Medicine & Science IN Sports & Exercise
The Official Journal of the American College of Sports Medicine
www.acsm-msse.org

. . . Published ahead of Print

# Genome-wide Association Study Identifies New Genetic Determinants of Cardiorespiratory Fitness: The HUNT Study

Marie Klevjer[1,2], Ada N Nordeidet[1,2], Ailin F Hansen[3], Erik Madssen[2], Ulrik Wisløff[1,4], Ben M Brumpton[3], and Anja Bye[1,2]

[1]Cardiac Exercise Research Group (CERG), Department of Circulation and Medical Imaging, Faculty of Medicine and Health Sciences, Norwegian University of Science and Technology (NTNU), Trondheim, NORWAY; [2]Department of Cardiology, St. Olavs Hospital, Trondheim University Hospital, Trondheim, NORWAY; [3]K.G. Jebsen Center for Genetic Epidemiology, Department of Public Health and Nursing, Faculty of Medicine and Health Sciences, Norwegian University of Science and Technology (NTNU), Trondheim, NORWAY; [4]Centre for Research on Exercise, Physical Activity and Health, School of Human Movement and Nutrition Sciences, University of Queensland, St. Lucia, Brisbane, Queensland, AUSTRALIA

Accepted for Publication: 4 April 2022

Medicine & Science in Sports & Exercise, Publish Ahead of Print
DOI: 10.1249/MSS.0000000000002951

# Genome-wide Association Study Identifies New Genetic Determinants of Cardiorespiratory Fitness: The HUNT Study

Marie Klevjer[1,2], Ada N. Nordeidet[1,2], Ailin F. Hansen[3], Erik Madssen[2], Ulrik Wisløff[1,4], Ben M. Brumpton[3], and Anja Bye[1,2]

[1]Cardiac Exercise Research Group (CERG), Department of Circulation and Medical Imaging, Faculty of Medicine and Health Sciences, Norwegian University of Science and Technology (NTNU), Trondheim, NORWAY; [2]Department of Cardiology, St. Olavs Hospital, Trondheim University Hospital, Trondheim, NORWAY; [3]K.G. Jebsen Center for Genetic Epidemiology, Department of Public Health and Nursing, Faculty of Medicine and Health Sciences, Norwegian University of Science and Technology (NTNU), Trondheim, NORWAY; [4]Centre for Research on Exercise, Physical Activity and Health, School of Human Movement and Nutrition Sciences, University of Queensland, St. Lucia, Brisbane, Queensland, AUSTRALIA

**Address for Correspondence:** Marie Klevjer, Ph.D., Department of Circulation and Medical Imaging, Norwegian University of Science and Technology, Prinsesse Ragnhilds gate 3, Trondheim, 7033, Norway; Telephone: +47 47364121; E-mail: marie.klevjer@ntnu.no.

**Conflict of Interest and Funding Source:**

This work was supported by Central Norway Regional Health Authority and Norwegian Health Association. No conflict of interest was reported by the authors. The results of the study are

presented clearly, honestly, and without fabrication, falsification, or inappropriate data manipulation. The results of this study do not constitute endorsement by the American College of Sports Medicine.

## ABSTRACT

**Purpose:** Low cardiorespiratory fitness (CRF) is a major risk factor for CVD and a stronger predictor of CVD morbidity and mortality than established risk factors. The genetic component of CRF, quantified as peak oxygen uptake ($VO_{2peak}$), is estimated to be ~60%. Unfortunately, current studies on genetic markers for CRF have been limited by small sample sizes and using estimated CRF. To overcome these limitations, we performed a large-scale systematic screening for genetic variants associated with $VO_{2peak}$. **Methods:** A genome-wide association study (GWAS) was performed with BOLT-LMM including directly measured $VO_{2peak}$ from 4,525 participants in the HUNT3 Fitness study and 14 million single-nucleotide polymorphisms (SNPs). For validation, similar analyses were performed in the United Kingdom Biobank (UKB), where CRF was assessed through a submaximal bicycle test, including ~60,000 participants and ~60 million SNPs. Functional mapping and annotation of the GWAS results was conducted using FUMA. **Results:** In HUNT, two genome-wide significant SNPs associated with $VO_{2peak}$ were identified in the total population, two in males, and 35 in females. Two SNPs in the female population showed nominally significant association in the UKB. One of the replicated SNPs is located in *PIK3R5*, shown to be of importance for cardiac function and CVD. Bioinformatic analyses of the total and male population revealed candidate SNPs in *PPP3CA*, previously associated with CRF. **Conclusions:** We identified 38 novel SNPs associated with $VO_{2peak}$ in HUNT. Two SNPs were nominally replicated in UKB. Several interesting genes emerged from the functional analyses, among them one previously reported to be associated with CVD and another with CRF.

**Key Words:** UKB, GWAS, CRF, SNP, $VO_{2max}$

## INTRODUCTION

Cardiovascular disease (CVD) is the leading cause of death worldwide (1). The World Health Organization estimates that nearly 18 million people die from CVD every year globally. Low cardiorespiratory fitness (CRF) quantified as low maximal oxygen uptake ($VO_{2max}$) is one of the most important risk factors for CVD and premature death, and suggested to be a stronger predictor of morbidity and mortality than established CVD risk factors like obesity, diabetes, smoking, and hypercholesterolemia (2). $VO_{2max}$ is also shown to a be strong and independent predictor of all-cause and diseases-specific mortality (3). Low $VO_{2max}$ is shown to increase risk of dementia, cancer, and other lifestyle-related diseases (4, 5). The research on $VO_{2max}$ as a risk factor is still in its infancy, hence, better understanding of how $VO_{2max}$ is related to different life-style related diseases has a great potential in the search for improved prevention strategies.

$VO_{2max}$ is a complex trait determined by both genetic and environmental factors. The HEalth, RIsk factors, exercise Training And GEnetics (HERITAGE) Family Study, estimated the genetic component of $VO_{2max}$ to account for ~60% of the trait variability (6). Directly measured $VO_{2max}$ is an objective measure and considered the gold standard for measuring CRF. Directly measured $VO_{2max}$ is quantified as the maximal amount of oxygen the body can utilize during dynamic work with large muscle mass. As measuring $VO_{2max}$ is time-consuming and costly, most of the previous studies are based on estimated CRF. Only two previous studies report findings with genome-wide significant *p*-value ($5\times10^{-8}$) on CRF, but neither used directly measured $VO_{2max}$ (one using an endurance run test and the other using estimated CRF from a submaximal bicycle test (7, 8)).

Many of the previous studies on the genetic contribution to $VO_{2max}$ have been candidate gene studies based on a hypothesis-driven approach including the search for associations in pre-defined genomic regions excluding discovery of novel genetic loci (9). To our knowledge, only four previous genome-wide association studies (GWAS) have been conducted using directly measured $VO_{2max}$, and these studies were limited by few participants and few single-nucleotide polymorphisms (SNPs), (473 participants and 324,611 SNPs (10), 80 participants and 1,140,419 SNPs (11), 3470 participants and 123,545 SNPs (12), and 79 participants and 2,391,739 SNPs (13)). Because of the limited power in these previous studies, they have failed to reach the genome-wide significance, reporting *p*-values ~$5\times10^{-4}$ - $5\times10^{-5}$. To overcome these limitations, we performed a large-scale systematic screening for genetic variants associated with $VO_{2max}$ in a population which serves as the European reference material on CRF (14). The sample size in this study is ~10-57 times larger than previous GWAS on directly measured $VO_{2max}$, except for the study by Bye et al., that uses a subset of the same study population (12). In addition, the number of SNPs resulting from the GWAS in our study is ~6-114 times larger than the previous GWAS on directly measured $VO_{2max}$.

Women are often underrepresented or lacking in studies on CRF (15, 16). As there are large differences in the physiology of men and women, it may be plausible that the genetic variants can influence $VO_{2max}$ differently between sexes. In the research on cardiovascular diseases, major bias has been introduced by focusing on male subjects (17). Alarmingly, although improved understanding and the development of prevention of coronary artery diseases has led to more than 60% reduction in deaths from heart attack and stroke, the lowest rate of improvement has been in younger women. In addition, genotype by sex interactions are thought to account for some of the

sex differences in complex traits and in the risk of a wide array of diseases (18). Bernabeu et. al studied genotype by sex among more than 450 000 participants and 446 binary and 84 quantitative traits from UK Biobank (UKB). Surprisingly, they found that ~50% of the binary and 7% of the quantitative traits exhibited significant genetic heritability differences between the sexes, and that few of the identified differences in heritability were due to environmental variance, supporting the role for genetic variation in the observed sex differences in heritability.

The aim of our study is to identify genetic markers associated with $VO_{2max}$. This might provide awaited insight into this complex trait, that in the future could contribute to the discovery of possible shared genetic background between $VO_{2max}$ and life-style related diseases like CVD. By focusing on sex, we hope to learn more of the sex-specific biology that in the future could lead to more precise risk prediction, especially for women.

## MATERIAL AND METHODS

*Study participants*

The Trøndelag Health Study (HUNT) is one of the largest health studies ever performed (19). The HUNT study has collected questionnaire data, clinical measurements, and biological samples from more than 125,000 participants through four sub-studies throughout 35 years. DNA extracted from blood samples have been subject to large-scale genotyping and imputation. The genetic data from the HUNT study consists of 25 million SNPs from ~70,000 participants (20). The HUNT3 Fitness Study, involving 4,656 participants that completed a directly measured cardiopulmonary oxygen test, serves as the largest European reference material on CRF in a healthy adult population (14). Participants in the HUNT3 Fitness Study were healthy adults (20-90 years) free from

cardiovascular disease, respiratory symptoms, cancer and the use of blood pressure medication. The present study involves 4,525 participants from HUNT3 Fitness Study with available genetic data, with 2,218 males and 2,307 females in the sex-specific analyses.

*$VO_{2max}$*

$VO_{2max}$ has been measured using an individualized protocol by trained personnel (14). All participants used a heart rate monitor and a tight facemask, which performed continuous measurement of exhaled gases by a mixing chamber gas analyser (MetaMax II; Cortex Biophysik Gmbh, Leipzig, Germany). The participants were familiarized with the treadmill and performed a 10-minute warm-up. This was followed by a gradual increase in pace and incline until the participant reached a plateau in oxygen consumption despite increased workload, and the respiratory exchange ratio reached above 1.05, which is defined as true $VO_{2max}$. 12.6% of the subjects did not reach a plateau in oxygen consumption, hence we will use the term $VO_{2peak}$ in the following. $VO_{2peak}$ was measured as liters of oxygen per minute (L/min). This value was used to calculate $VO_{2peak}$ relative to body weight (mL/kg/min) and $VO_{2peak}$ scaled by body weight ($mL/kg^{0.75}$/min). $VO_{2peak}$ relative to body weight penalizes heavy individuals too strongly, hence $VO_{2peak}$ scaled by body weight is preferred (21). Consequently, the $VO_{2peak}$ scaled by body weight was used in the GWAS.

*Questionnaire-based information*

Physical activity (PA) is one of the behavioral factors that influences $VO_{2peak}$ the most, and to isolate the genetic contribution to $VO_{2peak}$, analyses were adjusted for PA. A self-reported questionnaire assessed information about the PA of the participants. Based on three questions

regarding PA, the points per answer (numbers in brackets) were multiplied to make a PA index score (Kurtze score) (22). *Question 1*: "How frequently do you exercise?", with the response options: "Never" (0), "Less than once a week" (0), "Once a week" (1), "2-3 times per week" (2.5), and "Almost every day" (5). *Question 2*: "If you exercise as frequently as once or more times a week: How hard do you push yourself?", with response options: "I take it easy without breaking a sweat or losing my breath" (1), "I push myself so hard that I lose my breath and break into sweat" (2), and "I push myself to near exhaustion" (3). *Question 3*: "How long do each session last?", with the response options: "Less than 15 minutes" (0.1), "16-30 minutes" (0.38), "30 minutes to 1 hour" (0.75) and "More than 1 hour" (1.0).

