Genetic Studies of Elite Athlete Status

by

Guan Wang, M.Res.

A Doctoral Thesis

Submitted in fulfilment of the requirements for the Degree of

Doctor of Philosophy

August 2013

Institute of Cardiovascular and Medical Sciences
College of Medical, Veterinary, and Life Sciences
University of Glasgow

Author's Declaration

I declare that this thesis is the result of my own work, except that subject recruitment in each country was orchestrated by the relevant PIs as follows: Dr. Yannis Pitsiladis (Jamaican sprint cohort), Dr. Krista Austin (USA sprint cohort), Dr. Rob Lee (Caucasian swim cohort), Dr. Noriyuki Fuku (Japanese swim and track-and-field athlete cohorts), and Dr. Sandy Hsieh (Taiwanese swim cohort); genotyping of Japanese and Taiwanese swim cohort was conducted by Dr. Eri Mikami and Dr. Li-Ling Chiu; genome-wide genotyping of Jamaican and USA sprint cohorts as well as Japanese track-and-field athlete cohort was overseen by Dr. Noriyuki Fuku and carried out in Japan; raw genotype data of additional 350 African-American controls was received from Dr. Braxton Mitchell at University of Maryland, USA. This work has not been submitted previously for any other degree at the University of Glasgow or any other institution.

Guan Wang

Acknowledgement

I would like to express my sincere thanks to my supervisors Dr. Yannis Pitsiladis and Dr. Sandosh Padmanabhan for their continuous support and guidance. Dr. Pitsiladis inspired me on the path to the completion of my PhD study with his enthusiasm to the field, guided and reminded me to reflect on my weaknesses, and provided me opportunities to be involved in diverse projects to improve my knowledge and skills in the field I am interested in. Dr. Padmanabhan provided me with his immense knowledge and persistent help on the GWAS project. Without the training from Dr. Padmanabhan on GWAS analysis, the completion of my PhD and this thesis would have not been possible.

Also, I am truly thankful to Dr. Mark Bailey for his patience and time in helping me understand the essence of genetics, and his insightful comments and questions kept me thinking about the right way through a genetic research. Besides, I would also like to send "many thanks" to Dr. Claire Hastie for her generous help during the initial setting up of the GWAS analysis.

Now, it is a great pleasure to thank everyone who provided their help and kindness to assist me at certain stage of my PhD: Dr. Noriyuki Fuku, Dr. Alessandra de Perini, Dr. Eri Mikami, Dr. Wai Kwong Lee, Dr. Jim McCulloch, Ms. Morvern Campbell, Mr. Michael Deason, Dr. Carlos Celis, Dr. Anna Christina Koni, Mr. David Hughes, Mrs. Anne Keenan. I would also like to thank the Great Britain Sasakawa Foundation for providing an award to me for visiting the lab in Japan on genome-wide genotyping technology.

Finally, it is the "thanks" that beyond words that can adequately convey. These are to my family, my parents and uncle, for the roles they play in my life and their encouragement to

me at every step, and to my boyfriend, his sense of humour, tolerance and kindness helped me through every difficult moment over the past few years.

Table of Contents

List of Figures

List of Tables

List of Abbreviations

AA	African-American
ACE	angiotensin converting enzyme
ACTN3	α-actinin-3
ACVR1B	activin receptor 1B
ADAMTS18	ADAM metallopeptidase with thrombospondin type 1 motif, 18
AGT	angiotensinogen (serpin peptidase inhibitor, clade A, member 8)
AMD	age-related macular degeneration
AUC	area under the curve
BDKRB2	bradykinin receptor B2
BMD	bone mineral density
BP	base pair
CA	Chinese ancestry
CaTS	Power Calculator for Two Stage Association Studies
CCNY	cyclin Y
CDCV	common disease-common variant
CEU	Caucasians in Utah, USA
CHB	Han Chinese in Beijing, China
Chr	chromosome
CI	confidence interval
CNTF	ciliary neurotrophic factor
CNV	copy number variation
CFH	Complement Factor H
CREM	cAMP responsive element modulator
CSF2	colony stimulating factor 2 (granulocyte-macrophage)

CUL2	cullin 2
D	the linkage disequilibrium determinant
D'	unit of measurement of linkage disequilibrium
dbSNP	The Single Nucleotide Polymorphism database
DCTN4	dynactin 4
df	degree of freedom
DIO1	deiodinase, iodothyronine, type I
DNA	deoxyribonucleic acid
DRD2	dopamine D2 receptor
DREW	Dose Response to Exercise
EA	European ancestry
eQTL	expresson quantitative trait locus
FHS	Framingham Heart Study
FTO	fat mass and obesity associated
FZD8	frizzled family receptor 8
GALNT13	UDP-N-acetyl-alpha-D-galactosamine:polypeptide N-acetylgalactosaminyltransferase 13
GC	genomic control
Genevar	GENe Expression VARiation
GJD4	gap junction protein, delta 4, 40.1kDa
GLUT4	glucose transporter type 4
GRC	Genome Reference Consortium
GTEx	Genotype-Tissue Expression
GWAS	Genome-wide association study
h^2	narrow-sense heritability
HERITAGE	HEalth, RIsk factors, exercise Training And GEnetics

HLA	Human Leukocyte Antigen System
HWE	Hardy-Weinberg Equilibrium
I^2	I^2 heterogeneity index
IBD	identity by descent
IBS	identity by state
IL15RA	interleukin 15 receptor, alpha
IL6	interleukin 6 (interferon, beta 2)
JAM	Jamaicans
JAP	Japanese
JPT	Japanese in Tokyo, Japan
Kb	Kilobase
KO	knockout
λ	genomic inflation factor
LD	linkage disequilibrium
LD	long distance (Chapter 3)
LR	logistic regression
MAF	minor allele frequency
Mb	Megabase
MCT1	monocarboxylate transporter 1
MD	middle distance
MGB	minor groove binders
MIM	Mendelian Inheritance in Man
miR	microRNA
MIR4683	microRNA 4683
miRNA	micro ribonucleic acid
mRNA	messenger ribonucleic acid

MSTN	myostatin
NCBI	National Center for Biotechnology Information
NFQ	nonfluorescent quencher
NOS3	nitric oxide synthase 3 (endothelial cell)
NR	not reported
NRXN3	neurexin 3
OMIM	Online Medelian Inheritance in Man
OR	odds ratio
PARD3	par-3 partitioning defective 3 homolog (C. elegans)
PC	principle component
PCA	principle component analysis
PCR	polymerase chain reaction
PheGenl	Phenotype-Genotype Integrator
PPARGC	peroxisome proliferator-activated receptor gamma, coactivator 1 alpha
PT	permutation test
Q	*p*-value for Cochrane's Q statistic
QC	quality control
r^2	unit of measurement of linkage disequilibrium
RAS	renin-angiotensin system
RNA	ribonucleic acid
ROC	receiver operating characteristic
RPRM	reprimo
SCAN	SNP and CNV Annotation database
SD	short distance (Chapter 3)
SE	standard error

SMD	short and middle distance
SNP	single nucleotide polymorphism
STRRIDE	Studies of a Targeted Risk Reduction Intervention Through Defined Exercise
T2DM	Type 2 Diabetes Mellitus
TDT	transmission-disequilibrium test
TGS	total genotype score
TNF	tumor necrosis factor
TRHR	thyrotropin-releasing hormone receptor
TSI	Tuscans in Italy
UTR	untranslated region
VEGFR	vascular endothelial growth factor receptor
$\dot{V}O_{2max}$	maximal oxygen uptake
WTCCC	Wellcome Trust Case Control Consortium
YRI	Yoruba in Ibadan, Nigeria
χ^2	the chi-squared test statistic
1kGP	1,000 Genomes Project

Publications, Awards and Presentations

Publications:

Pitsiladis Y, **Wang G**, Wolfarth B, Scott R, Fuku N, Mikami E, He Z, Fiuza-Luces C, Eynon N, Lucia A. Genomics of elite sporting performance: what little we know and necessary advances. *British Journal of Sports Medicine,* 2013; 47(9):550-555.

Wang G, Mikami E, Chiu LL, de Perini A., Deason M, Fuku N, Miyachi M, Kaneoka K, Murakami H, Tanaka M, Hsieh LL, Hsieh SS, Caporossi D, Pigozzi F, Hilley A, Lee R, Galloway SD, Gulbin J, Rogozkin VA, Ahmetov II, Yang N, North KN, Ploutarhos S, Montgomery HE, Bailey ME, Pitsiladis YP. Association analysis of ACE and ACTN3 in Elite Caucasian and East Asian Swimmers. *Medicine and Science in Sports and Exercise,* 2013; 45(5):892-900.

Pitsiladis YP, and **Wang G.** Necessary advances in exercise genomics and likely pitfalls. *Journal of Applied Physiology,* 2011; 110(5): 1150-1151.

Pitsiladis Y, **Wang G**, Wolfarth B (Invited contributors, 2011). Genomics of aerobic capacity and endurance performance. Molecular and Translational Medicine Series, Volume: Exercise Genomics, Edited by Linda S. Pescatello and Stephen M. Roth, 179-229. New York, NY: Springer.

Koni AC, Scott RA, **Wang G**, Bailey ME, Peplies J, Pitsiladis YP on behalf of the IDEFICS Consortium. DNA yield and quality of saliva samples and suitability for large scale epidemiological studies in children. *International Journal of Obesity,* 2011; 35: S113-S118.

Lagou V, Scott RA, Manios Y, Chen TL, **Wang G**, Grammatikaki E, Kortsalioudaki C, Liarigkovinos T, Moschonis G, Roma-Giannikou E, Pitsiladis YP. Impact of perxisome proliferator-activated receptor gamma and delta on adiposity in preschoolers. *Obesity*, 2008; 16(4):913-918.

Awards:

Great Britain SASAKAWA Foundation Award for a PhD research trip to Japan on genotyping technology in 2011.

Oral Presentations:

Wang G, Mikami E, de Perini A., Deason M, Fuku N, Miyachi M, Kaneoka K, Murakami H, Tanaka M, Caporossi D, Pigozzi F, Hilley A, Lee R, Galloway SD, Gulbin J, Rogozkin VA, Ahmetov II, Yang N, North KN, Ploutarhos S, Montgomery HE, Bailey ME, Pitsiladis YP. ACTN3 and ACE Genotypes in Elite Caucasian and Japanese Swimmers. International Convention on Science, Education and Medicine in Sport, Glasgow, UK, 19/07/12.

Wang G, de Perini A, Caporossi D, Pigozzi F, Bailey ME, Deason M, Hilley A, Lee R, Ahmetov II, Rogozkin VA, Galloway S, Gulbin J, Yang N, North K, Montgomery HE, Pitsiladis YP. ACTN3 and ACE Genotypes in Elite Caucasian Swimmers. The American College of Sports Medicine Annual Meeting, Denver, Colorado, USA, 31/05/11.

Posters and Abstracts:

Wang G, Fuku N, Padmanabhan P, Mikami E, Tanaka M, Miyachi M, Murakami H, Morrison E, Pitsiladis Y. Genome-wide association study of world class athlete status

using elite Jamaican sprinters. The Malawi-Liverpool-Wellcome Trust Clinical Research Programme, Glasgow, UK, 21/03/13.

Wang G, Fuku N, Padmanabhan P, Mikami E, Tanaka M, Miyachi M, Murakami H, Morrison E, Pitsiladis Y. Genome-Wide Association Study (GWAS) of elite sprinters of West African Ancestry. XXXII World Congress of Sports Medicine, Rome, Italy, 27/09/12.

Wang G, Fuku N, Padmanabhan P, Mikami E, Tanaka M, Miyachi M, Murakami H, Morrison E, Pitsiladis Y. Genome-wide association study approach in elite Jamaican athletes to identify common genetic variations associated with world class athlete status. Glasgow Polyomics Launch Symposium, Glasgow, UK, 14/09/12.

Wang G, Fuku N, Padmanabhan P, Mikami E, Tanaka M, Miyachi M, Murakami H, Morrison E, Pitsiladis Y. A genome wide association study of elite Jamaican athletes identifies multiple putative single nucleotide polymorphisms associated with world class athlete status. The West of Scotland Health and Ethnicity Network conference, Glasgow, UK, 20/08/12.

Deason ML, **Wang G,** Fuku F, Mikami E, Scott RA, Irwin L, Irving RR, Charlton V, Morrison E, Austin AK, Tladi D, Headley SA, Kolkhhorst FW, Yamada Y, Tanaka M, Pitsiladis YP. Genetic Ancestry and Elite Sprinting in groups of West African descent. International Convention on Science, Education and Medicine in Sport, Glasgow, UK, 19/07/12.

Abstract

In the past decade, limited progress has been made in identifying genetic associations with performance and health-related fitness phenotypes due to the use primarily of the traditional candidate-gene approach involving small sample sizes and few coordinated research efforts. Much of the genetic data relating to human performance has been generated while exploring the aetiology of lifestyle-related disorders such as obesity and type 2 diabetes mellitus (T2DM). As of 2008, over 200 autosomal gene entries and quantitative trait loci have been reported to be significantly associated with performance and health-related fitness. However, most genetic findings to date have been inconclusive due to studies employing relatively small sample sizes and predominantly single-gene approaches which are especially prone to type I errors. It is widely accepted that there will be many genes involved in sporting performance and health-related fitness phenotypes, and hence it is timely that genetic research has moved to the genomics era with the use of a genome-wide approach (e.g. genotyping a large number of variants simultaneously across the entire human genome) in a well-phenotyped, large cohort. This thesis summarizes the recent findings of genetic predisposition to elite human performance by using the conventional candidate-gene approach as well as the unbiased genome-wide approach (i.e. genome-wide association studies, GWASs).

The current candidate gene study focused on investigating whether polymorphisms in the angiotensin-converting enzyme (*ACE*) and α-actinin-3 (*ACTN3*) genes are associated with elite swimmer status (stratified by swimming distance) in Caucasians and East Asians. *ACE* I/D and *ACTN3* p.R577X polymorphisms were genotyped for 200 elite Caucasian swimmers (short and middle distance, ≤ 400 m, n = 130; long distance, > 400 m, n = 70) and 326 elite Japanese and Taiwanese swimmers (short distance, ≤ 100 m, n = 166; middle

distance, 200–400 m, n = 160). Logistic regression and multiple-testing adjustment were applied to test for these genetic associations. *ACE* I/D was found to be associated with swimmer status in Caucasians, with the D allele being overrepresented in short-and-middle-distance swimmers with the largest effect being observed for the I-allele-dominant model (odds ratio = 1.90; logistic regression $p = 0.001$; permutation test $p = 0.0005$). In East Asians, however, the I allele was overrepresented in the short-distance swimmer group under the D-allele-dominant model (odds ratio = 1.52; logistic regression $p = 0.012$; permutation test $p = 0.0098$). The *ACE* I/D association findings in the elite swimmer cohorts showed that different risk alleles responsible for the associations were observed in swimmers of different ethnicities. *ACTN3* p.R577X was not statistically significantly associated with swimmer status in either Caucasian or East Asian population. The lack of associations between the functional *ACTN3* p.R577X polymorphism and elite swimmer status in both cohorts were in contrast to many associations with power-/sprint-performance in other sports previously reported. Since current sample size is relatively modest, larger studies will be required to further confirm these results, which, however, have highlighted that it is probable that the genes studied here are not the resulting variants responsible for the phenotypes of interest, despite the associations reported by previous candidate-gene studies in other sports.

The present GWAS were conducted in an attempt to identify common polymorphisms associated with elite sprint and endurance status in Jamaicans, African-Americans and Japanese, respectively. These unique athlete cohorts comprised of athletes of the highest standard including world record holders, world champions, Olympians and winners of other international events. Following exclusion of individuals and markers failing the quality control filters, 609,801 autosomal SNPs in 88 Jamaican sprint athletes and 87 Jamaican controls, 637,991 autosomal SNPs in 79 African-American sprint athletes and 391 African-American controls, and 541,179 autosomal SNPs in 114 Japanese athletes

(including 60 endurance and 54 sprint athletes) and 116 Japanese controls, were available for association analyses. 17, 7, 36 and 21 SNPs were associated with elite athlete status at a $p < 5 \times 10^{-5}$ threshold of significance in elite Jamaican sprint, African-American sprint, Japanese sprint and Japanese endurance GWAS sets, respectively. Meta-analyses were performed for SNPs with unadjusted association $p < 5 \times 10^{-5}$ across the sprint GWAS sample sets (i.e. Jamaican sprint, African-American sprint, Japanese sprint GWAS cohorts), using the fixed-effects model. The top 17 SNPs (unadjusted $p < 5 \times 10^{-5}$) from the Jamaican sprint cohort were extracted from the association results of African-American sprint, Japanese sprint cohorts, respectively, for the combined effects to be calculated using a meta-analysis method. The same procedure was also applied to the top hits in African-American and Japanese cohorts. The combined odds ratio for the top meta-analysis hit (rs10196189) was 2.61 ($p = 4.66 \times 10^{-7}$) with the allele G associated with elite sprint status in Jamaicans, African-Americans and Japanese. Although meta-analysis has increased the sample size and power to detect associations in the current GWAS, independent replication of these associations followed by functional studies of replicated SNPs are required.

The results of the association studies presented here are the very first positive findings from GWAS involving world-class athletes and these encouraging findings provide further evidence of the importance of genetic predisposition to elite human performance. GWAS of athletes of the highest performance caliber as well as the application of meta-analysis across several initial GWASs seemed to help to circumvent the need for very large cohort of elite athletes and increase the study power. Nevertheless, future GWAS involving large well-funded collaborations using larger cohorts of elite athletes will be necessary in order to explore further the genetic architecture underlying elite human performance. Such initiatives may also allow gene x gene and gene x environment interactions to be explored to some extent, as well as the predictive utility of this genomic research.

1 Introduction

1.1 Elite human performance

The success of Jamaican and USA athletes (i.e. African-Americans) in sprint events is phenomenal. In the recent London Olympics, Jamaican and African-American sprinters won all medals in the men's and women's 100 and 200 m events. Moreover, USA swimmers (i.e. European-Americans) or other swimmers of Caucasian background took the medals in the men's 50 m and 100 m and in the women's 100 m, 200 m and 400 m swimming events. Asians are also good at swimming, dominating the men's 400 m and the women's 400 m individual medley. They were so successful compared to most of their competitors, why is this? These successes have now been considered as a result of a combined action of multiple genes, socio-culture, training, diet as well as other environmental factors (1).

Numerous studies have been conducted to identify the responsible genetic variants in relation to elite human performance. Historically, heritability estimate (h^2) was used to indirectly assess the genetic basis of human performance (2-4). Direct approach that tests a genotype-phenotype association by genotyping a well-phenotyped population has now been widely used. Nevertheless, genes related to increased athletic performance have not yet been fully identified, and the vast majority of the genes/polymorphisms that have been found have not been replicated due to various reasons. Several animal studies (5-7) exploring the mechanisms of gene-performance associations, however, have been successfully established and described below.

Mosher et al. (2007) (5) demonstrated a favourable effect of a mutation (2bp deletion in exon 3) in the myostatin (*MSTN*) gene on enhanced athletic performance in the whippet.

This gene has been mapped to chromosome two in humans. The heterozygote whippets (i.e. having one copy of the mutated allele) showed greater muscle growth compared to the wild type and advantages in racing performance. This is likely to be caused by the reduction of the myostatin protein in the *MSTN* heterozygote, resulting in increased muscle mass. It has also been reported that a higher proportion of fast-twitch muscle fibres (generating power more efficiently) were observed in *Mstn* knockout (KO) mice (6). All of these supported a role of *MSTN* gene in regulating muscle differentiation and growth, hence the potential to increase individual athletic ability in humans. In addition, *MSTN* mutation was found to be associated with muscle hypertrophy in a child, whose mother was a former professional athlete, healthy and has one copy of the mutated allele (8). No health problems were reported for the child at the time of the study (8).

Another example of a well-known KO mice model was used to investigate the mechanisms underlying α-actinin-3 deficiency to athletic performance (7). The authors found that the KO mice had the reduced muscle mass (i.e. due to reduced fibre diameter of the fast-twitch muscle), increased activity of aerobic enzymes, longer muscle contracting time and shorter recovery period from fatigue, which were attributed to the characteristics of the slow-twitch fibres. This KO mice model supported the idea of increased endurance but reduced muscle strength in homozygous carriers of the *ACTN3* null allele (complete deficiency of ACTN3).

To date, limited progress has been made in elucidating the genetic effects on human athletic performance. This discovery process is summarized in following sections in detail.

1.2 Genetic variations in elite human performance

1.2.1 Family-based studies: genetic evidence

Genetic variants related to phenotypic outcomes have been traditionally assessed on sets of related individuals by using twin studies, linkage analysis or other family-based designs, providing an indirect measure of the relationship between genes and the traits. Aggregation studies are generally the first step to study whether there is a genetic component to the trait, typically, in families with affected individuals as opposed to the general population. Heritability estimates the degree to which genes contribute to the phenotype variability, whereas the mode of inheritance (e.g. dominant or recessive) can be tested in pedigrees with known affection status using segregation analysis. Linkage analysis can be used to locate the genes by assessing pedigrees and genotype information for relevant markers, which tend to be linked and inherited together for the presence of a phenotype in related individuals.

Family-based studies support a genetic basis both for continuously measured performance-related traits (e.g. maximal oxygen uptake, $\dot{V}O_{2max}$) and athletic ability itself. Hereditary influence on endurance and muscle strength include studies (reviewed in (9)) of $\dot{V}O_{2max}$ and $\dot{V}O_{2max}$ trainability (h^2 = ~23% - 71%), cardiac mass, structure and function (h^2 = ~6% - 93%), pulmonary function (h^2 = ~28% - 77%), muscle strength and power (h^2 = ~30% - 83%), lactic acid concentrations (h^2 = ~76% - 93%), muscle fibre distributions (h^2 = ~25% - 99.5%, (4,10,11)), and performance time for a 1000 m run (h^2 = 98% and 69% for monozygotic and dizygotic twins respectively, (12)). The widely differed genetic contribution estimation from literature is evident, and has received considerable criticism, including small number of twin pairs in some of the studies, different phenotype evaluation process, and uneven controlled environmental factors (1,9). The largest twin study so far in exercise related traits, comprised of 37,051 twin pairs from seven countries: Australia, Denmark, Finland, the Netherlands, Norway, Sweden, and United Kingdom, suggested an additive effect of the genetic variants contribute significantly to exercise participation (h^2 =

62%) (13) and emphasised the importance of genetic variation in explaining individual differences in exercise behaviour. The evidence for a genetic component to human performance, as revealed by familial aggregation and heritability studies, is incontrovertible; however, questions such as putative genes & magnitude, gene locations and allele frequencies remain unanswered. As briefly mentioned above, a step further would be to map the genes on the chromosomes through linkage analysis.

A small number of linkage based genome-wide scans have been used to identify chromosomal regions associated with human performance phenotypes (14). Initial linkage-based reports were mainly generated using the cohort from the HEalth, RIsk factors, exercise Training And Genetics (HERITAGE) family study (14), which is a large family intervention study involving 742 participants from approx. 300 families who were subjected to a 20-wk controlled endurance training sessions to assess the genetic effects on cardiovascular, metabolic and hormonal responses to regular exercise (http://www.pbrc.edu/heritage/home.htm; (15)). These studies revealed a number of genetic regions associated with $\dot{V}O_{2max}$ and its response to training (16-19), maximal power output (18,19), exercise blood pressure (20), stroke volume (21,22) and cardiac output (22), sub-maximal exercise heart rate (23) and changes in body composition (24-27), glucose and insulin metabolism-related phenotypes (28,29) in response to training. Other studies have reported genomic regions related to physical activity levels (30-33) and muscle strength related-traits (34,35). Notably, fine mapping and follow-up replication studies of a previously identified linkage peak for knee strength on chromosome 12q12-14 identified activin receptor 1B (*ACVR1B*) rs2854464 AA genotype associated with increased muscle strength (36). Moreover, rs2854464 locates in miR-24 binding site, but no change of mRNA expression level in quadriceps muscle was observed with this genotypic variation (36).

A main advantage of family studies lies in its robustness to population stratification, but the resulting genes may be localized in a broad genetic interval since the DNA section is usually delimited by a crossover between a nearby marker and the actual causal locus in linkage analysis (37). While linkage studies have revealed some interesting genomic regions that may harbour genes contributing to human physical performance traits as described above, linkage analysis has been much less successful in detecting common genetic variations contributing to complex traits comparing with its success in identifying genes related to human monogenic traits (38). Recent advances in molecular technologies allow dense of markers to be genotyped simultaneously. This has accelerated the move from family studies to population-based case-control association studies, which are expected to have greater power in identifying common genetic variants through studying unrelated individuals (affected cases vs. healthy controls) from a population.

1.2.2 Association studies

Association studies are usually conducted in unrelated case-control samples by comparing the allele frequencies of a single marker or a set of markers in candidate regions even spanning the human genome. Single nucleotide polymorphism (SNP) is the most common form of DNA sequence variation, where a single nucleotide A, T, G or C is replaced by another, and this occurs more frequently in non-coding regions of the genome (39). SNP appears once in about every 290 nucleotides and approx. 11 million SNPs present in roughly 3.2 billion DNA base pairs across the entire human genome (39). Most of the differences exist among individuals, owing largely to the substitutions at a SNP locus (39). The least frequent allele of a SNP needs to be above 1% in a population to be effectively assessed by association studies.

There are two types of associations: direct and indirect (Figure 1.1 (40)). The direct method focuses on causal polymorphism of a phenotype; such studies can be referred to as

candidate polymorphism studies. The identification of the casual polymorphism is challenging, because many of these causal variants for common and complex traits may be non-coding and yet insufficient and unclear information on the causation of the complex traits for such polymorphisms to be accurately identified. The indirect candidate gene association studies require prior knowledge of known function of the candidate regions involving a number of SNPs, which may be the causal variants themselves or in linkage disequilibrium (LD) with the causal polymorphisms. LD refers to the non-random correlation between alleles at two loci. LD may rise from linkage, but it is also possible that two alleles physically unlinked are associated due to e.g. non-random mating or selection (41). *D* is one of the first used methods to predict the extent of LD and it depends on allele frequencies (42). In a simplified manner, for a two-locus haplotype, *D* refers to the difference between the observed haplotype frequency at two loci and the expected frequency when they segregate randomly. Other common measurements of LD include *D'* and r^2. $D' = 1$ indicates complete LD (no historical recombination), but $D' < 1$ cannot be meaningfully interpreted because *D'* will be inflated in small samples with high *D'* values possibly obtained from markers in linkage equilibrium (37). r^2 method is defined as D^2 divided by four allele frequencies at the two loci, taking into account the allele frequency differences (37). The value of r^2 varies from 0 to 1 and is in inverse proportion to sample size. $r^2 > 0.3$ might be taken as the minimum value for useful mapping (37). Patterns of LD can be influenced by genetic drift, population growth, admixture, population structure, natural selection, gene conversion, and mutation and recombination rates (37). GWASs are indirect association studies at the genome scale. In a GWAS, hundreds of thousands of tagSNPs (served as proxies for real causal variants based on level of LD) across the entire genome will be interrogated simultaneously, and this is thought to be more effective to discover common genetic variants underlying complex traits.

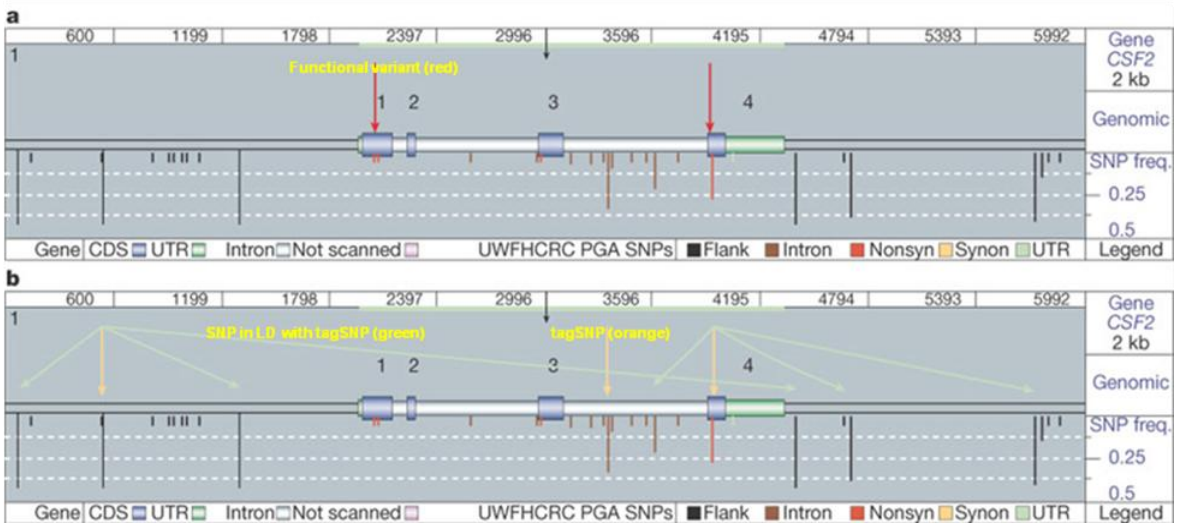


Figure 1.1 Using SNPs image of the CSF2 gene to elucidate the direct and indirect associations. a. Direct association study, in which two functional variants (red arrows) are genotyped directly for association analysis. b. Indirect association study, in which only a subset of tagSNPs (orange arrows) is typed, other SNPs in LD (green arrows) with those tagSNPs can therefore be indirectly detected for association analysis. (adapted from (40)).

1.2.3 Candidate gene association studies

Since 2000, a group of researchers devoted themselves to the annual update of the human gene map for fitness and performance-related phenotypes (43-48). The last version of the gene map (the 2006-2007 update) covered over 200 genes in both the nuclear and mitochondrial genomes reported to influence physical performance and health-related fitness (48). The weakness in these yearly updates was that the genomic entries included in the map were from a mix of good and weak studies. The authors have shifted the focus to only publications with the strongest evidences in the field of exercise genomics for drafting subsequent reviews (49-52). The candidate gene associations with elite human performance have been inconclusive to date. Genes or loci related to aerobic capacity, endurance performance and muscle metabolism with at least one positive study reported previously is summarized in Figure 1.2 (reproduced from (53)). However, bear in mind, genetic association studies need to be always interpreted with caution (54-56), as these discoveries may be heavily prone to chance and hence responsible for the non-reproducible associations often observed with human common/complex traits. The potential reasons (57,58) may include: (1) the variant genotyped is not causal and provides incomplete LD

with other potentially functional important variants, (2) studies are underpowered, (3) population stratification, (4) phenotypic and locus heterogeneity. Additionally, false positive discoveries are likely to occur in studies interrogating multiple genes or splitting the cohorts into sub-groups for separate analysis (54), while multiple testing has not been corrected or has been inappropriately dealt with. Candidate gene approach focuses on certain candidate gene regions and has advantage over GWAS (see section 1.3) if the candidate loci are precisely defined. However, the drawback is that candidate gene study precludes new biological pathways to be discovered.

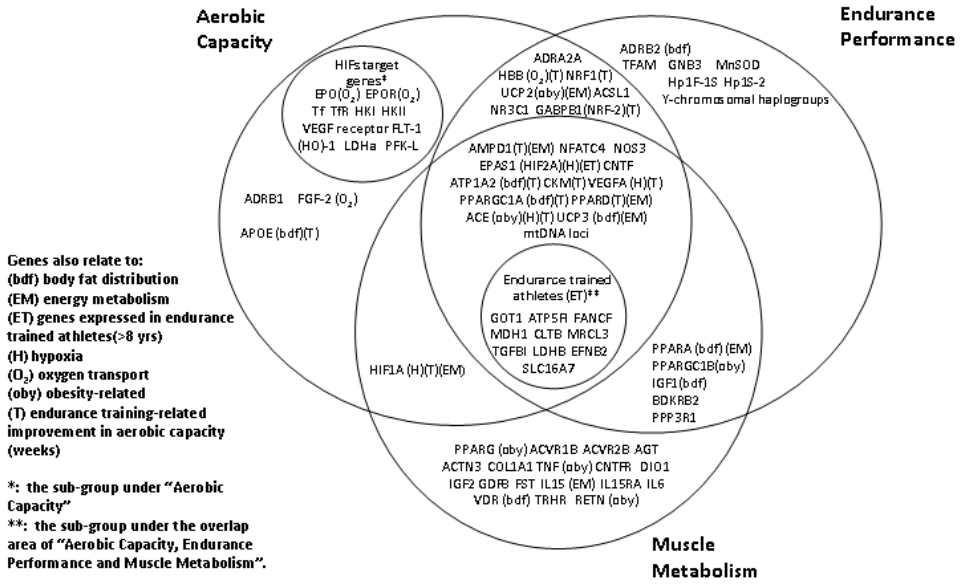


Figure 1.2 Genes or loci extracted from annual update/review of gene map related to aerobic capacity, endurance performance and muscle metabolism. (Reproduced from (53)).

1.3 Genome-wide association studies in complex traits

The first GWAS in age-related macular degeneration (AMD) (59) revealed an intronic and common variant significantly related to AMD in 96 cases and 50 controls and consequently a functional polymorphism in Complement Factor H (*CFH*) gene was identified by resequencing. Since then, numerous GWASs in various complex traits have emerged rapidly (Figure 1.3, (60)).

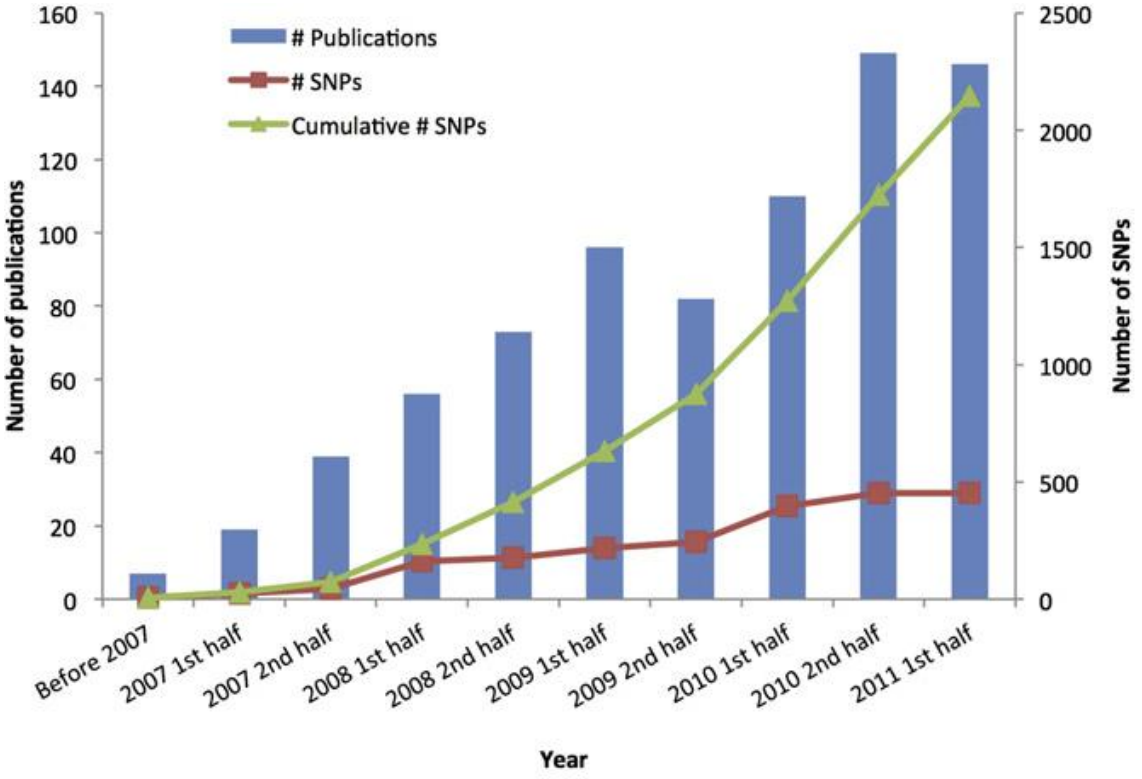


Figure 1.3 Number of GWAS publications and statistically significant SNPs exceeds $5x10^{-8}$ (60). ($5x10^{-8}$ - GWAS significance threshold, see 1.3.2 Statistical Significant Thresholds).

GWAS mapping relies on LD structure between tagSNPs and the causal variants, aimed to gain good coverage of the entire genome by only assaying selected tagSNPs, which is more economical. GWAS is hypothesis-free (no prior assumptions made regarding the location or function of the causal variant), and could identify potential variants that may not be involved in any previous pathway analyses, hence new biology underlying a given

trait may be uncovered. Commercial GWAS genotyping arrays contain hundreds of thousands of SNPs to be genotyped per sample, and this number has been dramatically increased to ~ 4.3 million by using Illumina® HumanOmni5-Quad beadchip (61). Data sets used for content selection of the GWAS chips are mainly obtained from the International HapMap Project and 1,000 Genomes Project. The International HapMap Project focused on the identification of the common patterns (haplotypes) of genetic variations and the determination of tagSNPs representative for these haplotypes in 11 populations of African, European and Asian ancestries with a total sample size of 1,184 (HapMap Phase I, II & III, called "HapMap 3") (62). In "HapMap 3", except that tagging can effectively capture variants/haplotypes with minor allele frequency (MAF) of ≥ 5%, imputation accuracy for variants with lower frequencies (MAF ≤ 5%) is also improved. This integrated dataset (i.e. HapMap 3) provides a robust reference panel to study human variations across various diseases or other complex traits (62). In October 2012, the 1,000 Genomes Project Consortium announced the sequencing data of 1,092 human genomes from 14 populations drawn from Europe, East Asia, sub-Saharan Africa and the Americas, respectively. Ultimately, ~2,500 individuals from 26 populations will be sequenced, and the aim of the 1,000 Genomes project is to provide a more comprehensive catalogue of genetic variations in human genome (63).

African populations have haplotypes much more divergent than Europeans or Asians (64-66). Based on the out-of-Africa hypothesis of human origins, low diversity in other continental populations is a result of population bottlenecks (67,68). Since the initial detection of SNPs was predominantly performed in populations of European ancestry, such as those typed in HapMap Phase I and II, ascertainment bias may arise when mapping fraction of the genome in different populations using the HapMap data (69). Fortunately, tagSNPs are often highly portable across different human populations, however, are less portable for low LD populations, i.e. Africans (65,70). GWAS (using current SNPs

database) in African populations remain to be a particular challenge (71). Genome coverage should be improved with new variants eventually identified from the 1,000 Genomes Project, and ultimately the fulfilment of individual genome-wide deep sequencing will help understand better the genetic architecture in African populations.

An alternative approach in addition to GWAS was suggested by some researchers (72). This is to examine the gene-centric regions of the genome, including synonymous and non-synonymous coding SNPs, and SNPs in 5′-UTR and 3′-UTR regions. They are more likely to have effects on RNA transcription compared to non-coding and intronic SNPs. Which approach to use depends on how the research group decides to balance the completeness and efficiency of a study with or without knowing where the causal variants lie for a particular phenotype in question. Gene-centric approach focuses solely on genes, and only a small set of genic SNPs will be studied. Therefore, multiple-testing burden is reduced. This can accelerate an initial and efficient genome-wide association scan aiming to identify biologically functional variants with reduced type I error rate (false positive rate), but the indirect GWAS can have greater overall power than genic studies.

1.3.1 Sample size and statistical power

Statistical power of a GWAS depends largely on sample size, genetic effect size, marker and causal allele frequencies as well as their correlation (73). Sample size relies on allele frequency and genetic effect of the risk allele in a sample set. The effect size in a case-control study is usually estimated using odds ratio, which implies the ratio of the odds of an individual being a case with a particular genotype/allele vs. the odds being a control with the same genotype/allele. An odds ratio is greater than 1 suggesting a positive association between the genetic variant and the phenotype, and an odds ratio equals 1 meaning no association. The first small GWAS in macular degeneration (59,74) demonstrated that SNPs, with large effect sizes meanwhile frequently enough to be

detected in a population, might be identified even though the sample size is small. For most common complex traits, small effects of common SNPs on trait/disease production require considerably large samples in order to obtain an adequate statistical power to detect these small effects. It is usually found that significant findings from GWAS of complex traits typically have the odds ratios ranged from 1.1 to 1.4 (majority between 1.12 and 1.20, (75)). In this range, a sample size of 8,000 – 12,000 cases and equivalent number of controls can generate sufficient power to detect a variant, if the effect size and allele frequency of the variant can reach the required levels (see Figure 1.4, (75)). Since GWAS is designed based on LD pattern to identify indirect association, the extent of LD between the marker SNP and the causal variant can influence the likelihood of detecting an association. For the causal variants with moderate effect (odds ratio < 2.0) which are likely to be observed in GWAS of complex traits, power will be retained if there are sufficient LD between the marker and the causal allele, and both variants are common and have similar frequencies. Finally, little power will be lost without screening the comparison subjects (i.e. controls) if the phenotype studied is relatively uncommon (< 5% prevalence), and power will be improved for traits with higher prevalence (> 5%) if subjects are predisposed to the trait of interest are excluded from the comparison group (75).

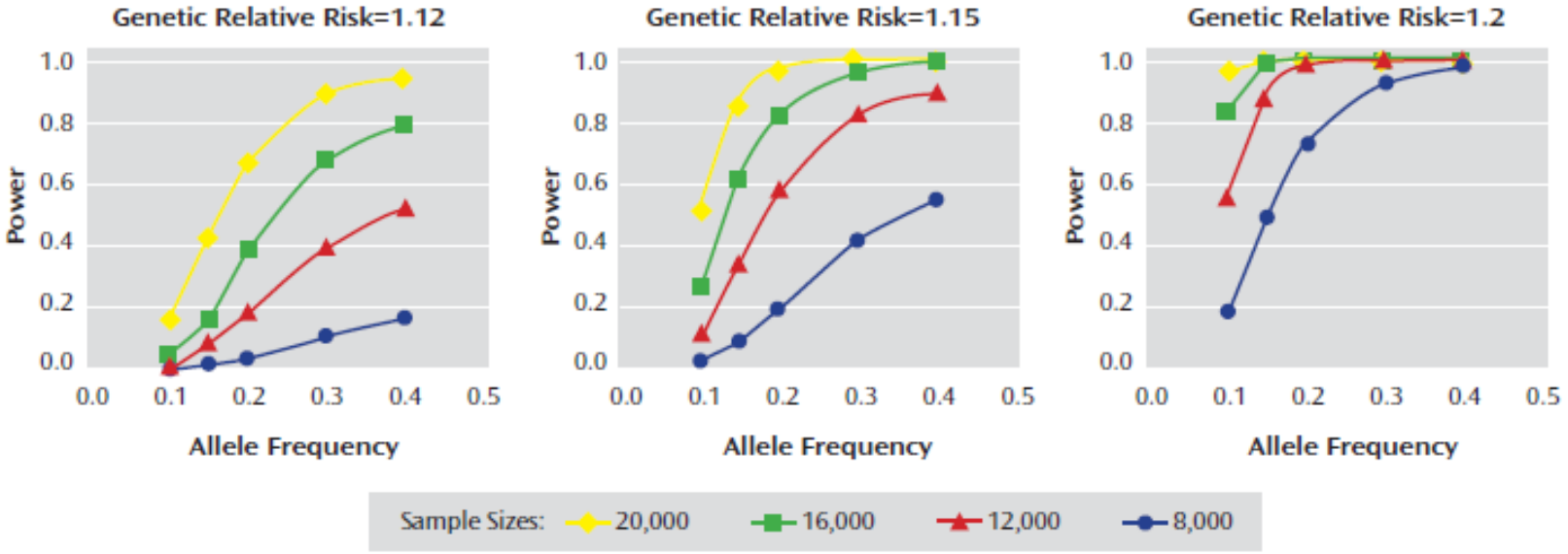


Figure 1.4 Relationship among allele frequency, genetic relative risk and power (75). The images show the expected power for a phenotype with 1% prevalence in a population (p = $5x10^{-8}$), given minor allele frequency, sample size (e.g. yellow represents 20,000 cases and 20,000 controls involved in the comparison), and genetic relative risk (following multiplicative inheritance mode; it is similar to O.R. here, O.R. is the estimate of relative risk in a case-control study). For example, 8,000/8,000 cases/controls will not be able to identify most of the SNPs that count for less than 20% increase in risk (genetic relative risk < 1.2), while 20,000/20,000 set has overall greater power compared with other sample sets (i.e. 8,000/8,000; 12,000/12,000; 16,000/16,000).

1.3.2 Statistical significance thresholds

GWAS produces hundreds of thousands genotype results for each sample simultaneously. Each SNP is tested for an association with the phenotype of interest. This creates the so-called multiple tests/multiple hypotheses problems. If each SNP test is treated as a "repeat", the conventional statistical significant threshold of 0.05 becomes too liberal to distinguish false positives from any true associations. The well-known Bonferroni correction is used to correct for multiple tests that have been carried out, but this is not ideal because it assumes SNPs tested in a GWAS data set are independent from each other (this is not true as SNPs are somewhat correlated in the genome, even among the tagSNPs). Based on Bonferroni correction, a conventional genome-wide significance threshold is set up at $5x10^{-8}$ in GWAS of non-Africans (73), it is calculated by dividing 0.05 by one million SNPs that are thought to be effectively independent across the genome based on LD patterns (76). Another theoretical concept is whether we should deal with this as an issue of multiple hypotheses instead of multiple tests (77). At one extreme of the spectrum, it is argued that there is no need to correct, but to report any significant finding altogether with the number of hypotheses tested (78), then more focused studies/experiments can be developed to validate these positive findings and discard those that are proved to be false. Other approaches to reduce the false positive risks have also been proposed, such as permutation tests (79) and false discovery rate correction (80).

1.3.3 Population stratification and its solution

Allele frequencies of many SNPs may vary from population to population that has different genetic structure from one another. Admixture occurs when genetically distinct groups begin interbreeding. The disparity in frequencies existing between cases and controls would result in population stratification, which might lead to spurious results in association studies. Differences in trait prevalence between cases and controls have also been noted as responsible for false discoveries (81,82). It is even clearly stated by

Wacholder et al (83) that both conditions (differences in allele frequency and trait prevalence) must be met to raise any substantial bias on genuine associations. However, there are only two often used empirical examples to illustrate the effect of population stratification on biased association outcomes (81). One was caused by mixing individuals of European and American Indian ancestries in the association study of an HLA (Human Leukocyte Antigen System) haplotype and diabetes. Both haplotype frequency and diabetes prevalence differed between White Europeans and Pima Indians (84). The other refers to association studies of alcoholism and the dopamine D2 receptor (DRD2), in which the significant associations were resulted from varied *DRD2* allele frequencies and alcoholism prevalence differences across different ethnic groups (85,86). Other studies in population structure, however, demonstrated that standard methods may not be sufficient to detect underlying population stratification, and markers with widely spread allele frequencies among ethnic (sub)groups indeed elevate false positive rates in association studies (82,87,88). Finally, other factors that may be relevant to population stratification include sample size (89,90) and the number of ethnic groups (91). The effect of stratification tends to increase when the sample size increases, but decrease when the number of ethnic groups increases (among non-Hispanic U.S. Caucasians of European origin, (91)). Despite matching cases and controls for ancestry, different statistical methods have also been developed for correcting population stratification in GWAS of unrelated cases and controls so as to minimize potential confounding effect owing to population stratification.

1.3.3.1 Genomic control

Genomic Control is an older method that is used for population stratification evaluation. Devlin and Roeder (1999) (92) first innovated this concept, in which the inflated chi-square test statistics due to population heterogeneity can be adjusted using randomly selected, “unlinked” markers (and unrelated to the phenotype in question) to estimate the

null distribution of the usual test statistic (89,93-95). More precisely, there will be, when substructure problems exist, a higher median of the actual χ^2 distribution than the median of a null distribution. The genomic control approach measures the inflation factor λ that is the median of χ^2 association statistic across SNPs (e.g. genome-wide χ^2 distribution in a GWAS) divided by the median of the normal χ^2 distribution (92). If λ value is ~ 1, the distribution is thought to be close to ideal, suggesting no evidence of population stratification or other confounders; for example, family structure of cryptic relatedness (96). Generally, $\lambda < 1.05$ is considered benign (96). When population stratification exists, genomic control correction can be applied by standardize the actual χ^2 statistic results over the λ value and the corresponding *p* values (corrected) become less significant. It also should be noted that the uniform λ adjustment may lead to overcorrection for markers that do not differentiate by allele frequencies, while insufficiently control for markers alleles that significantly differed across study populations (97).

1.3.3.2 Structure assessment

"Unlinked" genetic markers across the genome can also be used to define subgroups in the total sample set to homogeneous subgroups (94,98-101). Association tests will then be carried out in those matched subpopulations, independently, and these results will be combined statistically for the overall genotype-phenotype association (81). However, since this approach uses genotype data to estimate the probability that an individual belongs to a subgroup, it may not be always able to precisely define the exact number of subgroups, but capture the major structure in the data to infer the fewest number of subpopulations (98).

1.3.3.3 Principal component analysis

Another popular and commonly used approach for population stratification is the principal component analysis (PCA). PCA is used to identify patterns of a data set to visualize their similarities and differences, and by reducing the number of dimensions, data can be

compressed without losing much information (102). This data reduction procedure transforms variables into continuous axes of variation with the first axis (principal component) explaining the greatest variance in the data set, followed by an uncorrelated principal component accounting for the second greatest variance and so on. In population genetic association studies, PCA can be used to examine marker allele frequency variation and to assess population structure among cases, controls and selected reference populations (e.g. populations from the HapMap) along continuous axes of genetic variation, and these methods are implemented in EIGENSOFT package including EIGENSTAT (for population stratification, (97)) and SMARTPCA (for population structure, (103)). Eigenvalues on each axis (usually the first few principal components are thought to be able to capture the stratification patterns) can then be treated as covariates in logistic regression (for binary traits) or linear regression (for quantitative traits) analyses to correct for ancestry effects.

1.3.3.4 Family-based controlling approach

As mentioned above, an important advantage of family-based design over population-based studies in genetics is that it is immune to population stratification. This is because the matching of case and control within a family avoid the problem of allele frequency differences occurring at population level (81). The transmission-disequilibrium test (TDT) design is a popular family-based matching method used to protect from stratification. It requires an affected offspring and his/her parents and assumes 50% chance of inheriting each allele for a polymorphic marker from each parent. Therefore, alleles that are transmitted from the parents to the affected offspring are the cases, and controls are the alleles that are not transmitted. The frequency of an allele transmitted to the affected individual can be estimated by TDT, and an allele with 50% chance more of the time appeared in the affected may indicate a positive association between the allele and the trait. A few drawbacks, however, should be noted for family-based TDT method: 1) Sampling

efficiency. The recruitment of all family members (offspring and parents) may be very difficult or impossible particularly in individual family with late-onset disorders or other severe conditions; 2) Genotyping efficiency and sensitivity. Genotyping needs to be done in all three people (one affected offspring and the parents) to enable a case-control comparison (affected alleles vs. non-affected alleles) using TDT approach, however, one in fact can use full genotype information obtained from all three individuals (TDT design has a two-thirds of the genotyping efficiency, (104,105)). Furthermore, such design is more sensitive to random genotyping errors, which can cause inflated type I error rates (106); 3) Parents need to be heterozygous at a marker locus for allele transmission rate to be calculated, hence at least 50% of the parental genotype data will not be analyzed (81).

1.3.4 Genetic architecture of complex traits: CDCV hypothesis, infinitesimal model, rare allele model and the broad sense heritability model

In the field of genetic epidemiology studies of common and complex traits, identification of underlying genetic structure by GWAS initially relied on the Common Disease-Common Variant hypothesis (CDCV hypothesis), which assumed that the risk of common and complex traits are largely explained by a moderate number of common variants (107-110). However, the small fraction of genetic variation revealed by current GWAS loci raised the issue of missing heritability (111,112) Another three models were developed to look further into the problem: 1) The infinitesimal model, in which genetic variance is attributable to numerous common variants with small effects (113); 2) The rare allele model, in which genetic variance is attributable to many rare variants (allele frequency is typically less than 1%) with large effects (114); and 3) The broad sense heritability model (relative to the narrow sense heritability that refers to the additive portion of the genetic variance), in which genetic variance is attributable to the non-additive components: the gene x gene interactions (i.e. epistasis), gene x environment interactions and epigenetics (e.g. the effect of DNA methylation, histone modification, and microRNA (miRNA)

expression on a genotype without change of the DNA sequence) (115-117). An intuitive illustration of the above 3 models as well as the CDCV hypothesis model is displayed in Figure 1.5 (taken from (115)). There is no clear answer for which elements contribute to an inferred genetic variance and in what proportion, but a way forward is to think about how these proposed hypotheses work together and build the genetic foundation of a complex trait (118).

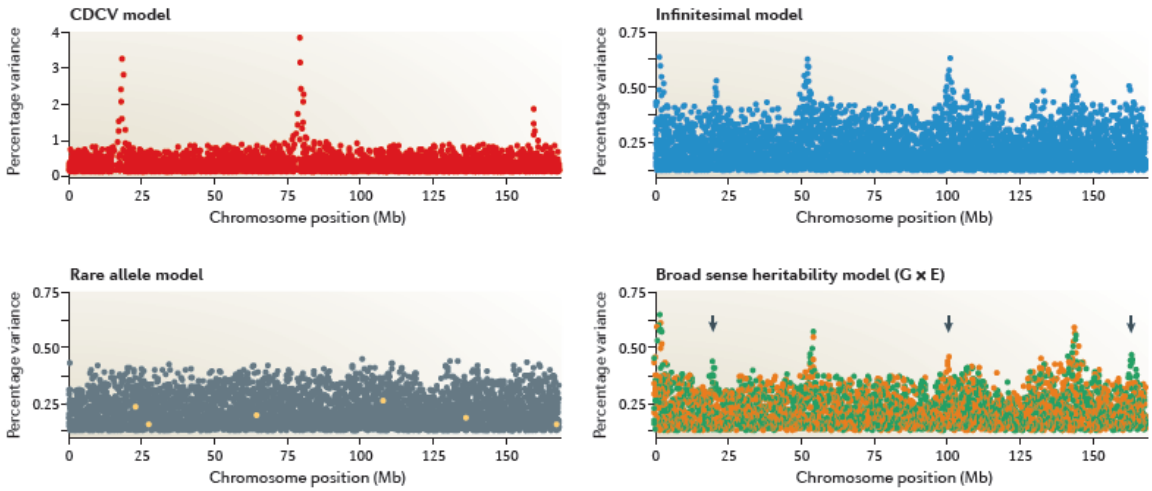


Figure 1.5 Genome-wide association signals for elucidation of four models related to common and complex traits (115). Each plot represents an expected distribution of SNP effects for a study of 2,000 cases and controls. The Y axis shows the percentage of genetic variance explained by each SNP for a trait in a population, and X axis refers to the chromosomal location for each SNP. In the plot of the CDCV model, a small number of SNPs show strong effects on trait being studied (i.e. the expended scale of the percentage of variance on the Y axis compared to other plots). In the plot of the infinitesimal model, the strongly associated signals are explained by a large number of SNPs with small effects. In the plot of the rare allele model, rare causal variants (shown in yellow) may have large effect in a few individuals, although they are not common in a population to explain a reasonable amount of variance and to result in genome-wide significance. In the plot of the broad sense heritability model, for associations that are only present under certain conditions (e.g. influenced by environmental factors, shown as green and orange signals), the overall effect will be reduced in a mixed population at such loci (see arrows, bottom right) and this would lead to few associations to be detected, hence less variance observed (115).

1.3.5 GWAS of exercise-/performance-related traits

A few GWASs of exercise/performance-related traits have been reported to date. These studies that have been published (summarized in Table 1.1) have identified several signals in relation to bone mineral density (BMD; which is thought to correlate with the increased likelihood of injury, therefore BMD was included as a trait of interest when searching the

GWAS Catalog (119)), lean body mass, left ventricular mass, exercise participation and heart rate variability related traits. In the summary table (Table 1.1), all eight studies have been performed in a small to moderate number of samples (i.e. from ~ 200 cases and controls to ~ 1,600 cases and controls in their GWAS discovery cohorts), authors from five of the studies attempted to replicate the results, and only two studies (after replication) reached the conventional GWAS significance threshold of $5x10^{-8}$. These findings are nevertheless interesting, e.g. warrant further replication (for those that have not been replicated) or zoom into the already identified GWAS regions (for those after replications).

Kiel et al (2007) (120) did not find any associations for BMD (called the Framingham Osteoporosis Study) exceeding $5x10^{-8}$ in 1,141 individuals from the Framingham Heart Study (FHS; the most comprehensively characterized multi-generational studies in the epidemiology of cardiovascular disease (121)) by analyzing 70,897 SNPs (the genomic coverage of these SNPs was low); the Framingham Osteoporosis Study is a derived study from the FHS. Xiong et al (2009) (122) found that rs11864477 in *ADAMTS18* gene was significantly associated with hip BMD ($p = 2 \times 10^{-8}$) after replications and meta-analyses across different ethnic groups (i.e. White U.S. samples, Chinese and west African ancestry samples), and future molecular studies are needed for better understanding of the mechanisms.

De Moor et al (2009) (123) identified three suggestive SNPs in association to exercise participation in the Dutch and American cohorts of 2,622 individuals using approx. 1.6 million imputed SNPs. Rankinen et al (2011) (124) performed a GWAS in 472 individuals of European ancestry from the HERITAGE Family Study, their data suggested that ten most significant SNPs could account for 35.9% of the variance in the submaximal exercise heart rate response, and 9 of these SNPs contribute 100% to the heritability of this heart rate variability trait. Nevertheless, these studies require either further replications or additional studies to confirm these identified loci.

Despite of small sample size (GWAS cohort) and limited genomic coverage provided by Affymetrix 100K chip used in the study of Arnett et al (2009) (125), the authors reported that five SNPs from the initial GWAS of left ventricular mass in Caucasians were subsequently validated in an independent Caucasian cohort with a much greater number of individuals compared to the GWAS samples. In addition, a SNP within the intron of a previously reported candidate gene (*KVNB1*) for left ventricular hypertrophy was replicated in the African-American cohort (a 2nd replication cohort used by the authors). Future fine mapping and functional studies are required to find the causal genetic variant and its functional relevance to left ventricular hypotrophy, which is also often seen in sprint/power-oriented athletes.

Two GWASs of lean body mass (one of the studied phenotypes) in Chinese populations were carried out (126,127). Although authors from both studies attempted to replicate their association results in larger cohorts of European ancestry, most significantly replicated SNPs did not reach genome wide significance of $5x10^{-8}$. The authors aware that further replications would be needed in other populations and in a larger scale, while these findings are interesting and deserve detailed investigation on the biological function of the putative loci once further confirmed.

Polymorphisms in the thyrotropin-releasing hormone receptor (*TRHR*) gene in relation to skeletal muscle trait, which were identified from a genome-wide association scan and confirmed in three additional replications and meta-analyses, seem the most promising (128). Another study used genomic predictor score to establish a panel of GWAS SNPs contributing to $\dot{V}O_{2max}$ trainability in response to standardized endurance exercise training followed by replications in other cohorts, offering an alternative way to conduct GWAS in exercise-related traits. These two studies are described in detail below.

Table 1.1 Summary of studies for exercise-/performance-related traits from the NHGRI GWAS Catalog (adapted from (119)).

First Autho	Date	Disease/Trait	Initial Sample Size	Replication Sample Size	Region	Reported Gene(s)	Mapped gene	Strongest SNP-Risk Allele	Context	Risk Allele Frequency	p-Value	p-Value (text)	OR or beta; 95% C.I.	Platform [SNPs passing QC]
Han Y	08/31/2012	Compressive strength and appendicular lean mass	825 CA females, 802 CA males	1,059 EA males, 2,227 EA females	11q12.2	FADS1, FADS2	FADS1	rs174547-C	intron	0.33	2.00E-07	(Males + Females)	NR; NR	Affymetrix [701,525]
Han Y	08/31/2012	Compressive strength and appendicular lean mass	825 CA females, 802 CA males	1,059 EA males, 2,227 EA females	11q12.2	FADS1, FADS2	FADS1	rs174549-A	intron	0.3	8.00E-07	(Males)	NR; NR	Affymetrix [701,525]
Hai R	06/29/2012	Lean body mass and age at menarche (combined)	801 Han Chinese women	1,692 EA women	1q23.2	DARC	DARC	rs3027009-?	nearGene-5	NR	7.00E-07	(Bivariate)	NR; NR	Affymetrix [909,622]
Rankinen T	12/15/2011	Heart rate variability t	472 EA individuals from 99 fami	NR	2p25.1	YWHAQ	ADAM17 - YWHAQ	rs6432018-?	Intergenic	NR	8.00E-07		NR; NR	Illumina [320,000]
Rankinen T	12/15/2011	Heart rate variability t	472 EA individuals from 99 fami	NR	8p12	RBPMS	RBPMS	rs2979481-?	intron	NR	4.00E-06		NR; NR	Illumina [320,000]
De Moor MH	2009-2-9	Exercise (leisure time)	1,644 Dutch individuals, 978 European individuals	NR	10q23.2	PAPSS2	PAPSS2	rs10887741-T	intron	NR	4.00E-06		1.32; 1.17-1.49	Affymetrix and Perlegen [~1.6 millio (imputed)
De Moor MH	2009-2-9	Exercise (leisure time)	1,644 Dutch individuals, 978 European individuals	NR	18p11.32	C18orf2	C18orf2 - METTL4	rs8097348-G	Intergenic	NR	7.00E-06		1.36; 1.19-1.56	Affymetrix and Perlegen [~1.6 millio (imputed)
De Moor MH	2009-2-9	Exercise (leisure time)	1,644 Dutch individuals, 978 European individuals	NR	2q33.1	DNAPTP6	C2orf47 - SPATS2L	rs12612420-A	Intergenic	NR	8.00E-06		1.43; 1.22-1.67	Affymetrix and Perlegen [~1.6 millio (imputed)
Arnett DK	05/19/2009	Left ventricular mass	101 EA cases, 101 EA controls	704 EA siblings, 1,467 AA ancestry sibli	7q21.11	CD36	CD36	rs10499859-?	intron	0.45	3.00E-06	(Caucasian)	0.09; NR (LVMI)	Affymetrix [96,258]
Arnett DK	05/19/2009	Left ventricular mass	101 EA cases, 101 EA controls	704 EA siblings, 1,467 AA ancestry sibli	5p13.2	RAI14	C1QTNF3 - RAI14	rs409045-?	Intergenic	0.38	8.00E-07	(Caucasian)	0; NR (LVMI)	Affymetrix [96,258]
Liu XG	2009-4-3	Body mass (lean)	1,000 individuals	1,488 individuals, 1,972 family members, 2,955 Chinese individuals	8q23.1	TRHR	TRHR	rs7832552-T	intron	0.32	4.00E-10		0.1; 0.04-0.16 kg increase	Affymetrix [379,319]
Xiong DH	02/25/2009	Bone mineral density	1,000 EA individuals	4,925 EA individuals, 350 CA hip fracture cases, 350 CA controls, 2,955 CA individuals, 908 WA ancestry men	1p22.1	TGFBR3	TGFBR3	rs17131547-A	intron	0.01	1.00E-06	(spine BMD)	1.2; % [NR] of variance explained	Affymetrix [379,319]
Xiong DH	02/25/2009	Bone mineral density	1,000 EA individuals	4,925 EA individuals, 350 CA hip fracture cases, 350 CA controls, 2,955 CA individuals, 908 WA ancestry men	16q23.1	ADAMTS18	ADAMTS18	rs11864477-C	intron	0.12	2.00E-08	(hip BMD)	1; % [NR] of variance explained	Affymetrix [379,319]
Kiel DP	09/19/2007	Bone mineral density	1,141 individuals(Framingham)	NR	13q21.31	Intergenic	RPL32P28 - OR7E156P	rs9317284-?	Intergenic	NR	2.00E-07	(FNBMDm)	NR; NR	Affymetrix[70,897]
Kiel DP	09/19/2007	Bone mineral density	1,141 individuals(Framingham)	NR	10p15.2	Intergenic	PITRM1 - KLF6	rs2165468-?	Intergenic	NR	1.00E-06	(FNBMDm)	NR; NR	Affymetrix[70,897]
Kiel DP	09/19/2007	Bone mineral density	1,141 individuals(Framingham)	NR	3p24.1	RBMS3	RBMS3	rs10510628-?	intron	NR	3.00E-06	(TRBMDm)	NR; NR	Affymetrix[70,897]
Kiel DP	09/19/2007	Bone mineral density	1,141 individuals(Framingham)	NR	20q11.23	CTNNBL1	CTNNBL1	rs4811196-?	intron	NR	1.00E-06	(TRBMDf)	NR; NR	Affymetrix[70,897]
Kiel DP	09/19/2007	Bone mineral density	1,141 individuals(Framingham)	NR	4p16.1	Intergenic	RAF1P1 - ZNF518B	rs9291683-?	Intergenic	NR	2.00E-06	(BUA)	NR; NR	Affymetrix[70,897]
Kiel DP	09/19/2007	Bone mineral density	1,141 individuals(Framingham)	NR	12q21.1	Intergenic	RPL31P48 - VENTXP3	rs10506701-?	Intergenic	NR	1.00E-06	(TRBMD)	NR; NR	Affymetrix[70,897]
Kiel DP	09/19/2007	Bone mineral density	1,141 individuals(Framingham)	NR	7q35	CNTNAP2	CNTNAP2	rs2214681-?	intron	NR	3.00E-06	(BUA)	NR; NR	Affymetrix[70,897]
Kiel DP	09/19/2007	Bone mineral density	1,141 individuals(Framingham)	NR	16q23.3	Intergenic	MPHOSPH6 - CDH13	rs4087296-?	Intergenic	NR	3.00E-07	(TRBMDf)	NR; NR	Affymetrix[70,897]

AA: African-American; CA: Chinese ancestry; EA: European ancestry; NR: not reported; OR: odds ratio; 95%CI: 95% confidence interval; QC: quality control; ?: risk allele not reported; *p*-Values are round to 1 significant digit (e.g. a reported *p*-Value of 4.8×10^{-7} is rounded to 5×10^{-7}); *p*-Value (text): information describing context of *p*-Value.

