*J Appl Physiol* 126: 1292–1314, 2019.
First published January 3, 2019; doi:10.1152/japplphysiol.00035.2018.

**RESEARCH ARTICLE**

# Exploring the underlying biology of intrinsic cardiorespiratory fitness through integrative analysis of genomic variants and muscle gene expression profiling

**Sujoy Ghosh,[1,2] Monalisa Hota,[2] Xiaoran Chai,[2] Jencee Kiranya,[2] Palash Ghosh,[3] Zihong He,[1,4] Jonathan J. Ruiz-Ramie,[5] Mark A. Sarzynski,[5] and Claude Bouchard[1]**

[1]*Human Genomics Laboratory, Pennington Biomedical Research Center, Baton Rouge, Louisiana;* [2]*Cardiovascular and Metabolic Disorders Program and Centre for Computational Biology, Duke-National University of Singapore Medical School, Singapore;* [3]*Center for Quantitative Medicine, Duke-National University of Singapore Medical School, Singapore;* [4]*Department of Biology, China Institute of Sport Science, Beijing, China; and* [5]*Department of Exercise Science, Arnold School of Public Health, University of South Carolina, Columbia, South Carolina*

Submitted 12 January 2018; accepted in final form 9 December 2018

**Ghosh S, Hota M, Chai X, Kiranya J, Ghosh P, He Z, Ruiz-Ramie JJ, Sarzynski MA, Bouchard C.** Exploring the underlying biology of intrinsic cardiorespiratory fitness through integrative analysis of genomic variants and muscle gene expression profiling. *J Appl Physiol* 126: 1292–1314, 2019. First published January 3, 2019; doi:10.1152/japplphysiol.00035.2018.—Intrinsic cardiorespiratory fitness (CRF) is defined as the level of CRF in the sedentary state. There are large individual differences in intrinsic CRF among sedentary adults. The physiology of variability in CRF has received much attention, but little is known about the genetic and molecular mechanisms that impact intrinsic CRF. These issues were explored in the present study by interrogating intrinsic CRF-associated DNA sequence variation and skeletal muscle gene expression data from the HERITAGE Family Study through an integrative bioinformatics guided approach. A combined analytic strategy involving genetic association, pathway enrichment, tissue-specific network structure, cis-regulatory genome effects, and expression quantitative trait loci was used to select and rank genes through a variation-adjusted weighted ranking scheme. Prioritized genes were further interrogated for corroborative evidence from knockout mouse phenotypes and relevant physiological traits from the HERITAGE cohort. The mean intrinsic $\dot{V}O_{2max}$ was 33.1 ml $O_2 \cdot kg^{-1} \cdot min^{-1}$ (SD = 8.8) for the sample of 493 sedentary adults. Suggestive evidence was found for gene loci related to cardiovascular physiology (*ATE1*, *CASQ2*, *NOTO*, and *SGCG*), hematopoiesis (*PICALM*, *SSB*, *CA9*, and *CASQ2*), skeletal muscle phenotypes (*SGCG*, *DMRT2*, *ADARB1*, and *CASQ2*), and metabolism (*ATE1*, *PICALM*, *RAB11FIP5*, *GBA2*, *SGCG*, *PRADC1*, *ARL6IP5*, and *CASQ2*). Supportive evidence for a role of several of these loci was uncovered via association between DNA variants and muscle gene expression levels with exercise cardiovascular and muscle physiological traits. This initial effort to define the underlying molecular substrates of intrinsic CRF warrants further studies based on appropriate cohorts and study designs, complemented by functional investigations.

**NEW & NOTEWORTHY** Intrinsic cardiorespiratory fitness (CRF) is measured in the sedentary state and is highly variable among sedentary adults. The physiology of variability in intrinsic cardiorespiratory fitness has received much attention, but little is known about the genetic and molecular mechanisms that impact intrinsic CRF. These issues were explored computationally in the present study, with further corroborative evidence obtained from analysis of phenotype data from knockout mouse models and human cardiovascular and skeletal muscle measurements.

bioinformatics; cardiovascular physiology; in silico exploration of the biology of cardiorespiratory fitness; intrinsic cardiorespiratory fitness; skeletal muscle biology

## INTRODUCTION

$\dot{V}O_{2max}$ is defined as the oxygen uptake attained during maximal exercise intensity that is not increased despite further increases in exercise workload (44). $\dot{V}O_{2max}$ is considered the best indicator of cardiorespiratory and physiological fitness. Cardiorespiratory fitness (CRF) is broadly defined as the ability of the circulatory and respiratory systems to deliver oxygen to the working muscles during maximal exercise. There are also submaximal indicators of CRF, but they are not considered in the present paper. Low CRF is a well-established risk factor for all-cause and disease-specific mortality (43) in blacks and whites (52), in both sexes (4), in various body mass index (BMI) groups (88), in different age groups (51, 66), in apparently healthy people (52), but also in patients with diabetes (16, 53), cardiovascular disease (52), or hypertension (15).

Although of considerable importance, the distinction between intrinsic CRF and acquired CRF is rarely made (6). Intrinsic CRF relates to the level of CRF exhibited in the sedentary state. In contrast, acquired CRF is the level that is achieved as a result of exposure to regular exercise or exercise training. Importantly, there is no substantive relationship between intrinsic CRF and acquired CRF, as evidenced by the lack of correlation between baseline $\dot{V}O_{2max}$ and the gains in $\dot{V}O_{2max}$ in response to a 20-wk exercise program (78).

There are large individual differences in intrinsic CRF among adults who are sedentary and who have a history of not engaging in regular exercise (6). For instance, among 726 sedentary subjects (17–65 yr of age) in whom $\dot{V}O_{2max}$ was measured twice (on separate days) at baseline in the HERITAGE Family Study (8), the mean value was 31 ml $O_2 \cdot kg^{-1} \cdot min^{-1}$ with an SD of 9 ml $O_2 \cdot kg^{-1} \cdot min^{-1}$. This represents an extraordinary degree of heterogeneity among people who were confirmed as sedentary with no substantial amount of exercise training in their past. As shown by the findings of the HERITAGE Family Study, there are many factors contributing to variation in CRF. Among

Address for reprint requests and other correspondence: C. Bouchard, Human Genomics Laboratory, Pennington Biomedical Research Center, 6400 Perkins Rd., Baton Rouge, LA 70808 (e-mail: claude.bouchard@pbrc.edu).

them, maximum likelihood estimation revealed a maximal heritability of 51% for intrinsic $\dot{V}O_{2max}$ adjusted for age, sex, body mass, fat-free mass, and fat mass (7).

The physiological systems known to be broad determinants of $\dot{V}O_{2max}$ relate to $O_2$ carrying capacity and skeletal muscle $O_2$ extraction (2, 46). Oxygen delivery to the working muscle is driven by heart size, cardiac output, blood volume, and total hemoglobin (36, 37, 49, 76, 86, 89), while $O_2$ transfer and utilization is influenced by muscle capillary density, membrane permeability, $O_2$ solubility, muscle fiber size, fiber type, and total myoglobin (11, 23, 84), as well as muscle mitochondria volume density and oxidative capacity (23, 76, 85). These multiple components are part of a conductance pathway forming an integrated system, where all components are interacting to achieve the highest possible maximal power output, thus defining $\dot{V}O_{2max}$ (56, 85). In the aggregate, however, CRF is primarily determined by the maximal $O_2$ carrying and delivery capacity in healthy individuals (10, 56). Estimates are that ~70% of the variability in $\dot{V}O_{2max}$ is accounted by oxygen transport capacity (26, 27).

Although the integrative physiology of variability in CRF has received some attention, little is known about potentially causal molecular mechanisms involving genes, pathways, and networks that contribute to variation in intrinsic CRF in adults. Given the high heritability of intrinsic CRF, there is continuing interest in investigating the roles that genetic variation might play in regulating trait variation. Genome-wide association studies (GWAS) are one way to address genetic contributions, but for common traits (such as intrinsic CRF), the standard analysis has been handicapped traditionally by the fact that most of the trait-associated variation resides in the regulatory, noncoding genome, thereby precluding any direct associations of genetic variation with protein function. Nevertheless, compared with other genome features, the putative genomic regulatory elements (e.g., enhancer and promoter regions) are also known to have, by far, the largest contribution towards the heritability of several human traits (41). Recently, however, the availability of large data sets encompassing genome-scale information on chromatin regulatory marks (1, 27b), expression quantitative trait loci (eQTL) (39a), tissue-specific gene-gene interaction networks (39), etc. has presented unprecedented opportunities to generate hypotheses linking trait-associated genetic variation to genetic and molecular events and to ultimately connect genotypes to phenotypes (69). Generally referred to as "integrative genetics," this approach has led to significant advances in our understanding of several cardiometabolic phenotypes including obesity, type 2 diabetes, cardiovascular disease (33, 34, 60, 63, 67), and exercise-dependent changes in CRF (35).

In the present study, we have taken an analogous integrative approach to understand the role of genetic variation in intrinsic CRF. Briefly, as a starting point, we used summary statistics from a GWAS on intrinsic CRF (HERITAGE Family Study) and interrogated the possible roles of genetic variants in affecting promoter and enhancer-marking histone binding and transcription factor binding and influencing nearby gene expression as eQTLs. We further investigated the possible enrichment of variant-associated candidate genes in tissue-specific functional networks and highly curated biological pathways to generate hypotheses on tissue-relevant biological mechanisms impacted by genetic variation. We sought further confirmation of the candidate genes implicated from in silico analyses by interrogating, when available, their knockout phenotypes in mouse models. The availability of whole genome expression data on skeletal muscle biopsies on a subset of the HERITAGE cohort subjects further allowed us to test whether the expression of the computationally identified genes was also correlated with intrinsic CRF and with skeletal morphological and metabolic indicators obtained from muscle biopsies. Finally, DNA sequence variants at the gene loci prioritized at the conclusion of this series of analyses were investigated for their associations with cardiovascular and muscle intermediate phenotypes of intrinsic CRF.

## MATERIALS AND METHODS

*HERITAGE Family Study.* The subjects and study design of the HERITAGE Family Study have been described elsewhere (8). Briefly, 834 subjects from 218 families were recruited to participate in an endurance exercise training study and have measurements of baseline $\dot{V}O_{2max}$. Among them, 493 adults (241 men and 252 women) from 99 families of European descent who were confirmed sedentary and had taken two maximal exercise tests to exhaustion at baseline constitute the population of the current study. Parents were 65 yr of age or less while offspring ranged in age from 17 to 41 yr. Participants were sedentary at baseline, normotensive or mildly hypertensive (<160/100 mmHg) without medications for hypertension, diabetes, or dyslipidemia (8). The study protocol had been approved by the Institutional Review Boards at each of the participating centers of the HERITAGE Family Study consortium. Written informed consent was obtained from each participant.

*Cardiorespiratory fitness measurement.* Three exercise tests were performed on separate days at baseline on a SensorMedics 800S (Yorba Linda, CA) cycle ergometer and a SensorMedics 2900 metabolic measurement cart (80). The tests were conducted at about the same time of day, with at least 48 h between two tests. In the first maximal exercise test, subjects exercised at a power output of 50 W for 3 min, followed by increases of 25 W each 2 min until volitional exhaustion. For older, smaller, or less fit individuals, the test was started at 40 W, with increases of 10–20 W each 2 min thereafter. In the second test, subjects performed a submaximal exercise test during which subjects exercised for 10 min each at an absolute (50 W) and at a relative power output equivalent to 60% $\dot{V}O_{2max}$. Finally, a submaximal/maximal exercise test was performed, starting with the same protocol as the submaximal test and then followed by 3 min of exercise at a relative power output that was 80% of their $\dot{V}O_{2max}$, after which resistance was increased to the highest power output attained in the first maximal test. If the subjects were able to pedal after 2 min, power output was increased each 2 min thereafter until they reached volitional fatigue. The average $\dot{V}O_{2max}$ from the maximal and submaximal/maximal tests was taken as the $\dot{V}O_{2max}$ for that subject and used in analyses if both values were within 5% of each other. If they differed by >5%, the higher $\dot{V}O_{2max}$ value was used. Submaximal and maximal exercise phenotypes of interest that were measured in the sedentary state include heart rate, stroke volume, cardiac output, systolic blood pressure, $\dot{V}O_2$, and other metabolic indicators at 50 W and at 60% of $\dot{V}O_{2max}$ (80, 90).

*Genome-wide genotyping.* Genome-wide genotyping was performed using the Illumina HumanCNV370-Quad v3.0 BeadChips on Illumina BeadStation 500GX platform. The genotype calls were determined via the Illumina GenomeStudio software, and all samples were called in the same batch to eliminate batch-to-batch variation. Monomorphic single-nucleotide polymorphisms (SNPs), SNPs with only one heterozygote, and SNPs with >30% missing data were filtered out with GenomeStudio. Twelve samples were genotyped twice with 100% reproducibility across all SNPs. All GenomeStudio genotype calls with a GenTrain score <0.885 were checked and

confirmed manually. Quality control of the GWAS SNP data confirmed all family relationships and found no evidence of DNA sample mix-ups.

Imputation was performed using a CEU reference panel (Northern and Western European ancestry) consisting of 120 haplotypes from HapMap Phase II data (release 22, build 36) and the MACH software (57). A total of 2,228,863 directly typed or imputed SNPs were tested for association with baseline CRF with adjustment for age and sex as previously described (9).

*Affymetrix microarray analysis.* Biopsies of vastus lateralis muscle were performed at baseline using the percutaneous needle biopsy technique in 78 HERITAGE Caucasian subjects from the Laval University (Québec) Clinical Center (72). Total RNA was isolated from frozen muscle biopsies preserved in Tissue-Tek using Trizol and mRNA amplified with Ambion MessageAmp Premier following manufacturer's instructions. A subset of 52 samples was processed on Affymetrix HG-U133+2 arrays to quantify global gene expression levels. Background-corrected, quantile-normalized, and $\log_2$-transformed expression data were obtained via the method of Robust Multichip Averages (5). Probe sets with normalized expression $<50$ units in $>90\%$ of samples were removed, resulting in 16,312 probes for further analysis. Partial correlations between gene expression levels and intrinsic CRF levels were calculated via the *ppcor* package in R after adjustments for age, sex, BMI, and scan date. The relation of gene expression to intrinsic CRF levels was further modeled via general linear models (adjusted for age, sex, BMI, and scan date) and visualized via partial residual regression plots in JMP (version 10.0.2; SAS Institute, Cary, NC). For both generalized linear model and partial correlations, a nominal $P \leq 0.05$ was considered as the threshold for statistical significance. The raw microarray data for this study have been deposited with Gene Expression Omnibus under accession no. GSE117070 (17).

*Measurement of muscle-related traits.* The measurement of muscle-related phenotypes in the subsample of HERITAGE participants that underwent muscle biopsies has been described previously (72). Briefly, based on the staining properties for myofibrillar ATP, fibers from the muscle biopsy samples were designated as type 1, type 2a, and type 2b. The mean muscle fiber area was determined by averaging the cross-sectional areas of 20 randomly selected fibers of each type. The number of capillaries around each of these fibers was counted to determine the capillary density and the area per capillary ratio in each fiber type. The maximal activities of the following enzymes were also determined on muscle homogenates from the biopsy samples (72): creatine kinase, phosphorylase, hexokinase, phosphofructokinase, glyceraldehyde phosphate dehydrogenase, 3-β-hydroxyacyl CoA dehydrogenase (HADH), carnitine palmitoyl transferase, citrate synthase, and cytochrome *c* oxidase. As described below, these muscle-related traits were tested for their association with SNPs from candidate genes identified from our integrative bioinformatic analyses.

*Analyses of associations with SNP and expression with cardiovascular and muscle traits.* The integrative bioinformatic analysis (described below) identified four genes related to cardiovascular physiology (*ATE1*, *CASQ2*, *NOTO*, and *SGCG*) and four genes to skeletal muscle phenotypes (*SGCG*, *DMRT2*, *ADARB1*, and *CASQ2*), for a total of six unique genes. For these genes, the association of genotyped or imputed SNPs located $\pm 50$ kb from the gene were tested for association with select cardiovascular and muscle-related traits (Supplemental Table S1; Supplemental Material for this article is available online at the Journal website) using general linear models with age, sex, and BMI as covariates (Proc GLM in SAS version 9.4). As noted above, the SNP-muscle trait association analyses were performed in the subset of subjects with muscle biopsy data ($n = 52$). For every gene, the list of all SNPs associated with a cardiovascular or muscle trait at a nominal $P < 0.05$ was further pruned via the SNPClip module of LDLink software (59) using the 1000 Genomes (Phase 3) CEU reference panel for linkage disequilibrium (LD) estimations. Biallelic SNPs with $r^2 \leq 0.2$ and minor allele frequency $\leq 0.01$ were considered as independent.

The associations of gene expression with muscle traits were examined via general linear models after adjustments for age, sex, BMI, and scan date (JMP, v 10.0.2). Note that no gene expression results were available for *NOTO* as the gene was not represented on the Affymetrix arrays used in the study. The relations of gene expression to muscle traits were visualized via partial residual regression plots, similar to those for intrinsic CRF levels.

*Bioinformatic studies.* The overall strategy for integrative bioinformatics analysis used in the present report is depicted in Fig. 1. Details of the analysis components are described below and summarized in the Fig. 1 legend.