*Blood samples and genotyping*

DNA from blood samples of 71,860 participants were extracted for genotyping. The genotyping was performed with one of three different Illumina HumanCoreExome arrays (HumanCoreExome12 version 1.0, HumanCoreExome12 version 1.1, and UM HUNT Biobank version 1.0) according to standard protocols. Sample and quality control were performed by standard protocols and is described in detail elsewhere (23). In short, the quality control excluded samples with a call-rate < 99% or that showed deviation from the Hardy-Weinberg equilibrium in unrelated samples of European ancestry ($p$-value < 0.0001). Imputation was performed on the remaining 69,716 samples of European ancestry using Minmac3 (v2.0.1, http://genome.sph.umich.edu/wiki/Minimac3). Imputed variants with $R^2$ < 0.3 were excluded, resulting in more than 24.9 million well-imputed variants.

*Validation in UKB*

The UKB study is a prospective study of more than 500,000 participants aged 40 to 69 years, enrolled between 2006 to 2010 in 22 different assessment centers across the United Kingdom (https://www.ukbiobank.ac.uk/). As the HUNT Study, the UKB has collected questionnaire data, clinical measures, and biological samples. All 500,000 participants have been genotyped and imputed providing information on more than 90 million SNPs. A submaximal stationary bicycle test was used to assess CRF in ~100.000 participants, the protocol is described elsewhere (24). In short, the participants were divided into 5 risk categories which were used for their individualized bicycle protocol. The participants in the "minimal risk"-category 1, "small risk"-category 2 and "medium risk"-category 3 were tested at 50%, 30% and at a constant level of the predicted maximum workload, respectively. In the "high risk"-category 4 measurements were made at rest, and subjects in category 5 were excluded as electrocardiography was to be avoided as either unsafe or pointless. In the current study, only participants that performed a bicycle test (category 1, 2 and 3) were included. The HUNT 3 Fitness Study consist of relatively healthy individuals, and for making the analyses more comparable, the categories 4 and 5 in the UKB were excluded. The CRF was calculated as the net oxygen consumption from the following equation: oxygen consumption (mL/kg per minute) = 7 + 10.8 × workload (W)/body mass (kg), and then expressed in terms of maximal metabolic equivalents (METs), where 1 MET = 3.5 mL/kg per minute. Covariates from the UKB include age (field ID 21002), sex (ID 31), PA as summed MET min/week for all activity (ID 22040), genotyping batch (ID 22000) and 10 first genetic principal components (ID 22009).

*Statistical analyses*

BOLT linear mixed model (BOLT-LMM v2.3.4) was used to estimate the association between $VO_{2peak}$ and the 24.9 million SNPs in the HUNT population (25). Linear mixed models account for population stratification and relatedness, increasing the statistical power as related samples do not need to be excluded as compared to other methods for association testing in GWAS (26). The covariates included age (years), sex, and PA index (Kurtze score). In addition, the first four genetic principal components (PC) of ancestry and the genotype measurement batch were used as genetic and technical covariates. GWAS analyses adjusting for age, sex, PC, and genotyping batch were conducted, as PA is a self-reported variable which often is reported higher than device-measured PA (27). In addition, sex-specific analyses were performed. The minor allele count (MAC) was set to a minimum of 10 for the total population and a minimum of 5 for the sex specific analyses, yielding results for >14 million SNPs in each population. Genetic loci with a genome-wide significant $p$-value $< 5\times10^{-8}$ were regarded as significant findings. The genome-wide significant loci identified in the HUNT cohort, were further examined in the replication cohort, UKB, using the same statistical methods. The $p$-value for the replication cohort was set to 0.05. To address that the $p$-value in the replication cohort was not Bonferroni corrected, the effect sizes and directions were examined. The BOLT-REML algorithm for variance components analysis was used to access the SNP heritability for $VO_{2peak}$.

*Gene identification and their functional role in $VO_{2peak}$*

Further validation of the GWAS results were performed using functional annotation and gene mapping by SNP2GENE and GENE2FUNC functions in FUMA v1.3.7, a web-based platform integrating information from multiple biological resources (28). In this analysis, GWAS summary

statistics from the HUNT cohort were used to identify independent suggestive SNPs with $p < 5\times10^{-6}$ and $r^2 > 0.6$. Since the number of SNPs reaching genome-wide significance ($p < 5\times10^{-8}$) in the GWAS was low, especially for the total- and male population, we applied a less stringent *p*-value threshold to explore signals below genome-wide significance threshold. Each genomic risk loci comprised independent significant SNPs and SNPs in linkage disequilibrium (LD, $r^2 > 0.6$). The maximum distance between LD-blocks to merge into a genomic locus was 250 kb. For LD structure, a reference panel of 503 participants and 25,063,419 SNPs from a European population from 1000 Genome Project Phase 3 was used. Positional gene mapping was performed based on ANNOVAR annotations, and tissue specificity of mapped genes was assessed through an enrichment test for DEGs (Differentially Expressed Genes) as implemented in GENE2FUNC. The Combined Annotation Dependent Depletion (CADD) score was used to investigate whether a mutation might be deleterious. This score is predicted by 63 functional annotations, and it enables prioritization of functional, deleterious and pathogenic variants. In addition, MAGMA gene analysis and gene-set analysis, as well as gene-property analysis for tissue specificity were performed, the two latter with data from MSigDB and GTEx v8, respectively. MAGMA uses the full distribution of *p*-values from the GWAS summary statistics. Lastly, expression quantitative trait loci (eQTL) mapping was performed to map SNPs to a gene if the SNP had a significant effect on gene expression in tissues from heart, large arteries, and skeletal muscle, using eQTL data from GTEx. All database references used in FUMA are found in the Appendix.

*Ethical Approval*

This study was approved by the Regional committee for medical research ethics (2019/29771), the Trøndelag Health Study, the Norwegian Data Inspectorate, and by the National Directorate of

Health. The study is in conformity with Norwegian laws and the Helsinki declaration, and a written informed consent was obtained from all participants in the HUNT study. The UKB research protocol and study design were approved by the NHS National Research Ethics Service, and all study participants provided written informed consent (Project ID 40135). Ethical approval was obtained from the Northwest Centre for Research Ethics Committee (MREC, 11/NW/0382). In Scotland, the UKB has approval from the Community Health Index Advisory Group (CHIAG).

## RESULTS

### *GWAS of $VO_{2peak}$ without adjusting for PA*

To scan for associated genetic markers for $VO_{2peak}$, we used a total of 4,525 samples from the HUNT3 Fitness study (characteristics in Table 1) with genetic data available. The raw and scaled distribution of $VO_{2peak}$, are presented in Supplemental Figure 1 (see Supplemental Digital Content, Appendix, http://links.lww.com/MSS/C604) along with distributions of the scaled $VO_{2peak}$ separated by sex. The GWAS analyses adjusting for age, sex, PC1-4, and genotyping batch resulted in no significant SNPs for the total population and no significant SNPs for males. In females, 15 significant SNPs were genome-wide significant, and the results from this analysis are presented in Supplemental Table 1 (see Supplemental Digital Content, Appendix, Table 1, GWAS results of the analysis adjusted for age, PC1-4, and genotyping batch from the female population in HUNT, http://links.lww.com/MSS/C604).

### *GWAS of $VO_{2peak}$ adjusted for PA*

The GWAS adjusted for age, sex, PA, PC1-4, and genotyping batch identified a total of 36 novel loci associated with $VO_{2peak}$ (Table 2). Figure 1 illustrates the Manhattan plot for the total

population, males, and females, respectively. The analyses include more than 14 million SNPs after filtering. The genomic inflation factor, λ, was 1.047 in the total population, and 1.000 in both the male and the female population, showing that possible population stratification was adequately corrected with the used analysis model. The imputation quality is > 0.8 for most of the SNPs, and only seven SNPs have imputation quality below 0.5. Two SNPs were genome-wide significant in the total population, whereas two showed association in the male population and 33 in females, including 13 of the same SNPs identified in the analysis not adjusting for PA, and twenty new SNPs. Table 2 is a presentation of all genome-wide significant SNPs identified in the analyses adjusting for PA in addition to the two SNPs exclusively identified in the analysis not adjusting for PA. All the identified SNPs are rare variants with MAF <1%, except two SNPs identified in females, rs73176470 (MAF = 9.28%) and rs35073776 (MAF = 22.5%). The effect allele was associated with higher $VO_{2peak}$ for all SNPs, except from rs1467361640 in the male population, where the direction was negative.

The BOLT-REML algorithm estimates the proportion of variance in the $VO_{2peak}$ explained by the variance in the genotyped SNPs. The heritability was estimated to be 37.5% (SE 4.6%), 52.4% (SE 8.7%) and 26.0% (SE 8.8%) for the total, male and female population, respectively, for the analyses adjusted for PA. In the analyses not adjusting for PA the heritability was estimated to be 40.4% (SE 4.6%), 51.3% (SE 8.7%) and 30.8% (SE 8.8%) for the total, male and female population, respectively.

In the total population, rs75672239 ($p = 2.4\times10^{-8}$) and rs746795207 ($p = 4.0\times10^{-8}$) were significantly associated with $VO_{2peak}$. For rs75672239, heterozygote carriers (CG, n=10) had

38.5% (CI: 17.6 – 59.3) higher $VO_{2peak}$ than the homozygote carriers (GG, n=4515). rs75672239 is a downstream variant located near the gene chromodomain Y Like (*CDYL*), and this SNP was also significantly associated with $VO_{2peak}$ in the female-specific analysis. CDYL is a protein implicated in transcription repression and epigenetic regulation (29). For rs746795207, heterozygote carriers (TC, n=15) had 24.2% (CI: 4.3 – 44.2) higher $VO_{2peak}$ than the homozygote carriers (CC, n=4510). rs746795207 is an intergenic variant located near *LOC105371536,* an RNA gene which is affiliated with the non-coding RNA class (29).

Among females, we identified 35 genome-wide significant SNPs associated with $VO_{2peak}$ when combining the two analyses. rs73176470 was the only SNP with homozygote carriers also for the minor allele, and one of the two SNPs that were common variants (MAF = 9.28%). A boxplot of $VO_{2peak}$ across the alleles for rs73176470 is presented in Supplemental Figure 2 (see Supplemental Digital Content, Appendix, http://links.lww.com/MSS/C604). For rs73176470, heterozygote carriers (TC, n=369) had 4.4% (CI: 2.0 – 6.8) higher $VO_{2peak}$ than the TT carriers (n=1919), and the CC carriers (n=19) had 3.4% (CI: -5.3 – 12.2) and 7.8% (-0.7 – 16.4) higher $VO_{2peak}$ compared to heterozygote- and TT carriers, respectively. The most interesting SNPs based on SNP location was rs149814343, which is a missense variant in *TOE1* (target of *EGR1* (Early Growth Response 1)), and rs373113184 which is a 3'UTR, untranslated region, variant in *MAF1* (MAF1 Homolog, Negative Regulator Of RNA Polymerase III). A boxplot of $VO_{2peak}$ across the alleles for rs149814343 is presented in Supplemental Figure 2 (see Supplemental Digital Content, Appendix, http://links.lww.com/MSS/C604), where heterozygote carriers (GC, n=12) had 24.2% (CI: 2.7 - 45.7) higher $VO_{2peak}$ than homozygote carriers (GG, n=2295). Among the genome-wide significant findings, this was the only SNP directly genotyped.

In males, we identified two SNPs significantly associated with $VO_{2peak}$, rs549637059 which is intronic in Dynein Axonemal Heavy Chain 14 (*DNAH14*) and rs1467361640 which is intergenic near *LOC105375599*. Heterozygote carriers (CA, n=7) had 30.3% (CI: -3.7 – 64.3) lower $VO_{2peak}$ than homozygote carriers (AA, n=2211) for rs1467361640, the only SNP where the effect allele was associated with a decrease in $VO_{2peak}$. Heterozygote carriers (TC, n=5) had 42.5% (CI: 11.3 - 73.7) higher $VO_{2peak}$ than homozygote carriers (CC, n=2213) for rs549637059.