1.3.5.1 GWAS of skeletal muscle trait

Loss of muscle function can cause a number of diseases, particularly in elder people (129,130). Lean body mass is an indicator for skeletal muscle quantity and quality (131). The authors have found strongly associated SNPs in *TRHR* gene with lean body mass. The study can be divided to three stages: initial GWAS scan, replication stage and the meta-analyses. The study flow is extracted from Liu et al (2009) (128) (Figure 1.6).

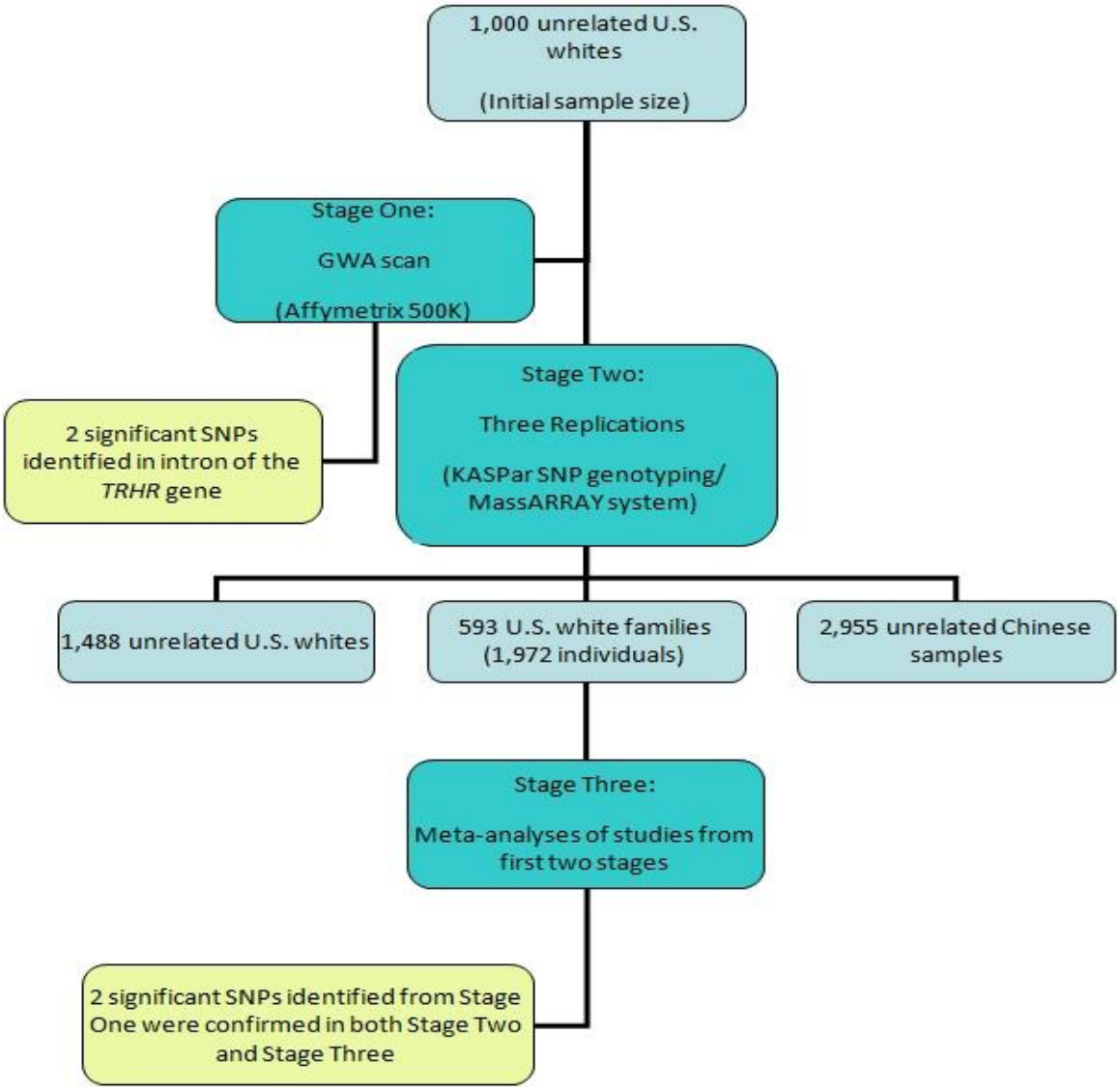


Figure 1.6 Study stages of Liu et al (2009) (128) to identify polymorphism related to lean body mass.

The authors set the significance level for their genome-wide association scan as 1.32 x 10^{-7}, and this was calculated by 0.05 divided by the number of available SNPs (n = 379,319) for association analysis. rs16892496 and rs7832552 in intron of the *TRHR* gene exceeded the threshold in the initial genome-wide associaiton scan. Both SNPs were subsequently replicated in three independent cohorts of multi-ethnic groups, and the strength of both signals was enhanced after the meta-analyses of the GWAS and replication studies. The combined *p* values were of 5.53 x 10^{-9} and 3.88 x 10^{-10} for rs 16892496 and rs7832552, respectively (see Table 1.2; (128)). The authors have argued that the findings are unlikely to be artificial, because 1) conservative Bonferroni correction was applied to claim for significant associations following genome-wide association scan, 2) other 15 SNPs in the *TRHR* gene region also made the same suggestive association signals along with rs16892496 and rs7832552, indicating that the two most significant signals are unlikely subjected to genotyping errors, 3) confounder, such as population stratification, was examined and strictly controlled, 4) independent replications confirmed the initial findings.

Table 1.2 Significant association results of the *TRHR* SNPs for lean body mass across the three stages. (128)

		Replication Studies[a]			
SNP	Discovery Scan (n = 973)	Sample 1 (n = 1488)	Sample 2 (n = 2955)	Sample 3[b] (n = 1972)	Combined[c] (n = 7415)
rs16892496	7.55×10^{-8} (0.107 ± 0.032)	0.018 (0.112 ± 0.038)	0.013 (0.074 ± 0.024)	0.083 (0.087 ± 0.029)	5.53×10^{-9} (0.090 ± 0.015)
rs7832552	7.58×10^{-8} (0.102 ± 0.030)	0.0056 (0.105 ± 0.036)	0.012 (0.076 ± 0.024)	0.015 (0.081 ± 0.028)	3.88×10^{-10} (0.061 ± 0.014)

a. Sample 1, unrelated U.S. white sample; Sample 2, unrelated Chinese sample; Sample 3, U.S. white families. b. Effect sizes in U.S. white families were calculated based on founders. c. Meta-analyses were computed under the random-effect model. The values in the table above are *p* values followed by effect sizes expressed by beta coefficients ± standard errors.

Thyroid hormone has an important role in skeletal muscle development (132-134). The *TRHR* gene is considered as a candidate for the study of muscle power and strength (49). Liu et al (2009) (128) published the *TRHR* findings in 2009, since then there is not any functional study has further looked into the role of TRHR in muscle metabolism or other related pathways. There is an urgent need for such studies to be conducted for establishing

functional/physiological relevance of TRHR to lean body mass variation and muscle strength.

1.3.5.2 GWAS of $\dot{V}O_{2max}$ trainability

This is the first GWAS of $\dot{V}O_{2max}$ changes in response to standardized exercise training. The authors (135) hoped to identify SNPs and genes using an intervention study design to precisely define the phenotype of interest (i.e. $\dot{V}O_{2max}$ was measured twice at baseline and post-training). GWAS association analysis was performed in a subset (n = 470) of whites in the HERITAGE Family Studies to identify genetic polymorphisms associated with the increase in $\dot{V}O_{2max,}$ the genotyping was done using Illumina HumanCNV370-Quad Beadchips. 39 SNPs (MAF ≥ 8%) were reported to be associated with $\dot{V}O_{2max}$ training response at the $p < 1.5x10^{-4}$. Among them, the redundant SNPs (mainly due to high LD) were eliminated using a backward regression model, 21 out of the 39 SNPs retained in the final model ($p < 0.05$). These 21 SNPs explained ~48.6% $\dot{V}O_{2max}$ response variation, notably, 6 individual SNPs each accounted for ≥ ~3% of the variance. The authors then computed a summary "predictor score" using the 21 SNPs, and the SNP was coded as "0 = low-response allele homozygote, 1 = heterozygote, and 2 = high-response allele homozygote". In theory, the predictor score would range from 0 to 42 (2 times 21), and the observed range was from 7 to 31. Individuals with a score of ≤ 9 had a mean increase of 221 ml/min in $\dot{V}O_{2max,}$ while individuals showed a much higher $\dot{V}O_{2max}$ response (mean = 604 ml/min) with a score of ≥19. 15 most significant SNPs of the 21 were re-tested for replication in a subgroup of blacks from the HERITAGE Family Study (n = 247), women in the Dose Response to Exercise (DREW) Study (n = 112), and men and women in the Studies of a Targeted Risk Reduction Intervention Through Defined Exercise (STRRIDE, n = 183), and they were genotyped by Illumina GoldenGate assay and Veracode technology on the BeadXpress platform. The concordance between GoldenGate and GWAS arrays were 100% through genotyping 20 HERITAGE white subjects using both

methods. SNPs were retained if reached the significance of $p < 0.05$ in the replication cohorts, and only 5 of the 15 SNPs were replicated. As the authors aware that a few explanations for the lack of replication exist: 1) low genomic coverage of the initial genome-wide association scan in the HERITAGE whites, genotyping a larger panel of SNPs may produce more significant signals for subsequent replications; 2) although the HERITAGE blacks (replication cohort) were selected and trained based on the same protocol as the HERITAGE whites, the irreproducible GWAS association results may be due to discrepancies of allele frequencies and LD structures in individuals of European and African descent; 3) the DREW and STRRIDE (the other two replication cohorts) were subjected to different training programmes (relative to the HERITAGE Family Study), subjects in these two studies were ~20 yr older than the HERITAGE whites and blacks and the increase in $\dot{V}O_{2max}$ response were lower than those in the HERITAGE subjects , these might result in the reduced strength of the replication signals to be detected; 4) lastly, sample size of the initial GWAS was moderate at most, and the three replications gave even smaller sample size. Despite these limitations, the phenotype of $\dot{V}O_{2max}$ training response was clearly defined and measured, this carefully designed intervention study might reduce the number of confounders and therefore the sample size required to detect SNPs with a significant effect size (136). Furthermore, the genomic predictors of $\dot{V}O_{2max}$ training response identified from the summary score analysis provided candidate markers for the new biology of aerobic fitness and adaptation to regular exercise.

1.3.6 Replication

In association studies, most reported association findings between genotypes and phenotypes failed in subsequent replication. Replication is recognized as a main tool to distinguish chance from true associations for both candidate gene and GWAS associations. This is even more true for GWAS replication, because the large number of SNPs analyzed reflect a significant amount of tests/hypotheses to be statistically tested, hence increasing

the likelihood of finding type I errors. The goals of replication for the initial association hits from GWAS can be summarized as following (137):

1) To provide convincing statistical evidence for association. In Bayes' theorem, a true association depends not only on the observed *p* value, but power of the study (as indicated before, power is a function of minor allele and the correlated allele frequencies, effect size and sample size), prior probability of the associated variant for a given trait, and the anticipated effect size (138-140). Given yet unclear genetic architecture underlying the common/complex traits (see section 1.3.4), the true effect of a variant causing a specific phenotype is unknown. Formal replication can be applied to genotype the most promising GWAS SNPs in an independent sample of sufficient size to re-confirm these signals before taking them forward for functional studies. An alternative approach to replication is to include a region in a predictive genomic score for the trait so as to provide an update for the prior for association in subsequent studies based on Bayesian probability theory, and a less stringent threshold may be applied.

2) To eliminate false association due to bias. Other artificial effect accounting for a significant association may be the present, for example, population stratification and technical bias owing to the differences in genotyping and analysis procedures between cases and controls.

3) To improve the reliability for effect estimation. The effect size is often to be inflated in the discovery GWAS due to inadequate power to detect the true effect with smaller magnitude, this is known as "winners' curse" (141-144). Replications in additional samples taking into account appropriate calculation for the number of samples required should be able to produce more accurate estimates of the genetic effect.

4) To generalize an association across different populations. Given discrepancies in allele frequencies and LD structure in various populations, the initial replications should be carried out in samples of similar genetic ancestry to the discovery cohort. After this, replications in different ethnic populations may be taken forward for potential common functional variants to be identified across these different populations.

1.4 Extreme phenotype

Complicated genetic architecture underlying common and complex traits have made the determination of the resulting genetic variants extremely challenging. For such genetic studies, particularly at genome-wide scale, an adequate power is difficult to achieve, because an insufficient sample size may be employed for the identification of common genetic variations with small and modest effects or rarer variants with larger effects. One strategy that may increase efficiency is to study individuals at the two extremes of a phenotype distribution, in which the allele frequency may be enriched in one or both phenotype extremes; therefore, the need for very large samples may be circumvented and potential candidate genes/SNPs may be uncovered (145). For example, in a GWAS of obesity and its related traits (146), the authors analyzed a few hundred (a classical GWAS typically requires a sample size of a few thousands at least) cases (extremely obese) and similar size of controls (never overweight) and then successfully replicated top ~500 GWAS SNPs in the combined cohort of cases, controls and family members using a different genotyping platform, they found genome-wide significant SNPs within the fat mass and obesity associated (*FTO*) gene (16 SNPs) and neurexin 3 (*NRXN3*) gene (1 SNP) associated with obesity (as a binary trait) and body fat distribution, respectively. Another research group used exome sequencing and an extreme phenotype study design in individuals with cystic fibrosis to identify genetic factors leading to *Pseudomonas*

aeruginosa infection in cystic fibrosis, the authors successfully performed exome sequencing in 91 participants (out of 96), including 43 individuals with early age of onset of chronic *P. aeruginosa* infection (treated as early *P. aeruginosa* extreme cases) and 48 oldest individuals who had not had chronic *P. aeruginosa* infection as late *P. aeruginosa* extreme controls (147). Genetic variants in the dynactin 4 (*DCTN4*) gene, encoding a dynactin protein, were significantly associated with time to chronic *P. aeruginosa* infection. The authors demonstrated that the success of this approach (using exome sequencing to discover genes responsible to a complex trait) was partially due to well phenotyped and matched extreme samples, relatively large estimated effect size for *DCTN4* and reasonably high MAF variants included in the analysis (147).

1.5 Aims

The overall aim of the studies described in this thesis is to identify genetic variants related to elite human performance by analysing athletes of the highest standard including world record holders, world champions, Olympians and winners of other international events by carrying out two types of association studies: the candidate gene association study and GWAS. Therefore, the specific aims of this thesis are:

- To investigate whether two genes encoding angiotensin converting enzyme (*ACE*) and actinin, alpha 3 (*ACTN3*), which have evident roles in the regulations of circulatory homeostasis and muscle metabolism, respectively, are associated with elite swimmer status in both Caucasians and East Asians, who consistently show high levels of performance in swimming, using the candidate gene association approach; limited efforts has been made to investigate the relationship between the two genes and elite swimmer status in the two populations;

- To identify common polymorphisms associated with elite sprint athlete status by carrying out GWAS in Jamaicans, African-Americans and Japanese, respectively. As an aside, potential common polymorphisms involving in elite endurance performance in Japanese are also examined using the genome-wide association approach;

    - To perform a meta-analysis of the combined three sprint GWAS cohorts, namely Jamaicans, African-Americans and Japanese;

- Additionally, to investigate whether sprint-related SNPs identified from published reports show predictive utility on elite sprint performance, assessed by using the genomic data generated from current GWAS cohorts.

2 Materials and methods

This chapter summarizes the procedures applied to sample collection, DNA preparation, storage and transportation, genotyping and data analyses.

2.1 Study samples

Elite athletes were sampled from the population of world record holders, world champions, Olympians and winners of other international events, or athletes who had at least either participated in international or national level competitions. The sport events mainly include swimming, running, jumping and throwing. Controls were drawn from geographically matched regions as the elite athletes.

2.1.1 Elite athlete cohorts

Elite Caucasian and East Asian Swim Cohort: The elite swimmer cohort consists of Caucasian and East Asian subjects. 200 elite Caucasian swimmers from European, Commonwealth, American and Russian sub-cohorts were sampled during swimming competitions during 2005 and 2006 and sub-divided into short and middle distance (SMD ≤ 400 m, n = 130) or long distance swimmers (LD > 400 m, n = 70). Caucasian swimmers of the European, Commonwealth and American sub-cohorts were of world-class status or highly competitive in international competitions (lifetime World Rankings in the top 50, averaging of the World Rankings of these swimmers in these events within the top 20; swimming World Rankings can be accessed through http://www.fina.org/H2O/). Caucasian swimmers of the Russian sub-cohort (n = 21) all had represented their country in international competitions at very long distances (5 – 25 km), and many were World Champions or World Championship prizewinners. Caucasian controls were drawn from previous published reports (*ACE*-C: n = 1248, (148); *ACTN3*-C: n = 1694, (149-153)).

Differences in *ACTN3* R577X allele frequency between East and West Europe were dealt with by including Russian controls in the same proportion as Russians made up in the swimmer cohort (21 Russians in a total of 193 genotyped swimmers = 10.9%). Therefore, 184 Russian controls were randomly selected for inclusion in the analysis. 158 elite Japanese and 168 elite Taiwanese swimmers were classified as short distance (SD ≤ 100 m; n = 166) and middle distance (MD: 200 – 400; n = 160), and all had either participated in international competitions such as the Olympics, World Championships and Asian Games, or were participants in national competitions. Controls were pooled from general Japanese (n = 649) and Taiwanese (n = 603) populations, respectively and were healthy adults of both sexes and not professionally connected with athletics/sport. Gender, ethnicity, event and level of performance were recorded for all subjects.

Elite Jamaican and USA sprint cohorts: These cohorts are comprised of elite Jamaican and African-American athletes representing the highest level of sprint performance and geographically matched controls. In the Jamaican cohort, 116 athletes (male = 60, female = 56) and 311 control subjects (throughout the whole island; male = 156, female = 155) were recruited (154). 71 and 35 athletes had participated in 100-200 m and 400 m sprint events, respectively; and 10 athletes were involved in the jump and throw events. These athletes can be further classified into national (n = 28) and international athletes (n = 88) who were competitive at the national level in Jamaica and the Caribbean or at major international competitions for Jamaica. Among the 88 international athletes, 46 had won medals at major international events or held world records in sprinting. In the African-American cohort, samples from 114 elite sprint athletes (male = 62, female = 52) and 191 controls (throughout the United States; male = 72, female = 119) were collected (154). Among these athletes, 48, 42 and 24 athletes participated 100-200 m, 400 m, and jump and throw events, respectively. Athletes can be subdivided into 28 national and 86 international

athletes; 35 of these athletes had won medals at international games and/or broken sprint world records.

Elite Japanese track-and-field (TF) athlete cohort: The elite Japanese TF athlete cohort involves 60 international endurance and 54 international sprint athletes, but not necessarily the medallists. 118 healthy controls of both sexes who were not involved in competitive sports were recruited from general Japanese population.

2.1.2 DNA collection, extraction, quantification, storage and transportation

DNA from elite Jamaican and USA sprint and elite Caucasian swim cohorts was isolated from buccal cells. These subjects were asked not to consume food or drink for at least 30 minutes before providing a sample. Buccal cell samples were collected by a trained individual by firmly rubbing a brush (Medical Packaging Corporation, Camarillo, CA, USA) against the inside of each subject's cheek for at least 15 seconds. The head of the brush was cut into a screw cap tube containing cell lysis solution (0.1 M Tris-HCl pH 8.0, 0.1 M EDTA; 1 % SDS). DNA was extracted using the QIAamp® DNA Mini kit (QIAgen, Hilden, Germany) according to the manufacturer's instructions (155) with minor adjustments. In brief, 500 µl of each sample was transferred to a clean 1.5 ml microcentrifuge tube, followed by adding in 15 µl proteinase K and then incubated at 55°C for 30 minutes to 1 hour in an air incubator (Binder B28, BINDER GmbH, Tuttlingen, Germany). 500 µl of 100% ethanol was added to the microcentrifuge tube, mixed by vortexing thoroughly and then was carefully transferred to a spin column with a pipette. The supernatant in the 2 ml collection tube after centrifuge was then discarded and the DNA was absorbed to the membrane of the spin column. The DNA was washed twice using 500 µl of two different buffers (AW1 and AW2) and finally was dissolved in 200 µl of AE buffer (10 mM Tris-Cl; 0.5mM EDTA; pH 9.0). The DNA samples were then quantified using the Nanodrop Technologies Nanodrop® ND-8000 Spectrophotometer

(Wilmington, DE, USA) measuring 8 samples at a time with a multichannel pipette to transfer 2 µl undiluted sample to the sample pedestals, and DNA concentration was determined using the absorbance method (i.e. absorbance at 260 nm (A_{260})).

There are also a few athletes DNA samples that were collected from whole saliva using Oragene DNA Self Collection Kit – disc format (OG-250) (DNA Genotek Inc., Canada). Participants were advised to rinse mouths with drinking water, and to wait for 5 minutes before saliva collection. About 2 ml of saliva was collected under appropriate supervision. The trained individual would then cover the disk with the cap tightly and invert the container repeatedly for 10 seconds in order to sufficiently mix the saliva sample with the Oragene chemistry (DNA Genotek Inc., Canada). DNA was extracted following manufacturer's instructions for manual purification of DNA from 0.5 ml of sample using DNA Genotek's prepIT·L2P DNA extraction kit (156) with minor adjustments. In brief, 500 µl of saliva sample was transferred into a 1.5 ml microcentrifuge tube. 20 µl Oragene DNA purifier solution was added to the microcentrifuge tube containing the sample, mixed by vortexing for 3 seconds and then placed on ice for 10 minutes. The sample mix was centrifuged using a microcentrifuge at room temperature for 10 minutes at 13,000 rpm (15,000 × *g*). The supernatant was carefully transferred into a fresh tube and an equal volume of 100% ethanol (500 µl) was added, mixed by inverting gently 10 times. The mix was incubated at room temperature for 10 minutes to allow full precipitation of the DNA, and was centrifuged at room temperature for 2 minutes at 13,000 rpm. The supernatant was removed and discarded this time, and pellets were dried in an air incubator at 50°C for about 20 minutes and dissolved in 300 µl of TE buffer (100 mM Tris, 10 mM EDTA, pH 8.0). Samples were then stored at room temperature overnight to allow complete rehydration of the DNA, followed by vortexing. Undiluted DNA samples were quantified by the absorbance method using the Nanodrop Technologies Nanodrop® ND-8000 Spectrophotometer (Wilmington, DE, USA).

Samples were held at 4°C while processing (for example, during DNA extraction and quantification). For longer term storage, the remained buccal cells/saliva samples as well as the extracted DNA were frozen at -20°C. For whole genome genotyping using Illumina Omni Whole Genome Assays, the purified genomic DNA samples of elite Jamaican and USA sprint cohort were then shipped to Tokyo Metropolitan Institute of Gerontology, Japan, where the samples were further processed and prepared for whole genome genotyping, by a logistic transportation courier (DHL, UK). These DNA samples were stored in 8 x 12 format sterile Thermo Scientific Matrix Storage Tubes (0.75 ml, Thermo Fisher Scientific, Hudson, New Hampshire, USA) and placed in Polystyrene Cold Boxes covered with dry ice during transportation, and arrived to Japan in a maximum of 72 hours. DNA quantity and quality were evaluated again in Japan using PicoGreen® Assay (a more precise measurement for quantifying double stranded DNA); samples with at least 50 ng of DNA were taken forward for whole-genome genotyping. For single SNP genotyping using Taqman Assays, the DNA was diluted and standardized to a working concentration of 3 $ng \cdot \mu l^{-1}$ and stored, during the genotyping, at 4 °C in Rigid Thin Wall 96 x 0.2 ml Skirted Microplates (Starlabs UK Ltd, Buckinghamshire, UK).

2.2 Genotyping

Taqman® SNP genotyping method was used for genotyping of *ACE* and *ACTN3* polymorphisms. The whole genome genotyping arrays were used for whole genome wide association analysis by interrogating > 700,000 markers/sample simultaneously across the entire human genome.

2.2.1 Taqman® SNP genotyping

2.2.1.1 Taqman® assay

Taqman® SNP genotyping assay contains sequence specific forward & reverse primers, and two Taqman® minor groove binders (MGB) probes. All Taqman® SNP genotyping

assays are designed to work with Taqman® universal master mix, containing AmpliTaq Gold® DNA polymerase and buffer components optimized for tight endpoint fluorescence cluster. More specifically, the two primers are used for DNA chain extension and amplification during the Polymerase Chain Reaction (PCR), and the two Taqman® MGB probes provide a fluorescence signal for each targeted allele. The two probes are labelled with two fluorescent dyes at their 5 prime ends with the VIC-dye attached to the 5′ end of the allele 1 probe and the FAM-dye to the 5′ end of the allele 2 probe, and a nonfluorescent quencher (NFQ) is attached to the 3 prime end of each probe. When the reporter dye (VIC-/FAM-dye) and the quencher dye on the probe is intact, the reporter fluorescence is absorbed by the quencher dye. AmpliTaq Gold® polymerase extends the primers bound to the template DNA in a 5′ to 3′ direction. Once a probe matched with a specific sequence, the AmpliTaq Gold® polymerase cleaves, resulting in separating the reporter dye from the NFQ and leading to the increase of the fluorescence emissions during the real-time PCR amplification using the ABI's StepOnePlus™ Real Time PCR system (Applied Biosystems, CA, USA). Genotypes were called from end-point reads using ABI's StepOne™ Software v2.1 (see 2.2.1.2 for more).

2.2.1.2 PCR conditions and endpoint reading using StepOne™ software v2.1

The recommended purified genomic DNA template as per the manufacturer's instructions is 1 – 20 ng with a uniformed concentration. A total volume of 20 µl reaction mix, including 9 ng gDNA (i.e. 3ng/µl for 3 µl), 1.0 µl 20 x Taqman genotyping assay, 10.0 µl TaqMan universal PCR master mix and 6.0 µl distilled water, was used for real-time PCR for each sample processed at Glasgow (i.e. Caucasian swimmers, see experimental Chapter 3). Different reaction design is possible as long as it meets the initial recommendation from the manufacturer (e.g. roughly equivalent DNA concentration/quantity for all samples across the whole PCR plate). For each 96-well PCR plate, it included 94 DNA reactions and 2 no template (or negative) controls. The genotyping was performed using ABI

StepOnePlus™ Real Time PCR system (Applied Biosystems, CA, USA) with the manufacturer's recommended PCR thermal cycling conditions: AmpliTaq Gold enzyme activation (10 minutes at 95°C), followed by 40 cycles at 92°C for 15 seconds (denature) and at 60°C for 1 minute (anneal/extend). Genotypes were then called from end-point reads using ABI StepOne™ Software v2.1 measuring the fluorescence during plate reading to plot the fluorescence values based on the signal for each DNA sample. The automatic allele calls were set up during calling. A successfully genotyped plate would show tightly clustered VIC- and FAM-dye homozygote and heterozygote clusters. The genotyping results can be exported as an excel file for further analysis.

2.2.2 Illumina whole-genome genotyping

2.2.2.1 Illumina Infinium DNA analysis BeadChips

The Infinium® high density array is designed to genotype a large number of SNPs across the whole human genome by interrogating genomic markers through two steps: hybridization of the 50-mer probes to the markers of interest and enzymatic single-base extension with labelled nucleotide on Beadchip (157). Subsequent fluorescent staining and signal intensities can then be detected using Illumina's HiScan/iScan imaging systems. Automated genotype calling can be analyzed using Illumina GenomeStudio Software, where three distinct colour specifies homozygotes and heterozygote for a given marker and individual.

Current analysis used two Beadchips: the HumanOmniExpress and HumanOmni1-Quad Beadchips (Illumina, San Diego, California, USA). The HumanOmniExpress Beadchip genotypes more than 700,000 markers per sample, tagSNPs from this Beadchip are selected from all three phases of the International HapMap Project and are able to capture the greatest amount of common SNP variation. An up to 200,000 markers can be added to the OmniExpress chip to allow researchers to carry out their unique studies. Full content of

the OmniExpress chip is contained within the HumanOmni1-Quad Beadchip, on which markers are derived from the 1,000 Genomes Project and all three HapMap phases for common variation and copy number variation (CNV) discovery by looking at 1,000,000 markers/sample at a time. Both assays consistently produce high average call rates (> 99%) and reproducibility (> 99.9%). All samples were genotyped in Japan. An additional 350 African-American controls were previously genotyped using the HumanOmni1-Quad Beadchip at University of Maryland, USA (published data (158)) and the genotype data was provided to us in order to boost the number of African-American controls for current analysis (i.e. Originally, only 47 African-American controls were genotyped).

2.2.2.2 Illumina GenomeStudio v2010.3/v1.8: genotyping module

Illumina GenomeStudio Software (v2010.3/v.18, Illumina, San Diego, California, USA) is used for data visualization and analysis generated by all Illumina's platforms. The GenomeStudio Software package comprises of seven different application modules for DNA/RNA/CHIP sequencing, genotyping, gene expression, methylation and protein analysis. The results can be easily exported as plain text files for use with a number of third-party software tools for further analysis.

Here, the Genotyping Module was applied to analyze SNP data across hundreds of thousands of markers on the chip and detect sample outliers. The genotype clusters displayed in GenomeStudio with data points marked in three colours (e.g. red=AA, purple=AB, blue=BB), and each data point represents an individual. Genotypes are called for each sample by signal intensity and allele frequency based upon information derived from a standard cluster file provided by Illumina, which uses a representative sample set of over 100 samples from the HapMap CEU (Caucasian), CHB+JPT (Chinese + Japanese) and YRI (Yoruban) populations; and this standard cluster file should represent the genetic diversity well in these populations (for a clustering example, see Figure 2.1). If needed, data quality can be optimized before generating a final report, by performing individual

locus analysis, for example, sorting on call rate or cluster separation, and then do re-clustering.

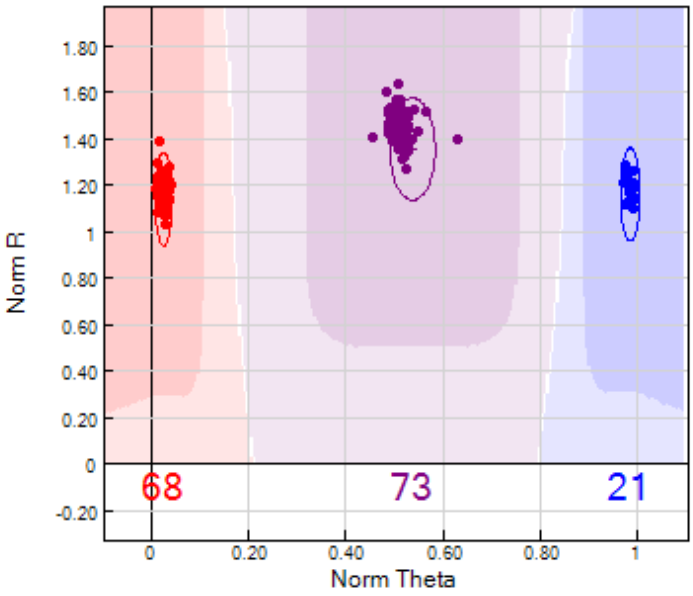


Figure 2.1 Image of genotyping clusters in GenomeStudio. Each data point corresponds to a sample with 3 colour coding (red, purple and blue). Norm R on Y-axis refers to signal intensity and Norm Theta on X-axis refers to allele frequency. Dark shading surrounding each genotype cluster implies the standard cluster position for a given marker. Three figures under each cluster are the number of samples for each genotype group.

2.3 Software

Standard analytical software, such as IBM® SPSS® statistics 19 (SPSS, Inc., Chicago, USA) and R (R Foundation for Statistical Computing, Vienna, Austria), was used for data sorting, visualization and basic statistical runs. Here the focus is to introduce a few other commonly used programmes for analyzing genotypic data for whole genome-wide association analysis.

2.3.1 PLINK

PLINK (159,160) is a free, open-source and command-line based toolset for whole genome wide association analysis. It is used to deal with the large scale genomic data in a computationally efficient manner, requiring all commands to start with typing “plink” at the command prompt followed by other options (e.g. --option) indicating files and methods

involved. PLINK output results are stored as plain text files; the extensions of the files are various and depend on the content of the results.

Main features implemented in PLINK include, basic data management (data recoding, reordering, merging, extracting and DNA-strand flipping), standard summary statistics (e.g. missing genotype rate, MAF, HWE failures), thresholds set-up (based on summary statistics), IBD estimation (identity-by-descent estimate, looking for individuals look too similar in a data set), association tests (for binary and quantitative traits etc.), meta-analysis (combining two or more generically-formatted files, and testing for fixed and random effects models) and other features (e.g. family-based association, permutation procedures, multimarker tests, imputation). For computational efficiency, GWAS analyses reported here were performed on a remote server via the open-source Telnet/SSH client PuTTY (http://www.chiark.greenend.org.uk/~sgtatham/putty/) and files were managed remotely using WinSCP (http://winscp.net/eng/index.php), an open-source SFTP, FTP and SCP client for Windows.

PLINK results files are often large; the genome-wide outputs may subsequently uploaded to other applications, such as R programme, Haploview (161) and EIGENSRAT (97), for data visualization and manipulation.

2.3.2 Haploview

Various functions that Haploview (161) may provide include LD & haplotype block analysis, single SNP & haplotype association tests, permutation tests, haplotype frequency estimation in a population, and PLINK GWAS results visualization as well as advanced filtering options.

In current studies, Haploview was used to take in PLINK outputs – a map file of markers containing information of SNP identifier, chromosome and base-pair position for each

marker and a PLINK association or quality control results file. SNPs in the map file and PLINK results file need not be in the same order, and the map file may contain SNPs that are not present in the PLINK results file. Once uploaded to Haploview, the PLINK results file is displayed as a sortable table with selective filter options based on parameters, for example, association *p*-value and MAF. If required, one can also use "Combine P-Values" and "plot" functions to produce combined *p*-values for 2-5 specified SNPs (note that this does not take into account effect direction using the Fisher's combined algorithm implemented in Haploview) and graphical plots given the uploaded PLINK outputs (the x-axis will always represent chromosomes, while the y-axis can be defined from the drop-down menu for possible parameters to base plots on; parameters refer to the "columns" available from the PLINK results file, such as "the column" of "*p*-values").

2.3.3 EIGENSTRAT

EIGENSTRAT (97) is based on principle component analysis for detection of ancestry differences between cases and controls in GWAS. The correction of population stratification using EIGENSATRAT is specific to a single marker's variation in frequency across ancestral populations along continuous axes of variation. This method can easily handle hundreds of thousands of markers in a GWAS data set. As of December 2006, EIGENSTRAT is included as a part of the EIGENSOFT package (97,103).

Three directories, CONVERTF, POPGENE and EIGENSTRAT are included in the EIGENSOFT package. The convertf programme is used to convert files among five formats (ANCESTRYMAP, EIGENSTRAT, PED, PACKEDPED and PACKEDANCESTRYMAP), the smartpca programme in POPGENE directory is for running PCA, and the EIGENSTRAT method for population stratification correction as well as a smartpca.perl script (to call the smartpca programme in POPGENE) are stored in the EIGENSTRAT directory.

The EIGENSTRAT method focuses on individual genotype data to infer continuous axes of genetic variation and aims to describe major data variation in as few dimensions as possible; then genotypes and phenotypes are continuously adjusted given amounts attributable to ancestry along each axis, by computing residuals of linear regressions; finally, association statistics are calculated based on ancestry-adjusted genotypes and phenotypes (97). The axes of variation (dimensions), explaining maximum data variability, are called principle components (PCs), and these can also be referred to as "eigenvectors (or the largest eigenvalues corresponding with the axes)". The axes are ranked according to the amount of variation represented by the axes.

By simulations, the EIGENSTRAT method is shown to be insensitive to the sample size (EIGENSTRAT effectively corrected for stratification in sample size as low as 100, and up to 1,000) (97). It is also suggested to use the EIGENSTRAT method for at least 100,000 SNPs, because the inclusion of a candidate marker in a smaller set of markers may lead to power loss (97). The authors also tested EIGENSTRAT in an admixed population using simulated data. The association statistics significances were computed using the Armitage trend χ^2 statistic (not corrected for stratification), genomic control and EIGENSTRAT. EIGENSTRAT was reported for achieving higher power than genomic control on correcting population stratification at highly differentiated SNPs as well as the causal SNPs.

The same authors (97) also pointed out that the insensitivity of EIGENSTRAT to the number of axes of variation used as long as a sufficient number of axes are involved to capture the true effect of population structure. By default, the number of axes is set at 10 for running EIGENSTRAT; on the other hand, the number of statistically significant axes of variation may be adopted. Lastly, despite of ancestry effect, assay artefacts, if present, can also be detected by EIGENSTRAT. For example, in the European American data set

explored by Price et al. (2006) (97), the top two axes describe population stratification effects, whereas the third axis explains the subtle differences present in laboratory treatment among samples.

2.4 Statistical analysis

In this section, statistical tests used for candidate gene association analysis in elite Caucasian and East Asian swimmers (Chapter 3, section 3.2.4) and GWAS in elite Jamaican and USA sprint cohort as well as in elite Japanese TF athlete cohort (Chapter 4, section 4.1) are described.

2.4.1 Candidate gene association analysis

Two genes (*ACE* and *ACTN3*) were studied in relation to elite swimmer status in Caucasians and East Asians. Genetic associations were evaluated by multinomial logistic regression, and PTest and MAX3 test were accommodated for multiple testing adjustment to investigate whether polymorphisms in *ACE* and *ACTN3* are associated with elite swimmer status in Caucasian and East Asian populations.

Three genetic models, additive allelic effects and two models assuming complete dominance of each allele in turn, were tested by multinomial logistic regression. To control for multiple testing across genetic models, two further tests (PTest and MAX3) were applied in parallel. PTest is a permutation test tool that generates association χ^2 test *p*-values effectively adjusted for multiple testing (http://rosalind.infj.ulst.ac.uk/Software.html#PTest; ref (162)). MAX3 implements an efficiency robust trend test implemented in the R Package Rassoc (163). The 'boot' method was used to compute simulation-derived empirical *p*-values based on data resampling to generate the null distribution for the test statistic, and inherently adjusted for multiple testing of three genetic models. In addition, the effect of ethnic subdivision within

the East Asian cohort was evaluated by including a genotype x ethnicity interaction term in the multinomial logistic regression model and assessing significance using a likelihood ratio test.

Analyses were carried out using IBM® SPSS® Statistics 19 software (SPSS, Inc., Chicago, USA), and R (R Foundation for Statistical Computing, Vienna, Austria). *p*-values for significance (α values) were defined as following: no adjustment was carried out for testing of the two genes, because the published literature supported a prior hypothesis that we would find association in each case; stratifying by ethnic group was then adjusted for and the α value for significance of the multinomial logistic regression test/permutation test in each ethnic subgroup was considered at $p < 0.025$; further multiple testing after stratification into two event distance groups (vs. controls) was handled by adjusting further for these pairwise comparisons, and the α value for significance of the pairwise logistic regression/permutation tests was therefore defined as $p < 0.0125$. Specific details of statistical analysis for the candidate gene association study are described in "Methods" of Chapter 3 (section 3.2.4).

2.4.2 Analysis of GWAS

2.4.2.1 Power calculations

Power for GWAS was estimated under multiplicative, additive, dominant and recessive models using CaTS Power Calculator version 0.0.2 (http://www.sph.umich.edu/csg/abecasis/cats/; ref(164)). CaTS is specifically designed for power calculations in large genetic association studies. For simulating a similar magnitude relative to current GWAS in elite athletes, 200 samples (half cases and half controls) were used here for power estimation. The relationship between power and effect sizes (or the odds ratios) for 100 cases and 100 controls, with a range of MAF varied from 0.05 to 0.5, assuming low prevalence of the trait at 0.1, was explored (Figure 2.2). It should be noted

that this estimation gives a rough idea of the power possibly achieved given varying MAFs and effect sizes as well as a pre-defined sample size and prevalence of the trait. Sensible judgement on power gained should be made accordingly upon one actual study.

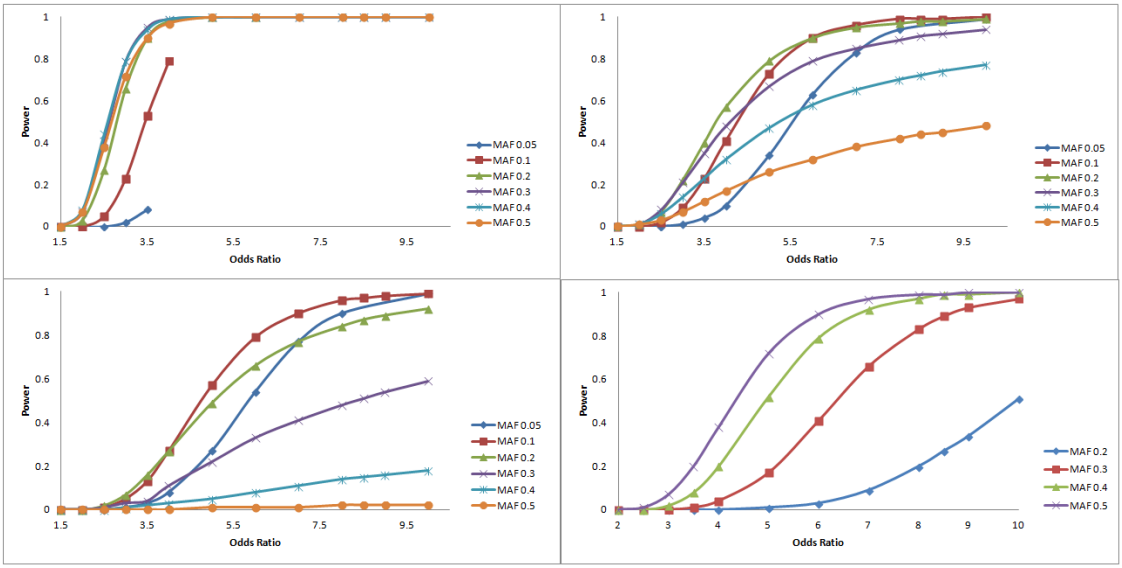


Figure 2.2 Power vs. effect size for 100 cases and 100 controls under the multiplicative (top left), additive (top right), dominant (bottom left) and recessive (bottom right) models, assuming a low prevalence of the trait at 0.1 for MAF varied from 0.05 to 0.5.

The simulated data showed the failure of detecting any association, when underlying genetic effect is smaller than and equal to 1.5, under any of the genetic models examined here in 200 samples, although the allele of a variant may be frequently present in the study samples. This is in line with previous reports showing typical significant association findings from GWAS of complex traits with an odds ratio ranging from 1.1 to 1.4 would require thousands of samples for obtaining power at a reasonable level (75). For 80% power, minimum effect sizes are required to be ~ 3.02, 5.04, 6.07 and 5.36 with the corresponded MAF of 0.3/0.4, 0.2, 0.1 and 0.5 under the multiplicative, additive, dominant and recessive models, respectively.

2.4.2.2 Formats converting from Illumina output files to PLINK formats

Using Illumina GenomeStudio software, genotype data was exported as plain text files, including a final report, a sample and a map files. The Illumina final report contains 10

header rows of descriptive information, and the data following the headers is represented as a row for each individual for each marker. The separate sample and map files (for samples' and markers' annotations, respectively) can be created in order to reduce the size of the final report. There are a number of fields can be included to the final report file, however only sample ID, SNP name, allele 1, and allele 2 are used for generating PLINK LGEN file. Because the samples involved in current studies are unrelated individuals, no family IDs are available. The family ID was therefore duplicated with the individual ID (or the sample ID) and added as a fifth column to meet PLINK LGEN format requirement. All header rows need to be removed when converting from Illumina output files (final report, sample and SNP files) to PLINK long-format filesets (LGEN, MAP – listing marker information and FAM – listing individual information). The long-format filesets can then be transformed to PED or BED format for downstream data analysis in PLINK.

The PED file can be a space, white-space, or tab delimited file, with first six columns (mandatory) specifying family ID, individual ID, paternal ID, maternal ID, sex and phenotype, followed by genotypes from column 7 onwards (markers must be biallelic). In current studies, family ID and individual ID were replaced with sample ID. Paternal and maternal IDs were coded as 0 for the unrelated individuals studied. The accompanied MAP file must contain the same set of markers (one row per marker) as in the PED file, and the order of markers in the MAP file should align with the order of the PED file markers. Three columns in the MAP file are expected, including chromosome, SNP identifier and base-pair position.

Another PLINK file format, binary PED files (the BED files) can be created, comprised of the binary file (*.bed; genotype information), a separate pedigree/phenotype FAM file (*.fam; first 6 columns of the PED file) and an extended MAP file (*.bim; with two extra columns containing information of the allele names). The BED fileset is smaller than the

PED files, and easier for analysis manipulation. The BED files were used in current studies.

2.4.2.3 Quality control

The basic quality control (QC) steps include sex check, missingness check, and examinations of marker MAFs and markers deviated from HWE.

X-chromosome data is used to determine sex based on X chromosome heterozygosity estimate, this information is compared to the recorded sex from the FAM/PED file. Individuals showing mis-matched sex information should be excluded unless a typo of the recorded sex or other human errors can be identified.

In the missingness analysis, two output files are generated: *.imiss and *.lmiss, implying missingness by individual and by locus, respectively. For each sample, proportion of missing SNPs for this sample is recorded by the column of "F_MISS" in the *.imiss file. For each SNP, proportion of sample missing for this SNP is indicated under the column of "F_MISS" in the *.lmiss file.

MAF for each SNP can be produced using the allele frequency command. SNPs that fail the Hardy-Weinberg test can be identified by specifying a significance threshold using the --hwe filter; by default, for family-based data, this test is based on founders; for case-control data, markers excluded given HWE is for controls only (165).

2.4.2.4 Detection of cryptic relatedness and population stratification

Cryptic relatedness may exist among seemingly unrelated individuals, it is important to rule out such confounding effect on a GWAS. PLINK allows to calculate genome-wide IBD given identity-by-state (IBS) and allele frequencies for every pair of individuals using genome-wide data (ideally, 100,000 independent SNPs or more). PLINK "--genome"

command creates a file called *.genome, in which several features are reported. Among them, fields of Z0 and Z1 refer to the proportion of markers identical by descent 0 and 1, respectively. Additionally, by adding another flag "--min 0.05", only the individual pairs showing high levels of IBD sharing (i.e. pairs with proportion IBD > 0.05) are displayed as a result in *.genome, and these pairs are of particular interest.

The relationship between Z0 and Z1 for every pair can be visualized using R programme (i.e. plot Z1 on y-axis against Z0 on x-axis), and each data point can be colour coded based on the relationship type from the PED file. The R code was provided at http://gettinggeneticsdone.blogspot.co.uk/2009/10/visualizing-sample-relatedness-in-gwas.html. Parent-offspring pairs share 100% of their alleles IBD=1. If a pair shows P(IBD=0) = 0 and P(IBD=1) = 0, this means this pair shares two alleles identical by descent at every locus across the genome, implying they are either duplicated samples or identical twins. Unrelated individual pairs are expected to show up at bottom right quadrant of the plot.

EIGENSTRAT was used to detect population stratification (i.e. to identify outliers lying away from the main population cluster of interest in a PCA plot) and subsequently to correct for it using eigenvectors as covariates in association analysis if needed. Prior to EIGENSTRAT, the study samples (i.e. cases + controls) were merged with a subset of individuals from the HapMap3, including 112 CEU; 84 CHB; 86 JPT; 113 YRI and 88 TSI, to allow for comparison with densely genotyped ancestral population groups.

2.4.2.5 Association tests

PLINK provides basic association tests for detecting an association between a variant and case/control status, via comparing allele frequencies. The asymptotic *p*-value for this test is usually reported. Fisher's Exact test can be called to generate exact *p*-value for significance. Alternative association tests (rather than these basic allelic tests), such as

Cochran-Armitage trend test, genotypic test (2 d.f.) and dominant/recessive gene action test (1 d.f.), can be also applied. Another two tests, the linear and logistic regressions allow for multiple covariates as well as their interactions to be included to the tests' models for association analyses. The covariates can be continuous or binary. For linear regression, the regression coefficient is returned; and for logistic regression, odds ratio is reported. The linear and logistic regression tests are more flexible than the basic association tests described above in terms of specifying covariates and interactions, genetic models and joint tests (e.g. jointly test a main effect and interaction effect against the null hypothesis). It is also important to be aware that the basic allelic tests assume Hardy-Weinberg equilibrium, SNPs show severe deviations from HWE in controls may reflect genotyping or stratification issue in a sample set, therefore we should be cautious with such SNPs and exclude them from following analyses.

For current studies, the standard case-control allelic association test using asymptotic *p*-value was run in PLINK by comparing allele frequencies between cases and controls. SNPs and individuals must meet following inclusion criterion: maximal proportion of missing SNPs per sample 0.05; minimum minor allele frequency 0.01; minimum Hardy-Weinberg disequilibrium frequency *p*-value 1×10^{-7}; maximum proportion of missing samples per SNP 0.05. Individuals with discordant sex (as identified by the QC filter of sex check), related individuals and individuals who were outliers when population stratification was assessed were also removed. Two arbitrary genome-wide significance thresholds were set up at $p < 5 \times 10^{-5}$ and 5×10^{-6}, and the threshold of 5×10^{-5} is used for prioritizing a set of SNPs to be taken forward into further investigations.

Adjustments for significance values were also done, such that genomic control corrected *p*-values and Bonferroni adjusted *p*-values amongst other parameters were produced. Logistic regression analysis was conducted, where population stratification or other

adjustments required, by entering relevant variables as covariates into a logistic regression model.

2.4.2.6 Meta-analysis

Meta-analyses were performed for elite Jamaican sprint, African-American sprint and Japanese sprint GWAS cohorts. The combined effects from SNPs with unadjusted association $p < 5x10^{-5}$ across these populations were computed using the meta-analysis option implemented in PLINK. The odds ratio and standard error of each SNP for each study was entered into analysis. SNPs across input files for meta-analysis need neither be in the same order nor featured in all files; only SNPs present in two or more files are reported by default. Both fixed- and random-effects odds ratios and *p*-values of the common SNPs across multiple input files are stored in *.meta PLINK output file, which also contains the fields of Cochran's Q test *p*-value and the I^2 index for assessing between-study heterogeneity.

Both within-study variability and between-study variability account for heterogeneity in a meta-analysis (166). The former is due to sampling error that would always present since each single study uses different samples. The latter is caused by varying characteristics among a set of studies (e.g. sample characteristics, variations in study design and quality), and this is known as the true heterogeneity (among the population effect sizes estimated by the individual studies) over sampling errors. The extent of true heterogeneity (or between-study variability) can be measured by I^2 index, which is calculated by dividing the difference between the Q test value and its degree of freedom (*k*-1) by the Q statistic itself, where *k* refers to the number of studies. The Cochran's Q test computes the Q statistic by summing the squared deviations of each single study effect estimate from the average effect estimate overall, and the contribution of each study is weighted by its inverse variance. The Q statistic follows a chi-square distribution with *k*-1 degree of freedom under

the null hypothesis of homogeneity (for the effect sizes). The power of Q statistic is sensitive to the number of samples. For example, this test has low power to detect true heterogeneity when a small number of studies included in the meta-analysis, but enlarges any negligible variability when the number of study is large. Additionally, the Q test does not tell the extent of true heterogeneity, but its statistical significance. Instead, the I^2 index as defined above can be interpreted as the percentage of the total variability among the effect sizes owing to true heterogeneity, therefore, it is treated as an assessment for the extent of true heterogeneity.

The fixed-effects model is usually adopted when only the within-study variability exists or the estimated effect sizes only differ because of sampling error (166). The random-effects model should otherwise be used to take into account the between-study variability. Finally, it should be noted – although I^2 can be used to infer the extent of true heterogeneity, it also suffers the same low power problem as the Q test and should be interpreted with caution with a small number of studies ($k < 20$) on deciding which statistical model (fixed- or random-effects model) to use in a meta-analysis (166).

2.4.2.7 Annotation

Regional association plots of the top signals (unadjusted $p < 5 \times 10^{-5}$) were created using LocusZoom Version 1.1 (http://csg.sph.umich.edu/locuszoom/; ref (167)), which is designed to view the local association results with information of the location and orientation of the genes, recombination rates and LD coefficients. GWAS summary results can be uploaded to the web-based form of LocusZoom to create a plot of region of interest at a time or can be submitted using batch mode to create multiple plots in one go. Each regional association plot was specified by the SNP of interest (i.e. the index SNP), which is treated as the key marker representing for that region. 500kb flanking region on each side of the index SNP was specified. Plots were generated based on Human Genome Build 19

(hg19). LD levels between the index SNP and its surrounding SNPs, as well as recombination rates, were estimated using samples from 1,000 Genomes Project (the version released in March 2012) as the reference populations.

The Single Nucleotide Polymorphism database (dbSNP) collects simple genetic polymorphisms, including SNPs, deletion insertion polymorphisms, retroposable element insertions and microsatellite repeat variations. Information for such variants contains the sequence context, frequency of the SNP (at population or individual level), and relevant experimental materials (e.g. methods and protocols) (168). The current dbSNP build for human is 137 based on GRCh37.p5 ("p" stands for patch that corrects sequence or adds sequence in a major release). OMIM, the Online Medelian Inheritance in Man (169), contains all known mendelian disorders and focuses on the relationship between genotype and phenotype. The database is updated daily and now has > 21,000 gene entries (assessed on 3rd January, 2013). The precursor of OMIM was the book edition archiving for mendelian traits and disorders, published between 1966 and 1998, called the Mendelian Inheritance in Man (MIM). The online version was created in 1985 and made available since 1987. Main information from OMIM includes gene phenotype relationships/correlations, cloning, gene structure and function, molecular genetics, animal models and allelic variants (mutations that are disease-causing). Both dbSNP and OMIM are hosted by the National Center for Biotechnology Information (NCBI) that provides access to biomedical and genomic information through a variety of housed databases. All these databases can be browsed through the Entrez cross-database search system in NCBI.

2.4.2.8 Genotype score analysis in addition to common variants for prediction of elite performance

Genotype score was constructed on the basis of the number of risk alleles inferred from previously identified sprint performance-related SNPs (primarily derived from candidate gene association studies), assuming an additive effect (Chapter 4, section 4.2). Receiver

Operating Characteristic (ROC) curve was used to interpret sensitivity and specificity levels of the genotype score approach in distinguishing elite sprint athletes from endurance athletes and/or controls. The area under the curve (AUC) and 95% C.I. were calculated for the overall diagnostic accuracy of a ROC curve. It aims to investigate whether SNPs identified from published reports show predictive utility on elite sprint performance, assessed by using the genomic data generated from current GWASs.

2.4.2.9 Flow of current and perspective studies

Before biological validation and clinical translation can be carried out, multiple stages are needed for the discovery and replication of associated genetic markers. A general flow from GWAS to clinical translation is elucidated in Figure 2.3. The discovery GWASs of elite performance in Jamaican, African-American and Japanese populations would reveal a number of potential associations. Several steps below may be used to protect against false positives:

- To apply a stringent GWAS significance threshold (i.e. 5×10^{-8}) to select markers to follow up for formal replication;

- And/or, to perform meta-analysis across several individual GWASs, thus, to improve power by increasing sample size;

- And/or, to adopt a less stringent threshold by analyzing SNPs in a predictive genomic score for subsequent studies (such permissive early thresholds may minimize false negative reports).

Details for the initial GWASs, which were carried out at the discovery stage, and marker annotation and selection for replication are described in Chapter 4. Fine mapping/sequencing studies could then be performed for the identification of risk

enhancing alleles across identified GWAS loci, and functional studies can help understand more of the biological basis of athleticism. Such findings are also likely to have implications on clinical research related to public health as genetic variants affecting elite performance are also expected to have impact on cardiac and skeletal muscle functions (53).

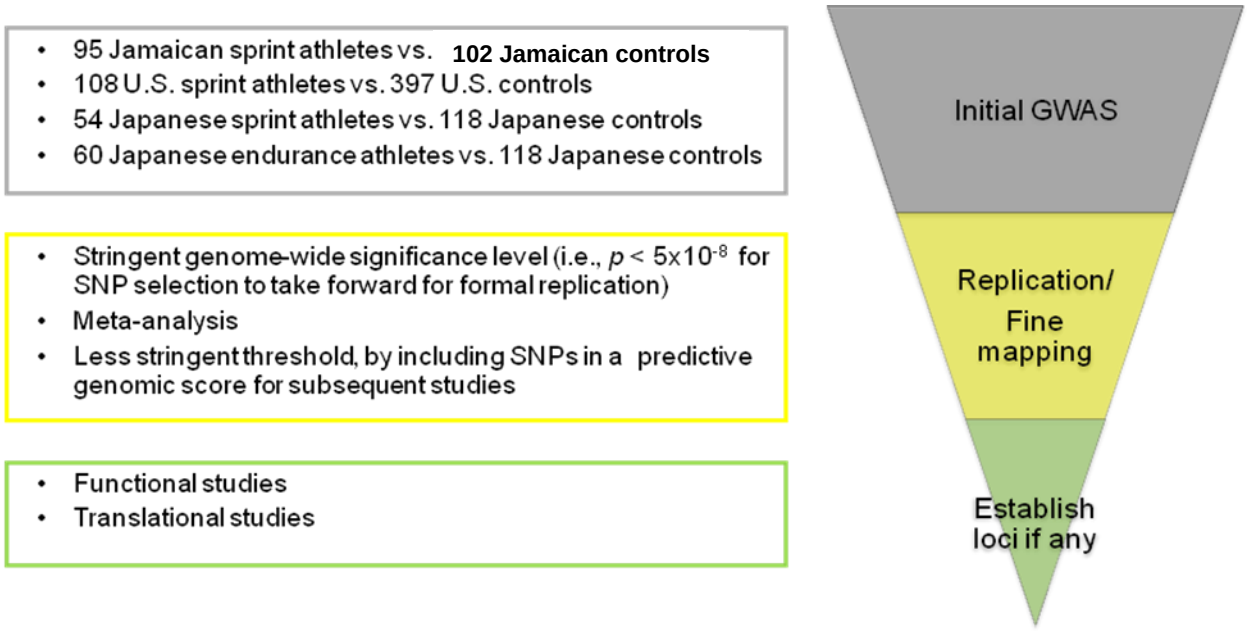


Figure 2.3 A summary of current and perspective studies.

3 Candidate gene association study in elite swimmers

In this chapter, the relationship between the two genes (*ACE* and *ACTN3*) and the elite swimmer status was explored in both Caucasian and East Asian populations. This work (170) has recently been published in the American College of Sports Medicine (ACSM) monthly journal - the Medicine & Science in Sports & Exercise (MSSE). The hardcopy publication is scheduled for the May 2013. This published paper is adapted into a thesis chapter and referenced where appropriate.

3.1 Introduction

Numerous candidate gene association studies have been carried out to identify genes related to elite human performance (see section 1.2.3), however, genetic contributions to high level performance in swimming have received little attention (148,171). Therefore, a candidate gene study was carried out to examine two genes encoding angiotensin converting enzyme (*ACE)* and actinin, alpha 3 (*ACTN3)* in relation to elite swimmer status in Caucasians and East Asians, respectively.

Variants in both *ACE* and *ACTN3* have been reported to be associated with elite athletic performance, and with normal, quantitative physical performance traits in the general population. Angiotensin converting enzyme plays a critical role in circulatory homeostasis as a component of the circulating renin-angiotensin system (RAS), catalysing the conversion of angiotensin I to the vasoconstrictor angiotensin II and the degradation of the vasodilator bradykinin. However, local (tissue or cellular) RAS in a variety of tissues subserve diverse roles, including the regulation of inflammation, cell growth, and aspects of metabolism (172). A 287bp *Alu* repeat insertion/deletion (I/D) polymorphism (rs4340) in intron 16 is associated with circulating and tissue ACE levels, with higher ACE activity

being associated with the D (deletion) variant in both Caucasians (173) and East Asians (174). In contrast, in populations of sub-Saharan African descent, the I/D polymorphism is associated with ACE activity to a considerably lower extent, reflecting the LD structure across the gene and the fact that the I/D variant is not thought to be the functional variant affecting ACE activity (175).

ACE I/D is associated with a variety of exercise-related phenotypes, including sporting performance (176), fatigue resistance in response to physical training, the cardiac growth response, differences in muscle efficiency and strength, hypoxic ventilatory drive, and skeletal muscle fibre distribution (reviewed in (172)). It has also been suggested, at least in part, that the associations between *ACE* and performance-related phenotypes are mediated through bradykinin, whose actions are mediated by the bradykinin receptor β_2 (*BDKRB2*) gene (177). Bradykinin, through *BDKRB2*, may lead to the increase of glucose uptake and GLUT-4 translocation in skeletal muscle in response to exercise (178), as well as lowered respiration in skeletal muscle and heart mitochondria (179,180) via the production of nitric oxide that inhibits cytochrome-c oxidase (the mitochondrial complex IV) (179). The *ACE* I/D polymorphism has been previously shown to modulate levels of bradykinin, with *ACE* I-allele associated with higher kinin activity (181). A haplotype analysis between *ACE* I/D and *BDKRB2* -9/+9 (a 9-base pair repeat in exon 1 of the *BDKRB2* gene, related to higher mRNA expression of the receptor) has found that the *ACE* I/*BDKRB2* -9 haplotype was significantly associated with higher skeletal muscle efficiency and endurance running performance in Caucasians, suggesting that the associations between *ACE* and human physical performance may in part be due to the elevation of kinin activity (177). In Caucasian populations, the I-allele has previously been reported to be associated with enhanced elite endurance performance in long-distance runners and rowers, and with enhanced performance at high altitude (172), all activities requiring endurance capabilities; the D-allele, on the other hand, has been reported to be associated with strength/power

sports, such as sprinting (182) and swimming events of ≤ 400 m (148,171). It should be noted that the discrepancies exist between studies of muscle strength and size in relation to *ACE* I/D genotype, for example, with gains of muscle strength in the I-allele carriers (e.g. (183)), the D-allele carriers (e.g. (184,185)) or no association (e.g. (186)). It might be explained as the result of the competing effect of higher ACE activity and production of angiotensin II (a cellular growth factor), associated with the D-allele, on muscle growth, and enhanced muscular contraction efficiency mediated through the I-allele in association with bradykinin activity (172,177,185). Data from populations of East Asian descent, however, have revealed conflicting results relative to the above *ACE* association findings in Caucasians, the D-allele being associated with elite Japanese long distance runner status (187) and the I-allele with elite Korean power-oriented athlete status (188). Another 4 studies of *ACE* I/D in sport have also been carried out in East and Southeast Asian populations (189-192), but the results of these studies are much less reliable due to very small sample sizes (ranging from 17 to 108) and/or multiple-testing problems that have not been properly dealt with. Both factors would result in inflated type I error rate.

ACTN3 encodes α-actinin-3, an actin-binding protein with a structural role at the sarcomeric Z-line in glycolytic (type II, fast-twitch) muscle fibres and an increasingly evident role in the regulation of muscle metabolism (reviewed in (193)). A common nonsense polymorphism, p.R577X (located in exon 16 of the gene), exists in many human populations. The 577X-allele is a protein-null allele, from which no ACTN3 is produced, so that XX homozygotes do not express ACTN3 at all in their muscles (193). In the knockout mouse, it is clear that Actn3 deficiency alters skeletal muscle function (193). The 577X-allele is found worldwide but at widely differing frequencies in different populations (193). Associations have been reported between R577X and physical performance both in elite athletes and in the general population, with the 577R-allele being associated with increased sprint performance (149,194). The 577XX null genotype has been reported to be

found at a reduced frequency in elite Australian Caucasian and Finnish sprinters and other sprint/power athletes (149,195). The 577X-allele is found at very low frequency in sub-Saharan Africa (193), and, in line with this, associations with sprint or power athletic status in Nigerians, Jamaicans and African-Americans (154,196), or with endurance athletic status in East Africans (196), have not been found. Only 2 studies of *ACTN3* polymorphism in elite East Asian athletes have been found. Among them, the first study was carried out in Chinese male and female endurance athletes, however, the findings are subject to high probability of type I errors (197), and the second study examined the role of *ACTN3* polymorphism in Taiwanese swimmers, the reason that this paper cannot be cited as an independent evidence for a genetic association is that the individuals used in this reference are a subset of those included in current bigger study population in order to boost overall sample numbers, therefore, it was rather cited purely for the purpose of stating the details of the genotyping assay (see section 3.2.3.3).

The aims of the study as indicated at the beginning of the chapter were two-fold:

- To explore further the associations of *ACE* and *ACTN3* genotype with elite swimmer status;

And,

- To investigate whether such associations differed by swimming event distance or by ethnicity, focusing on Caucasian and East Asian populations.