*Regulatory genome analysis.* To investigate the possible influence of regulatory, noncoding variation on intrinsic CRF levels, we interrogated GWAS-associated SNPs (index SNP $P \leq 1 \times 10^{-04}$, plus SNPs in LD at $r^2 \geq 0.8$, total of 1,596 SNPs; 1000G Phase 1 EUR population-based LD estimated via LDlink software) (59) for their overlap with chromatin states associated with active or repressed transcription. Specifically, we used ENCODE and Roadmap Epigenomics project-derived data in Haploreg (version 4.1) (87) to identify SNPs overlapping with binding sites for modified histones associated with enhancers (H3K4me1 and H2K27ac) and promoters (H3K4me3 and H3K9ac), respectively, in 28 selected tissues considered relevant for intrinsic CRF. These epigenomic signatures were derived from a 15-state hidden Markov model. Enrichment for enhancer coverage among the query SNPs, compared with a background of all common SNPs (minor allele frequency $>5\%$), was calculated in Haploreg by the binomial test across 28 tissues. Additional statistical analysis for SNP enrichment at other genomic features including long intergenic noncoding RNAs (lincRNAs), modified-histone binding regions, gene features (introns, exons, and untranslated regions), DNA

Fig. 1. Overall scheme for integrative bioinformatics. Diagram summarizes the integrative bioinformatics approach employed in this study. It consists of applying specific tools to examine the possible consequences of genetic association study results on genome regulation (DNA) or gene expression (RNA) (highlighted in yellow on the *left*). The *boxes 1–8* at the *left* summarize the various analytic approaches with brief descriptions. The *middle* shows a visual model of the analysis referred to. The blue boxes to the *right* list the various bioinformatic tools employed to perform the corresponding analyses. The numbers in the blue boxes to the right of the bioinformatic tools refer to the Pubmed IDs (PMIDs) of the publications describing the respective software tools. Beginning with genome-wide association study (GWAS) summary data (*bottom*, *box 1*), single-nucleotide polymorphisms (SNPs) are selected at a user-defined threshold (above the red line in the Manhattan plot) and queried, via several high-content online databases (e.g., ENCODE, Roadmap, GTEx, etc.), for their effects on genome cis-regulation [e.g., histone and transcription factor (TF) binding and expression of nearby genes] and their enrichment in functional interaction networks (*boxes 2–5*). Additional queries investigate the enrichment of SNP-associated genes in biological pathways (*box 6*). These combined analyses help generate a short list of candidate genes potentially linked to the trait of interest. Further validation of these gene candidates is sought by interrogating their effects in knockout animal models (*box 7*) and by correlating gene expression to molecular and physiologic end points relevant to the trait (*box 8*). The refined list of candidate genes is then ranked, via some appropriate metric, to select genes for functional validation against trait-relevant phenotypes such as those listed at the *top* of the diagram. For illustration purposes, a gene is indicated schematically in red as a line with 4 boxes to represent the exon-intron structure. A selected SNP, shown to reside in the gene promoter that overlaps with histone and transcription factor binding, functions as an expression quantitative trait loci (eQTL). The SNP-associated gene belongs to a gene network and biological pathway, shows a phenotype in mouse models and is transcriptionally correlated to intrinsic cardiorespiratory fitness levels (ascending order of analyses in the schematic, dashed line).

methylation regions, long-range chromatin interactions (HiC), and nuclear lamina-associated domains (LAD) were assessed via the Genomic Association Tester software (GAT) (45). For this analysis, genome-feature level data was retrieved from multiple data repositories including ENCODE, FANTOM, and UCSC (27b, 47). Random size-matched SNP sets were tested for feature overlap to generate the expected overlap for a null distribution. The magnitude of enrichment was estimated via fold changes (observed overlap vs. expected overlap), and significance of enrichment was assessed from simulation derived empirical *P* values after control for the false discovery rate (FDR). The predicted effects of a SNP on a histone binding site were quantified as the differences in chromatin feature probabilities (here

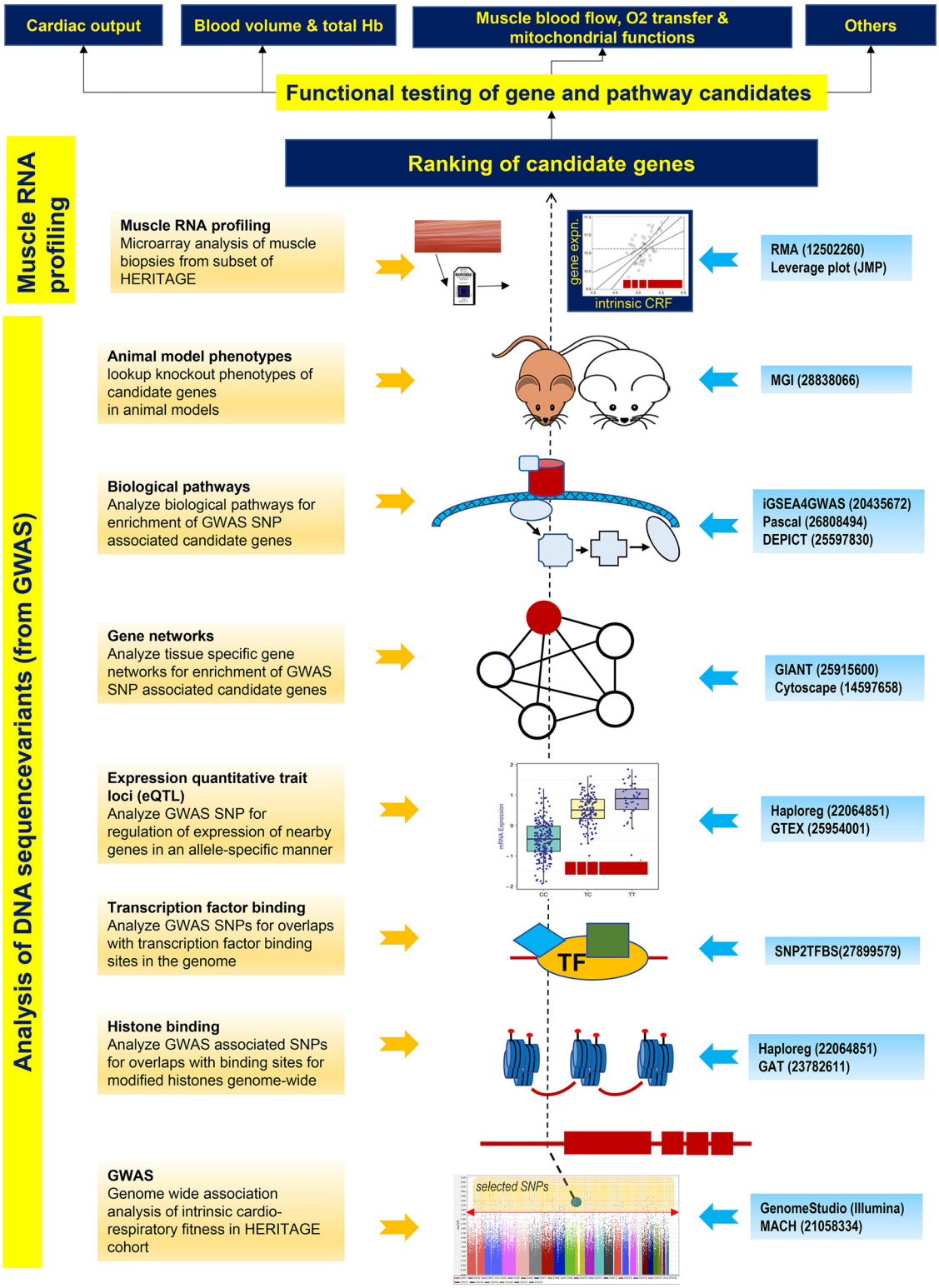


Fig. 1.

histone binding) between the reference and alternative SNP alleles, computed via a deep-learning-based algorithmic framework implemented in DeepSea (92). The analysis was conducted separately for the four major histone modifications (H3K4me1, H3K4me3, H3K9ac, and H3K27ac) using data available in ENCODE for seven cell types (H1-hESC, K562, CD14+ monocyte, NH-A, NHDF, NHEK, and NHLF). SNP-level probability differences for each histone mark were then converted to gene-level effects by taking the 90th percentile of the probability difference values for all SNPs mapping to a gene.

In addition to investigating the effects on histone binding, we also interrogated the overlap of the 1596 query SNPs with genome-wide eQTL reported in Haploreg from tissue-specific gene expression data from GTEX v6.0 (39c) and other eQTL databases (42, 71). Finally, we also examined the extent of SNP overlap with transcription factor binding sites through the SNP2TFBS tool by analyzing the effects of reference and alternate SNP alleles on transcription factor binding based on changes in the position weight matrix (PWM) scores for the binding specificity of the corresponding factor (JASPAR Core Vertebrate 2014 version).

*Tissue prioritization through network analysis.* We conducted genome-scale tissue-level network analysis on networks available from the Genome-scale Integrated Analysis of gene Networks in Tissues (GIANT) resource (39). These tissue networks are based on 987 genome-scale data sets encompassing ~38,000 experimental conditions covering gene expression and functional interaction measurements from multiple data sources (12, 62, 65). Together, we queried 145 tissue-specific networks via programmatic access through the GIANT API (74) utilizing 186 genes nominally associated to intrinsic CRF ($P \leq 0.01$ by Pascal max method) as positive controls (GWAS-associated genes). The connectivity of a given gene to the GWAS-associated genes within a tissue network was quantified by the NetWAS score (38). The NetWAS score is a weighted and predicted score that captures the extent to which a gene's network connectivity is associated with the GWAS-associated gene set in a tissue, with higher NetWAS scores indicating greater connectivity to the significant gene set (38). The distribution of the NetWAS scores in each tissue was visualized via boxplots, with positive scores indicating greater connectivity and, by extension, greater tissue relevance for intrinsic CRF. Some of the top-scoring tissues (based on median NetWAS scores) were further examined by analyzing subnetworks containing highly connected GWAS genes (NetWAS score $\geq$0.3, recommended cutoff by GIANT). From each of these subnetworks, the top 200 edges were visualized in Cytoscape 3.4.0 (https://cytoscape.org/). The selection of the top 200 edges was a subjective decision driven by considerations for clarity of network visualization.

*Pathway analysis.* For pathway enrichment analyses, we queried a customized pathway database consisting of Reactome-based pathway annotation from the Molecular Signature Database (v5.2) (81) plus custom gene sets, resulting in a total of 678 pathways ($\geq$10 and $\leq$250 genes per pathway). Although a wide number of gene-set repositories are available, we selected the expert-authored and manually curated Reactome database (24) due to its clear structural hierarchy and internally consistent "reaction-based" data model encompassing a wide variety of biological processes. Pathway enrichment analysis was conducted via two separate methods [iGSEA4GWAS V2 (91) and Pascal (55)] to reduce false discoveries by identifying the common pathways implicated in both. For both methods, a gene-wide association *P* value was obtained from the *P* values of individual SNP markers mapping to the gene body $\pm$50-kb upstream and downstream boundaries by using the max statistics (maximum of $\chi^2$-scores or –log *P* values). iGSEA4GWAS examines the enrichment of significantly associated variants within a priori-defined gene sets (pathways) by determining if a specific gene set ranks higher than a randomly distributed set, based on a running-sum statistic on the ranked list of genes (ranked by association *P* values or an equivalent statistic). The "improvement" in iGSEA4GWAS over traditional GSEA approaches is realized by focusing only on gene sets with high proportions of significant genes instead of relying solely on the overall gene set significance that may sometimes originate from only a few genes. In contrast, Pascal incorporates numerical and analytical solutions for *P* value estimation from the max statistics after adjusting for gene length and LD and adopts a modified Fisher's method (29) to compute parameter-free pathway scores (empirical or $\chi^2$-based *P* values) attaining rigorous type 1 error control. Pathways with both an iGSEA4GWAS-derived FDR $<$0.01 and a Pascal-derived $\chi^2$ $P <$ 0.05 were considered to be significantly enriched. Significantly enriched pathways identified in common between iGSEA4GWAS and Pascal were further analyzed via quantile-quantile plots in R (https://cran.r-project.org) to compare the observed distribution of pathway gene *P* values to the expected null. Hierarchical clustering of the significant pathways, based on their gene contents, was conducted via Ward's method in JMP Version 10 statistical package (SAS Institute, Cary, NC).

In addition to iGSEA4GWAS and Pascal, where pathway enrichment was computed on the full SNP-association data set, we also used the Data-driven Expression Prioritized Integration for Complex Traits (DEPICT; https://www.broadinstitute.org/depict) (68) tool for pathway analysis, using a prespecified list of 221 independent SNPs with intrinsic CRF association *P* values $<1 \times 10^{-04}$ (68). Independent loci were identified via PLINK (v1.90b3.42) (70) by clumping SNPs in LD ($r^2 > 0.1$) or $\leq$500 kb from the index SNP (LD calculations were based on 1000 Genomes Project Phase 1 Integrated Release V3 haplotypes). Additionally, DEPICT was used to predict most likely causal genes at associated loci and to identify target tissues where such genes are highly expressed.

*Weighted ranking of selected genes.* We developed a weighted rank-based approach to prioritize the candidate genes that were identified through the genetic, genomic, and bioinformatic analyses described above (evidence categories). The candidate genes were first ranked within each of the following evidence categories: strength of GWAS association, predictions from DEPICT, evidence for eQTL, overlap with histone marks or transcription factor binding sites, high connectivity in tissue networks, and correlation of gene expression to intrinsic CRF levels. Then, the relative influence of each evidence category was determined by the inverse of the coefficient of variation (CV) of the observed values in that category, such that categories with higher CVs (deemed to have larger information content than categories with smaller CVs) were accorded greater weight in the final ranking of the genes. We used the CV to make the weights comparable across all evidence categories. More specifically, to combine the ranks corresponding to a specific candidate gene ($G_i$), an inverse CV-weighted rank for gene G ($ICVWR_G$) metric was constructed as

$$ICVWR_{G_i} = \sum_{j=1}^{J} \frac{w_j}{\sum_{j=1}^{J} w_j} R_{ij},$$

where $R_{ij}$ is the rank of the *i*th gene in the *j*th evidence type, the weights, $w_j = 1/CV_j$, the coefficient of variation, $CV_j = \sigma_j/\mu_j$, $\mu_j$ is the mean, and $\sigma_j$ is the SD of the *j*th evidence type in their original scale (*P* values or some other measures) and $j = 1, ..., J$. In this sequence, an evidence type with the higher SD ($\sigma_j$) will give a higher $CV_j$ and a lower $w_j$, resulting in a lower value of $ICVWR_G$ when the mean ($\mu_j$) is fixed.

*Analysis of mouse phenotypes.* We queried the Mouse Genome Informatics (MGI; www.informatics.jax.org) database to identify and classify mouse phenotypes that are affected by knockout of candidate genes identified from the integrative bioinformatic analyses. A total of 38 genes were selected for this analysis, based on the extent of collective evidence for the following: *1*) evidence for joint eQTL and histone mark overlap at one or more SNPs near a gene, *2*) evidence for the gene being an eQTL target in any tissue tested, or *3*) DEPICT-based prioritization of the gene as associated with intrinsic CRF association ($P < 0.05$). Only mouse phenotypes arising from spontaneous mutations, targeted gene knockout, or gene trap studies were

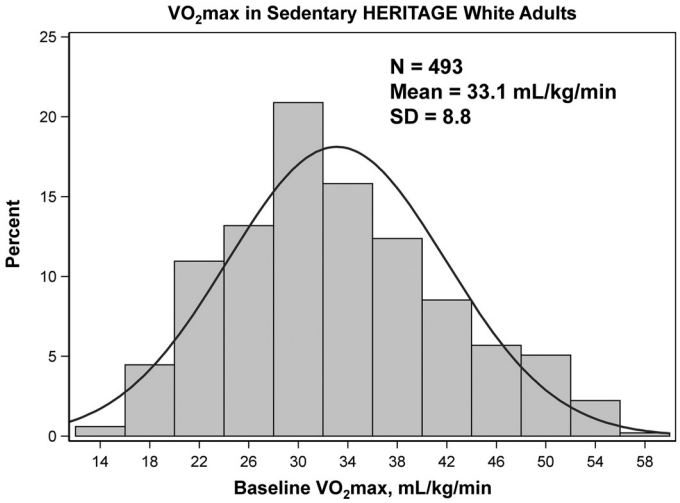


Fig. 2. Distribution of intrinsic $\dot{V}O_{2max}$ expressed by kg of body weight in 493 sedentary adults of both sexes of European ancestry. Maximal $O_2$ uptake was measured on a cycle ergometer on two different days. See text MATERIALS AND METHODS.

retrieved to avoid confounding effects from other mouse models (e.g., mutations in unrelated genes in chemical mutagenesis or multiple gene knockout type models). Phenotypes associated with the query genes were further grouped into their "root phenotypes" according to the hierarchical ontology utilized in the MGI V6.07 mammalian phenotype browser. For every gene, the percentage of phenotypes belonging to specific root phenotypes compared with all observed phenotypes was also estimated as a measure of root phenotype enrichment. As mouse data were not available for all candidate genes emerging from the prior analyses, we elected not to use the mouse knockout findings in our computation of the weighted ranking of genes as described above. Rather we interrogated MGI with the aim of extracting functional evidence in favor or against genes previously identified.