*Validation in UKB*

The GWAS on CRF in the UKB were adjusted for age, sex, PA, PC1-10, and genotyping batch. The GWAS were performed for the total population, males, and females, and after filtration the analyses includes 60,903,286 SNPs, 53,417,250 SNPs, and 55,045,706 SNPs, respectively. The two loci identified for the total HUNT-population were not replicated in the UKB. Neither were the two loci identified for the male population. In the female population, two loci were replicated to be nominally significant in the UKB, rs551942830 and rs376927175. rs551942830 ($p$ = 0.0381) is a 3'UTR variant in phosphoinositide-3-kinase regulatory subunit 5 (*PIK3R5*) (proxy for rs190675254 with LD = 1.0). *PIK3R5* encodes the regulatory subunit of one class of phosphoinositide-3-kinases (PI3Ks), a family of enzymes shown to be of importance in cardiac function and CVD. The direction of the effect was consistent between the exploration and the validation cohort. Further, the analysis that was adjusted for age, PC1-10, and genotyping batch, but not PA nominally validated rs376927175 ($p$ = 0.03) located near *APBA1* in the female population. The direction of the effect was consistent between the exploration and validation cohort.

*Exploration of the biological role of* $VO_{2peak}$ *-SNPs*

The GWAS summary statistics for PA-adjusted $VO_{2peak}$ in the total, male, and female population in HUNT were further explored using FUMA, to investigate potential functional roles of the candidate SNPs. The female population contains the highest number of identified SNPs and mapped genes, where 123 independent significant SNPs and 880 candidate SNPs were identified. Candidate SNPs are defined as all independent lead SNPs and SNPs in LD ($r^2 = 0.6$) in the FUMA platform. Further, 197 genes were mapped through positional mapping (Supplemental Table 2, see Supplemental Digital Content, Appendix, Summary of SNPs and mapped genes, from FUMA SNP2GENE analysis, http://links.lww.com/MSS/C604). The vast majority (~85%) of the candidate SNPs in females were annotated as either intronic or intergenic. Further, 13 SNPs were in coding regions (exonic) and 7 were UTR variants. The rest were either intronic in non-coding RNA (ncRNA), or downstream- or upstream variants.

The eQTL analysis in the female population revealed that multiple candidate SNPs on chromosomes 1 and 11 were eQTLs. 73 of the identified candidate SNPs were mapped to a region on chromosome 11 (11:100562544-100696230). The gene, Rho GTPase Activating Protein 42 (*ARHGAP42*), was mapped to this region, with two independent significant SNPs (Figure 2). eQTL analysis showed that multiple candidate SNPs in this region were statistically significant eQTLs (FDR < 0.05), affecting expression of *ARHGAP42* in skeletal muscle, blood vessels (aorta and tibial artery), and in the atrial appendage of the heart. *ARHGAP42* is a protein coding gene, and expression of this gene is enriched in vascular smooth muscle cells and is involved in regulation of vascular tone and control of blood pressure (29).

The MAGMA gene-based test mapped the input SNPs in GWAS summary statistics to protein coding genes, but none reached genome-wide significance in the female population. The MAGMA tissue expression analysis showed blood vessel-, muscle-, and heart tissue in top 6, but the findings were not statistically significant (Supplemental Figure 3, see Supplemental Digital Content, Appendix, MAGMA tissue expression, http://links.lww.com/MSS/C604). The MAGMA Gene-set Analysis yielded no significant gene sets, but one nominally significant ($p = 3\times10^{-4}$), that was the GO-term phosphatidylinositol (PI) dephosphorylation. PIs are involved in PI3K-Akt pathway, as PI3Ks phosphorylates lipid PIs, and the PI3K-Akt signaling pathway is associated with CVD.

To assess tissue specificity of the mapped genes, the enrichment test for DEGs implemented in GENE2FUNC was used. This test evaluates whether the mapped genes are overrepresented in differentially expressed gene sets in 54 specific tissue types. All 197 mapped genes were analyzed, and the results revealed significant differential gene expression in the cerebellum (Figure 3).

The MAGMA gene-based test assigned the GWAS SNPs to genes and in the total population 20 genes reached genome-wide significance (Figure 4). The most significant gene was CUB and sushi multiple domains 1 (*CSMD1*), a protein coding gene believed to function as a control protein in the complement system, which is part of the immune system. *CSMD1* is thought to play a role in cancer and schizophrenia, but its function is not established (30). The following MAGMA tissue expression analysis, based on input from the gene-based test, showed high expression profile in the blood, brain and lung for the total population, although not statistically

significant (Supplemental Figure 3, see Supplemental Digital Content, Appendix, MAGMA tissue expression, http://links.lww.com/MSS/C604). In males there was a high expression profile in heart, although not statistically significant (Supplementary Figure 3, see Supplemental Digital Content, Appendix, MAGMA tissue expression, http://links.lww.com/MSS/C604).

One genomic risk locus on chromosome 4 (4:102132472-102309597) was identified in both the total- and male population. 48 of the candidate SNPs in the total population were mapped to this region, and the *p*-value for the independent lead SNP was $2.2\times10^{-6}$ (Figure 5). In males, a total of 67 SNPs were mapped to the region (Figure 5), two of which were independent SNPs with *p*-value $3.3\times10^{-7}$. The gene Protein Phosphatase 3 Catalytic Subunit Alpha (*PPP3CA*) was mapped to this region. *PPP3CA* encodes the catalytic subunit α of protein phosphatase 3, that plays an important role in calcium signaling. *PPP3CA* has previously been found to be associated with human endurance (31).

Two of the candidate SNPs identified by SNP2GENE were identified as eQTLs affecting the expression of *AP001816.1* in the male and total population located in the presented risk loci on chromosome 4. Both eQTLs are intron variants in *PPP3CA*, where rs28660297 affects *AP001816.1* expression in heart, skeletal muscle, and artery tissue, and rs7698954 affects expression of *AP001816.1* in tibial artery tissue. A third eQTL, rs78824621, only identified as candidate SNP in the male population affects *AP001816.1* expression in left ventricle tissue. This eQTL is also an intron variant in *PPP3CA*. *AP001816.1*, also called *LOC90024*, is a less studied long non-coding RNA (lncRNA) gene located 250 bp downstream of *PPP3CA*.

## DISCUSSION

To our knowledge this is the first study to identify genome-wide significant genetic determinants of directly measured $VO_{2peak}$ within a large-scale healthy population. This study identified a total of 38 novel genetic determinants associated with $VO_{2peak}$ in the HUNT cohort. Two of these SNPs were nominally validated in the female analysis in UKB. Further, one of the validated SNPs is of functional interest as it is a 3'UTR variant located in *PIK3R5* previously reported to be related to heart function and CVD. In addition, the bioinformatic follow-up analyses in the total- and male population revealed candidate SNPs in a gene previously found to be associated with endurance, *PPP3CA.* All significant SNPs, but two, were rare variants which requires caution in interpretation of the strength of their association with $VO_{2peak}$. Fortunately, most SNPs had high imputation quality, making us more confident in the findings.

$VO_{2peak}$ is a strong cardiovascular health predictor, and genetic variants associated with $VO_{2peak}$ can point towards links between $VO_{2peak}$ and CVD. In the present GWAS, one of the identified genome-wide significant lead SNPs in females (rs190675254) is intergenic with Myosin Heavy Chain 10 (*MYH10*) as the nearest gene. *MYH10* was one of the mapped genes from FUMA analysis. *MYH10* is required for coronary vessel formation and epicardial function during development (32). Further, this SNP is in complete LD with three proxies located in *PIK3R5*, in a European reference sample (1000 Genome Project Phase 3, reference in Appendix). One of the proxies (rs551942830) was nominally validated in the UKB, thus strengthening its role as a candidate SNP. *PIK3R5* encodes PI3-kinase regulatory subunit 5 (*PIK3R5*, also known as p101), a subunit in Class IB phosphoinositide 3-kinases, named PI3Kγs. PI3Ks are widely recognized as important modulators in cardiac function and CVD (33, 34). One of the proxies (rs551942830)

resides in the 3 prime UTR region of *PIK3R5*, potentially affecting post-transcriptional gene expression. It has been suggested that PI3Kγ is involved in cardiac remodeling and contractility (35, 36). When studied in a mice model, PIK3R5 was suggested as a mediator of apoptosis and contractile dysfunction in cardiomyocytes (37). Further, multiple studies suggests that PI3Kγ can be a promising therapeutic target in different CVDs, like hypertension, heart failure and atherosclerosis. PI3Ks mediates intracellular signaling pathways in response to extracellular signals, and PI3Kγ is the only PI3K that plays and important role in G-protein dependent regulation of cell signaling in health and disease (38). The PI3Kγ catalytic subunit can associate with two possible regulatory subunits, either p101 or p87 (*PIK3R6*). p101 and p87 serves as adaptors assigning different functions to the PI3Kγ variants and the two isoforms are hence hypothesized to exhibit separate cellular functions. Inhibition of PI3Kγ has been suggested as a therapeutic measure in inflammatory disease and cancer, but it has been proposed that future studies must discriminate between the two PI3Kγ isoforms (p101 vs p87) to understand the biological role(s) of PI3Kγ. PI3Kγ is mostly studied in relation to the immune system and the physiological role of p101 is best characterized in neutrophils (39). Whether genetic variation in *PIK3R5*, including the candidate SNP rs551942830, influences $VO_{2peak}$ requires additional functional studies. It then remains to investigate how the regulatory subunit *PIK3R5* might play a role in the physiology of $VO_{2peak}$, and if it is a candidate link between $VO_{2peak}$ and CVD.

The second female SNP validated in UKB, rs376927175, is located near *APBA1* on chromosome 9. There was also another genome-wide significant SNP in the intron of *APBA1*, rs112987080, in the female population in HUNT. APBA1, also known as X11, is a neuronal adapter protein that interacts with the amyloid precursor protein (APP), best known as a precursor

to amyloid-beta in Alzheimer's disease. Further, the enrichment analyses on the 193 mapped genes showed significant DEGs in cerebellum (Figure 3), the brain region where beta amyloid plaque is deposited in Alzheimer's disease. As $VO_{2peak}$ is inversely associated with the risk of dementia (4, 40), it could be interesting to further investigate possibly shared genetics between $VO_{2peak}$ and dementia.

Another interesting gene identified in the female population using SNP2GENE in FUMA was *ARHGAP42,* which encodes Rho GTPase activating protein 42, a protein found to be participating in the regulation of vascular tone and control blood pressure (41). The expression of *ARHGAP42* is enriched in vascular smooth muscle cells, and our eQTL analyses revealed that among the candidate SNPs there were eQTLs affecting the expression of *ARHGAP42* in the heart, skeletal muscle, aorta, and tibial artery. It is suggested that expression of *ARHGAP42* might act as a negative feedback mechanism to limit excessive vessel constriction (41). The true functional consequences of SNPs located in or near *ARHGAP42* would need to be further assessed with additional functional studies.

There were two common variants (MAF > 5%) among the GWAS significant variants. One from the model adjusting for PA (rs73176470, MAF = 9.28%) and one from the model not adjusting for PA (rs35073776, MAF = 22.49%). The likelihood of false positives decreases with increased MAF, thus strengthening these findings. rs73176470 is an intronic variant in *LRRC31* (Leucine-rich repeat-containing protein 31). It has been shown that mutations in genes encoding LRR-containing proteins are associated with over sixty human diseases, including different cancers, neurological disorders, and dilated cardiomyopathy (42). rs35073776 is in an intron of

*IPCEF1* (interaction protein for cytohesin exchange factors 1), and it has six proxies in high LD (R2 > 0.95), all intronic variants in *IPCEF1*. It has previously been shown that IPCEF1 is involved in enhancing motility of epithelial cells through its involvement in Arf6 signaling events (43). An alternative name for IPCEF1 is Phosphoinositide-binding protein PIP3-E, and PI3K-AKT signaling is among its related pathways (29). The exact role of *LRRC32* and *IPCEF1* is not well established yet, and how these two SNPs might affect $VO_{2max}$ remains unknown. However, as these are genome-wide significant common variants with high imputation quality these findings could be interesting to further investigate in relation to CRF.

