
72 Nejentsev, S., Walker, N., Riches, D., Egholm, M. & Todd, J. A. Rare variants of IFIH1, a gene implicated in antiviral responses, protect against type 1 diabetes. *Science* **324**, 387-389, doi:10.1126/science.1167728 (2009).
73 Bonnefond, A. *et al.* Rare MTNR1B variants impairing melatonin receptor 1B function contribute to type 2 diabetes. *Nat. Genet.* **44**, 297-301, doi:10.1038/ng.1053 (2012).
74 Zhang, W. *et al.* Extended haplotype association study in Crohn's disease identifies a novel, Ashkenazi Jewish-specific missense mutation in the NF-kappaB pathway gene, HEATR3. *Genes Immun* **14**, 310-316, doi:10.1038/gene.2013.19 (2013).
75 Ichimura, A. *et al.* Dysfunction of lipid sensor GPR120 leads to obesity in both mouse and human. *Nature* **483**, 350-354, doi:10.1038/nature10798 (2012).
76 Agarwala, V., Flannick, J., Sunyaev, S. & Altshuler, D. Evaluating empirical bounds on complex disease genetic architecture. *Nat. Genet.* **45**, 1418-1427, doi:10.1038/ng.2804 (2013).
77 Janssens, A. C. *et al.* Predictive testing for complex diseases using multiple genes: fact or fiction? *Genet Med* **8**, 395-400, doi:10.109701.gim.0000229689.18263.f4 (2006).

78 Roberts, N. J. *et al.* The predictive capacity of personal genome sequencing. *Sci Transl Med* **4**, 133ra158, doi:10.1126/scitranslmed.3003380 (2012).
79 Begg, C. B. & Pike, M. C. Comment on "the predictive capacity of personal genome sequencing". *Sci Transl Med* **4**, 135le133; author reply 135lr133, doi:10.1126/scitranslmed.3004162 (2012).
80 Golan, D. & Rosset, S. Comment on "the predictive capacity of personal genome sequencing". *Sci Transl Med* **4**, 135le134; author reply 135lr133, doi:10.1126/scitranslmed.3004197 (2012).
81 Topol, E. J. Comment on "the predictive capacity of personal genome sequencing". *Sci Transl Med* **4**, 135le135; author reply 135lr133, doi:10.1126/scitranslmed.3004126 (2012).
82 Patel, C. J. *et al.* Whole Genome Sequencing in support of Wellness and Health Maintenance. *Genome Med* **5**, 58, doi:10.1186/gm462 (2013).
83 Chen, R. *et al.* Personal omics profiling reveals dynamic molecular and medical phenotypes. *Cell* **148**, 1293-1307, doi:10.1016/j.cell.2012.02.009 (2012).
84 Wheeler, H. E. *et al.* Poly-Omic Prediction of Complex Traits: OmicKriging. *Genet. Epidemiol.*, doi:10.1002/gepi.21808 (2014).
85 Pasic, M. D., Samaan, S. & Yousef, G. M. Genomic medicine: new frontiers and new challenges. *Clin. Chem.* **59**, 158-167, doi:10.1373/clinchem.2012.184622 (2013).
86 Blazquez, A., Mas, S., Plana, M. T., Lafuente, A. & Lazaro, L. Fluoxetine pharmacogenetics in child and adult populations. *Eur. Child Adolesc. Psychiatry* **21**, 599-610, doi:10.1007/s00787-012-0305-6 (2012).
87 Klotz, U. The role of pharmacogenetics in the metabolism of antiepileptic drugs: pharmacokinetic and therapeutic implications. *Clin. Pharmacokinet.* **46**, 271-279, doi:10.2165/00003088-200746040-00001 (2007).
88 Johnson, J. A. Warfarin pharmacogenetics: a rising tide for its clinical value. *Circulation* **125**, 1964-1966, doi:10.1161/CIRCULATIONAHA.112.100628 (2012).
89 Sconce, E. A. *et al.* The impact of CYP2C9 and VKORC1 genetic polymorphism and patient characteristics upon warfarin dose requirements: proposal for a new dosing regimen. *Blood* **106**, 2329-2333, doi:10.1182/blood-2005-03-1108 (2005).
90 Rieder, M. J. *et al.* Effect of VKORC1 haplotypes on transcriptional regulation and warfarin dose. *N. Engl. J. Med.* **352**, 2285-2293, doi:10.1056/NEJMoa044503 (2005).
91 Klein, T. E. *et al.* Estimation of the warfarin dose with clinical and pharmacogenetic data. *N. Engl. J. Med.* **360**, 753-764, doi:10.1056/NEJMoa0809329 (2009).
92 Epstein, R. S. *et al.* Warfarin genotyping reduces hospitalization rates results from the MM-WES (Medco-Mayo Warfarin Effectiveness study). *J. Am. Coll. Cardiol.* **55**, 2804-2812, doi:10.1016/j.jacc.2010.03.009 (2010).
93 Anderson, J. L. *et al.* A randomized and clinical effectiveness trial comparing two pharmacogenetic algorithms and standard care for individualizing warfarin dosing (CoumaGen-II). *Circulation* **125**, 1997-2005, doi:10.1161/CIRCULATIONAHA.111.070920 (2012).
94 Gage, B. F. *et al.* Use of pharmacogenetic and clinical factors to predict the therapeutic dose of warfarin. *Clin. Pharmacol. Ther.* **84**, 326-331, doi:10.1038/clpt.2008.10 (2008).
95 Cavallari, L. H., Shin, J. & Perera, M. A. Role of pharmacogenomics in the management of traditional and novel oral anticoagulants. *Pharmacotherapy* **31**, 1192-1207, doi:10.1592/phco.31.12.1192 (2011).