## RESULTS

*Study population and intrinsic CRF distribution.* A total of 493 sedentary white adults from the HERITAGE Family Study were available for the study; the mean intrinsic (sedentary) $\dot{V}O_{2max}$ was 33.1 ml $O_2 \cdot kg^{-1} \cdot min^{-1}$ with a SD of 8.8 $O_2 \cdot kg^{-1} \cdot min^{-1}$. The distribution of intrinsic $\dot{V}O_{2max}$ ranged from 14 to 58 ml $O_2 \cdot kg^{-1} \cdot min^{-1}$ based on two maximal tests, as displayed in Fig. 2. Some of the variance in intrinsic $\dot{V}O_{2max}$ can be accounted for by measurement error and day-to-day variability in the trait. Attempts to quantify these variance components in HERITAGE have been based on three strategies: $\dot{V}O_{2max}$ was measured twice at baseline in all subjects; 60 sedentary subjects meeting all HERITAGE inclusion criteria were tested 3 times over several weeks; and 8 subjects traveled across the 4 HERITAGE clinical sites to be tested within a 2-wk period (8, 77, 79). The intraclass coefficient for repeated $\dot{V}O_{2max}$ measures ranged from 0.96 to 0.98 across these 3 conditions, with CVs extending from 4.1 to 5.0%. The SD for repeated $\dot{V}O_{2max}$ measures ranged from 108 to 137 or ~1.4 to 1.8 ml $O_2 \cdot kg^{-1} \cdot min^{-1}$ (mean body weight of 75 kg), which represents ~15 to 20% of the variability (as the SD = 8.8) in intrinsic $\dot{V}O_{2max}$ as defined in HERITAGE.

*Steps in bioinformatic analyses.* The different components of the integrative bioinformatic analysis undertaken with the genetic association and gene expression data are outlined in Fig. 1. Briefly, our analytic strategy explored the impact of intrinsic CRF-associated sequence variation in influencing nearby gene expression (cis-eQTLs) or perturbations of transcription factor binding sites or chromatin marks associated with the noncoding, regulatory genome (e.g., histone binding sites at gene enhancers and promoters). Additionally, we examined the impact of sequence variation on biological mechanisms and tissue-specific functional interactions in the form of well-documented biological pathways and genome-scale tissue interaction networks, respectively. Finally, we investigated potential genotype-phenotype associations through an analysis of the functional consequences of candidate intrinsic CRF-associated genes in knockout mouse models and also interrogated the skeletal muscle specific transcriptomic association of candidate genes with intrinsic CRF levels. Thus, starting from SNP-level effects on the regulatory genome, we expanded our analysis to cellular events involving biological pathways, networks, and gene transcription and finally explored phenotypic implications at the organismal level.

*Analysis of effects from noncoding SNPs.* We interrogated data from Haploreg (v4.1) to identify SNPs that overlapped with enhancer and promoter regions in various relevant tissues, based on histone modification signals from the Roadmap Epigenomics and ENCODE projects (27a, 73). We considered signals arising from a subset of the intrinsic CRF-associated SNPs ($P \leq 1 \times 10^{-04}$), plus SNPs in high LD ($r^2 \geq 0.8$), in

Fig. 3. Effects of noncoding single-nucleotide polymorphisms (SNPs) on histone and transcription factor occupancy and cis-gene expression. *A*: enhancer-enrichment analysis across selected tissues, based on the overlap of intrinsic cardiorespiratory fitness (CRF)-associated SNPs with enhancer regions identified in Haploreg; the significance of enhancer enrichment is indicated by the negative logarithm of the binomial *P* value on the *y*-axis. *B*: distribution of the overlap of intrinsic CRF-associated SNPs (plus SNPs in linkage disequilibrium, $r^2 > 0.8$) with regions of active promoters and enhancers identified by modified histone binding across selected tissues. The four histone marks representing active promoter and enhancer elements are shown in columns and selected tissues in rows. Data are column normalized and color coded from blue (low overlap) to red (high overlap). Gray cells indicate absence of data from ENCODE. *C*: analysis of enrichment for intrinsic CRF-associated SNPs with genomic features via permutation testing in Genomic Association Tester software (GAT). The genomic features and tissues tested for feature overlap (where applicable) are represented on the *y*-axis. The *x*-axis displays the negative logarithm of the empirical *P* value observed from association testing based on 1,000 simulations. Points in the scatterplot are colored by the type of feature tested and sized by the fold change of observed vs. expected overlaps. *D*: top scoring expression quantitative trait loci (eQTL) across tissue categories. Genes with allele-dependent expression patterns are shown for 6 tissue categories which are aggregated over 16 different tissues. For each gene, the negative logarithm of the most significant regression *P* values obtained from its eQTL SNPs are plotted with deeper shades of red indicating greater significance. *E*: SNPs displaying joint eQTL and histone-mark overlap properties. SNPs with joint behavior in at least 1 tissue were selected. eQTL and histone-site overlap results are shown side-by-side for each tissue in columns. A positive association is indicated in gray. *F*: overlap of SNPs with transcription factor binding sites (TFBS) predicted by SNP2TFBS tool. The %change in position weight matrix scores relative to the reference allele is shown on the *x*-axis, and SNPs, along with their nearest genes and their genomic annotation, are indicated on the *y*-axis. Plot is restricted to SNPs that *1*) overlap a predicted TFBS with a high binding score ($P < 3 \times 10^{-06}$) in at least 1 of the two alleles, and *2*) overlap a modified histone binding site and/or function as an eQTL.

6 broad tissue types (adipose, brain, heart, lung, skeletal muscle, and pancreas) encompassing 16 different tissues; Supplemental Table S2). Out of a total of 1596 SNPs queried, a statistically significant enrichment (binomial $P \leq 0.01$) for enhancer-associated SNPs was observed in tissues specifically related to cardiorespiratory function such as aorta, right ventricle, and skeletal muscle (including myoblasts and satellite cells; Fig. 3*A*). Furthermore, as shown in Fig. 3*B*, we observed increased overlap with active enhancers (H3K4me1 and H3K27ac marks) and active promoters (H3K4me3 and

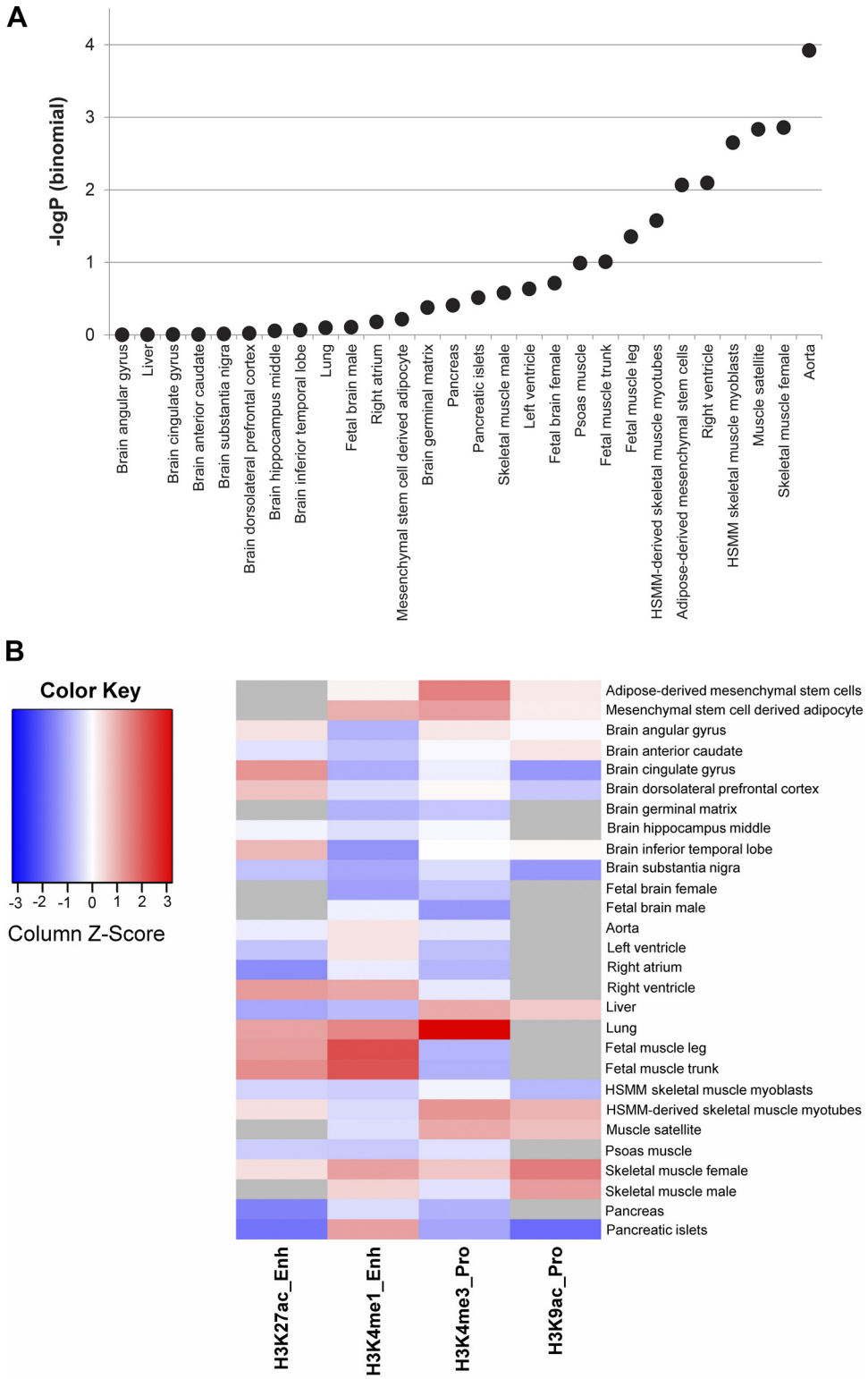


Fig. 3.

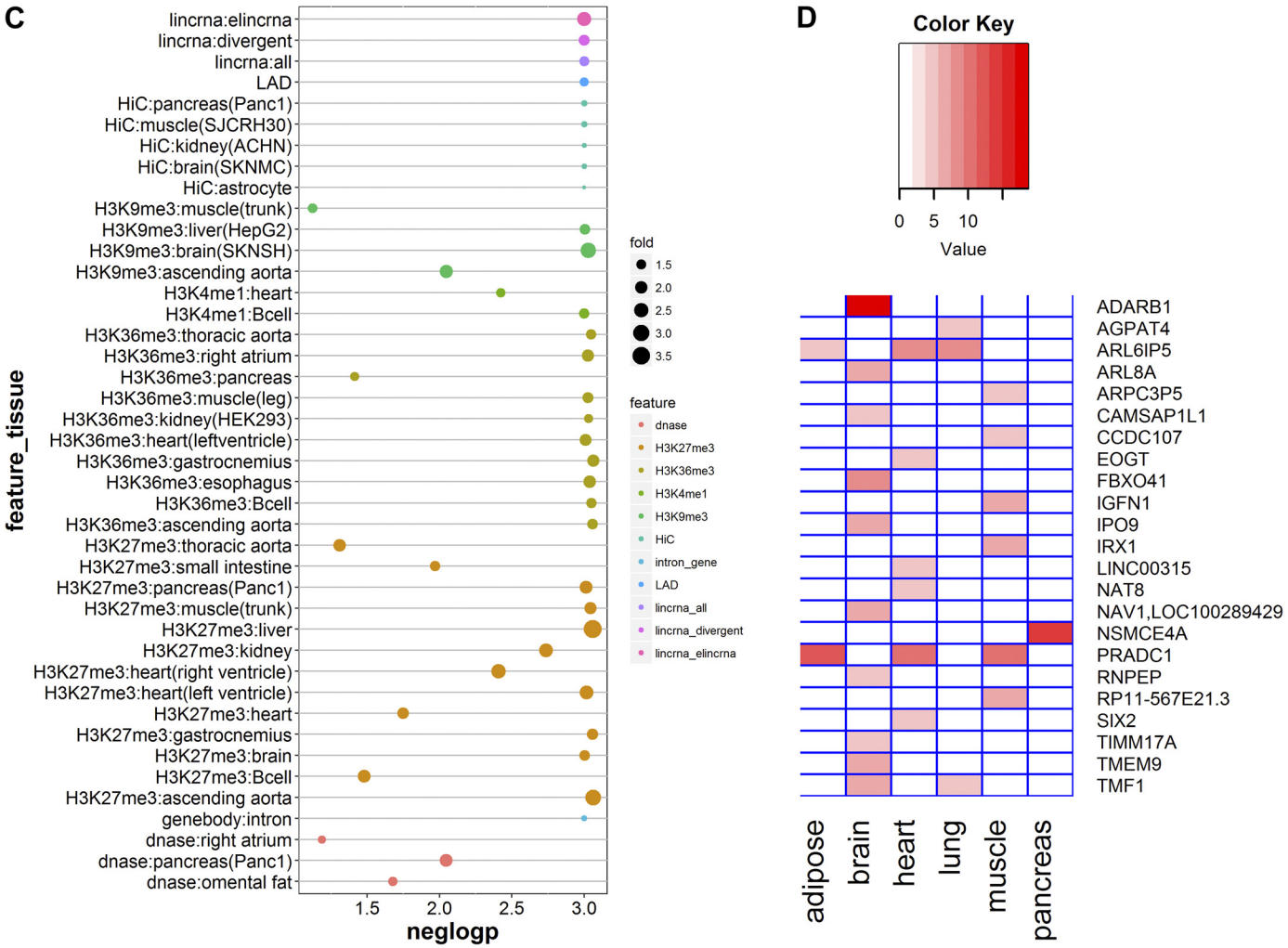


Fig. 3. *Continued*

H3K9ac marks) in multiple biological samples related to skeletal muscle (including myoblasts and myotubes) as well as increased levels of active enhancer associated H3K4me1-site overlaps in aorta. Adipose tissue generally demonstrated increased levels of overlap with enhancer and promoter sites, whereas sites overlapping with H3K4me1 were particularly underrepresented in a variety of brain sections tested. The enrichment of associated SNPs with enhancers was further corroborated by GAT analysis which showed a statistically significant overlap ($P < 8.7 \times 10^{-04}$ to $7.5 \times 10^{-02}$) of query SNPs in modified-histone (H3K9me3, H3K4me1, H3K36me3, and H3K27me3) binding regions of the genomes in tissues related to cardiorespiratory function including heart, ascending aorta, skeletal muscle, and others (Fig. 3*C*). Interestingly, the intrinsic CRF-associated SNPs were also enriched in other regions of the genome associated with introns ($P < 0.001$), lincRNAs ($P < 0.001$), DNAse hypersensitivity regions ($P < 9 \times 10^{-03}$ to $6.510^{-02}$), regions of long-range chromatin interactions ($P < 0.001$), and regions of lamina associated domains ($P < 0.001$) (additional details in Supplemental Table S3). Conversely, we did not observe statistical enrichment in transcription start sites, 5′/3′-untranslated regions, or DNA hypermethylation regions (FDR >0.2). These findings provide important insights into the roles played by elements of the noncoding regulatory genome in influencing intrinsic CRF.

We next interrogated cis-eQTL data from Haploreg (v4.1) to determine if a subset of the 1596 SNPs was also predicted to be eQTLs in relevant tissues. A total of 216 SNPs were significantly associated with cis-gene expression (eQTL $P < 1 \times 10^{-05}$). The top three most significant eQTL associations were observed for SNP rs2838815 on chromosome 21 influencing *ADARB1* gene expression in brain ($P < 1 \times 10^{-18}$), a cluster of highly linked SNPs on chromosome 10 (intronic to *ATE1*) regulating nonstructural maintenance of chromosome element 4 homolog A (*NSMCE4A*) gene expression in pancreas ($P < 1 \times 10^{-14}$), and another cluster of tightly linked SNPs on chromosome 2 affecting expression of the protease-associated domain containing 1 (*PRADC1*) gene ($P < 1 \times 10^{-12}$) in adipose, heart, and skeletal muscle. Figure 3*D* depicts all significant eQTLs ($P < 1 \times 10^{-05}$) observed across the tissue categories (additional SNP and tissue details are provided in Supplemental Table S4).