One of the genome-wide significant SNPs in the female population was the missense variant in *TOE1* (rs149814343). This SNP was genotyped, adding more confidence to the finding. The consequence of the missense mutation is valine to leucine amino acid change. The CADD-score for rs149814343 is 18.98, suggesting deleteriousness as it is above the proposed deleteriousness threshold (>12.37). *TOE1* encodes an unconventional deadenylase and was first identified as a target of *EGR1*. *EGR1* is a transcription factor that regulates transcription of several genes related to vascular diseases (44). *EGR1* serves a role as a key immediate-early gene in CVD, its expression is rapidly induced by pathologic stimuli, like hypoxia and pro-inflammatory cytokines (45).

Exploration of the GWAS summary statistics in the total and male population with the SNP2GENE and GENE2FUNC implemented in the FUMA platform revealed a genomic risk locus on chromosome 4 with a cluster of candidate SNPs located in *PPP3CA*, a gene previously found to be associated with endurance capacity in humans (31). This gene encodes CnAα, the

calmodulin-binding catalytic subunits for calcineurin, a protein phosphatase that plays an essential role in calcium signaling in the heart where it affects cardiac function and adaption (46). Calcineurin serves as a link between increased intracellular calcium concentration and the phosphorylation state of several target proteins with relevance to cardiovascular health. Inhibition of calcineurin in cardiomyocytes can reduce hypertrophic growth and protect the heart from oxidative damage and pathological remodeling (47). After the discovery that *PPP3CA* is associated with human endurance, it was later shown that gain-of-function of *PPP3CA* increased endurance performance in mice (48). The transgenic mice with overexpression of *PPP3CA* showed a 33% increase in endurance performance, as quantified by running distance on a treadmill, compared to wild-type mice (49). The observation that $VO_{2peak}$ GWAS SNPs were mapped to *PPP3CA*, a gene previously linked to endurance, gives confidence to the GWAS findings. In addition, since this gene is also established as important in cardiac function, it might be a candidate for further studies into the functional link between CRF and CVD.

There are large differences in the physiology of women and men, and it is plausible that genetic variants can influence $VO_{2peak}$ differently between the sexes. Most of the genome-wide significant SNPs were found in the analysis of the female HUNT population, and most were exclusively significant for women, which is an interesting finding in itself. In addition, the estimated heritability showed great difference in the proportion of variance of $VO_{2peak}$ explained by the genotyped SNPs in the male and female population, with values 52.4% and 26.0%, respectively. Future studies should investigate whether the identified genetic variants and genes might elucidate sex-specific factors in pathways and mechanisms linking $VO_{2peak}$ and CVD. This would be awaited insight as women often are underrepresented or lacking in studies on CRF (15,

16). Sex-specific studies of the genetics of $VO_{2peak}$ and how it relates to CVD (and other diseases) might be the key to more precise preventive and therapeutic strategies.

The significant SNPs from this GWAS are not previously reported. As the sample sizes in previous GWAS on $VO_{2peak}$ are extremely small, replication of these results was expected to be difficult. Also, previous GWAS report results with *p*-values ~$5\times10^{-4}$ - $5\times10^{-5}$, hence those results may have similar *p*-values in this GWAS, not reported here due to *p*-values below genome-wide significant threshold.

*Limitations*

The HUNT study consists of directly measured $VO_{2peak}$ on a treadmill, while the UKB have estimated CRF based on a submaximal bicycle test. Hence the phenotype is not identical between the exploration cohort and the validation cohort. However, in a preprint, the estimated CRF was correlated with Pearson's r range: 0.68-0.74 to directly measured $VO_{2max}$ in a sub-study of the UKB CRF (n=133) (50). Furthermore, the MAC was low for most of the identified SNPs, making the results more prone to false positive. Not lowering the MAC in the sex-specific analyses would have resulted in a MAF threshold of 0.0022, resulting in no genome-wide significant SNPs in males and only six SNPs in females. Fortunately, most of the identified SNPs had an imputation quality >0.8, making us more confident in the strength of our findings. There were seven SNPs with imputation quality below 0.5, which combined with low MAF could affect the reliability of these SNPs. Ideally, there should be another large database on directly measured $VO_{2peak}$ for validation of the HUNT results. In the FUMA analyses we used a more relaxed *p*-value threshold

of $5\times10^{-6}$, that increases the number of lead SNPs and enable a more extensive SNP-to-gene mapping, again with the cost of increasing the potential of false positive discoveries.

The HUNT cohort consist of participants with European ancestry only, while the UKB have ~95% Europeans. Ideally, the analyses in UKB should only have included participants with European ancestry. Excluding non-Europeans in the UKB analyses would have reduced the sample size and hence the power, but the power would at the same time increase as both cohorts involves participants with same ethnicity. The analyses in the UKB were adjusted for the 10 first PCs, that could capture some of the variability from the ~5% non-Europeans. Since the genetic variants in this GWAS are identified in a European population (HUNT cohort) the findings may not necessarily be valid for other ethnicities.

Further, one of the models includes PA as a covariate. In the HUNT study this covariate is calculated by the Kurtze score, while UKB uses the variable Summed MET minutes per week for all activity. The Kurtze score and the summed MET minutes per week for all activity are measures on PA, but not calculated the same way. Both are based on self-reported questionnaire data, making both exposed to the same recall- and social desirability bias. It is shown that self-reported PA is higher than device-measured PA, hence the PA variable included here might not reflect the true PA. However, PA is considered as one of the behavioral factors that influence the $VO_{2peak}$ the most, therefore it is believed necessary to include in the analysis, as we are interested in genetics related to inborn $VO_{2peak}$. To supplement the results, we have also included analyses without PA as covariate, and these analyses elude the potential bias from self-reported PA. As PA is heritable, adjusting the analysis for PA could introduce collider bias, if the SNPs identified to have a

significant association with $VO_{2peak}$ also is associated with PA. To our knowledge, the identified SNPs are not previously reported to be associated with PA, hence the analyses should not include collider bias. 13 of 15 SNPs identified in the female population in the model not adjusted for PA were also identified in the analysis adjusting for PA, making us confident that at least these SNPs are free from potential false positives due to collider bias.

This is the largest sample size of a GWAS performed on directly measured $VO_{2peak}$. However, the sample size is quite small in a GWAS context. This results in the limitations discussed above regarding the low MAF and the liberal *p*-value for the replication cohort. The limitations required careful interpretation of the implication of the statistically significant associations. These limitations are partially met with functional and biological analyses of the results. As a result, we are discussing the SNPs we think has the strongest potential of being true findings despite the limitations of the study design. This includes the two SNPs with high MAF, the two validated SNPs and the SNPs and genes identified in FUMA.

## CONCLUSIONS

This is the first large-scale GWAS on directly measured $VO_{2peak}$ in a population which serves as the European reference material on CRF. We have identified a total of 38 novel SNPs associated with $VO_{2peak}$ in the HUNT cohort. Two of these SNPs were nominally validated in females in UKB (rs551942830 and rs376927175). One of the validated SNPs, rs551942830, resides within a gene previously reported to be related to heart function and CVD. In addition, the bioinformatic analyses in the total- and male population revealed candidate SNPs in a gene previously found to be associated with endurance. The bioinformatic analyses pointed to several interesting genes and

biological pathways, that possibly could bring more insight to the physiology of the complex trait $VO_{2peak}$. Further analyses using bioinformatic approaches may provide more information on the physiological importance of these findings and their relation to CVD.

**Acknowledgements**

The genotyping was financed by the National Institute of health (NIH), University of Michigan, The Norwegian Research council, and Central Norway Regional Health Authority and the Faculty of Medicine and Health Sciences, Norwegian University of Science and Technology (NTNU). The genotype quality control and imputation has been conducted by the K.G. Jebsen center for genetic epidemiology, Department of public health and nursing, Faculty of medicine and health sciences, Norwegian University of Science and Technology (NTNU).

The HUNT Study is a collaboration between the HUNT Research Centre, Faculty of Medicine and Health Sciences, NTNU, Norwegian University of Science and Technology NTNU, the Nord-Trøndelag County Council, Central Norway Health Authority and the Norwegian Institute of Public Health. We are appreciative of the participants from the HUNT study, and the management of the study for using these data.

**Conflicts of Interest and Source of Funding**

This work was supported by Central Norway Regional Health Authority and Norwegian Health Association.

No conflict of interest was reported by the authors. The results of the study are presented clearly, honestly, and without fabrication, falsification, or inappropriate data manipulation. The results of this study do not constitute endorsement by the American College of Sports Medicine.

**Data availability**

Data generated or analyzed during this study are available from the corresponding authors upon reasonable request.

## REFERENCES


1. Organization WH. Cardiovascular diseases 2021. Available from: https://www.who.int/health-topics/cardiovascular-diseases/#tab=tab_1.
2. Nauman J, Nes BM, Lavie CJ, et al. Prediction of cardiovascular mortality by estimated cardiorespiratory fitness independent of traditional risk factors: The HUNT Study. *Mayo Clin Proc.* 2017;92(2):218-27.
3. Strasser B, Burtscher M. Survival of the fittest: VO(2)max, a key predictor of longevity? *Front Biosci (Landmark Ed).* 2018;23(8):1505-16.
4. Kurl S, Laukkanen JA, Lonnroos E, Remes AM, Soininen H. Cardiorespiratory fitness and risk of dementia: a prospective population-based cohort study. *Age Ageing.* 2018;47(4):611-4.
5. Marshall CH, Al-Mallah MH, Dardari Z, et al. Cardiorespiratory fitness and incident lung and colorectal cancer in men and women: Results from the Henry Ford Exercise Testing (FIT) cohort. *Cancer*. 2019;125(15):2594-601.
6. Schutte NM, Nederend I, Hudziak JJ, Bartels M, Geus EJCd. Twin-sibling study and meta-analysis on the heritability of maximal oxygen consumption. *Physiol Genomics.* 2016;48(3):210-9.
7. Zhao Y, Huang G, Chen Z, et al. Four loci are associated with cardiorespiratory fitness and endurance performance in young chinese females. *Sci Rep.* 2020;10(1):10117.
8. Hanscombe KB, Persyn E, Traylor M, et al. The genetic case for cardiorespiratory fitness as a clinical vital sign and the routine prescription of physical activity in healthcare. *Genome Med.* 2021;13(1):180.

9. Williams CJ, Williams MG, Eynon N, et al. Genes to predict VO(2max) trainability: a systematic review. *BMC Genomics.* 2017;18(Suppl 8):831.

10. 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 (1985).* 2011;110(5):1160-70.

11. Ahmetov I, Kulemin N, Popov D, et al. Genome-wide association study identifies three novel genetic markers associated with elite endurance performance. *Biol Sport.* 2015;32(1):3-9.

12. Bye A, Klevjer M, Ryeng E, et al. Identification of novel genetic variants associated with cardiorespiratory fitness. *Prog Cardiovasc Dis.* 2020;63(3):341-9.

13. Yoo J, Kim BH, Kim SH, Kim Y, Yim SV. Genetic polymorphisms to predict gains in maximal O2 uptake and knee peak torque after a high intensity training program in humans. Eur *J Appl Physiol.* 2016;116(5):947-57.

14. Loe H, Steinshamn S, Wisløff U. Cardio-respiratory reference data in 4631 healthy men and women 20-90 years: The HUNT 3 Fitness Study. *PLoS One.* 2014;9(11):e113884.