3.2 Methods

3.2.1 Subjects

Two elite swimmer cohorts, comprising Caucasian and East Asian subjects, respectively, were studied with the approval of the respective local ethics committees (the Sports

Studies Ethics Committee (SSEC) at the University of Stirling, Scotland; the Institutional Review Board of Tokyo Metropolitan Institute of Gerontology, National Institute of Health and Nutrition, Japan; and the Institutional Review Board of Chang Gung Memorial Hospital, Taiwan). Written informed consent was obtained from all subjects. Parental consent was sought for subjects under 16 years of age in both cohorts.

3.2.1.1 Caucasians

A total of 200 elite Caucasian swimmers from European, Commonwealth, American and Russian sub-cohorts were sampled during swimming competitions during 2005 and 2006 and categorized as short and middle distance (SMD ≤ 400 m, n = 130) or long distance swimmers (LD > 400 m, n = 70) (Table 3.1). Distances of 400 m and below have been previously used to study swimmers excelling in a shorter swimming duration (148). Competitive swimmers are generally unable to excel (i.e. win world-class competitions) in events in both short-distance and longer-distance categories, but several swimmers had taken part in events spanning this 400 m cutoff. Caucasian swimmers were classified as LD if their event range included distances of 500 m and above, or if they were described as competing in "Middle and Distance" or "Distance" events; all other swimmers in the Caucasian sample were classified as SMD. For the European, Commonwealth and U.S. sub-cohorts, the swimmers were of world-class status or highly competitive in international competitions (lifetime World Rankings in the top 50, averaging within the top 20; swimming World Rankings can be accessed through http://www.fina.org/H2O/). For the Russian sub-cohort (n = 21) all had represented their country in international competitions at very long distances (5 – 25 km), and many were World Champions or World Championship prizewinners. Controls for this cohort comprised individuals of known genotype from the general population reported in previous studies (*ACE*-C: n = 1248, (148); *ACTN3*-C: n = 1694, (149-153)). Differences in *ACTN3* R577X allele frequency between East and West Europe were dealt with by including Russian controls in the same

proportion as Russians made up in the swimmer cohort (21 Russians in a total of 193 genotyped swimmers = 10.9%) to minimize any potential stratification effects. Thus, 184 Russian controls were randomly selected for inclusion in the analysis.

3.2.1.2 East Asians

Elite Japanese (n = 158) and Taiwanese (n = 168) swimmers were recruited and classified as short distance (SD; n = 166) if their best event in competition was below 200 m and middle distance (MD; n = 160) if their best event was at 200 m or 400 m (Table 3.1). None of these swimmers excelled at distances greater than 400m. All had either participated in international competitions such as the Olympics, World Championships and Asian Games, or were participants in national competitions. Controls for this group came from two sources - Japanese controls were recruited for this study from the general population in Tokyo and its environs (n = 649); Taiwanese controls were a randomly selected subset (n = 603) of a larger cohort (n = 3000) recruited from the general Taiwanese population, as previously described (198). All controls were healthy adults of both sexes and were not professionally connected with athletics/sport. These Japanese and Taiwanese subgroups were combined in the analysis as a single control group except in models testing ethnicity x genotype interactions (see below). It should be noted that the Taiwanese samples and data for *ACTN3* have been previously published (see section 3.2.3.3; (199)) and they were included in the current study to boost overall sample numbers.

Table 3.1 Total numbers of elite Caucasian and East Asian swimmers recruited in this study.

Cohort	Event	Male	Female	Total
Caucasians	SMD (≤ 400 m)	74	56	130
	LD (> 400 m)	42	28	70
East Asians	SD (≤100 m)	101	65	166
	MD (200 – 400 m)	95	65	160

The Caucasian and East Asian swimmers reported here include all elite swimmers used for genotyping. The number of Caucasian swimmers differed from those included in the analysis as some samples were not successfully genotyped.

3.2.2 DNA collection/extraction/quantification

3.2.2.1 Caucasians

Sample collection, DNA extraction and quantification for Caucasians have been previously described in 2.1.2.

3.2.2.2 East Asians

East Asians swimmers and controls were collected, extracted and quantified by Japanese and Taiwanese collaborators in Japan and Taiwan, respectively. For the Japanese swimmers and controls, genomic DNA was isolated from either 7 ml venous blood or 2 ml saliva using QIAamp® DNA Blood Mini or Maxi Kits (QIAgen, Hilden, Germany) or Oragene® DNA Self-Collection Kit (DNA Genotek Inc., Ottawa, Ontario, Canada). DNA was then quantified using either a Nanodrop or a GeneQuant Pro (Amersham Biosciences, Amersham, UK) Spectrophotometer. For the Taiwanese swimmers and controls, 5 ml venous blood was collected into heparinized tubes (Vacutainer). The whole blood was centrifuged within 24 hours, and buffy coat cells stored at -70 °C until extraction of genomic DNA as previously described (200).

3.2.3 Genotyping

3.2.3.1 TaqMan SNP genotyping method

Genotypes were determined using TaqMan® assays (Applied Biosystems, Warrington, UK; Applied Biosystems, CA, USA). For the Caucasian swimmers, and for the Japanese swimmers and controls, genotypes were obtained at *ACE* SNP rs4341 (ABI assay ID: C__29403047_10) and at *ACTN3* p.R577X (rs1815739; ABI assay ID: C____590093_1_). rs4341 is known to be in perfect LD with I/D (rs4340) in Caucasian and Asian populations (201,202). For Caucasian swimmers (as described in section 2.2.1.2), amplifications were carried out in 20 μl reactions containing 10 μl universal master mix, 1.0 μl ABI assay mix (20 ×), 6 μl distilled water and 9 ng genomic DNA. For Japanese subjects, amplifications

were carried out in 5 μl reactions containing 2.5 μl Taqman® GTXpress™ master mix, 0.125 μl ABI assay mix (40 ×), 1.375 μl sterile water and 10 ng genomic DNA. Amplifications were carried out using ABI's StepOnePlus™ Real Time PCR system (Applied Biosystems, CA, USA). Genotypes were called from end-point reads using ABI's StepOne™ Software v2.1. *ACE* I/D genotypes were calculated from rs4341 genotypes as follows: rs4341 G/G was called as D/D; C/G was called as I/D; C/C was called as I/I.

3.2.3.2 Allele discriminatory PCR method

Genotyping of *ACE* I/D (rs4340) in Taiwanese swimmers and controls was performed using a standard gel-based allelic discrimination assay method as previously described (203). The PCR primers for *ACE* I/D were Forward: 5′-CTGGAGACCACTCCCATCCTTTCT-3′ and Reverse: 5′-GATGTGGCCATCACATTCGTCAGAT-3′. The PCR was performed in a 25 μl reaction using a Mastercycler gradient thermocycler (Eppendorf, Hamburg, Germany). The PCR constituents were 100 ng genomic DNA, 3.5 mm $MgCl_2$, 200 μM dNTPs, 1 unit of Taq polymerase, and 400 nM of each primer in 1× PCR buffer for 35 cycles under the following conditions: 95 °C for 1 min, 58 °C for 30 s, and 72 °C for 40 s. PCR products were electrophoresed through an 8% polyacrylamide gel, stained with ethidium bromide, and photographed under UV light. The I- and D-alleles yielded fragments of approx. 480 bp and 190 bp, respectively. Because amplification of the I-allele can be suppressed in ID heterozygotes, resulting in allelic dropout and miscalling of heterozygotes as DD homozygotes, all samples classified as DD genotype were subjected to a second PCR using an I-allele-specific primer pair: Forward: 5′-TGGGACCACAGCGCCCGCCACTAC-3′ and Reverse: 5′-TCGCCAGCCCTCCCATGCCCATAA-3′ (185) (using 30 PCR cycles of 1 min at 95 °C, 40 s at 67 °C, and 2 min at 72 °C). Products were detected by 6% polyacrylamide gel electrophoresis. A 335 bp fragment indicated the presence of the I-

allele, and samples positive for both this 335 bp fragment and the 190 bp fragment in the first PCR were called as ID heterozygotes.

3.2.3.3 PCR-RFLP genotyping

Genotyping of Taiwanese swimmers and controls at *ACTN3* R577X was carried out after PCR amplification across the polymorphic site and restriction digestion, as previously described (199).

3.2.4 Statistical analysis

Genotype and allele frequencies were calculated for both *ACE* and *ACTN3* polymorphisms and Hardy-Weinberg equilibrium (HWE) assessed using a χ^2 test. Three separate tests were performed to investigate associations between genotype and swimmer (case/control) status. In the first test, multinomial logistic regression tests were applied, with three outcome states were used in the models, analyzing the regional cohorts separately. For Caucasians, the outcome states were SMD swimmer, LD swimmer and control. For East Asians, the outcome states were SD swimmer, MD swimmer and control. Associations of genotype with outcome were modeled using three genetic models - additive allelic effects and two models assuming complete dominance of each allele in turn. To control for multiple testing across genetic models, two further tests were applied in parallel, PTest and MAX3. PTest (http://rosalind.infj.ulst.ac.uk/Software.html#PTest; ref. (162)) is a permutation test tool that generates association χ^2 test *p*-values effectively adjusted for multiple testing of, in this case, the three genetic models being examined while maintaining an experiment-wide type I error rate of 0.05. For each calculation, 99,999 permutations were computed (Appendix A1.1 for permutation input and output files; (170)). Separate tests were run for each regional cohort at each gene tested. MAX3 implements an efficiency robust trend test implemented in the R Package Rassoc (163). We used the 'boot' option, in which the program reports simulation-derived empirical *p*-values based on data resampling to

generate the null distribution for the test statistic, which adjusts for the inherent multiple testing of three genetic models. Additionally, the effect of ethnic subdivision within the East Asian cohort was evaluated by including a genotype x ethnicity interaction term in the multinomial logistic regression models and assessing significance using a likelihood ratio test. Analyses were carried out using IBM® SPSS® Statistics 19 software (SPSS, Inc., Chicago, USA), and R (R Foundation for Statistical Computing, Vienna, Austria). *p*-values for significance (α values) were defined in relation to the number of tests done in the following way: no adjustment was carried out for testing two genes, as the published literature supported a prior hypothesis that we would find association in each case; stratification by event distance and the inherent multiple testing that thus arose due to having more than two 'outcome' categories was handled via the use of a single, multinomial logistic regression test and a permutation-based test; stratifying by ethnic group was explicitly adjusted for; thus the α value for significance of the multinomial logistic regression test/permutation test in each ethnic subgroup was $p < 0.025$; further multiple testing after stratification into event distance groups was handled by adjusting further for the two pairwise comparisons (i.e. SMD vs control, and LD vs control), thus the α value for significance of the pairwise logistic regression/permutation tests was $p < 0.0125$.

3.3 Results

In the Caucasian cohort, genotype data were available for 191 cases (swimmers) and 1248 controls for *ACE*, and 193 cases and 1694 controls for *ACTN3*. For East Asians, data were available for 326 cases and 1244 controls for *ACE*, and 326 cases and 1252 controls for *ACTN3*. Both polymorphisms were in HWE in both cases and controls for both Caucasian and East Asian cohorts (Appendix Tables A1.2 and A1.3; (170)). In addition, *ACE* and *ACTN3* genotype frequencies in the two East Asian ethnic sub-cohorts have also been provided for both swimmers and controls (Appendix Tables A1.4 and A1.5; (170)). Allele

frequencies for *ACE* I/D (as measured using rs4341 in Caucasian swimmers and Japanese swimmers and controls) in the control groups were consistent with the published literature, with the I-allele being at relatively higher frequency in East Asians (187,188). For the Caucasian cohort, *ACE* I/D control group was based on a UK study in which the D-allele frequency was 0.51. The average D-allele frequency across Europe (148,204,205), Australia (206) and the U.S. (207) is 0.52, and it was therefore concluded that effects of population stratification due to use of swimmers from several separate populations of European origin are unlikely to be large. Allele frequencies for *ACTN3* showed only small differences between the regional subgroups and were in line with expectation (149-153,194,197). The *ACTN3* data in Caucasian controls were obtained from 5 separate studies. There were no differences in allele frequencies or genotype distributions between these studies (genotype comparisons: $\chi^2 = 6.03$, $p = 0.64$; Appendix Table A1.6; (170)).

ACE I/D genotype was associated with elite swimmer status in Caucasians. The multinomial logistic regression models were significant (Table 3.2; (170); $p = 0.017$ for the additive model and $p = 0.005$ for the I-allele-dominant model;). This association was mediated by effects in SMD swimmers (Figure 3.1, Table 3.2; (170)), with the largest effect size observed for the I-allele-dominant model (D-allele homozygotes vs. I-allele carriers: odds ratio = 1.90; logistic regression $p = 0.001$; permutation test $p = 0.0005$; MAX3 test statistic = 3.37, $p = 0.0017$), with the D-allele being over-represented in the swimmers. The D-allele is associated, in recessive fashion, with elite SMD swimmer status in Caucasians. No significant association was found between the *ACE* I/D polymorphism and Caucasian LD swimmer status (Figure 3.1, Table 3.2; (170)).

Table 3.2 Multinomial logistic regression and other analyses of associations between ACE and ACTN3 polymorphisms and elite Caucasian and East Asian swimmer status.

Gene	Cohort	Group	Risk allele[#]	Additive Model					Dominant Model[†]					
				L.R. Model *p*[&]	PT model *p**	O.R.[$] (95% C.I.)	pairwise *p*[¶]	PT pairwise *p*[§]	Dom. allele	L.R. Model *p*	PT model *p*	O.R.[††] (95% C.I.)	pairwise *p*	PT pairwise *p*
ACE	Caucasians	SMD	D	0.017	0.021	1.46 (1.12 - 1.90)	0.005	0.003	I	0.005	0.0033	1.90 (1.30 –2.78)	0.001	0.0005
		LD	(D)			1.04 (0.74 - 1.47)	0.82	>0.05	I			1.12 (0.65 - 1.93)	0.70	>0.05
	East Asians	SD	I	0.085	0.043	1.33 (1.03 - 1.72)	0.029	0.041	D	0.031	0.0299	1.52 (1.10 - 2.11)	0.012	0.0098
		MD	(I)			1.04 (0.81 - 1.34)	0.74	>0.05	D			0.93 (0.67 – 1.29)	0.65	>0.05
ACTN3	Caucasians	SMD	(X)	0.27	>0.05	1.12 (0.86 – 1.44)	0.41	>0.05	X	0.12	>0.05	1.20 (0.80 – 1.80)	0.37	>0.05
		LD	(R)			0.78 (0.55 – 1.12)	0.18	>0.05	X			0.63 (0.39 – 1.03)	0.065	>0.05
	East Asians	SD	R	0.082	>0.05	1.30 (1.03 – 1.65)	0.026	>0.05	X	0.069	>0.05	1.50 (1.07 – 2.12)	0.02	0.015
		MD	(R)			1.03 (0.82 – 1.31)	0.78	>0.05	X			1.13 (0.78 - 1.63)	0.53	>0.05

risk allele is designated as the allele whose frequency is higher in the relevant swimmer group than in controls; it has no meaning where tests reveal no significant association (see parentheses).

& *p*-values are given for the multinomial logistic regression (L.R.) model, in which two swimmer groups (e.g. SMD and LD, for Caucasians) are compared against controls in a single test.

* *p*-value calculated using a single model permutation test (PT) based on a χ^2 test implemented in PTest, inputs for this test were "Class" – swimmer group e.g. SMD, LD, Control, and "Feature" – in this case separate variables denoting genotype under 3 genetic models, additive (genotypes coded as "0, 1, 2") and two dominant models (genotypes coded as "0,0,1" and "0,1,1", respectively). These p-values are inherently adjusted for the multiple genetic models included in the overall model. *p*-values > 0.05 are not reported by PTest.

$ O.R. - odds ratio; 95% C.I. - 95% confidence interval. Odds ratios are reported for the designated risk allele for *ACE*, and for the *ACTN3* 577X-allele in Caucasians and the 577R-allele in East Asians.

¶ *p*-value for the estimate of β for the effect of genotype on the pertinent pairwise outcome comparison (e.g. SMD vs Control) embedded within the multinomial L.R. test.

§ *p*-value calculated using a pairwise PT approach implemented in PTest, in which a single swimmer group (e.g. SMD, for Caucasians) is compared against controls. Inputs were otherwise as described above.

† for each cohort, the model *p* value is given for the dominant model (as indicated in the 'Dom. allele' column) with the lowest *p* value.

†† O.R. - odds ratio; 95% C.I. - 95% confidence interval. Odds ratios are reported for the designated homozygous of the risk allele for *ACE*, and for the *ACTN3* 577X-allele in Caucasians and 577RR genotype in East Asians.

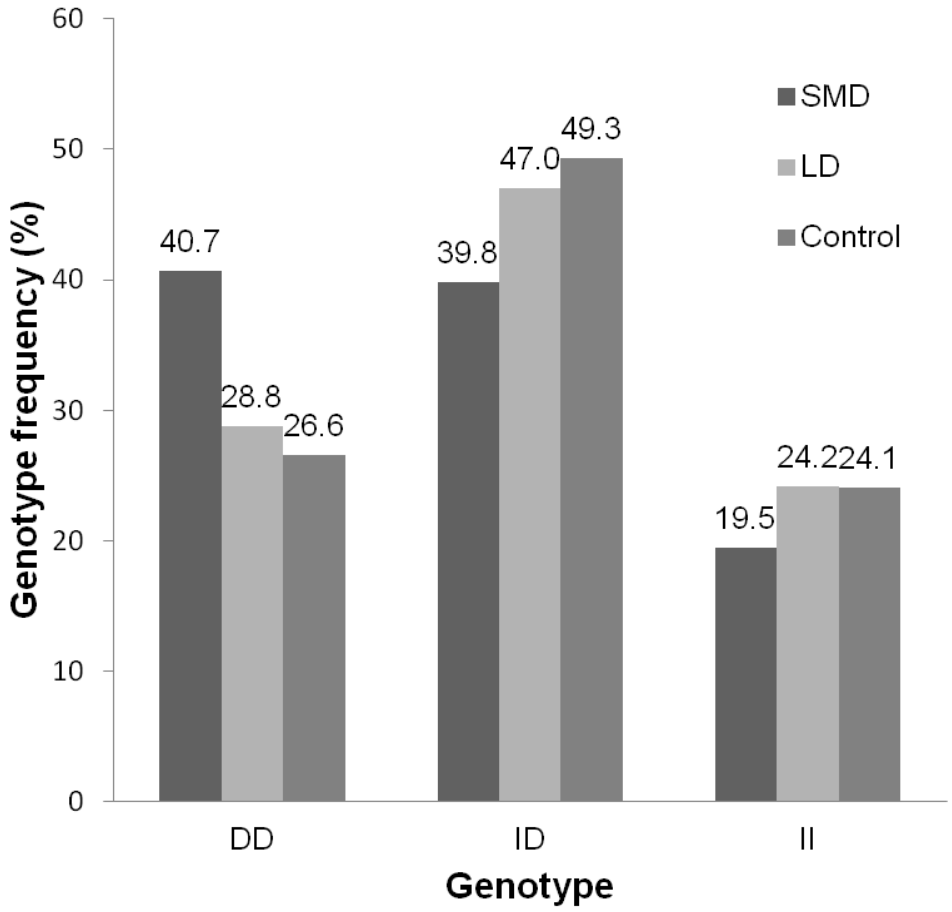


Figure 3.1 Genotype frequency distribution for ACE I/D in elite Caucasian swimmers and controls.

Before deciding how to treat the East Asian sample in the association analyses, multinomial logistic regression models were evaluated for 'genotype by ethnicity' interactions (i.e. the models were Outcome = genotype + ethnicity*(Japanese/Taiwanese)* + (genotype x ethnicity) + error) to determine whether effects on outcome differed between the two ethnic sub-cohorts. In models evaluated for both *ACE* and *ACTN3* under all three genetic models (additive and both dominant models), the interaction term was not significant ($p \geq 0.11$; Appendix Table A1.7; (170)). As a result, the Japanese and Taiwanese subgroups were treated as a single East Asian cohort in all subsequent analyses.

In East Asian SD swimmers, *ACE* I/D genotype was also associated with swimmer status (Figure 3.2, Table 3.2; (170)). The multinomial logistic regression models approached significance (Table 3.2; (170); $p = 0.085$ for the additive model and $p = 0.031$ for the D-allele dominant model). This tendency towards association was mediated by effects in SD

swimmers (I-allele homozygotes vs D-allele carriers: odds ratio = 1.52; logistic regression $p = 0.012$; permutation test $p = 0.0098$; MAX3 test statistic = 2.53, $p = 0.026$), with the I-allele being over-represented in the swimmers. It is therefore concluded that I-allele predisposes, in recessive fashion, to elite SD swimmer status in East Asians. No significant association was found between the *ACE* I/D polymorphism and East Asian MD swimmer status (Figure 3.2, Table 3.2; (170)).

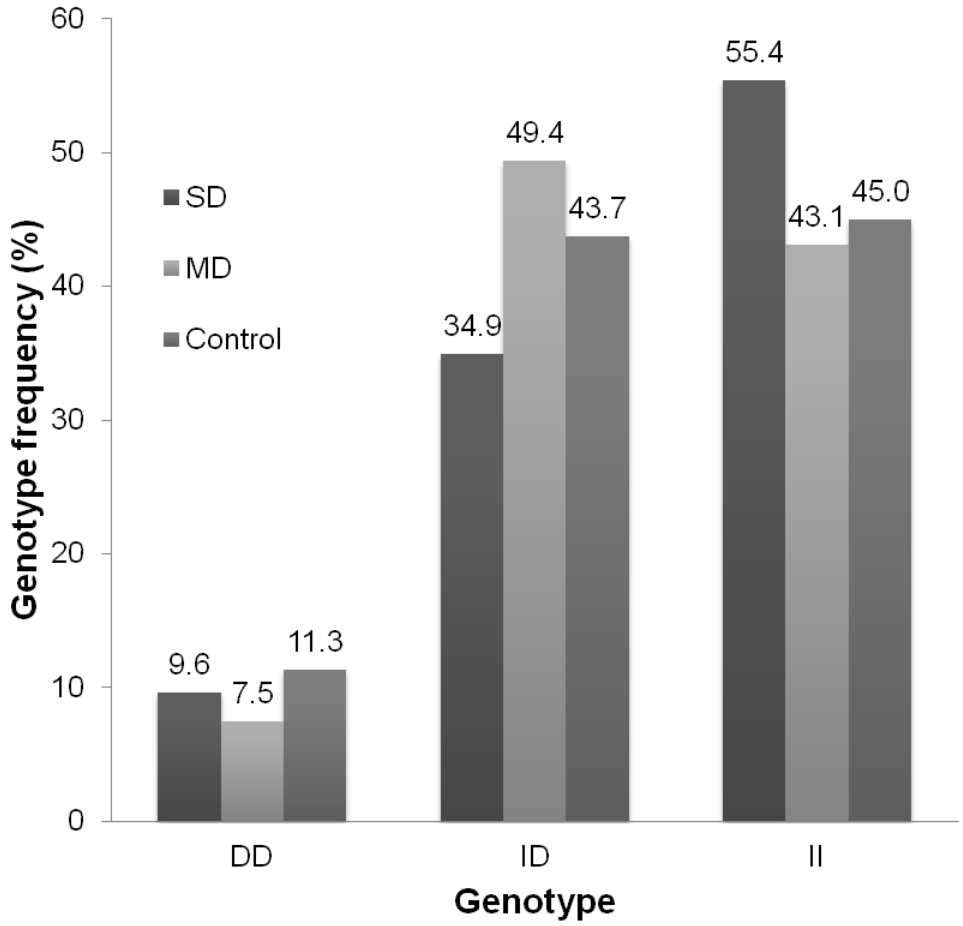


Figure 3.2 Genotype frequency distribution for ACE I/D in elite East Asian swimmers and controls.

For *ACTN3* R577X, no statistically significant associations were observed in either regional subgroup for any of the swim distance subgroups (Table 3.2, Appendix Table A1.3, Figures A1.8 and A1.9; (170)). The multinomial logistic regression model for East Asian swimmers approached significance ($p = 0.082$ and 0.069 under the additive and 577X-allele dominant models, respectively), with most of the effect coming from the SD

swimmers (logistic regression p = 0.02, permutation test p = 0.015) in whom the 577R-allele would be the performance-enhancing allele.

3.4 Discussion

The results show that the *ACE* I/D polymorphism is associated with elite swimmer status in both Caucasians and East Asians. The association is not seen in the longer distance events in each group, but only in SMD swimmers in Caucasians and only in SD swimmers in East Asians. *ACTN3* p.R577X genotype was not significantly associated with swimmer status in these samples.

The findings for *ACE* I/D need to be interpreted in the context of population differences in I/D allele frequency. Previous reports have highlighted the fact that allele frequencies at this locus differ somewhat between regional populations, with the D-allele occurring at lower frequency in Asian populations than in individuals of African or European descent (208,209). The frequency of the D-allele has been reported as 0.3 and 0.4 in Chinese and Japanese population samples, respectively (210,211), while in Caucasians, the average frequency of the D-allele is 0.52 (Appendix Figure A1.10; (170)). *ACE* I/D allele frequencies observed in the control samples employed here either came from these previously published studies or were entirely consistent with those previous reports. The lower minor allele frequency in East Asians reduces power to detect associations somewhat, but did not prevent an association being detected here, at least in the shorter distance swimmers. Despite the association in Caucasians being observed in swimmers of combined SD and MD designation (the SMD swimmer subgroup), there was no tendency for genotypes of East Asian MD swimmers to differ from controls in the same direction as in the significantly associated SD swimmers. Limited power is unlikely to explain this lack of trend, and the possibility should therefore be entertained that the populations differ in the extent to which *ACE* I/D affects swimmers at different distances.

In terms of direction of effect, the observation that the D-allele was associated with SMD swimmer status in Caucasians while the I-allele was associated with SD swimmer status in East Asians is particularly notable. The pattern of association of *ACE* I/D across ethnic groups is, however, in line with previous reports based on studies of other sporting events. Previous studies, though using smaller samples, have reported associations between the D-allele and elite SMD swimming status in Caucasians (148,171). The direction of effect in East Asians is consistent with previous reports if *ACE* affects other endurance/power-related sports in the same way as it does swimming - the D-allele has been reported to be associated with endurance performance in elite Japanese marathon runners (187), whereas the I-allele has been reported to be associated with elite power athlete status in Koreans (188). No associations with longer distance events were observed in the current study, but it is not always the case that complementary associations must be observed in opposing phenotypes and whether in fact these genotype effects operate across the entire phenotypic distribution in the whole population is not known.

While associations of opposite direction in different ethnic groups can be a result of type I error, there are several other possible explanations consistent with real association. Firstly, it may be that, although the causative variant(s) are identical in Caucasians and East Asians, the *ACE* haplotype networks found in Caucasians and East Asians are sufficiently different in the environs of these variants that different I/D alleles are on the predisposing haplotype more of the time in each group. Secondly, it may be that there are different causative variants in Caucasians and East Asians, with I- and D-alleles being on different haplotypes with respect to these more of the time in each regional subgroup. While the idea that common polymorphisms show association with phenotypes because of so-called ‘synthetic associations’ - where a number of different, individually rare causative alleles are all captured by a single tagging variant - is popular at present (212), there is remarkably little evidence for associations between a single complex phenotype and different

predisposing alleles in different populations (213). In addition, the relatively simple haplotype structure around I/D and relatively deep haplotype branching pattern (175), suggesting haplotype divergence predating the separation of the Caucasian and East Asian populations approx. 30-50,000 years ago, would argue against these first two explanations here. A third possible explanation is that ACE affects the relevant physiology differently in Caucasians and East Asians as a result of other changes in physiology appearing since the two population subgroups diverged. Thus, for example, higher ACE activity may predispose to short distance swimming performance in one population and lower ACE activity have the same effect in the other population.

The failure of previous studies to observe associations between *ACE* I/D and power-related performance in sub-Saharan African and African American/Jamaican samples (154), or indeed with endurance-related performance (214,215), is easier to explain. The I/D polymorphism is not thought to be the causative site influencing serum ACE activity, which is thought to be located between intron 18 and the 3′ UTR (175), with potential additional functional sites located in the 5′ region of the gene (175,216)). The haplotype structure in Caucasians and East Asians means that I/D is in very strong LD with at least one of these functional sites, almost certainly as a result of the out-of-Africa bottleneck. In Africa, however, there is much greater haplotype diversity across *ACE* and the LD structure means that I/D is not strongly associated with serum ACE activity (175); this is likely to be a large part of the explanation for the lack of association with sporting performance in African populations. An alternative explanation, however, may be that serum ACE activity is not the important factor influencing associations with performance and that local actions of ACE within skeletal or cardiac muscle that influence blood flow or other determinants of muscle performance over the life course or during performance tasks, for example, are more important (172,176). Other commonly genotyped *ACE*

variants may capture the effects of functional variants on such local ACE actions more effectively than the I/D polymorphism does.

The lack of clear association between *ACTN3* genotype and swimmer status is interesting in light of previous studies. In Caucasians, multiple studies have reported the *ACTN3* 577X-allele to be under-represented in elite sprint/power event athletes (reviewed in (217)). Few studies have focused on this polymorphism in East Asian elite athletes (197,199). Of the two ethnic groups studied here, the East Asians came closest to showing an association, this effect being in the same direction as in previous studies (with the 577R-allele being moderately over-represented in SD swimmers). Although ACTN3 deficiency has a modest effect on muscle fibre distribution (50), its impact on ability to perform in elite power events may have just as much to do with the role of ACTN3 in muscle metabolism as it does with muscle structure or fibre-type distribution *per se* (193,194). It may be that none of these roles sub-served by ACTN3 are of particular importance in swimming, or it may be that the aspect of power performance affected by the polymorphism is under-engaged in swimming relative to other sports, possibly because of the relatively lower stress put on muscles supported in water and lack of eccentric contractions (218). It may also be that swimming performance has a much greater component of technique than other power events. Lastly, there is the possibility that type II error accounts for the fact of lack of association here. Although a meta-analysis of associations between *ACTN3* and sprint/power athlete status has been published and does find evidence for a real association (219), many studies, mainly with small sample sizes, have failed to observe any association between *ACTN3* variants and sporting performance.

The 50-m and 100-m swimming events require speed and power for a short duration (< 2 mins; mainly anaerobic) (220). As the swimming distance increases, the ability for a swimmer to maintain the same speed for a longer duration (2-5 mins, 200 – 400 m; mixed anaerobic and aerobic) becomes more important; that is to say that muscular endurance

starts to make more contribution to swimming performance (220). The reason for studying Caucasian swimmers who were competitive at 400 m and less together rather than further subdividing it into the SD (50 – 100 m) and MD (200 – 400 m) groups as done for East Asian swimmers is that 36 Caucasian swimmers out of 200 total swimmers were competitive for swimming events ranging from 50 – 200 m, 100 – 200 m or 100 – 400 m. It would be reasonable to include and analyze all swimmers excelling in 400 m and below as the SMD group to possibly maximize the power of the study. Moreover, a previous study reported the decreased frequency of the D-allele beyond the 200-m swimming event in elite Portuguese swimmers (171). It is unlikely that the inclusion of the middle distance swimmers (e.g. at 200/400 m) has given rise to the significant association found between the *ACE* D-allele and Caucasian SMD swimmer status in the current study.

The limitations of the study reported here relate primarily to this issue of sample size, and also to the consequences of this for study design. Although this study was carried out using the largest elite swimmer sample yet assembled, it is still a relatively modest sample size for a genetic association study, and conclusions should be drawn with caution since small samples are more prone to other hidden biases, such as population stratification and cryptic relatedness, both of which may lead to increased type I error rates. Such effects have been controlled to the greatest extent possible using current study design, in which only two candidate SNPs were selected for genotyping. If indeed a role for *ACE* gene variation in elite swimming ability has been identified using the candidate gene approach adopted here, these findings prompt a number of questions that require further study; for example, it would be interesting to know whether such associations are observed in cohorts of swimmers adhering to different training regimes/intensities, whether history of injury affects the association, or whether the range of event distances in which the effect is observed is wide or narrow. The findings from this study should be interpreted with caution until confirmed by future studies, but are interesting nonetheless.

4 Genome-wide association study of elite performance

Despite extensive research in the investigation of the genetic basis of common/complex human traits over the last decade, linkage and candidate gene association studies have often failed to produce informative results. This statement is also true for the complex phenotype, elite human performance (43-52), studied here. Recent advances enable unbiased genome-wide approaches to be developed and applied for unlocking the genetic make-up of human common/complex traits. The international HapMap project documents common patterns (i.e. haplotypes) of genome-wide variations in 11 populations of 3 ancestral populations (i.e. African, European and Asian ancestries) (62), accelerating development of a robust reference panel to study human variations. Moreover, the 1,000 Genomes Project provides a more comprehensive catalogue of genetic variations across the human genome by sequencing 11 populations of African, European and Asian descent (63). Dense genotyping arrays, containing hundreds of thousands of SNPs selected from above reference panels (i.e. the International HapMap and 1,000 Genome Projects), would provide good coverage of the genome. These have facilitated a GWAS approach to become feasible.

In this chapter, GWASs in elite Jamaican sprint, African-American sprint, Japanese sprint and Japanese endurance athletes are summarized, including details on quality control measures, association analyses, and meta-analyses that have been carried out. The main focus of this chapter is to identify SNPs related to sprint performance by analyzing these 3 regional populations. The inclusion of GWAS of Japanese endurance athletes is an addition in the reason that (1) relevant genotype data was also available at the time of analysis and (2) it is thought that it would also be interesting to explore endurance-related SNPs, when possible. Finally, a genotype score approach was briefly introduced, which aims to investigate whether sprint-related SNPs identified from published reports show predictive

utility on elite sprint performance, assessed by using the genomic data generated from current GWASs.

4.1 GWAS of elite human performance

This subsection describes per-sample and per-marker QCs that are undertaken in 3 regional cohorts (cases + controls) of West African (i.e. Jamaicans and African-Americans) and East Asian (i.e. Japanese) ancestries. Participants and SNPs failing the QC thresholds are excluded as a result. Population stratification is assessed and corrected, and specific methods for this assessment and correction are outlined. Population characteristics of the final sample set for the formal association analysis are then summarized. Finally, association results (including meta-analyses results) are presented and discussed. For reference, general information for GWAS genotyping methods, statistical packages and relevant statistical analysis applied are provided in detail in Chapter 2 sections 2.2.2, 2.3 and 2.4.2. (i.e. 2.4.2.1 – 2.4.2.7).

All subjects supplied written informed consent, which was approved by the UHWI/UWI/FMS Ethics Committee, University of West Indies, Jamaica; participating institutions in the United States; the Institutional Review Board of Tokyo Metropolitan Institute of Gerontology; and National Institute of Health and Nutrition, Japan.

4.1.1 Per-individual QC

4.1.1.1 Gender inspection: estimated sex vs. recorded sex

X chromosome homozygosity estimate across all X-chromosome markers is used to detect discrepancies between sex estimated by genotype information and sex assigned in the phenotype file for each individual. Males have only one copy of the X chromosome; in theory, they are expected to have a homozygosity rate of 1. Individuals were inspected if they were recorded as males but showed an excess amount of heterozygous genotypes for

X-chromosome markers (estimated homozygosity rate < 0.2), or recorded female samples had a higher than expected homozygosity rate (> 0.8). If sample sex discordance cannot be identified confidently based on the genotype data or the recorded phenotype data, problematic individuals with sex mismatch are then needed to be excluded from further analyses.

4.1.1.2 Missingness and heterozygosity rate

Samples with low DNA quality would often result in lower than average call rates. Samples with genotype missing rate > 5% were therefore removed.

Sample quality can also be measured by autosomal heterozygosity rates. Samples with excessive heterozygous genotypes may imply potential sample contamination, whereas reduced heterozygous rates may indicate inbreeding. Mean heterozygosity per individual is calculated given the difference between the number of non-missing genotypes and the number of observed homozygous genotypes divided by the number of non-missing genotypes. Samples with heterozygosity rate > ±3 standard deviations from mean heterozygosity were removed.

4.1.1.3 Cryptic relatedness

Unexpected relatedness between study samples were examined for every pair of individuals with proportion IBD > 0.05. The relationship between Z0 and Z1 were inspected. For problematic pairs (i.e. duplicated samples/identical twins), sample with lower call rate in each pair were excluded from further analyses.

4.1.1.4 Outliers of PCA

Ancestry detection was examined using principle component analysis for current GWAS cohorts along with the HapMap reference populations (Europe: CEU + TSI; Asia: CHB + JPT; Africa: YRI). 10 PCs were extracted using EIGENSTRAT (implemented in the

EIGENSOFT package). By default, extreme outliers were firstly removed (sigma 6.0 – ≥ 6 standard deviations from the mean along one of the top 10 PCs, by 5 iterations) (97).

The first two PCs were able to separately cluster individuals from the 3 HapMap populations given large genetic differences among these 3 ancestral groups, and allowing individuals of the GWAS cohorts to be clustered along with these HapMap groups. Current GWAS samples that were not clustered with the rest were subsequently removed.

Individuals removed at each sample QC step are shown in Table 4.1. Examples of outliers removed based on PCA are shown in Figure 4.1-4.3.

Table 4.1 Individuals failed sample QCs and excluded from further association analyses in current GWAS cohorts.

	Jamaica	**USA**	**Japan**
Sex-check	2 athletes 2 controls	6 athletes 2 controls	-
Call rate	1 athlete 1 control	-	-
Outliers of heterozygosity	4 controls	6 athletes 2 controls	2 controls
Cryptic relatedness	1 athlete 4 controls	2 athletes 1 control	-
Outliers of PCA	3 athletes 4 controls	13 athletes	-

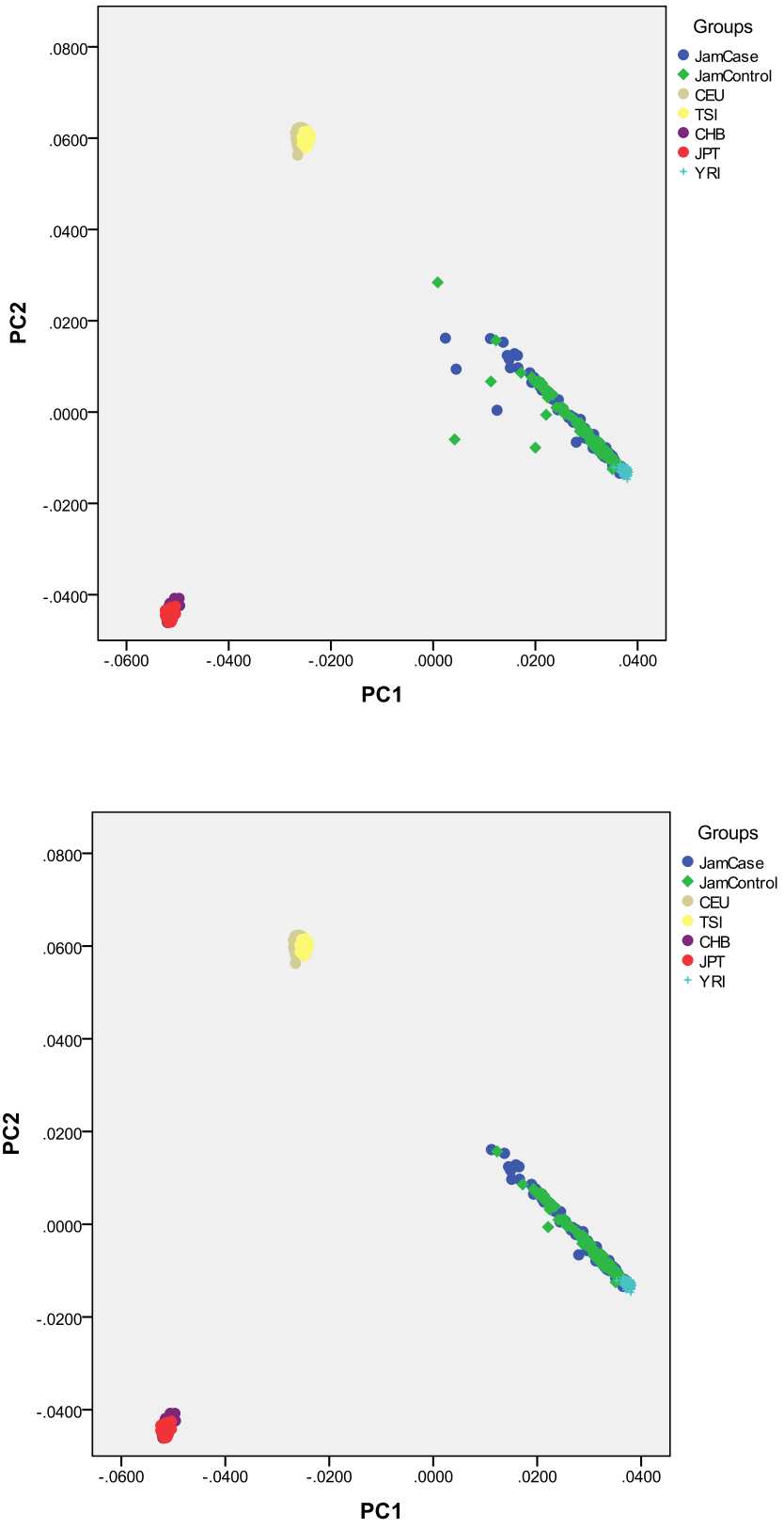


Figure 4.1 First two PCs for ancestry clustering of Jamaican sprint athletes and Jamaican controls alongside with 5 Hapmap3 reference populations (CEU: Utah residents with Northern and Western European ancestry; TSI: Toscani in Italy; CHB: Han Chinese in Beijing, China; JPT = Japanese in Tokyo, Japan; YRI = Yoruba in Ibadan, Nigeria). Top graph: before removal of outliers; bottom graph: after removal.

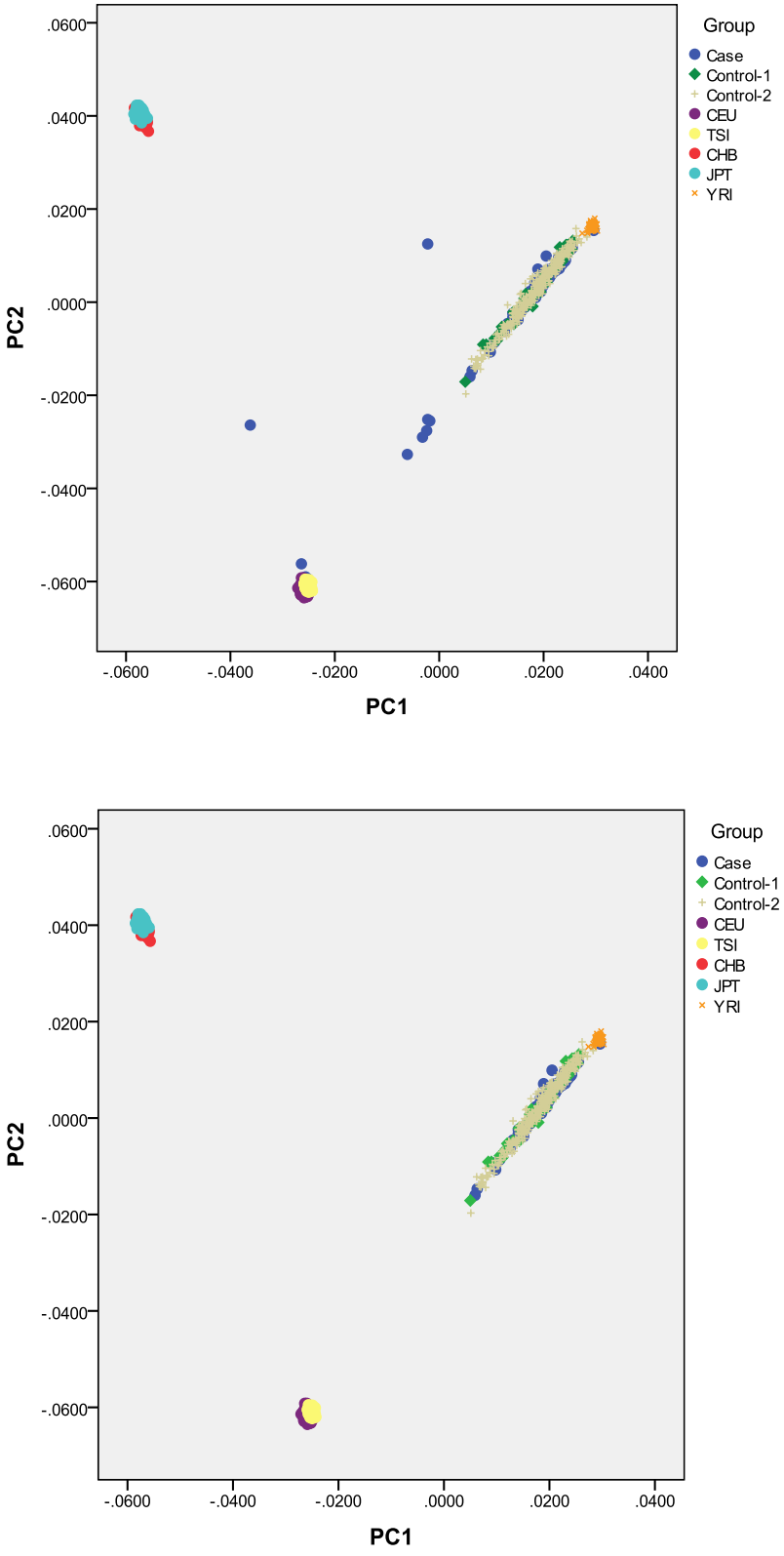


Figure 4.2 First two PCs for ancestry clustering of African-American sprint athletes and African-American controls alongside with 5 Hapmap3 reference populations (CEU: Utah residents with Northern and Western European ancestry; TSI: Toscani in Italy; CHB: Han Chinese in Beijing, China; JPT = Japanese in Tokyo, Japan; YRI = Yoruba in Ibadan, Nigeria). Top graph: before removal of outliers; bottom graph: after removal. Control-1: originally collected African-American controls; Control-2: African-American controls' genotype data received from another group at the University of Maryland.

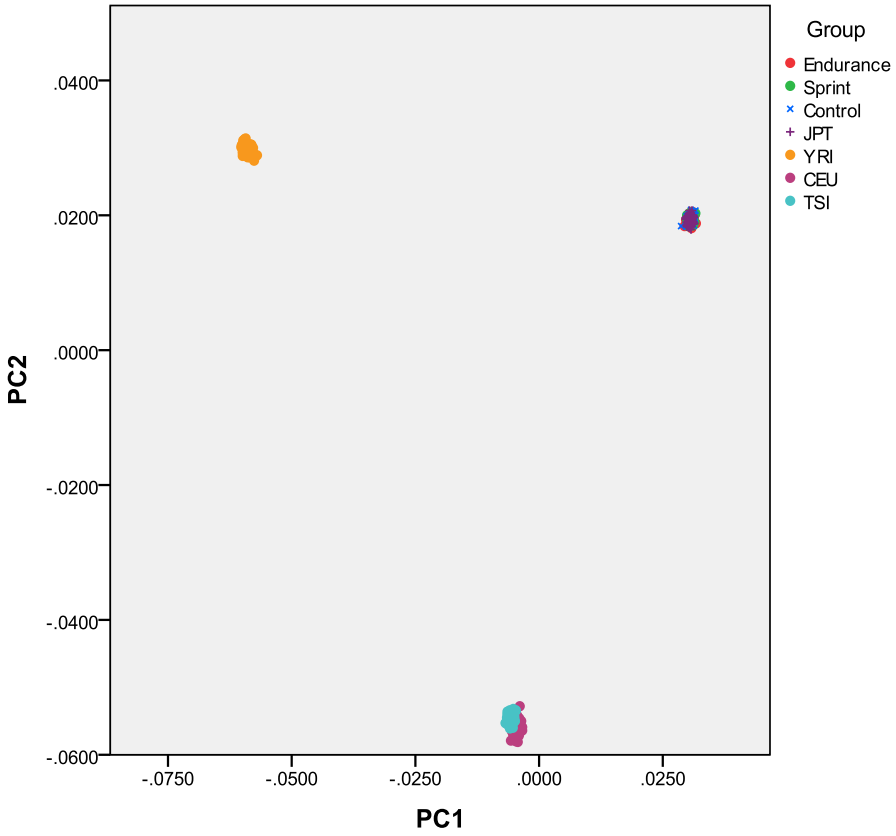


Figure 4.3 First two PCs for ancestry clustering of Japanese endurance, sprint athletes and Japanese controls alongside with 4 Hapmap3 reference populations (CEU: Utah residents with Northern and Western European ancestry; TSI: Toscani in Italy; JPT = Japanese in Tokyo, Japan; YRI = Yoruba in Ibadan, Nigeria).

4.1.2 Per-marker QC

For Jamaican and African-American samples, both Illumina® HumanOmniExpress and HumanOmni1-Quad Beadchips were used for the whole genome genotyping. Data was firstly merged and cleaned (i.e. by QCs) for genotypes generated using the same genotyping platform, and then the cleaned genotype data from these 2 platforms was merged and QCs were repeated and the numbers of markers failed are reported below. For Japanese samples, all was genotyped using Illumina® HumanOmniExpress Beadchip; however, genotyping data was received multiple times and merged prior to data cleaning. The numbers of markers excluded from the final merged data set are listed below.

4.1.2.1 Missingness

267,969, 297,404 and 48,490 SNPs with a genotype missing rate more than 5% were excluded from GWAS cohorts of Jamaicans, African-Americans and Japanese, respectively.

4.1.2.2 Minor allele frequency

9 SNPs with a MAF less than 1% were removed from GWAS of African-Americans. No SNPs were excluded from Jamaican and Japanese GWAS cohorts given MAF < 1%.

4.1.2.3 Hardy-Weinberg Equilibrium

Control samples were used for testing markers deviated from HWE, and the significance threshold was declared at $p \leq 1 \times 10^{-7}$ for excluding SNPs showing deviations from HWE. 31 SNPs were removed from further investigation for Japanese GWAS cohort, whereas no SNPs exceeded HWE p of 1×10^{-7} for Jamaican and African-American GWAS sample sets.

After SNP quality controls, a final set of 609,801, 637,991 and 541,179 autosomal SNPs were available for Jamaican, African-American and Japanese GWAS association analyses.

4.1.2.4 Population stratification adjustments using PCA in Jamaican and African-American GWAS

For Jamaicans, the top PCs extracted from PCA were entered into logistic regression models as covariates (athlete/control status ~ SNPs + PC1 + PC3 +… + PC10) to account for population structure within a sample set and subsequently correct for population stratification. For African-Americans, except involving PCs, a covariant of "genotyping centre" was also included in the model (athlete/control status ~ SNPs + PC1 + PC3 + … + PC10 + Genotyping Centre) to minimize the between-centre effect on genotype calling, which may induce false associations.

4.1.3 Association analysis

The final sample set available for formal analysis for each regional cohort is presented in Table 4.2. After removing individuals and SNPs failing QC, 609,801 autosomal SNPs in 88 Jamaican sprint athletes and 87 Jamaican controls, 637,991 autosomal SNPs in 79 African-American sprint athletes and 391 African-American controls, and 541,179 autosomal SNPs in 114 Japanese athletes (including 60 endurance and 54 sprint athletes) and 116 Japanese controls, were available for analysis.

Table 4.2 Number of samples available in the formal analysis for each regional cohort.

	Jamaican cohort	African-American cohort	Japanese cohort
Athletes	88 (sprinting)	79 (sprinting)	60 (endurance) 54 (sprinting)
Controls	87	391	116

Genetic associations were evaluated by logistic regression, assuming an additive effect, for Jamaicans and African-Americans, respectively. For association analysis of Jamaicans, population stratification was adjusted by entering PCs into the logistic regression model. For association analysis of African-Americans, both population stratification and genotyping-centre effect were corrected by including PCs and "genotyping centre" variables as covariates in the logistic regression model. Standard allelic association analysis was performed by comparing allele-frequency differences between Japanese endurance athletes and controls, and Japanese sprint athletes and controls, respectively.

The Quantile-Quantile *p*-value plots of observed versus expected $-\log_{10}(p)$ values for each comparison are shown in Figure 4.4. Genomic inflation factor (λ) values are 1.075 and 1.070 for GWAS of Jamaicans and African-Americans, respectively, after necessary

corrections (see paragraph above). λ values in Japanese GWAS cohort are of 1.002 and 1.031 for endurance and sprinting sub-cohorts, respectively, and are smaller than 1.05 indicating that there is no substantial evidence of population stratification. Manhattan plots of $-\log_{10}(p)$ values for association of elite athletic status with markers in 22 autosomes are shown in Figure 4.5-4.8.

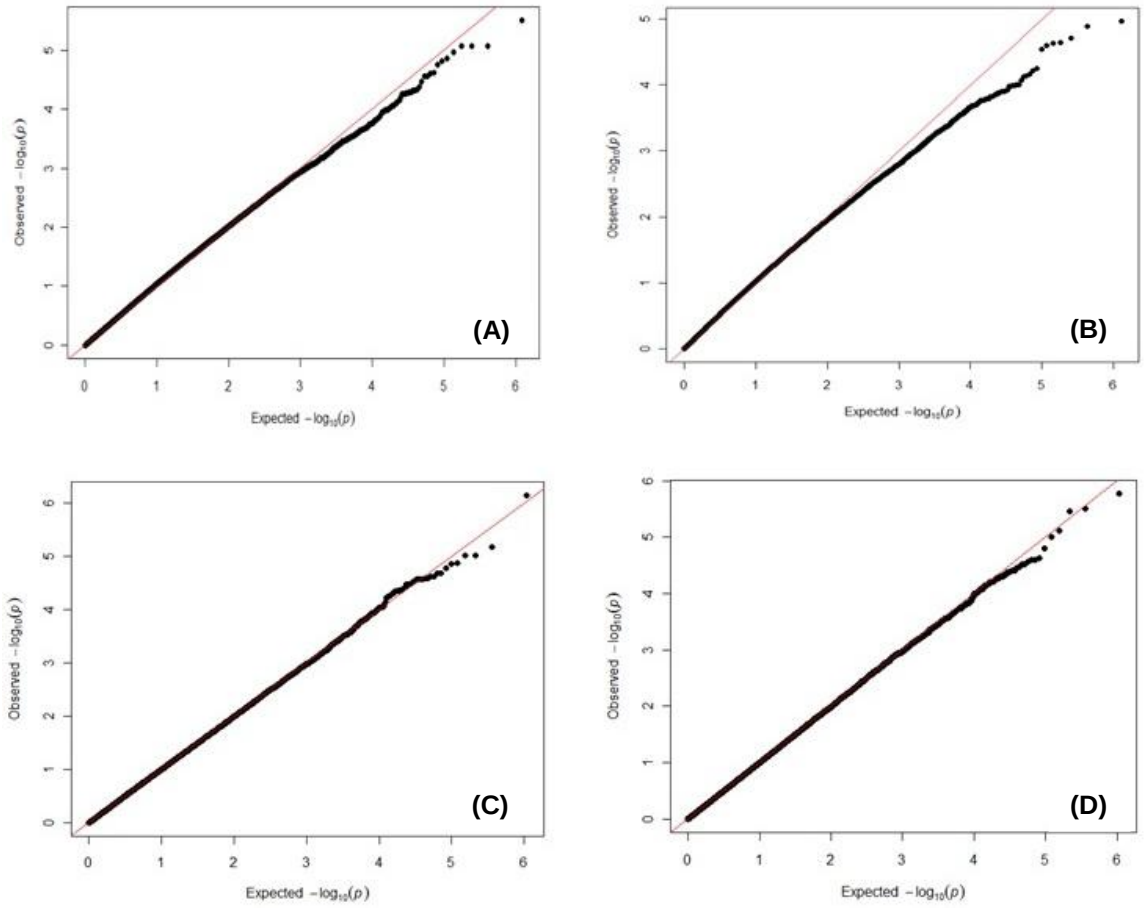


Figure 4.4 Quantile-Quantile plots of observed vs. expected $-\log_{10}(p)$ values for genome-wide data. Red line indicates the null line of no association. (A) Jamaican sprint cohort; (B) African-American sprint cohort; (C) Japanese sprint cohort; (D) Japanese endurance cohort. λ = 1.075, 1.070, 1.031, 1.002 for A, B, C and D, respectively.

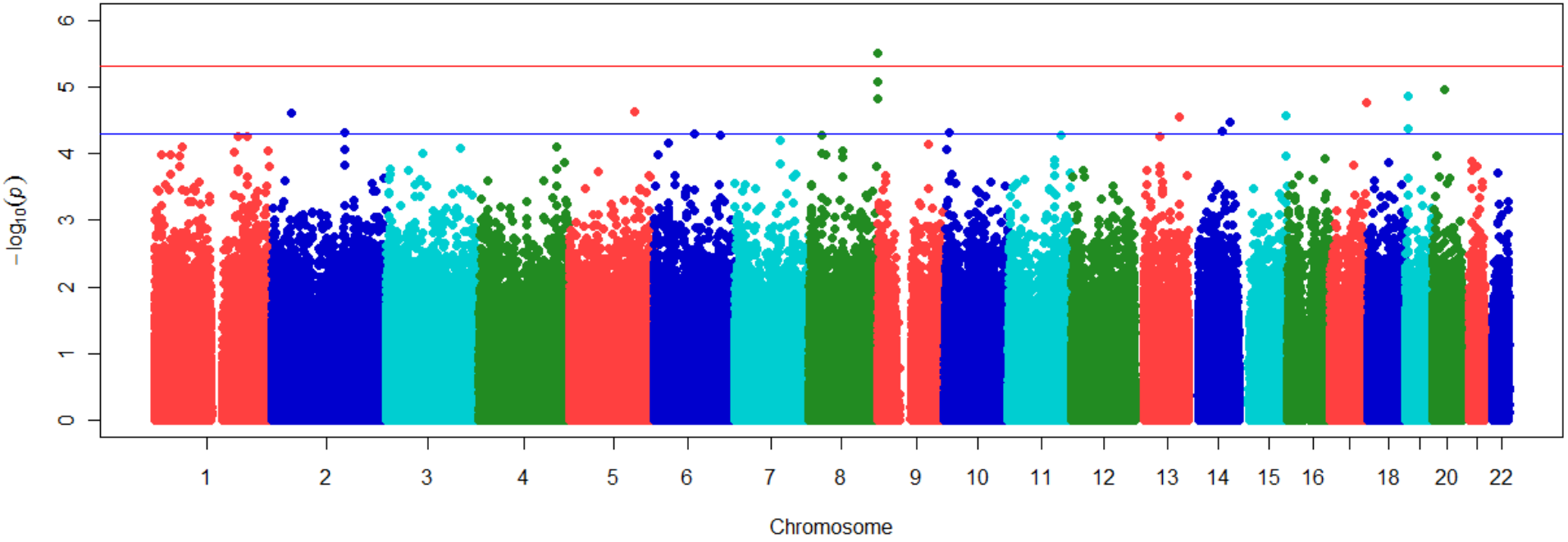


Figure 4.5 Manhattan plot of $-\log_{10}(p)$ values against genomic position for association of elite sprint status with markers in 22 autosomes in Jamaicans. Red line refers to $p = 5\text{x}10^{-6}$; blue line refers to $p = 5\text{x}10^{-5}$.

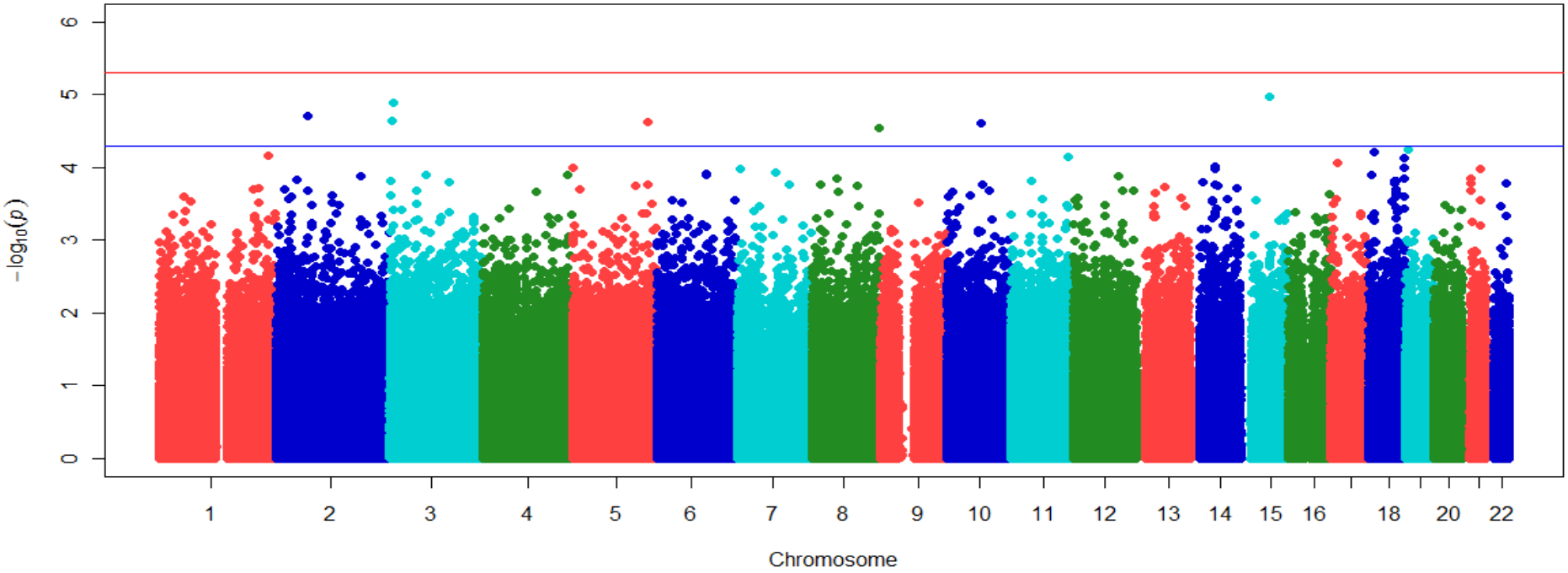


Figure 4.6 Manhattan plot of $-\log_{10}(p)$ values against genomic position for association of elite sprint status with markers in 22 autosomes in African-Americans. Red line refers to $p = 5\text{x}10^{-6}$; blue line refers to $p = 5\text{x}10^{-5}$.

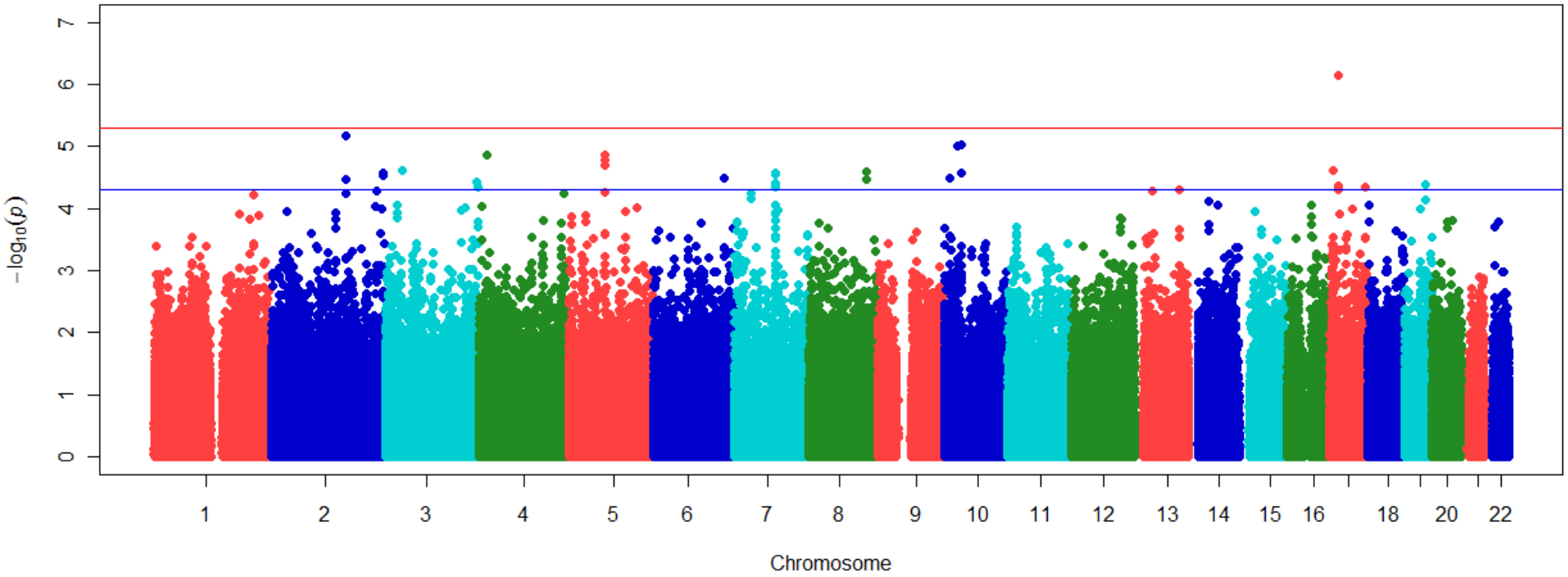


Figure 4.7 Manhattan plot of $-\log_{10}(p)$ values against genomic position for association of elite sprint status with markers in 22 autosomes in Japanese. Red line refers to $p = 5\text{x}10^{-6}$; blue line refers to $p = 5\text{x}10^{-5}$.