96 Kimmel, S. E. *et al.* A pharmacogenetic versus a clinical algorithm for warfarin dosing. *N. Engl. J. Med.* **369**, 2283-2293, doi:10.1056/NEJMoa1310669 (2013).
97 Pirmohamed, M. *et al.* A randomized trial of genotype-guided dosing of warfarin. *N. Engl. J. Med.* **369**, 2294-2303, doi:10.1056/NEJMoa1311386 (2013).
98 Furie, B. Do pharmacogenetics have a role in the dosing of vitamin K antagonists? *N. Engl. J. Med.* **369**, 2345-2346, doi:10.1056/NEJMe1313682 (2013).
99 Baigent, C. *et al.* Efficacy and safety of cholesterol-lowering treatment: prospective meta-analysis of data from 90,056 participants in 14 randomised trials of statins. *Lancet* **366**, 1267-1278, doi:10.1016/S0140-6736(05)67394-1 (2005).
100 Bowman, L., Armitage, J., Bulbulia, R., Parish, S. & Collins, R. Study of the effectiveness of additional reductions in cholesterol and homocysteine (SEARCH): characteristics of a randomized trial among 12064 myocardial infarction survivors. *Am. Heart J.* **154**, 815-823, 823 e811-816, doi:10.1016/j.ahj.2007.06.034 (2007).
101 Link, E. *et al.* SLCO1B1 variants and statin-induced myopathy--a genomewide study. *N. Engl. J. Med.* **359**, 789-799, doi:10.1056/NEJMoa0801936 (2008).
102 Hicks, J. K. *et al.* Clinical Pharmacogenetics Implementation Consortium guideline for CYP2D6 and CYP2C19 genotypes and dosing of tricyclic antidepressants. *Clin. Pharmacol. Ther.* **93**, 402-408, doi:10.1038/clpt.2013.2 (2013).
103 Sim, S. C. *et al.* Association between CYP2C19 polymorphism and depressive symptoms. *Am J Med Genet B Neuropsychiatr Genet* **153B**, 1160-1166, doi:10.1002/ajmg.b.31081 (2010).
104 Mrazek, D. A. *et al.* CYP2C19 variation and citalopram response. *Pharmacogenet Genomics* **21**, 1-9 (2011).
105 Thase, M. E. *et al.* Remission rates following antidepressant therapy with bupropion or selective serotonin reuptake inhibitors: a meta-analysis of original data from 7 randomized controlled trials. *J. Clin. Psychiatry* **66**, 974-981 (2005).
106 Olesen, O. V. & Linnet, K. Studies on the stereoselective metabolism of citalopram by human liver microsomes and cDNA-expressed cytochrome P450 enzymes. *Pharmacology* **59**, 298-309, doi:28333 (1999).
107 von Moltke, L. L. *et al.* Citalopram and desmethylcitalopram in vitro: human cytochromes mediating transformation, and cytochrome inhibitory effects. *Biol. Psychiatry* **46**, 839-849 (1999).
108 Rudberg, I., Mohebi, B., Hermann, M., Refsum, H. & Molden, E. Impact of the ultrarapid CYP2C19*17 allele on serum concentration of escitalopram in psychiatric patients. *Clin. Pharmacol. Ther.* **83**, 322-327, doi:10.1038/sj.clpt.6100291 (2008).
109 Sindrup, S. H. *et al.* Pharmacokinetics of citalopram in relation to the sparteine and the mephenytoin oxidation polymorphisms. *Ther. Drug Monit.* **15**, 11-17 (1993).
110 Kirchheiner, J. *et al.* Pharmacogenetics of antidepressants and antipsychotics: the contribution of allelic variations to the phenotype of drug response. *Mol. Psychiatry* **9**, 442-473, doi:10.1038/sj.mp.4001494 (2004).
111 Feero, W. G., Guttmacher, A. E. & Collins, F. S. Genomic medicine--an updated primer. *N. Engl. J. Med.* **362**, 2001-2011, doi:10.1056/NEJMra0907175 (2010).
112 Frueh, F. W. *et al.* Pharmacogenomic biomarker information in drug labels approved by the United States food and drug administration: prevalence of related drug use. *Pharmacotherapy* **28**, 992-998, doi:10.1592/phco.28.8.992 (2008).