15. Khan H, Jaffar N, Rauramaa R, Kurl S, Savonen K, Laukkanen JA. Cardiorespiratory fitness and nonfatalcardiovascular events: A population-based follow-up study. *Am Heart J.* 2017;184:55-61.

16. Laukkanen JA, Lakka TA, Rauramaa R, et al. Cardiovascular fitness as a predictor of mortality in men. *Arch Intern Med.* 2001;161(6):825-31.

17. Barc J, Erdmann J. Sex matters? Sex matters! *Cardiovasc Res.* 2021;118(1):e1-e3.

18. Bernabeu E, Canela-Xandri O, Rawlik K, Talenti A, Prendergast J, Tenesa A. Sex differences in genetic architecture in the UK Biobank. *Nat Genet.* 2021;53(9):1283-9.

19. Krokstad S, Langhammer A, Hveem K, et al. Cohort profile: The HUNT Study, Norway. *Int J Epidemiol.* 2013;42(4):968-77.

20. Åsvold BO, Langhammer A, Rehn TA, et al. Cohort Profile Update: The HUNT Study, Norway. medRxiv [Internet]. 2021:[2021.10.12.21264858 p.]. Available from: https://www.medrxiv.org/content/medrxiv/early/2021/10/29/2021.10.12.21264858.full.pdf .

21. Lolli L, Batterham AM, Weston KL, Atkinson G. Size exponents for scaling maximal oxygen uptake in over 6500 humans: A systematic review and meta-analysis. *Sports Med.* 2017;47(7):1405-19.

22. Kurtze N, Rangul V, Hustvedt BE, Flanders WD. Reliability and validity of self-reported physical activity in the Nord-Trøndelag Health Study (HUNT 2). *Eur J Epidemiol.* 2007;22(6):379-87.

23. Ferreira MA, Vonk JM, Baurecht H, et al. Shared genetic origin of asthma, hay fever and eczema elucidates allergic disease biology. *Nat Genet.* 2017;49(12):1752-7.

24. Laukkanen JAMDP, Kunutsor SKMDP, Yates TP, et al. Prognostic relevance of cardiorespiratory fitness as assessed by submaximal exercise testing for all-cause mortality: A UK Biobank prospective study. *Mayo Clin Proc.* 2020;95(5):867-78.

25. Loh P-R. BOLT-LMM v2.3.4 User Manual. 2019.

26. Loh PR, Tucker G, Bulik-Sullivan BK, et al. Efficient Bayesian mixed-model analysis increases association power in large cohorts. *Nat Genet.* 2015;47(3):284-90.

27. Dyrstad SM, Hansen BH, Holme IM, Anderssen SA. Comparison of self-reported versus accelerometer-measured physical activity. *Med Sci Sports Exerc.* 2014;46(1):99-106.

28. Watanabe K, Taskesen E, van Bochoven A, Posthuma D. Functional mapping and annotation of genetic associations with FUMA. *Nat Commun.* 2017;8(1):1826.

29. Stelzer G, Rosen N, Plaschkes I, et al. The GeneCards Suite: From gene data mining to disease genome sequence analyses. *Curr Protoc Bioinformatics.* 2016;54(1):1.30.1-1.30.33.

30. Gialeli C, Gungor B, Blom AM. Novel potential inhibitors of complement system and their roles in complement regulation and beyond. *Molec Immunol.* 2018;102:73-83.

31. Ahmetov, II, Egorova ES, Gabdrakhmanova LJ, Fedotovskaya ON. Genes and athletic performance: An update. *Med Sport Sci.* 2016;61:41-54.

32. Ridge LA, Mitchell K, Al-Anbaki A, et al. Non-muscle myosin IIB (Myh10) is required for epicardial function and coronary vessel formation during mammalian development. *PLoS Genet.* 2017;13(10):e1007068.

33. Pretorius L, Owen KL, McMullen JR. Role of phosphoinositide 3-kinases in regulating cardiac function. *Front Biosci (Landmark Ed).* 2009;14(6):2221-9.

34. Durrant TN, Hers I. PI3K inhibitors in thrombosis and cardiovascular disease. *Clin Transl Med.* 2020;9(1):8.

35. Patrucco E, Notte A, Barberis L, et al. PI3Kgamma modulates the cardiac response to chronic pressure overload by distinct kinase-dependent and -independent effects. *Cell*. 2004;118(3):375-87.

36. Guo D, Thiyam G, Bodiga S, Kassiri Z, Oudit GY. Uncoupling between enhanced excitation-contraction coupling and the response to heart disease: lessons from the PI3Kγ knockout murine model. *J Mol Cell Cardiol.* 2011;50(4):606-12.

37. Brosinsky P, Bornbaum J, Warga B, et al. PI3K as mediator of apoptosis and contractile dysfunction in TGFβ(1)-stimulated cardiomyocytes. *Biology (Basel).* 2021;10(7):670.

38. Nürnberg B, Beer-Hammer S. Function, Regulation and biological roles of PI3Kγ variants. *Biomolecules*. 2019;9(9):427.

39. Rathinaswamy MK, Burke JE. Class I phosphoinositide 3-kinase (PI3K) regulatory subunits and their roles in signaling and disease. *Adv Biol Regul.* 2020;75:100657.

40. Tari AR, Nauman J, Zisko N, et al. Temporal changes in cardiorespiratory fitness and risk of dementia incidence and mortality: a population-based prospective cohort study. *Lancet Public Health.* 2019;4(11):e565-e74.

41. Bai X, Lenhart KC, Bird KE, et al. The smooth muscle-selective RhoGAP GRAF3 is a critical regulator of vascular tone and hypertension. *Nat Commun.* 2013;4(1):2910.

42. Matsushima N, Takatsuka S, Miyashita H, Kretsinger RH. Leucine rich repeat proteins: Sequences, mutations, structures and diseases. *Protein Pept Lett.* 2019;26(2):108-31.

43. Attar MA, Salem JC, Pursel HS, Santy LC. CNK3 and IPCEF1 produce a single protein that is required for HGF dependent Arf6 activation and migration. *Exp Cell Res.* 2012;318(3):228-37.

44. Youreva V, Srivastava AK. Early growth response protein-1 expression by insulin-like growth factor-1 requires ROS-dependent activation of ERK1/2 and PKB Pathways in vascular smooth muscle cells. *J Cell Biochem.* 2016;117(1):152-62.

45. Khachigian LM. Early growth response-1 in the pathogenesis of cardiovascular disease. *J Mol Med (Berl).* 2016;94(7):747-53.

46. Rusnak F, Mertz P. Calcineurin: form and function. *Physiol Rev.* 2000;80(4):1483-521.

47. Parra V, Rothermel BA. Calcineurin signaling in the heart: The importance of time and place. *J Mol Cell Cardiol.* 2017;103:121-36. 10.1016/j.yjmcc.2016.12.006. PubMed PMID: 28007541; PubMed Central PMCID: PMCPMC5778886.

48. Yaghoob Nezhad F, Verbrugge SAJ, Schönfelder M, Becker L, Hrabě de Angelis M, Wackerhage H. Genes whose gain or loss-of-function increases endurance performance in mice: A systematic literature review. *Front Physiol.* 2019;10:262.

49. Jiang LQ, Garcia-Roves PM, Barbosa TdC, Zierath JR. Constitutively active calcineurin in skeletal muscle increases endurance performance and mitochondrial respiratory capacity. *Am J Physiol Endocrinol Metab.* 2010;298(1):E8-E16.

50. Gonzales TI, Westgate K, Strain T, et al. The UK Biobank submaximal cycle ergometer test for assessment of cardiorespiratory fitness: Validity, reliability, and association with disease outcomes. medRxiv [Internet]. 2020:[2020.09.29.20203828 p.]. Available from: https://www.medrxiv.org/content/medrxiv/early/2020/09/29/2020.09.29.20203828.full.pdf

## FIGURE LEGENDS

**Figure 1:** Manhattan summary plot of the genome-wide association study (GWAS) of peak oxygen uptake ($VO_{2peak}$). Panel A is the total population, panel B is the male population, and panel C is the female population. The genomic position in each chromosome is on the x-axis and the negative logarithm of the associated *p*-value for each single nucleotide polymorphisms (SNPs) is at the y-axis. The red bar indicated the genome-wide significance level $p = 5\times10^{-8}$.

**Figure 2:** Regional plot for risk loci (11:100562544-100696230) on chromosome 11 for females. 73 candidate SNPs and one gene were mapped to this region. The two independent significant SNPs are represented by the purple dot and the red dot framed by black.

**Figure 3:** Tissue specificity of mapped genes for the female population. Red indicates significant DEGs in the tissue. Different tissues are expressed on the x-axis and the negative logarithm of the associated *p*-value for each tissue is on the y-axis. DEGs: Differentially Expressed Genes.

**Figure 4:** The Manhattan plot for the MAGMA gene-based test in the total population. The genomic position in each chromosome is on the x-axis and the negative logarithm of the associated *p*-value for each single nucleotide polymorphisms (SNPs) is at the y-axis. The red bar indicated the Bonferroni corrected significance $p = 2.574\times10^{-6}$, for testing 19428 protein coding genes.

**Figure 5:** Regional plot of genomic risk locus on chromosome 4 in the total population in panel A, and the male population in panel B. The two independent significant SNPs, including the lead SNP, are circled in black, the lead SNP is colored dark purple.

**SUPPLEMENTAL DIGITAL CONTENT**

**SDC 1:** Appendix.docx

**Appendix:** References of databases used in FUMA.

**Supplemental Figure 1:** The distribution of raw and scaled $VO_{2peak}$ values.

**Supplemental Table 1:** GWAS results of the analysis that was adjusted for age, PC1-4, and genotyping batch from the female population in HUNT.