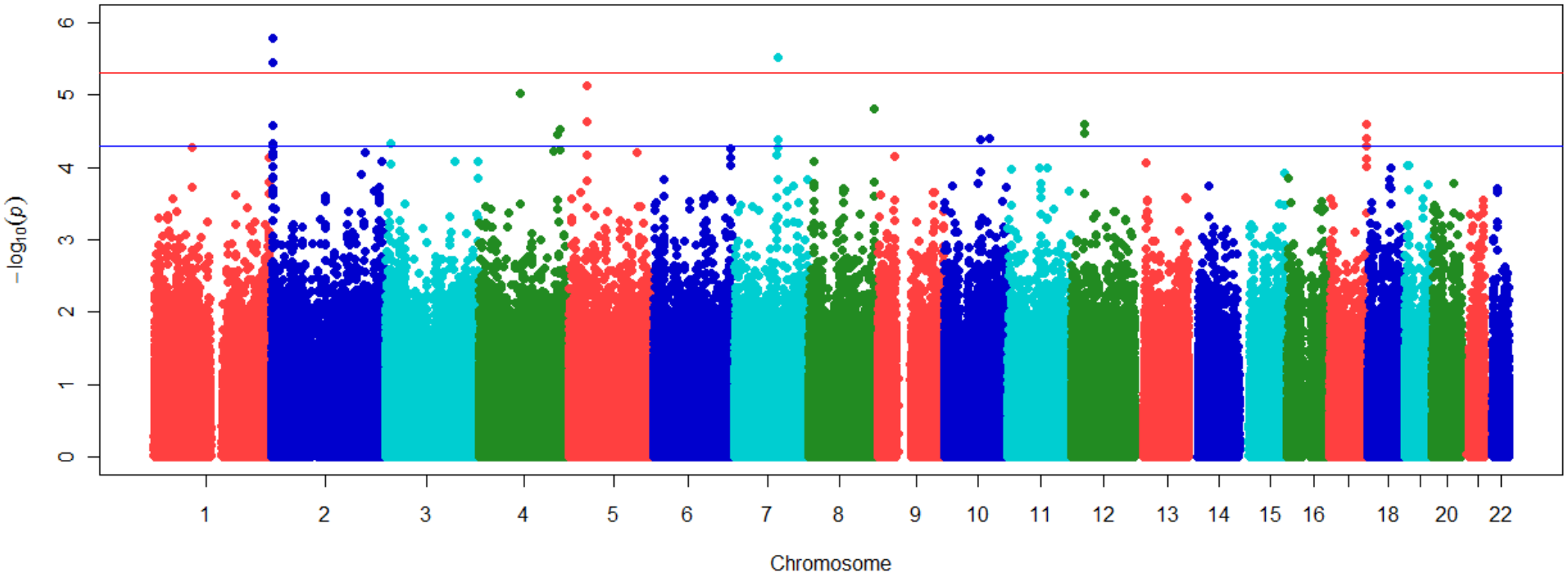


Figure 4.8 Manhattan plot of $-\log_{10}(p)$ values against genomic position for association of elite endurance status with markers in 22 autosomes in Japanese. Red line refers to $p = 5\text{x}10^{-6}$; blue line refers to $p = 5\text{x}10^{-5}$.

Table 4.3 - Table 4.6 present the association results for markers with an unadjusted $p < 5 \times 10^{-5}$. Out of the total 609,801 SNPs entered into the association analysis in the Jamaican sprint cohort, 17 met this threshold of significance; similarly, 7 out of 637,991 SNPs, 36 out of 541,179 SNPs and 21 out of 541,179 SNPs were found exceeding the same threshold in the African-American sprint, Japanese sprint and Japanese endurance cohorts, respectively. Furthermore, 1, 0, 1 and 3 SNPs attained $p < 5 \times 10^{-6}$ (unadjusted) for each of these cohorts.

GC adjusted *p* values were also reported. 10, 6, 24, and 21 SNPs with GC adjusted $p < 5 \times 10^{-5}$ are present in the association results of Jamaican sprint, African-American sprint, Japanese sprint and Japanese endurance cohorts, respectively; and 1 and 3 SNPs are remained significant in Japanese sprint and endurance cohorts at a GC adjusted *p* value of 5×10^{-6}.

Table 4.3 Association results for markers with unadjusted $p < 5 \times 10^{-5}$ in Jamaican sprint cohort.

SNP	Chr.	Position	O.R.	95% C.I.	Unadjusted *p*	GC adjusted *p*
rs4557742	8	145508113	3.27	1.99 - 5.37	3.11E-06	6.86E-06
rs4977203	8	145513753	2.92	1.82 - 4.69	8.58E-06	1.77E-05
rs11998675	8	145514420	2.92	1.82 - 4.69	8.58E-06	1.77E-05
rs4977219	8	145516698	2.92	1.82 - 4.69	8.58E-06	1.77E-05
rs4815390	20	25158241	0.32	0.19 - 0.53	1.10E-05	2.23E-05
rs2303115	19	7708214	0.33	0.2 - 0.54	1.37E-05	2.74E-05
rs35253356	8	145519034	2.86	1.78 - 4.6	1.51E-05	3.00E-05
rs2606193	17	77211481	0.23	0.12 - 0.45	1.76E-05	3.46E-05
rs187167	5	139029000	5.19	2.42 - 11.13	2.41E-05	4.64E-05
rs2374482	2	43305926	0.33	0.2 - 0.55	2.51E-05	4.81E-05
rs8027231	15	98968160	0.30	0.17 - 0.53	2.73E-05	5.21E-05
rs7336411	13	96014509	0.37	0.23 - 0.59	2.78E-05	5.30E-05
rs804944	14	86817161	3.40	1.91 - 6.08	3.42E-05	6.42E-05
rs10415518	19	7763917	0.33	0.19 - 0.56	4.25E-05	7.87E-05
rs17107388	14	70357182	0.25	0.13 - 0.49	4.72E-05	8.68E-05
rs10196189	2	154826491	2.98	1.76 - 5.05	4.89E-05	8.97E-05
rs734366	10	10968124	0.31	0.18 - 0.55	4.89E-05	8.98E-05

Chr. – Chromosome
O.R. – Odds Ratio
95% C.I. – 95% Confidence Interval
GC – Genomic Control

Table 4.4 Association results for markers with unadjusted $p < 5 \times 10^{-5}$ in African-American sprint cohort.

SNP	Chr.	Position	O.R.	95% C.I.	Unadjusted *p*	GC adjusted *p*
rs7175629	15	60393976	0.21	0.11 – 0.42	1.08E-05	2.10E-05
rs3864067	3	7677215	0.19	0.09 – 0.4	1.30E-05	2.50E-05
rs17034251	2	67955919	0.21	0.1 – 0.43	1.93E-05	3.63E-05
rs4054851	3	5193135	0.22	0.11 – 0.44	2.26E-05	4.21E-05
rs7716847	5	161662347	0.20	0.09 – 0.42	2.35E-05	4.37E-05
rs4747094	10	72499334	0.16	0.07 – 0.37	2.50E-05	4.62E-05
rs10111342	8	142730519	5.03	2.36 – 10.71	2.91E-05	5.33E-05

Chr. – Chromosome
O.R. – Odds Ratio
95% C.I. – 95% Confidence Interval
GC – Genomic Control

Table 4.5 Association results for markers with unadjusted $p < 5 \times 10^{-5}$ in Japanese sprint cohort.

SNP	Chr.	Position	O.R.	95% C.I.	Unadjusted *p*	GC adjusted *p*
rs12450878	17	17084269	9.08	3.27 - 25.19	7.04E-07	1.04E-06
rs10497155	2	157875846	5.09	2.37 - 10.95	6.73E-06	9.33E-06
rs7921820	10	35554947	3.27	1.91 - 5.62	9.56E-06	1.31E-05
rs10763704	10	29586483	2.89	1.79 - 4.67	9.63E-06	1.32E-05
rs1715747	5	76274537	2.81	1.75 - 4.51	1.34E-05	1.83E-05
rs1400938	4	18725182	3.06	1.82 - 5.12	1.40E-05	1.90E-05
rs2046046	5	76644962	3.24	1.87 - 5.62	1.63E-05	2.20E-05
rs1875999	5	76264982	2.75	1.72 - 4.42	2.03E-05	2.72E-05
rs1053989	5	76265035	2.75	1.72 - 4.42	2.03E-05	2.72E-05
rs356045	17	6451937	2.71	1.70 - 4.33	2.37E-05	3.17E-05
rs17033272	3	35509452	0.08	0.02 - 0.35	2.38E-05	3.18E-05
rs13439619	8	121948544	2.81	1.72 - 4.58	2.52E-05	3.37E-05
rs6942407	7	86861313	0.36	0.22 - 0.59	2.61E-05	3.48E-05
rs3740082	10	35502533	3.06	1.79 - 5.21	2.65E-05	3.52E-05
rs16935888	10	35432405	3.06	1.79 - 5.21	2.65E-05	3.52E-05
rs1531550	10	35464778	3.06	1.79 - 5.21	2.65E-05	3.52E-05
rs4503948	2	237758257	2.77	1.71 - 4.50	2.70E-05	3.59E-05
rs17766292	7	86830190	2.71	1.69 - 4.35	2.81E-05	3.73E-05
rs6431485	2	237760783	2.94	1.75 - 4.93	2.97E-05	3.94E-05
rs4750319	10	13268710	2.72	1.69 - 4.40	3.20E-05	4.24E-05
rs789481	6	150594363	2.66	1.67 - 4.26	3.30E-05	4.36E-05
rs16894449	8	121950167	2.77	1.70 - 4.52	3.32E-05	4.39E-05
rs6747313	2	157873775	3.33	1.85 - 6.00	3.34E-05	4.41E-05
rs6806282	3	194060008	2.74	1.68 - 4.45	3.67E-05	4.84E-05
rs7778976	7	86944642	2.67	1.66 - 4.30	3.98E-05	5.23E-05
rs11880216	19	44444416	2.87	1.72 - 4.80	4.08E-05	5.37E-05
rs7223686	17	16960911	3.33	1.84 - 6.05	4.23E-05	5.56E-05
rs12452303	17	16968670	3.33	1.84 - 6.05	4.23E-05	5.56E-05
rs10852845	17	17017584	3.33	1.84 - 6.05	4.23E-05	5.56E-05
rs9840798	3	197136335	3.86	1.95 - 7.62	4.44E-05	5.82E-05
rs7220712	17	74782513	3.86	1.95 - 7.62	4.44E-05	5.82E-05
rs7812191	7	86767689	2.65	1.65 - 4.25	4.51E-05	5.92E-05
rs8075751	17	16941236	3.20	1.80 - 5.69	4.54E-05	5.94E-05
rs11658904	17	16944622	3.20	1.80 - 5.69	4.54E-05	5.94E-05
rs4773783	13	94929371	0.29	0.15 - 0.54	4.88E-05	6.38E-05
rs4985714	17	16953884	3.30	1.82 - 6.00	4.95E-05	6.47E-05

Chr. – Chromosome
O.R. – Odds Ratio
95% C.I. – 95% Confidence Interval
GC – Genomic Control

Table 4.6 Association results for markers with unadjusted $p < 5x10^{-5}$ in Japanese endurance cohort.

SNP	Chr.	Position	O.R.	95% C.I.	Unadjusted *p*	GC adjusted *p*
rs921665	2	3174321	0.32	0.20 - 0.52	1.68E-06	1.72E-06
rs11975386	7	93705033	3.14	1.92 - 5.13	3.07E-06	3.15E-06
rs4854131	2	3173024	0.34	0.21 - 0.54	3.52E-06	3.60E-06
rs2910756	5	37860074	0.28	0.16 - 0.50	7.58E-06	7.75E-06
rs10007111	4	88749701	2.87	1.79 - 4.62	9.72E-06	9.93E-06
rs16906888	8	138243618	2.97	1.79 - 4.92	1.58E-05	1.62E-05
rs2973033	5	37839633	0.31	0.17 - 0.54	2.34E-05	2.39E-05
rs4541108	17	77328609	4.00	2.03 - 7.90	2.51E-05	2.56E-05
rs7975710	12	26534066	0.34	0.20 - 0.57	2.54E-05	2.59E-05
rs6548153	2	3326045	0.37	0.23 - 0.59	2.69E-05	2.75E-05
rs494219	4	172758442	0.20	0.09 - 0.45	3.01E-05	3.07E-05
rs558129	4	172751111	0.20	0.09 - 0.45	3.01E-05	3.07E-05
rs12582235	12	26537202	0.33	0.19 - 0.56	3.36E-05	3.42E-05
rs7668194	4	168472046	4.71	2.14 - 10.37	3.49E-05	3.55E-05
rs8081466	17	77330461	3.62	1.91 - 6.87	3.97E-05	4.04E-05
rs7209293	17	77333274	3.62	1.91 - 6.87	3.97E-05	4.04E-05
rs2761291	10	95088180	3.54	1.89 - 6.64	4.03E-05	4.10E-05
rs10245760	7	93746499	3.01	1.75 - 5.15	4.12E-05	4.20E-05
rs17690338	10	77117556	2.63	1.65 - 4.21	4.21E-05	4.29E-05
rs2887311	2	3190384	2.54	1.61 - 3.99	4.73E-05	4.82E-05
rs7650685	3	11702456	2.54	1.61 - 3.99	4.73E-05	4.82E-05

Chr. – Chromosome
O.R. – Odds Ratio
95% C.I. – 95% Confidence Interval
GC – Genomic Control

Meta-analyses were performed for SNPs with unadjusted association $p < 5 \times 10^{-5}$ across the sprint GWAS sample sets (i.e. Jamaican sprint, African-American, Japanese sprint and Japanese GWAS cohorts), using the fixed-effects model (221,222). For example, the top 17 SNPs (unadjusted $p < 5 \times 10^{-5}$) from the Jamaican sprint cohort were extracted from the association results of African-American sprint, Japanese sprint cohorts, respectively, for the combined effects to be calculated using a meta-analysis method. The same procedure was applied to the top hits in African-American and Japanese cohorts in turn. The new significance levels after meta-analyses were defined as 3×10^{-6}, 8×10^{-6} and 2×10^{-6}, which were calculated given 5×10^{-5} divided by the number of extra meta-analysis tests carried out in each meta-analysis. After the meta-analysis, rs10196189 which was initially identified from Jamaican sprint GWAS remained significant ($p = 4.66 \times 10^{-7}$, exceeding 3

x 10^{-6}), and rs1531550 from Japanese sprint GWAS attained at 2 x 10^{-6} (p = 1.88 x 10^{-6}) (see Table 4.7).

Regional association plots of the SNPs crossed an unadjusted $p < 5 \times 10^{-5}$ were further inspected for four GWAS cohorts (including the Japanese endurance GWAS samples) respectively, leading to the exclusion of 10, 0, 18, and 9 problematic SNPs, which are considered as either redundant or false signals. The regional association plots for the remaining SNPs, served as the key markers representing for those regions, are presented in the Appendix A2 to A5. In general, these SNPs that are not filtered out could be taken forward for future studies. Among the top SNPs discovered from the three initial sprint GWAS cohorts, no common SNPs are found. Nevertheless, given the results of the meta-analyses of combined sprint GWAS, rs10196189 is considered as the first candidate SNP for validation and replication, with the allele G (odds ratio = 2.61, p = 4.66 x 10^{-7}; Table 4.7) associated with elite sprint status in Jamaicans, African-Americans and Japanese. The regional association plots for this hit in the 3 sprint GWAS cohorts are presented in Figure 4.9 – 4.11, respectively. The second significantly associated signal is from rs1531550, and the same effect direction of the allele A is only observed in Jamaican and Japanese sprint GWAS cohorts (odds ratio = 3.09, p = 1.88 x 10^{-6}; Table 4.7). Plots that demonstrate the regional association relationships between rs1531550 and the surrounding SNPs are displayed in Figure 4.12 and 4.13 for Japanese and Jamaicans, respectively.

Table 4.7 Significant meta-analyses results of the top SNPs with an unadjusted fixed-effects *p*-value < 5x10^{-5} across Jamaican sprint, African-American sprint, Japanese sprint GWAS samples.

Initial GWAS	Chr	SNP	BP	A1	A2	MAF JAM	O.R. JAM	95% C.I. JAM	*p* JAM	MAF AA	O.R. AA	95% C.I. AA	*p* AA	MAF JAP	O.R. JAP	95% C.I. JAP	*p* JAP	N	*p* (F)	O.R. (F)	Q	I^2
JAM	2	rs10196189	154826491	G	A	0.36	2.98	1.76 – 5.05	4.89E-05	0.32	2.16	1.12-4.15	0.02	0.06	2.52	1.04-6.12	0.036	3	4.66E-07	2.61	0.75	0
JAP	10	rs1531550	35464778	A	G	0.07	3.19	1.25-8.14	0.015	-	-	-	-	0.21	3.06	1.79-5.21	2.65E-05	2 (JAM, JAP)	1.88E-06	3.09	0.94	0

JAM - Jamaican, AA – African-American, JAP – Japanese;
Chr – chromosome; BP – base pair; A1 – minor allele; A2 – the alternative allele;
MAF – minor allele frequency; O.R. – odds ratio (with respect to allele A1); 95% C.I.: 95% confidence interval; the shaded area refers to the initial association results of the top SNPs from each GWAS cohort;
N – number of valid studies for this SNP in meta-analysis; *p* (F) – fixed-effects meta-analysis p value; O.R. (F) - fixed-effects odds ratio estimate; Q – P-value for Cochrane's Q statistic; I^2 – I^2 heterogeneity index (0-100).

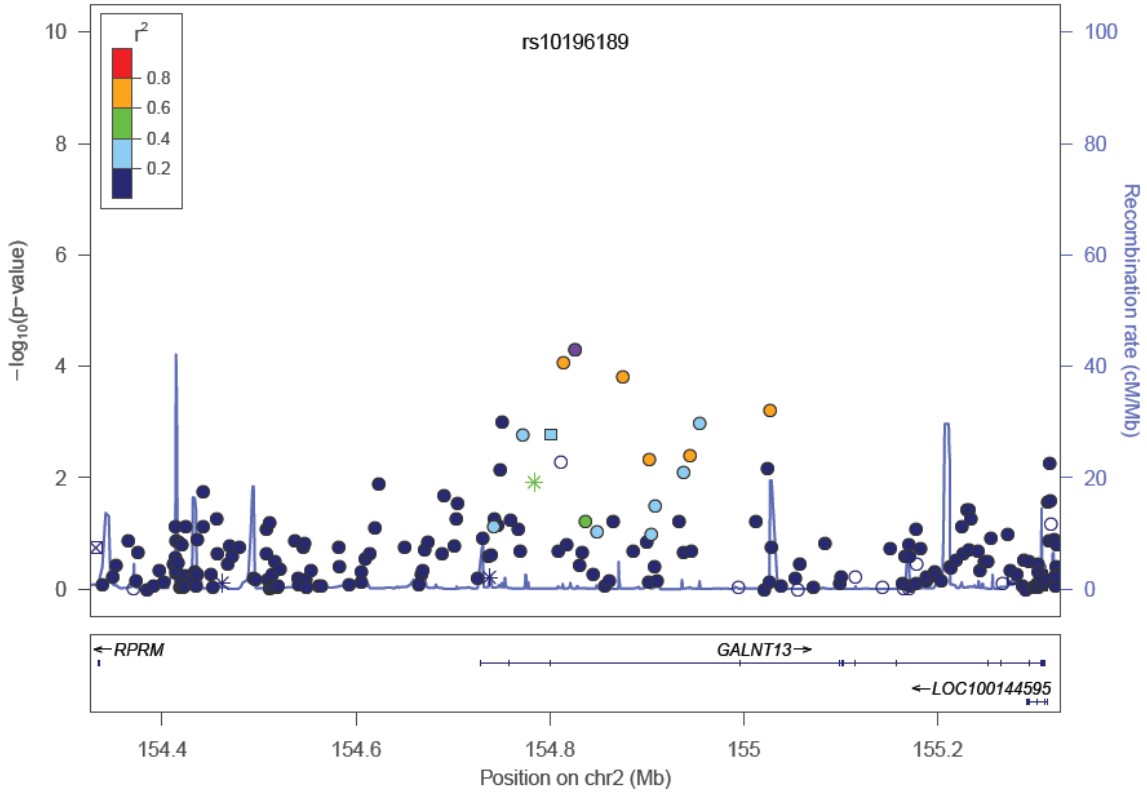


Figure 4.9 Regional association plot of rs10196189 (purple filled circle) from the Jamaican sprint GWAS with 500Kb flanking region on each side. –log10 transformed *p*-values on the Y-axis indicate the strength of the association with elite sprint status in the Jamaican cohort. The level of LD between rs10196189 and its surrounding SNPs as well as the recombination rate are estimated using 1000 Genomes AFR samples (Mar 2012). The level of LD is indicated by the colour key with red corresponding to high LD, and the recombination rate is represented by the blue line. Functional annotation key: triangle = framestop/splice, inverted triangle = non-synonymous, square = synonymous/UTR, star = conserved transcription factor binding site, square with diagonal lines = region is highly conserved in placental mammals, circle = no annotation. RPRM: reprimo; GALNT13: UDP-N-acetyl-alpha-D-galactosamine:polypeptide N-acetylgalactosaminyltransferase 13.

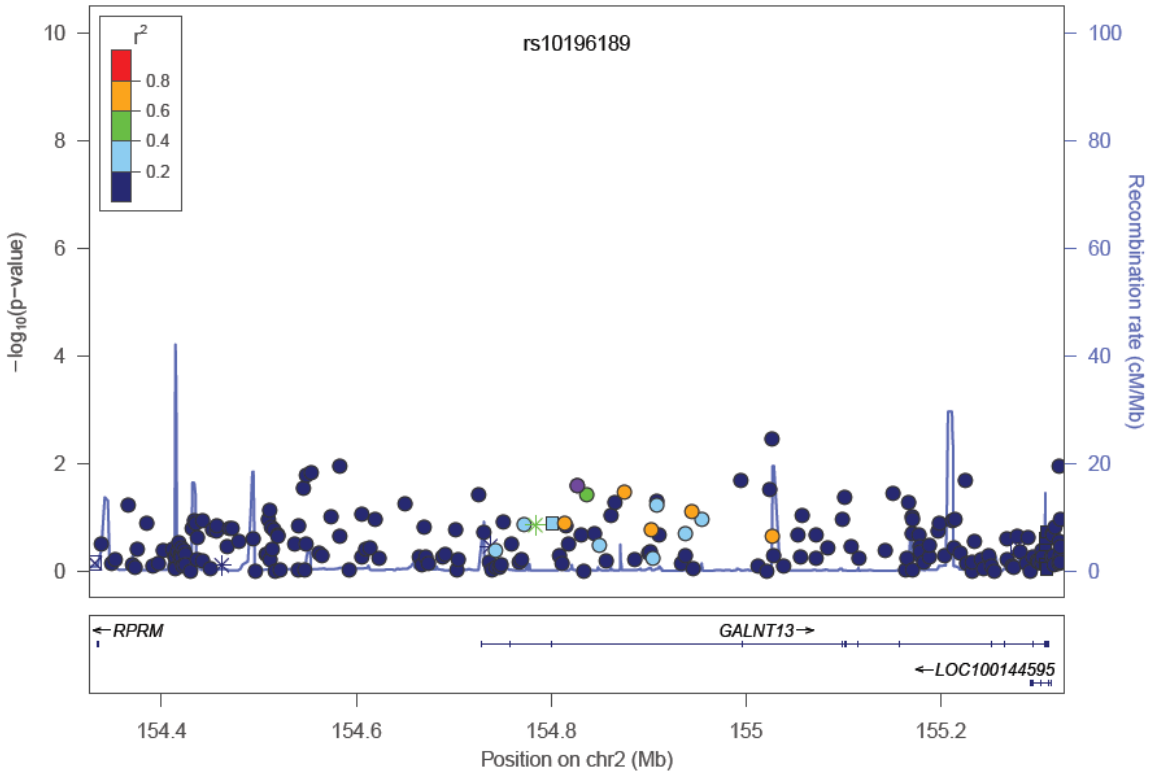


Figure 4.10 Regional association plot of rs10196189 (purple filled circle) from the African-American sprint GWAS with 500Kb flanking region on each side. –log10 transformed *p*-values on the Y-axis indicate the strength of the association with elite sprint status in the African-American cohort. The level of LD between rs10196189 and its surrounding SNPs as well as the recombination rate are estimated using 1000 Genomes AFR samples (Mar 2012). The level of LD is indicated by the colour key with red corresponding to high LD, and the recombination rate is represented by the blue line. Functional annotation key: triangle = framestop/splice, inverted triangle = non-synonymous, square = synonymous/UTR, star = conserved transcription factor binding site, square with diagonal lines = region is highly conserved in placental mammals, circle = no annotation. RPRM: reprimo; GALNT13: UDP-N-acetyl-alpha-D-galactosamine:polypeptide N-acetylgalactosaminyltransferase 13.

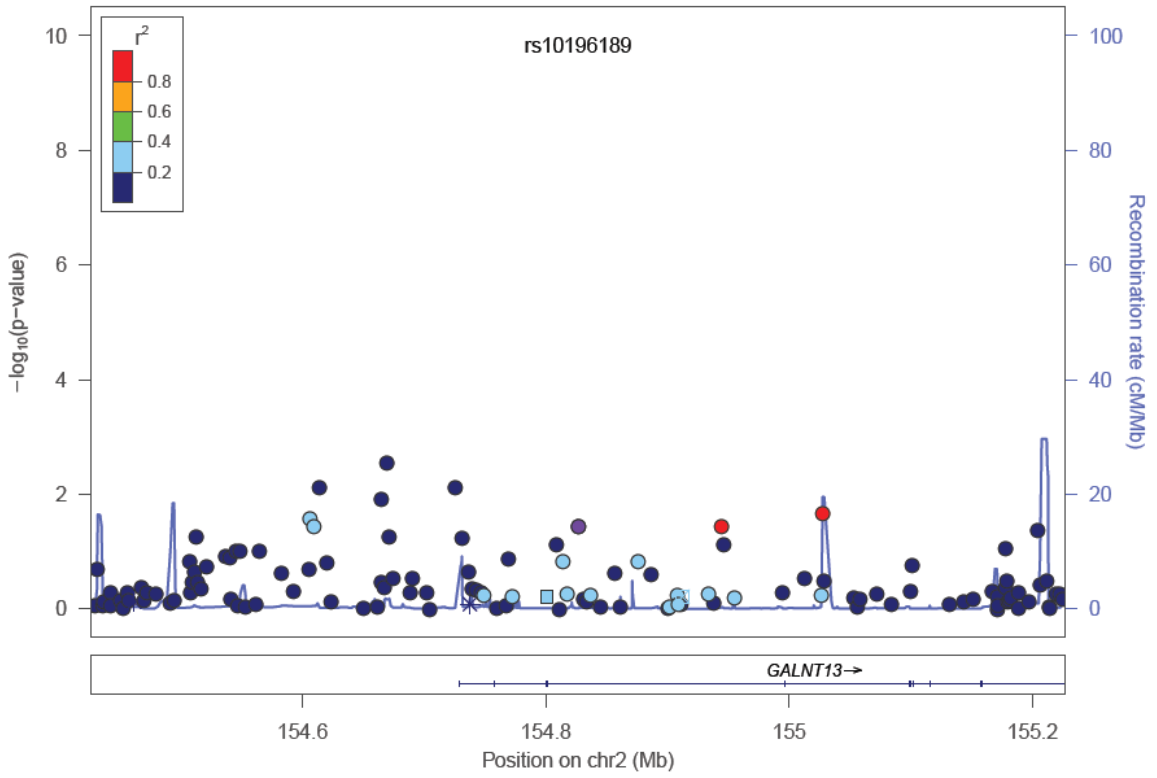


Figure 4.11 Regional association plot of rs10196189 (purple filled circle) from the Japanese sprint GWAS with 500Kb flanking region on each side. –log10 transformed *p*-values on the Y-axis indicate the strength of the association with elite sprint status in the Japanese cohort. The level of LD between rs10196189 and its surrounding SNPs as well as the recombination rate are estimated using 1000 Genomes ASN samples (Mar 2012). The level of LD is indicated by the colour key with red corresponding to high LD, and the recombination rate is represented by the blue line. Functional annotation key: triangle = framestop/splice, inverted triangle = non-synonymous, square = synonymous/UTR, star = conserved transcription factor binding site, square with diagonal lines = region is highly conserved in placental mammals, circle = no annotation. GALNT13: UDP-N-acetyl-alpha-D-galactosamine:polypeptide N-acetylgalactosaminyltransferase 13.

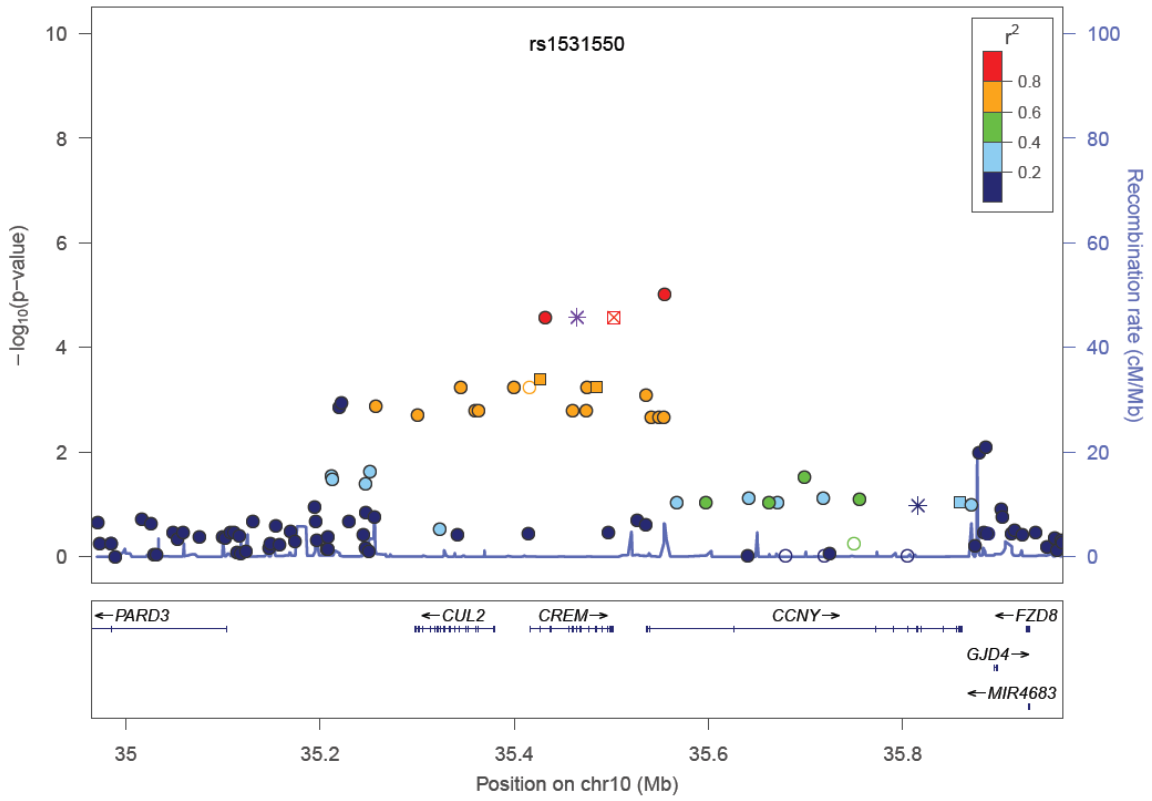


Figure 4.12 Regional association plot of rs1531550 (purple star) from the Japanese sprint GWAS with 500Kb flanking region on each side. –log10 transformed *p*-values on the Y-axis indicate the strength of the association with elite sprint status in the Japanese sprint GWAS cohort. The level of LD between rs1531550 and its surrounding SNPs as well as the recombination rate are estimated using 1000 Genomes ASN samples (Mar 2012). The level of LD is indicated by the colour key with red corresponding to high LD, and the recombination rate is represented by the blue line. Functional annotation key: triangle = framestop/splice, inverted triangle = non-synonymous, square = synonymous/UTR, star = conserved transcription factor binding site, square with diagonal lines = region is highly conserved in placental mammals, circle = no annotation. PARD3: par-3 partitioning defective 3 homolog (C. elegans); CUL2: cullin 2; CREM: cAMP responsive element modulator; CCNY: cyclin Y; FZD8: frizzled family receptor 8; GJD4: gap junction protein, delta 4, 40.1kDa; MIR4683: microRNA 4683.

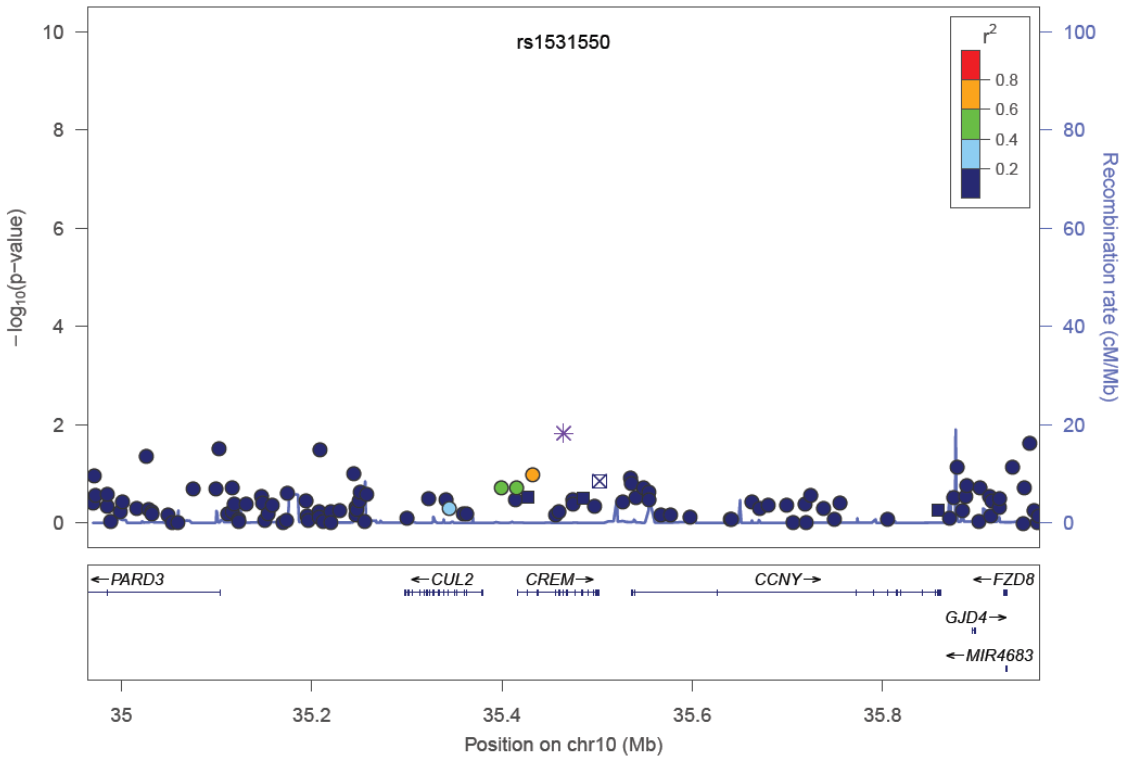


Figure 4.13 Regional association plot of rs1531550 (purple star) from the Jamaican sprint GWAS with 500Kb flanking region on each side. –log10 transformed *p*-values on the Y-axis indicate the strength of the association with elite sprint status in the Jamaican sprint GWAS cohort. The level of LD between rs1531550 and its surrounding SNPs as well as the recombination rate are estimated using 1000 Genomes AFR samples (Mar 2012). The level of LD is indicated by the colour key with red corresponding to high LD, and the recombination rate is represented by the blue line. Functional annotation key: triangle = framestop/splice, inverted triangle = non-synonymous, square = synonymous/UTR, star = conserved transcription factor binding site, square with diagonal lines = region is highly conserved in placental mammals, circle = no annotation. PARD3: par-3 partitioning defective 3 homolog (C. elegans); CUL2: cullin 2; CREM: cAMP responsive element modulator; CCNY: cyclin Y; FZD8: frizzled family receptor 8; GJD4: gap junction protein, delta 4, 40.1kDa; MIR4683: microRNA 4683.

4.1.4 Discussion

Four GWASs were carried out to identify common variations associated with elite sprint/power and endurance athlete status by use of elite sprint/power athletes from Jamaica, USA and Japan, and high level endurance athletes from Japan. Sample characteristics for these samples were specified in section 2.1.1. The association results from the four initial GWAS sample sets showed no overlap among the SNPs exceeding an unadjusted significance threshold of 5×10^{-5}. The meta-analysis results of combined effects in sprint GWAS cohorts revealed that 2 SNPs remained significant after taking into account the extra tests performed.

The phenotype of interest here is elite human performance. The true genetic architecture under this complex trait is unknown. Except that common variants of small to modest effect and rare variants of large effect may contribute to this phenotype variation, common variants of large effect as well as rare variants of small effect may also exist (Figure 4.14; ref(112)).

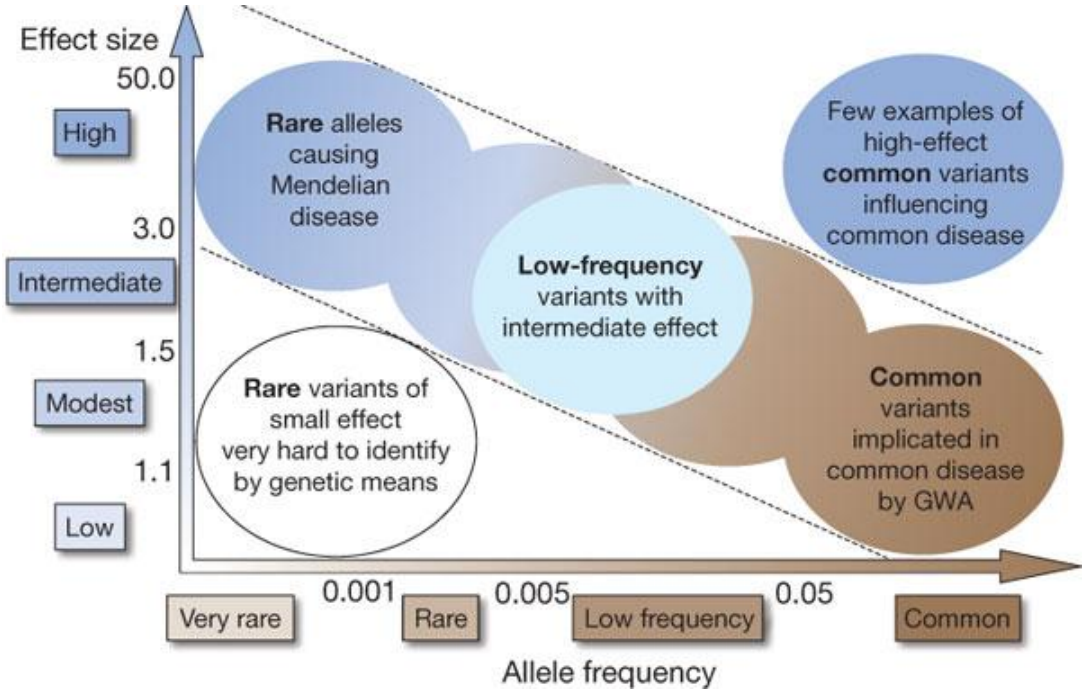


Figure 4.14 Feasibility of identifying genetic variants by risk allele frequency and strength of genetic effect (odds ratio) (112).

Unlike most of other common and complex traits, genetic study of a large and elite athlete cohort with the number of thousands is practically unachievable due to limited resources. Therefore, current GWASs were carried out in an attempt to increase the efficiency of identifying genetic variants in relation to elite athlete status by comparing allele frequencies in elite athletes who are at one extreme of the phenotype distribution to their geographically matched controls from the general populations. The allele frequency may be enriched in one or both phenotype extremes to circumvent the need for very large samples (136,145). In addition, the first GWAS in AMD revealed an intronic and common variant (with an effect size of 7.4) significantly related to AMD by comparing 96 cases to 50 controls, and subsequently a functional polymorphism (in LD with the risk allele of this common variant) in the *CFH* gene was identified by resequencing (59), suggesting that it is not unreasonable to expect that variants of large effect may be detected in a small study. As described in section 1.3.1, power is a function of sample size, effect size, correlation between the marker and the causal variant as well as their allele frequencies. When effect size of a variant is large and the variant is frequent enough to be detected in a population, the power for detecting it is also increased. In section 2.4.2.1, power was estimated for current initial GWAS sample sets based on simulation on 200 samples (100 cases and 100 controls), in which the relationship between power and effect sizes, with marker MAF varying from 0.05 to 0.5, was examined, assuming low prevalence of the trait at 0.1 (see Figure 2.2). Under these conditions, to achieve 80% power, minimum effect sizes range from 3.02 to 6.07 under the four different genetic models (i.e. the multiplicative, additive, dominant and recessive models). While power calculation is an important measure to ensure true discoveries in an association study, thorough data cleaning at the discovery phase and meta-analysis for combing several discovery studies (to increase the sample size, where applicable) are also very important and would help improve power. The estimated odds ratios from current discovery GWASs are likely to be inflated due to the inadequate power (e.g. due to small samples) in these initial GWASs (141-144). However,

as stated above, meta-analyses of these independent GWASs have greatly increased the sample size and statistical power. For example, the association of rs10196189 to elite sprint performance has been significantly improved after meta-analyses of the 3 initial sprint GWASs, with an odds ratio of 2.61 (p = 4.66 x10^{-7}) in 221 sprint athletes and 594 controls (i.e. size of meta-analyses samples). rs10196189 is an intron-variant located in the *GALNT13* gene on chromosome 2, which encodes glycosyltransferase that initiates mucin-type O-glycosylation (223). The *GALNT13* gene is also conserved across species, e.g. chimpanzee, dog, cow, mouse, rat, chicken and zebrafish (see: http://www.ncbi.nlm.nih.gov/gene/114805). Variations in introns of this gene have been reported to be associated so far with menopause, sudden cardiac death, echocardiography, erythrocyte indices, and blood pressure, and these associations are catalogued in the Phenotype-Genotype Integrator (PheGenl) database housed by NCBI (see: http://www.ncbi.nlm.nih.gov/gap/phegeni?tab=1&gene=114805). Several intergenic SNPs have been found to be related to coronary artery disease, tunica media and electrocardiography (see: http://www.ncbi.nlm.nih.gov/gap/phegeni?tab=1&gene=114805). There is the possibility that rarer causal variants of large effect correlate with common variants identified by GWAS, and because the *GALNT13* is not a large gene (i.e. 506Kb, see: http://omim.org/entry/608369), it is probably sensible to have deep sequencing for the whole gene region, on the premise that the GWAS tagSNP(s) (i.e. rs10196189 in this case) is replicated, in order to capture more detailed variation structure in this region that may influence the trait of interest (i.e. elite sprint performance in this case). Another meta-analysis-signal of interest is rs1531550, which is a conserved transcription factor binding site on chromosome 10 in the *CREM* gene (see: http://omim.org/entry/123812), which regulates the transcription of cAMP-responsive genes and may be involved in the regulation of cardiac gene expression (224). Non-coding SNPs in the *CREM* gene or the

surrounding SNPs have been found to associate with Parkinson disease (see: http://www.ncbi.nlm.nih.gov/gap/phegeni?tab=1&gene=1390).

Despite successful GWAS in identifying novel genetic variants for AMD (59), T2DM (225), the interleukin 23 pathway in Crohn's disease(226) and obesity-related traits (227), important limitations have also been noticed. SNPs identified by GWAS have typically explained a small fraction of the heritability. For example, human height is a highly heritable quantitative trait (up to 90% of variation most likely explained by genetic factors) (228-231) as well as stable and easy to measure. One of the largest studies (n = 183,727) to date identified at least 180 loci associated with adult height (explaining only 10% of the phenotypic variation in height) (232). There have been suggestions that common variants do explain up to 45% of the variance in height (233), but the small effect size of these variants may render these variants undetectable by common study designs (112). Furthermore, the identified common variants associated with most complex diseases do not show predictive utility (234). Thus, much of the heritability of complex traits remains hidden or missing (112,235), and several explanations have been proposed to account for this missing heritability (115) – the CDCV model, the infinitesimal model, the rare allele model, the broad sense heritability model (as previously described). Additionally, structural variation (e.g. CNVs, inversions, translocations etc.) that has not been substantially investigated in relation to complex traits are also responsible for the missing heritability (112). Although common variants of small to modest effect size are likely to be detected by GWAS, there are no published papers to date on GWAS of elite human performance and true genetic architecture underlying elite athlete status is not known. Despite a study of 4488 adult British female twins showing athlete status is highly heritable (h^2 = 66%) (33), the proportion of phenotypic variation that can be explained by GWAS markers is currently unclear. For example, the heritability of elite sprint athlete status that can be explained by the two intronic meta-analysis hits, rs10196189 and

rs1531550 (yet to be proven to be true by independent studies), as discovered from current GWASs may be extremely limited given the polygenic nature of elite performance.

Nevertheless, GWAS has been able to detect many loci that implicate biologically related genes and pathways (232). At present, GWAS remains to be an effective way of investigating genetic variants associated with common diseases and complex traits. The association signals identified from GWAS may highlight the genomic regions harbouring other forms of variation, such as rare and structural variants. Following GWAS, additional approaches, such as fine mapping and sequencing, may be used to find other common variants with larger effect sizes than GWAS tagging SNPs and to identify rarer variants across GWAS loci. Through these efforts, it is hoped to fill up much missing part of heritability for the common/complex traits.

Elite athlete status is a phenotype with multiple genes contributing to a set of advantageous performance-related phenotypes, for example, genetic variants that are positively related to cardiorespiratory fitness (assessed by $\dot{V}O_{2max}$ obtained during a progressive intensity test before exhaustion) and muscle fibre type distribution (higher percentage of type I muscle fibres) would together contribute significantly to the make of an elite endurance athlete. Many GWASs conducted to date are based on qualitative diagnoses of cases and controls (236,237). It is acknowledged that genetic component is distributed quantitatively in these common/complex traits. In other words, common/complex traits such as elite athlete status are the extremes of quantitative traits (238), representing the quantitative extremes of continuous distributions of genetic risk. Genes found for elite athlete status in a case-control study will not only be associated with differences between athletes and controls, but with individual differences in sporting ability throughout the entire range of variation.

Most GWAS has been primarily focused on European populations. Research beyond European populations could also contribute significantly to the determination of genetic

variants in relation to multifactorial traits. For example, study of genetic variation in populations of African ancestry (i.e. with greater genetic diversity, (67)) may help understand why some diseases have a greater impact on some groups compared to others and find ways to deal with them more effectively. The methodological challenges of GWAS in African and African-derived populations are present owing to the high levels of genome variation and population structure in these populations (71). The Illumina OmniExpress and Omni1-Quad Beadchips were used for current GWAS. The company Illumina launched the Omni family of microarrays in 2009 using tagSNPs derived from the International HapMap and 1,000 Genomes Projects. Coverage for the two beadchips used here are better for CHB + JPT samples (HumanOmni1-Qaud chip: 76% and HumanOmniExpress chip: 74% with 1kGP marker MAF > 5%; HumanOmni1-Qaud chip: 63% and HumanOmniExpress chip: 62% with 1kGP marker MAF > 1%) and CEU samples (HumanOmni1-Qaud chip: 76% and HumanOmniExpress chip: 73% with 1kGP marker MAF > 5%; HumanOmni1-Qaud chip: 63% and HumanOmniExpress chip: 58% with 1kGP marker MAF > 1%). For YRI samples, the HumanOmni1-Qaud chip has a coverage of 48% and 31% for 1kGP markers with MAF > 5% and > 1%, respectively, and the two coverage values are 40% and 25% for the HumanOmniExpress chip. The shorter LD present in African populations produces higher genetic diversity compared to the non-Africans and results in the lower variation coverage rates as identified using the two Illumina chips described above, with a maximum number of markers just above 1 million selected from the whole genome. The newest version of HumanOmni5-Quad Beadchip provides a much more extensive coverage of the genome using > 4 millions of selected tagSNPs, the variation captured by this chip is thus significantly improved for YRI samples (i.e. 71% for 1kGP markers with MAF > 5%, and 58% for 1kGP markers with MAF > 1%). The two Illumina genotyping platforms used for current GWAS are the most recent releases at the time of study. The full content of the OmniExpress chip is contained within the HumanOmni1-Quad Beadchip. Most Jamaican and African-American athletes and

controls were randomly mixed and genotyped on the beadchips, following the central rule that cases and controls should be treated and analyzed under the same experimental conditions. However, at the end, only more athletes were genotyped in order to increase the number of athletes being studied and spend the limited budget on analysing more athlete samples, including 18 Jamaican and 58 African-American sprint athletes. After QCs, no substantial substructure was observed, and logistic regression model involving top PCs as covariates was used to further minimize any confounding effects that may be caused by different batches as well as population stratification. Japanese control and athlete samples were genotyped sequentially rather than in parallel. No evidence of strong effects caused by these experiments was found after QCs that were carefully conducted.

In summary, the genome-wide association analyses in Jamaican, African-American and Japanese cohorts uncovered a few subsets of SNPs associated with elite sprint/power athlete status at a significance level of $p < 5 \times 10^{-5}$ (unadjusted). The combined effects of the top sprint-related signals revealed that 2 SNPs, rs10196189 and rs1531550, remained significant after the meta-analysis. They are considered the most informative signals and could be the focus for follow-up replications in large number of subjects in the cohorts of the same ethnicity or carrying out multi-ethnic replications for both SNPs that present in two or more different ethnic groups.

4.2 A GWAS-derived investigation: genotype score approach in addition to common variations for prediction of elite athlete status

4.2.1 Characteristics of sprint-related SNPs from published reports and current GWAS data

The 2006-2007 update of the human gene map for fitness and performance-related phenotypes (48) and subsequent reviews in 2008-2009, 2010, 2011 and 2012 (49-52) with the focus on the strongest evidences in the field of exercise genomics were reviewed. 25 SNPs at 22 loci associated with muscle power/strength were identified in at least one previously published study and looked up in Jamaican sprint, African-American sprint, Japanese sprint and Japanese endurance GWAS data for current genotype score analysis. The common loci between the literature-reported SNPs and the current GWAS SNPs associated with elite sprint performance are listed in Table 4.8.

Table 4.8 Common loci associated with elite sprint performance-related phenotypes between the literature-identified SNPs and the GWAS SNPs.

Rs Number	Gene Symbol	Risk Allele (literature)*	Related Traits	Minor Allele[#] (HapMap CEU)	MAF[#] (HapMap CEU)	Minor Allele[$] (GWAS)	MAF (Jamaican sprint GWAS)	MAF (African-American sprint GWAS)	MAF (Japanese sprint/endurance GWAS)
rs1815739	ACTN3	C/T	Sprint athlete status	C	0.49	T	0.16	0.20	0.47/0.47
rs2854464	ACVR1B	A/G	Muscle strength	G	0.22	G	0.39	0.44	0.43/0.47
rs699	AGT	C/T	Sprint athlete status	C	0.41	T	0.16	0.16	0.19/0.18
rs1800169	CNTF	A/G	Muscle strength	A	0.16	A	0.05	0.04	0.21/0.19
rs11206244; rs12095080	DIO1 haplotype	TA	Muscle strength	T; G	T: 0.36; G: 0.10	T; G	T: 0.16; G: 0.18	T: 0.21; G: 0.20	-
rs2296135	IL15RA	A/C	Trainability of Lean mass	A	0.50	A (C)	0.23	0.25	0.45/0.46
rs1049434	MCT-1	A/T	Lactate transport capability	T	0.36	T	0.13	-	-
rs1800629	TNF	G/A	Muscle strength	A	0.17	A	0.15	0.13	0.02/0.02
rs7832552	TRHR	T/C	Lean body Mass	T	0.33	T (C)	0.18	0.19	0.44/0.45

***** Risk allele is underscored for each SNP identified from literature.

[#] Minor allele and MAF of HapMap CEU population are reported above, since most study populations from literature are of European ancestry so to provide some information on marker allele frequency variation relative to current GWAS discovery populations.

[$] Minor allele of each common SNP across current GWAS datasets (i.e. Jamaicans, African-Americans and Japanese). Note that for *IL15RA* and *TRHR* SNPs, the alternative allele is the minor allele (shown in bracket) for Japanese GWAS cohort.

4.2.2 Genotype score analysis

Genotype score was constructed on the basis of the number of risk alleles, which were inferred from previously identified sprint-related SNPs (derived primarily from the candidate gene association studies), assuming an additive effect. A score of 2 is assigned to the "optimal" genotype contributing to power-related phenotypes, and a score of 0 is assigned for the alternative homozygous genotype of the less "optimal" allele. For example, a genotype score of 2 would be assigned for CC genotype of *ACTN3* (see Table 4.8), a score of 0 would be assigned for *ACTN3* TT genotype and finally, a score of 1 would be for the heterozygous genotype of TC. The sum of genotype scores from the SNPs for each individual is then standardized, namely total genotype score (TGS). An equation for TGS calculation (239) is :

"TGS = (sum of all genotype scores / (the number of studied polymorphisms*2)) X 100"

Except the sampling distribution of Japanese sprint athlete is normal (p = 0.09), other sampled athletes and controls in the GWAS cohorts are not normally distributed for the TGS data, assessed by the Kolmogorov-Smirnov test ($p \leq 0.002$). Therefore, the differences of mean TGS between GWAS athletes and controls were examined using the non-parametric tests. The mean TGS across athletes and controls did not differ significantly in any of the GWAS sample sets (data shown in Figure 4.15 - 4.17 for Jamaicans, African-Americans and Japanese, respectively; asymptotic two-sided $p \geq 0.43$).

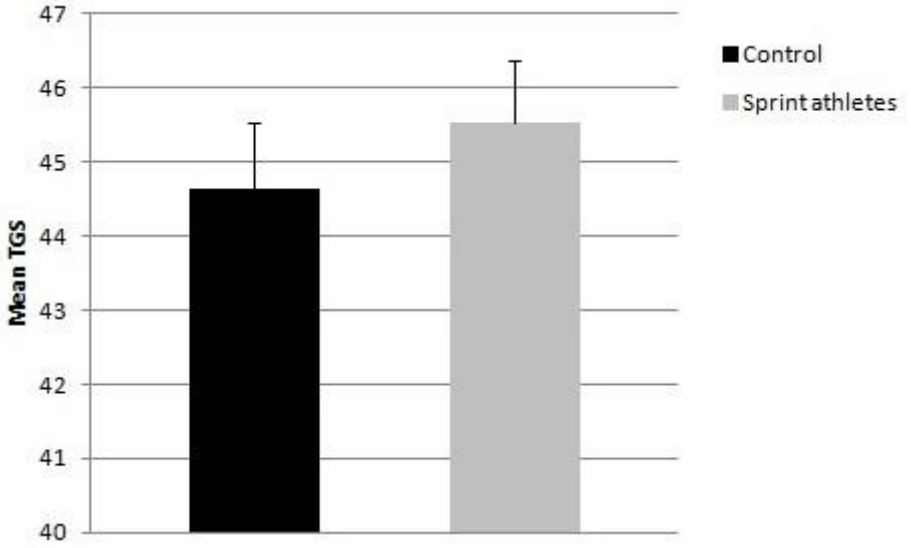


Figure 4.15 Mean TGS (± standard error) in Jamaican sprint athletes (45.5±0.8) and controls (44.6±0.9); *p* = 0.43 (two-sided).

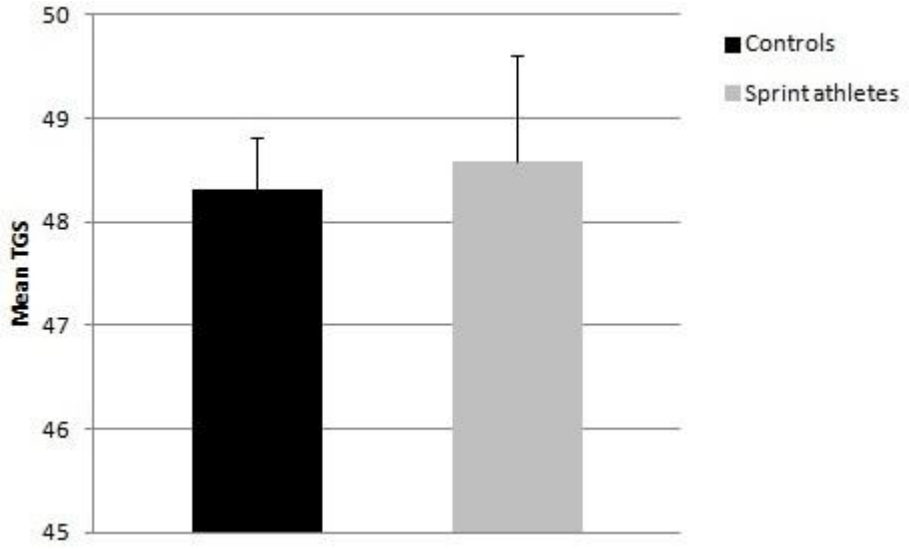


Figure 4.16 Mean TGS (± standard error) in African-American sprint athletes (48.6±1.0) and controls (48.3±0.5); *p* = 0.81 (two-sided).

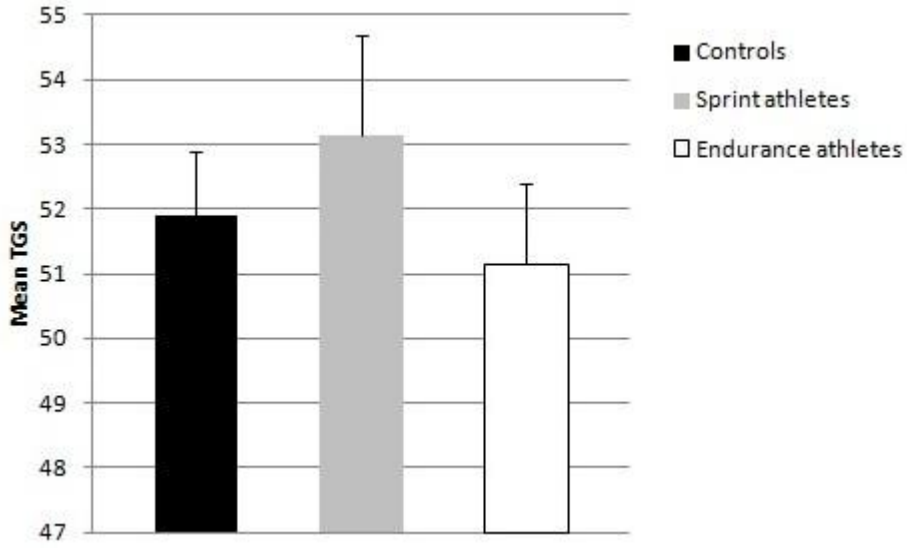


Figure 4.17 Mean TGS (± standard error) in Japanese sprint (53.1±1.5), endurance (51.2±1.2) athletes and controls (51.9±1.0); *p* = 0.62 (two-sided).

The frequency distributions of TGS across athletes and controls from each GWAS cohort were depicted in Figure 4.18. By visual inspection, in both Jamaican and African-American samples, frequency of TGS tends to be higher in controls than sprint athletes at "the lower score" end; and in African-Americans, a higher frequency of TGS is observed in sprint athletes relative to controls at "the higher score" end. In Japanese, at the lower scores, a higher frequency of TGS is noticed in endurance athletes compared to sprint athletes and controls; at the higher scores, sprint athletes have a higher frequency of TGS in comparison to that in controls and endurance athletes. In addition, Japanese control group shows a wider range of TGS towards the lower scores. However, the TGS distribution did not reveal any distinguishable pattern (not statistically significant) between athletes and controls of any of the studied GWAS cohorts, in other words, the distribution of TGS is the same across athletes and controls.

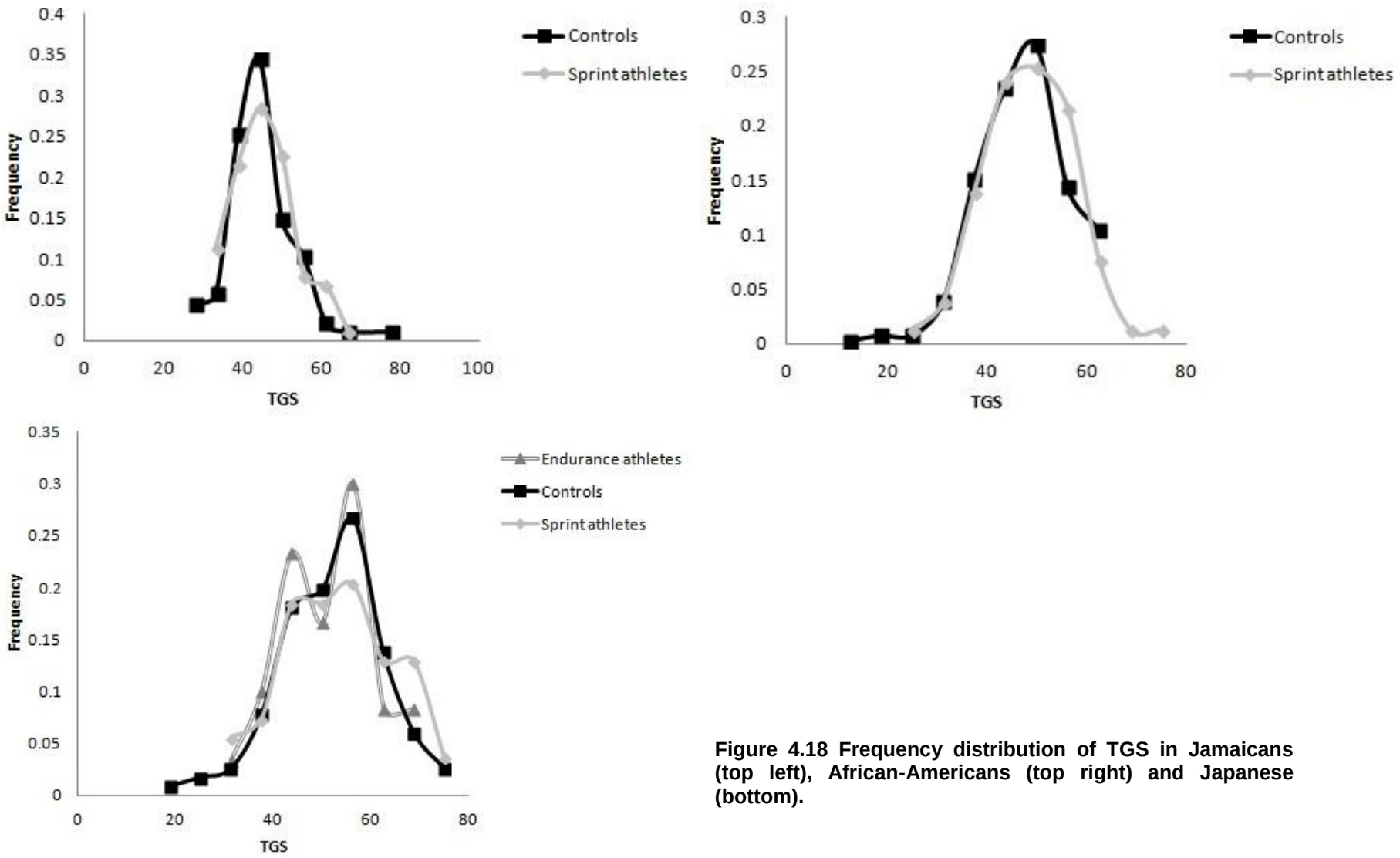


Figure 4.18 Frequency distribution of TGS in Jamaicans (top left), African-Americans (top right) and Japanese (bottom).

The proportion of advantageous genotypes based on TGS was also calculated. In Jamaicans, 3% or 22% of the elite sprint athletes possess a minimum of 5 or 4 (out of 9) optimal genotypes, respectively, relative to 3% or 24% in controls. In African-Americans, 24% of the elite sprint athletes had at least 4 (out of 8) optimal genotypes and this percentage is 21% in controls. In Japanese, 24% of the elite sprint athletes had more than or equal to 4 (out of 8) optimal genotypes, and these figures are 25% and 19% in Japanese endurance athletes and controls, respectively. Furthermore, none of athletes had a TGS of 100 in any of the GWAS cohorts. The ranges of TGS are of 33 – 67, 25 – 75, 31 – 75, and 38 – 69 in Jamaican sprint, African-American sprint, Japanese sprint and Japanese endurance athletes, relative to the score range of 28 – 78, 25 – 75, and 19 – 75 in their respective controls.

4.2.3 ROC curve

The ROC curve was used to interpret sensitivity (true positive rate) and specificity (true negative rate) levels of the genotype score approach in distinguishing elite sprint athletes from elite endurance athletes and/or controls. The AUC and 95% C.I. were calculated for the overall diagnostic accuracy of a ROC curve. A few features in a ROC curve to be noted include:

- Y-axis: Sensitivity = true positive rate; X-axis: 1- Specificity (true negative rate) = false positive rate.

- Red line: null line (null hypothesis: true AUC = 0.5). The closer to 1 of the AUC, the better. The point (0,1) indicates perfect classification.

The ROC analyses indicate that the TGS approach using literature-identified sprint-related SNPs has no predictive power in identification of an elite sprint athlete using the genomic data derived from current GWAS cohorts (Figure 4.19-4.21).