113 Chute, C. G. & Kohane, I. S. Genomic medicine, health information technology, and patient care. *JAMA* **309**, 1467-1468, doi:10.1001/jama.2013.1414 (2013).
114 Statement of the American Society of Clinical Oncology: genetic testing for cancer susceptibility, Adopted on February 20, 1996. *J. Clin. Oncol.* **14**, 1730-1736; discussion 1737-1740 (1996).
115 Robson, M. E., Storm, C. D., Weitzel, J., Wollins, D. S. & Offit, K. American Society of Clinical Oncology policy statement update: genetic and genomic testing for cancer susceptibility. *J. Clin. Oncol.* **28**, 893-901, doi:10.1200/JCO.2009.27.0660 (2010).
116 Offit, K. BRCA mutation frequency and penetrance: new data, old debate. *J. Natl. Cancer Inst.* **98**, 1675-1677, doi:10.1093/jnci/djj500 (2006).
117 Clark, A. E. Sequence thyself: personalized medicine and therapies for the future: 2012 Yale Healthcare Conference. *Yale J. Biol. Med.* **85**, 421-424 (2012).
118 McDermott, U., Downing, J. R. & Stratton, M. R. Genomics and the continuum of cancer care. *N. Engl. J. Med.* **364**, 340-350, doi:10.1056/NEJMra0907178 (2011).
119 Mischel, P. S. *et al.* Identification of molecular subtypes of glioblastoma by gene expression profiling. *Oncogene* **22**, 2361-2373, doi:10.1038/sj.onc.1206344 (2003).
120 Perou, C. M. *et al.* Molecular portraits of human breast tumours. *Nature* **406**, 747-752, doi:10.1038/35021093 (2000).
121 Sotiriou, C. *et al.* Breast cancer classification and prognosis based on gene expression profiles from a population-based study. *Proc. Natl. Acad. Sci. U. S. A.* **100**, 10393-10398, doi:10.1073/pnas.1732912100 (2003).
122 Esteller, M. Epigenetics in cancer. *N. Engl. J. Med.* **358**, 1148-1159, doi:10.1056/NEJMra072067 (2008).
123 Feinberg, A. P. Phenotypic plasticity and the epigenetics of human disease. *Nature* **447**, 433-440, doi:10.1038/nature05919 (2007).
124 Calin, G. A. *et al.* Human microRNA genes are frequently located at fragile sites and genomic regions involved in cancers. *Proc. Natl. Acad. Sci. U. S. A.* **101**, 2999-3004, doi:10.1073/pnas.0307323101 (2004).
125 Sethi, S., Ali, S. & Sarkar, F. H. MicroRNAs in personalized cancer therapy. *Clin. Genet.*, doi:10.1111/cge.12362 (2014).
126 Jones, P. A. & Baylin, S. B. The epigenomics of cancer. *Cell* **128**, 683-692, doi:10.1016/j.cell.2007.01.029 (2007).
127 Esteller, M. *et al.* Promoter hypermethylation and BRCA1 inactivation in sporadic breast and ovarian tumors. *J. Natl. Cancer Inst.* **92**, 564-569 (2000).
128 Esteller, M. *et al.* Inactivation of the DNA-repair gene MGMT and the clinical response of gliomas to alkylating agents. *N. Engl. J. Med.* **343**, 1350-1354, doi:10.1056/NEJM200011093431901 (2000).
129 McCabe, M. T., Brandes, J. C. & Vertino, P. M. Cancer DNA methylation: molecular mechanisms and clinical implications. *Clin. Cancer Res.* **15**, 3927-3937, doi:10.1158/1078-0432.CCR-08-2784 (2009).
130 Fernandez, A. F. *et al.* A DNA methylation fingerprint of 1628 human samples. *Genome Res.* **22**, 407-419, doi:10.1101/gr.119867.110 (2012).
131 He, L. & Hannon, G. J. MicroRNAs: small RNAs with a big role in gene regulation. *Nat Rev Genet* **5**, 522-531, doi:10.1038/nrg1379 (2004).
132 Bartel, D. P. MicroRNAs: genomics, biogenesis, mechanism, and function. *Cell* **116**, 281-297 (2004).