**Supplemental Figure 2:** Boxplots of $VO_{2peak}$ by alleles for rs73176470 and rs149814343.

**Supplemental Table 2:** Summary of SNPs and mapped genes, from FUMA SNP2GENE analysis.

**Supplemental Figure 3:** Magma tissue expression analysis.

**Figure 1**

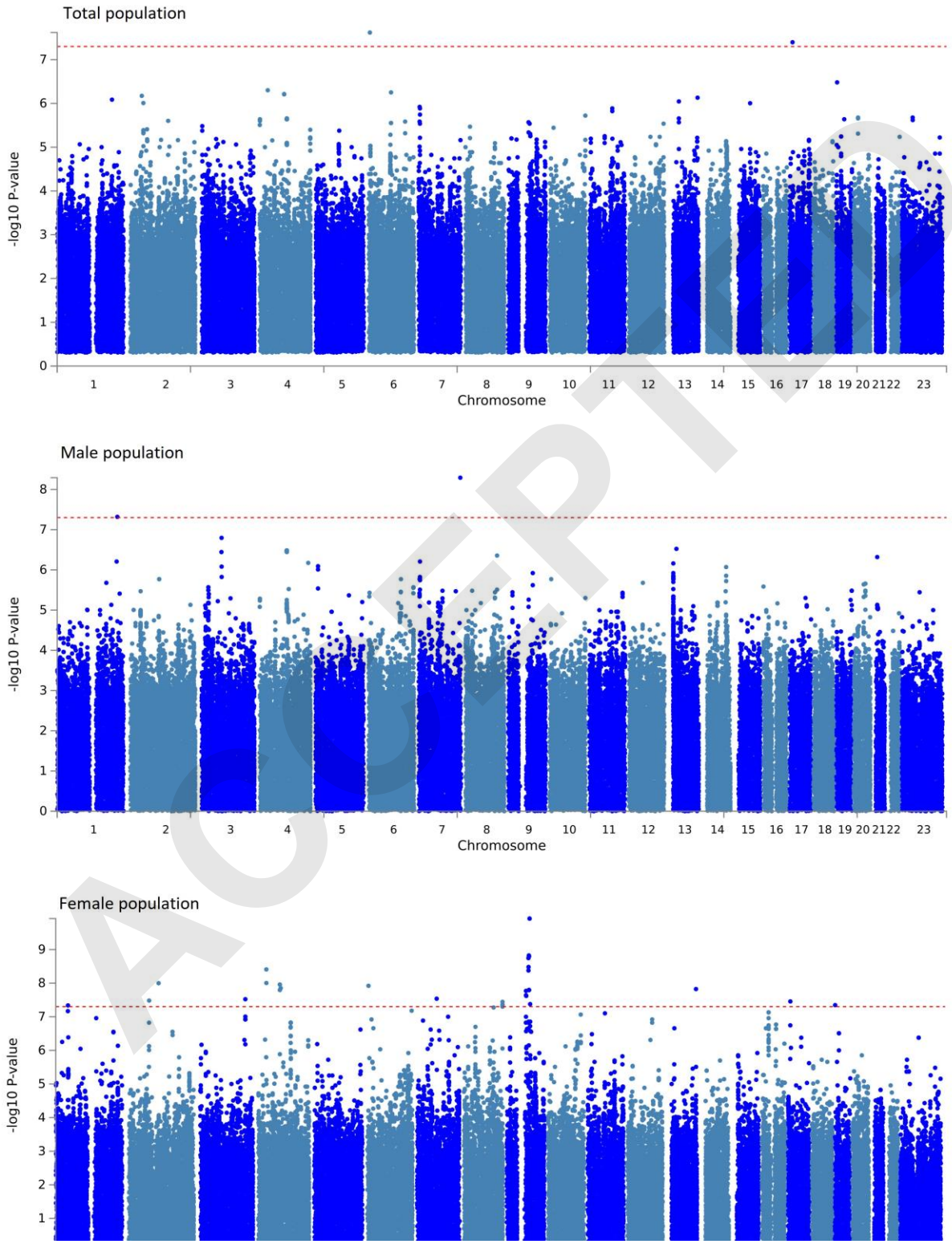


**Figure 2**

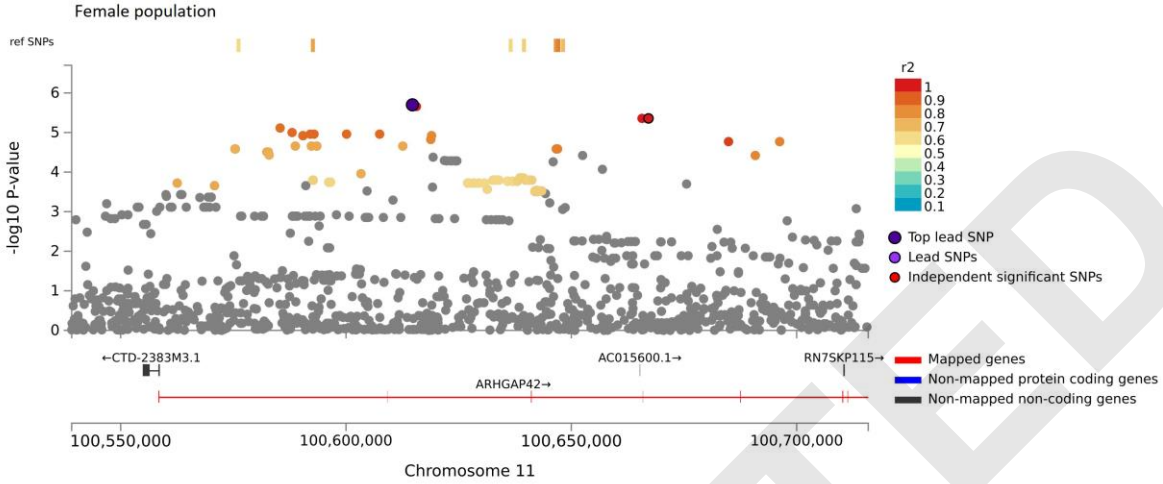


**Figure 3**

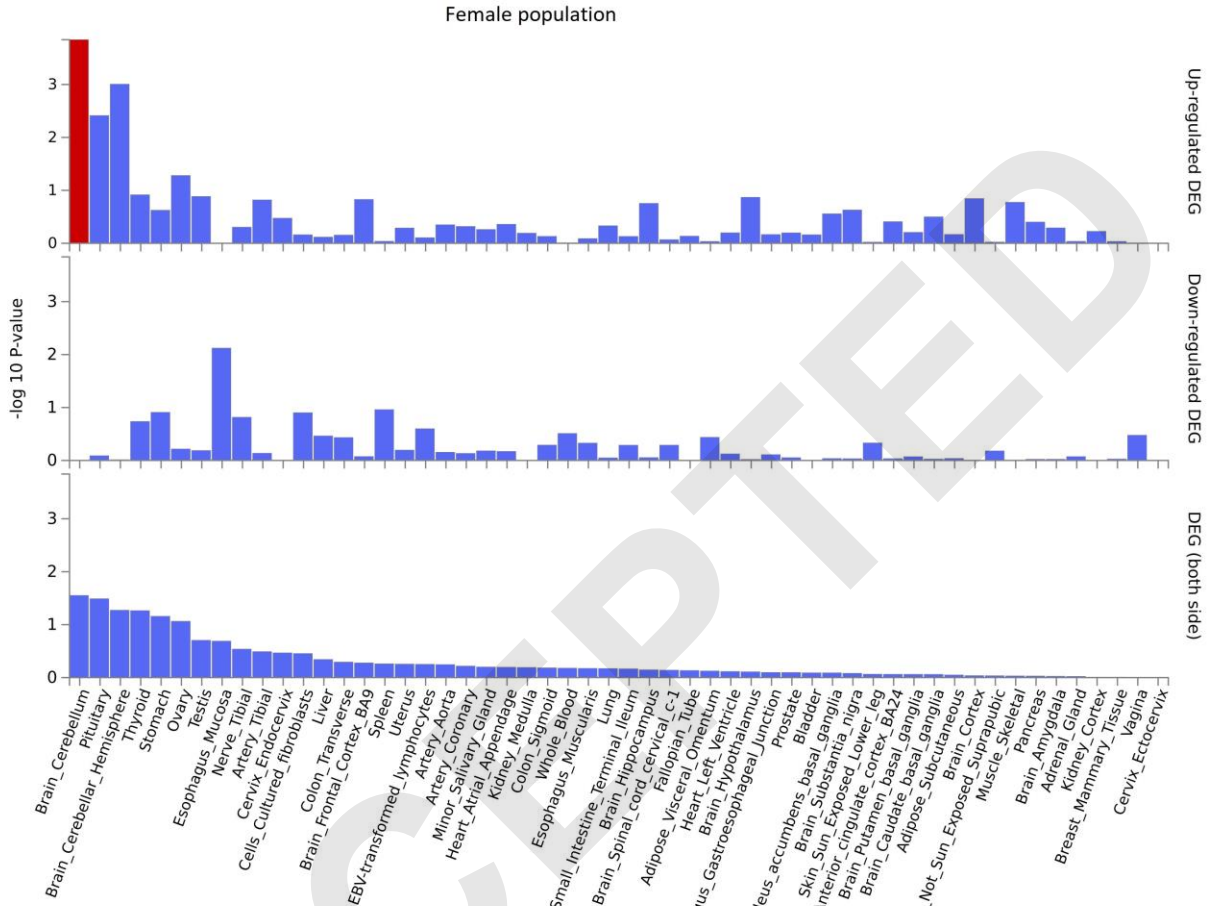


**Figure 4**

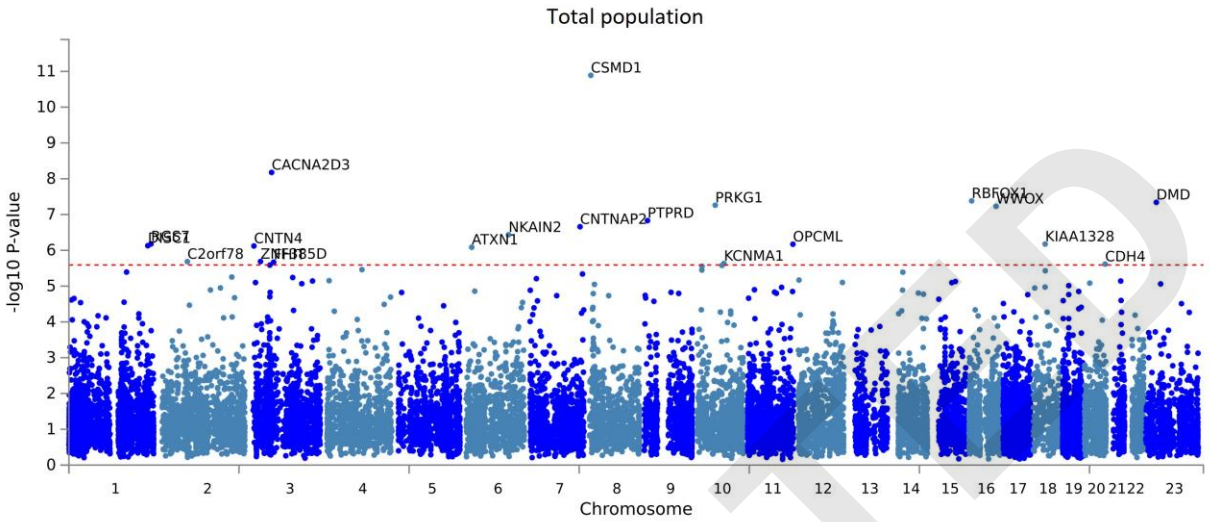


**Figure 5**

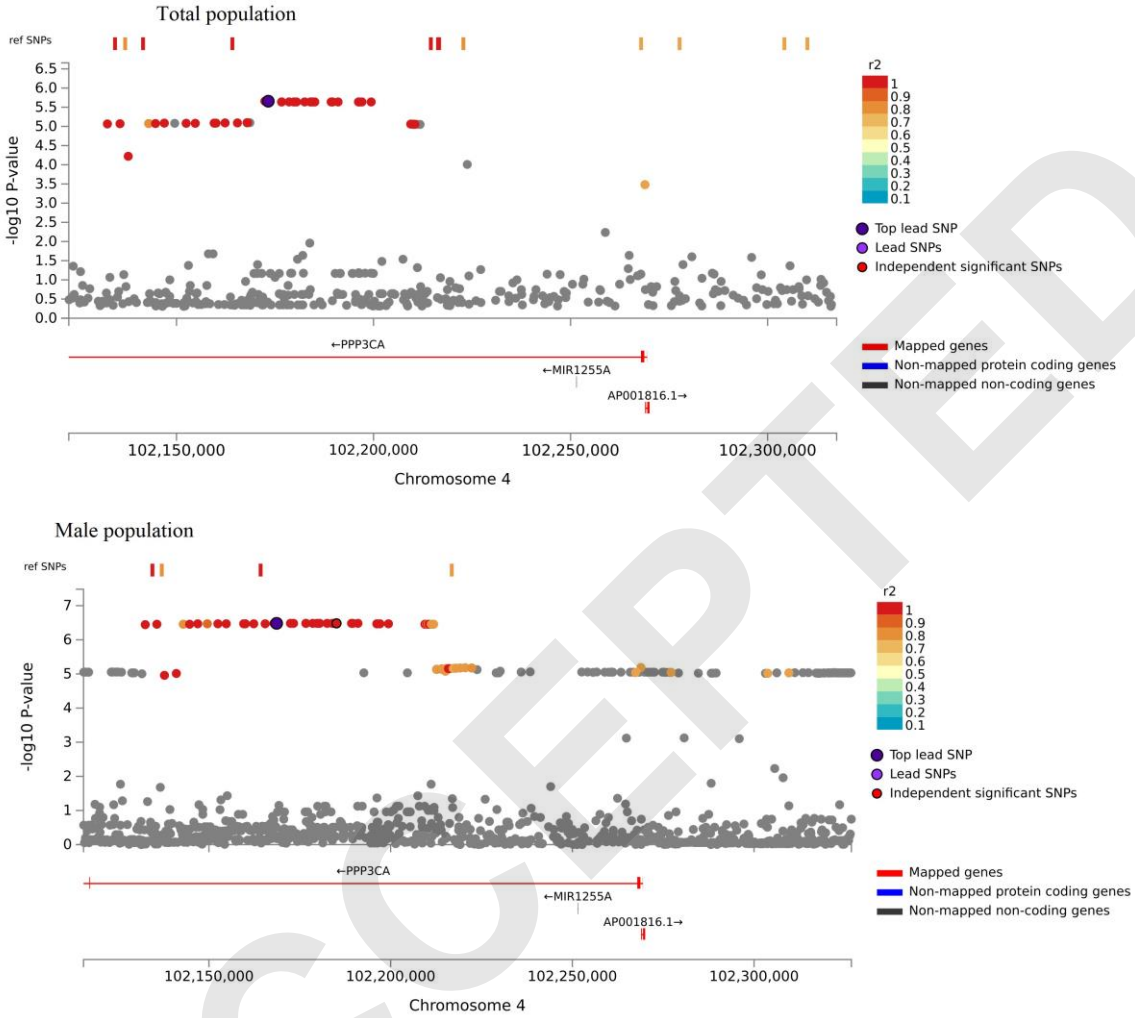


**Table 1:** Participant characteristics.

| | Total population (N=4,656) | Males (n=2,271) | Females (n=2,385) |
|---|---|---|---|
| $VO_{2peak}$, mL/min | 3.1 (0.9) | 3.7 (0.8) | 2.5 (0.5) |
| $VO_{2peak}$ scaled, mL/kg$^{0.75}$/min | 115.3 (28.4) | 133.4 (26.7) | 101.6 (20.9) |
| Weight, kg | 77.5 (13.8) | 85.6 (11.5) | 69.8 (11.2) |
| Age, years | 48.9 (13.9) | 49.3 (13.7) | 48.4 (14.2) |
| Physical activity index score | 3.8 (2.6) | 2.5 (2.8) | 3.8 (2.5) |

Data is presented as mean (standard deviation, SD); $VO_{2peak}$, peak oxygen uptake.

**Table 2:** Genome-wide significant SNPs associated with $VO_{2peak}$ in HUNT.

| Chr | BP | SNP | Gene | SNP location | Effect allele/ reference allele | MAF | Estimated Imputation Accuracy ($R^2$) | Beta | Std.error | P-value |
|---|---|---|---|---|---|---|---|---|---|---|
| **Total Population** | | | | | | | | | | |
| 6 | 4957866 | rs75672239 | CDYL* | downstream | C/G | 0.00160733 | 0.59834 | 31.478 | 5.6417 | 2.4e-08 |
| 17 | 10910855 | rs746795207 | LOC105371536* | intergenic | T/C | 0.00125375 | 0.65245 | 38.9309 | 7.0887 | 4E-08 |
| **Males** | | | | | | | | | | |
| 1 | 225530517 | rs549637059 | DNAH14 | intron | T/C | 0.00128034 | 0.80568 | 50.344 | 9.2235 | 4.8e-08 |
| 7 | 156150246 | rs1467361640 | LOC105375599* | intergenic | C/A | 0.00166081 | 0.94577 | -43.1544 | 7.38267 | 5.1e-09 |
| **Females** | | | | | | | | | | |
| 1 | 45806750 | rs149814343 | TOE1 | missense | C/G | 0.00264201 | genotyped | 25.0357 | 4.58001 | 4.6e-08 |
| 2 | 76050313 | rs868340477 | GCFC2* | intergenic | A/T | 0.00147886 | 0.35846 | 66.882 | 12.1102 | 3.3e-08 |
| 2 | 111556833 | rs190319501 | ACOXL | intron | C/T | 0.00135117 | 0.58459 | 45.8334 | 8.00466 | 1E-08 |
| 3 | 169578004 | rs73176470 | LRRC31 | intron | C/T | 0.0928234 | 0.93161 | 4.63866 | 0.83735 | 3E-08 |
| 4 | 30474174 | rs190373655 | PCDH7* | intergenic | A/T | 0.00121466 | 0.4164 | 57.2064 | 9.71859 | 3.9e-09 |
| 4 | 30572841 | rs145899094 | PCDH7* | intergenic | G/A | 0.00143021 | 0.42865 | 51.877 | 9.05009 | 9.9e-09 |
| 4 | 81273087 | rs563842883 | CFAP299 | intron | T/C | 0.00122787 | 0.48934 | 58.1069 | 10.2793 | 1.6e-08 |
| 4 | 81293181 | rs576426289 | CFAP299 | intron | C/G | 0.0012151 | 0.47773 | 59.6485 | 10.4395 | 1.1e-08 |
| 4 | 84470626 | rs1464440 | GPAT3 | intron | G/A | 0.00144562 | 0.46088 | 51.2648 | 9.02753 | 1.4e-08 |
| 6 | 4957866 | rs75672239 † | CDYL* | downstream | C/G | 0.00122787 | 0.59834 | 48.9615 | 8.59415 | 1.2e-08 |
| 7 | 73169125 | rs561785764 † | CLDN3* | intergenic | G/A | 0.0012915 | 0.69183 | 41.38 | 7.46151 | 2.9e-08 |
| 8 | 145131837 | rs565705049 | EXOSC4 | intron | A/G | 0.00146631 | 0.79789 | 35.7731 | 6.52816 | 4.3e-08 |