We next tested whether some of the genomic regions identified as harboring eQTLs were also coincident with regions overlapping histone binding sites (“joint” SNPs). Such an association would suggest the likelihood that SNP-dependent

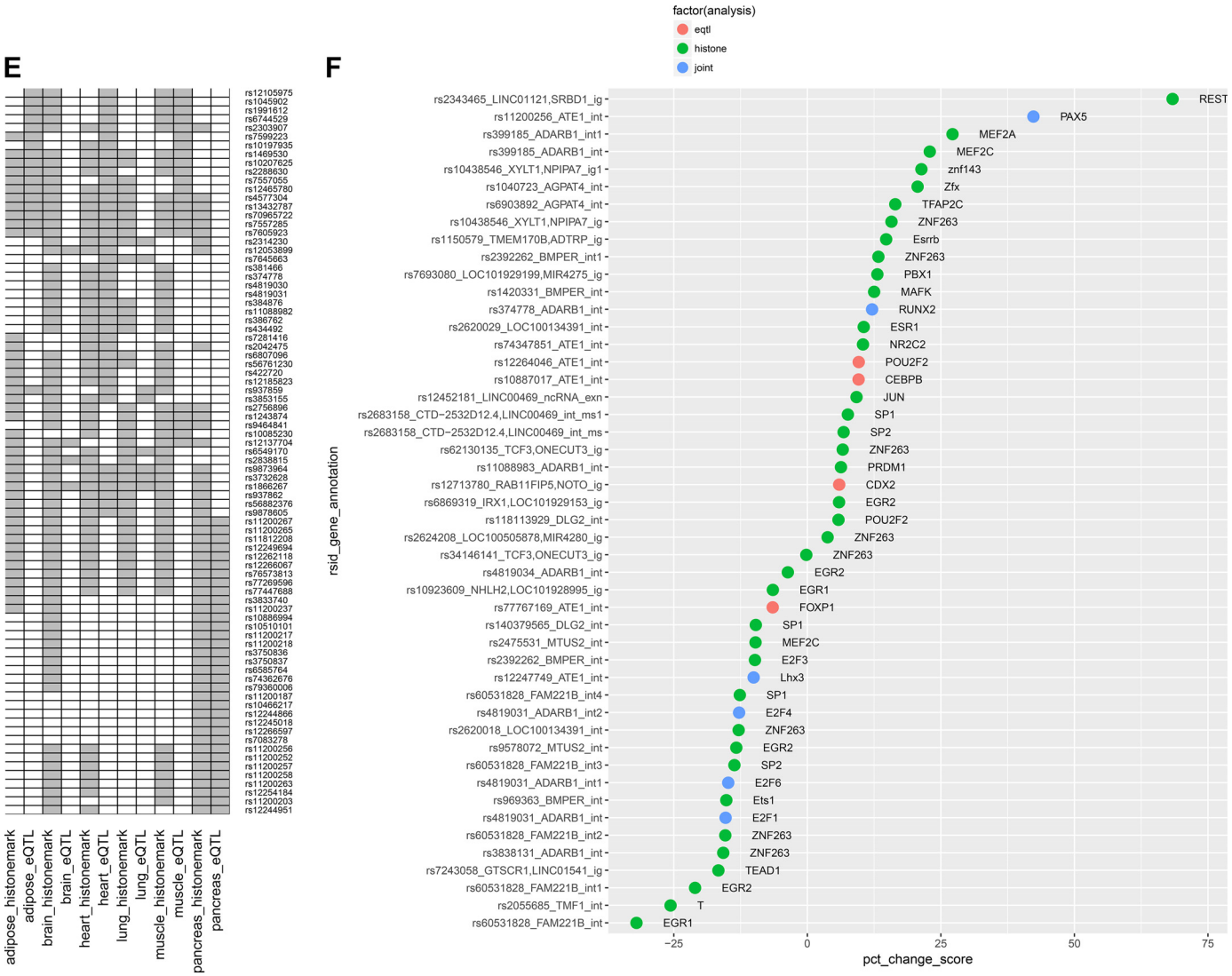


Fig. 3. *Continued*

alterations in histone binding are affecting enhancer/promoter activity and in turn affecting downstream gene expression and partly explain the mechanism behind SNP(s) from the same region functioning as eQTLs. The results for a subset of such "joint" SNPs are shown in Fig. 3*E* where eQTL and histone-binding site overlaps are compared across six tissue categories (full details in Supplemental Table S5). The largest overlap between histone-binding and eQTL SNPs was observed in heart tissue, followed by pancreas, muscle, and adipose, whereas the overlap was much lower for brain and lung.

We further interrogated the effects of noncoding SNPs on gene transcription by computing their predicted impact on transcription factor binding sites based on predicted allele-dependent changes in PWM scores. Several SNPs were predicted to significantly alter transcription factor binding based on large, significant changes in PWM scores between the SNP alleles (e.g., rs2343465 for *REST* and rs11200256 for *PAX5* binding) as summarized in Fig. 3*F*. We then investigated possible overlap among the previously identified eQTL or histone binding site overlapping SNPs and SNPs predicted to significantly alter transcription factor binding. Several instances of overlap between eQTL, histone binding, and transcription factor binding were observed (Supplemental Table S6). In several cases, the transcription factor binding motifs were identified in the first intron of the transcribed gene which is consistent with the observation that transcription regulatory elements appear to be enriched within the 5′-most introns (13).

*Analysis of tissue-specific networks.* Tissue-specific genome-scale gene networks were queried with a list of 186 genes nominally associated to intrinsic CRF ($P < 0.01$ by Pascal) with the goal of identifying relevant tissues based on overall connectivity of tissue-expressed genes to the set of GWAS-significant genes and to explore the subnetwork neighborhood for a subset of the highly connected genes in such tissues. The boxplots (Fig. 4*A*) depict the distribution of gene connectivity in 44 tissues based on the connectivity score (NetWAS score) obtained for each gene in each tissue (individual gene-level scores per tissue and boxplots for all 145 tissues are shown in Supplemental Table S7 and Supplemental Fig. S1, respectively). The top five tissues with high median network connectivity included placenta, heart, skeletal muscle,

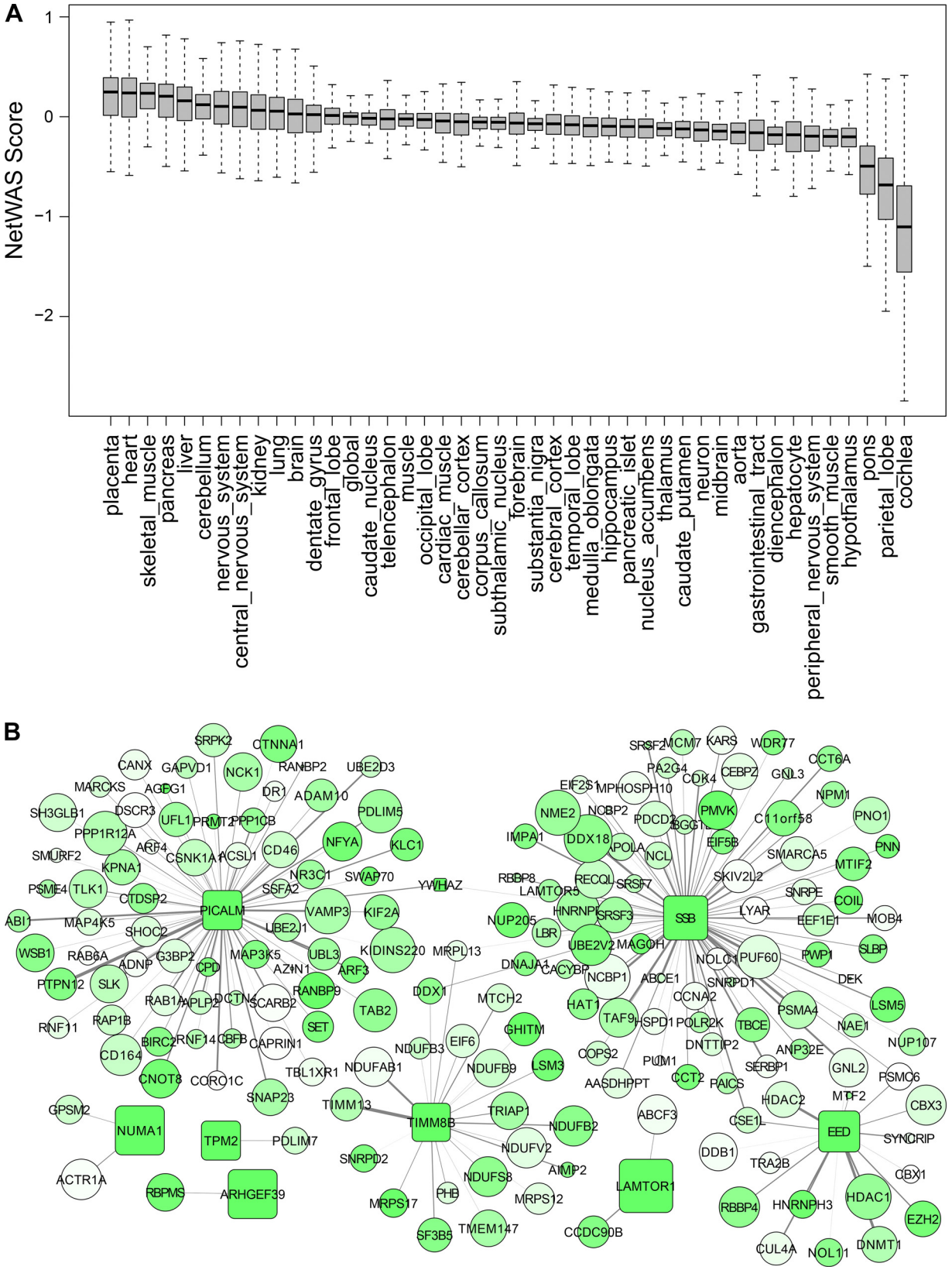


Fig. 4.

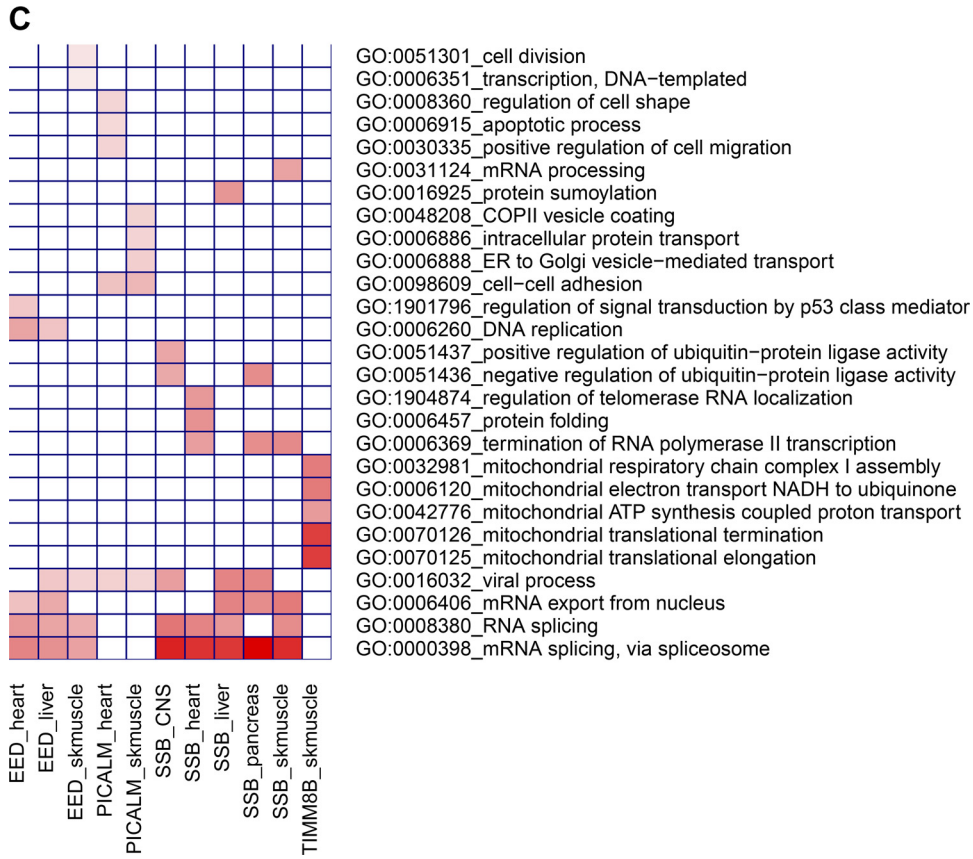


Fig. 4. *Continued*

pancreas, and liver. We further extracted subnetworks from these tissue networks to visualize the gene neighbors around some of the strongly connected intrinsic CRF-associated genes (NetWAS score $\geq 0.3$). The skeletal muscle subnetwork is shown in Fig. 4*B* and demonstrates a high degree of connectivity for the intrinsic CRF-associated genes *SSB*, *EED*, *PICALM*, and *TIMM8B* (hub genes). Some of these genes (*SSB* and *EED*) were also found in the subnetworks from other tissues (Supplemental Fig. S2) suggesting common mechanisms, whereas other genes were more tissue selective (e.g., *PICALM* primarily in heart and skeletal muscle, and *TIMM8B* in skeletal muscle). We next investigated whether any biological functions were overrepresented among the genes associated with these hub genes using the gene overrepresentation tool, DAVID (48) (Fig. 4*C* and Supplemental Table S8). Genes connected to *SSB* and *EED* were enriched for functions related to RNA biology (mRNA splicing, mRNA export, etc.), whereas *PICALM*-associated genes mostly functioned in cell adhesion and intracellular transport processes. Finally, the *TIMM8B* subnetwork was enriched for genes involved in mitochondrial functions including mitochondrial translation and mitochondrial electron transport.

*Pathway enrichment analysis.* Pathway enrichment analysis was conducted using iGSEA4GWAS and Pascal, and 34 pathways identified as significant by both methods (FDR $< 0.01$ for iGSEA, and $P_{\chi^2} < 0.05$ for Pascal) were further analyzed (the full list of significant pathways provided in Supplemental Table S9). Hierarchical clustering of the significant pathways based on gene content identified groups of pathways with shared genes, reflecting common biological processes (Fig.

Fig. 4. Analysis of genetic associations on genome-scale tissue networks. *A*: distribution of network connectivity to intrinsic cardiorespiratory fitness (CRF)-associated genes in tissue-specific gene networks. A NetWAS analysis was conducted to estimate the extent of connectivity of all genes to intrinsic CRF-associated genes (Pascal $P \leq 0.01$) in tissue-specific interactomes obtained from Genome-scale Integrated Analysis of gene Networks in Tissues (GIANT; giant.princeton.edu). The distribution of the connectivity scores (NetWAS score) for all genes across 44 selected tissues are shown as boxplots. Tissues with greater connectivities involving intrinsic CRF-associated genes tend to have higher median scores. *B*: analysis of a skeletal muscle sub-network centered on intrinsic CRF-associated genes. The top connected genes from the skeletal muscle NetWAS analysis (NetWAS score $\geq 0.3$) were extracted, and the network structure around the intrinsic CRF-associated genes were visualized. The genome-wide association study (GWAS)-associated genes are shown as boxes, whereas genes interacting with them (but not nominally GWAS-associated, $P > 0.01$) are shown as circles. Genes are color coded by the negative logarithm of their GWAS-association *P* values, with deeper shades of green indicating stronger associations. Additionally, the node size is proportional to the gene NetWAS score, and edge width is proportional to the posterior probability of network connectivity as determined in GIANT. *C*: pathway overrepresentation analysis among the skeletal muscle gene subnetworks shown in *B*. The sets of genes interacting with each GWAS-associated hub gene (*EED*, *SSB*, *PICALM*, and *TIMM8B*) were separately queried for enrichment of biological function via DAVID. The top 5 enriched Gene Ontology pathways among each hub gene neighbors are depicted as a heatmap. Tissue-specific hub gene subnetworks are indicated in columns and significant pathways in rows. Heatmap is color coded according to the negative logarithm of the significance of pathway enrichment (deeper red indicates greater significance).

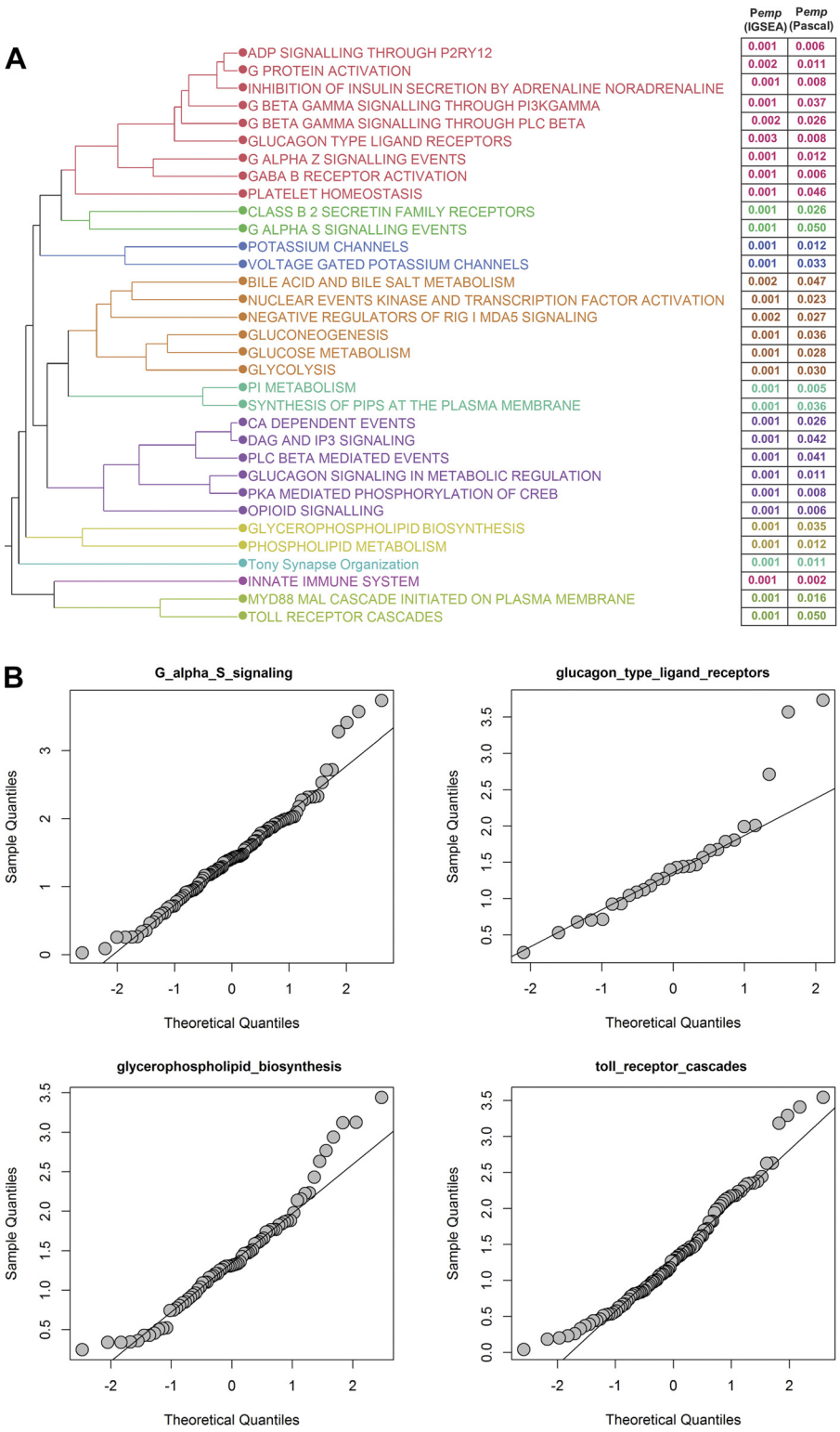


Fig. 5. Pathway enrichment analysis of intrinsic cardiorespiratory fitness (CRF)-associated genes. *A*: hierarchical clustering of common pathways found to be significantly enriched for intrinsic CRF-associated genes by iGSEA and Pascal. Pathways were clustered based on the shared number of significantly associated genes among pathways. Pathway significance levels from each tool (empirical *P* values) are indicated to the *right*. Pathways are color coded based on their similarities after cutting the dendrogram at 10 clusters. *B*: quantile-quantile plots for a subset of the significant pathways. For each pathway, the expected distribution of gene association *P* values are plotted on the *x*-axis and the observed distributions shown on the *y*-axis. Deviations from the diagonal indicate enrichment of significantly associated genes in a pathway.