133 Fabbri, M. MicroRNAs and cancer: towards a personalized medicine. *Curr Mol Med* **13**, 751-756 (2013).
134 Stefansson, O. A. & Esteller, M. Epigenetic modifications in breast cancer and their role in personalized medicine. *Am. J. Pathol.* **183**, 1052-1063, doi:10.1016/j.ajpath.2013.04.033 (2013).
135 Slamon, D. J. *et al.* Human breast cancer: correlation of relapse and survival with amplification of the HER-2/neu oncogene. *Science* **235**, 177-182 (1987).
136 Vogel, C. L. *et al.* Efficacy and safety of trastuzumab as a single agent in first-line treatment of HER2-overexpressing metastatic breast cancer. *J. Clin. Oncol.* **20**, 719-726 (2002).
137 Romond, E. H. *et al.* Trastuzumab plus adjuvant chemotherapy for operable HER2-positive breast cancer. *N. Engl. J. Med.* **353**, 1673-1684, doi:10.1056/NEJMoa052122 (2005).
138 Lynch, T. J. *et al.* Activating mutations in the epidermal growth factor receptor underlying responsiveness of non-small-cell lung cancer to gefitinib. *N. Engl. J. Med.* **350**, 2129-2139, doi:10.1056/NEJMoa040938 (2004).
139 Mok, T. S. *et al.* Gefitinib or carboplatin-paclitaxel in pulmonary adenocarcinoma. *N. Engl. J. Med.* **361**, 947-957, doi:10.1056/NEJMoa0810699 (2009).
140 van 't Veer, L. J. *et al.* Gene expression profiling predicts clinical outcome of breast cancer. *Nature* **415**, 530-536, doi:10.1038/415530a (2002).
141 van de Vijver, M. J. *et al.* A gene-expression signature as a predictor of survival in breast cancer. *N. Engl. J. Med.* **347**, 1999-2009, doi:10.1056/NEJMoa021967 (2002).
142 Slodkowska, E. A. & Ross, J. S. MammaPrint 70-gene signature: another milestone in personalized medical care for breast cancer patients. *Expert Rev Mol Diagn* **9**, 417-422, doi:10.1586/erm.09.32 (2009).
143 Wang, Y. *et al.* Gene-expression profiles to predict distant metastasis of lymph-node-negative primary breast cancer. *Lancet* **365**, 671-679, doi:10.1016/S0140-6736(05)17947-1 (2005).
144 Sotiriou, C. & Pusztai, L. Gene-expression signatures in breast cancer. *N. Engl. J. Med.* **360**, 790-800, doi:10.1056/NEJMra0801289 (2009).
145 Stankiewicz, P. & Lupski, J. R. Structural variation in the human genome and its role in disease. *Annu. Rev. Med.* **61**, 437-455, doi:10.1146/annurev-med-100708-204735 (2010).
146 Liu, L. *et al.* Comparison of next-generation sequencing systems. *J Biomed Biotechnol* **2012**, 251364, doi:10.1155/2012/251364 (2012).
147 Medvedev, P., Stanciu, M. & Brudno, M. Computational methods for discovering structural variation with next-generation sequencing. *Nat Methods* **6**, S13-20, doi:10.1038/nmeth.1374 (2009).
148 Mardis, E. R. The $1,000 genome, the $100,000 analysis? *Genome medicine* **2**, 84, doi:10.1186/gm205 (2010).
149 Kohane, I. S., Masys, D. R. & Altman, R. B. The incidentalome: a threat to genomic medicine. *JAMA* **296**, 212-215, doi:10.1001/jama.296.2.212 (2006).
150 Solomon, B. D. Incidentalomas in genomics and radiology. *N. Engl. J. Med.* **370**, 988-990, doi:10.1056/NEJMp1310471 (2014).
151 Ioannidis, J. P. & Khoury, M. J. Improving validation practices in "omics" research. *Science* **334**, 1230-1232, doi:10.1126/science.1211811 (2011).