| 8 | 145141845 | rs1480758367 | GPAA1* | intergenic | T/C | 0.00144672 | 0.79267 | 35.7011 | 6.52926 | 4.6e-08 |
| 8 | 145162054 | rs373113184 | MAF1 | 3 Prime UTR Variant | G/A | 0.00149031 | 0.7746 | 35.907 | 6.51606 | 3.6e-08 |
| 9 | 72249673 | rs112987080 † | APBA1 | intron | T/C | 0.00125473 | 0.91494 | 40.889 | 7.3151 | 2.3e-08 |
| 9 | 72298696 | rs376927175 † | APBA1* | intergenic | C/A | 0.00122127 | 0.9364 | 40.9888 | 7.26617 | 1.7e-08 |
| 9 | 73180155 | rs190370577 | TRPM3; KLF9-DT | intron | A/G | 0.00178974 | 0.93769 | 32.5177 | 5.82979 | 2.4e-08 |
| 9 | 81328394 | † | MTND2P8* | intergenic | C/T | 0.00132783 | 0.91666 | 42.8323 | 7.12402 | 1.8e-09 |

| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| 9 | 81537070 | rs1476580894 † | LOC105376097* | intergenic | T/A | 0.00201365 | 0.56172 | 40.6245 | 6.91128 | 4.2e-09 |
| 9 | 81763296 | rs929134931 † | LOC101927450* | intergenic | G/A | 0.0013701 | 0.8896 | 41.7095 | 7.05294 | 3.3e-09 |
| 9 | 82806133 | rs1027466565 † | LINC01507* | intergenic | T/G | 0.00177653 | 0.94738 | 33.8797 | 5.61673 | 1.6e-09 |
| 9 | 82971623 | rs974429814 † | LOC105376103 | Non-coding transcript | A/G | 0.00176574 | 0.98952 | 33.8839 | 5.61009 | 1.5e-09 |
| 9 | 83338814 | rs769545479 † | LOC107987084* | intergenic | C/T | 0.00178952 | 0.93283 | 33.8258 | 5.6098 | 1.6e-09 |
| 9 | 83492791 | | ENSG00000226798* | intergenic | CTATAG/C | 0.00221753 | 0.95848 | 28.3403 | 5.01855 | 1.6e-08 |
| 9 | 83495172 | rs111916021 | LOC107987084 | intron | A/G | 0.00221753 | 0.95864 | 28.3403 | 5.01855 | 1.6e-08 |
| 9 | 83503451 | rs1017411201 † | LOC107987084; LOC107987085 | intron | T/A | 0.0017686 | 0.97356 | 33.8807 | 5.61064 | 1.6e-09 |
| 9 | 83640958 | rs952685889 | LOC107987084* | intergenic | A/G | 0.00221885 | 0.97155 | 28.3349 | 5.01852 | 1.6e-08 |
| 9 | 85634079 | rs1001359314 † | RASEF | intron | A/G | 0.00156121 | 0.96722 | 38.8068 | 6.02239 | 1.2e-10 |
| 9 | 85750165 | rs1244784023 † | RASEF* | intergenic | G/A | 0.00155064 | 0.97175 | 38.8729 | 6.03519 | 1.2e-10 |
| 9 | 87805316 | rs762195991 | UBE2V1P10* | intergenic | G/A | 0.0017052 | 0.83795 | 34.4319 | 6.28138 | 4.2e-08 |
| 13 | 111123645 | rs574355537 | COL4A2 | intron | A/G | 0.00154271 | 0.97242 | 33.9001 | 5.99338 | 1.5e-08 |
| 17 | 8557211 | rs190675254 | MYH10* | intergenic | A/G | 0.00152664 | 0.75728 | 42.6319 | 7.73035 | 3.5e-08 |
| 19 | 1703623 | rs184422695 | ONECUT3* | intergenic | T/C | 0.00116931 | 0.44299 | 60.3896 | 11.0386 | 4.5e-08 |
| **Females without adjusting for PA** | | | | | | | | | | |
| 6 | 154591411 | rs35073776 | IPCEF1 | intron | G/T | 0.224869 | 0.99133 | 3.30825 | 0.59745 | 3.1e-08 |
| 9 | 72195971 | rs112610325 | APBA1 | Intron | G/A | 0.00140178 | 0.92391 | 39.347 | 7.19263 | 4.5e-08 |

The range of $VO_{2peak}$ scaled is from 38.45 to 222.06 ($mL/kg^{0.75}/min$), explaining the high beta values. Chr: Chromosome, BP: Base pair, SNP: Single-nucleotide polymorphism, MAF: Minor allele frequency, Std.error: Standard error, * = Nearest gene, † = SNP also identified in the female analysis not adjusting for PA.

**Appendix**:

References of databases used in FUMA:

1000 Genome Project Phase 3:

Auton A, Abecasis GR, Altshuler DM, Durbin RM, Abecasis GR, Bentley DR, et al. A global reference for human genetic variation. Nature. 2015;526(7571):68-74.