5*A*). For example, the purple cluster in Fig. 5*A* refers to pathways involved in calcium and inositol 1,4,5-trisphosphate signaling, whereas the red cluster involves pathways involved in G-protein signaling. Additional pathway analysis conducted in DEPICT further identified Reactome-based pathways related to insulin and PI3 signaling as nominally significant ($P < 0.002$; Supplemental Table S10). The distribution of intrinsic CRF association *P* values among genes in the significant pathways was analyzed via quantile-quantile plots. SDs from the expected null, due to enrichment for small *P* values, were observed for a subset of the pathways, e.g., $G\alpha_s$ signaling, glycerophospholipid signaling and glucagon type II ligand receptor-associated pathways (Fig. 5*B*; full analysis details on all pathways presented in Supplemental Fig. S3).

*Analysis of knockout mouse phenotypes.* We summarized findings from all preceding bioinformatic analyses and arrived at a list of 38 genes that were identified by at least 1 of the analyses (Supplemental Table S11). These genes were further interrogated for their effects in knockout mouse models (detailed results in Supplemental Table S12). Of the 38 genes selected, knockout phenotypes were available for 21 genes. Figure 6 summarizes the distribution of the observed root-level mouse phenotypes among the gene knockouts. From this list, we focused on cardiovascular, hematopoietic, muscle, and metabolism related root phenotypes due to their greater relevance for intrinsic CRF. The individual phenotypes underlying each of these root phenotypes and the genes where they are observed are presented in Fig. 7. Notably, knockout of the *Casq2* gene affected all four root phenotype categories, with the majority of its observed phenotypes belonging to the cardiovascular system (60% of all reported *Casq2* phenotypes). Knockout of *Picalm* was associated mostly with hematopoietic phenotypes (62% of all reported phenotypes). *Sgcg* knockout produced effects on cardiovascular and muscle phenotypes (17.5 and 41.1% of all reported phenotypes, respectively), and *Pradc1* knockout led to only one specific metabolic effect (increased circulating glucose). Analysis of the relationship between the number of publications per gene and the number of knockout phenotypes ascribed to that gene did not show evidence for publication bias (Supplemental Fig. S4).

*Analysis of gene expression in skeletal muscle biopsies.* Whole genome expression profiling data of vastus lateralis muscle biopsies from a subset of 52 genotyped participants was analyzed to identify genes that were transcriptionally correlated with $\log_2$ intrinsic CRF levels, after adjustments for age, sex, BMI, and scan date. We identified 47 Affymetrix probes that were significant in analysis of covariance (ANCOVA) analysis ($P_{\text{ancova}} \leq 0.05$) and also significantly correlated to intrinsic CRF levels (absolute partial correlation $\geq 0.3$, $P_{\text{partialcorrel}} \leq 0.05$; Supplemental Table S13). In some instances, multiple probes were associated to the same gene; e.g., two probes corresponding to the glutamine transporter *SLC38A1* and two probes mapping to the eukaryotic translation initiation factor 5B (*EIF5B*) were among the top positively and negatively associated probes, respectively (Fig. 8). Notably, this list also included three genes (*CASQ2*, *COX7A2L*, and *PRADC1*) that were earlier identified as potentially relevant for intrinsic CRF through the integrative bioinformatic analyses.

*Ranking of genes based on cumulative evidence.* In Fig. 9, we have summarized our findings from the combined bioinformatic analysis and ranked the 38 candidate genes based on the cumulative evidence supporting their association with intrinsic CRF levels. The evidence categories included *1*) strength of genetic association in the GWAS; *2*) predictions of gene prioritization via DEPICT; *3*) evidence for eQTL function in relevant tissues; *4* and *5*) overlap with histone marks or transcription factor binding sites; *6*) connectivity to GWAS-associated genes in tissue networks, and *7*) correlation of gene expression to intrinsic CRF levels. Due to the large number of missing values in the knockout mouse phenotypes (data was available for only 21 of the 38 candidate genes), we did not include this information when calculating the weighted sum of ranks but analyzed it more qualitatively to identify candidate

| Gene | adipose tissue phenotype | behavior/neurological phenotype | cardiovascular system phenotype | cellular phenotype | craniofacial phenotype | digestive/alimentary phenotype | embryo phenotype | endocrine/exocrine gland phenotype | growth/size/body region phenotype | hearing/vestibular/ear phenotype | Hematopoietic system phenotype | homeostasis/metabolism phenotype | immune system phenotype | integument phenotype | limbs/digits/tail phenotype | Liver/biliary system phenotype | mortality/aging | muscle phenotype | neoplasm | nervous system phenotype | no abnormal phenotype detected | renal/urinary system phenotype | reproductive system phenotype | respiratory system phenotype | skeleton phenotype | vision/eye phenotype | TOTAL |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Ate1 | 5 | 6 | 13 | 3 | 1 | 1 | 1 | 1 | 4 | | | 5 | | 3 | | | 3 | | | 2 | | 1 | 4 | | 2 | | 55 |
| Picalm | | | | 1 | | | 3 | | 4 | | 28 | 4 | | 1 | | 1 | 3 | | | | | | | | | | 45 |
| Six2 | | | | | | | | | | | | | | | | | 3 | | | | 1 | 31 | | | | | 35 |
| Sgcg | | 6 | 6 | 2 | | | | | 2 | | | 1 | | | | | 2 | 14 | | | | | | | 1 | | 34 |
| Dmrt2 | | | | | | | 1 | | | | | | | | 3 | | 1 | 3 | | | | | | 3 | 14 | | 25 |
| Adarb1 | | 5 | | | | | | | 1 | | | | | | 1 | | 3 | 3 | | 9 | 1 | | | | | | 23 |
| Casq2 | 1 | | 14 | | | | 1 | | 1 | | 1 | 1 | | | | | 1 | 2 | | | | | | | | 1 | 23 |
| Fbxo41 | | 8 | | | | | | | 3 | 1 | | | | | | | 2 | | | 4 | | | | | | | 18 |
| Ssb | | | | | | | 3 | | 1 | | 6 | | | | | | 1 | | | 7 | | | | | | | 18 |
| Noto | | | 1 | 3 | | | 5 | | | | | | | | 5 | | 1 | | | | | | | | 1 | | 16 |
| Arl6ip5 | | 4 | | 1 | | | | | 1 | | | 1 | 1 | 1 | | | | | 1 | | 1 | | | | 2 | | 13 |
| Gba2 | | | | 4 | | | | 3 | | | | 3 | | | | | | | | | | | 3 | | | | 13 |
| Car9 | | | | | | 9 | | | | | 1 | | | | | | 1 | | | | | | | | | | 11 |
| Tmf1 | | | | 5 | | | | 1 | | | | | | | | | | | | | | | 4 | | | | 10 |
| Uba3 | | | | 3 | | | 7 | | | | | | | | | | | | | | | | | | | | 10 |
| Rab11fip5 | 1 | 2 | | | | | | | 2 | | | 3 | | | | | | | | | 1 | | | | | | 9 |
| Cables1 | | | | | | | | 2 | | | | | | | | | | | | | | | 6 | | | | 8 |
| Lamtor1 | | | | 1 | | | 4 | | | | | | | | | | | | | | | | | | | | 5 |
| Onecut3 | | | | | | | | 3 | | | | | | | | | | | | | | | | | | | 3 |
| Pradc1 | | | | | | | | | | | | 1 | | | | | | | | | | | | | | | 1 |
| Smyd5 | | 1 | | | | | | | | | | | | | | | | | | | | | | | | | 1 |

Fig. 6. *Column 1*, gene symbol; *columns 2–27*, different root phenotypes; *column 28*, total number of root phenotypes observed for each gene. For any root phenotype, number of observed subphenotypes is indicated in individual cells. Cardiovascular, hematopoietic, metabolic, and muscle-related root phenotypes are highlighted in gray.

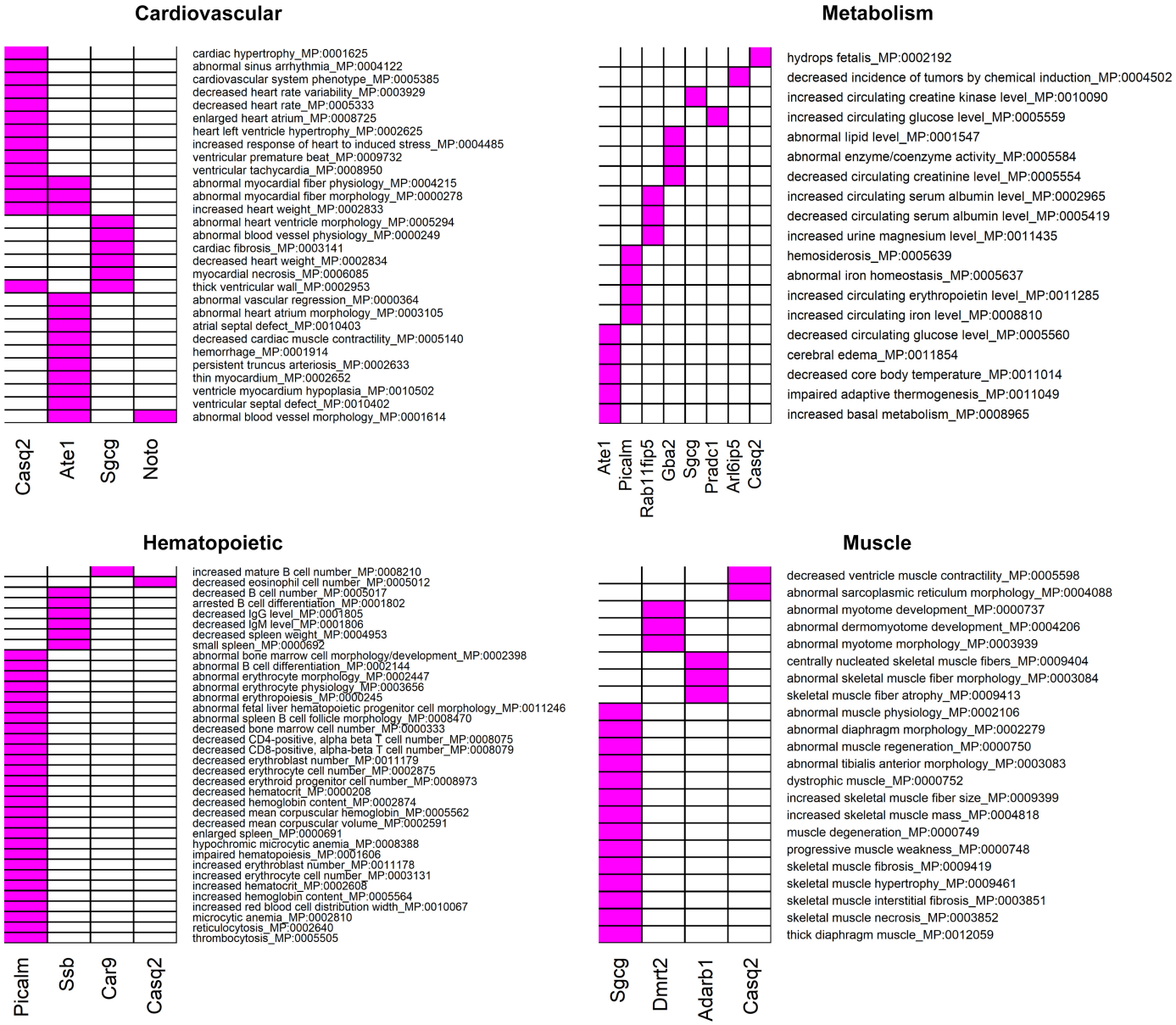


Fig. 7. Effect of candidate gene knockouts on phenotypes potentially relevant to cardiorespiratory fitness. Candidate intrinsic CRF-associated genes were used to query the Mouse Genome Informatics (MGI) database for phenotypes arising from targeted gene knockouts or gene trap models. Heatmaps show the reported individual phenotypes under four root phenotype categories (cardiovascular, hematopoietic, metabolic, and muscle) in gene knockouts. Phenotypes are indicated in rows and genes in columns. In each heatmap, genes displaying at least one knockout phenotype are considered. The presence of an association between a gene and a phenotype is indicated in magenta.

genes with relevant mouse phenotypes. In the bioinformatic analysis, genes such as *ADARB1*, *CASQ2*, *IGFN1*, and *PRADC1* displayed evidence over multiple categories, whereas the category evidence for genes such as *PICALM*, *NSMCE4A*, and *SLC38A1* was more restricted. However, as the evidence categories are unlikely to have an equal influence in the association of a gene to intrinsic CRF, we weighted the categories by the inverse CV of the observed values in each category and used this information to compute a weighted sum of ranks for each gene based on the $ICVWR_G$ metric described in MATERIALS AND METHODS. The *ADARB1* and *PRADC1* genes scored high on the bioinformatic analysis and demonstrated phenotypic evidence for muscle and metabolism, respectively, from MGI data. Notably, the *CASQ2* gene was also ranked high in the bioinformatic explorations (rank of 7) and demonstrated a positive phenotype in all 4 root phenotype classes, with 14/23 phenotypes related to cardiovascular function (Fig. 6). Brief descriptions of the biological functions of the genes listed in Fig. 9 are described in Supplemental Table S14.

*Association of genetic variants in identified candidate genes with traits related to underlying physiology of intrinsic CRF.* The combined bioinformatic analysis found suggestive evidence for four gene loci related to cardiovascular physiology (*ATE1*, *CASQ2*, *NOTO*, and *SGCG*) and four loci linked to skeletal muscle phenotypes (*SGCG*, *DMRT2*, *ADARB1*, and *CASQ2*). To further explore the link between these loci and the major physiological determinants of $\dot{V}O_{2max}$ (i.e., central and

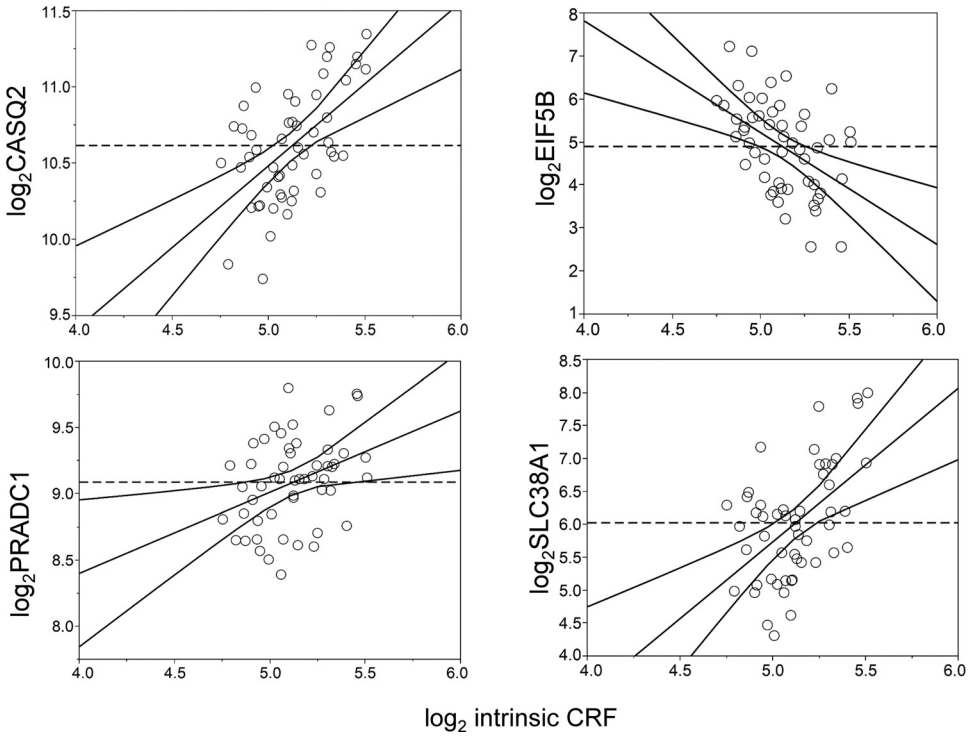


Fig. 8. Association of gene expression with intrinsic cardiorespiratory fitness (CRF) levels in vastus lateralis muscle biopsies in a subset of the HERITAGE cohort. Partial regression residual leverage plots based on partial correlations of gene expression (*y*-axis) to intrinsic CRF levels (*x*-axis) after adjustments for age, sex, body mass index, and scan date are shown for four genes (*CASQ2*, *EIF5B*, *PRADC1*, and *SLC38A1*). Both gene expression and intrinsic CRF are plotted in the $\log_2$ scale for ease of interpretation. The dashed horizontal line corresponds to a gene expression partial residual = 0 and represents the model where the hypothesized value of gene expression is constrained to 0. The least squares line through the plotted points and its 95% confidence curves are shown. Significant effects are indicated when the confidence curve crosses the horizontal line.

muscle components), we examined the association of SNPs located in and near *ATE1*, *CASQ2*, *NOTO*, and *SGCG* with cardiovascular phenotypes and SNPs in or near *SGCG*, *DMRT2*, *ADARB1*, and *CASQ2* with muscle-related phenotypes. Furthermore, we examined the association of skeletal muscle gene expression of *SGCG*, *DMRT2*, *ADARB1*, and *CASQ2* with muscle-related phenotypes.