152 Goozner, M. Duke scandal highlights need for genomics research criteria. *J. Natl. Cancer Inst.* **103**, 916-917, doi:10.1093/jnci/djr231 (2011).

153 Bloss, C. S., Darst, B. F., Topol, E. J. & Schork, N. J. Direct-to-consumer personalized genomic testing. *Hum. Mol. Genet.* **20**, R132-141, doi:10.1093/hmg/ddr349 (2011).

154 Bloss, C. S., Wineinger, N. E., Darst, B. F., Schork, N. J. & Topol, E. J. Impact of direct-to-consumer genomic testing at long term follow-up. *J. Med. Genet.* **50**, 393-400, doi:10.1136/jmedgenet-2012-101207 (2013).

155 Ng, P. C., Murray, S. S., Levy, S. & Venter, J. C. An agenda for personalized medicine. *Nature* **461**, 724-726, doi:10.1038/461724a (2009).

156 Fleming, N. *Rival genetic tests leaves buyers confused*, 2008).

157 Janssens, A. C. *et al.* A critical appraisal of the scientific basis of commercial genomic profiles used to assess health risks and personalize health interventions. *Am J Hum Genet* **82**, 593-599, doi:10.1016/j.ajhg.2007.12.020 (2008).

158 Downing, N. S. & Ross, J. S. Innovation, risk, and patient empowerment: the FDA-mandated withdrawal of 23andMe's Personal Genome Service. *JAMA* **311**, 793-794, doi:10.1001/jama.2014.148 (2014).

159 Marshall, E. Human genome 10th anniversary. Waiting for the revolution. *Science* **331**, 526-529, doi:10.1126/science.331.6017.526 (2011).

160 Meric-Bernstam, F., Farhangfar, C., Mendelsohn, J. & Mills, G. B. Building a personalized medicine infrastructure at a major cancer center. *J. Clin. Oncol.* **31**, 1849-1857, doi:10.1200/JCO.2012.45.3043 (2013).

161 Pandey, A. A piece of my mind. Preparing for the 21st-century patient. *JAMA* **309**, 1471-1472, doi:10.1001/jama.2012.116971 (2013).

162 Najafzadeh, M., Davis, J. C., Joshi, P. & Marra, C. Barriers for integrating personalized medicine into clinical practice: a qualitative analysis. *Am J Med Genet A* **161**, 758-763, doi:10.1002/ajmg.a.35811 (2013).

163 McGuire, A. L., Diaz, C. M., Wang, T. & Hilsenbeck, S. G. Social networkers' attitudes toward direct-to-consumer personal genome testing. *Am J Bioeth* **9**, 3-10, doi:10.1080/15265160902928209 (2009).

164 Lo, S. S. *et al.* Prospective multicenter study of the impact of the 21-gene recurrence score assay on medical oncologist and patient adjuvant breast cancer treatment selection. *J. Clin. Oncol.* **28**, 1671-1676, doi:10.1200/JCO.2008.20.2119 (2010).

165 Gura, T. Rare diseases: Genomics, plain and simple. *Nature* **483**, 20-22, doi:10.1038/483020a (2012).

166 Chuang, C. L., DT. The pros and cons of gene patents. *Publications* (2010).

167 Maxmen, A. *Personalized medicine enters a new era*, <http://www.pbs.org/wgbh/nova/next/body/gene-patents-and-personalized-medicine/> (2013).

168 Hudson, K. L., Holohan, M. K. & Collins, F. S. Keeping pace with the times--the Genetic Information Nondiscrimination Act of 2008. *N. Engl. J. Med.* **358**, 2661-2663, doi:10.1056/NEJMp0803964 (2008).

169 Bragazzi, N. L. Children, adolescents, and young adults participatory medicine: involving them in the health care process as a strategy for facing the infertility issue. *Am J Bioeth* **13**, 43-44, doi:10.1080/15265161.2012.760674 (2013).