ANNOVAR annotations and DEGs (Differentially Expressed Genes):

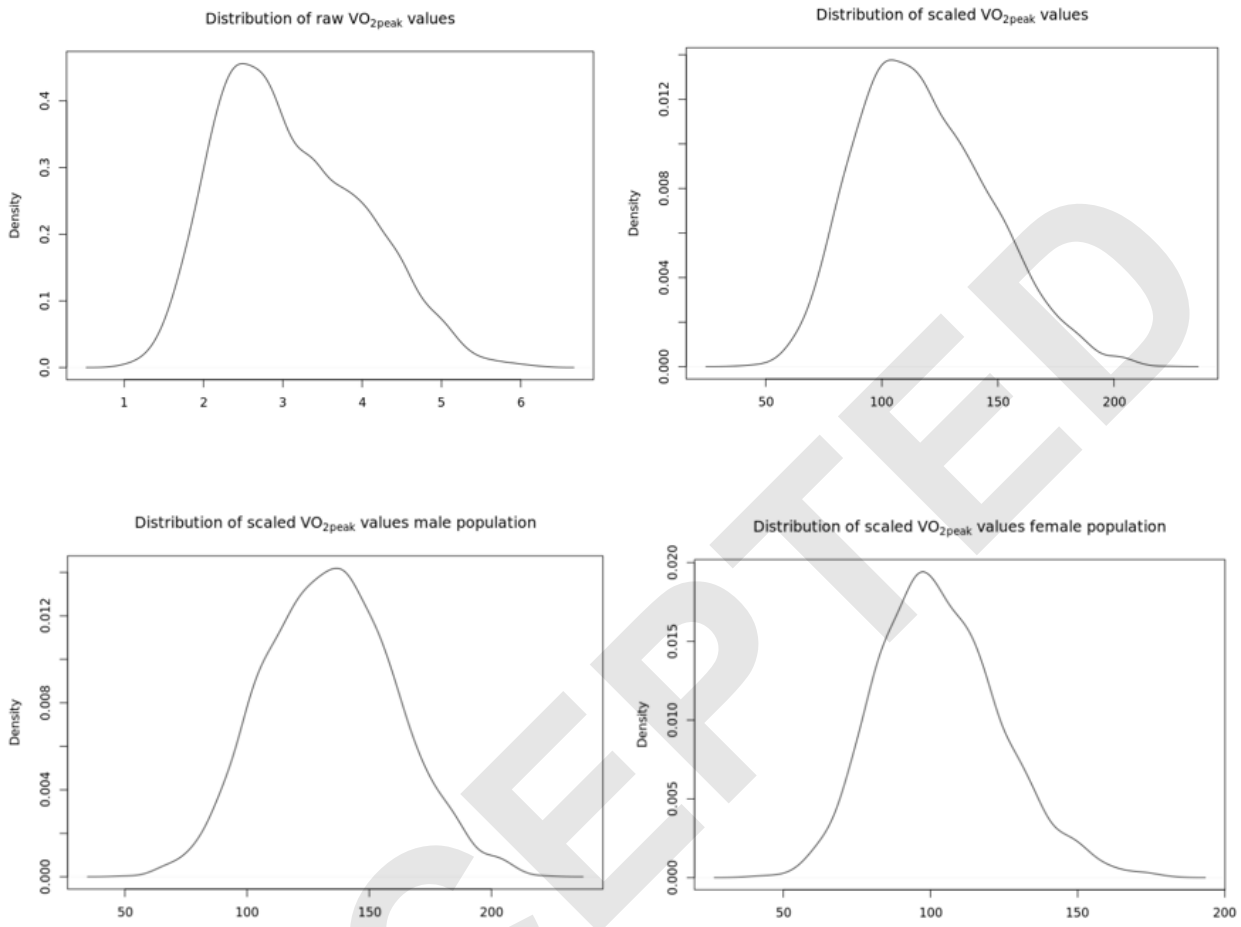
Wang K, Li M, Hakonarson H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. Nucleic Acids Res. 2010;38(16):e164.

The Combined Annotation Dependent Depletion (CADD) score and MSigDB and GTEx v8 databases:

Võsa U, Claringbould A, Westra H-J, Bonder MJ, Deelen P, Zeng B, et al. Unraveling the polygenic architecture of complex traits using blood eQTL metaanalysis. bioRxiv; 2018.

Liberzon A, Subramanian A, Pinchback R, Thorvaldsdóttir H, Tamayo P, Mesirov JP. Molecular signatures database (MSigDB) 3.0. Bioinformatics. 2011;27(12):1739-40.

**Figures and tables to Supplementary:**

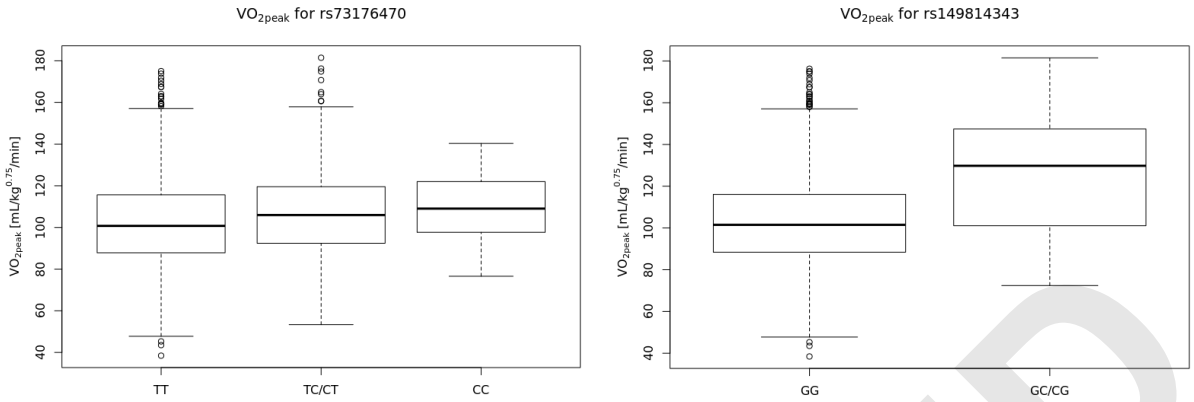



Supplementary Figure 1: The distribution of the raw $VO_{2peak}$ in the total population at the top left corner, the distribution of the scaled $VO_{2peak}$ in the total population at the top right corner, the distribution of the scaled $VO_{2peak}$ in the male population in the lower left corner and the distribution of the scaled $VO_{2peak}$ in the female population in the lower right corner.

Supplementary Table 1: GWAS results of the analysis that was adjusted for age, PC1-4, and genotyping batch from the female population in HUNT.

| Chr | BP | SNP | Gene | SNP location | Effect allele/ reference allele | MAF | Beta | Std. error | P-value |
|---|---|---|---|---|---|---|---|---|---|
| 6 | 4957866 | rs75672239 | CDYL* | Downstream | C/G | 0.00121295 | 50.1606 | 9.14902 | 4.2e-08 |
| 6 | 154591411 | rs35073776 | IPCEF1 | Intron | G/T | 0.224869 | 3.30825 | 0.59745 | 3.1e-08 |
| 7 | 73169125 | rs561785764 | CLDN3* | Intergenic | G/A | 0.00127488 | 43.4964 | 7.93558 | 4.2e-08 |
| 9 | 72195971 | rs112610325 | APBA1 | Intron | G/A | 0.00140178 | 39.347 | 7.19263 | 4.5e-08 |
| 9 | 72249673 | rs112987080 | APBA1 | Intron | T/C | 0.00123837 | 46.0953 | 7.78054 | 3.1e-09 |
| 9 | 72298696 | rs376927175 | APBA1* | Intergenic | C/A | 0.00120535 | 46.2356 | 7.72824 | 2.2e-09 |
| 9 | 81328394 | rs34124007 | MTND2P8* | Indel | C/T | 0.00131052 | 44.7228 | 7.57946 | 3.6e-09 |
| 9 | 81537070 | rs1476580894 | LOC105376097* | Intergenic | T/A | 0.00200022 | 42.2179 | 7.35335 | 9.4e-09 |
| 9 | 81763296 | rs929134931 | LOC101927450* | Intergenic | G/A | 0.00135224 | 43.4195 | 7.50413 | 7.2e-09 |
| 9 | 82806133 | rs1027466565 | LINC01507* | Intergenic | T/G | 0.00175337 | 33.928 | 5.97522 | 1.4e-08 |
| 9 | 82971623 | rs974429814 | LOC105376103 | Non-coding transcript | A/G | 0.00174272 | 33.9349 | 5.96818 | 1.3e-08 |
| 9 | 83338814 | rs769545479 | LOC107987084* | Intergenic | C/T | 0.00176619 | 33.8727 | 5.9679 | 1.4e-08 |
| 9 | 83503451 | rs1017411201 | LOC107987084; LOC107987085 | Intron | T/A | 0.00174555 | 33.9357 | 5.96879 | 1.3e-08 |
| 9 | 85634079 | rs1001359314 | RASEF | Intron | A/G | 0.00154129 | 37.6687 | 6.41038 | 4.2e-09 |
| 9 | 85750165 | rs1244784023 | RASEF* | intergenic | G/A | 0.00153042 | 37.7185 | 6.42401 | 4.3e-09 |

Chr: Chromosome, BP: Base pair, SNP: Single-nucleotide polymorphism, MAF: Minor allele frequency, * = Nearest gene.

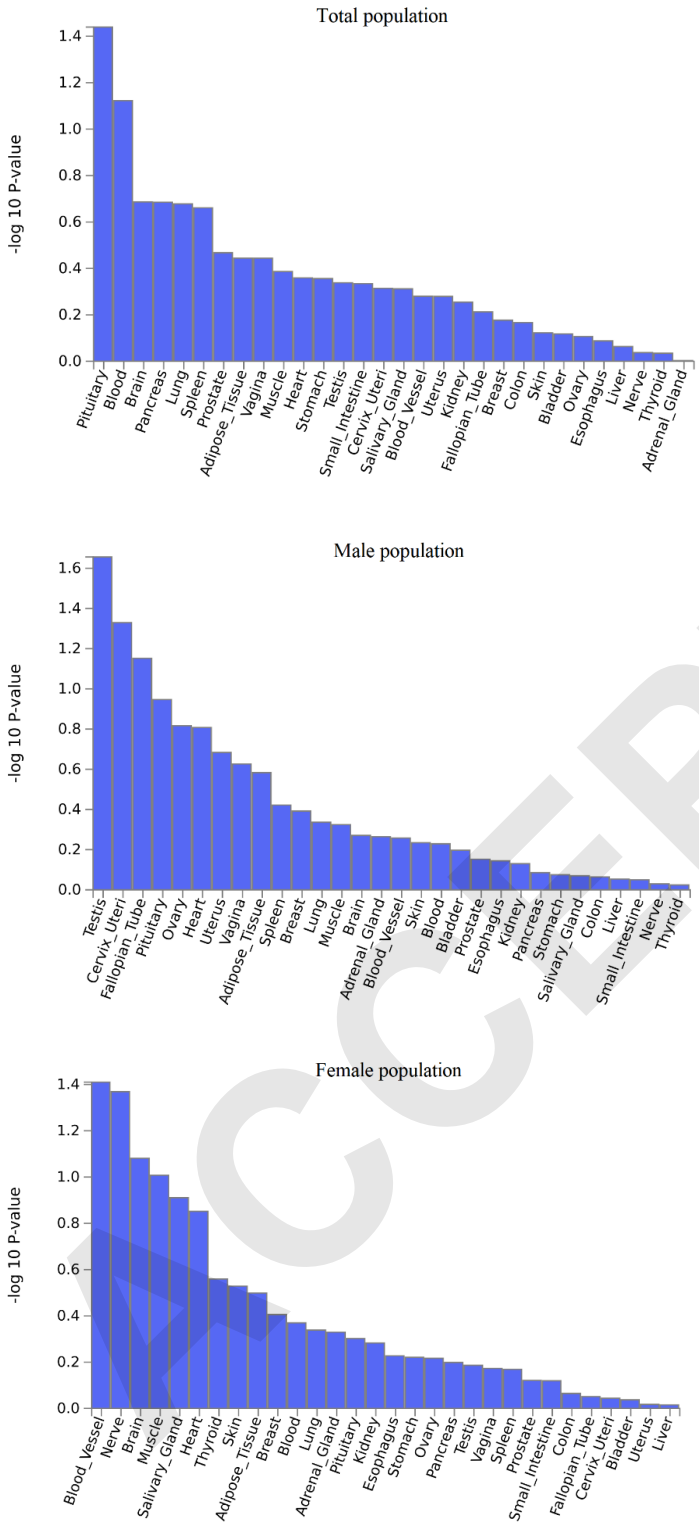



Supplementary Figure 2: Boxplots of $VO_{2peak}$ by alleles for rs73176470 and rs149814343, respectively.

Supplementary Table 2: Summary of SNPs and mapped genes, from FUMA SNP2GENE analysis.

| | Number of genomic risk loci | Number of lead SNPs ($p < 5\times10^{-6}$) | Number of independent significant SNPs | Number of candidate SNPs | Number of mapped genes |
|---|---|---|---|---|---|
| Total | 22 | 25 | 30 | 223 | 43 |
| Male | 28 | 30 | 34 | 535 | 61 |
| Female | 95 | 104 | 123 | 880 | 197 |

SNP: Single-nucleotide polymorphism.




Supplementary Figure 3: Magma tissue expression analysis for the total population on top, the male population in the middle, and the female population at bottom. The x-axis shows the different tissues and the negative logarithm of the associated *p*-value for each tissue is on the y-axis.