Multiple SNPs in *ATE1*, *CASQ2*, *NOTO*, and *SGCG* were nominally associated ($P < 0.05$) with cardiovascular traits measured at rest and during submaximal and maximal exercise, including heart rate, stroke volume, cardiac output, and systolic blood pressure (Supplemental Table S15). Similarly, multiple SNPs (Supplemental Table S15) and gene expression (Supplemental Table S16) of *SGCG*, *DMRT2*, *ADARB1*, and *CASQ2* were nominally associated with muscle-related traits, including percent fiber type, capillary area per fiber, and muscle enzyme activities. For example, *CASQ2* SNP rs7523715 was associated with cardiac output and heart rate during submaximal and maximal exercise, respectively (Fig. 10*A*). Similarly, *CASQ2* SNP rs2999460 was associated with the percentage of type 1 (Fig. 10*A*, *left*) and type 2b muscle fibers (middle panel) and hydroxyacylcoA-dehydrogenase activity (Fig. 10*A*, *right*), with TT homozygotes having significantly higher proportions of type 1 and lower proportions of type 2b fibers and higher HADH levels (Fig. 10*B*). *CASQ2* gene expression was positively associated with the percentage, area, and capillarization of type 1 fibers and negatively associated with the percentage and area of type 2b fibers (Fig. 11*A*). Moreover, *CASQ2* gene expression was positively associated with enzyme activities related to aerobic cellular respiration (cytochrome oxidase, citrate synthase, and HADH) and negatively associated with the rate-limiting enzyme of glycolysis (phosphofructokinase) (Fig. 11*B*).

## DISCUSSION

A large body of data from the exercise physiology literature supports the conclusion that heart size, stroke volume, cardiac output, blood volume, and total hemoglobin content are the most critical determinants of maximal $O_2$ uptake or CRF (56, 58). We assumed that this also applies to intrinsic CRF in completely sedentary individuals. Even though there is some clarity in the physiology underlying intrinsic CRF, there is a gap in our understanding of the genetic and molecular determinants that underlie and regulate the observed physiology. In this study, we have sought to address this gap by exploring the potential genetic regulation of intrinsic CRF through integration of summary level GWAS association and muscle gene expression data with an integrative analysis of functional information relating to biological pathways, tissue-specific networks, noncoding genome regulation, and knockout mouse phenotypes.

To summarize the key findings from our study and put them into perspective, we begin with an examination of the impact of intrinsic CRF-associated genetic variants on the noncoding, regulatory genome. As the majority of trait-associated SNPs are located in the noncoding genome, exploratory analyses have been conducted in prior publications to examine the overlap of genome-wide significant (and LD-associated) variants with specific regulatory signatures such as DNAse 1 hypersensitivity sites, promoter-associated *H3K4me3/H3K9ac*, and enhancer-associated *H3K4me1/H3K27ac* marks in data generated through the ENCODE consortium and Roadmap Epigenomics Project (3). An examination of 426 GWAS data sets further demonstrated that disease-associated variants are significantly more likely to overlap with regulatory domains such as strong enhancers (28). Notably, analysis of the current

| | Genetic Evidence | | | | | | | | | | | Phenotypic Evidence | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Candidate Gene | GWAS assn. ($P_{Pascal}$) | DEPICT gene prioritization ($P_{Depict}$) | eQTL association ($P_{besteQTL}$) | H3K4me1 binding (DiffProb) | H3K4me3 binding (DiffProb) | H3K9ac binding (DiffProb) | H3K27ac binding (DiffProb) | TF binding (Percent change$_{TFBS}$) | Tissue Network (max NetWAS score) | Gene Expression ($P_{regression}$) | Weighted rank (ICVWR$_G$) | Cardiovascular | Hematopoietic | Muscle | Metabolism | Other | Unknown |
| ADARB1 | 0.003 | 0.49 | 1.90E-19 | 0.02 | 0.01 | 0.01 | 0.01 | 38.5 | 0.69 | 0.04 | 1 | | | | | | |
| PRADC1 | 0.001 | 0.05 | 7.87E-13 | 0.01 | 0.00 | 0.00 | 0.00 | 0.0 | 0.45 | 0.03 | 2 | | | | | | |
| TPM2 | 0.002 | 0.01 | 2.00E-05 | 0.07 | 0.03 | 0.04 | 0.04 | 0.0 | 0.37 | 0.24 | 3 | | | | | | |
| FBXO41 | 0.001 | 0.7 | 4.19E-09 | 0.02 | 0.01 | 0.01 | 0.01 | 0.0 | 0.39 | 0.13 | 4 | | | | | | |
| IGFN1 | 0.002 | 0.01 | 9.05E-07 | 0.01 | 0.00 | 0.00 | 0.00 | 2.0 | 0.34 | 0.04 | 5 | | | | | | |
| NOTO | 0.001 | 0.99 | 2.00E-05 | 0.02 | 0.03 | 0.01 | 0.01 | 3.4 | 0.44 | NA | 6 | | | | | | |
| CASQ2 | 0.001 | 0.04 | 2.00E-05 | 0.03 | 0.00 | 0.00 | 0.01 | 3.8 | 0.21 | 3.10E-05 | 7 | | | | | | |
| ARL6IP5 | 1.8E-04 | 0.99 | 8.00E-09 | 0.05 | 0.04 | 0.07 | 0.07 | 0.0 | 0.38 | 0.73 | 8 | | | | | | |
| RAB11FIP5 | 0.010 | 0.08 | 2.00E-05 | 0.02 | 0.02 | 0.03 | 0.02 | 17.1 | 0.50 | 0.38 | 9 | | | | | | |
| TMF1 | 0.000 | 0.9 | 6.75E-07 | 0.01 | 0.00 | 0.01 | 0.01 | 25.6 | 0.32 | 0.08 | 10 | | | | | | |
| LAMTOR1 | 0.006 | 0.99 | 2.00E-05 | 0.00 | 0.00 | 0.00 | 0.00 | 0.0 | 1.03 | 0.03 | 11 | | | | | | |
| NSMCE4A | 0.043 | 0.94 | 1.80E-15 | 0.00 | 0.00 | 0.00 | 0.00 | 0.0 | 0.48 | 0.23 | 12 | | | | | | |
| ATE1 | 0.006 | 0.79 | 2.00E-05 | 0.02 | 0.02 | 0.01 | 0.02 | 42.3 | 0.29 | 0.25 | 13 | | | | | | |
| CA9 | 0.001 | 0.03 | 2.00E-05 | 0.00 | 0.00 | 0.00 | 0.00 | 0.0 | 0.48 | 0.56 | 14 | | | | | | |
| SSB | 0.003 | 0.95 | 2.00E-05 | 0.00 | 0.00 | 0.00 | 0.00 | 0.0 | 0.63 | 0.08 | 15 | | | | | | |
| EOGT | 1.9E-04 | 0.99 | 7.12E-06 | 0.04 | 0.01 | 0.03 | 0.02 | 2.1 | 0.11 | 0.17 | 16 | | | | | | |
| ONECUT3 | 0.003 | 0.05 | 2.00E-05 | 0.03 | 0.03 | 0.03 | 0.03 | 0.0 | 0.08 | 0.31 | 17 | | | | | | |
| SMYD5 | 0.001 | 0.85 | 2.00E-05 | 0.02 | 0.01 | 0.01 | 0.02 | 5.5 | 0.28 | 0.35 | 18 | | | | | | |
| IPO9 | 0.608 | 0.99 | 8.32E-07 | 0.00 | 0.00 | 0.00 | 0.00 | 0.0 | 0.43 | 0.07 | 19 | | | | | | |
| CABLES1 | 0.004 | 0.02 | 2.00E-05 | 0.02 | 0.01 | 0.01 | 0.01 | 0.0 | 0.06 | 0.11 | 20 | | | | | | |
| SGCG | 0.265 | 0.02 | 2.00E-05 | 0.00 | 0.00 | 0.00 | 0.00 | 0.0 | 0.50 | 0.78 | 21 | | | | | | |
| COX7A2L | 0.005 | 0.99 | 2.00E-05 | 0.00 | 0.00 | 0.00 | 0.00 | 0.0 | 0.37 | 0.001 | 22 | | | | | | |
| TMEM9 | 0.002 | 0.99 | 6.33E-08 | 0.00 | 0.00 | 0.00 | 0.00 | 0.0 | 0.38 | 0.32 | 23 | | | | | | |
| GBA2 | 0.001 | 0.15 | 2.00E-05 | 0.00 | 0.00 | 0.00 | 0.00 | 0.0 | 0.36 | 0.52 | 24 | | | | | | |
| PICALM | 0.005 | 0.99 | 2.00E-05 | 0.00 | 0.00 | 0.00 | 0.00 | 0.0 | 0.62 | 0.51 | 25 | | | | | | |
| DMRT2 | 0.221 | 0.04 | 2.00E-05 | 0.00 | 0.00 | 0.00 | 0.00 | 0.3 | 0.10 | 0.05 | 26 | | | | | | |
| TIMM8B | 0.002 | 0.99 | 2.00E-05 | 0.00 | 0.00 | 0.00 | 0.00 | 0.0 | 0.32 | 0.07 | 27 | | | | | | |
| UBA3 | 1.1E-04 | 0.84 | 2.00E-05 | 0.01 | 0.01 | 0.00 | 0.01 | 0.0 | 0.26 | 0.54 | 28 | | | | | | |
| NAV1 | 0.683 | 0.99 | 9.82E-06 | 0.00 | 0.00 | 0.00 | 0.00 | 0.0 | 0.40 | 0.15 | 29 | | | | | | |
| CCT7 | 0.001 | 0.86 | 2.00E-05 | 0.04 | 0.01 | 0.01 | 0.01 | 0.0 | 0.11 | 0.73 | 30 | | | | | | |
| EIF5B | 0.200 | 0.99 | 2.00E-05 | 0.00 | 0.00 | 0.00 | 0.00 | 0.0 | 0.12 | 0.002 | 31 | | | | | | |
| CCDC107 | 0.002 | 0.05 | 4.23E-06 | 0.00 | 0.00 | 0.00 | 0.00 | 0.0 | 0.09 | 0.84 | 32 | | | | | | |
| SLC38A1 | 0.877 | 0.99 | 2.00E-05 | 0.00 | 0.00 | 0.00 | 0.00 | 0.0 | -0.10 | 2.00E-04 | 33 | | | | | | |
| ARL8A | 0.181 | 0.99 | 6.42E-08 | 0.00 | 0.00 | 0.00 | 0.00 | 0.0 | 0.18 | 0.67 | 34 | | | | | | |
| TIMM17A | 0.438 | 0.99 | 2.77E-06 | 0.00 | 0.00 | 0.00 | 0.00 | 0.0 | 0.21 | 0.58 | 35 | | | | | | |
| RNPEP | 0.392 | 0.99 | 3.08E-06 | 0.00 | 0.00 | 0.00 | 0.00 | 0.0 | 0.24 | 0.74 | 36 | | | | | | |
| CAMSAP2 | 0.746 | 0.99 | 3.60E-06 | 0.00 | 0.00 | 0.00 | 0.00 | 0.0 | -0.11 | 0.17 | 37 | | | | | | |
| SIX2 | 0.371 | 0.99 | 1.25E-05 | 0.00 | 0.00 | 0.00 | 0.00 | 2.5 | -0.01 | 0.67 | 38 | | | | | | |

Fig. 9. Summary of bioinformatic and phenotypic analysis of intrinsic cardiorespiratory fitness (CRF) associated candidate genes. Genes are listed in rows and the various genetic and phenotypic evidence categories are listed in columns. Values in each column are derived from genetic or bioinformatic analysis. *Column 1*, gene name; *column 2*, gene-level genome-wide association study (GWAS) association *P* value from Pascal analysis; *column 3*, Data-driven Expression Prioritized Integration for Complex Traits (DEPICT)-predicted gene prioritization *P* value; *column 4*, best expression quantitative trait loci (eQTL) association *P* value observed for gene in any tissue tested; *columns 5–8*, DeepSea predicted difference in probabilities for major modified-histone binding (H3K4me1, H3K4me3, H3K9ac, and H3K27ac) between reference and alternate alleles; *column 9*, transcription factor binding sites (TFBS) predicted max. percent change in allele-dependent position weight matrix scores for any transcription factor; *column 10*, max. observed NetWAS score for gene across tissues; *column 11*, regression *P* value for muscle gene expression changes with intrinsic CRF for a subset of HERITAGE cohort (no Affymetrix probe mapped to NOTO gene); *Column 12*, relative ranking of genes based on ICVWRG method described in the text; and *columns 13–18*, presence of a root phenotype effect for gene knockout in mouse models. CRF relevant root phenotypes are shown in red, other phenotypes in blue, and absence of mouse phenotype data is shown in green.

data set also indicated significant enrichment of active enhancer regions in heart and skeletal muscle related tissues, providing a possible basis for regulation of relevant tissue function via the noncoding genome. Additionally, significant overlap of the query SNPs was also observed for other regulatory genomic features such as several modified-histone binding sites, lincRNA coding regions, long-range chromatin interaction regions, and nuclear lamina-associated chromosome domains. Although not further explored in this study, these findings provide intriguing leads into future studies to probe the consequences of genetic variation in these regulatory domains for the control of intrinsic CRF.

Several recent studies have also indicated that trait-associated SNPs are often enriched for loci that alter the expression of nearby genes, possibly via their impact on the binding of transcription regulators at gene promoters and enhancers (39b, 40, 75). Known as cis-eQTLs, the association of sequence variation with proximal gene expression may often provide a functional interpretation for the associated variants that could be biologically more meaningful than the conventional assumption of a SNP impacting its most proximally located gene (64). By comparing GWAS-associated SNP variants with eQTL data from published sources, we observed strong eQTL predictions for several genes, notably *PRADC1* (predicted by a cluster of SNPs in and near the gene on Chr 2, eQTL association *P* values ranging from $1.6 \times 10^{-07}$ to $1.3 \times 10^{-11}$) in heart, adipose, and skeletal muscle, and a tightly linked cluster of SNPs intronic to the *ATE1* gene predicted as eQTLs for *NSMCE4A* specifically in the pancreas (eQTL *P* values between $2.5 \times 10^{-08}$ to $6.3 \times 10^{-15}$). The roles of either of these genes in intrinsic CRF are currently unknown; *PRADC1* (also known as *HPAP21*) is a secreted glycoprotein with no reported function (93), whereas *NSMCE4A* is a component of the SMC5-SMC6 complex (82) involved in homologous recombination and telomere maintenance. Interestingly, a recent publication studying microdeletion at 10q26.1 reported on *ATE1* and *NSMCE4A* as candidate genes for heart defects and growth cessation, among other phenotypes (14). We should point out,