170 Hiraki, S. & Green, N. S. Newborn screening for treatable genetic conditions: past, present and future. *Obstet. Gynecol. Clin. North Am.* **37**, 11-21, doi:10.1016/j.ogc.2010.01.002 (2010).
171 Goldenberg, A. J. & Sharp, R. R. The ethical hazards and programmatic challenges of genomic newborn screening. *JAMA* **307**, 461-462, doi:10.1001/jama.2012.68 (2012).
172 Collins, F. S. & Hamburg, M. A. First FDA authorization for next-generation sequencer. *N. Engl. J. Med.* **369**, 2369-2371, doi:10.1056/NEJMp1314561 (2013).
173 Green, R. C. *et al.* ACMG recommendations for reporting of incidental findings in clinical exome and genome sequencing. *Genet Med* **15**, 565-574, doi:10.1038/gim.2013.73 (2013).
174 Kocarnik, J. M. & Fullerton, S. M. Returning pleiotropic results from genetic testing to patients and research participants. *JAMA* **311**, 795-796, doi:10.1001/jama.2014.369 (2014).
175 Hood, L. & Friend, S. H. Predictive, personalized, preventive, participatory (P4) cancer medicine. *Nat Rev Clin Oncol* **8**, 184-187, doi:10.1038/nrclinonc.2010.227 (2011).
176 Weston, A. D. & Hood, L. Systems biology, proteomics, and the future of health care: toward predictive, preventative, and personalized medicine. *J Proteome Res* **3**, 179-196 (2004).
177 Tian, Q., Price, N. D. & Hood, L. Systems cancer medicine: towards realization of predictive, preventive, personalized and participatory (P4) medicine. *J. Intern. Med.* **271**, 111-121, doi:10.1111/j.1365-2796.2011.02498.x (2012).
178 Khoury, M. J., Gwinn, M. L., Glasgow, R. E. & Kramer, B. S. A population approach to precision medicine. *Am. J. Prev. Med.* **42**, 639-645, doi:10.1016/j.amepre.2012.02.012 (2012).
179 Kindig, D. & Stoddart, G. What is population health? *Am. J. Public Health* **93**, 380-383 (2003).
180 Subramanian, J. & Simon, R. Gene expression-based prognostic signatures in lung cancer: ready for clinical use? *J. Natl. Cancer Inst.* **102**, 464-474, doi:10.1093/jnci/djq025 (2010).
181 Gorini, A. & Pravettoni, G. P5 medicine: a plus for a personalized approach to oncology. *Nat Rev Clin Oncol* **8**, 444, doi:10.1038/nrclinonc.2010.227-c1 (2011).
182 Pravettoni, G. & Gorini, A. A P5 cancer medicine approach: why personalized medicine cannot ignore psychology. *J. Eval. Clin. Pract.* **17**, 594-596, doi:10.1111/j.1365-2753.2011.01709.x (2011).
183 Bragazzi, N. L. From P0 to P6 medicine, a model of highly participatory, narrative, interactive, and "augmented" medicine: some considerations on Salvatore Iaconesi's clinical story. *Patient Prefer Adherence* **7**, 353-359, doi:10.2147/PPA.S38578 (2013).
184 Hood, L. & Flores, M. A personal view on systems medicine and the emergence of proactive P4 medicine: predictive, preventive, personalized and participatory. *N Biotechnol* **29**, 613-624, doi:10.1016/j.nbt.2012.03.004 (2012).
185 Qin, S. *et al.* SRM targeted proteomics in search for biomarkers of HCV-induced progression of fibrosis to cirrhosis in HALT-C patients. *Proteomics* **12**, 1244-1252, doi:10.1002/pmic.201100601 (2012).
186 Varmus, H. Ten years on--the human genome and medicine. *N. Engl. J. Med.* **362**, 2028-2029, doi:10.1056/NEJMe0911933 (2010).

187 Snyder, M., Weissman, S. & Gerstein, M. Personal phenotypes to go with personal genomes. *Mol Syst Biol* **5**, 273, doi:10.1038/msb.2009.32 (2009).
188 Phillips, K. A. Closing the evidence gap in the use of emerging testing technologies in clinical practice. *JAMA* **300**, 2542-2544, doi:10.1001/jama.2008.754 (2008).
189 Lyssenko, V. *et al.* Clinical risk factors, DNA variants, and the development of type 2 diabetes. *N. Engl. J. Med.* **359**, 2220-2232, doi:10.1056/NEJMoa0801869 (2008).