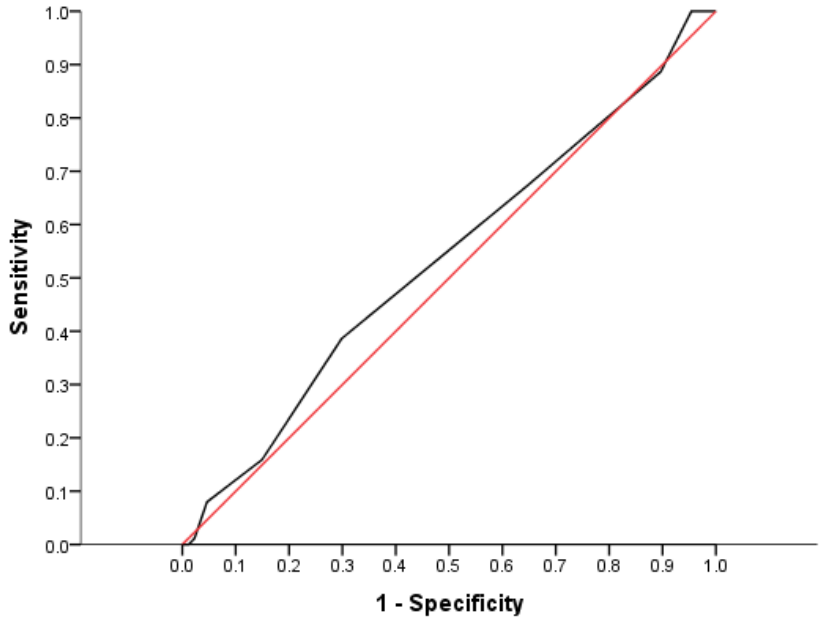


Figure 4.19 The ROC curve analysis for the reliability of TGS in distinguishing elite Jamaican sprint athletes from controls. AUC = 0.53, 95% C.I. = 0.45 – 0.62, *p* = 0.45.

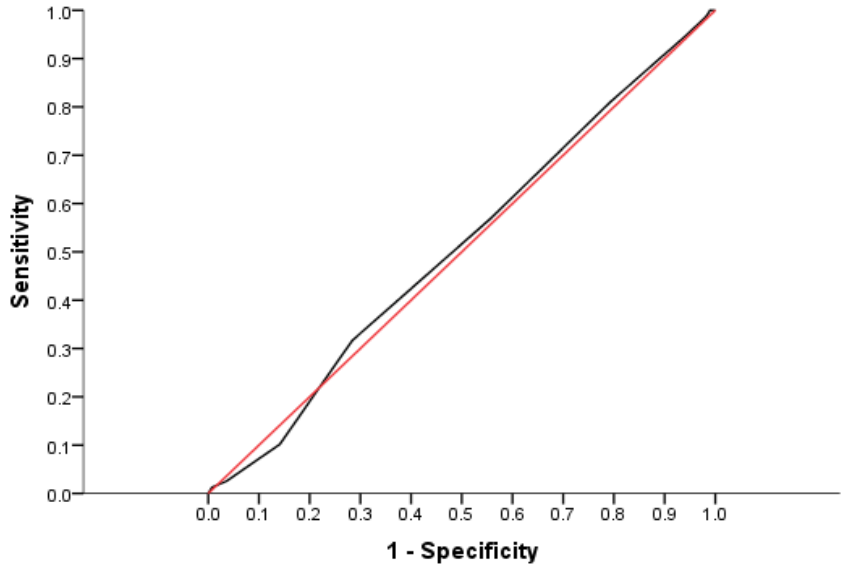


Figure 4.20 The ROC curve analysis for the reliability of TGS in distinguishing elite African-American sprint athletes from controls. AUC = 0.51, 95% C.I. = 0.44 – 0.58, *p* = 0.81.

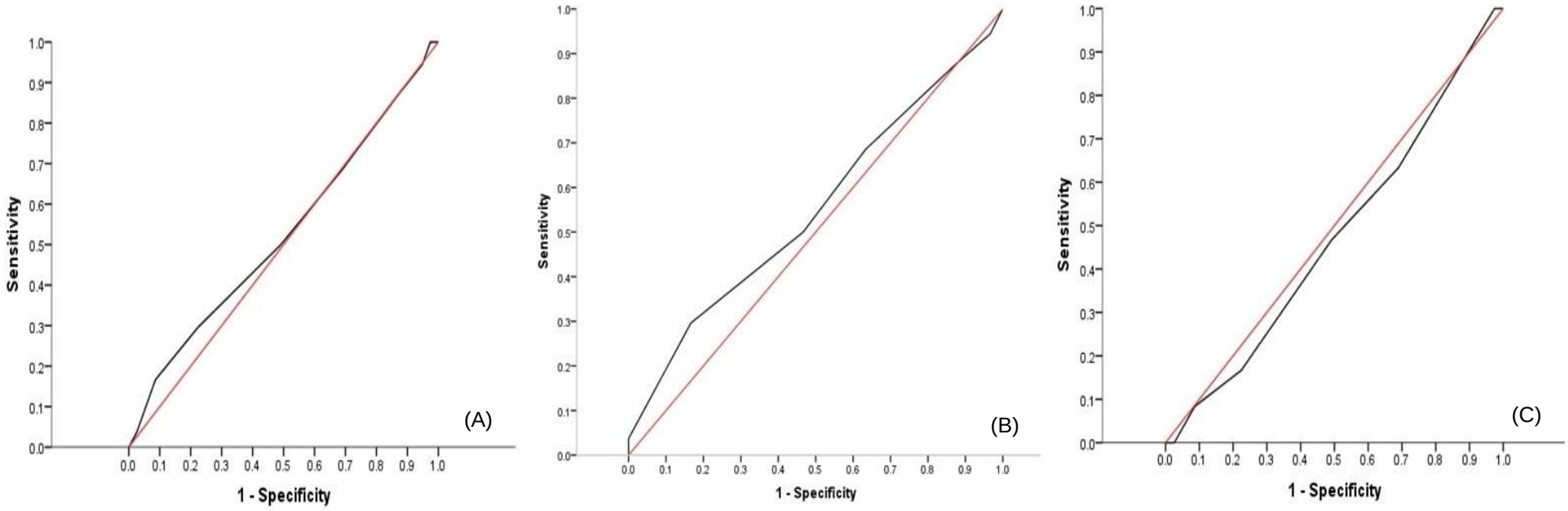


Figure 4.21 The ROC curve analysis for the reliability of TGS in distinguishing elite Japanese sprint athletes from controls (A), endurance athletes (B) as well as endurance athletes from controls (C). (A): AUC = 0.53, 95% C.I. = 0.43 – 0.62, *p* = 0.61; (B): AUC = 0.55, 95% C.I. = 0.44 – 0.66, *p* = 0.35; (C): AUC = 0.47, 95% C.I. = 0.38 – 0.56, *p* = 0.54.

4.2.4 Discussion

As reviewed in previous chapters, it is acknowledged that elite performance is a complex multifactorial trait, and genetic endowment is commonly perceived as one of the important factors contributing to it. A number of genes have been reported to be associated with elite performance by the primarily employed candidate gene approach. Knowing that multiple variants may contribute small amount of effects to elite performance, it is natural to combine alleles of several genes in order to explore further the underlying genetic architecture of the trait. Thus, the polygenic profile of the elite athletes currently being analysed under GWAS is assessed using the sprint-performance-associated genetic variants inferred from previously published reports.

The main findings are 1) mean TGS is not significantly differentiated between any GWAS athlete-control sample sets, or the distribution of TGS is the same across athletes and controls in each individual GWAS cohort. 2) ROC curves confirm that the TGS approach has no predictive power in talent identification as assessed by using current GWAS samples. 3) 3% and 22% of the elite Jamaican sprint athletes had more than or equal to 5 and 4 (out of 9) optimal genotypes, respectively, and these two number are 3% and 24% in Jamaican controls. 4) 24% of the elite African-American sprint athletes had more than or equal to 4 (out of 8) optimal genotypes, this percentage is 21% in controls. 5) 24% of the elite Japanese sprint athletes had more than or equal to 4 (out of 8) optimal genotypes, and 25% and 19% in Japanese endurance athletes and controls. 6) Finally, the optimum polygenic profile does not differ between sprint and endurance-oriented Japanese athletes in current analysis, which is, in fact, against a conventional notion that the genetic profiles of endurance and sprint performance are likely to be different owing to the variations in the phenotypic traits that determine performance in both types of events (240,241).

The use of a genotype score for polygenic profiling is not new. Williams and Folland (2008) (239) firstly introduced the TGS model in 2008; in this study, genotype score of 23 endurance-related genetic polymorphisms in 19 genes were computed to form an optimal polygenic endurance profile. However, 99% of the population (Caucasians) exhibited scores in the range of 37 – 65 and it has been estimated that the chance for one Caucasian individual (on the globe) possessing all 23 optimal alleles is extremely low (0.0005%). Similar studies carried out by Ruiz et al in 2009 (242) and 2010 (241) aimed to identify optimum endurance and power polygenic profiles using elite Spanish endurance and power athletes. In the 2009 study (242), 7 polymorphisms in 7 genes associated with endurance phenotypic traits were included in the TGS calculation for 46 elite endurance athletes and 123 controls. The average TGS is 70 in endurance athletes versus 62 in controls. In the 2010 study (241), TGS analysis of 6 power-oriented polymorphisms in 6 genes was performed in 53 elite power, 100 elite endurance athletes and 100 controls. The elite power athletes had a higher average TGS of 71 compared to 60 in endurance athletes and 63 in controls. Both studies have partially supported the concept of TGS on distinguishing different athletes groups (as well as controls). At the individual level, 60% of the elite power athletes had less than or equal to 3 (out of 6) optimal genotypes and 20% of the elite endurance athletes had 4 or 5 optimal genotypes (50). The small number of polymorphisms included in the TGS and the genuineness of these polymorphisms identified from literature may account for the high false negative rate at the individual TGS data of power athletes, and more importantly the low power of the study (50). Despite of a similar magnitude in terms of sample size between current GWAS cohorts used for TGS analyses and Ruiz et al studies (241,242), the negative observations from current TGS analyses using 9 previously reported sprint-related genetic loci may be explained, in part, by 1) the small number and validity of the power-related polymorphisms from previous studies, 2) varied allele frequencies and LD block structure across populations. Therefore, a genetic variant identified from one ethnic group (e.g.

Caucasians, from whom the selected SNPs are derived) might not well explain the underlying biology for other specific populations (e.g. Africans and Asians).

Biotechnology companies, such as Sports X factor and Atlas Sports Genetics, offered for genetic testing of athletic ability using a limited number of genes. For example, Sports X factor selects a panel of 7 genes, including *ACTN3*, *ACE*, *PPARGC*, *DIO1*, *VEGFR*, *NOS3*, and *IL6*, as performance indicators for individual genetic testing, whereas Atlas Sports Genetics looks at specifically the *ACTN3* gene. The development of current genetic profiling in elite performance is still in its infancy. The complex nature of this trait is yet to be understood and this would prevent any potential direct and effective genetic testing from using previously "identified" genetic markers from published reports. The lack of predictive utility of the TGS approach adopted here may also reflect the low reliability of the reported genetic variants predisposing to elite performance. Therefore, genetic testing in athletes is currently not recommended for coaches and athletes, who want to get such tests run (243).

To summarize, the results of current TGS analyses are not statistically significant, and inconsistent with previous investigations, in which TGS approach did show some predictive power on distinguishing athletes and controls. Interestingly, the genetic profiles (based on the 8 polymorphisms currently tested) of Japanese elite endurance and sprint athletes are not statistically significantly differed. This might partially be due to those reported optimum polymorphisms (examined here) derived mainly from studies of Caucasian populations may not well represent the exact genetic causation for a different population, and the small number of polymorphisms analyzed are not the causative SNPs or in good LD with the casual ones. Or perhaps, these SNPs studied here may be involved in certain biological pathways that would result in certain common phenotypes contributing evenly to both endurance and sprint performance; however, such mechanisms are yet to be found and understood.

5 General discussion and prospects

This thesis involved two types of association studies (i.e. candidate gene association study and GWAS) in an attempt to examine the genetic component of elite human performance. As such, a candidate gene association study of *ACE* and *ACTN3* was conducted in elite Caucasian and East Asian swimmers (see Chapter 3, starting page 77). The *ACE* I/D polymorphism was found to associate with elite swimmer status in both Caucasian SMD swimmers and East Asian SD swimmers. More specifically, DD homozygotes were found to be associated with SMD swimmer status in Caucasians, with the largest effect size observed for the I-allele dominant model (i.e. odds ratio was calculated for DD homozygotes relative to the I-allele carriers), while the I-allele homozygotes were found to be over-represented in East Asian SD swimmers under the D-allele dominant model (i.e. odds ratio was calculated for I-allele homozygotes with respect to D-allele carriers). *ACTN3* p.R577X was not significantly associated with swimmer status in any of the sample sets. However, there was a trend for *ACTN3* R-allele being modestly over-represented in the SD swimmers in East Asians, which is line with previous studies in sprint-/power-oriented sports. Detailed results for *ACE* and *ACTN3* associations with elite swimmer status can be found in section 3.3, page 86.

Notably, allele frequencies for *ACE* I/D vary across different populations, with a lower frequency of the D allele present in Asians (0.3 in Chinese, (210) and 0.4 in Japanese, (211)) relative to populations of African (0.56, (209)) and European (the average: 0.52, (148,204-206)) descent. *ACE* control data for the present thesis was obtained from a previous study (148), and the D allele frequency was 0.51 and consistent with above European populations (i.e. 0.52). Unlike the significant *ACE* association with Caucasian SMD swimmer (SD + MD) status, similar findings were not found in East Asian MD

swimmers despite of over-representation of I-allele homozygotes in East Asian SD swimmers. This finding may be because *ACE* I/D affects swimmers of varying swimming distances differently across populations. Alternatively and less likely, this lack of association in East Asian swimmers may be due to insufficient power to detect a difference. As already discussed in Chapter 3, page 78/79, *ACE* associations in opposing direction in the two ethnic groups are in line with previous studies in the same populations. For example, in Caucasians, the I allele has been reported to be associated with elite endurance performance in long-distance runners and rowers, and mountaineers at high altitude (172), and the D allele has been found to be associated with strength/power sports(e.g. sprinting (182) and swimming events of ≤ 400 m (148,171)). However, data from populations of East Asian descent have revealed that the D allele was associated with elite Japanese long distance runner status (187) and the I-allele with elite Korean power-oriented athlete status (188). Several explanations for the opposing effects of the *ACE* I/D alleles include: the same causative variant with different I-/D-alleles being on the predisposing haplotype more of the time in each group; different causative variants with I- and D-alleles being on different haplotypes more of the time in each group; or *ACE* affects related physiology differently in the two groups. There is less evidence currently to support the first two explanations (see Chapter 3, page 92), while the third explanation exists, where higher ACE activity may have an impact on swimming performance of shorter distance in one population and lower ACE activity may have the same impact in the other population since the two populations diverged approx. 30-35,000 yrs ago.

The *ACTN3* polymorphism has a modest effect on muscle fibre distribution (50), and carriers of the XX genotype do not express *ACTN3* in the muscles (193). The XX genotype is reported to be at a lower frequency in sprint/power athletes in previous studies (149,195). It is worth noting that there is a lack of relationship between *ACTN3* and elite swimmer status in the current study. It may be due to *ACTN3* polymorphism not being of

particular importance in swimming or type II errors. The sample size of this study is relatively modest, although it is the largest elite swimmer sample yet assembled (current: 200 Caucasian swimmers and 326 East Asian swimmers vs. other studies: 35–120, (148,171,244)). Despite numerous candidate gene studies carried out in the field of exercise science, candidate gene approach has produced inconsistent results in studies employing relatively small sample sizes and neglecting multiple testing adjustments. Other confounding factors may also exist, such as wrong causative SNPs selected, population stratification, phenotypic and locus heterogeneity (discussed in section 1.2.3, page 28). Here, only two SNPs (*ACE* I/D and *ACTN3* p.R577X) were selected for association analysis and multiple-testing issue was carefully controlled, therefore the reliability of current results is believed to be improved to the greatest extent possible. Again, the findings from this study are interesting in light of the opposing effects of *ACE* on elite swimmer status in Caucasians and East Asians (pointing out the possibility that this SNP may not be the key candidate seriously implicated in human performance in general and the causal variant(s) that affect(s) *ACE* activity through the I/D polymorphism might locate outside of the *ACE* gene region, perhaps in a nearby gene) as well as the lack of association between *ACTN3* and elite swimmer status in both populations (possibly because swimmers may require different components to excelling at power-dominated swimming events relative to other sprint/power sports that may be largely influenced by the *ACTN3* polymorphism). Nevertheless, these findings should be interpreted with caution until confirmed by larger studies.

GWAS is a preferred unbiased approach to identify genes contributing to elite sporting ability; however, a traditional GWAS would require very large “case” cohorts, which would preclude a study design involving established world-class athletes from similar sporting disciplines. On the other hand, the use of elite athletes who are at one extreme of the phenotype distribution might circumvent the need for very large cohorts since the allele

frequency may be enriched in the group of high level athletes, hence increased chance of finding common variants of larger effect in relatively small samples (145). The application of meta-analysis across several independent GWASs is also a good practice for increasing the overall sample size and improving statistical power to identify true associations (73). In terms of this thesis, the present GWASs were carried out in an attempt to identify common variations associated with elite sprint/power and endurance status in Jamaicans, African-Americans and Japanese, respectively (see Chapter 4, starting page 96). Meta-analyses were then performed for SNPs with unadjusted association $p < 5 \times 10^{-5}$ across the sprint GWAS sample sets (i.e. Jamaican sprint, African-American sprint, Japanese sprint GWAS cohorts). 2 SNPs remained significant after adjusting for the additional tests done given the meta-analysis (see section 4.1.3, page 113). Both SNPs are common intronic-variant with an intermediate meta-analysis fixed-effect size between 2.6 and 3 (see section 4.1.3, Table 4.7). The "synthetic associations" theory implies that several rare causal alleles may be tagged by the same common variant, and therefore the true effect size and proportion of variance explained by the set of rare variants may be underestimated by a common tagSNP identified from a GWAS (212). This theory would have implications on interpretation of GWAS signals as well as on the design of follow-up studies. For example, fine-mapping studies would be straightforward for identifying common casual variants of large effect in small sample sizes by zooming into the candidate region that is prioritized by GWAS association signals of small to modest effect (212). However, if common polymorphisms show associations with phenotypes because of the "synthetic associations", targeted long range resequencing extending beyond the LD block of GWAS-identified common variants would be helpful in finding multiple causal variants of low frequency. This is because the "synthetic associations" could be due to rare variants that lie megabases (Mb) away from the common variants being identified by GWAS (245). Therefore, if rs10196189 and rs1531550 (see above) can be replicated by further studies, the next step would be to sequence large regions (e.g. ideally 10 Mb as recommended by Dickson et al 2010 (245))

around the discovery hits so as to obtain comprehensive sequence data on both common and rare casual variants potentially contributing to elite human performance. Furthermore, the cost of large-scale sequencing has dramatically dropped, from the first complete human genome costing $3 billion to sequence in 2000 to $1,000 per genome as promised by the company Ion Torrent (a division of Life Technologies) using the new benchtop Ion Proton sequencer with the Ion PI chip in 2012, and the cost of large-sequencing will become even cheaper over the next years. Undoubtedly sequencing will become the most important and widely used approach in the next generation of GWAS through targeted sequencing in the candidate genomic region in large cohorts, whole exome sequencing, and ultimately whole genome sequencing in a large number of subjects.

Despite key advances in molecular biology, heritability remains largely unexplained for most complex traits. In the GWASs of elite performance (see Chapter 4), out of > 1 million common SNPs, only two meta-analysis hits were identified (with some level of confidence) to be associated with sprint performance. Even if both hits can be replicated, the heritability that could be explained by these two SNPs would be extremely limited given the multiofactorial nature of sporting performance. Additional discoveries would be required on rarer associated variants, better understanding of the modes of inheritance and interactions of gene x gene and gene x environment, and refined heritability estimates (73). Moreover, it is highly likely that many common variants with small effect that cannot be captured using a standard GWAS design account for much of the "missing heritability". Some researchers (246) have tried to investigate whether genomic data from GWAS could be used to improve discrimination of complex disease affection status by applying the genomic score approach (or can be said as doing “genomic profiling”) to the Wellcome Trust Case Control Consortium (WTCCC) genome-wide data (237) of seven common diseases and looking at multiple genetic loci of small effect, which is very likely possessed by many common alleles, simultaneously. Interestingly, the authors found that profiling

using GWAS data tended to show the greatest prediction utility when a less stringent threshold was adopted for the additional SNPs to be included for the score calculation. This is in line with the possibility that there may be many variants of small effect spread widely across the genome reflecting true loci that do not meet the stringent levels required for the genome-wide statistical significance and these loci would contribute to at least some of the "missing heritability". In contrast, for some disease traits studied by WTCCC, the discriminative ability is the most reliable for SNPs at stringent thresholds, suggesting that most loci influencing these traits have been discovered, therefore, the remaining genome-wide data has little added value. The genome-wide scores are not constructed for GWAS cohorts studied in this thesis since the sample sizes are small, which would prohibit the many variants with smaller effects from being reliably detected. However, the genotype score approach has been used to examine the ability of previously reported sprint-associating SNPs (from literature) on discrimination of athlete-control status using the genome-wide data available from the present GWASs. Based on the genotype score analysis (see section 4.2, page 127), discrimination is very poor. The selected variants from published reports showed no predictive value at all in discriminating athlete-control status in current GWAS sample sets; therefore, there is little utility of such variants on athletic talent identification at the present time, at least this would apply to populations of West African and East Asian ancestries.

Future research involving large well-funded collaborations/consortia using large cohorts would be required to better understand genetic fundamentals of exercise performance. For instance, the IDEFICS (Identification and prevention of dietary- and lifestyle-induced health effects in children and infants) is an integrated project initiated with funding from the sixth Framework Programme of the European Commission (247). It is one of the largest single studies to investigate genetic and environmental factors in childhood obesity in 16,224 young children (ages 2–9 years). The IDEFICS project has been successful in

generating one of the largest DNA biobanks with multiple, high quality phenotype datasets collected from a large cohort of young children (including objectively measured physical activity levels using accelerometry). From the total number of samples collected using the whole saliva and the sponge collection methods, 4,678 samples were randomly selected for extraction (248). Both collection methods provided sufficient DNA yields and quality for large-scale genetic epidemiological studies (248). The saliva-based/buccal-cell-based collection method is relatively inexpensive, convenient, and noninvasive in comparison to blood sampling for DNA collection (249). It is worth noting that, in the present GWASs described in this thesis, DNA isolated from buccal cells have also shown a high utility for genotyping at a genome-wide scale with high and consistent successful genotyping rate present (typically, > 99%) across the genotyped samples.

In order to understand the function of loci underlying complex traits, the genomic region harbouring a genetic variant contributing to gene expression variation is of particular interest in recent years (250-254). This is achieved by studying variable transcription levels among individuals through expression association mapping (eQTLs, expression quantitative trait loci) by treating transcript abundance as a quantitative trait (73). The complex trait associated SNPs identified from GWAS that are also associated with quantitative transcript levels of eQTL variants may help to shed light on mechanisms of the underlying biology. Findings from previous studies contribute to evidence that there is the overlap between the genetic and eQTL variants (255,256), and that GWAS trait-associated SNPs are significantly enriched for eQTLs comparing to MAF-matched SNPs (257). Transcription is also limited by structural variations (i.e. CNVs), insertion-deletion polymorphisms and short tandem repeats (258). Their (including SNPs) relations to the genome, transcripts and other functional data can be annotated through the VarySysDB database. Other publicly available datasets for eQTL studies include GENe Expression VARiation (Genevar, http://www.sanger.ac.uk/resources/software/genevar/, ref (259)) and

Genotype-Tissue Expression (GTEx; https://commonfund.nih.gov/GTEx/). Genevar is a database aimed for analysing SNP-gene associations in eQTL studies, through studying eQTL association patterns within a genomic region of interest. However, only three tissue and cell types are available from a limited number of study cohorts that may not be relevant to other specific traits studies. The GTEx programme focuses on gene expression and regulation in multiple tissues, with the potential to be developed into the most comprehensive tissue bank for numerous studies in the future. Correlations between genetic variation and tissue-specific gene expression levels will be examined to provide insights into the mechanisms of gene regulation.

Other systems genetics approaches to integration of large sets of genetic variants (e.g. from GWAS results) with other data, such as methylation and miRNA regulatory networks, can also be used to aid identification of biology of a complex trait. DNA methylation (one of the epigenetic mechanisms) regulates gene expression. For example, more methylation near gene promoters correlates with no or low transcription, and this process significantly depends on cell type (260). The integration of methylation data with GWAS genotyping data may therefore help to understand the interplay between methylation state and genetic variations in driving the traits of interest. Most variants identified from GWAS do not appear to be functional themselves. It is very probable that the functional polymorphisms in LD with the GWAS SNPs are yet to be discovered. Some researchers (261), who utilized GWAS SNPs that alter miRNA seed sites (the most important region for binding and repression of mRNA by a miRNA), have successfully identified functional candidate SNPs in relation to traits/diseases. This has helped to prioritize GWAS candidate SNPs for follow-up functional studies. The investigation of gene x environment interaction contributing to a complex trait has not been widely carried out, due primarily to lack of information on environmental exposure variables. Also, large perspective cohorts are needed in order to facilitate the testing for gene x environment interaction. The gene x gene

interaction study would suffer from the increased multiple testing burden (262), hence it has not been examined by most GWAS or has only been tested on a limited number of well-established SNPs (263).

GWASs of complex traits have been mainly conducted in populations of European descent previously. GWASs in e.g. Africans and Asians are also emerging in recent years. GWASs in different populations are required as they would help to identify population-specific associations with causative mutations occurring after major ethnic groups migrated (73). Notably, many GWAS signals have been replicated across different ethnic populations (264,265). However, significant differences in allele frequency and lack of effect in one population relative to the other have also resulted in differences on GWAS signals between populations (266). Interestingly, it is suggested that signal mapping across multi-ethnic groups may greatly increase the power to detect associations (267). This is consistent with the current findings of the meta-analyses hit – rs10196189, for example. MAF of rs10196189 is low in Japanese sprint GWAS cohort (i.e. 0.06), whereas rs10196189 is frequently present in Jamaicans and African-Americans with MAF of 0.36 and 0.32, respectively. The statistical power is boosted for the detection of the association effect of rs10196189 in the meta-analysis samples in comparison to the effect observed in the Japanese samples only. This may be caused by genetic drift elevating allele frequencies of certain variants across different populations (267).

These molecular-based approaches will improve our understanding of the factors that limit physical performance in both health and disease. Gene transfer technology has been successfully applied to life-threatening diseases such as tumors (268,269), cardiomyopathies and muscular dystrophy (270,271), human severe combined immunodeficiency (272), and Parkinson's disease (273). Genes related to muscle metabolism may be used in the context of gene-based therapy for treating patients suffering from muscle atrophy or other skeletal-muscular diseases. On the other hand,

enhanced muscle function improved by gene therapy may be misused by individuals who try to obtain a competitive edge at all costs in sports. Similarly, genes related to oxygen or energy delivery that may be manipulated to enhance sports performance also have clinical implications in treating dialysis patients or lifestyle diseases, such as obesity and T2DM. As discussed throughout this thesis, current genetic architecture underlying elite athletic performance is unclear, despite great effort made and being made in search for performance-related markers. The predictive utility of established SNPs (from literature) is extremely low, preventing drawing any conclusions too strongly given current state of knowledge. To fight with misuse of genetic information in gene doping, it is preferable to be proactive and to develop substantial understanding of the biological mechanism underlying high level performance.

Finally, both candidate gene and genome-wide association studies using current study designs have shed some light on the genetics of elite human performance. The very first positive findings through using an unbiased genome-wide approach, i.e. GWAS, are encouraging. Further studies are required to validate/replicate these findings. Functional annotation studies could be then carried out to explore further the fundamental biology underlying elite athlete performance before this knowledge can be used for translational medicine.

Appendix

A1.1 Explanatory notes for *ACE* and *ACTN3* association results in elite Caucasian and East Asian swimmers

Input and output files for the PTest permutation-based association analysis, to allow for replication of the analysis. Data under each model are presented in collapsed from, with one line per genotype type, indicating how many individuals were used in the input file by reporting 'n' for each genotype.

A1.1.1 PTest input and output files for a model testing *ACE* associations with swimmer status (SMD vs LD vs controls) in Caucasians.

Data in the input and output files are represented by tables. In the table for the input file, "Class" (1st column) indicates swimmer group, and the columns after that represent different "Features" - each indicates genotype under one of the three genetic models tested - 2nd column = additive allelic effects (genotypes coded as "0, 1, 2"), 3rd and 4th columns = two dominant models (genotypes coded as "0,0,1" and "0,1,1", respectively); 5th column (not required for the actual input file) 'Notes' = no. of individuals with each coding pattern (i.e. with each underlying genotype); these lines are replicated 'n' times in the actual input file, but are shown collapsed here for the sake of brevity. In the table for the output file, PTest only reports *P*-values for 'Features' with $P < 0.05$

PTest input:

Class	Add.	Idom.	Ddom.	Notes
SMD	0	0	0	n=24
SMD	1	0	1	n=50
SMD	2	1	1	n=51
LD	0	0	0	n=16
LD	1	0	1	n=31
LD	2	1	1	n=19
Control	0	0	0	n=301
Control	1	0	1	n=615
Control	2	1	1	n=332

PTest output:

----------Input information----------
Total number of features: 3
Number of classes: 3
Test statistic: Chi-square - Categorical data
Number of permutations: 99999
Significance level: 0.05
----------Permutation results----------

Feature	test_statistic	raw_P-value	Adj_P-value	Adj_P-value B&H	Adj_P-value_(PT)
Add	1.14E+01	2.21E-02	6.63E-02	3.31E-02	2.12E-02
Idom	1.14E+01	3.34E-03	1.00E-02	1.00E-02	3.28E-03

Number of features whose P-values were below significance level (0.05): 2
Number of features whose P-value was below significance level according to Bonferroni correction: 1
Number of features whose P-value was below significance level according to Benjamini and Hochberg (B&H) correction: 2
Number of features whose P-value was below significance level according to permutation test: 2
Adj P-value: adjusted P-value, based on Bonferroni multiple testing correction.
Adj P-value B&H: adjusted P-value, based on Benjamini and Hochberg multiple testing correction.
Adj P-value (PT): adjusted P-value, based on permutation test.

A1.1.2 PTest input and output files for a model testing *ACE* associations with swimmer status (SD vs MD vs controls) in East Asians

PTest input:

Class	Add	Idom	Ddom	Notes
SD	0	0	0	n=16
SD	1	1	0	n=58
SD	2	1	1	n=92
MD	0	0	0	n=12
MD	1	1	0	n=79
MD	2	1	1	n=69
Control	0	0	0	n=140
Control	1	1	0	n=544
Control	2	1	1	n=560

PTest output:

----------Input information----------
Total number of features: 3
Number of classes: 3
Test statistic: Chi-square - Categorical data
Number of permutations: 99999
Significance level: 0.05
----------Permutation results----------

Feature	test_statistic	raw_P-value	Adj_P-value	Adj_P-value B&H	Adj_P-value_(PT)
Add	9.91E+00	4.20E-02	1.26E-01	6.30E-02	4.31E-02
Ddom	6.95E+00	3.10E-02	9.30E-02	9.30E-02	2.99E-02

Number of features whose P-values were below significance level (0.05): 2
Number of features whose P-value was below significance level according to Bonferroni correction: 0
Number of features whose P-value was below significance level according to Benjamini and Hochberg (B&H) correction: 0
Number of features whose P-value was below significance level according to permutation test: 2
Adj P-value: adjusted P-value, based on Bonferroni multiple testing correction.
Adj P-value B&H: adjusted P-value, based on Benjamini and Hochberg multiple testing correction.
Adj P-value (PT): adjusted P-value, based on permutation test.

A1.1.3 PTest input and output files for a model testing *ACTN3* associations with swimmer status (SMD vs LD vs controls) in Caucasians

PTest input:

Class	Add	Rdom	Xdom	Notes
SMD	0	0	0	n=35
SMD	1	0	1	n=65
SMD	2	1	1	n=25
LD	0	0	0	n=29
LD	1	0	1	n=27
LD	2	1	1	n=12
Control	0	0	0	n=540
Control	1	0	1	n=840
Control	2	1	1	n=314

PTest output:

----------Input information----------
Total number of features: 3
Number of classes: 3
Test statistic: Chi-square - Categorical data
Number of permutations: 99999
Significance level: 0.05
----------Permutation results----------

A1.1.4 PTest input and output files for a model testing *ACTN3* associations with swimmer status (SD vs MD vs controls) in East Asians

PTest input:

Class	Add	Rdom	Xdom	Notes
SD	0	0	0	n=31
SD	1	1	0	n=78
SD	2	1	1	n=57
MD	0	0	0	n=39
MD	1	1	0	n=76
MD	2	1	1	n=45
Control	0	0	0	n=289
Control	1	1	0	n=640
Control	2	1	1	n=323

PTest output:

----------Input information----------
Total number of features: 3
Number of classes: 3
Test statistic: Chi-square - Categorical data
Number of permutations: 99999
Significance level: 0.05
----------Permutation results----------

A1.1.5 PTest input and output files for pairwise comparison *ACE* association with swimmer status (SMD vs controls) in Caucasians

PTest input:

Class	Add	Idom	Ddom	Notes
SMD	0	0	0	n=24
SMD	1	0	1	n=50
SMD	2	1	1	n=51
Control	0	0	0	n=301
Control	1	0	1	n=615
Control	2	1	1	n=332

PTest output:

----------Input information----------
Total number of features: 3
Number of classes: 2
Test statistic: Chi-square - Categorical data
Number of permutations: 99999
Significance level: 0.05
----------Permutation results----------

Feature	test_statistic	raw_P-value	Adj_P-value	Adj_P-value B&H	Adj_P-value_(PT)
Add	1.14E+01	3.36E-03	1.01E-02	5.04E-03	3.00E-03
Idom	1.14E+01	7.40E-04	2.22E-03	2.22E-03	5.40E-04

Number of features whose P-values were below significance level (0.05): 2
Number of features whose P-value was below significance level according to Bonferroni correction: 2
Number of features whose P-value was below significance level according to Benjamini and Hochberg (B&H) correction: 2
Number of features whose P-value was below significance level according to permutation test: 2
Adj P-value: adjusted P-value, based on Bonferroni multiple testing correction.
Adj P-value B&H: adjusted P-value, based on Benjamini and Hochberg multiple testing correction.
Adj P-value (PT): adjusted P-value, based on permutation test.

A1.1.6 PTest input and output files for pairwise comparison *ACE* association with swimmer status (LD vs controls) in Caucasians

PTest input:

Class	Add	Idom	Ddom	Notes
LD	0	0	0	n=16
LD	1	0	1	n=31
LD	2	1	1	n=19
Control	0	0	0	n=301
Control	1	0	1	n=615
Control	2	1	1	n=332

PTest output:

```
----------------------------------Input information-----------------------------
Total number of features: 3
Number of classes: 2
Test statistic: Chi-square - Categorical data
Number of permutations: 99999
Significance level: 0.05
------------------------------Permutation results-------------------------------
--------------------------------------------------------------------------------
```

A1.1.7 PTest input and output files for pairwise comparison *ACE* association with swimmer status (SD vs controls) in East Asians

PTest input:

Class	Add	Idom	Ddom	Notes
SD	0	0	0	n=16
SD	1	1	0	n=58
SD	2	1	1	n=92
Control	0	0	0	n=140
Control	1	1	0	n=544
Control	2	1	1	n=560

PTest output:

----------Input information----------
Total number of features: 3
Number of classes: 2
Test statistic: Chi-square - Categorical data
Number of permutations: 99999
Significance level: 0.05
----------Permutation results----------

Feature	test_statistic	raw_P-value	Adj_P-value	Adj_P-value B&H	Adj_P-value_(PT)
Add	6.43E+00	4.02E-02	1.21E-01	6.04E-02	4.06E-02
Ddom	6.38E+00	1.15E-02	3.46E-02	3.46E-02	9.79E-03

Number of features whose P-values were below significance level (0.05): 2
Number of features whose P-value was below significance level according to Bonferroni correction: 1
Number of features whose P-value was below significance level according to Benjamini and Hochberg (B&H) correction: 1
Number of features whose P-value was below significance level according to permutation test: 2
Adj P-value: adjusted P-value, based on Bonferroni multiple testing correction.
Adj P-value B&H: adjusted P-value, based on Benjamini and Hochberg multiple testing correction.
Adj P-value (PT): adjusted P-value, based on permutation test.

A1.1.8 PTest input and output files for pairwise comparison *ACE* association with swimmer status (MD vs controls) in East Asians

PTest input:

Class	Add	Idom	Ddom	Notes
MD	0	0	0	n=12
MD	1	1	0	n=79
MD	2	1	1	n=69
Control	0	0	0	n=140
Control	1	1	0	n=544
Control	2	1	1	n=560

PTest output:

```
----------------------------------Input information-----------------------------
Total number of features: 3
Number of classes: 2
Test statistic: Chi-square - Categorical data
Number of permutations: 99999
Significance level: 0.05
-----------------------------Permutation results-------------------------------
-------------------------------------------------------------------------------
```

A1.1.9 PTest input and output files for pairwise comparison *ACTN3* association with swimmer status (SMD vs controls) in Caucasians

PTest input:

Class	Add	Rdom	Xdom	Notes
SMD	0	0	0	n=35
SMD	1	0	1	n=65
SMD	2	1	1	n=25
Control	0	0	0	n=540
Control	1	0	1	n=840
Control	2	1	1	n=314

PTest output:

```
----------------------------------Input information-----------------------------
Total number of features: 3
Number of classes: 2
Test statistic: Chi-square - Categorical data
Number of permutations: 99999
Significance level: 0.05
------------------------------Permutation results-------------------------------
--------------------------------------------------------------------------------
```

A1.1.10 PTest input and output files for pairwise comparison *ACTN3* association with swimmer status (LD vs controls) in Caucasians

PTest input:

Class	Add	Rdom	Xdom	Notes
LD	0	0	0	n=29
LD	1	0	1	n=27
LD	2	1	1	n=12
Control	0	0	0	n=540
Control	1	0	1	n=840
Control	2	1	1	n=314

PTest output:

----------Input information----------
Total number of features: 3
Number of classes: 2
Test statistic: Chi-square - Categorical data
Number of permutations: 99999
Significance level: 0.05
----------Permutation results----------

A1.1.11 PTest input and output files for pairwise comparison *ACTN3* association with swimmer status (SD vs controls) in East Asians

PTest input:

Class	Add	Rdom	Xdom	Notes
SD	0	0	0	n=31
SD	1	1	0	n=78
SD	2	1	1	n=57
Control	0	0	0	n=289
Control	1	1	0	n=640
Control	2	1	1	n=323

PTest outputs:

----------Input information----------
Total number of features: 3
Number of classes: 2
Test statistic: Chi-square - Categorical data
Number of permutations: 99999
Significance level: 0.05
----------Permutation results----------

Feature	test_statistic	raw_P-value	Adj_P-value	Adj_P-value B&H	Adj_P-value_(PT)
Xdom	5.45E+00	1.96E-02	5.88E-02	5.88E-02	1.54E-02

Number of features whose P-values were below significance level (0.05): 1
Number of features whose P-value was below significance level according to Bonferroni correction: 0
Number of features whose P-value was below significance level according to Benjamini and Hochberg (B&H) correction: 0
Number of features whose P-value was below significance level according to permutation test: 1
Adj P-value: adjusted P-value, based on Bonferroni multiple testing correction.
Adj P-value B&H: adjusted P-value, based on Benjamini and Hochberg multiple testing correction.
Adj P-value (PT): adjusted P-value, based on permutation test.

A1.1.12 PTest input and output files for pairwise comparison *ACTN3* association with swimmer status (MD vs controls) in East Asians

PTest input:

Class	Add	Rdom	Xdom	Notes
MD	0	0	0	n=39
MD	1	1	0	n=76
MD	2	1	1	n=45
Control	0	0	0	n=289
Control	1	1	0	n=640
Control	2	1	1	n=323

PTest output:

```
---------------------------------Input information----------------------------
Total number of features: 3
Number of classes: 2
Test statistic: Chi-square - Categorical data
Number of permutations: 99999
Significance level: 0.05
----------------------------Permutation results-------------------------------
--------------------------------------------------------------------------------
```

A1.2 Observed *ACE* genotypes and allele frequencies in Caucasians and East Asians.

		Caucasian cohort			East Asian cohort		
Groups		**SMD**	**LD**	**Controls** [a]	**SD**	**MD**	**Controls**
Observed Genotype Counts, n (%)	**D/D**	51 (40.7)	19 (28.8)	332 (26.6)	16 (9.6)	12 (7.5)	140 (11.3)
	I/D	50 (39.8)	31 (47.0)	615 (49.3)	58 (34.9)	79 (49.4)	544 (43.7)
	I/I	24 (19.5)	16 (24.2)	301 (24.1)	92 (55.4)	69 (43.1)	560 (45.0)
Total		125	66	1248	166	160	1244
Allele Frequency	**D**	0.61	0.52	0.51	0.27	0.32	0.33
	I	0.39	0.48	0.49	0.73	0.68	0.67
HWE *P*-value		0.07	0.63	0.63	0.096	0.079	0.65

a. Caucasian control data were drawn from a previous published study (148).

A1.3 Observed *ACTN3* genotypes and allele frequencies in Caucasians and East Asians.

		Caucasian cohort			East Asian cohort		
Groups		**SMD**	**LD**	**Controls** [a]	**SD**	**MD**	**Controls**
Observed Genotype Counts, n (%)	**R/R**	35 (28)	29 (42.6)	540 (31.9)	57 (34.3)	45 (28.1)	323 (25.8)
	R/X	65 (52)	27 (39.7)	840 (49.6)	78 (47.0)	76 (47.5)	640 (51.1)
	X/X	25 (20)	12 (17.6)	314 (18.5)	31 (18.7)	39 (24.4)	289 (23.1)
Total		125	68	1694	166	160	1252
Allele Frequency	**R**	0.54	0.625	0.57	0.58	0.52	0.51
	X	0.46	0.375	0.43	0.42	0.48	0.49
HWE *P*-value		0.60	0.21	0.69	0.76	0.55	0.41

a. The total controls combined from five published *ACTN3* Caucasian controls (see A1.6).

A1.4 Observed *ACE* genotypes and allele frequencies in Japanese and Taiwanese, respectively.

		Japanese			Taiwanese		
Groups		**SD**	**MD**	**Controls**	**SD**	**MD**	**Controls**
Observed Genotype Counts, n (%)	**D/D**	7 (10)	8 (9.1)	79 (12.2)	9 (9.4)	4 (5.6)	61 (10.3)
	I/D	24 (34.3)	42 (47.7)	301 (46.4)	34 (35.4)	37 (51.4)	243 (40.8)
	I/I	39 (55.7)	38 (43.2)	269 (41.4)	53 (55.2)	31 (43.1)	291 (48.9)
Total		70	88	649	96	72	595
Allele Frequency	**D**	0.27	0.33	0.35	0.27	0.31	0.31
	I	0.73	0.67	0.65	0.73	0.69	0.69
HWE *P*-value		0.27	0.45	0.71	0.31	0.10	0.33

A1.5 Observed *ACTN3* genotypes and allele frequencies in Japanese and Taiwanese, respectively.

		Japanese			Taiwanese		
Groups		**SD**	**MD**	**Controls**	**SD**	**MD**	**Controls**
Observed Genotype Counts, n (%)	**R/R**	20 (28.6)	19 (21.6)	132 (20.3)	37 (38.5)	26 (36.1)	191 (31.7)
	R/X	37 (52.9)	41 (46.6)	346 (53.3)	41 (42.7)	36 (50)	294 (48.8)
	X/X	13 (18.6)	28 (31.8)	171 (26.3)	18 (18.8)	10 (13.9)	118 (19.6)
Total		70	88	649	96	72	603
Allele Frequency	**R**	0.55	0.45	0.47	0.60	0.61	0.56
	X	0.45	0.55	0.53	0.40	0.39	0.44
HWE *P*-value		0.57	0.58	0.07	0.28	0.66	0.80

A1.6 Observed *ACTN3* genotypes and allele frequencies in Caucasian controls drawn from 5 published studies.

		ACTN3 Caucasian controls				
Studies		**Yang et al. 2003**	**Lucia et al. 2006**	**Roth et al. 2008**	**Santiago et al. 2010**	**Ahmetov et al. 2010**
Observed Genotypes Counts, n (%)	**R/R**	130 (29.8)	35 (28.5)	218 (32.6)	90 (31.8)	67 (36.4)
	R/X	226 (51.8)	66 (53.7)	317 (47.5)	141 (49.8)	90 (48.9)
	X/X	80 (18.3)	22 (17.9)	133 (19.9)	52 (18.4)	27 (14.7)
Total		436	123	668	283	184
Allele Frequency	**R**	0.56	0.55	0.56	0.57	0.61
	X	0.44	0.45	0.44	0.43	0.39
HWE *P*-value		0.29	0.34	0.36	0.80	0.72
Chi-squared *P*-value		0.64 (Chi-squared statistic = 6.03)				

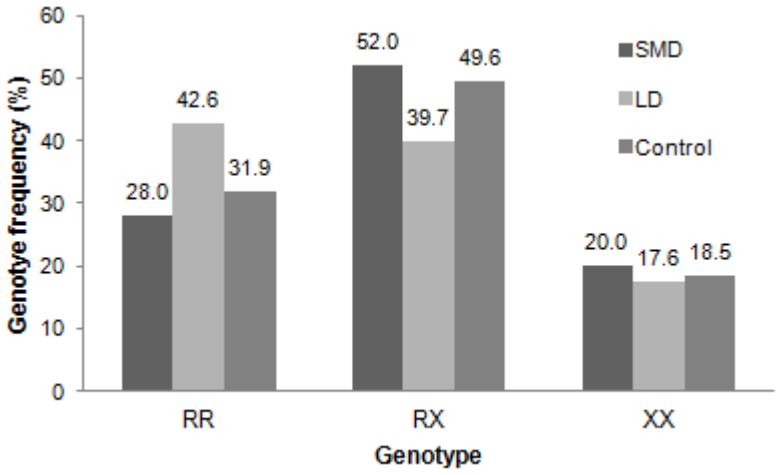
A1.7 Results of likelihood ratio tests examining the effect of 'genotype x ethnicity' interaction within association analysis models in the East Asian cohort.

	Likelihood Ratio Test		
	Chi-square	**d.f.**	**Significance *p***
ACE			
Intercept	.000	0	.
I-ADD	.000	0	.
Ethnicity	3.34	2	0.19
Ethnicity x I-ADD	**0.89**	**2**	**0.64**
Intercept	.000	0	.
D-DOM	.000	0	.
Ethnicity	6.91	2	0.032
Ethnicity x D-DOM	**1.74**	**2**	**0.42**
Intercept	.000	0	.
I-DOM	.000	0	.
Ethnicity	1.54	2	0.46
Ethnicity x I-DOM	**0.17**	**2**	**0.92**
ACTN3			
Intercept	.000	0	.
R-ADD	.000	0	.
Ethnicity	7.67	2	0.022
Ethnicity x R-ADD	**2.30**	**2**	**0.32**
Intercept	.000	0	.
R-DOM	.000	0	.
Ethnicity	9.71	2	0.008
Ethnicity x R-DOM	**4.34**	**2**	**0.11**
Intercept	.000	0	.
X-DOM	.000	0	.
Ethnicity	7.38	2	0.025
Ethnicity x X-DOM	**0.18**	**2**	**0.92**

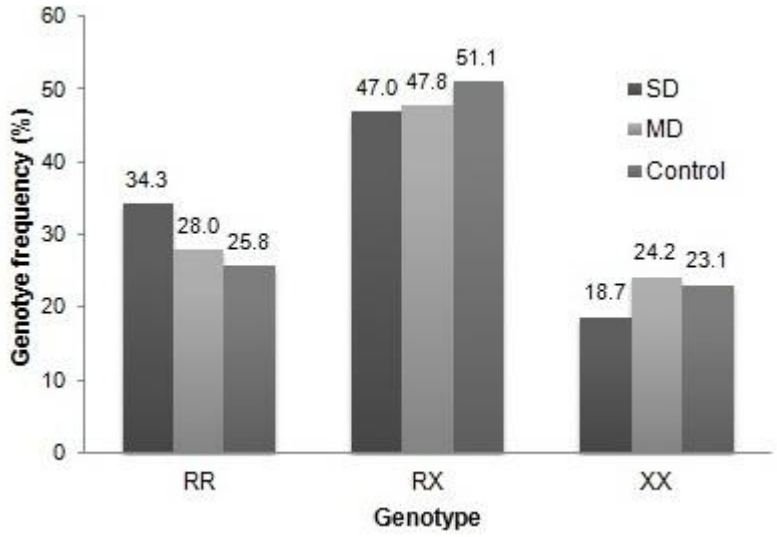
Models were of the form: Outcome (swimmer status) = error + genotype + ethnicity*(Japanese/Taiwanese)* + 'genotype x ethnicity'

Models with genotype coded for additive allelic effects and for both dominant effects were run separately.

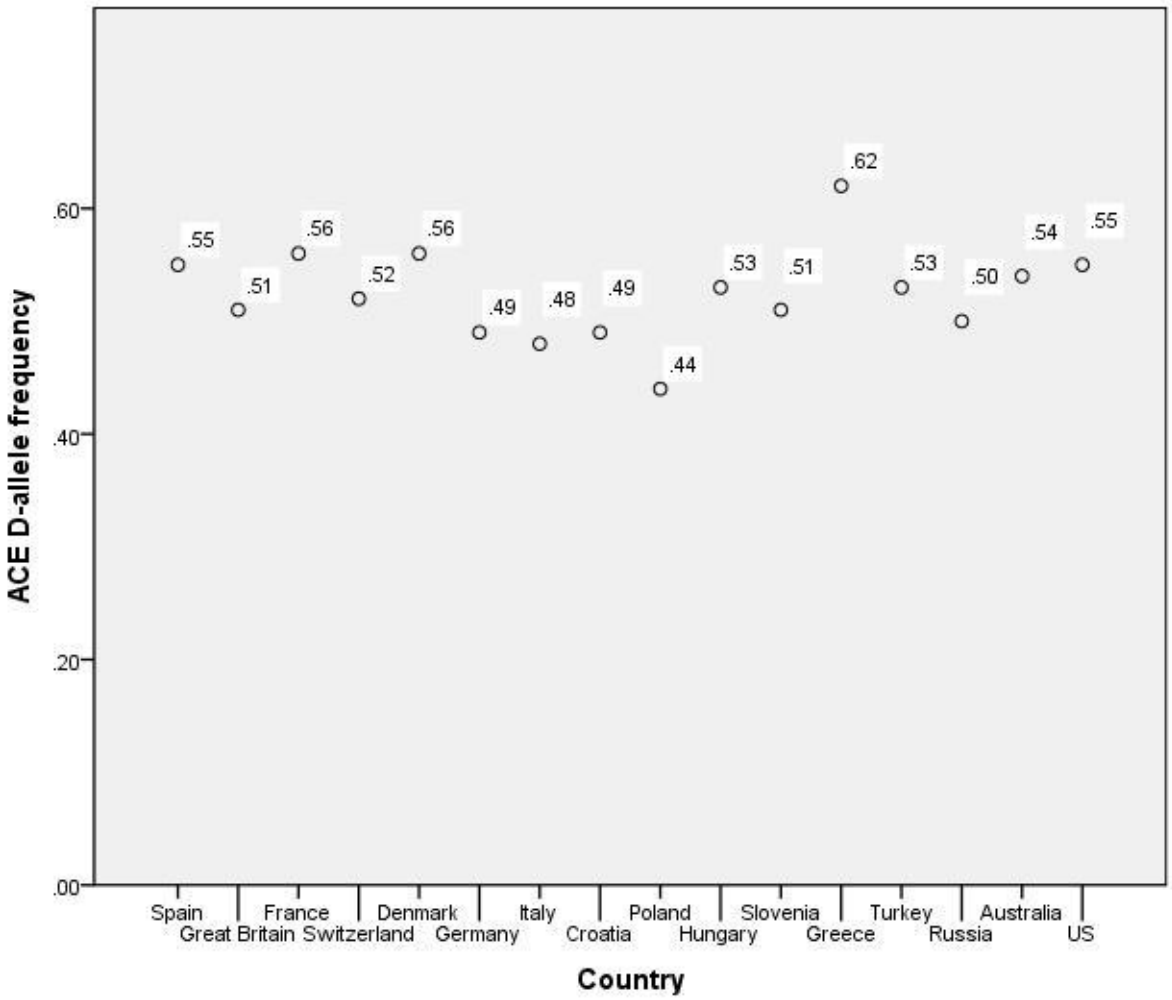
p-values are reported for all terms in the model for completeness; genotype and ethnicity *p*-values can be ignored as they are misleading in the presence of an interaction term. No significance values were generated for covariants (i.e. genotype) involved in higher-order interaction and returned as a '.'



A1.8 Genotype frequency distribution for *ACTN3* R577X in elite Caucasian swimmers and controls.

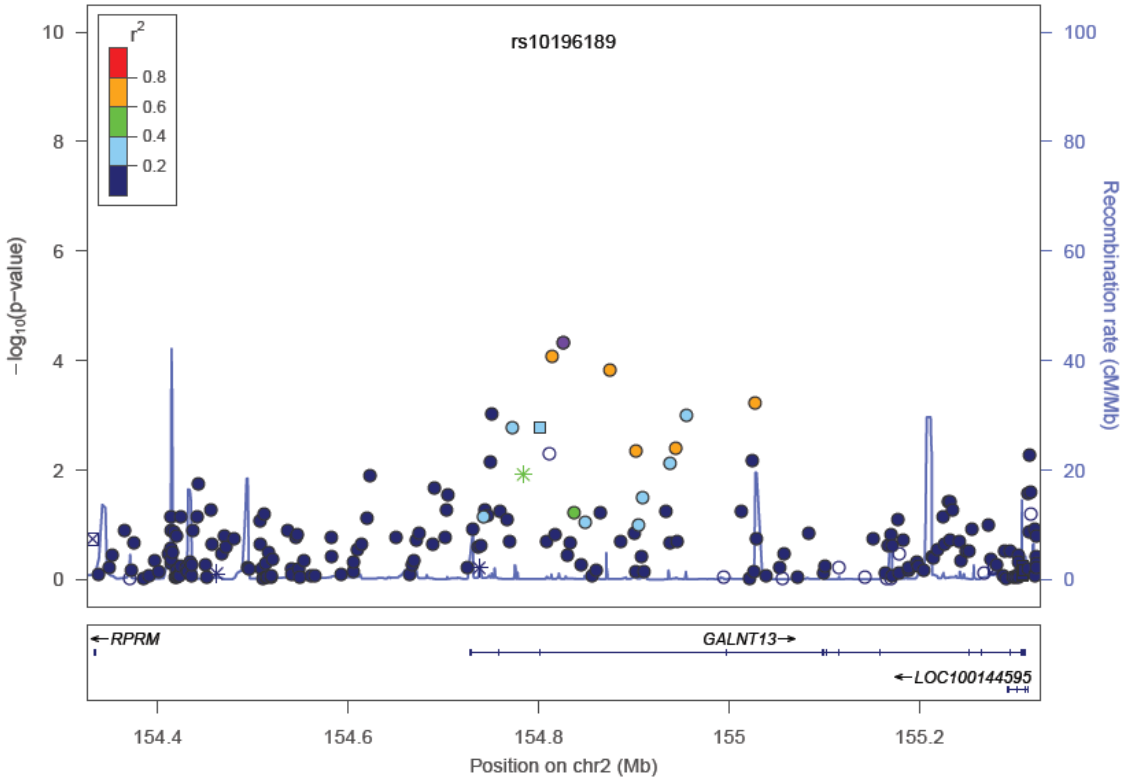


A1.9 Genotype frequency distribution for *ACTN3* R577X in elite East Asian swimmers and controls.

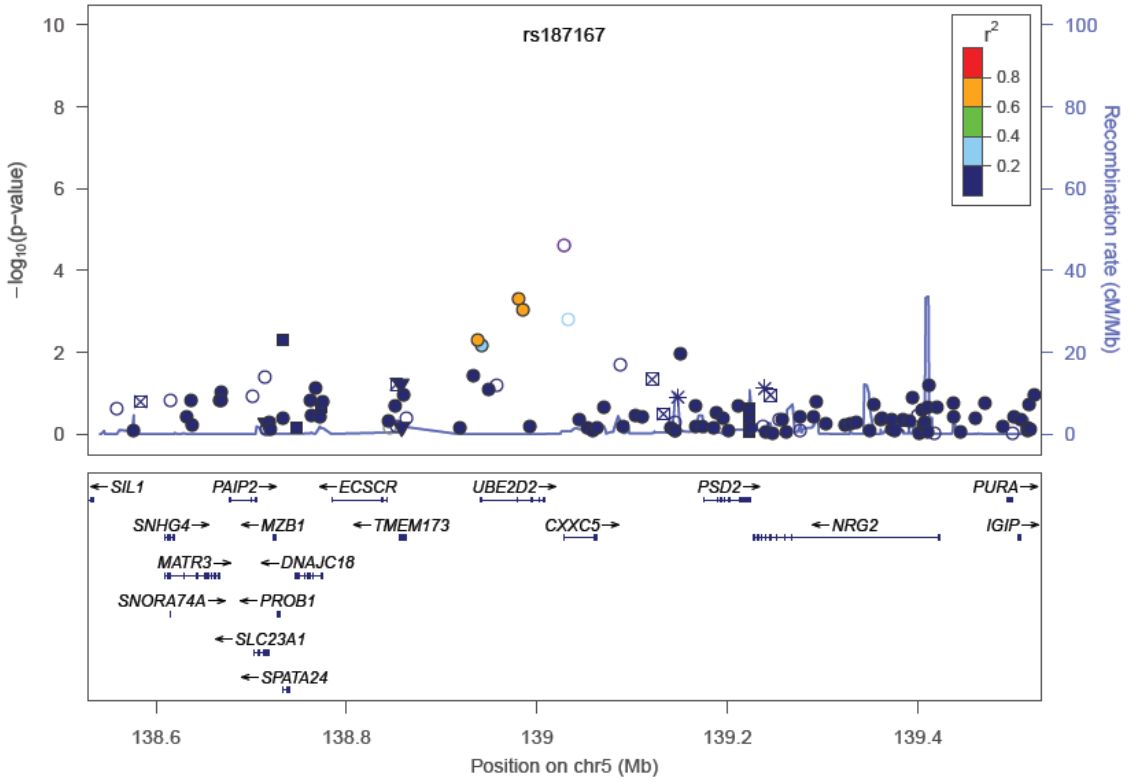


A1.10 *ACE* D-allele frequency distribution across Europe, the U.S. and Australia. Countries in Europe are displayed in longitudinal order from West to East.

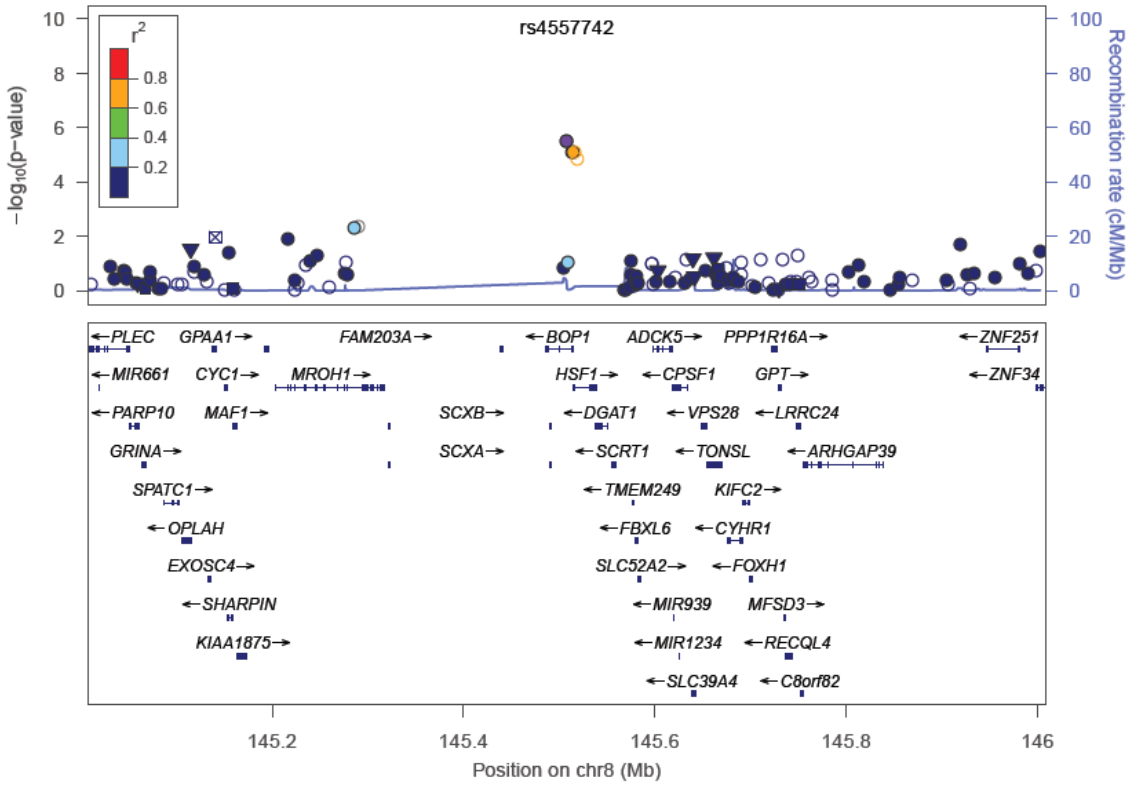
A2. Regional association plots of key markers (or index SNPs, in purple; see A2.1-A2.7) and 500Kb flanking region on each side of the markers for the Jamaican sprint cohort. –log10 transformed P values on the Y-axis indicate the strength of the association with elite sprint status in the Jamaican cohort. The level of LD between the index SNP and its surrounding SNPs as well as the recombination rate are estimated using 1000 Genomes AFR samples (Mar 2012). The level of LD is indicated by the colour key with red corresponding to high LD, and the recombination rate is represented by the blue line. Functional annotation key: triangle = framestop/splice, inverted triangle = non-synonymous, square = synonymous/UTR, star = conserved transcription factor binding site, square with diagonal lines = region is highly conserved in placental mammals, circle = no annotation.



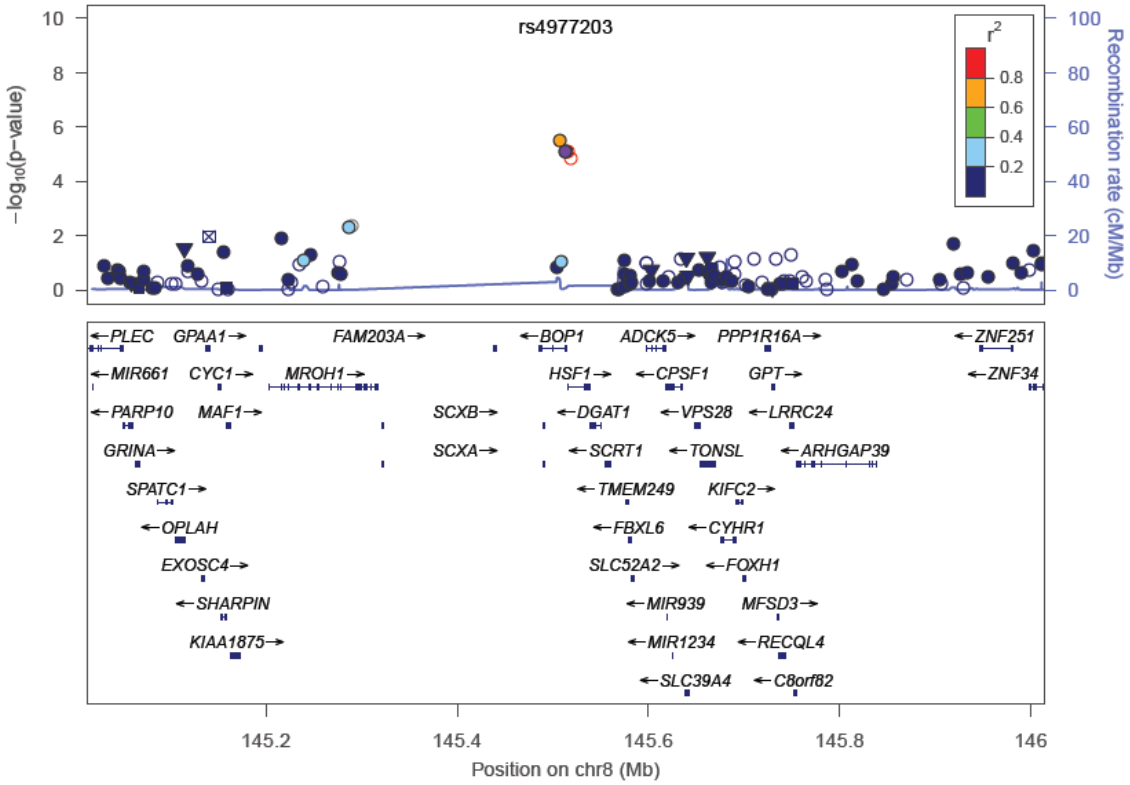
A2.1 Regional association plot of the index SNP - rs10196189. RPRM: reprimo; GALNT13: UDP-N-acetyl-alpha-D-galactosamine:polypeptide N-acetylgalactosaminyltransferase 13.