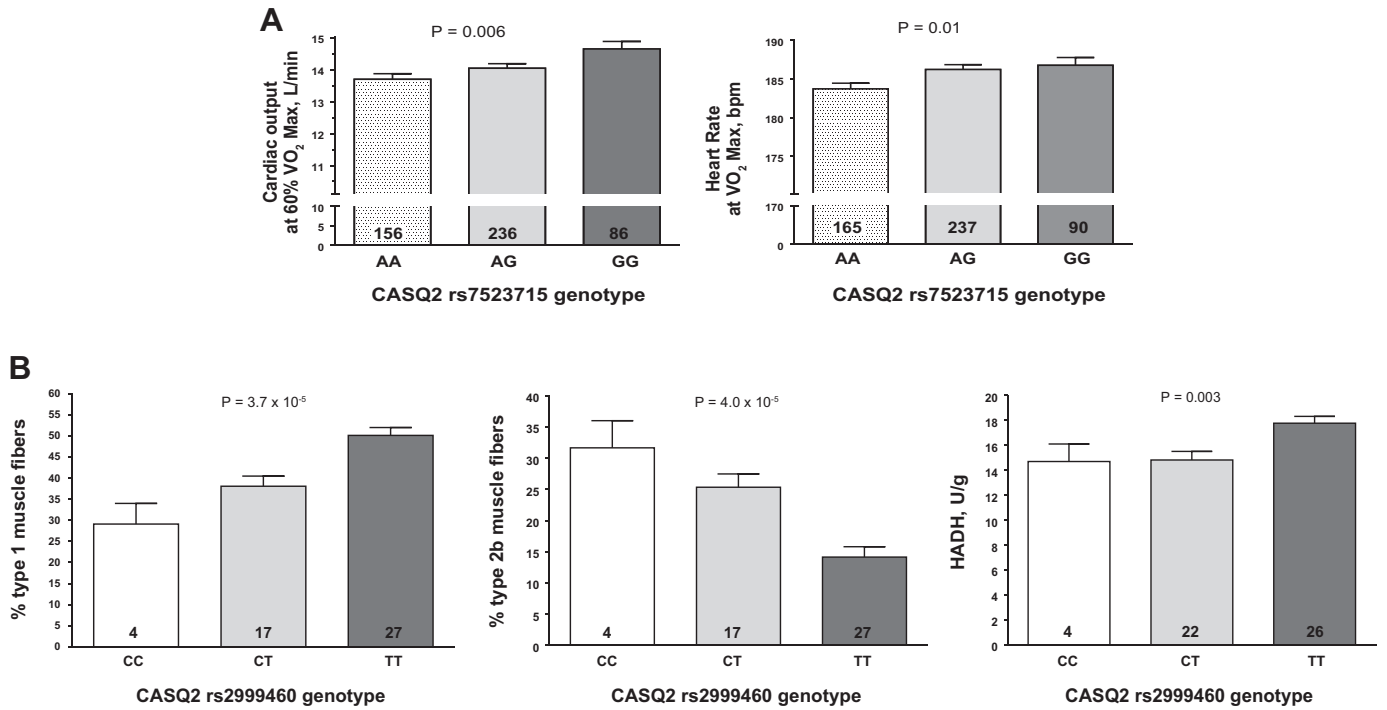


Fig. 10. *A*: association of *CASQ2* single-nucleotide polymorphism (SNP) rs7523715 genotype with cardiovascular traits measured during submaximal (*left*) and maximal (*right*) exercise. *B*: association of *CASQ2* SNP rs2999460 genotype with muscle-related traits measured in a subset of the HERITAGE cohort. Adjusted mean trait values shown for each genotype after adjustment for age, sex, and body mass index. *P* value for the main effect of genotype on each trait is shown at the top of each graph. Number of subjects with each genotype is indicated inside each histogram bar.

however, that the eQTL associations do not shed light on the causality of the SNPs altering gene expression; i.e., the predicted eQTL SNPs may be tagging other SNPs via strong LD that are the bona fide causal regulators of gene expression.

We further reasoned that the altered binding of regulatory proteins (e.g., modified histones and transcription factors) in active promoters and enhancers can offer a possible mechanism of function for an eQTL SNP, for example, when such histone/transcription factor binding sites overlap with eQTL sites. Thus we investigated joint associations of eQTL and histone/transcription factor binding motifs to narrow the list of potentially functional SNPs in CRF relevant tissues. Notably, the SNP clusters that were eQTLs for *PRADC1* and *NSMCE4A* also overlapped with modified histones tagging promoter and enhancer elements. Gene expression analysis on a subset of the genotyped cohort further suggested a significant correlation of expression for genes such as *PRADC1*, *CASQ2*, *DMRT2*, *IGFN1*, and *SLC38A1* with intrinsic CRF levels (adjusted for age, sex, BMI, and chip scan date).

While the above analyses elucidated the impact of individual SNPs and genes, we undertook a parallel investigation to identify the effect of genetic variation on sets of genes representing biological processes and tissue-specific interactomes. By definition, pathway enrichment analysis is not so much dependent on high genetic association signals from individual genes, but evidence is instead accumulated over multiple genes, even with individually small to modest effect sizes, that operate within a pathway. Our analysis identified 34 Reactome-based pathways that were significantly enriched with intrinsic CRF-associated variants. These pathways could be further categorized into broader categories including pathways related to calcium and inositol-phosphate signaling, G-protein activation and signaling, and glucose and phospholipid metabolism.

Tissue-specific network analysis provides valuable complementary insights into the biological pathways impacted by sequence variation. Whereas pathway enrichment analysis is focused on function (but not tissue-specific gene expression), tissue networks largely inform on tissue-relevant connectivities based on gene coexpression and protein-protein interactions. Analysis of tissue-specific networks in our study identified the heart and skeletal muscle as two top tissues where gene networks were enriched for gene variants associated with intrinsic CRF. The convergence between the findings from pathway analysis and network analysis (e.g., calcium and inositol 1,4,5-trisphosphate signaling as top pathways from pathway analysis and heart and skeletal muscle as top tissues in network analysis) provides an opportunity for more refined hypotheses regarding biological processes and the site of their action in the regulation of CRF.

Results from knockout animal models provide a starting point for evaluation of evidence for convergence (or divergence) of genotypic and phenotypic findings. We queried the MGI database (www.informatics.jax.org) to identify and classify intrinsic CRF-relevant mouse phenotypes that are affected by knockout of candidate genes identified through the integrative bioinformatic approach. Knockout phenotype information was available for a subset of the queried genes and was further classified according to the root phenotypes as described in the MGI phenotype browser. From this list, we focused on genes that displayed phenotypes related to one or more of cardiovascular, hematopoietic, muscular, or metabolic traits. This anal-

ysis identified the *CASQ2* (calsequestrin) gene with phenotypes in each of the four categories, most notable with effects on cardiac muscle contractility, increased heart weight, cardiac hypertrophy, sinus arrhythmia, etc. (cardiovascular phenotypes), reduced eosinophil number (hematopoietic phenotype), abnormal sarcoplasmic reticulum morphology, and reduced ventricle muscle contractility (muscle phenotypes). Interestingly, *CASQ2* is a moderate affinity calcium binding protein that is localized to the sarcoplasmic reticulum in cardiac and skeletal muscle and functions as an internal calcium store and regulator of luminal calcium release triggering muscle contraction. At least three separate rare variant mutations in *CASQ2* (*rs121434549*, *rs786205106*, and *rs121434550*) have been associated with catecholaminergic polymorphic ventricular

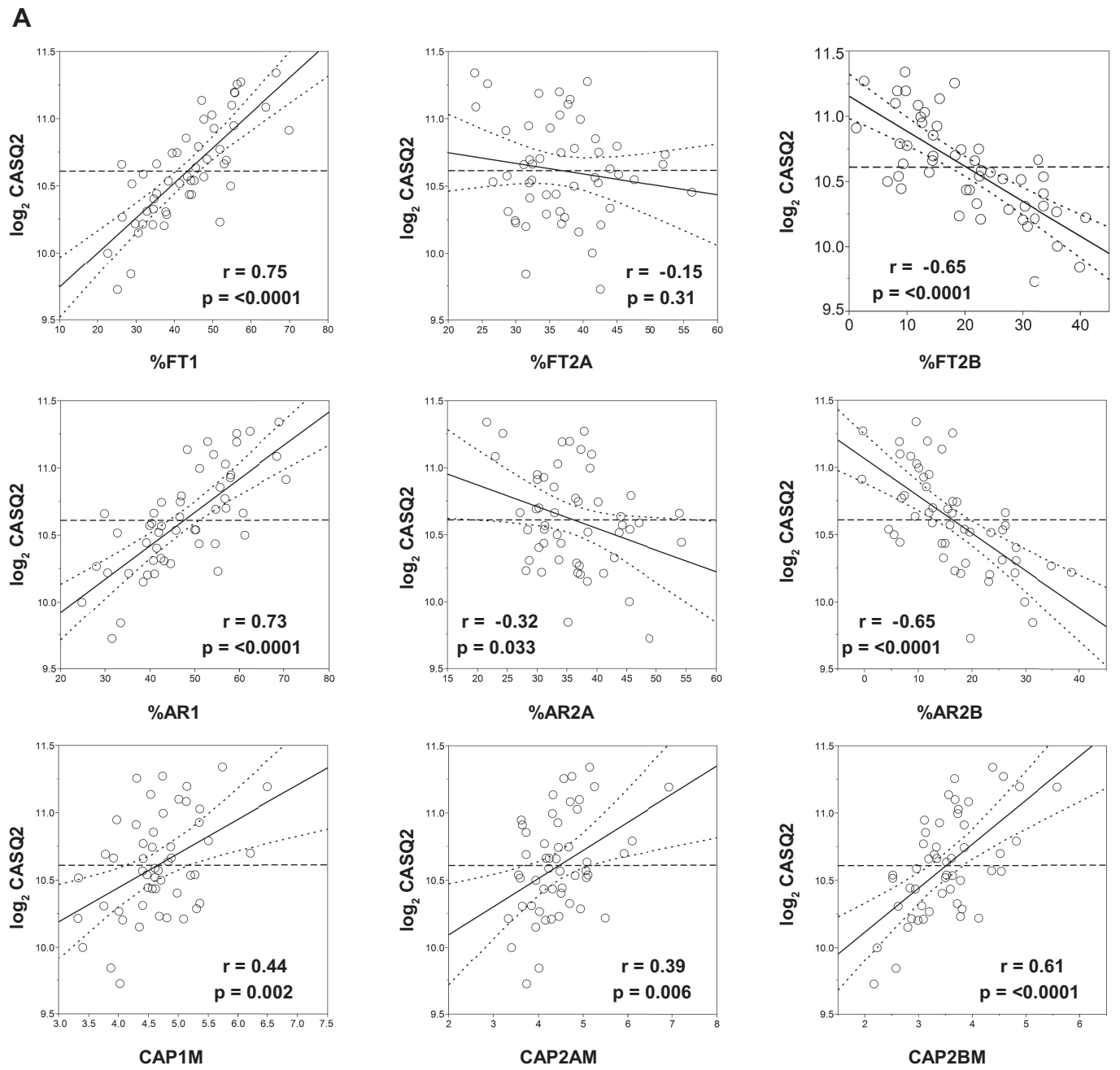


Fig. 11. Association of skeletal muscle *CASQ2* gene expression to selected muscle-related traits in a subset of the HERITAGE cohort. Partial regression residual leverage plots were constructed for selected muscle trait values (*x*-axis) against *CASQ2* gene expression from muscle biopsies ($log_2$ transformed, *y*-axis), after adjustments for age, sex, body mass index, and scan date. In all plots, the least squares line through the plotted points and its 95% confidence curves are shown. Significant effects are indicated when the confidence curve crosses the horizontal line. *A*: regression of *CASQ2* expression to muscle fiber-related phenotypes: *%FT1*, %type 1 fibers; *%FT2A*, %type 2A fibers; *%FT2B*, %type 2B fibers; *%AR1*, type 1 percentage area; *%AR2A*, type 2A percentage area; *%AR2B*, type 2B percentage area; *CAP1M*, capillary per fiber type 1 mean; *CAP2AM*, capillary per fiber type 2A mean; *CAP2BM*, capillary per fiber type 2B mean. *B*: regression of *CASQ2* expression on muscle enzyme activities (all reported as units/g): *COX*, cytochrome oxidase; *CS*, citrate synthase; *HADH*, hydroxyacyl-CoA dehydrogenase; *PFK*, phosphofructokinase.

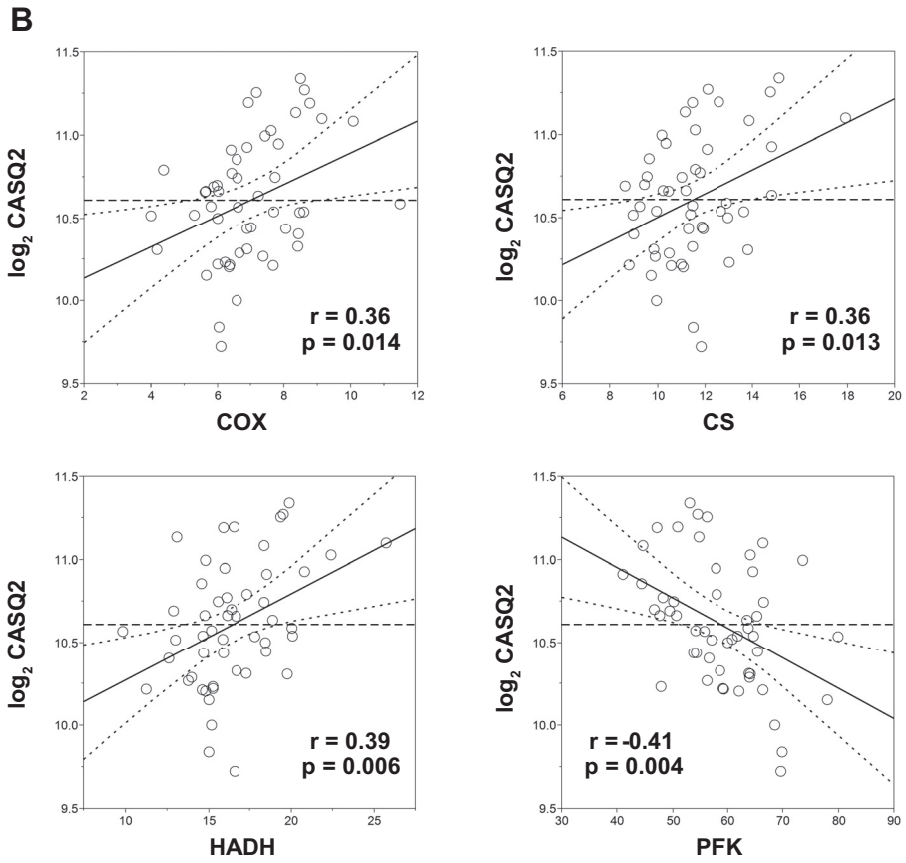


Fig. 11. *Continued*

tachycardia 2 (CPVT2) (25, 50, 54). Our findings would suggest that common variants in *CASQ2* may also influence cardiac function and, eventually, CRF. Knockout of arginyltransferase 1 (*ATE1*) in mouse models impacted several cardiovascular functions involving morphological defects in heart and metabolic phenotypes including impairments in adaptive thermogenesis and reduced circulating glucose. However, no human phenotypes of *ATE1* mutants have been reported yet. Phosphatidylinositol binding clathrin assembly protein (*PICALM*) gene knockout displayed a wide range of hematopoietic and metabolic effects (abnormal erythropoiesis, decreased erythroblast number, impaired hematopoiesis, abnormal iron homeostasis, etc.), whereas the largest number of muscle-related phenotypes were observed in sarcoglycan gamma (*SGCG*) knockouts (skeletal muscle hypertrophy, skeletal muscle necrosis, muscle degeneration, etc.). Notably, *SGCG* is a sarcolemmal transmembrane glycoprotein and a component of the sarcoglycan complex linking F-actin cytoskeleton and the extracellular matrix in muscle cells. Impairments in *SGCG* function result in early onset recessive muscular dystrophy, especially limb-girdle muscular dystrophy, in humans (30, 32, 61). Knockout of the protease-associated domain containing 1 (*PRADC1*) gene resulted in a metabolic phenotype characterized by increased circulating glucose. It is thus tempting to speculate that genes that direct metabolic fuel delivery to critical tissues, such as skeletal muscle and heart, may play an additional role in regulating CRF.

Using the HERITAGE resource of extensive exercise cardiovascular and muscle physiology phenotypes, we were able to extend our exploration of the potential role of the prioritized gene loci and muscle transcript levels to direct testing of their associations with traits known to be of relevance to intrinsic CRF. SNPs in the newly identified gene targets (*ADRB1*, *ATE1*, *CASQ2*, *DMRT2*, *NOTO*, and *SGCG*) for the central cardiovascular and $O_2$ delivery determinants of intrinsic CRF were found to be associated with relevant exercise traits such as heart rate at 60% of $\dot{V}O_{2max}$, stroke volume at 60% of $\dot{V}O_{2max}$, systolic blood pressure at $\dot{V}O_{2max}$, etc. Multiple SNPs in *CASQ2*, *SGCG*, *ADRB1*, and *DMRT2* were also associated with muscle fiber type distribution, capillarity, and enzyme activities in a manner concordant with the expected physiological relevance. For instance, when a given allele was associated positively with traits indicative of higher oxidative potential, the same allele was shown to be negatively correlated with traits indicative of greater muscle glycolytic potential. Globally, more research is clearly warranted on the new candidate loci identified in the present report and the role of DNA sequence variants on gene expression in relevant tissues in relation to maximal exercise capacity in sedentary people and in model organisms. Our findings strongly suggest that sets of genes, mostly unrecognized until now, have the potential to contribute to our understanding of the connection among the genetic, regulatory, and integrative physiology of cardiorespiratory fitness.