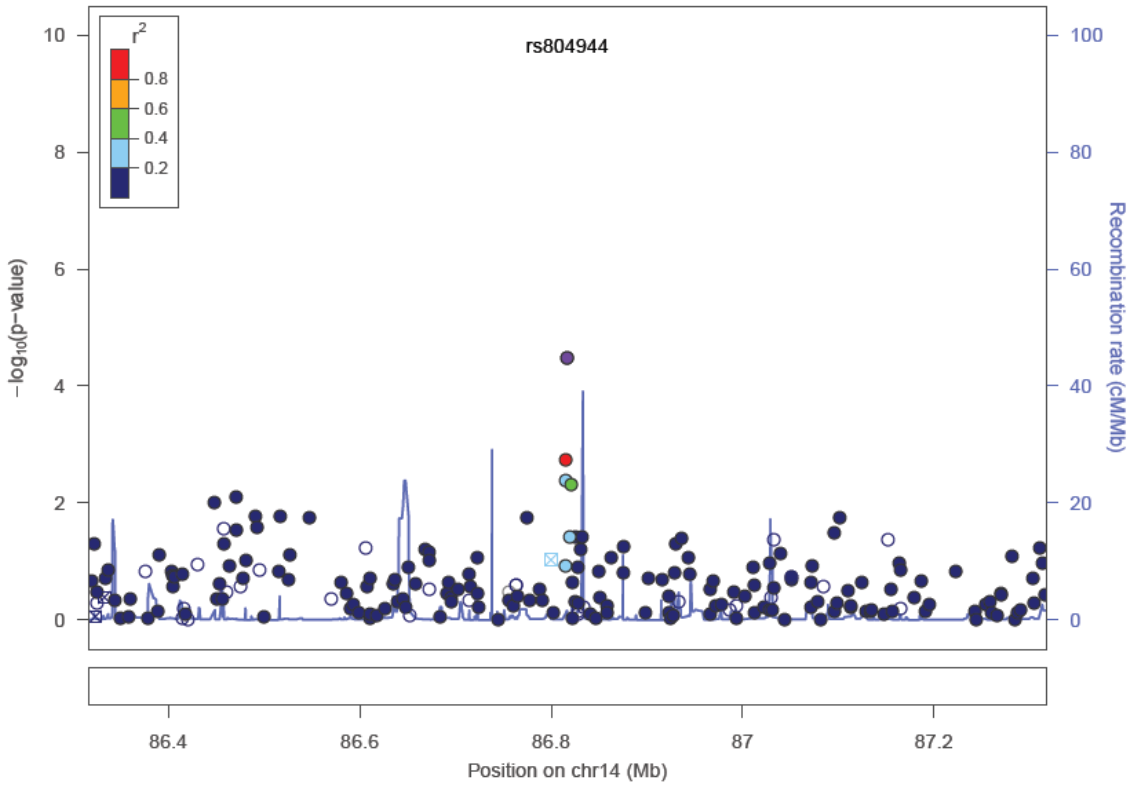
A2.2 Regional association plot of the index SNP - rs187167. SIL1: SIL1 homolog, endoplasmic reticulum chaperone (S. cerevisiae); PAIP2: poly(A) binding protein interacting protein 2; ECSCR: endothelial cell surface expressed chemotaxis and apoptosis regulator; UBE2D2: ubiquitin-conjugating enzyme E2D 2; PSD2: phosphatidylserine decarboxylase 2; PURA: purine-rich element binding protein A; SNHG4: small nucleolar RNA host gene 4 (non-protein coding); MZB1: marginal zone B and B1 cell-specific protein; TMEM173: transmembrane protein 173; CXXC5: CXXC finger protein 5; NRG2: neuregulin 2; IGIP: IgA-inducing protein homolog (Bos taurus); MATR3: matrin 3; DNAJC18: DnaJ (Hsp40) homolog, subfamily C, member 18; SNORA74A: small nucleolar RNA, H/ACA box 74A; PROB1: proline-rich basic protein 1; SLC23A1: solute carrier family 23 (nucleobase transporters), member 1; SPATA24: spermatogenesis associated 24.



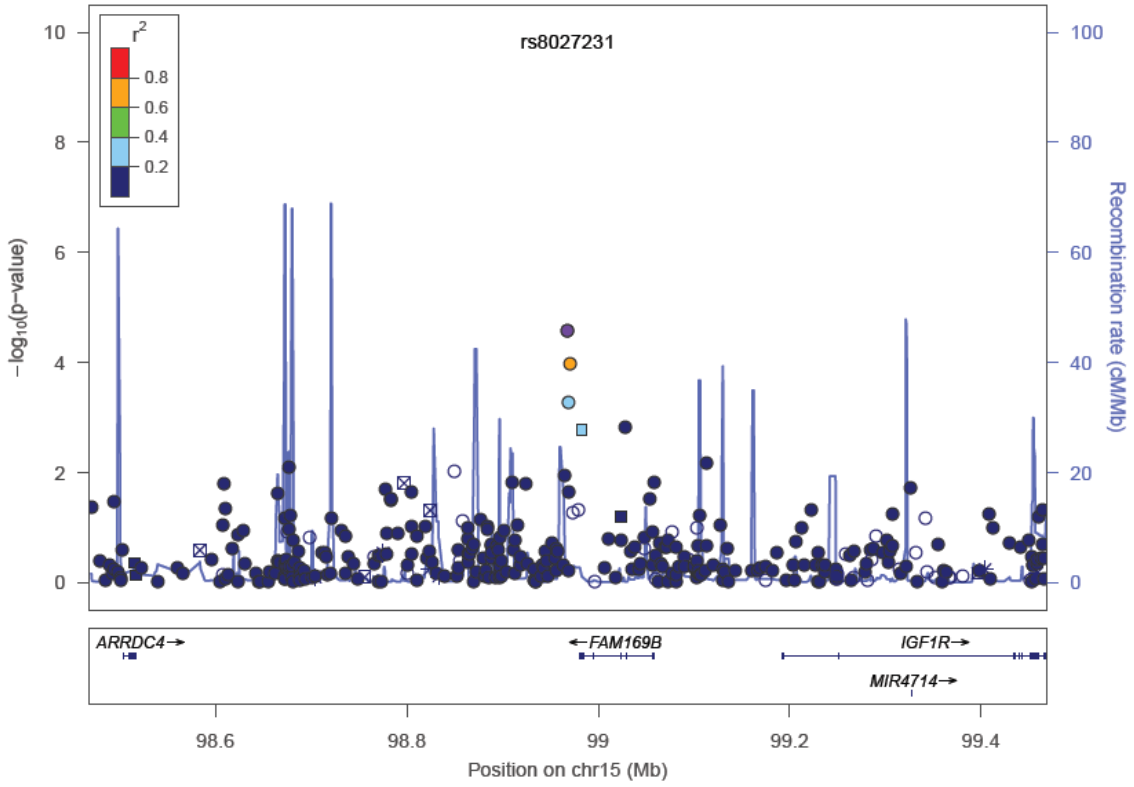
A2.3 Regional association plot of the index SNP – rs4557742. PLEC: plectin; GPAA1: glycosylphosphatidylinositol anchor attachment 1; FAM203A: family with sequence similarity 203, member A; BOP1: block of proliferation 1; ADCK5: aarF domain containing kinase 5; PPP1R16A: protein phosphatase 1, regulatory subunit 16A; ZNF251: zinc finger protein 251; MIR661: microRNA 661; CYC1: cytochrome c-1; MROH1: maestro heat-like repeat family member 1; HSF1: heat shock transcription factor 1; CPSF1: cleavage and polyadenylation specific factor 1, 160kDa; GPT: glutamic-pyruvate transaminase (alanine aminotransferase); ZNF34: zinc finger protein 34; PARP10: poly (ADP-ribose) polymerase family, member 10; MAF: v-maf musculoaponeurotic fibrosarcoma oncogene homolog (avian); SCXB: scleraxis homolog B (mouse); DGAT1: diacylglycerol O-acyltransferase 1; VPS28: vacuolar protein sorting 28 homolog (S. cerevisiae); LRRC24: leucine rich repeat containing 24; GRINA: glutamate receptor, ionotropic, N-methyl D-aspartate-associated protein 1 (glutamate binding); SCXA: scleraxis homolog A (mouse); SCRT1: scratch homolog 1, zinc finger protein (Drosophila); TONSL: tonsoku-like, DNA repair protein ; ARHGAP39: Rho GTPase activating protein 39; SPATC1: spermatogenesis and centriole associated 1; TMEM249: transmembrane protein 249; KIFC2: kinesin family member C2; OPLAH: 5-oxoprolinase (ATP-hydrolysing); FBXL6: F-box and leucine-rich repeat protein 6; CYHR1: cysteine/histidine-rich 1; EXOSC4: exosome component 4; SLC52A2: solute carrier family 52, riboflavin transporter, member 2; FOXH1: forkhead box H1; SHARPIN: SHANK-associated RH domain interactor; MIR939: microRNA 939; MFSD3: major facilitator superfamily domain containing 3; KIAA1875: KIAA1875; MIR1234: microRNA 1234; RECQL4: RecQ protein-like 4; SLC39A4: solute carrier family 39 (zinc transporter), member 4; C8orf82: chromosome 8 open reading frame 82.



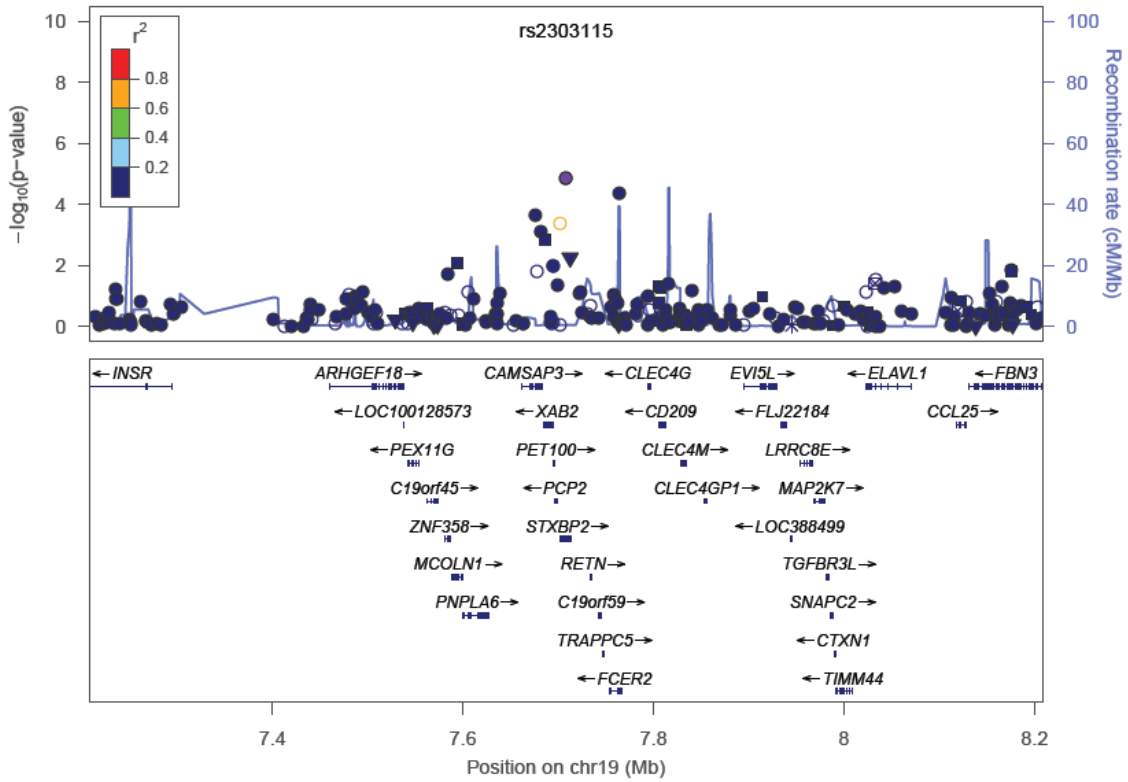
A2.4 Regional association plot of the index SNP – rs4977203. PLEC: plectin; GPAA1: glycosylphosphatidylinositol anchor attachment 1; FAM203A: family with sequence similarity 203, member A; BOP1: block of proliferation 1; ADCK5: aarF domain containing kinase 5; PPP1R16A: protein phosphatase 1, regulatory subunit 16A; ZNF251: zinc finger protein 251; MIR661: microRNA 661; CYC1: cytochrome c-1; MROH1: maestro heat-like repeat family member 1; HSF1: heat shock transcription factor 1; CPSF1: cleavage and polyadenylation specific factor 1, 160kDa; GPT: glutamic-pyruvate transaminase (alanine aminotransferase); ZNF34: zinc finger protein 34; PARP10: poly (ADP-ribose) polymerase family, member 10; MAF: v-maf musculoaponeurotic fibrosarcoma oncogene homolog (avian); SCXB: scleraxis homolog B (mouse); DGAT1: diacylglycerol O-acyltransferase 1; VPS28: vacuolar protein sorting 28 homolog (S. cerevisiae); LRRC24: leucine rich repeat containing 24; GRINA: glutamate receptor, ionotropic, N-methyl D-aspartate-associated protein 1 (glutamate binding); SCXA: scleraxis homolog A (mouse); SCRT1: scratch homolog 1, zinc finger protein (Drosophila); TONSL: tonsoku-like, DNA repair protein ; ARHGAP39: Rho GTPase activating protein 39; SPATC1: spermatogenesis and centriole associated 1; TMEM249: transmembrane protein 249; KIFC2: kinesin family member C2; OPLAH: 5-oxoprolinase (ATP-hydrolysing); FBXL6: F-box and leucine-rich repeat protein 6; CYHR1: cysteine/histidine-rich 1; EXOSC4: exosome component 4; SLC52A2: solute carrier family 52, riboflavin transporter, member 2; FOXH1: forkhead box H1; SHARPIN: SHANK-associated RH domain interactor; MIR939: microRNA 939; MFSD3: major facilitator superfamily domain containing 3; KIAA1875: KIAA1875; MIR1234: microRNA 1234; RECQL4: RecQ protein-like 4; SLC39A4: solute carrier family 39 (zinc transporter), member 4; C8orf82: chromosome 8 open reading frame 82.



A2.5 Regional association plot of the index SNP – rs804944.

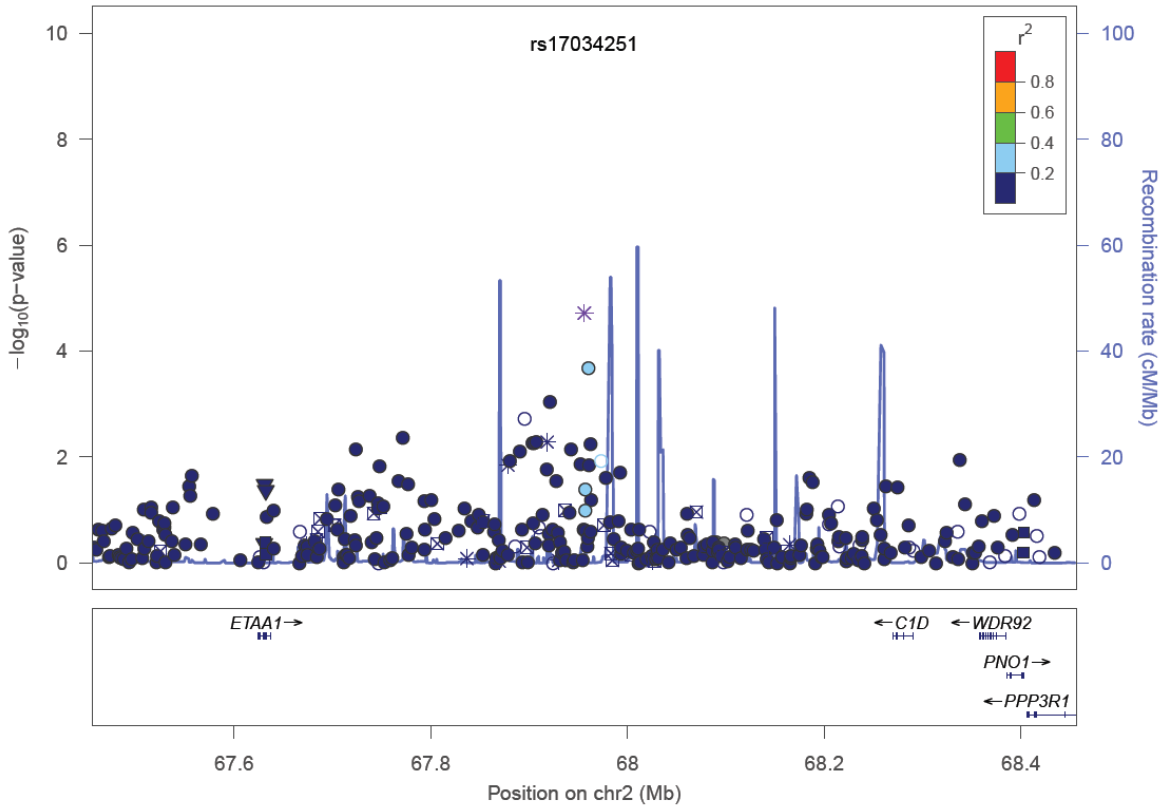


A2.6 Regional association plot of the index SNP – rs8027231. ARRDC4: arrestin domain containing 4; FAM169B: family with sequence similarity 169, member B; IGF1R: insulin-like growth factor 1 receptor ; MIR4714: microRNA 4714.

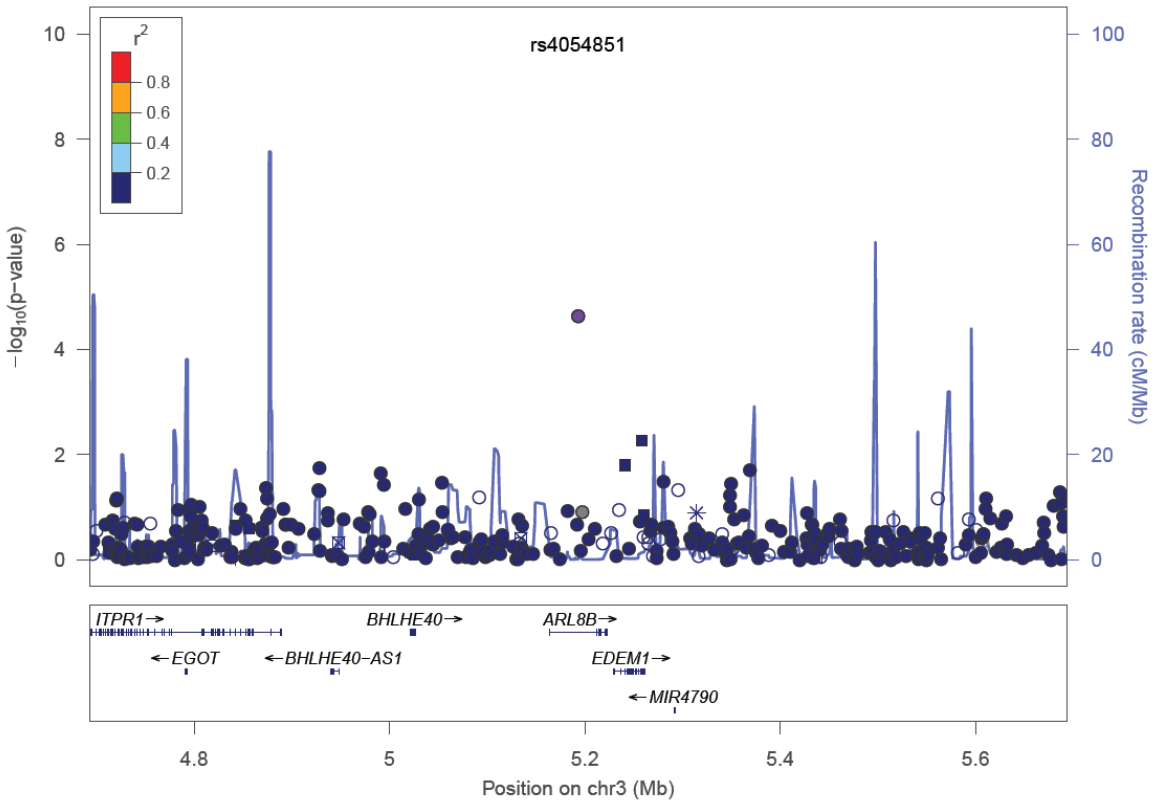


A2.7 Regional association plot of the index SNP – rs2303115. INSR: insulin receptor ; ARHGEF18: Rho/Rac guanine nucleotide exchange factor (GEF) 18; CAMSAP3: calmodulin regulated spectrin-associated protein family, member 3; CLEC4G: C-type lectin domain family 4, member G; EVI5L: ecotropic viral integration site 5-like; ELAVL1: ELAV (embryonic lethal, abnormal vision, Drosophila)-like 1 (Hu antigen R); FBN3: fibrillin 3; XAB2: XPA binding protein 2; CD209: CD209 molecule; FLJ22184: putative uncharacterized protein FLJ22184; CCL25: chemokine (C-C motif) ligand 25; PEX11G: peroxisomal biogenesis factor 11 gamma; PET100: PET100 homolog (S. cerevisiae); CLEC4M: C-type lectin domain family 4, member M ; LRRC8E: leucine rich repeat containing 8 family, member E ; C19orf45: chromosome 19 open reading frame 45; PCP2: Purkinje cell protein 2; CLEC4GP1: C-type lectin domain family 4, member G pseudogene 1; MAP2K7: mitogen-activated protein kinase kinase 7; ZNF358: zinc finger protein 358; STXBP2: syntaxin binding protein 2; MCOLN1: mucolipin 1; RETN: resistin; TGFBR3L: transforming growth factor, beta receptor III-like; PNPLA6: patatin-like phospholipase domain containing 6; C19orf59: chromosome 19 open reading frame 59; SNAPC2: small nuclear RNA activating complex, polypeptide 2, 45kDa; TRAPPC5: trafficking protein particle complex 5; CTXN1: cortexin 1; FCER2: Fc fragment of IgE, low affinity II, receptor for (CD23); TIMM44: translocase of inner mitochondrial membrane 44 homolog (yeast).

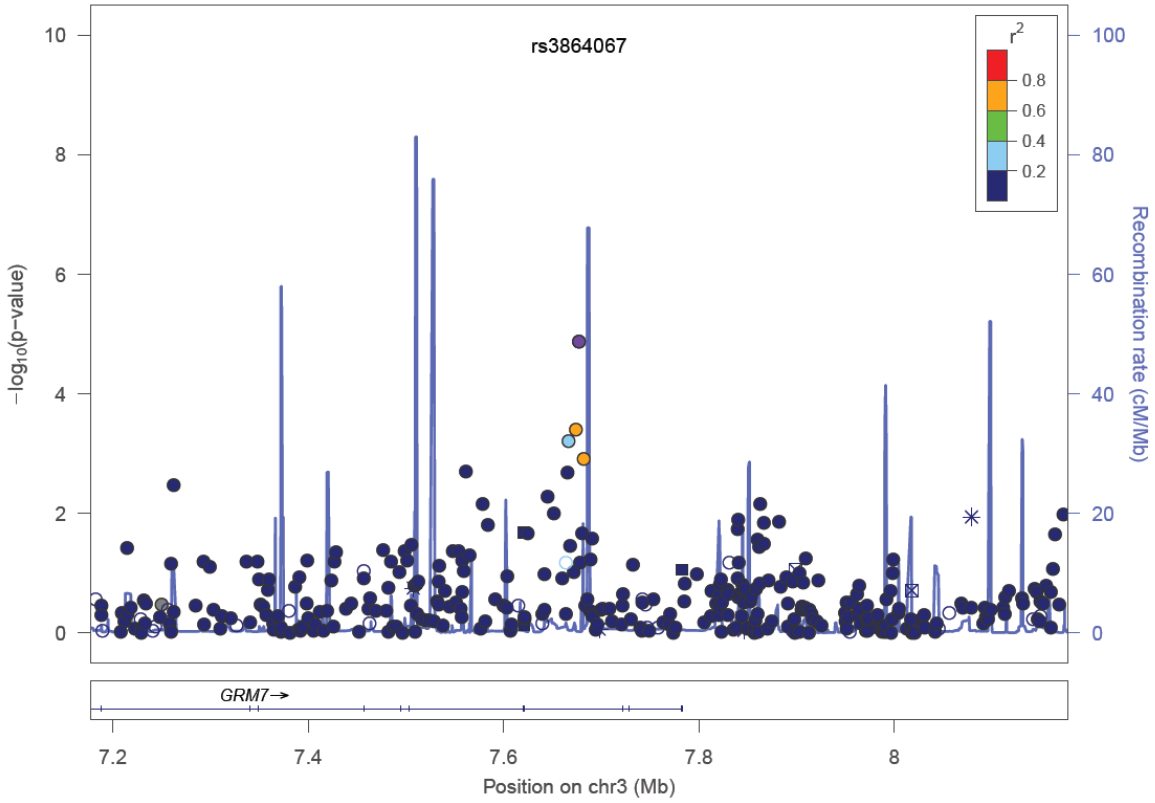
A3. Regional association plots of key markers (or index SNPs, in purple; see A3.1-A3.7) and 500Kb flanking region on each side of the markers for the African-American sprint cohort. –log10 transformed P values on the Y-axis indicate the strength of the association with elite sprint status in the African-American cohort. The level of LD between the index SNP and its surrounding SNPs as well as the recombination rate are estimated using 1000 Genomes AFR samples (Mar 2012). The level of LD is indicated by the colour key with red corresponding to high LD, and the recombination rate is represented by the blue line. Functional annotation key: triangle = framestop/splice, inverted triangle = non-synonymous, square = synonymous/UTR, star = conserved transcription factor binding site, square with diagonal lines = region is highly conserved in placental mammals, circle = no annotation.



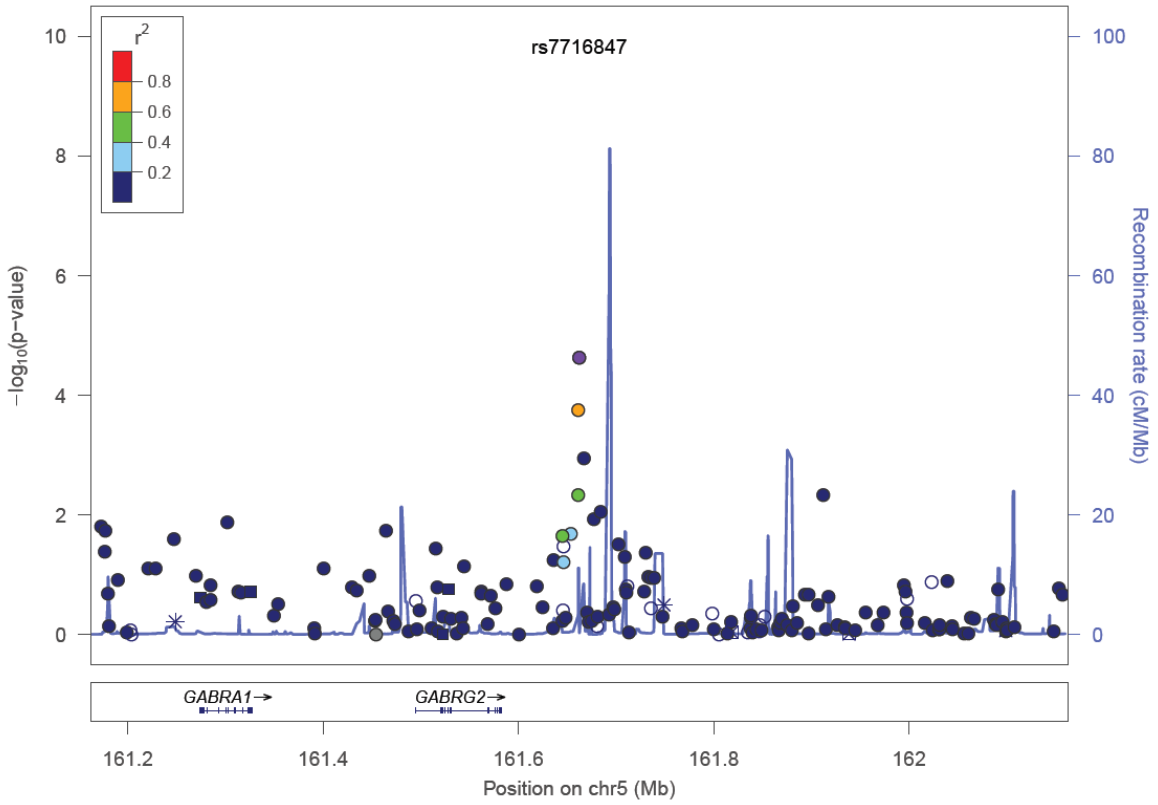
A3.1 Regional association plot of the index SNP – rs17034251. ETAA1: Ewing tumor-associated antigen 1; C1D: C1D nuclear receptor corepressor; WDR92: WD repeat domain 92; PNO1: partner of NOB1 homolog (S. cerevisiae); PPP3R1: protein phosphatase 3, regulatory subunit B, alpha.



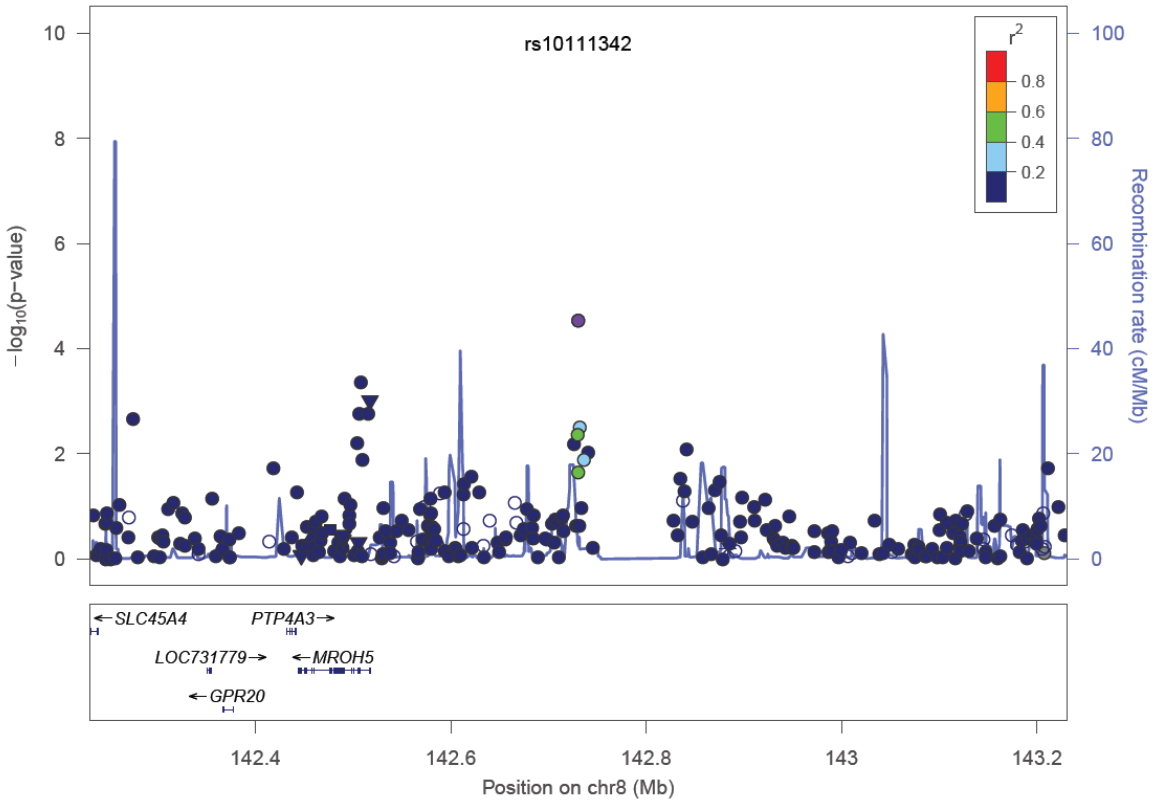
A3.2 Regional association plot of the index SNP – rs4054851. ITPR1: inositol 1,4,5-trisphosphate receptor, type 1; BHLHE40: basic helix-loop-helix family, member e40; ARL8B: ADP-ribosylation factor-like 8B; EGOT: eosinophil granule ontogeny transcript (non-protein coding); BHLHE40-AS1: BHLHE40 antisense RNA 1; EDEM1: ER degradation enhancer, mannosidase alpha-like 1; MIR4790: microRNA 4790.



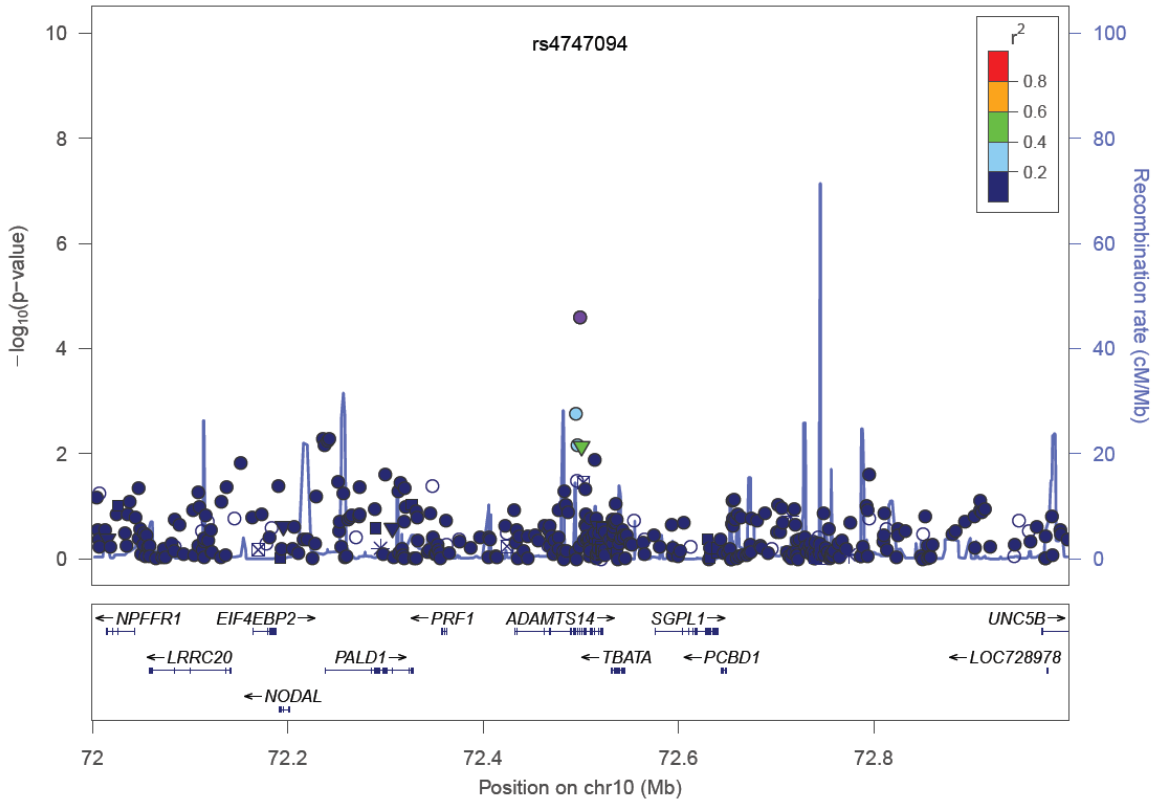
A3.3 Regional association plot of the index SNP – rs3864067. GRM7: glutamate receptor, metabotropic 7.



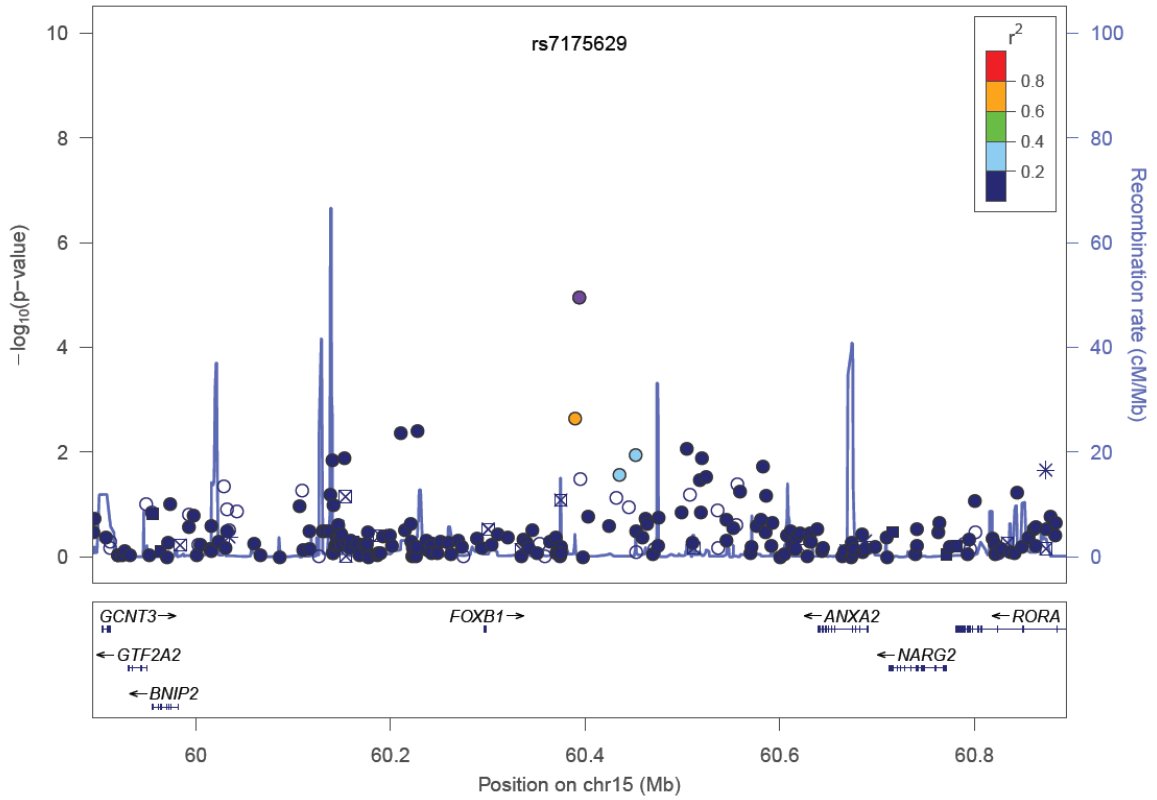
A3.4 Regional association plot of the index SNP – rs7716847. GABRA1: gamma-aminobutyric acid (GABA) A receptor, alpha 1; GABRG2: gamma-aminobutyric acid (GABA) A receptor, gamma 2.



A3.5 Regional association plot of the index SNP – rs10111342. SLC45A4: solute carrier family 45, member 4; PTP4A3: protein tyrosine phosphatase type IVA, member 3; MROH5: maestro heat-like repeat family member 5; GPR20: G protein-coupled receptor 20.

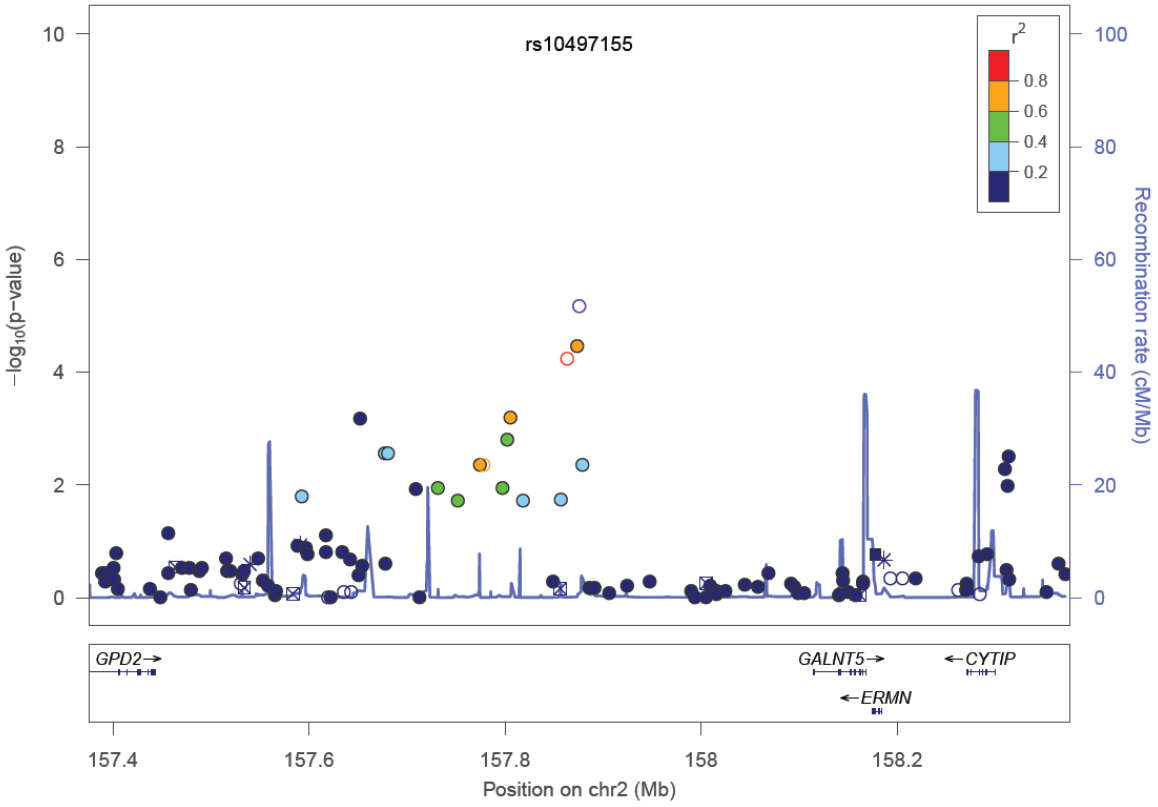


A3.6 Regional association plot of the index SNP – rs4747094. NPFFR1: neuropeptide FF receptor 1; EIF4EBP2: eukaryotic translation initiation factor 4E binding protein 2; PRF1: perforin 1 (pore forming protein); ADAMTS14: ADAM metallopeptidase with thrombospondin type 1 motif, 14; SGPL1: sphingosine-1-phosphate lyase 1; UNC5B: unc-5 homolog B (C. elegans); LRRC20: leucine rich repeat containing 20; PALD1: phosphatase domain containing, paladin 1; TBATA: thymus, brain and testes associated; PCBD1: pterin-4 alpha-carbinolamine dehydratase/dimerization cofactor of hepatocyte nuclear factor 1 alpha; NODAL: nodal growth differentiation factor.

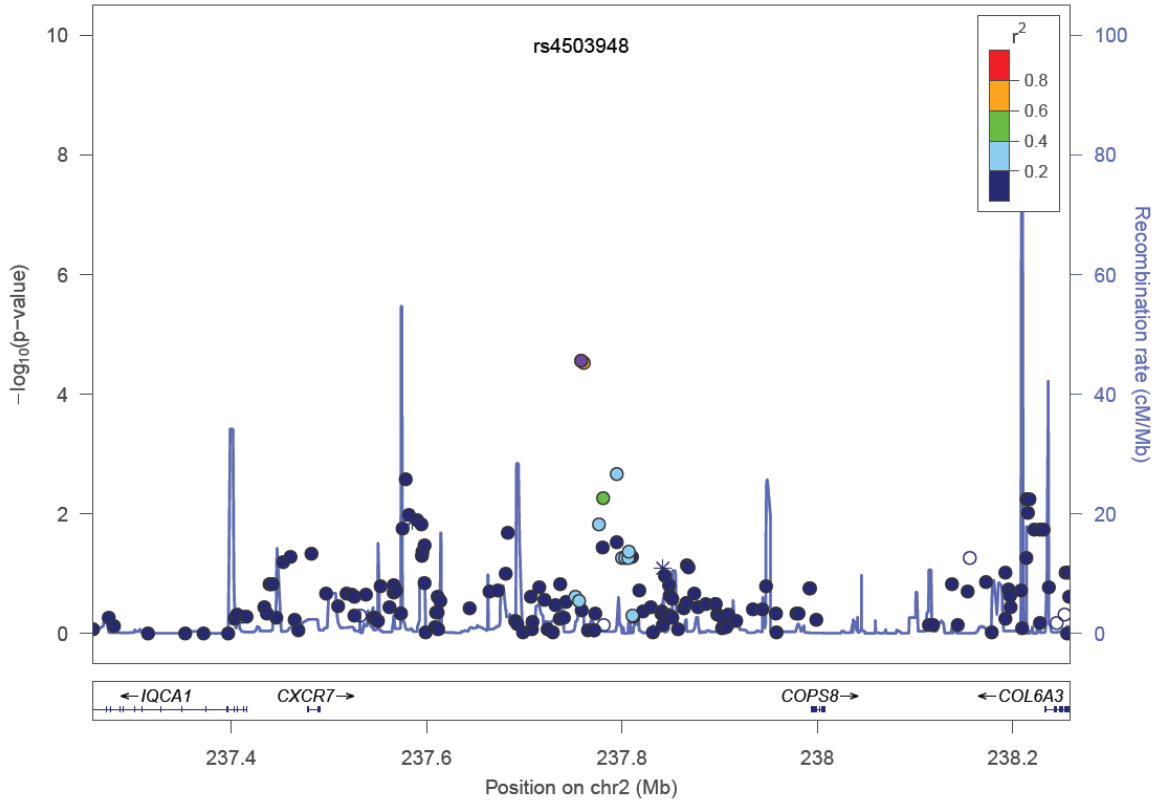


A3.7 Regional association plot of the index SNP – rs7175629. GCNT3: glucosaminyl (N-acetyl) transferase 3, mucin type; FOXB1: forkhead box B1; ANXA2: annexin A2; RORA: RAR-related orphan receptor A; GTF2A2: general transcription factor IIA, 2, 12kDa; NARG2: NMDA receptor regulated 2; BNIP2: BCL2/adenovirus E1B 19kDa interacting protein 2.

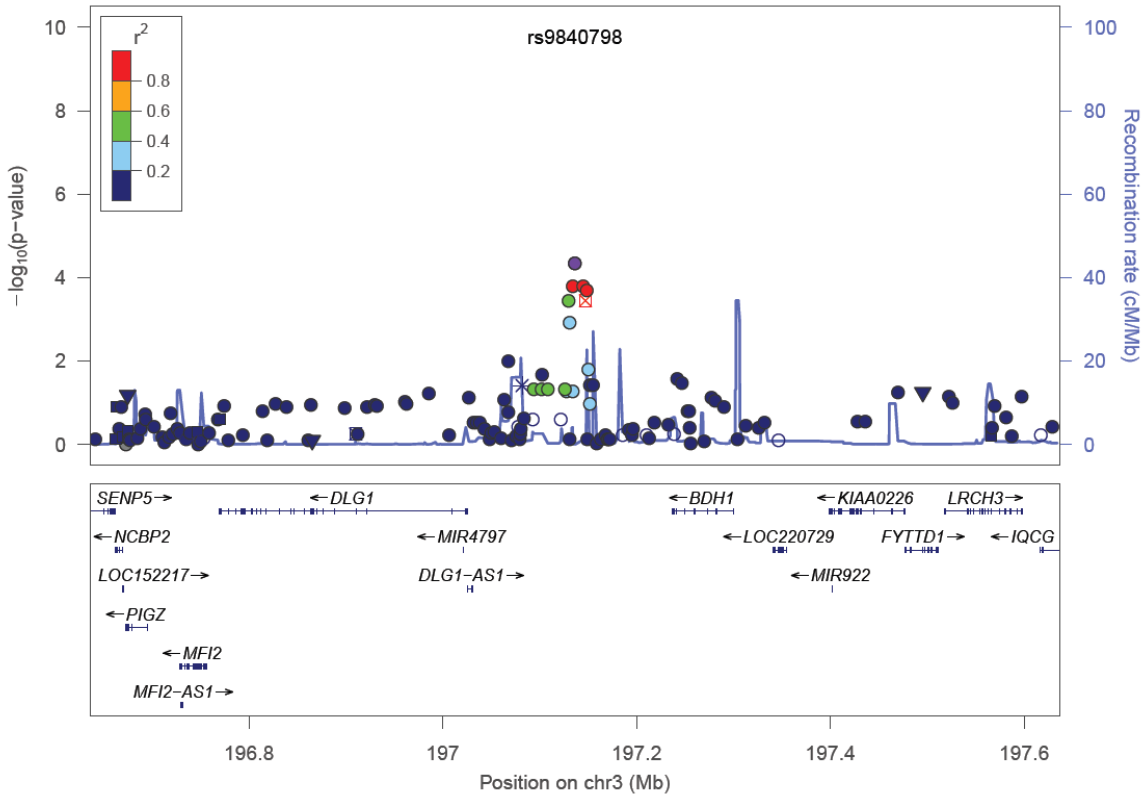
A4. Regional association plots of key markers (or index SNPs, in purple; see A4.1-A4.18) and 500Kb flanking region on each side of the markers for the Japanese sprint cohort. –log10 transformed P values on the Y-axis indicate the strength of the association with elite sprint status in the Japanese cohort. The level of LD between the index SNP and its surrounding SNPs as well as the recombination rate are estimated using 1000 Genomes ASN samples (Mar 2012). The level of LD is indicated by the colour key with red corresponding to high LD, and the recombination rate is represented by the blue line. Functional annotation key: triangle = framestop/splice, inverted triangle = non-synonymous, square = synonymous/UTR, star = conserved transcription factor binding site, square with diagonal lines = region is highly conserved in placental mammals, circle = no annotation.



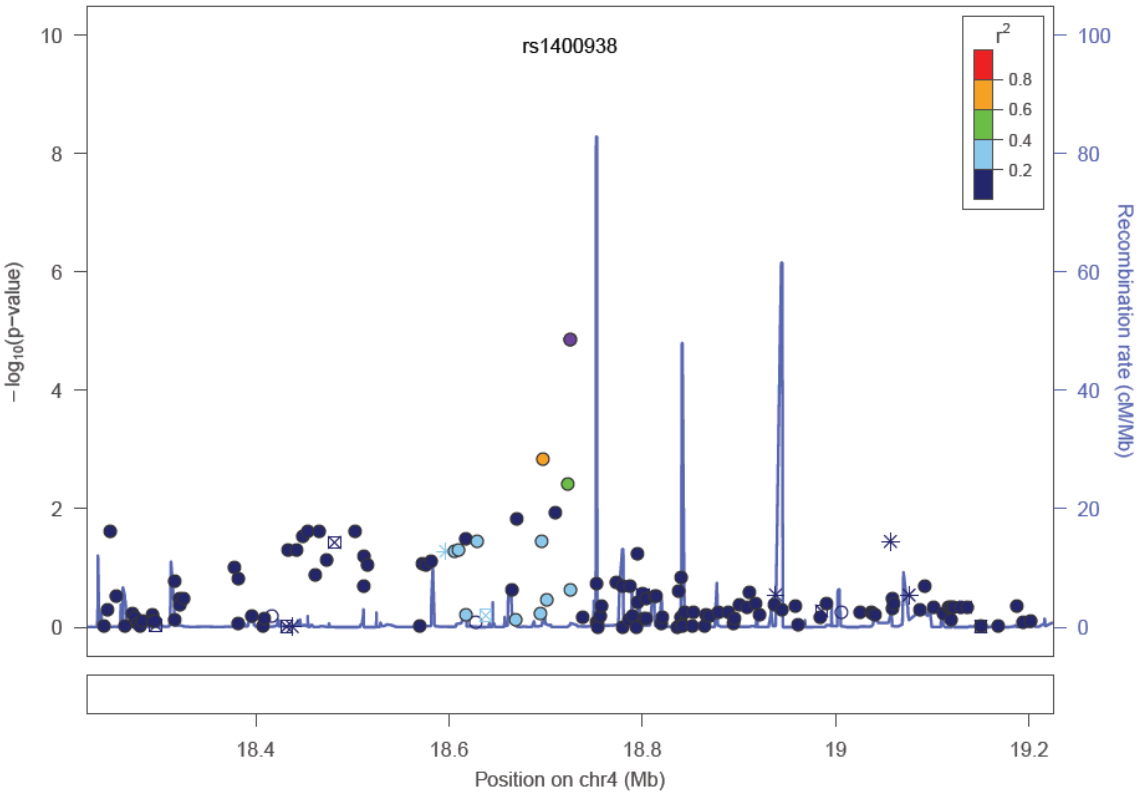
A4.1 Regional association plot of the index SNP - rs10497155. GPD2: glycerol-3-phosphate dehydrogenase 2 (mitochondrial); GALNT5: UDP-N-acetyl-alpha-D-galactosamine:polypeptide N-acetylgalactosaminyltransferase 5 (GalNAc-T5); CYTIP: cytohesin 1 interacting protein; ERMN: ermin, ERM-like protein**.**



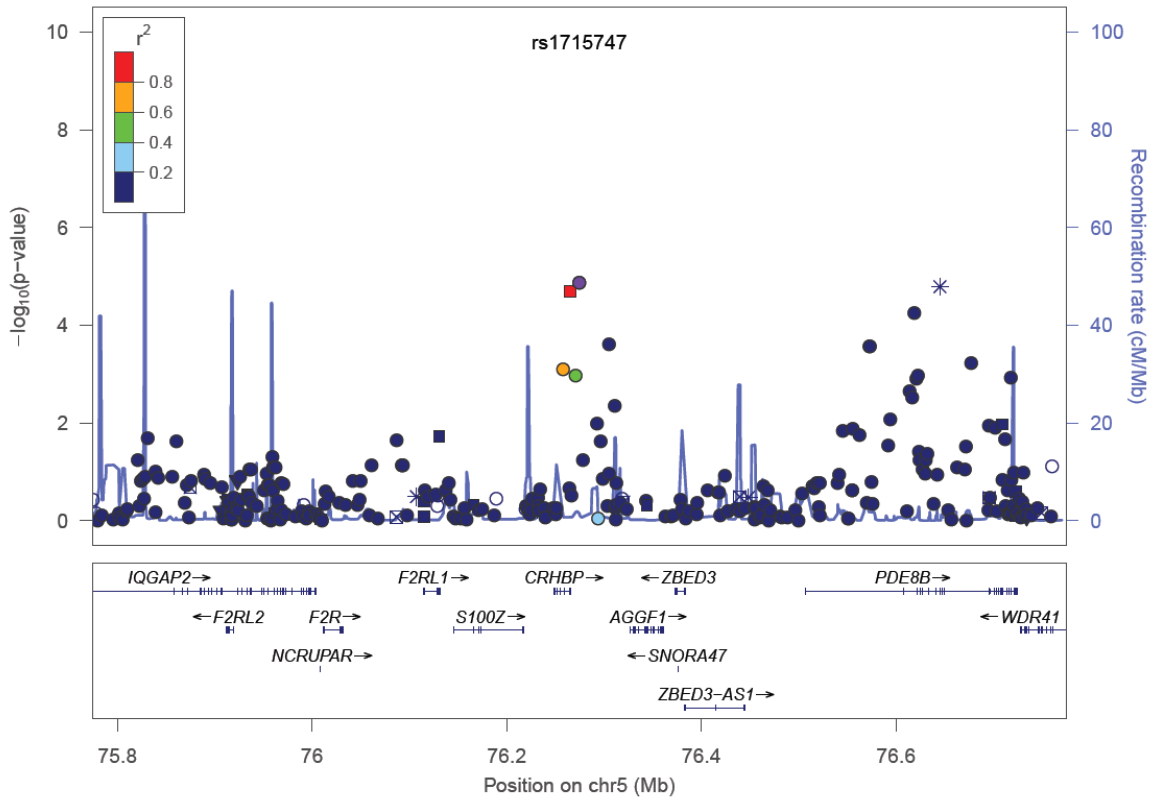
A4.2 Regional association plot of the index SNP – rs4503948. IQCA1: IQ motif containing with AAA domain 1; CXCR7: chemokine (C-X-C motif) receptor 7; COPS8: COP9 signalosome subunit 8; COL6A3: collagen, type VI, alpha 3.



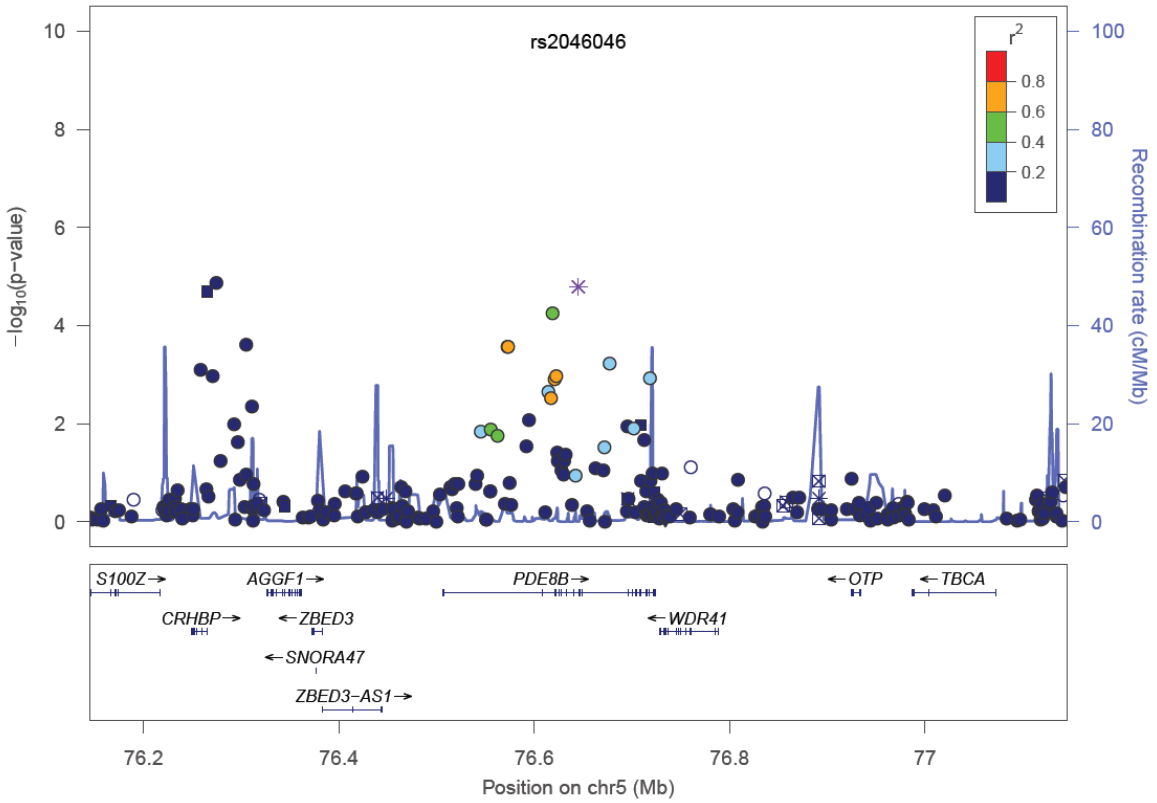
A4.3 Regional association plot of the index SNP – rs9840798. SENP5: SUMO1/sentrin specific peptidase 5; DLG1: discs, large homolog 1 (Drosophila); BDH1: 3-hydroxybutyrate dehydrogenase, type 1; KIAA0226: KIAA0226; LRCH3: leucine-rich repeats and calponin homology (CH) domain containing 3; NCBP2: nuclear cap binding protein subunit 2, 20kDa; MIR4797: microRNA 4797; FYTTD1: forty-two-three domain containing 1; IQCG: IQ motif containing G; DLG1-AS1: DLG1-AS1 DLG1 antisense RNA 1; MIR922: microRNA 922; PIGZ: phosphatidylinositol glycan anchor biosynthesis, class Z; MFI2: antigen p97 (melanoma associated) identified by monoclonal antibodies 133.2 and 96.5; MFI2-AS1: MFI2-AS1 MFI2 antisense RNA 1.



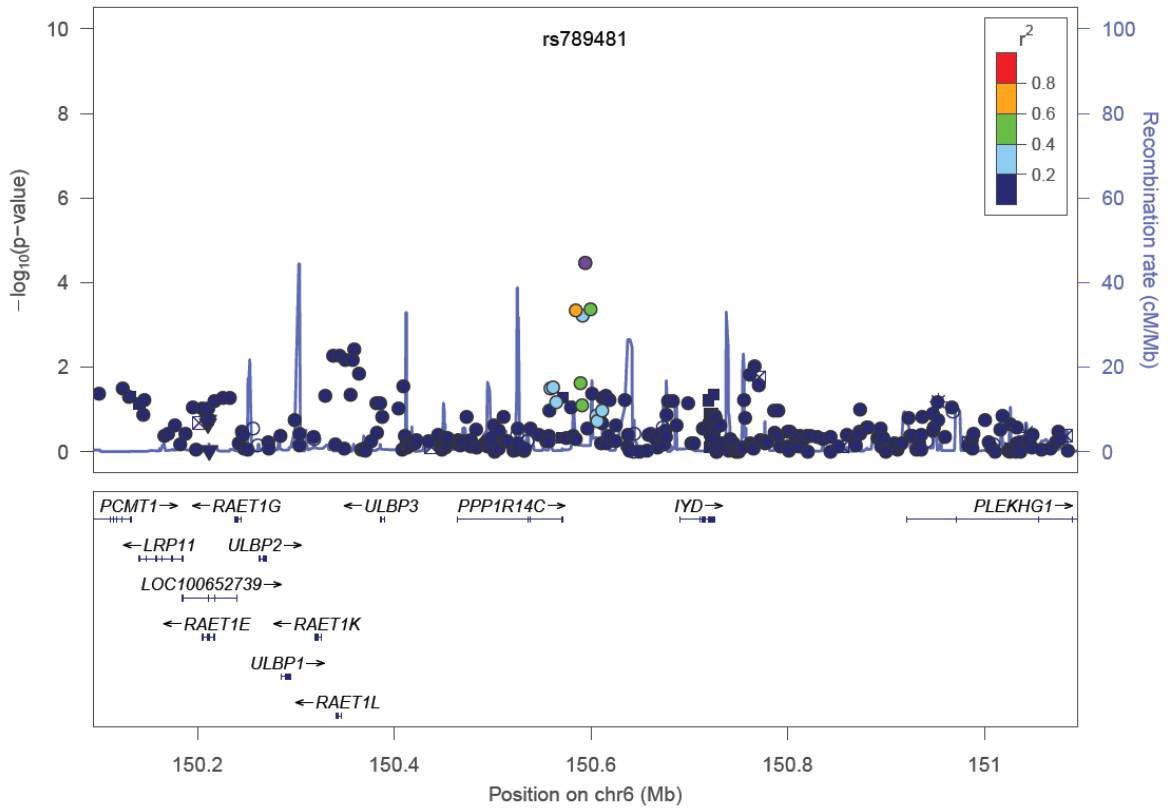
A4.4 Regional association plot of the index SNP - rs1400938.



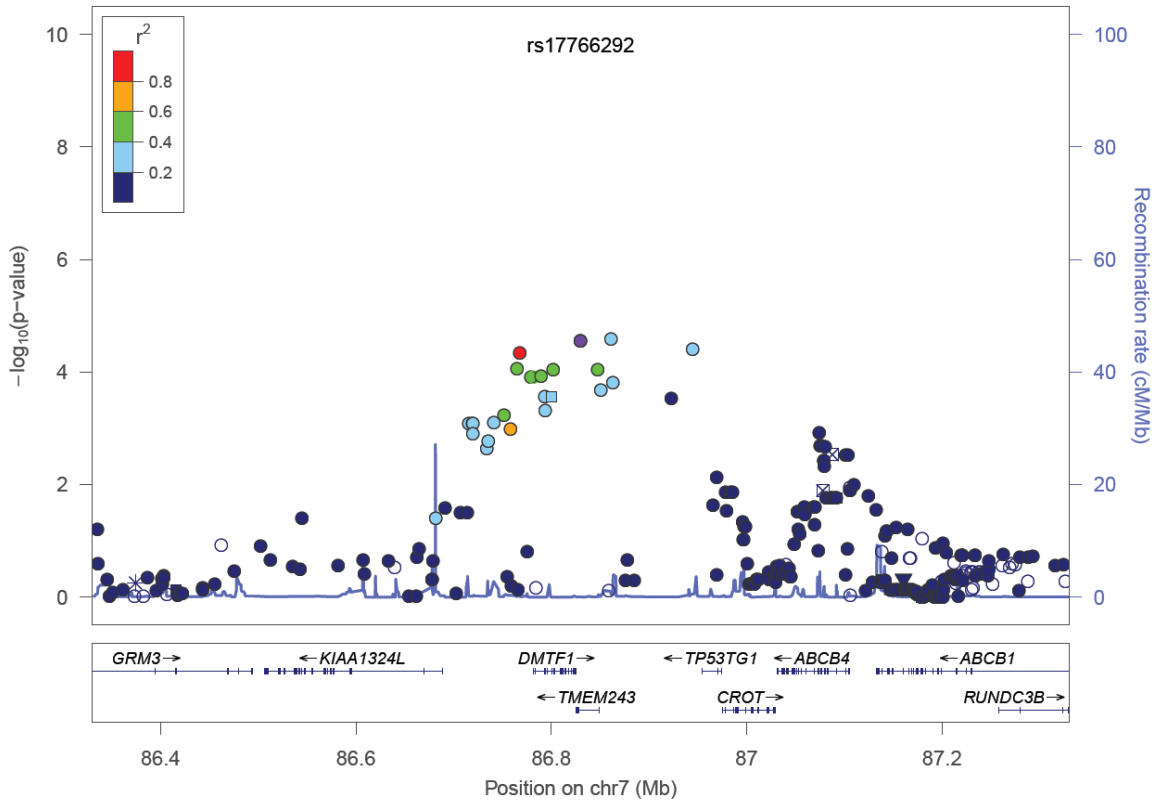
A4.5 Regional association plot of the index SNP - rs1715747. IQGAP2: IQ motif containing GTPase activating protein 2; F2RL1: coagulation factor II (thrombin) receptor-like 1; CRHBP: corticotropin releasing hormone binding protein; ZBED3: zinc finger, BED-type containing 3; PDE8B: phosphodiesterase 8B; F2RL2: coagulation factor II (thrombin) receptor-like 2; F2R: coagulation factor II (thrombin) receptor; S100Z: S100 calcium binding protein Z; AGGF1: angiogenic factor with G patch and FHA domains 1; WDR41: WD repeat domain 41; NCRUPAR: NCRUPAR non-protein coding RNA, upstream of F2R/PAR1; SNORA47: small nucleolar RNA, H/ACA box 47; ZBED3-AS1: ZBED3-AS1 ZBED3 antisense RNA 1.



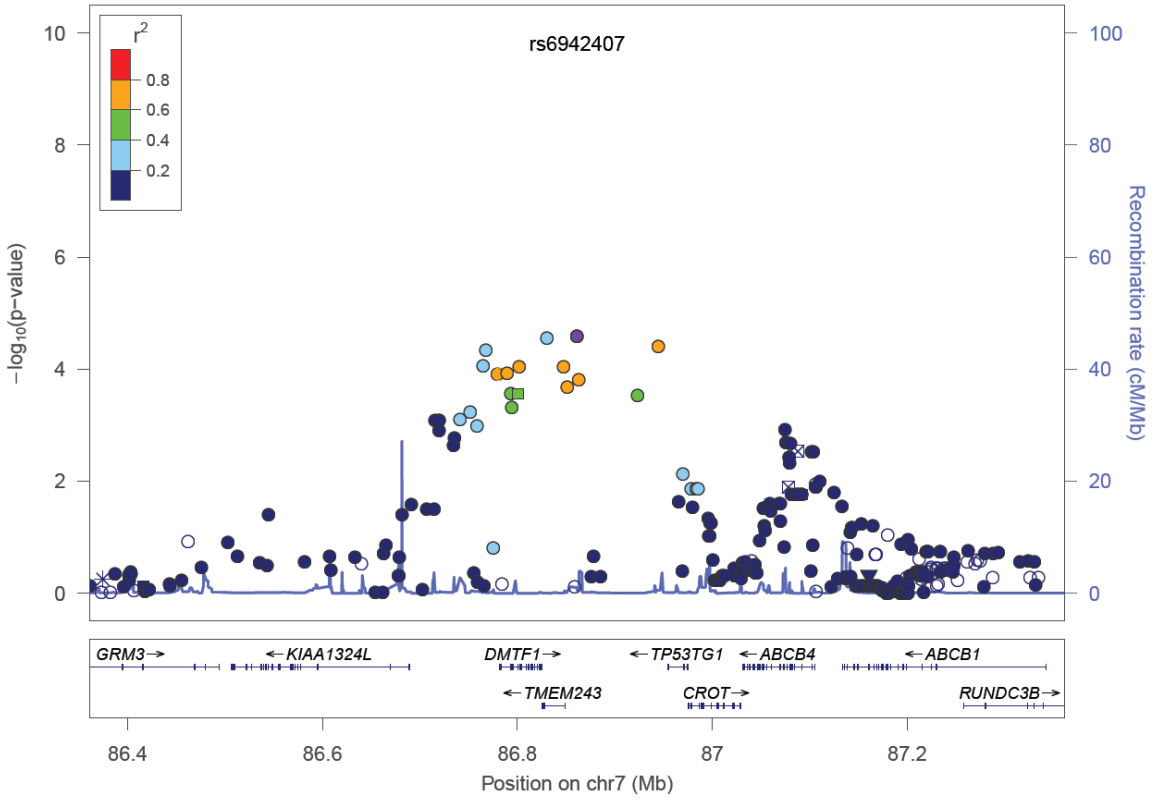
A4.6 Regional association plot of the index SNP – rs2046046. S100Z: S100 calcium binding protein Z ; AGGF1: angiogenic factor with G patch and FHA domains 1; PDE8B: phosphodiesterase 8B; OTP: orthopedia homeobox; TBCA: tubulin folding cofactor A; CRHBP: corticotropin releasing hormone binding protein; ZBED3: zinc finger, BED-type containing 3; WDR41: WD repeat domain 41; SNORA47: small nucleolar RNA, H/ACA box 47; ZBED3-AS1: ZBED3-AS1 ZBED3 antisense RNA 1.



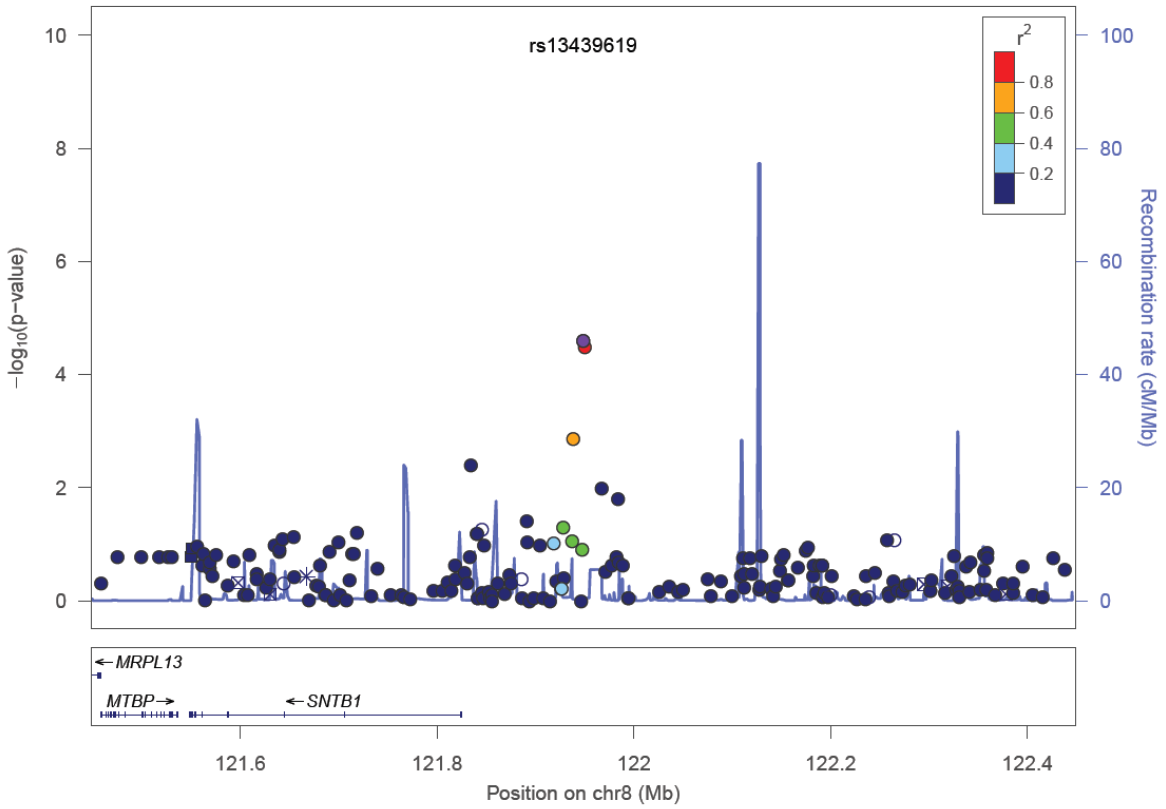
A4.7 Regional association plot of the index SNP – rs789481. PCMT1: protein-L-isoaspartate (D-aspartate) O-methyltransferase; RAET1G: retinoic acid early transcript 1G; ULBP3: UL16 binding protein 3; PPP1R14C: protein phosphatase 1, regulatory (inhibitor) subunit 14C; IYD: iodotyrosine deiodinase; PLEKHG1: pleckstrin homology domain containing, family G (with RhoGef domain) member 1; LRP11: low density lipoprotein receptor-related protein 11; ULBP2: UL16 binding protein 2; RAET1E: retinoic acid early transcript 1E; RAET1K: RAET1K retinoic acid early transcript 1K pseudogene; ULBP1: UL16 binding protein 1; RAET1L: retinoic acid early transcript 1L.



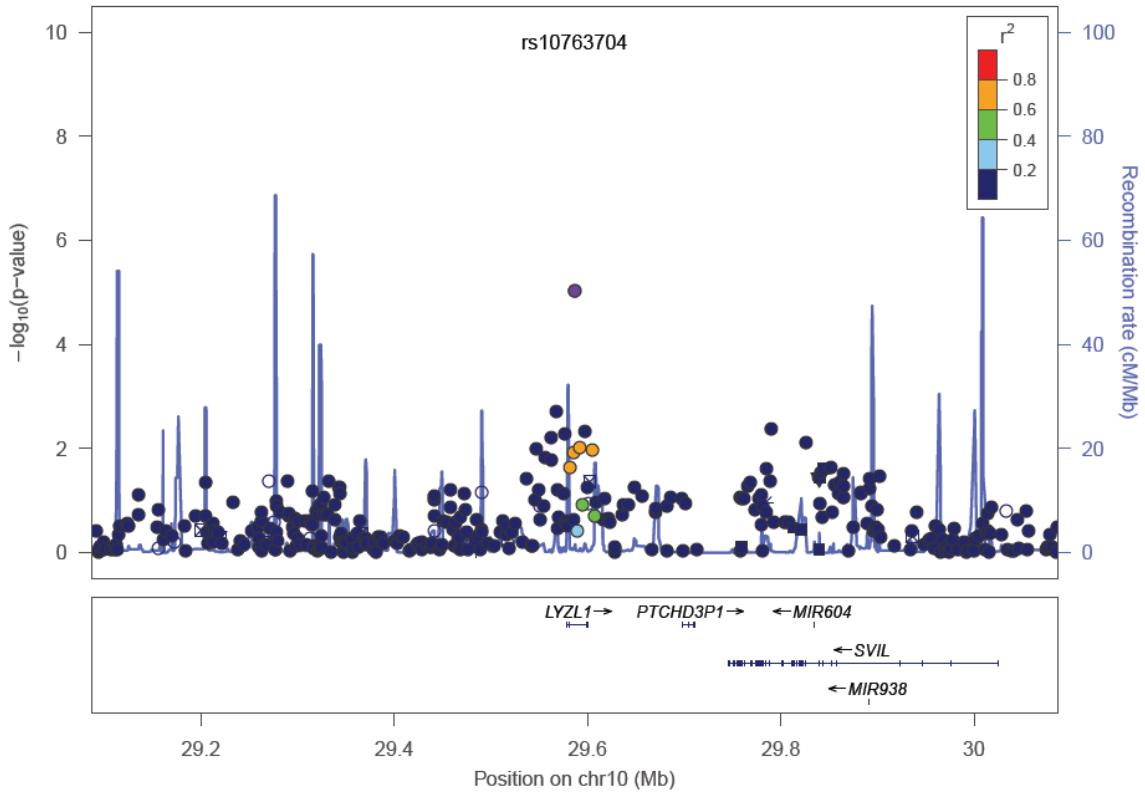
A4.8 Regional association plot of the index SNP - rs17766292. GRM3: glutamate receptor, metabotropic 3; KIAA1324L: KIAA1324-like; DMTF1: cyclin D binding myb-like transcription factor 1; TP53TG1: TP53 target 1 (non-protein coding); ABCB4: ATP-binding cassette, sub-family B (MDR/TAP), member 4; ABCB1: ATP-binding cassette, sub-family B (MDR/TAP), member 1; TMEM243: transmembrane protein 243, mitochondrial; CROT: carnitine O-octanoyltransferase; RUNDC3B: RUN domain containing 3B.



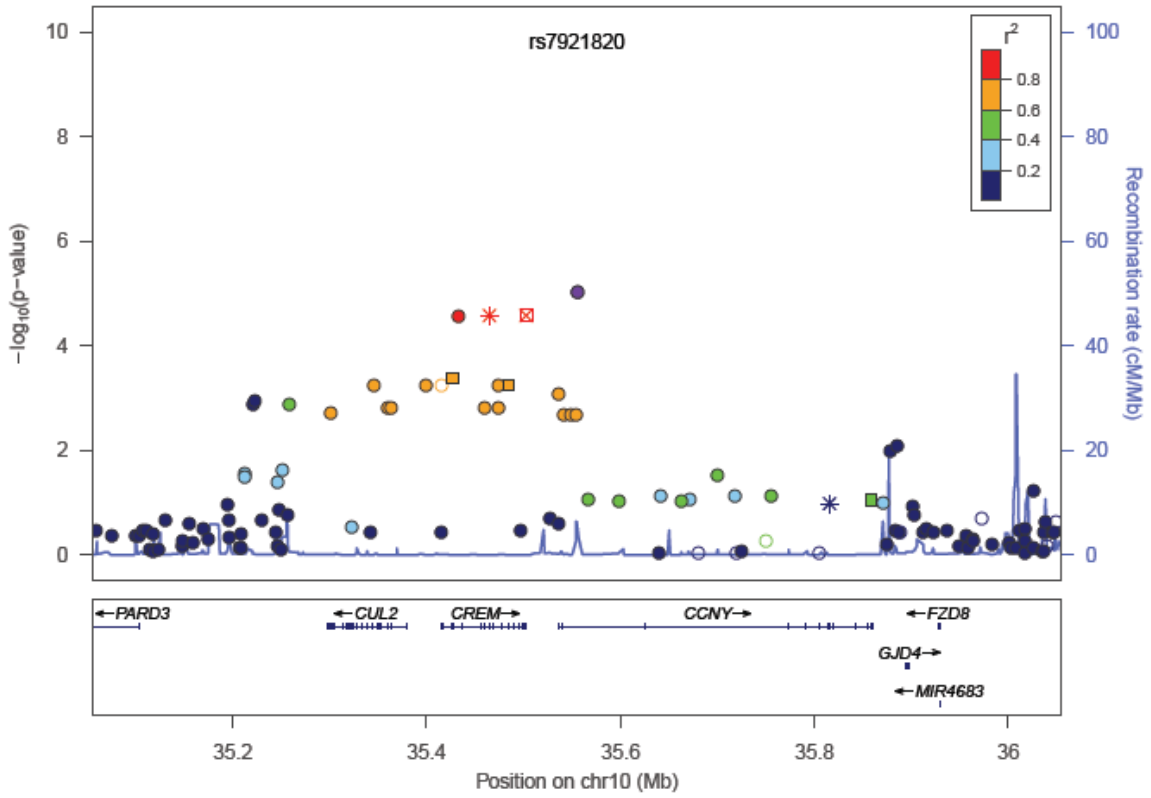
A4.9 Regional association plot of the index SNP – rs6942407. GRM3: glutamate receptor, metabotropic 3; KIAA1324L: KIAA1324-like; DMTF1: cyclin D binding myb-like transcription factor 1; TP53TG1: TP53 target 1 (non-protein coding); ABCB4: ATP-binding cassette, sub-family B (MDR/TAP), member 4; ABCB1: ATP-binding cassette, sub-family B (MDR/TAP), member 1; TMEM243: transmembrane protein 243, mitochondrial; CROT: carnitine O-octanoyltransferase; RUNDC3B: RUN domain containing 3B.



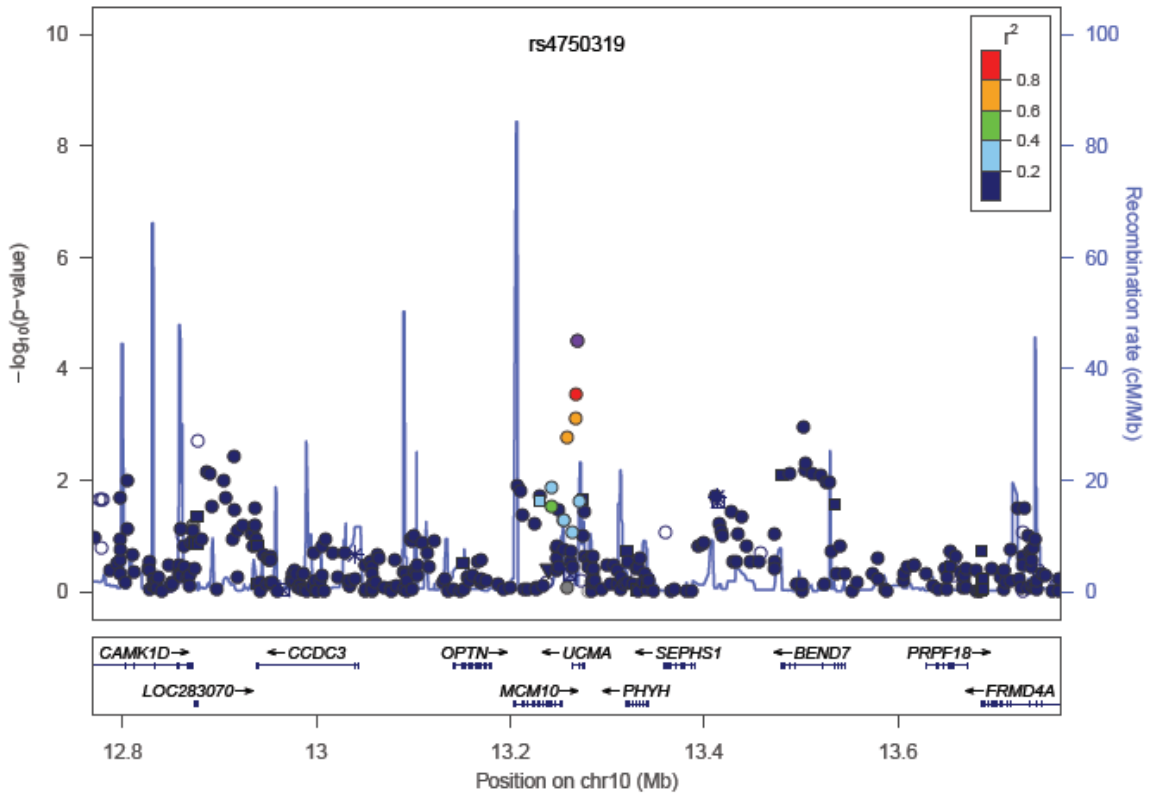
A4.10 Regional association plot of the index SNP – rs13439619. MRPL13: mitochondrial ribosomal protein L13; MTBP: Mdm2, transformed 3T3 cell double minute 2, p53 binding protein (mouse) binding protein, 104kDa; SNTB1: syntrophin, beta 1 (dystrophin-associated protein A1, 59kDa, basic component 1).



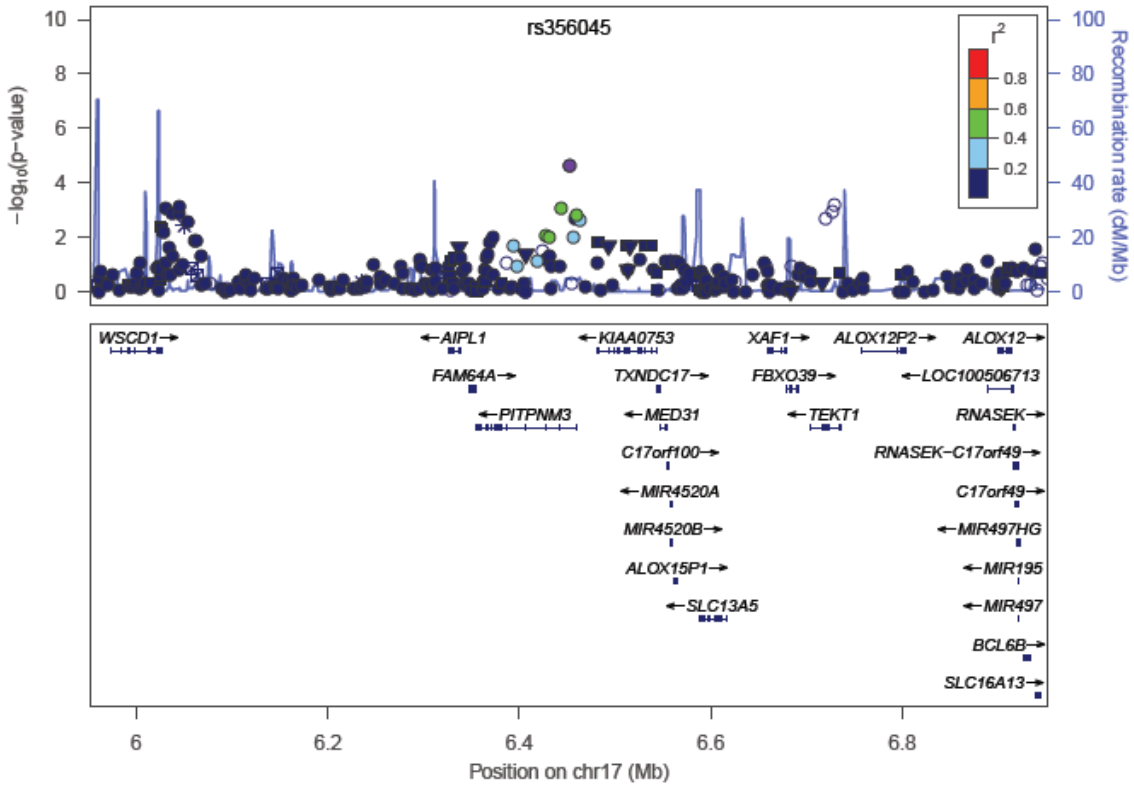
A4.11 Regional association plot of the index SNP - rs10763704. LYZL1: lysozyme-like 1; PTCHD3P1: patched domain containing 3 pseudogene 1; MIR604: microRNA 604; SVIL: supervillin; MIR938: microRNA 938.



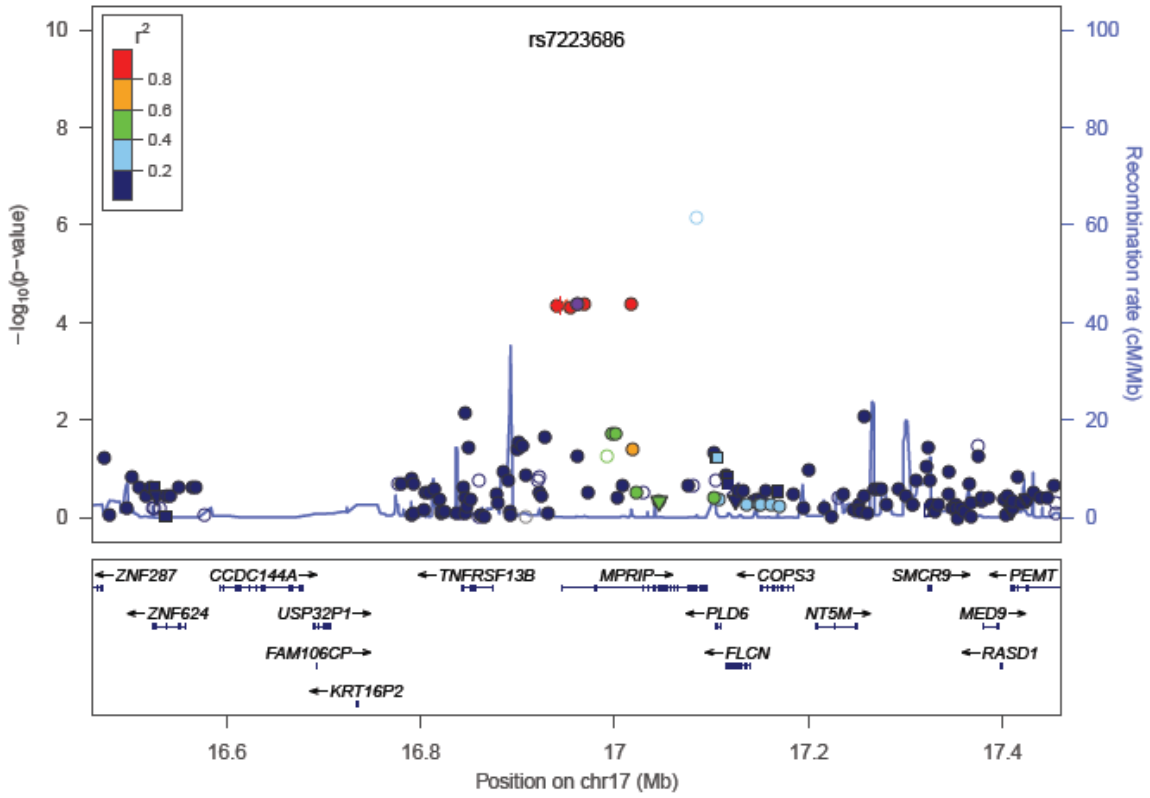
A4.12 Regional association plot of the index SNP – rs7921820. PARD3: par-3 partitioning defective 3 homolog (C. elegans); CUL2: cullin 2; CREM: cAMP responsive element modulator; CCNY: cyclin Y; FZD8: frizzled family receptor 8; GJD4: gap junction protein, delta 4, 40.1kDa; MIR4683: microRNA 4683.



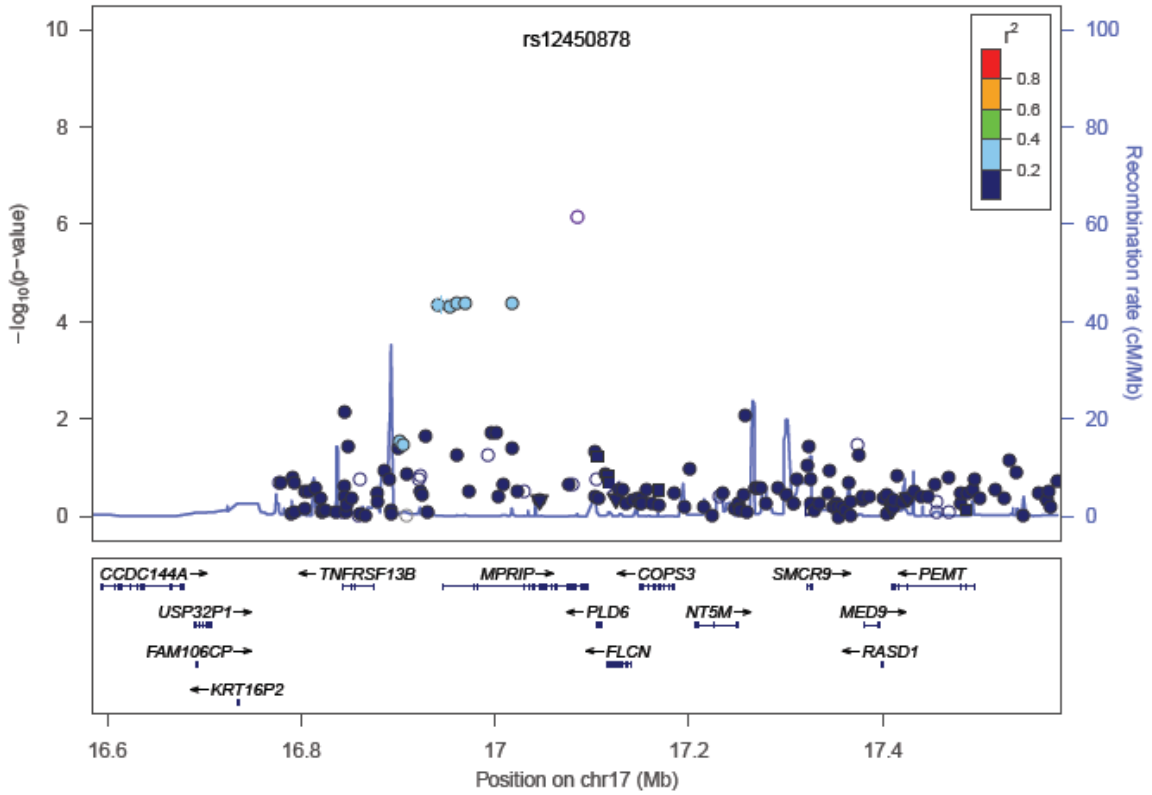
A4.13 Regional association plot of the index SNP – rs4750319. CAMK1D: calcium/calmodulin-dependent protein kinase ID; CCDC3: coiled-coil domain containing 3; OPTN: optineurin; UCMA: upper zone of growth plate and cartilage matrix associated; SEPHS1: selenophosphate synthetase 1; BEND7: BEN domain containing 7; PRPF18: PRP18 pre-mRNA processing factor 18 homolog (S. cerevisiae); MCM10: minichromosome maintenance complex component 10; PHYH: phytanoyl-CoA 2-hydroxylase; FRMD4A: FERM domain containing 4A.



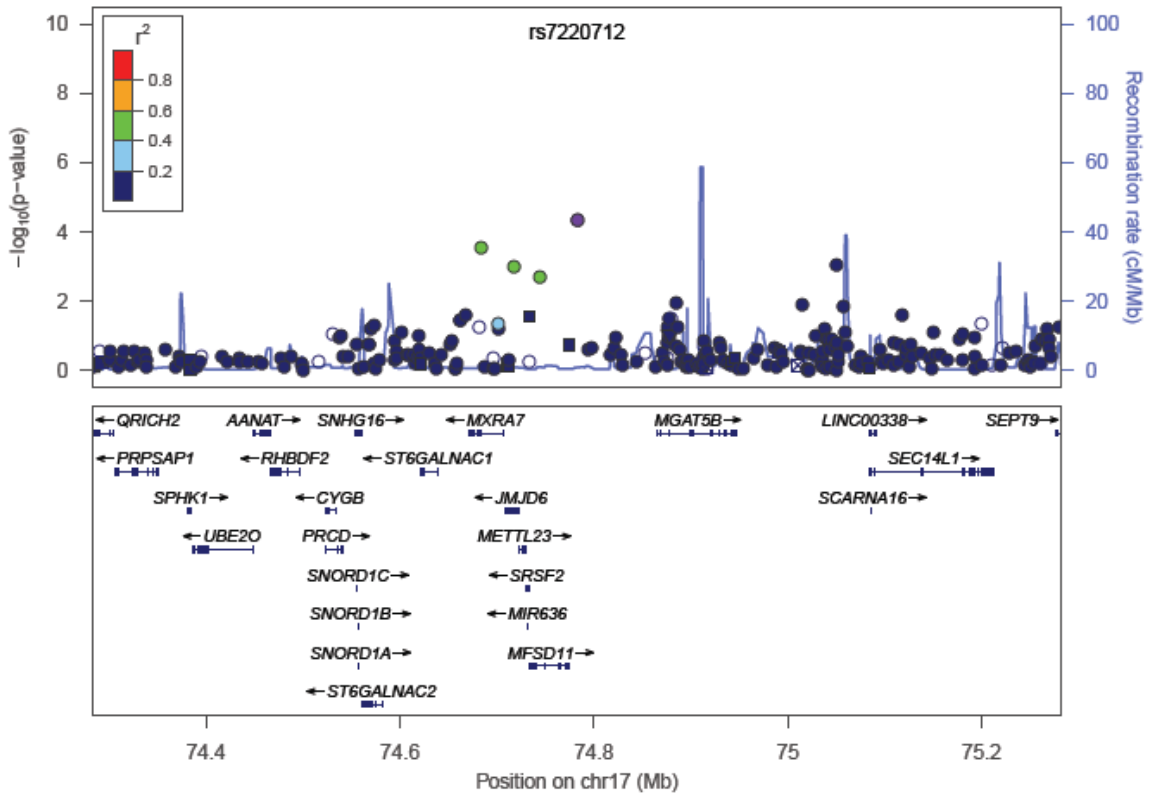
A4.14 Regional association plot of the index SNP – rs356045. WSCD1: WSC domain containing 1; AIPL1: aryl hydrocarbon receptor interacting protein-like 1; KIAA0753: KIAA0753; XAF1: XIAP associated factor 1; ALOX12P2: arachidonate 12-lipoxygenase pseudogene 2; ALOX12: arachidonate 12-lipoxygenase; FAM64A: family with sequence similarity 64, member A; TXNDC17: thioredoxin domain containing 17; FBXO39: F-box protein 39; PITPNM3: PITPNM family member 3; MED31: mediator complex subunit 31; TEKT1: tektin 1; RNASEK: ribonuclease, RNase K; C17orf100: chromosome 17 open reading frame 100; RNASEK-C17orf49: RNASEK-C17orf49 readthrough (non-protein coding); MIR4520A: microRNA 4520a; C17orf49: chromosome 17 open reading frame 49; MIR4520B: microRNA 4520b; MIR497HG: mir-497-195 cluster host gene (non-protein coding); ALOX15p1: arachidonate 15-lipoxygenase pseudogene 1; MIR195: microRNA 195; SLC13A5: solute carrier family 13 (sodium-dependent citrate transporter), member 5; MIR497: microRNA 497; BCL6B: B-cell CLL/lymphoma 6, member B; SLC16A13: solute carrier family 16, member 13 (monocarboxylic acid transporter 13).



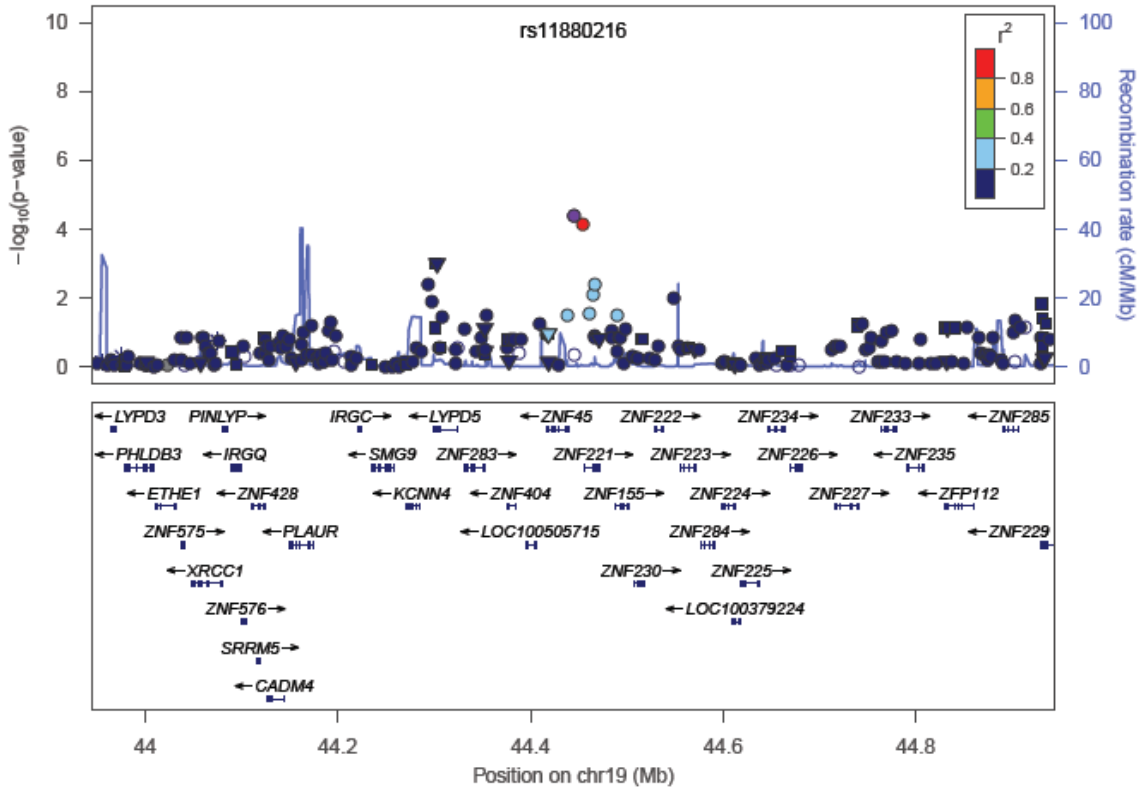
A4.15 Regional association plot of the index SNP – rs7223686. ZNF287: zinc finger protein 287; CCDC144A: coiled-coil domain containing 144A; TNFRSF13B: tumor necrosis factor receptor superfamily, member 13B; MPRIP: myosin phosphatase Rho interacting protein; COPS3: COP9 signalosome subunit 3; SMCR9: Smith-Magenis syndrome chromosome region, candidate 9; PEMT: phosphatidylethanolamine N-methyltransferase; ZNF624: zinc finger protein 624; USP32P1: ubiquitin specific peptidase 32 pseudogene 1; PLD6: phospholipase D family, member 6; NT5M: 5',3'-nucleotidase, mitochondrial; MED9: mediator complex subunit 9; FAM106CP: family with sequence similarity 106, member C, pseudogene; FLCN: folliculin; RASD1: RAS, dexamethasone-induced 1; KRT16P2: keratin 16 pseudogene 2.



A4.16 Regional association plot of the index SNP - rs12450878. CCDC144A: coiled-coil domain containing 144A; TNFRSF13B: tumor necrosis factor receptor superfamily, member 13B; MPRIP: myosin phosphatase Rho interacting protein; COPS3: COP9 signalosome subunit 3; SMCR9: Smith-Magenis syndrome chromosome region, candidate 9; PEMT: phosphatidylethanolamine N-methyltransferase; USP32P1: ubiquitin specific peptidase 32 pseudogene 1; PLD6: phospholipase D family, member 6; NT5M: 5',3'-nucleotidase, mitochondrial; MED9: mediator complex subunit 9; FAM106CP: family with sequence similarity 106, member C, pseudogene; FLCN: folliculin; RASD1: RAS, dexamethasone-induced 1; KRT16P2: keratin 16 pseudogene 2.

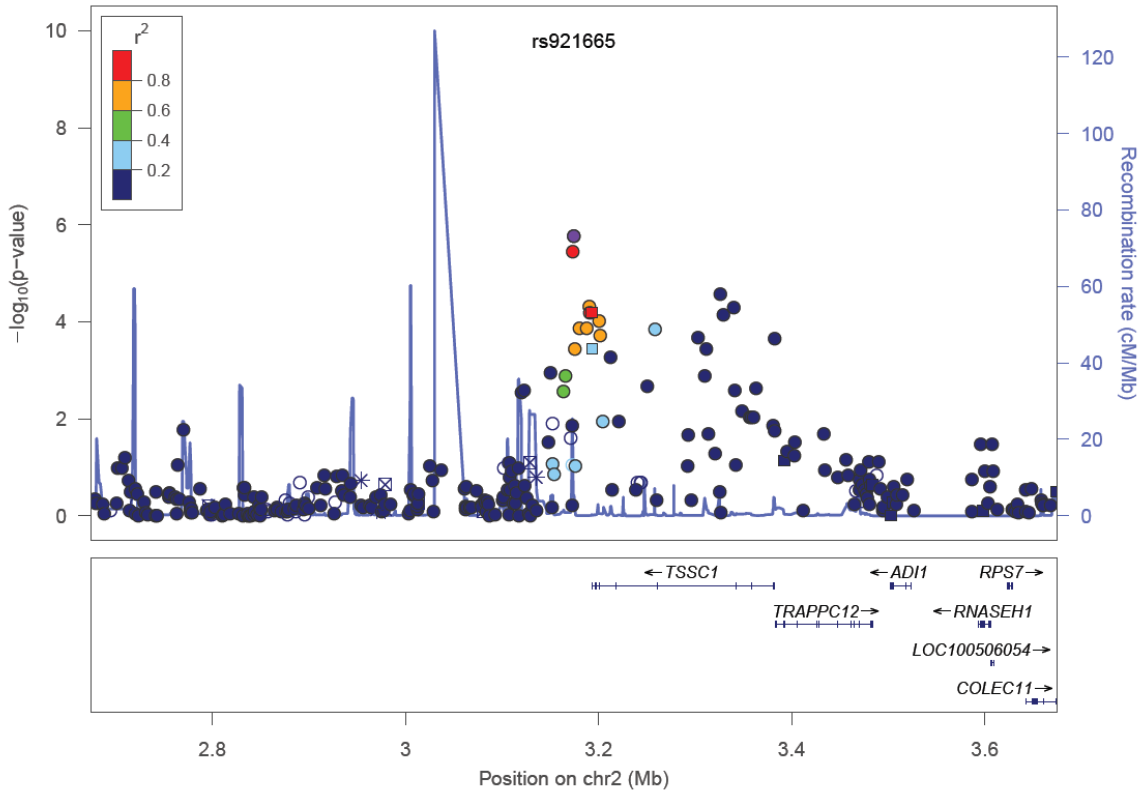


A4.17 Regional association plot of the index SNP – rs7220712. QRICH2: glutamine rich 2; AANAT: aralkylamine N-acetyltransferase; SNHG16: small nucleolar RNA host gene 16 (non-protein coding); MXRA7: matrix-remodelling associated 7; MGAT5B: mannosyl (alpha-1,6-)-glycoprotein beta-1,6-N-acetyl-glucosaminyltransferase, isozyme B; LINC00338: long intergenic non-protein coding RNA 338; SEPT9: septin 9; PRPSAP1: phosphoribosyl pyrophosphate synthetase-associated protein 1; RHBDF2: rhomboid 5 homolog 2 (Drosophila); ST6GALNAC1: ST6 (alpha-N-acetyl-neuraminyl-2,3-beta-galactosyl-1,3)-N-acetylgalactosaminide alpha-2,6-sialyltransferase 1; SEC14L1: SEC14-like 1 (S. cerevisiae); SPHK1: sphingosine kinase 1; CYGB: cytoglobin; JMJD6: jumonji domain containing 6; SCARNA16: small Cajal body-specific RNA 16; UBE2O: ubiquitin-conjugating enzyme E2O; PRCD: progressive rod-cone degeneration; METTL23: methyltransferase like 23; SNORD1C: small nucleolar RNA, C/D box 1C; SRSF2: serine/arginine-rich splicing factor 2; SNORD1B: small nucleolar RNA, C/D box 1B; MIR636: microRNA 636; SNORD1A: small nucleolar RNA, C/D box 1A; MFSD11: major facilitator superfamily domain containing 11; ST6GALNAC2: ST6 (alpha-N-acetyl-neuraminyl-2,3-beta-galactosyl-1,3)-N-acetylgalactosaminide alpha-2,6-sialyltransferase 2.

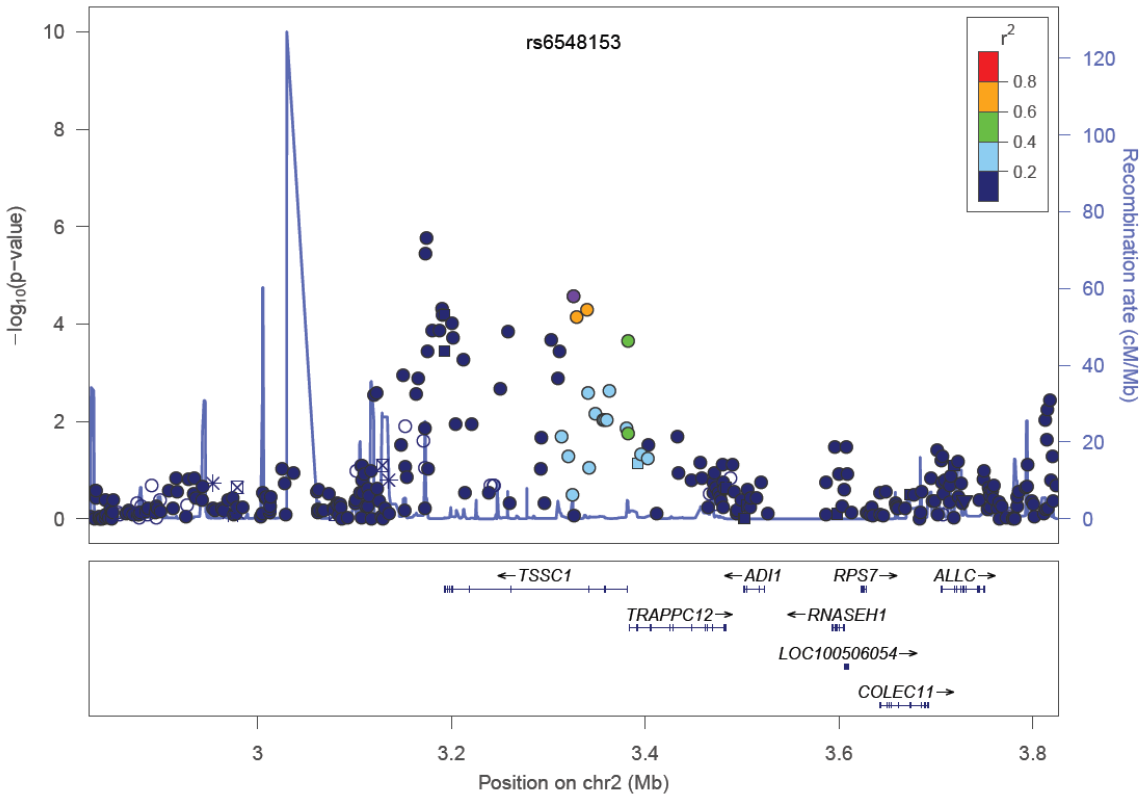


A4.18 Regional association plot of the index SNP - rs11880216. LYPD3: LY6/PLAUR domain containing 3; PINLYP: phospholipase A2 inhibitor and LY6/PLAUR domain containing; IRGC: immunity-related GTPase family, cinema; LYPD5: LY6/PLAUR domain containing 5; ZNF45: zinc finger protein 45; ZNF222: zinc finger protein 222; ZNF234: zinc finger protein 234; ZNF233: zinc finger protein 233; ZNF285: zinc finger protein 285; PHLDB3: pleckstrin homology-like domain, family B, member 3; IRGQ: immunity-related GTPase family, Q; SMG9: smg-9 homolog, nonsense mediated mRNA decay factor (C. elegans); ZNF283: zinc finger protein 283; ZNF 221: zinc finger protein 221; ZNF223: zinc finger protein 223; ZNF226: zinc finger protein 226; ZNF235: zinc finger protein 235; ETHE1: ethylmalonic encephalopathy 1; ZNF428: zinc finger protein 428; KCNN4: potassium intermediate/small conductance calcium-activated channel, subfamily N, member 4; ZNF404: zinc finger protein 404; ZNF155: zinc finger protein 155; ZNF224: zinc finger protein 224; ZNF227: zinc finger protein 227; ZFP112: ZNF575: zinc finger protein 575; PLAUR: plasminogen activator, urokinase receptor; ZNF284: zinc finger protein 284; ZNF229: zinc finger protein 229; XRCC1: ZNF230: zinc finger protein 230; ZNF225: zinc finger protein 225; ZNF576: zinc finger protein 576; SRRM5: serine/arginine repetitive matrix 5; CADM4: cell adhesion molecule 4.

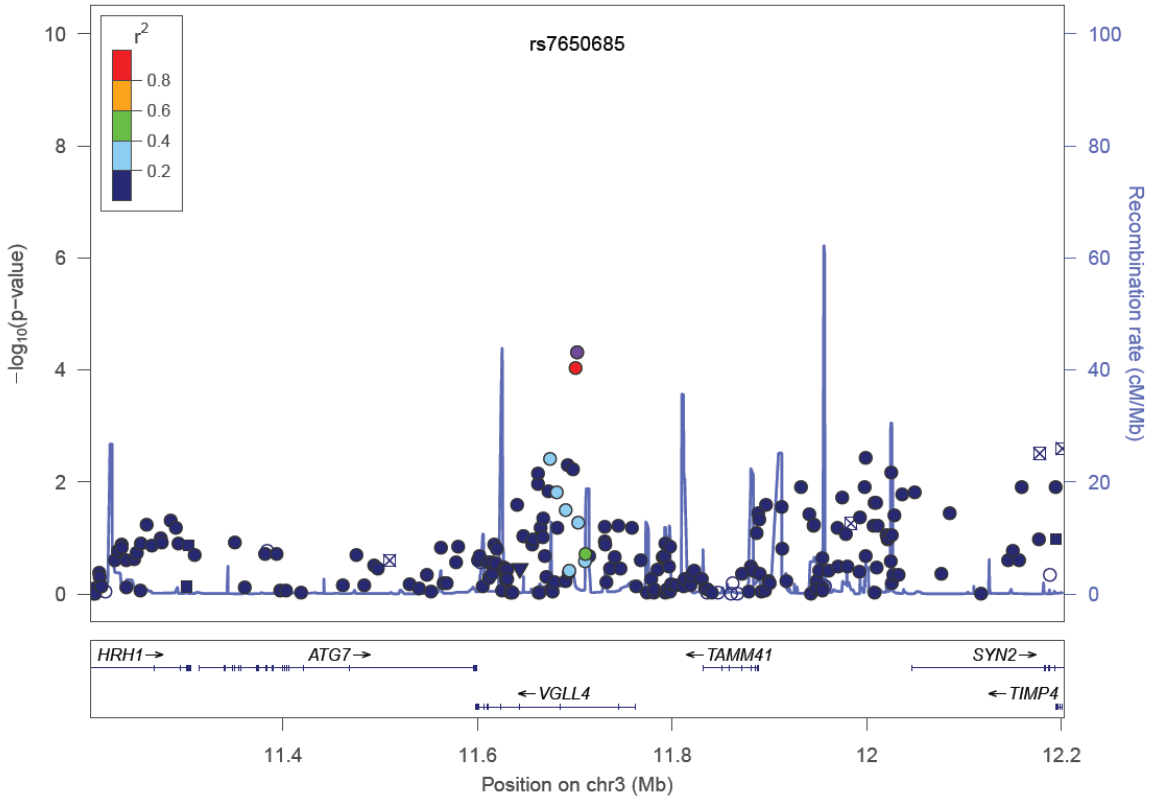
A5. Regional association plots of key markers (or index SNPs, in purple; see A5.1-A5.12) and 500Kb flanking region on each side of the markers for the Japanese endurance cohort. –log10 transformed P values on the Y-axis indicate the strength of the association with elite endurance status in the Japanese cohort. The level of LD between the index SNP and its surrounding SNPs as well as the recombination rate are estimated using 1000 Genomes ASN samples (Mar 2012). The level of LD is indicated by the colour key with red corresponding to high LD, and the recombination rate is represented by the blue line. Functional annotation key: triangle = framestop/splice, inverted triangle = non-synonymous, square = synonymous/UTR, star = conserved transcription factor binding site, square with diagonal lines = region is highly conserved in placental mammals, circle = no annotation.



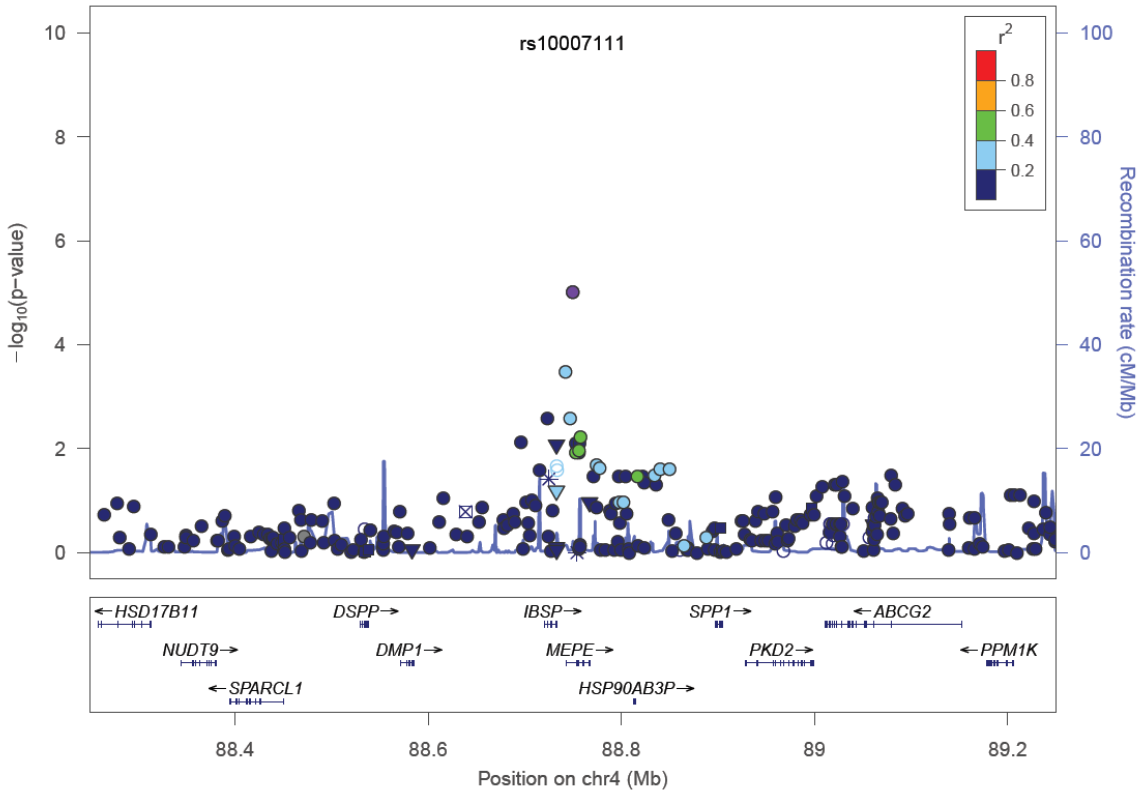
A5.1 Regional association plot of the index SNP – rs921665. TSSC1: tumor suppressing subtransferable candidate 1; ADI1: acireductone dioxygenase 1; RPS7: ribosomal protein S7; TRAPPC12: trafficking protein particle complex 12; RNASEH1: ribonuclease H1; COLEC11: collectin sub-family member 11.



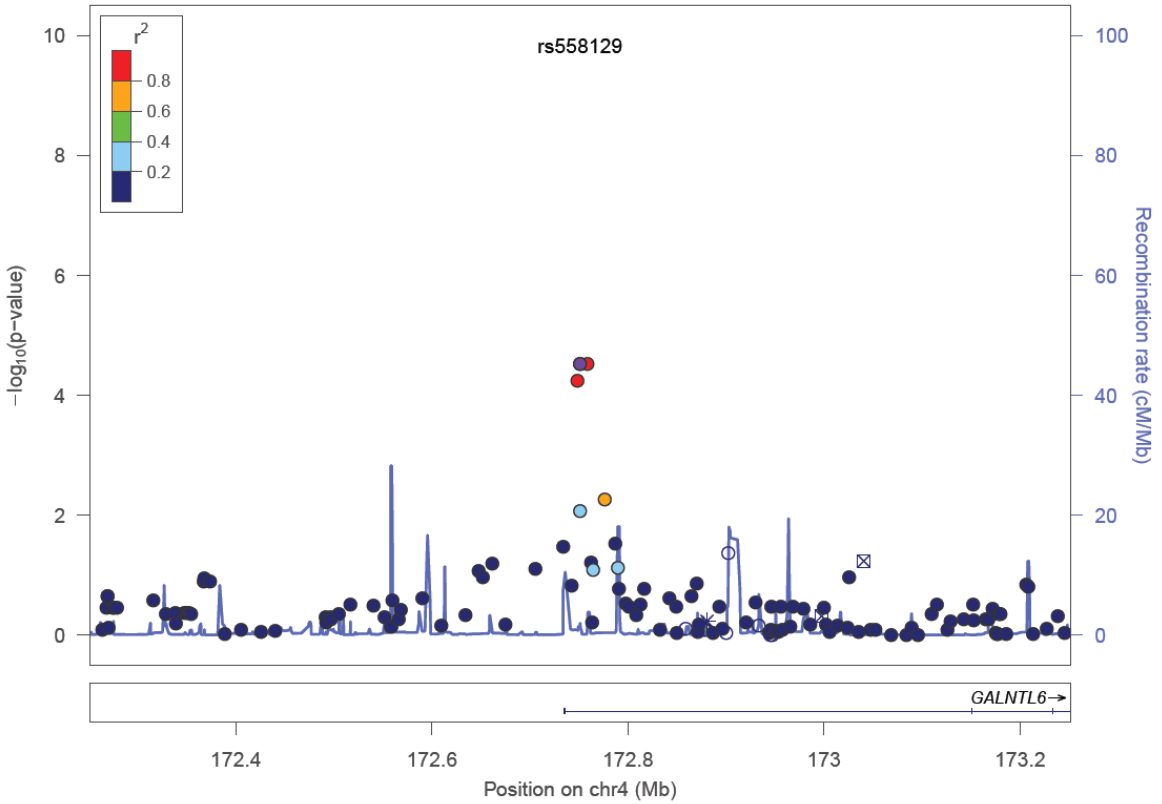
A5.2 Regional association plot of the index SNP – rs6548153. TSSC1: tumor suppressing subtransferable candidate 1; ADI1: acireductone dioxygenase 1; RPS7: ribosomal protein S7; TRAPPC12: trafficking protein particle complex 12; RNASEH1: ribonuclease H1; COLEC11: collectin sub-family member 11.



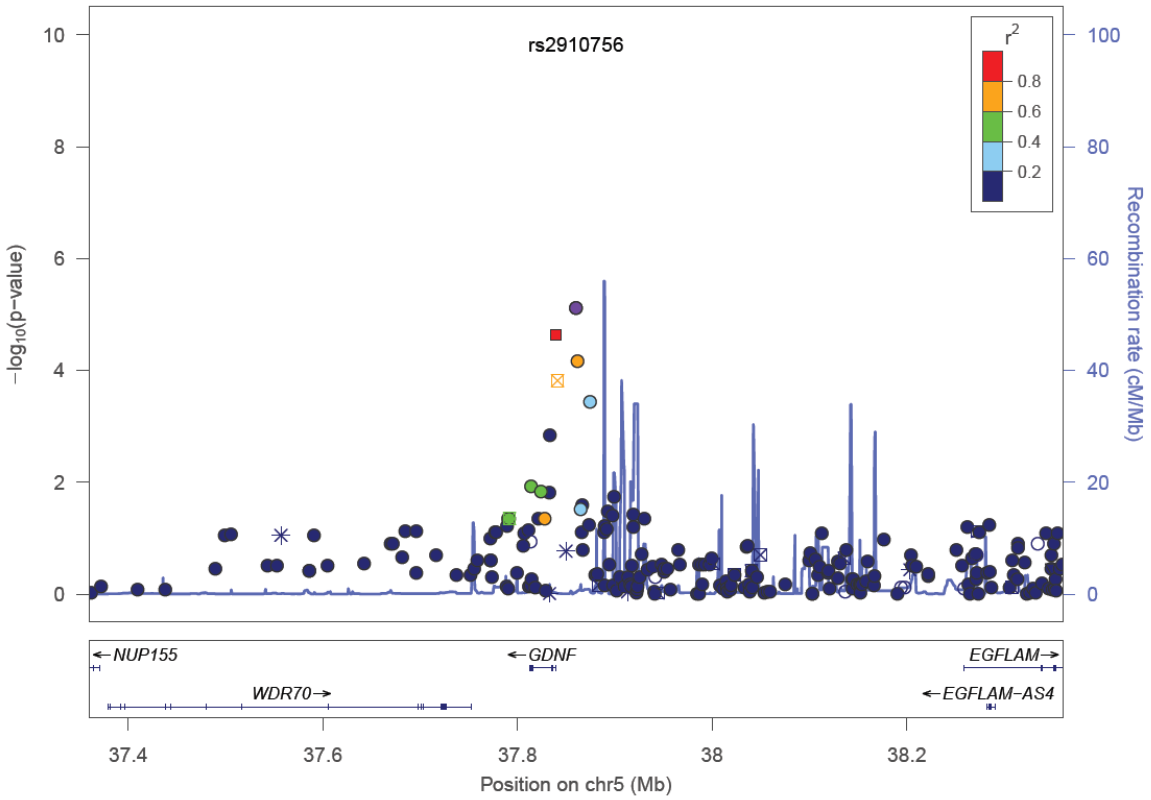
A5.3 Regional association plot of the index SNP – rs7650685. HRH1: histamine receptor H1; ATG7: autophagy related 7; TAMM41: TAM41, mitochondrial translocator assembly and maintenance protein, homolog (S. cerevisiae); SYN2: synapsin II; VGLL4: vestigial like 4 (Drosophila); TIMP4: TIMP metallopeptidase inhibitor 4.



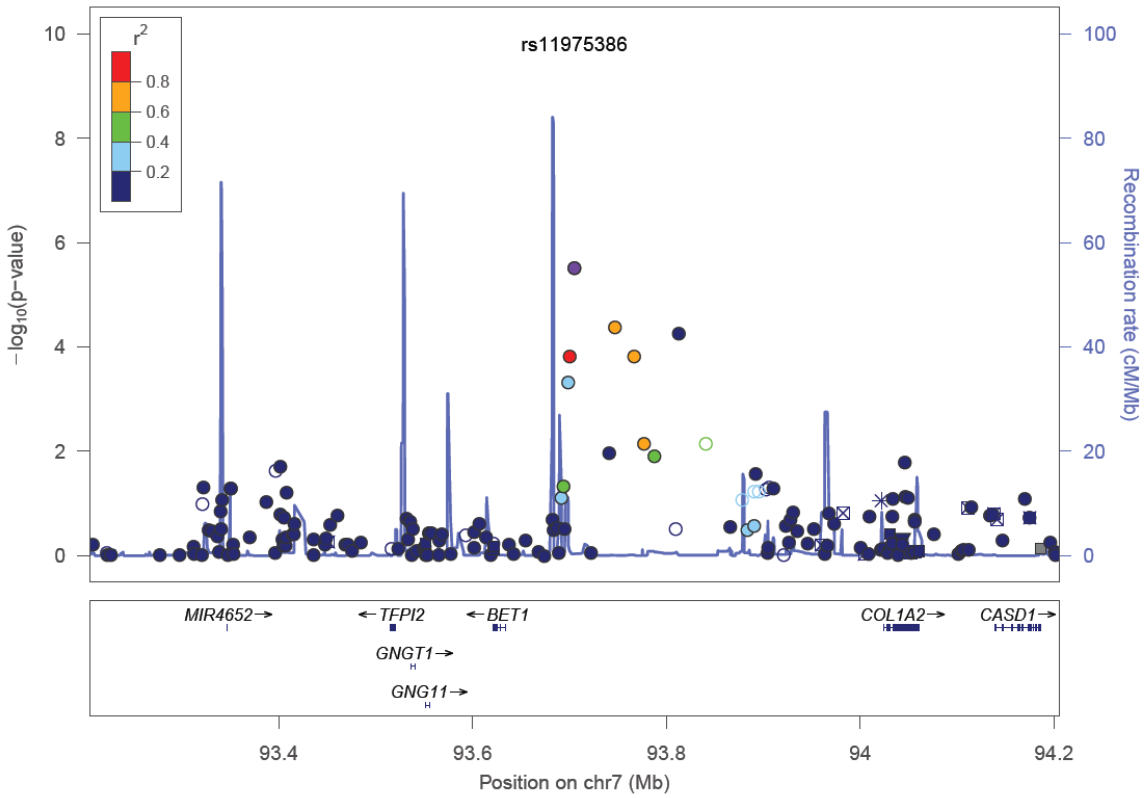
A5.4 Regional association plot of the index SNP – rs10007111. HSD17B11: hydroxysteroid (17-beta) dehydrogenase 11; DSPP: dentin sialophosphoprotein; IBSP: integrin-binding sialoprotein; SPP1: secreted phosphoprotein 1; ABCG2: ATP-binding cassette, sub-family G (WHITE), member 2; NUDT9: nudix (nucleoside diphosphate linked moiety X)-type motif 9; DMP1: dentin matrix acidic phosphoprotein 1; MEPE: matrix extracellular phosphoglycoprotein; PKD2: polycystic kidney disease 2 (autosomal dominant); PPM1K: protein phosphatase, Mg2+/Mn2+ dependent, 1K; SPARCL1: SPARC-like 1 (hevin); HSP90AB3P: heat shock protein 90kDa alpha (cytosolic), class B member 3, pseudogene.



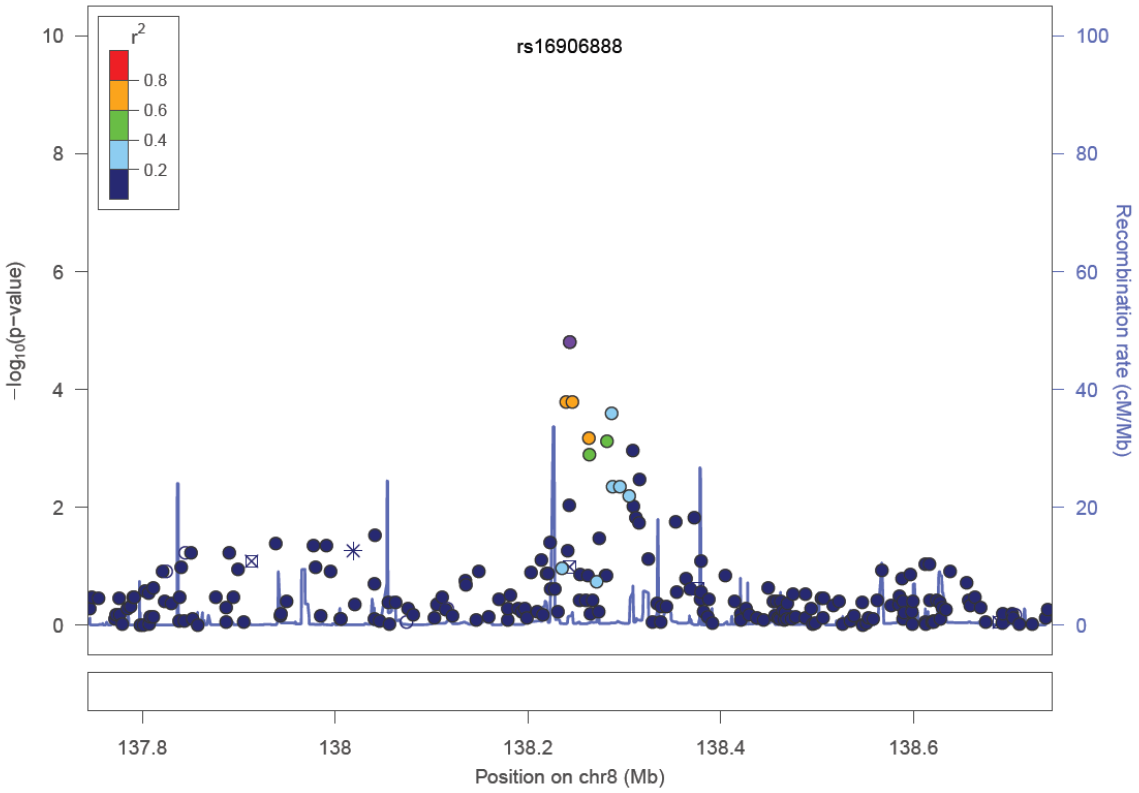
A5.5 Regional association plot of the index SNP – rs558129. GALNTL6: UDP-N-acetyl-alpha-D-galactosamine:polypeptide N-acetylgalactosaminyltransferase-like 6.



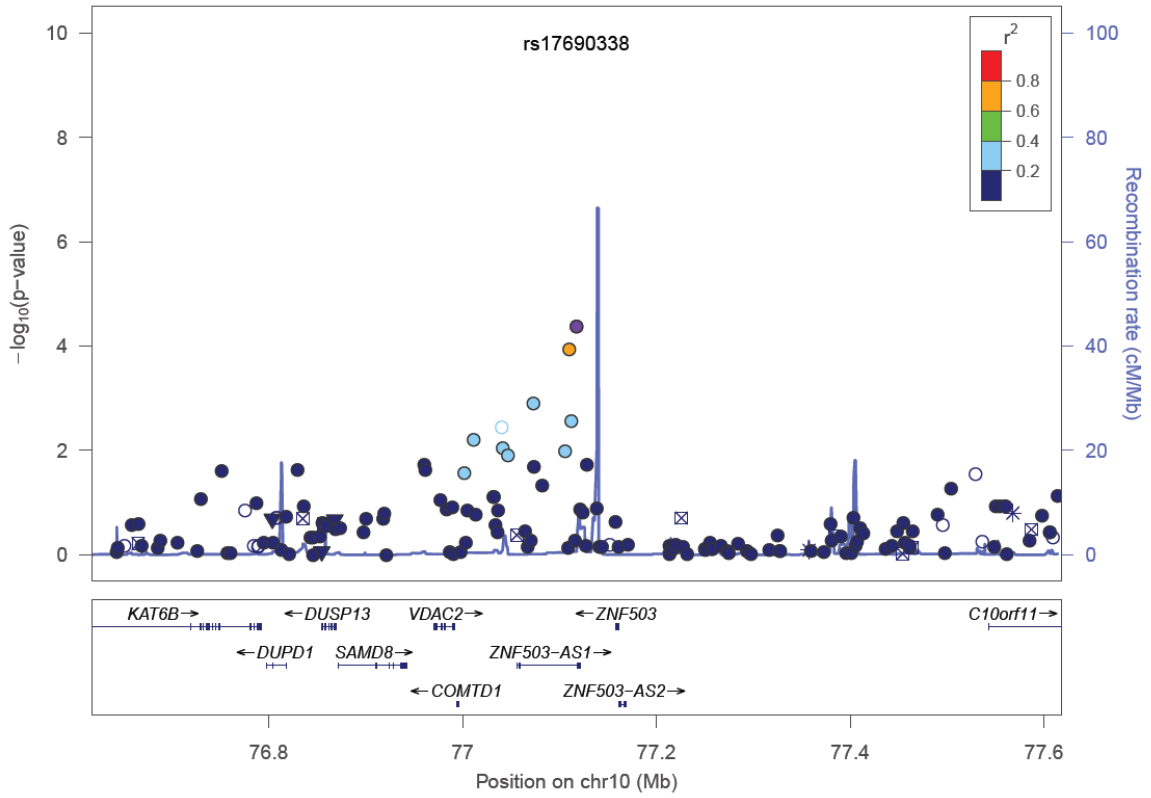
A5.6 Regional association plot of the index SNP – rs2910756. NUP155: nucleoporin 155kDa; GDNF: glial cell derived neurotrophic factor; EGFLAM: EGF-like, fibronectin type III and laminin G domains; WDR70: WD repeat domain 70; EGFLAM-AS4: EGFLAM antisense RNA 4.



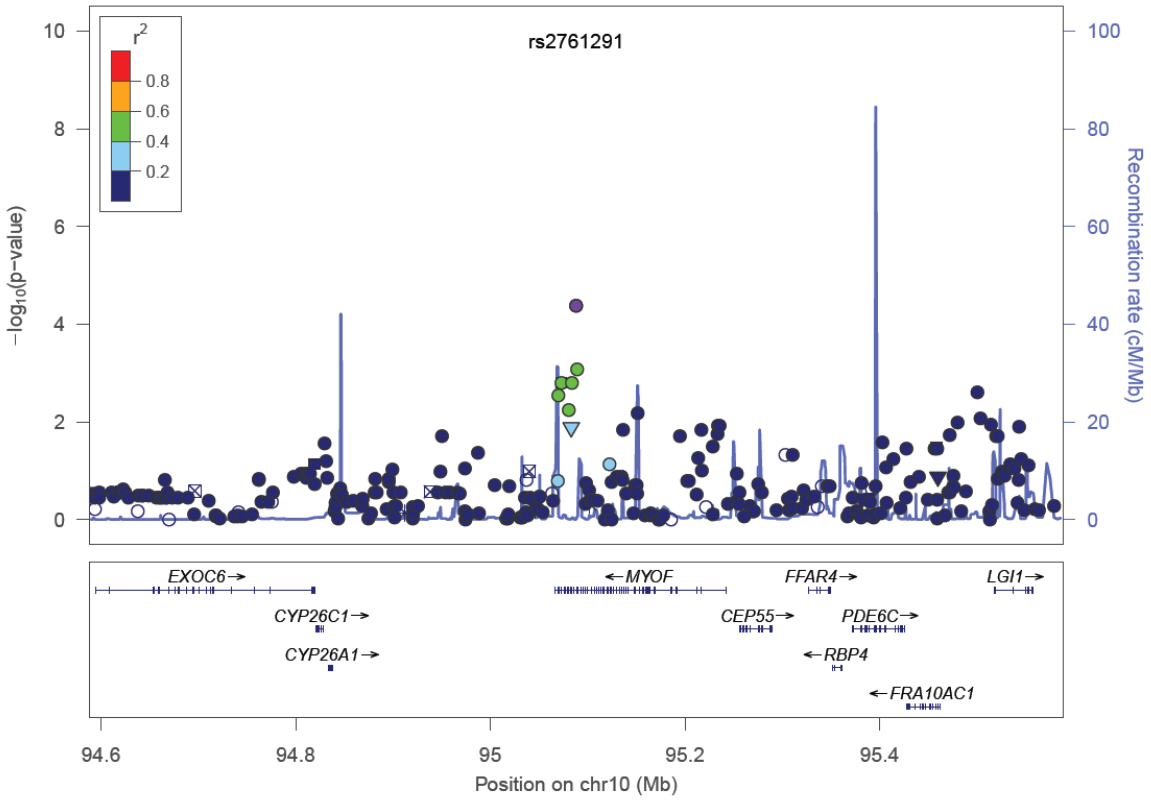
A5.7 Regional association plot of the index SNP – rs11975386. MIR4652: microRNA 4652; TFPI2: tissue factor pathway inhibitor 2; BET1: Bet1 golgi vesicular membrane trafficking protein; COL1A2: collagen, type I, alpha 2; CASD1: CAS1 domain containing 1; GNGT1: guanine nucleotide binding protein (G protein), gamma transducing activity polypeptide 1; GNG11: guanine nucleotide binding protein (G protein), gamma 11.



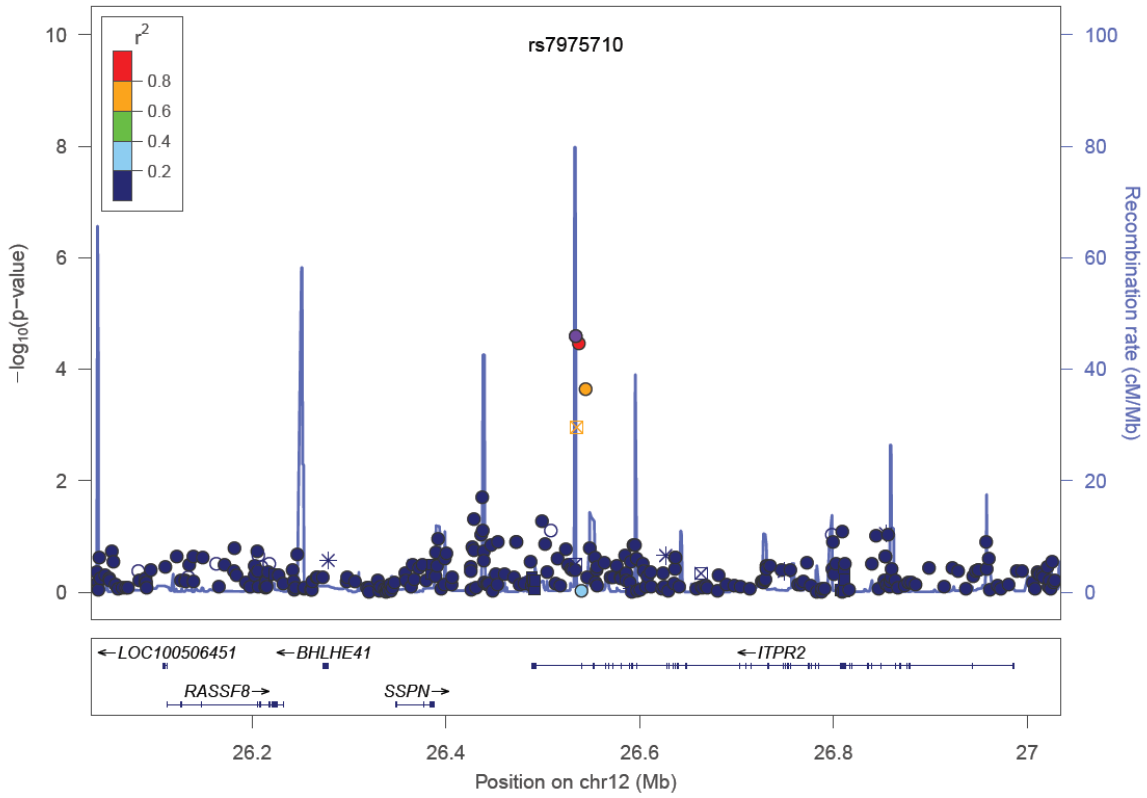
A5.8 Regional association plot of the index SNP – rs16906888.



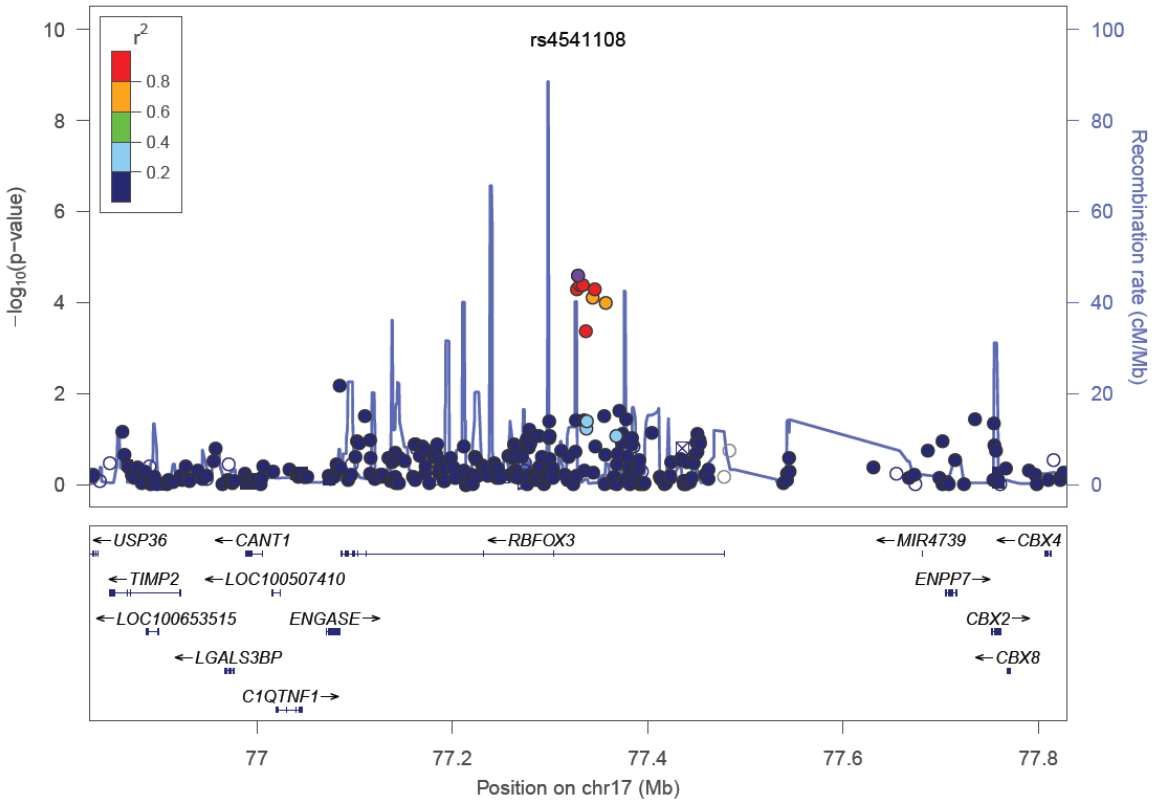
A5.9 Regional association plot of the index SNP – rs17690338. KAT6B: K(lysine) acetyltransferase 6B; DUSP13: dual specificity phosphatase 13; VDAC2: voltage-dependent anion channel 2; ZNF503: zinc finger protein 503; C10orf11: chromosome 10 open reading frame 11; DUPD1: dual specificity phosphatase and pro isomerase domain containing 1; SAMD8: sterile alpha motif domain containing 8; ZNF503-AS1: ZNF503 antisense RNA 1; COMTD1: catechol-O-methyltransferase domain containing 1; ZNF503-AS2: ZNF503 antisense RNA 2.



A5.10 Regional association plot of the index SNP – rs2761291. EXOC6: exocyst complex component 6; MYOF: myoferlin; FFAR4: free fatty acid receptor 4; LGI1: leucine-rich, glioma inactivated 1; CYP26C1: cytochrome P450, family 26, subfamily C, polypeptide 1; CEP55: centrosomal protein 55kDa; PDE6C: phosphodiesterase 6C, cGMP-specific, cone, alpha prime; CYP26A1: cytochrome P450, family 26, subfamily A, polypeptide 1; RBP4: retinol binding protein 4, plasma; FRA10AC1: fragile site, folic acid type, rare, fra(10)(q23.3) or fra(10)(q24.2) candidate 1.



A5.11 Regional association plot of the index SNP – rs7975710. BHLHE41: basic helix-loop-helix family, member e41; ITPR2: inositol 1,4,5-trisphosphate receptor, type 2; RASSF8: Ras association (RalGDS/AF-6) domain family (N-terminal) member 8; SSPN: sarcospan.



A5.12 Regional association plot of the index SNP – rs4541108. USP36: ubiquitin specific peptidase 36; CANT1: calcium activated nucleotidase 1; RBFOX3: RNA binding protein, fox-1 homolog (C. elegans) 3; MIR4739: miroRNA 4739; CBX4: chromobox homolog 4; TIMP2: TIMP metallopeptidase inhibitor 2; ENPP7: ectonucleotide pyrophosphatase/phosphodiesterase 7; ENGASE: endo-beta-N-acetylglucosaminidase; CBX2: chromobox homolog 2; LGALS3BP: lectin, galactoside-binding, soluble, 3 binding protein; CBX8: chromobox homolog 8; C1QTNF1: C1q and tumor necrosis factor related protein 1.

References

1. Davids K, Baker J. Genes, environment and sport performance: why the nature-nurture dualism is no longer relevant. *Sports Med.* 2007;37(11):961-80.

2. Klissouras V. Heritability of adaptive variation. *J Appl Physiol.* 1971;31(3):338-44.

3. Klissouras V, Pirnay F, Petit JM. Adaptation to maximal effort: genetics and age. *J Appl Physiol.* 1973;35(2):288-93.

4. Komi PV, Viitasalo JH, Havu M, et al. Skeletal muscle fibres and muscle enzyme activities in monozygous and dizygous twins of both sexes. *Acta Physiol Scand.* 1977;100(4):385-92.

5. Mosher DS, Quignon P, Bustamante CD, et al. A mutation in the myostatin gene increases muscle mass and enhances racing performance in heterozygote dogs. *PLoS Genet.* 2007;3(5):e79.

6. Girgenrath S, Song K, Whittemore LA. Loss of myostatin expression alters fiber-type distribution and expression of myosin heavy chain isoforms in slow- and fast-type skeletal muscle. *Muscle Nerve.* 2005;31(1):34-40.

7. MacArthur DG, Seto JT, Chan S, et al. An Actn3 knockout mouse provides mechanistic insights into the association between a-actinin-3 deficiency and human athletic performance. *Hum Mol Genet.* 2008;17(8):1076-86.

8. Schuelke M, Wagner KR, Stolz LE, et al. Myostatin mutation associated with gross muscle hypertrophy in a child. *N Engl J Med.* 2004;350(26):2682-8.

9. Costa AM, Breitenfeld L, Silva AJ, et al. Genetic inheritance effects on endurance and muscle strength: an update. *Sports Med.* 2012;42(6):449-58.

10. Bouchard C, Simoneau JA, Lortie G, et al. Genetic effects in human skeletal muscle fiber type distribution and enzyme activities. *Can J Physiol Pharmacol.* 1986;64(9):1245-51.

11. Suwa M, Nakamura T, Katsuta S. Heredity of muscle fiber composition and correlated response of the synergistic muscle in rats. *Am J Physiol.* 1996;271(2 Pt 2):R432-6.

12. Klissouras V. Prediction of potential performance with special reference to heredity. *J Sports Med Phys Fitness.* 1973;13(2):100-7.

13. Stubbe JH, Boomsma DI, Vink JM, et al. Genetic influences on exercise participation in 37,051 twin pairs from seven countries. *PLoS One.* 2006;1:e22.

14. Brutsaert TD, Parra EJ. What makes a champion? Explaining variation in human athletic performance. *Respir Physiol Neurobiol.* 2006;151(2-3):109-23.

15. The HERITAGE Family Study. http://www.pbrc.edu/heritage/home.htm. Accessed 11/12/2012, 2012.

16. Rivera MA, Pérusse L, Simoneau JA, et al. Linkage between a muscle-specific CK gene marker and VO2max in the HERITAGE Family Study. *Med Sci Sports Exerc.* 1999;31(5):698-701.

17. Bouchard C, Rankinen T, Chagnon YC, et al. Genomic scan for maximal oxygen uptake and its response to training in the HERITAGE Family Study. *J Appl Physiol.* 2000;88(2):551-9.

18. Rankinen T, Pérusse L, Borecki I, et al. The Na(+)-K(+)-ATPase alpha2 gene and trainability of cardiorespiratory endurance: the HERITAGE family study. *J Appl Physiol.* 2000;88(1):346-51.

19. Rico-Sanz J, Rankinen T, Rice T, et al. Quantitative trait loci for maximal exercise capacity phenotypes and their responses to training in the HERITAGE Family Study. *Physiol Genomics.* 2004;16(2):256-60.

20. Rankinen T, An P, Rice T, et al. Genomic scan for exercise blood pressure in the Health, Risk Factors, Exercise Training and Genetics (HERITAGE) Family Study. *Hypertension.* 2001;38(1):30-7.

21. Rankinen T, Rice T, Boudreau A, et al. Titin is a candidate gene for stroke volume response to endurance training: the HERITAGE Family Study. *Physiol Genomics.* 2003;15(1):27-33.

22. Rankinen T, An P, Pérusse L, et al. Genome-wide linkage scan for exercise stroke volume and cardiac output in the HERITAGE Family Study. *Physiol Genomics.* 2002;10(2):57-62.

23. Spielmann N, Leon AS, Rao DC, et al. Genome-wide linkage scan for submaximal exercise heart rate in the HERITAGE family study. *Am J Physiol Heart Circ Physiol.* 2007;293(6):H3366-71.

24. Sun G, Gagnon J, Chagnon YC, et al. Association and linkage between an insulin-like growth factor-1 gene polymorphism and fat free mass in the HERITAGE Family Study. *Int J Obes Relat Metab Disord.* 1999;23(9):929-35.

25. Chagnon YC, Rice T, Pérusse L, et al. Genomic scan for genes affecting body composition before and after training in Caucasians from HERITAGE. *J Appl Physiol.* 2001;90(5):1777-87.

26. Lanouette CM, Chagnon YC, Rice T, et al. Uncoupling protein 3 gene is associated with body composition changes with training in HERITAGE study. *J Appl Physiol.* 2002;92(3):1111-8.

27. Rice T, Chagnon YC, Pérusse L, et al. A genomewide linkage scan for abdominal subcutaneous and visceral fat in black and white families: The HERITAGE Family Study. *Diabetes.* 2002;51(3):848-55.

28. An P, Hong Y, Weisnagel SJ, et al. Genomic scan of glucose and insulin metabolism phenotypes: the HERITAGE Family Study. *Metabolism.* 2003;52(2):246-53.

29. Lakka TA, Rankinen T, Weisnagel SJ, et al. A quantitative trait locus on 7q31 for the changes in plasma insulin in response to exercise training: the HERITAGE Family Study. *Diabetes.* 2003;52(6):1583-7.

30. Simonen RL, Rankinen T, Perusse L, et al. Genome-wide linkage scan for physical activity levels in the Quebec Family study. *Med Sci Sports Exerc.* 2003;35(8):1355-9.

31. Cai G, Cole SA, Butte N, et al. A quantitative trait locus on chromosome 18q for physical activity and dietary intake in Hispanic children. *Obesity (Silver Spring).* 2006;14(9):1596-604.

32. De Moor MH, Posthuma D, Hottenga JJ, et al. Genome-wide linkage scan for exercise participation in Dutch sibling pairs. *Eur J Hum Genet.* 2007;15(12):1252-9.

33. De Moor MH, Spector TD, Cherkas LF, et al. Genome-wide linkage scan for athlete status in 700 British female DZ twin pairs. *Twin Res Hum Genet.* 2007;10(6):812-20.

34. Huygens W, Thomis MA, Peeters MW, et al. Linkage of myostatin pathway genes with knee strength in humans. *Physiol Genomics.* 2004;17(3):264-70.

35. Huygens W, Thomis MA, Peeters MW, et al. Quantitative trait loci for human muscle strength: linkage analysis of myostatin pathway genes. *Physiol Genomics.* 2005;22(3):390-7.

36. Windelinckx A, De Mars G, Huygens W, et al. Comprehensive fine mapping of chr12q12-14 and follow-up replication identify activin receptor 1B (ACVR1B) as a muscle strength gene. *Eur J Hum Genet.* 2011;19(2):208-15.

37. Ardlie KG, Kruglyak L, Seielstad M. Patterns of linkage disequilibrium in the human genome. *Nat Rev Genet.* 2002;3(4):299-309.

38. Dean M. Approaches to identify genes for complex human diseases: lessons from Mendelian disorders. *Hum Mutat.* 2003;22(4):261-74.

39. Kruglyak L, Nickerson DA. Variation is the spice of life. *Nat Genet.* 2001;27(3):234-6.

40. Carlson CS, Eberle MA, Kruglyak L, et al. Mapping complex disease loci in whole-genome association studies. *Nature.* 2004;429(6990):446-52.

41. Balding DJ. A tutorial on statistical methods for population association studies. *Nat Rev Genet.* 2006;7(10):781-91.

42. Lewontin RC. The Interaction of Selection and Linkage. I. General Considerations; Heterotic Models. *Genetics.* 1964;49(1):49-67.

43. Rankinen T, Pérusse L, Rauramaa R, et al. The human gene map for performance and health-related fitness phenotypes: the 2001 update. *Med Sci Sports Exerc.* 2002;34(8):1219-33.

44. Pérusse L, Rankinen T, Rauramaa R, et al. The human gene map for performance and health-related fitness phenotypes: the 2002 update. *Med Sci Sports Exerc.* 2003;35(8):1248-64.

45. Rankinen T, Pérusse L, Rauramaa R, et al. The human gene map for performance and health-related fitness phenotypes: the 2003 update. *Med Sci Sports Exerc.* 2004;36(9):1451-69.

46. Wolfarth B, Bray MS, Hagberg JM, et al. The human gene map for performance and health-related fitness phenotypes: the 2004 update. *Med Sci Sports Exerc.* 2005;37(6):881-903.

47. Rankinen T, Bray MS, Hagberg JM, et al. The human gene map for performance and health-related fitness phenotypes: the 2005 update. *Med Sci Sports Exerc.* 2006;38(11):1863-88.

48. Bray MS, Hagberg JM, Pérusse L, et al. The human gene map for performance and health related phenotypes: the 2006-2007 update. *Med Sci Sports Exerc.* 2009;41(1):35-73.

49. Rankinen T, Roth SM, Bray MS, et al. Advances in exercise, fitness, and performance genomics. *Med Sci Sports Exerc.* 2010;42(5):835-46.

50. Hagberg JM, Rankinen T, Loos RJ, et al. Advances in exercise, fitness, and performance genomics in 2010. *Med Sci Sports Exerc.* 2011;43(5):743-52.

51. Roth SM, Rankinen T, Hagberg JM, et al. Advances in exercise, fitness, and performance genomics in 2011. *Med Sci Sports Exerc.* 2012;44(5):809-17.

52. Pérusse L, Rankinen T, Hagberg JM, et al. Advances in Exercise, Fitness, and Performance Genomics in 2012. *Med Sci Sports Exerc.* 2013;45(5):824-31.

53. Pitsiladis Y, Wang G, Wolfarth B. Genomics of aerobic capacity and endurance performance: clinical implications. In: Pescatello LS, Roth SM, eds. *Exercise Genomics*: Humana Press; 2011:179-229.

54. MacArthur DG, North KN. Genes and human elite athletic performance. *Hum Genet.* 2005;116(5):331-9.

55. Lewis CM. Genetic association studies: design, analysis and interpretation. *Brief Bioinform.* 2002;3(2):146-53.

56. Romero R, Kuivaniemi H, Tromp G, et al. The design, execution, and interpretation of genetic association studies to decipher complex diseases. *Am J Obstet Gynecol.* 2002;187(5):1299-312.

57. Tabor HK, Risch NJ, Myers RM. Candidate-gene approaches for studying complex genetic traits: practical considerations. *Nat Rev Genet.* 2002;3(5):391-7.

58. Padmanabhan S, Melander O, Hastie C, et al. Hypertension and genome-wide association studies: combining high fidelity phenotyping and hypercontrols. *J Hypertens.* 2008;26(7):1275-81.

59. Klein RJ, Zeiss C, Chew EY, et al. Complement factor H polymorphism in age-related macular degeneration. *Science.* 2005;308(5720):385-9.

60. Visscher PM, Brown MA, McCarthy MI, et al. Five years of GWAS discovery. *Am J Hum Genet.* 2012;90(1):7-24.

61. Whole-genome genotyping and copy number variation analysis. http://www.illumina.com/applications/detail/snp_genotyping_and_cnv_analysis/whole_genome_genotyping_and_copy_number_variation_analysis.ilmn. Accessed 12/12/2012.

62. International HapMap 3 Consortium, Altshuler DM, Gibbs RA, et al. Integrating common and rare genetic variation in diverse human populations. *Nature.* 2010;467(7311):52-8.

63. 1000 Genomes Project Consortium, Abecasis GR, Auton A, et al. An integrated map of genetic variation from 1,092 human genomes. *Nature.* 2012;491(7422):56-65.

64. Tishkoff SA, Reed FA, Friedlaender FR, et al. The genetic structure and history of Africans and African Americans. *Science.* 2009;324(5930):1035-44.

65. Conrad DF, Jakobsson M, Coop G, et al. A worldwide survey of haplotype variation and linkage disequilibrium in the human genome. *Nat Genet.* 2006;38(11):1251-60.

66. Jakobsson M, Scholz SW, Scheet P, et al. Genotype, halotype and copy-number variation in worldwide human populations. *Nature.* 2008;451(7181):998-1003.

67. Campbell MC, Tishkoff SA. African genetic diversity: implications for human demographic history, modern human origins, and complex disease mapping. *Annu Rev Genomics Hum Genet.* 2008;9:403-33.

68. DeGiorgio M, Jakobsson M, Rosenberg NA. Out of Africa: modern human origins special feature: explaining worldwide patterns of human genetic variation using a coalescent-based serial founder model of migration outward from Africa. *Proc Natl Acad Sci U.S.A.* 2009;106(38):16057-62.

69. Rosenberg NA, Huang L, Jewett EM, et al. Genome-wide association studies in diverse populations. *Nat Rev Genet.* 2010;11(5):356-66.

70. González-Neira A, Ke X, Lao O, et al. The portability of tagSNPs across populations: a worldwide survey. *Genome Res.* 2006;16(3):323-30.

71. Teo YY, Small KS, Kwiatkowski DP. Methodological challenges of genome-wide association analysis in Africa. *Nat Rev Genet.* 2010;11(2):149-60.

72. Jorgenson E, Witte JS. A gene-centric approach to genome-wide association studies. *Nat Rev Genet.* 2006;7(11):885-91.

73. Stranger BE, Stahl EA, Raj T. Progress and promise of genome-wide association studies for human complex trait genetics. *Genetics.* 2011;187(2):367-83.

74. Dewan A, Liu M, Hartman S, et al. HTRA1 promoter polymorphism in wet age-related macular degeneration. *Science.* 2006;314(5801):989-92.

75. Psychiatric GWAS Consortium Coordinating Committee, Cichon S, Craddock N, et al. Genomewide association studies: history, rationale, and prospects for psychiatric disorders. *Am J Psychiatry.* 2009;166(5):540-56.

76. Pe'er I, Yelensky R, Altshuler D, et al. Estimation of the multiple testing burden for genomewide association studies of nearly all common variants. *Genet Epidemiol.* 2008;32(4):381-5.

77. Potkin SG, Turner JA, Guffanti G, et al. Genome-wide strategies for discovering genetic influences on cognition and cognitive disorders: methodological considerations. *Cogn Neuropsychiatry.* 2009;14(4-5):391-418.

78. Rothman KJ. Statistics in nonrandomized studies. *Epidemiology.* 1990;1(6):417-8.

79. Westfall PH, Young SS. Resampling-based multiple testing: examples and methods for p-value adjustment. New York: John Wiley & Sons; 1993.

80. Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J Roy Statist Soc Ser B (Methodological).* 1995;57:289-300.

81. Cardon LR, Palmer LJ. Population stratification and spurious allelic association. *Lancet.* 2003;361(9357):598-604.

82. Tian C, Gregersen PK, Seldin MF. Accounting for ancestry: population substructure and genome-wide association studies. *Hum Mol Genet.* 2008;17(R2):R143-50.

83. Wacholder S, Hartge P, Palmer LJ. Case-control study. In: Elston R, Olson JM, Palmer LJ, eds. *Biostatistical genetics and genetic epidemiology*. Chichester: Wiley; 2002:95-109.

84. Knowler WC, Williams RC, Pettitt DJ, et al. Gm3;5,13,14 and type 2 diabetes mellitus: an association in American Indians with genetic admixture. *Am J Hum Genet.* 1988;43(4):520-6.

85. Gelernter J, Goldman D, Risch N. The A allele at the D2 dopamine receptor gene and alcoholism. A reappraisal. *JAMA.* 1993;269(13):1673-7.

86. Pato CN, Macciardi F, Pato MT, et al. Review of the putative association of dopamine D2 receptor and alcoholism: a meta-analysis. *Am J Med Genet.* 1993;48(2):78-82.

87. Campbell CD, Ogburn EL, Lunetta KL, et al. Demonstrating stratification in a European American population. *Nat Genet.* 2005;37(8):868-72.

88. Tian C, Plenge RM, Ransom M, et al. Analysis and application of European genetic substructure using 300 K SNP information. *PLoS Genet.* 2008;4(1):e4.

89. Freedman ML, Reich D, Penney KL, et al. Assessing the impact of population stratification on genetic association studies. *Nat Genet.* 2004;36(4):388-93.

90. Marchini J, Cardon LR, Phillips MS, et al. The effects of human population structure on large genetic association studies. *Nat Genet.* 2004;36(5):512-7.

91. Wacholder S, Rothman N, Caporaso N. Population stratification in epidemiologic studies of common genetic variants and cancer: quantification of bias. *J Natl Cancer Inst.* 2000;92(14):1151-8.

92. Devlin B, Roeder K. Genomic control for association studies. *Biometrics.* 1999;55(4):997-1004.

93. Pritchard JK, Rosenberg NA. Use of unlinked genetic markers to detect population stratification in association studies. *Am J Hum Genet.* 1999;65(1):220-8.

94. Hoggart CJ, Parra EJ, Shriver MD, et al. Control of confounding of genetic associations in stratified populations. *Am J Hum Genet.* 2003;72(6):1492-504.

95. Hao K, Li C, Rosenow C, et al. Detect and adjust for population stratification in population-based association study using genomic control markers: an application of Affymetrix Genechip Human Mapping 10K array. *Eur J Hum Genet.* 2004;12(12):1001-6.

96. Price AL, Zaitlen NA, Reich D, et al. New approaches to population stratification in genome-wide association studies. *Nat Rev Genet.* 2010;11(7):459-63.

97. Price AL, Patterson NJ, Plenge RM, et al. Principal components analysis corrects for stratification in genome-wide association studies. *Nat Genet.* 2006;38(8):904-9.

98. Pritchard JK, Stephens M, Donnelly P. Inference of population structure using multilocus genotype data. *Genetics.* 2000;155(2):945-59.

99. Pritchard JK, Stephens M, Rosenberg NA, et al. Association mapping in structured populations. *Am J Hum Genet.* 2000;67(1):170-81.

100. Pritchard JK, Donnelly P. Case-control studies of association in structured or admixed populations. *Theor Popul Biol.* 2001;60(3):227-37.

101. Satten GA, Flanders WD, Yang Q. Accounting for unmeasured population substructure in case-control studies of genetic association using a novel latent-class model. *Am J Hum Genet.* 2001;68(2):466-77.

102. Smith LI. A Tutorial on Principal Component Analysis. 2002. http://www.iro.umontreal.ca/~pift6080/H08/documents/papers/pca_tutorial.pdf. Accessed 17/12/12.

103. Patterson N, Price AL, Reich D. Population structure and eigenanalysis. *PLoS Genet.* 2006;2(12):e190.

104. McGinnis R, Shifman S, Darvasi A. Power and efficiency of the TDT and case-control design for association scans. *Behav Genet.* 2002;32(2):135-44.

105. Nielsen DM, Weir BS. Association studies under general disease models. *Theor Popul Biol.* 2001;60(3):253-63.

106. Mitchell AA, Cutler DJ, Chakravarti A. Undetected genotyping errors cause apparent overtransmission of common alleles in the transmission/disequilibrium test. *Am J Hum Genet.* 2003;72(3):598-610.

107. Lander ES. The new genomics: global views of biology. *Science.* 1996;274(5287):536-9.

108. Reich DE, Lander ES. On the allelic spectrum of human disease. *Trends Genet.* 2001;17(9):502-10.

109. Pritchard JK, Cox NJ. The allelic architecture of human disease genes: common disease-common variant...or not? *Hum Mol Genet.* 2002;11(20):2417-23.

110. Botstein D, Risch N. Discovering genotypes underlying human phenotypes: past successes for mendelian disease, future approaches for complex disease. *Nat Genet.* 2003;33:Suppl:228-37.

111. Maher B. Personal genomes: The case of the missing heritability. *Nature.* 2008;456(7218):18-21.

112. Manolio T, A., Collins FS, Cox NJ, et al. Finding the missing heritability of complex diseases. *Nature.* 2009;461(7265):747-53.

113. Visscher PM, Hill WG, Wray NR. Heritability in the genomics era--concepts and misconceptions. *Nat Rev Genet.* 2008;9(4):255-66.

114. Cirulli ET, Goldstein DB. Uncovering the roles of rare variants in common disease through whole-genome sequencing. *Nat Rev Genet.* 2010;11(6):415-25.

115. Gibson G. Rare and common variants: twenty arguments. *Nat Rev Genet.* 2012;13(2):135-45.

116. Feldman MW, Lewontin RC. The heritability hang-up. *Science.* 1975;190(4220):1163-8.

117. Eichler EE, Flint J, Gibson G, et al. Missing heritability and strategies for finding the underlying causes of complex disease. *Nat Rev Genet.* 2010;11(6):446-50.

118. Schork NJ, Murray SS, Frazer KA, et al. Common vs. rare allele hypotheses for complex diseases. *Curr Opin Genet Dev.* 2009;19(3):212-9.

119. Hindorff LA, MacArthur J, Morales J, et al. A Catalog of Published Genome-Wide Association Studies. www.genome.gov/gwastudies. Accessed 18/12/2012.

120. Kiel DP, Demissie S, Dupuis J, et al. Genome-wide association with bone mass and geometry in the Framingham Heart Study. *BMC Med Genet.* 2007;8:Suppl 1:S14.

121. Cupples LA, Arruda HT, Benjamin EJ, et al. The Framingham Heart Study 100K SNP genome-wide association study resource: overview of 17 phenotype working group reports. *BMC Med Genet.* 2007;8 Suppl 1:S1.

122. Xiong DH, Liu XG, Guo YF, et al. Genome-wide association and follow-up replication studies identified ADAMTS18 and TGFBR3 as bone mass candidate genes in different ethnic groups. *Am J Hum Genet.* 2009;84(3):388-98.

123. De Moor MH, Liu YJ, Boomsma DI, et al. Genome-wide association study of exercise behavior in Dutch and American adults. *Med Sci Sports Exerc.* 2009;41(10):1887-95.

124. Rankinen T, Sung YJ, Sarzynski MA, et al. Heritability of submaximal exercise heart rate response to exercise training is accounted for by nine SNPs. *J Appl Physiol.* 2012;112(5):892-7.

125. Arnett DK, Li N, Tang W, et al. Genome-wide association study identifies single-nucleotide polymorphism in KCNB1 associated with left ventricular mass in humans: the HyperGEN Study. *BMC Med Genet.* 2009;10:43.

126. Hai R, Zhang L, Pei Y, et al. Bivariate genome-wide association study suggests that the DARC gene influences lean body mass and age at menarche. *Sci China Life Sci.* 2012;55(6):516-20.

127. Han Y, Pei Y, Liu Y, et al. Bivariate genome-wide association study suggests fatty acid desaturase genes and cadherin DCHS2 for variation of both compressive strength index and appendicular lean mass in males. *Bone.* 2012;51(6):1000-7.

128. Liu XG, Tan LJ, Lei SF, et al. Genome-wide association and replication studies identified TRHR as an important gene for lean body mass. *Am J Hum Genet.* 2009;84(3):418-23.

129. Sipilä S, Heikkinen E, Cheng S, et al. Endogenous hormones, muscle strength, and risk of fall-related fractures in older women. *J Gerontol A Biol Sci Med Sci.* 2006;61(1):92-6.

130. Karakelides H, Nair KS. Sarcopenia of aging and its metabolic impact. *Curr Top Dev Biol.* 2005;68:123-48.

131. Hansen RD, Raja C, Aslani A, et al. Determination of skeletal muscle and fat-free mass by nuclear and dual-energy x-ray absorptiometry methods in men and women aged 51-84 y (1-3). *Am J Clin Nutr.* 1999;70(2):228-33.

132. Larsson L, Li X, Teresi A, et al. Effects of thyroid hormone on fast- and slow-twitch skeletal muscles in young and old rats. *J Physiol.* 1994;481(Pt 1):149-61.

133. Norenberg KM, Herb RA, Dodd SL, et al. The effects of hypothyroidism on single fibers of the rat soleus muscle. *Can J Physiol Pharmacol.* 1996;74(4):362-7.

134. Soukup T, Jirmanová I. Regulation of myosin expression in developing and regenerating extrafusal and intrafusal muscle fibers with special emphasis on the role of thyroid hormones. *Physiol Res.* 2000;49(6):617-33.

135. Bouchard C, Sarzynski MA, Rice TK, et al. Genomic predictors of the maximal O_2 uptake response to standardized exercise training programs. *J Appl Physiol.* 2011;110(5):1160-70.

136. Pitsiladis Y, Wang G. Necessary advances in exercise genomics and likely pitfalls. *J Appl Physiol.* 2011;110(5):1150-1.

137. Kraft P, Zeggini E, Ioannidis JP. Replication in genome-wide association studies. *Stat Sci.* 2009;24(4):561-73.

138. Wacholder S, Chanock S, Garcia-Closas M, et al. Assessing the probability that a positive report is false: an approach for molecular epidemiology studies. *J Natl Cancer Inst.* 2004;96(6):434-42.

139. Ioannidis JP. Why most published research findings are false. *PLoS Med.* 2005;2(8):e124.

140. Wakefield J. A Bayesian measure of the probability of false discovery in genetic epidemiology studies. *Am J Hum Genet.* 2007;81(2):208-27.

141. Ioannidis JP. Why most discovered true associations are inflated. *Epidemiology.* 2008;19(5):640-8.

142. Xiao R, Boehnke M. Quantifying and correcting for the winner's curse in genetic association studies. *Genet Epidemiol.* 2009;33(5):453-62.

143. Yu K, Chatterjee N, Wheeler W, et al. Flexible design for following up positive findings. *Am J Hum Genet.* 2007;81(3):540-51.

144. Zhong H, Prentice RL. Bias-reduced estimators and confidence intervals for odds ratios in genome-wide association studies. *Biostatistics.* 2008;9(4):621-34.

145. Lanktree MB, Hegele RA, Schork NJ, et al. Extremes of unexplained variation as a phenotype: an efficient approach for genome-wide association studies of cardiovascular disease. *Circ Cardiovasc Genet.* 2010;3(2):215-21.

146. Wang K, Li WD, Zhang CK, et al. A genome-wide association study on obesity and obesity-related traits. *PLoS One.* 2011;6(4):e18939.

147. Emond MJ, Louie T, Emerson J, et al. Exome sequencing of extreme phenotypes identifies DCTN4 as a modifier of chronic Pseudomonas aeruginosa infection in cystic fibrosis. *Nat Genet.* 2012;44(8):886-9.

148. Woods D, Hickman M, Jamshidi Y, et al. Elite swimmers and the D allele of the ACE I/D polymorphism. *Hum Genet.* 2001;108(3):230-2.

149. Yang N, MacArthur DG, Gulbin JP, et al. ACTN3 genotype is associated with human elite athletic performance. *Am J Hum Genet.* 2003;73(3):627-31.

150. Santiago C, Rodríguez-Romo G, Gómez-Gallego F, et al. Is there an association between ACTN3 R577X polymorphism and muscle power phenotypes in young, non-athletic adults? *Scand J Med Sci Sports.* 2010;20(5):771-8.

151. Lucia A, Gómez-Gallego F, Santiago C, et al. ACTN3 genotype in professional endurance cyclists. *Int J Sports Med.* 2006;27(11):880-4.

152. Roth SM, Walsh S, Liu D, et al. The ACTN3 R577X nonsense allele is under-represented in elite-level strength athletes. *Eur J Hum Genet.* 2008;16(3):391-4.

153. Ahmetov II, Druzhevskaya AM, Astratenkova IV, et al. The ACTN3 R577X polymorphism in Russian endurance athletes. *Br J Sports Med.* 2010;44(9):649-52.

154. Scott RA, Irving R, Irwin L, et al. ACTN3 and ACE genotypes in elite Jamaican and US sprinters. *Med Sci Sports Exerc.* 2010;42(1):107-12.

155. QIAamp® DNA Mini and Blood Mini Handbook. April 2010 (3rd edition). Accessed 06/01/2013.

156. Laboratory protocol for manual purification of DNA from 0.5 mL of sample. 2012. Accessed 06/01/2013.

157. Infinium Assay Workflow. Accessed 06/01/2013.

158. Cheng YC, O'Connell JR, Cole JW, et al. Genome-wide association analysis of ischemic stroke in young adults. *G3 (Bethesda).* 2011;1(6):505-14.

159. Purcell S. PLINK (version 1.05). http://pngu.mgh.harvard.edu/purcell/plink/.

160. Purcell S, Neale B, Todd-Brown K, et al. PLINK: a toolset for whole-genome association and population-based linkage analysis. *Am J Hum Genet.* 2007;81(3):559-75.

161. Barrett JC, Fry B, Maller J, et al. Haploview: analysis and visualization of LD and haplotype maps. *Bioinformatics.* 2005;21(2):263-5.

162. Camargo A, Azuaje F, Wang H, et al. Permutation - based statistical tests for multiple hypotheses. *Source Code Biol Med.* 2008;3:15.

163. Zang Y, Fung WK, Zheng G. Simple algorithms to calculate asymptotic null distributions of robust tests in case-control genetic association studies in R. *J Stat Software.* 2010;33(8):1-24.

164. Skol AD, Scott LJ, Abecasis GR, et al. Joint analysis is more efficient than replication-based analysis for two-stage genome-wide association studies. *Nat Genet.* 2006;38(2):209-13.

165. Anderson CA, Pettersson FH, Clarke GM, et al. Data quality control in genetic case-control association studies. *Nat Protoc.* 2010;5(9):1564-73.

166. Huedo-Medina TB, Sanchez-Meca J, Marin-Martinez F, et al. Assessing heterogeneity in meta-analysis: Q statistic or I2 index? *Psychol Methods.* 2006;11(2):193-206.

167. Pruim RJ, Welch RP, Sanna S, et al. LocusZoom: Regional visualization of genome-wide association scan results. *Bioinformatics.* 2010;26(18):2336-7.

168. Kitts A, Sherry S. The Single Nucleotide Polymorphism Database (dbSNP) of Nucleotide Sequence Variation. In: McEntyre J, Ostell J, eds. The NCBI Handbook [Internet]. Bethesda (MD): National Center for Biotechnology Information (US); 2002-. Chapter 5. 2002 Oct 9 [Updated 2011 Feb 2]: http://www.ncbi.nlm.nih.gov/books/NBK21088/.

169. Online Mendelian Inheritance in Man, OMIM®. McKusick-Nathans Institute of Genetic Medicine, Johns Hopkins University (Baltimore, MD), 03/03/2013. http://omim.org/.

170. Wang G, Mikami E, Chiu LL, et al. Association analysis of ACE and ACTN3 in Elite Caucasian and East Asian Swimmers. *Med Sci Sports Exerc.* 2013;45(5):892-900.

171. Costa AM, Silva AJ, Garrido ND, et al. Association between ACE D allele and elite short distance swimming. *Eur J Appl Physiol.* 2009;106(6):785-90.

172. Puthucheary Z, Skipworth JR, Rawal J, et al. The ACE gene and human performance: 12 years on. *Sports Med.* 2011;41(6):433-48.

173. Rigat B, Hubert C, Alhenc-Gelas F, et al. An insertion/deletion polymorphism in the angiotensin I-converting enzyme gene accounting for half the variance of serum enzyme levels. *J Clin Invest.* 1990;86(4):1343-6.

174. Yamamoto K, Kataoka S, Hashimoto N, et al. Serum level and gene polymorphism of angiotensin I converting enzyme in Japanese children. *Acta Paediatr Jpn.* 1997;39(1):1-5.

175. Zhu X, McKenzie CA, Forrester T, et al. Localization of a small genomic region associated with elevated ACE. *Am J Hum Genet.* 2000;67(5):1144-53.

176. Moran CN, Vassilopoulos C, Tsiokanos A, et al. The associations of ACE polymorphisms with physical, physiological and skill parameters in adolescents. *Eur J Hum Genet.* 2006;14(3):332-9.

177. Williams AG, Dhamrait SS, Wootton PT, et al. Bradykinin receptor gene variant and human physical performance. *J Appl Physiol.* 2004;96(3):938-42.

178. Taguchi T, Kishikawa H, Motoshima H, et al. Involvement of bradykinin in acute exercise-induced increase of glucose uptake and GLUT-4 translocation in skeletal muscle: studies in normal and diabetic humans and rats. *Metabolism.* 2000;49(7):920-30.

179. Cleeter MW, Cooper JM, Darley-Usmar VM, et al. Reversible inhibition of cytochrome c oxidase, the terminal enzyme of the mitochondrial respiratory chain, by nitric oxide. Implications for neurodegenerative diseases. *FEBS Lett.* 1994;345(1):50-4.

180. Poderoso JJ, Carreras MC, Lisdero C, et al. Nitric oxide inhibits electron transfer and increases superoxide radical production in rat heart mitochondria and submitochondrial particles. *Arch Biochem Biophys.* 1996;328(1):85-92.

181. Murphey LJ, Gainer JV, Vaughan DE, et al. Angiotensin-converting enzyme insertion/deletion polymorphism modulates the human in vivo metabolism of bradykinin. *Circulation.* 2000;102(8):829-32.

182. Myerson S, Hemingway H, Budget R, et al. Human angiotensin I-converting enzyme gene and endurance performance. *J Appl Physiol.* 1999;87(4):1313-6.

183. Thomis MA, Huygens W, Heuninckx S, et al. Exploration of myostatin polymorphisms and the angiotensin-converting enzyme insertion/deletion genotype in responses of human muscle to strength training. *Eur J Appl Physiol.* 2004;92(3):267-74.

184. Folland J, Leach B, Little T, et al. Angiotensin-converting enzyme genotype affects the response of human skeletal muscle to functional overload. *Exp Physiol.* 2000;85(5):575-9.

185. Pescatello LS, Kostek MA, Gordish-Dressman H, et al. ACE ID genotype and the muscle strength and size response to unilateral resistance training. *Med Sci Sports Exerc.* 2006;38(6):1074-81.

186. Williams AG, Day SH, Folland JP, et al. Circulating angiotensin converting enzyme activity is correlated with muscle strength. *Med Sci Sports Exerc.* 2005;37(6):944-8.

187. Tobina T, Michishita R, Yamasawa F, et al. Association between the angiotensin I-converting enzyme gene insertion/deletion polymorphism and endurance running speed in Japanese runners. *J Physiol Sci.* 2010;60(5):325-30.

188. Kim CH, Cho JY, Jeon JY, et al. ACE DD genotype is unfavorable to Korean short-term muscle power athletes. *Int J Sports Med.* 2010 31(1):65-71.

189. Min SK, Takahashi K, Ishigami H, et al. Is there a gender difference between ACE gene and race distance? *Appl Physiol Nutr Metab.* 2009;34(5):926-32.

190. Goh KP, Chew K, Koh A, et al. The relationship between ACE gene ID polymorphism and aerobic capacity in Asian rugby players. *Singapore Med J.* 2009;50(10):997-1003.

191. Xi Y, Wu YQ, Zhang XL, et al. Research on the relation between ACE gene I/D polymorphisms and sensitivity to endurance training of Han nationality male. *Zhongguo Ying Yong Sheng Li Xue Za Zhi.* 2008;24(3):262-7.

192. Liu T, Sun X. An association study between the insertion/deletion polymorphism of angiotensin I converting enzyme gene and human speed endurance. *Sheng Wu Yi Xue Gong Cheng Xue Za Zhi.* 2006;23(5):1045-7.

193. Berman Y, North KN. A gene for speed: the emerging role of alpha-actinin-3 in muscle metabolism. *Physiology (Bethesda).* 2010;25(4):250-9.

194. Moran CN, Yang N, Bailey MES, et al. Association analysis of the ACTN3 R577X polymorphism and complex quantative body composition and performance phenotypes in adolescent Greeks. *Eur J Hum Genet.* 2007;15(1):88-93.

195. Niemi A, Majamaa K. Mitochondrial DNA and ACTN3 genotypes in Finnish elite endurance and sprint athletes. *Eur J Hum Genet.* 2005;13(8):965-9.

196. Yang N, MacArthur DG, Wolde B, et al. The ACTN3 R577X polymorphism in East and West African athletes. *Med Sci Sports Exerc.* 2007;39(11):1985-8.

197. Shang X, Huang C, Chang Q, et al. Association between the ACTN3 R577X polymorphism and female endurance athletes in China. *Int J Sports Med.* 2010;31(12):913-6.

198. Liou SH, Wu TN, Chiang HC, et al. Blood lead levels in the general population of Taiwan, Republic of China. *Int Arch Occup Environ Health.* 1994;66(4):255-60.

199. Chiu LL, Wu YF, Tang MT, et al. ACTN3 genotype and swimming performance in Taiwan. *Int J Sports Med.* 2011;32(6):476-80.

200. Hsieh LL, Liou SH, Chen YH, et al. Association between aminolevulinate dehydrogenase genotype and blood lead levels in Taiwan. *J Occup Environ Med.* 2000;42(2):151-5.

201. Tanaka C, Kamide K, Takiuchi S, et al. An alternative fast and convenient genotyping method for the screening of angiotensin converting enzyme gene polymorphisms. *Hypertens Res.* 2003;26(4):301-6.

202. Glenn KL, Du ZQ, Eisenmann JC, et al. An alternative method for genotyping of the ACE I/D polymorphism. *Mol Biol Rep.* 2009;36(6):1305-10.

203. Alvarez R, Reguero JR, Batalla A, et al. Angiotensin-converting enzyme and angiotensin II receptor 1 polymorphisms: association with early coronary disease. *Cardiovasc Res.* 1998;40(2):375-9.

204. Eleni S, Dimitrios K, Vaya P, et al. Angiotensin-I converting enzyme gene and I/D polymorphism distribution in the Greek population and a comparison with other European populations. *J Genet* 2008;87(1):91-3.

205. Nazarov IB, Woods DR, Montgomery HE, et al. The angiotensin converting enzyme I/D polymorphism in Russian athletes. *Eur J Hum Genet.* 2001;9(10):797-801.

206. van Bockxmeer FM, Mamotte CD, Burke V, et al. Angiotensin-converting enzyme gene polymorphism and premature coronary heart disease. *Clin Sci (Lond).* 2000;99(3):247-51.

207. U.S. Genome Variation Estimates. ACE Allele and Genotype Frequencies. http://www.cdc.gov/genomics/population/file/print/genvar/ACE.pdf. Assessed 30 April 2013.

208. Barley J, Blackwood A, Carter ND, et al. Angiotensin converting enzyme insertion/deletion polymorphism: association with ethnic origin. *J Hypertens.* 1994;12(8):955-7.

209. Sagnella GA, Rothwell MJ, Onipinla AK, et al. A population study of ethnic variations in the angiotensin-converting enzyme I/D polymorphism: relationships with gender, hypertension and impaired glucose metabolism. *J Hypertens.* 1999;17(5):657-64.

210. Lee EJ. Population genetics of the angiotensin-converting enzyme in Chinese. *Br J Clin Pharmacol.* 1994;37:212-4.

211. Ishigami T, Iwamoto T, Tamura K, et al. Angiotensin I converting enzyme (ACE) gene polymorphism and essential hypertension in Japan. Ethnic difference of ACE genotype. *Am J Hypertens.* 1995;8(1):95-7.

212. Wang K, Dickson SP, Stolle CA, et al. Interpretation of association signals and identification of causal variants from genome-wide association studies. *Am J Hum Genet.* 2010;86(5):730-42.

213. Fu J, Festen EA, Wijmenga C. Multi-ethnic studies in complex traits. *Hum Mol Genet.* 2011 20(R2):R206-13.

214. Scott RA, Moran C, Wilson RH, et al. No association between Angiotnesin Converting Enzyme (ACE) gene variation and endurance athlete status in Kenyans. *Comp Biochem Physiol A Mol Integr Physiol.* 2005;141(2):169-75.

215. Ash GI, Scott RA, Deason M, et al. No association between ACE gene variation and endurance athlete status in Ethiopians. *Med Sci Sports Exerc.* 2011;43(4):590-7.

216. McKenzie CA, Abecasis GR, Keavney B, et al. Trans-ethnic fine mapping of a quantitative trait locus for circulating angiotensin I-converting enzyme (ACE). *Hum Mol Genet.* 2001;10(10):1077-84.

217. MacArthur DG, North KN. ACTN3: A genetic influence on muscle function and athletic performance. *Exerc Sport Sci Rev.* 2007;35(1):30-4

218. Demura S, Aoki H, Yamamato Y, et al. Comparison of strength values and laterality in various muscle contractions between competitive swimmers and untrained persons. *Health.* 2010;2:1249-54.

219. Alfred T, Ben-Shlomo Y, Cooper R, et al. ACTN3 genotype, athletic status, and life course physical capability: meta-analysis of the published literature and findings from nine studies. *Hum Mutat.* 2011;32(9):1008-18.

220. Bompa TO, Carrera MC. *Periodization Training for Sports*. 2nd ed: Human Kinetics; 2005.

221. Pereira TV, Patsopoulos NA, Salanti G, et al. Discovery properties of genome-wide association signals from cumulatively combined data sets. *Am J Epidemiol.* 2009;170(10):1197-206.

222. Gögele M, Minelli C, Thakkinstian A, et al. Methods for meta-analyses of genome-wide association studies: critical assessment of empirical evidence. *Am J Epidemiol.* 2012;175(8):739-49.

223. Berois N, Blanc E, Ripoche H, et al. ppGalNAc-T13: a new molecular marker of bone marrow involvement in neuroblastoma. *Clin Chem.* 2006;52(9):1701-12.

224. Müller FU, Bokník P, Knapp J, et al. Identification and expression of a novel isoform of cAMP response element modulator in the human heart. *FASEB J.* 1998;12(12):1191-9.

225. McCarthy MI, Zeggini E. Genome-wide association studies in type 2 diabetes. *Curr Diab Rep.* 2009;9(2):164-71.

226. Massey D, Parkes M. Genome-wide association scanning highlights two autophagy genes, ATG16L1 and IRGM, as being significantly associated with Crohn's disease. *Autophagy.* 2007;3:649-51.

227. Frayling TM, Timpson NJ, Weedon MN, 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.

228. Preece MA. The genetic contribution to stature. *Horm Res.* 1996;45 (Suppl 2):56-8.

229. Silventoinen K, Kaprio J, Lahelma E, et al. Relative effect of genetic and environmental factors on body height: differences across birth cohorts among Finnish men and women. *Am J Public Health.* 2000;90(4):627-30.

230. Macgregor S, Cornes BK, Martin NG, et al. Bias, precision and heritability of self-reported and clinically measured height in Australian twins. *Hum Genet.* 2006;120(4):571-80.

231. Perola M, Sammalisto S, Hiekkalinna T, et al. Combined genome scans for body stature in 6,602 European twins: evidence for common Caucasian loci. *PLoS Genet.* 2007;3(6):e97.

232. Lango Allen H, Estrada K, Guillaume L, et al Hundreds of variants clustered in genomic loci and biological pathways affect human height. *Nature.* 2010;467:832-8.

233. Yang J, Benyamin B, McEvoy BP, et al. Common SNPs explain a large proportion of the heritability for human height. *Nat Genet.* 2010;42(7):565-9.

234. Talmud PJ, Hingorani AD, Cooper JA, et al. Utility of genetic and non-genetic risk factors in prediction of type 2 diabetes: Whitehall II prospective cohort study. *BMJ.* 2010;340:b4838.

235. Gibson G. Hints of hidden heritablity in GWAS. *Nat Genet.* 2010;42(7):558-60.

236. Donnelly P. Progress and challenges in genome-wide association studies in humans. *Nature.* 2008;456(7223):728-31.

237. WTCCC. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature.* 2007;447(7145):661-78.

238. Plomin R, Haworth CM, Davis OS. Common disorders are quantitative traits. *Nat Rev Genet.* 2009;10(12):872-8.

239. Williams AG, Folland JP. Similarity of polygenic profiles for elite human physical performance. *J Physiol.* 2008;586:113-21.

240. Van Damme R, Wilson RS, Vanhooydonck B, et al. Performance constraints in decathletes. *Nature.* 2002;415(6873):755-6.

241. Ruiz JR, Arteta D, Buxens A, et al. Can we identify a power-oriented polygenic profile? *J Appl Physiol.* 2010;108(3):561-6.

242. Ruiz JR, Gómez-Gallego F, Santiago C, et al. Is there an optimum endurance polygenic profile? *J Physiol.* 2009;587(P17):1527-34.

243. Pitsiladis Y, Wang G, Wolfarth B, et al. Genomics of elite sporting performance: what little we know and necessary advances. *Br J Sports Med.* 2013;47(9):550-5.

244. Tsianos G, Sanders J, Dhamrait S, et al. The ACE gene insertion/deletion polymorphism and elite endurance swimming. *Eur J Appl Physiol.* 2004;92:360-2.

245. Dickson SP, Wang K, Krantz I, et al. Rare variants create synthetic genome-wide associations. *PLoS Biol.* 2008;8(1):e1000294.

246. Evans DM, Visscher PM, Wray NR. Harnessing the information contained within genome-wide association studies to improve individual prediction of complex disease risk. *Hum Mol Genet.* 2009;18(18):3525-31.

247. Bammann K, Peplies J, Sjöström M, et al. Assessment of diet, physical activity and biological, social and environmental factors in a multi-centre European project on diet- and lifestyle-related disorders in children (IDEFICS). *J Public Health.* 2006;14(5):279-89.

248. Koni AC, Scott RA, Wang G, et al. DNA yield and quality of saliva samples and suitability for large-scale epidemiological studies in children. *Int J Obes (Lond).* 2011;35 Suppl 1:S113-8.

249. McMichael GL, Gibson CS, O'Callaghan ME, et al. DNA from buccal swabs suitable for high-throughput SNP multiplex analysis. *J Biomol Tech.* 2009;20(5):232-5.

250. Dixon AL, Liang L, Moffatt MF, et al. A genome-wide association study of global gene expression. *Nat Genet.* 2007;39(10):1202-7.

251. Göring HH, Curran JE, Johnson MP, et al. Discovery of expression QTLs using large-scale transcriptional profiling in human lymphocytes. *Nat Genet.* 2007;39(10):1208-16.

252. Stranger BE, Forrest MS, Dunning M, et al. Relative impact of nucleotide and copy number variation on gene expression phenotypes. *Science.* 2007;315(5813):848-53.

253. Stranger BE, Nica AC, Forrest MS, et al. Population genomics of human gene expression. *Nat Genet.* 2007;39(10):1217-24.

254. Dimas AS, Deutsch S, Stranger BE, et al. Common regulatory variation impacts gene expression in a cell type-dependent manner. *Science.* 2009;325(5945):1246-50.

255. Emilsson V, Thorleifsson G, Zhang B, et al. Genetics of gene expression and its effect on disease. *Nature.* 2008;452(7186):423-8.

256. Nica AC, Montgomery SB, Dimas AS, et al. Candidate causal regulatory effects by integration of expression QTLs with complex trait genetic associations. *PLoS Genet.* 2010;6(4):e1000895.

257. Nicolae DL, Gamazon E, Zhang W, et al. Trait-associated SNPs are more likely to be eQTLs: annotation to enhance discovery from GWAS. *PLoS Genet.* 2010;6(4):e1000888.

258. Shimada MK, Matsumoto R, Hayakawa Y, et al. VarySysDB: a human genetic polymorphism database based on all H-InvDB transcripts. *Nucleic Acids Res.* 2009;37(Database issue):D810-5.

259. Yang TP, Beazley C, Montgomery SB, et al. Genevar: a database and Java application for the analysis and visualization of SNP-gene associations in eQTL studies. *Bioinformatics.* 2010;26(19):2474-6.

260. Suzuki MM, Bird A. DNA methylation landscapes: provocative insights from epigenomics. *Nat Rev Genet.* 2008;9(6):465-76.

261. Richardson K, Lai CQ, Parnell LD, et al. A genome-wide survey for SNPs altering microRNA seed sites identifies functional candidates in GWAS. *BMC Genomics.* 2011;12:504.

262. Cordell HJ. Detecting gene-gene interactions that underlie human diseases. *Nat Rev Genet.* 2009;10(6):392-404.

263. Barrett JC, Hansoul S, Nicolae DL, et al. Genome-wide association defines more than 30 distinct susceptibility loci for Crohn's disease. *Nat Genet.* 2008;40(8):955-62.

264. Waters KM, Le Marchand L, Kolonel LN, et al. Generalizability of associations from prostate cancer genome-wide association studies in multiple populations. *Cancer Epidemiol Biomarkers Prev.* 2009;18(4):1285-9.

265. Teslovich TM, Musunuru K, Smith AV, et al. Biological, clinical and population relevance of 95 loci for blood lipids. *Nature.* 466;466(7307):707-13.

266. Kochi Y, Suzuki A, Yamada R, et al. Genetics of rheumatoid arthritis: underlying evidence of ethnic differences. *J Autoimmun.* 2009;32(3-4):158-62.

267. Pulit SL, Voight BF, de Bakker PI. Multiethnic genetic association studies improve power for locus discovery. *PLoS One.* 2010;5(9):e12600.

268. Hernandez J, Cooper J, Babel N, et al. TNFalpha gene delivery therapy for solid tumors. *Expert Opin Biol Ther.* 2010;10(6):993-9.

269. López-Lázaro M. A new view of carcinogenesis and an alternative approach to cancer therapy. *Mol Med.* 2010;16(3-4):144-53.

270. Muntoni F, Wells D. Genetic treatments in muscular dystrophies. *Curr Opin Neurol.* 2007;20(5):590-4.

271. Bushby K, Lochmüller H, Lynn S, et al. Interventions for muscular dystrophy: molecular medicines entering the clinic. *Lancet.* 2009;374(9704):1849-56.

272. Kulkarni M. Gene therapy for human severe combined immunodeficiency disease. http://www.buzzle.com/articles/gene-therapy-for-human-severe-combined-immunodeficiency-scid-disease.html. Accessed 30 April 2013.

273. Gene therapy for Parkinson's disease is safe and some patients benefit, according to study. http://www.sciencedaily.com/releases/2007/06/070622101037.htm. Assessed 30 April 2013.