Some potential limitations of the study are now discussed. Due to the relatively modest cohort size, analyses requiring prior SNP selection were based on a less stringent GWAS association $P$ value cutoff ($P < 1 \times 10^{-04}$) to partly offset the generally weaker association signals observed in smaller cohorts. For larger GWAS studies, we recommend filtering SNPs at $P < 1 \times 10^{-05}$ or $1 \times 10^{-06}$ for the types of integrative analyses performed here. Second, since the association of SNPs with regulatory marks (eQTL and histone and transcription factor binding) are based on data from heterogeneous sources (ENCODE, GTEx, etc.), one should see these results primarily as hypothesis generating and subsequently conduct focused experiments to test such hypotheses in subjects intensively phenotyped for relevant exercise and CRF traits, as well in appropriate experimental systems (e.g., heart and muscle induced pluripotent stem cells derived from individuals at the extremes of intrinsic CRF). Third, in addition to the inverse CV-based weighted ranking scheme implemented in this study for the ranking of genes from combined bioinformatic evidence, other approaches for evidence weighting (e.g., Bayesian models) are conceptually possible and warrant further exploration (31, 83). Lastly, the interpretation of results from the knockout mouse database analysis may be susceptible to inference bias; while the presence of a phenotype upon knockout of a gene is usually a good reason to consider the gene's importance toward the phenotype, the absence of such a phenotype is not necessarily evidence for the gene's noninvolvement. For example, a CRF-relevant phenotype may simply not have been tested so far for a specific gene knockout or a phenotype may only become apparent upon appropriate stimulation (e.g., treadmill test).

We conclude that since human heterogeneity in intrinsic CRF, as evaluated by repeated $\dot{V}O_{2max}$ tests in confirmed sedentary adults, is considerable and because CRF is strongly associated with several common chronic diseases and human longevity, there is a need to fully understand the underlying physiology and molecular biology that governs variation in intrinsic CRF. In this paper, we have explored some possible mechanisms for the genetic regulation of intrinsic CRF by combining integrative bioinformatic analyses with evidence for phenotypic effects identified in knockout mouse models. Our analysis suggests four gene loci related to cardiovascular physiology (*ATE1*, *CASQ2*, *NOTO*, and *SGCG*), four loci related to hematopoiesis (*PICALM*, *SSB*, *CASQ2*, and *CA9*), four loci related to skeletal muscle function (*SGCG*, *DMRT2*, *ADARB1*, and *CASQ2*), and eight loci related to metabolism (*ATE1*, *PICALM*, *RAB11FIP5*, *GBA2*, *SGCG*, *PRADC1*, *ARL6IP5*, and *CASQ2*) as possible candidates for functional follow-up based on combined bioinformatic evidence and phenotypic associations. We were able to provide preliminary supportive evidence for a role of DNA sequence variants and muscle transcript expression levels on exercise cardiovascular physiology and skeletal muscle morphology and metabolism traits for some of these genes in sedentary adults (*ADRB1*, *ATE1*, *CASQ2*, *DMRT2*, *NOTO*, and *SGCG*). Replication studies based on appropriate cohorts and study designs are warranted.

## ACKNOWLEDGMENTS

We thank Drs. Arthur S. Leon, D.C. Rao, James S. Skinner, Tuomo Rankinen, Jacques Gagnonm, and the late Jack H. Wilmore for contributions to the planning, data collection, and conduct of the HERITAGE project.

## GRANTS

This research was partially funded by National Heart, Lung, and Blood Institute Grants HL-45670, HL-47317, HL-47321, HL-47323, and HL-47327, all in support of the HERITAGE Family Study). C. Bouchard is partially funded by the John W. Barton Sr. Chair in Genetics and Nutrition. Z. He is funded by the China Scholarship Council (File No. 201603620001) and China Institute of Sport Science (2015-01, 2016-01). S. Ghosh and C. Bouchard are partially supported by the National Institute of General Medical Sciences (NIGMS)-funded COBRE Grant 8-P30-GM-118430-01. S. Ghosh is supported in part by NIGMS Grant 2-U54-GM-104940, which funds the Louisiana Clinical and Translational Science Center. M. A. Sarzynski is partially supported by NIGMS Grant P20-GM-103499, which funds the South Carolina IDeA Network of Biomedical Research Excellence. This research was also supported by the National Medical Research Council, Ministry of Health, Singapore (WBS R913200076263; to S. Ghosh).

## DISCLOSURES

M.A. Sarzynski is a consultant for Genetic Direction. The other authors have no conflicts of interest to declare.

## AUTHOR CONTRIBUTIONS

S.G. and C.B. conceived and designed research; S.G., M.H., X.C., J.K., P.G., and J.J.R.-R. analyzed data; Z.H., M.A.S., and C.B. interpreted results of experiments; S.G., J.J.R.-R., M.A.S., and C.B. prepared figures; S.G., M.A.S., and C.B. drafted manuscript; S.G., Z.H., M.A.S., and C.B. edited and revised manuscript; S.G., Z.H., M.A.S., and C.B. approved final version of manuscript.

## REFERENCES

1. **Albert FW, Kruglyak L.** The role of regulatory variation in complex traits and disease. *Nat Rev Genet* 16: 197–212, 2015. doi:10.1038/nrg3891.
2. **Bassett DR Jr, Howley ET.** Limiting factors for maximum oxygen uptake and determinants of endurance performance. *Med Sci Sports Exerc* 32: 70–84, 2000. doi:10.1097/00005768-200001000-00012.
3. **Bernstein BE, Stamatoyannopoulos JA, Costello JF, Ren B, Milosavljevic A, Meissner A, Kellis M, Marra MA, Beaudet AL, Ecker JR, Farnham PJ, Hirst M, Lander ES, Mikkelsen TS, Thomson JA.** The NIH Roadmap Epigenomics Mapping Consortium. *Nat Biotechnol* 28: 1045–1048, 2010. doi:10.1038/nbt1010-1045.
4. **Blair SN, Kampert JB, Kohl HW 3rd, Barlow CE, Macera CA, Paffenbarger RS Jr, Gibbons LW.** Influences of cardiorespiratory fitness and other precursors on cardiovascular disease and all-cause mortality in men and women. *JAMA* 276: 205–210, 1996. doi:10.1001/jama.1996.03540030039029.
5. **Bolstad BM, Irizarry RA, Astrand M, Speed TP.** A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *Bioinformatics* 19: 185–193, 2003. doi:10.1093/bioinformatics/19.2.185.
6. **Bouchard C, Blair SN, Katzmarzyk PT.** Less sitting, more physical activity, or higher fitness? *Mayo Clin Proc* 90: 1533–1540, 2015. doi:10.1016/j.mayocp.2015.08.005.
7. **Bouchard C, Daw EW, Rice T, Pérusse L, Gagnon J, Province MA, Leon AS, Rao DC, Skinner JS, Wilmore JH.** Familial resemblance for $VO_2$max in the sedentary state: the HERITAGE family study. *Med Sci Sports Exerc* 30: 252–258, 1998. doi:10.1097/00005768-199802000-00013.
8. **Bouchard C, Leon AS, Rao DC, Skinner JS, Wilmore JH, Gagnon J.** The HERITAGE family study. Aims, design, and measurement protocol. *Med Sci Sports Exerc* 27: 721–729, 1995. doi:10.1249/00005768-199505000-00015.
9. **Bouchard C, Sarzynski MA, Rice TK, Kraus WE, Church TS, Sung YJ, Rao DC, Rankinen T.** Genomic predictors of the maximal $O_2$ uptake response to standardized exercise training programs. *J Appl Physiol (1985)* 110: 1160–1170, 2011. doi:10.1152/japplphysiol.00973.2010.
10. **Calbet JA, Lundby C, Sander M, Robach P, Saltin B, Boushel R.** Effects of ATP-induced leg vasodilation on $\dot{V}O_{2peak}$ and leg $O_2$ extraction during maximal exercise in humans. *Am J Physiol Regul Integr Comp Physiol* 291: R447–R453, 2006. doi:10.1152/ajpregu.00746.2005.
11. **Ceaser T, Hunter G.** Black and White race differences in aerobic capacity, muscle fiber type, and their influence on metabolic processes. *Sports Med* 45: 615–623, 2015. doi:10.1007/s40279-015-0318-7.

12. **Chatr-Aryamontri A, Breitkreutz BJ, Oughtred R, Boucher L, Heinicke S, Chen D, Stark C, Breitkreutz A, Kolas N, O'Donnell L, Reguly T, Nixon J, Ramage L, Winter A, Sellam A, Chang C, Hirschman J, Theesfeld C, Rust J, Livstone MS, Dolinski K, Tyers M.** The BioGRID interaction database: 2015 update. *Nucleic Acids Res* 43, *D1*: D470–D478, 2015. doi:10.1093/nar/gku1204.
13. **Chorev M, Carmel L.** The function of introns. *Front Genet* 3: 55, 2012. doi:10.3389/fgene.2012.00055.
14. **Choucair N, Abou Ghoch J, Fawaz A, Mégarbané A, Chouery E.** 10q26.1 Microdeletion: Redefining the critical regions for microcephaly and genital anomalies. *Am J Med Genet A* 167A: 2707–2713, 2015. doi:10.1002/ajmg.a.37211.
15. **Church TS, Kampert JB, Gibbons LW, Barlow CE, Blair SN.** Usefulness of cardiorespiratory fitness as a predictor of all-cause and cardiovascular disease mortality in men with systemic hypertension. *Am J Cardiol* 88: 651–656, 2001. doi:10.1016/S0002-9149(01)01808-2.
16. **Church TS, LaMonte MJ, Barlow CE, Blair SN.** Cardiorespiratory fitness and body mass index as predictors of cardiovascular disease mortality among men with diabetes. *Arch Intern Med* 165: 2114–2120, 2005. doi:10.1001/archinte.165.18.2114.
17. **Clarke K, Ricciardi S, Pearson T, Bharudin I, Davidsen PK, Bonomo M, Brina D, Scagliola A, Simpson DM, Beynon RJ, Khanim F, Ankers J, Sarzynski MA, Ghosh S, Pisconti A, Rozman J, Hrabe de Angelis M, Bunce C, Stewart C, Egginton S, Caddick M, Jackson M, Bouchard C, Biffo S, Falciani F.** The role of Eif6 in skeletal muscle homeostasis revealed by endurance training co-expression networks. *Cell Rep* 21: 1507–1520, 2017. doi:10.1016/j.celrep.2017.10.040.
23. **Costill DL, Daniels J, Evans W, Fink W, Krahenbuhl G, Saltin B.** Skeletal muscle enzymes and fiber composition in male and female track athletes. *J Appl Physiol* 40: 149–154, 1976. doi:10.1152/jappl.1976.40.2.149.
24. **Croft D, Mundo AF, Haw R, Milacic M, Weiser J, Wu G, Caudy M, Garapati P, Gillespie M, Kamdar MR, Jassal B, Jupe S, Matthews L, May B, Palatnik S, Rothfels K, Shamovsky V, Song H, Williams M, Birney E, Hermjakob H, Stein L, D'Eustachio P.** The Reactome pathway knowledgebase. *Nucleic Acids Res* 42, *D1*: D472–D477, 2014. doi:10.1093/nar/gkt1102.
25. **di Barletta MR, Viatchenko-Karpinski S, Nori A, Memmi M, Terentyev D, Turcato F, Valle G, Rizzi N, Napolitano C, Gyorke S, Volpe P, Priori SG.** Clinical phenotype and functional characterization of CASQ2 mutations associated with catecholaminergic polymorphic ventricular tachycardia. *Circulation* 114: 1012–1019, 2006. doi:10.1161/CIRCULATIONAHA.106.623793.
26. **di Prampero PE.** Metabolic and circulatory limitations to $VO_2$ max at the whole animal level. *J Exp Biol* 115: 319–331, 1985.
27. **di Prampero PE, Ferretti G.** Factors limiting maximal oxygen consumption in humans. *Respir Physiol* 80: 113–127, 1990. doi:10.1016/0034-5687(90)90075-A.
27a. **ENCODE Project Consortium.** The ENCODE (ENCyclopedia Of DNA Elements) Project. *Science* 306: 636–640, 2004. doi:10.1126/science.1105136.
27b. **ENCODE Project Consortium.** An integrated encyclopedia of DNA elements in the human genome. *Nature* 489: 57–74, 2012. doi:10.1038/nature11247.
28. **Ernst J, Kheradpour P, Mikkelsen TS, Shoresh N, Ward LD, Epstein CB, Zhang X, Wang L, Issner R, Coyne M, Ku M, Durham T, Kellis M, Bernstein BE.** Mapping and analysis of chromatin state dynamics in nine human cell types. *Nature* 473: 43–49, 2011. doi:10.1038/nature09906.
29. **Evangelou M, Smyth DJ, Fortune MD, Burren OS, Walker NM, Guo H, Onengut-Gumuscu S, Chen WM, Concannon P, Rich SS, Todd JA, Wallace C.** A method for gene-based pathway analysis using genomewide association study summary statistics reveals nine new type 1 diabetes associations. *Genet Epidemiol* 38: 661–670, 2014. doi:10.1002/gepi.21853.
30. **Fanin M, Hoffman EP, Angelini C, Pegoraro E.** Private beta- and gamma-sarcoglycan gene mutations: evidence of a founder effect in Northern Italy. *Hum Mutat* 16: 13–17, 2000. doi:10.1002/1098-1004(200007)16:1<13::AID-HUMU3>3.0.CO;2-V.
31. **Gagliano SA, Barnes MR, Weale ME, Knight J.** A Bayesian method to incorporate hundreds of functional characteristics with association evidence to improve variant prioritization. *PLoS One* 9: e98122, 2014. doi:10.1371/journal.pone.0098122.
32. **Georgieva B, Todorova A, Tournev I, Mitev V, Kremensky I.** C283Y gamma-sarcoglycan gene mutation in the Bulgarian Roma (Gypsy) population: prevalence study and carrier screening in a high-risk community. *Clin Genet* 66: 467–472, 2004. doi:10.1111/j.1399-0004.2004.00335.x.
33. **Ghosh S, Bouchard C.** Convergence between biological, behavioural and genetic determinants of obesity. *Nat Rev Genet* 18: 731–748, 2017. doi:10.1038/nrg.2017.72.
34. **Ghosh S, Vivar J, Nelson CP, Willenborg C, Segrè AV, Mäkinen VP, Nikpay M, Erdmann J, Blankenberg S, O'Donnell C, März W, Laaksonen R, Stewart AF, Epstein SE, Shah SH, Granger CB, Hazen SL, Kathiresan S, Reilly MP, Yang X, Quertermous T, Samani NJ, Schunkert H, Assimes TL, McPherson R.** Systems genetics analysis of genome-wide association study reveals novel associations between key biological processes and coronary artery disease. *Arterioscler Thromb Vasc Biol* 35: 1712–1722, 2015. doi:10.1161/ATVBAHA.115.305513.
35. **Ghosh S, Vivar JC, Sarzynski MA, Sung YJ, Timmons JA, Bouchard C, Rankinen T.** Integrative pathway analysis of a genome-wide association study of $\dot{V}O_{2max}$ response to exercise training. *J Appl Physiol (1985)* 115: 1343–1359, 2013. doi:10.1152/japplphysiol.01487.2012.
36. **Gledhill N, Warburton D, Jamnik V.** Haemoglobin, blood volume, cardiac function, and aerobic power. *Can J Appl Physiol* 24: 54–65, 1999. doi:10.1139/h99-006.
37. **González-Alonso J, Calbet JA.** Reductions in systemic and skeletal muscle blood flow and oxygen delivery limit maximal aerobic capacity in humans. *Circulation* 107: 824–830, 2003. doi:10.1161/01.CIR.0000049746.29175.3F.
38. **Greene CS, Himmelstein DS.** Genetic association-guided analysis of gene networks for the study of complex traits. *Circ Cardiovasc Genet* 9: 179–184, 2016. doi:10.1161/CIRCGENETICS.115.001181.
39. **Greene CS, Krishnan A, Wong AK, Ricciotti E, Zelaya RA, Himmelstein DS, Zhang R, Hartmann BM, Zaslavsky E, Sealfon SC, Chasman DI, FitzGerald GA, Dolinski K, Grosser T, Troyanskaya OG.** Understanding multicellular function and disease with human tissue-specific networks. *Nat Genet* 47: 569–576, 2015. doi:10.1038/ng.3259.
39a. **GTEx Consortium.** Human genomics. The Genotype-Tissue Expression (GTEx) pilot analysis: multitissue gene regulation in humans. *Science* 348: 648–660, 2015. doi:10.1126/science.1262110.
39b. **GTEx Consortium; Laboratory, Data Analysis &Coordinating Center (LDACC)—Analysis Working Group; Statistical Methods groups—Analysis Working Group; Enhancing GTEx (eGTEx) groups; NIH Common Fund; NIH/NCI; NIH/NHGRI; NIH/NIMH; NIH/NIDA; Biospecimen Collection Source Site—NDRI; Biospecimen Collection Source Site—RPCI; Biospecimen Core Resource—VARI; Brain Bank Repository—University of Miami Brain Endowment Bank; Leidos Biomedical—Project Management; ELSI Study; Genome Browser Data Integration &Visualization—EBI; Genome Browser Data Integration &Visualization—UCSC Genomics Institute, University of California Santa Cruz; Lead analysts; Laboratory, Data Analysis & Coordinating Center (LDACC); NIH program management; Biospecimen collection; Pathology; eQTL manuscript working group, Battle A, Brown CD, Engelhardt BE, Montgomery SB.** Genetic effects on gene expression across human tissues. *Nature* 550: 204–213, 2017. [Erratum in Nature 553: 530, 2018] doi:10.1038/nature24277.
39c. **GTEx Consortium.** Human genomics. The Genotype-Tissue Expression (GTEx) pilot analysis: multitissue gene regulation in humans. *Science* 348: 648–660, 2015. doi:10.1126/science.1262110.
40. **Gusev A, Ko A, Shi H, Bhatia G, Chung W, Penninx BW, Jansen R, de Geus EJ, Boomsma DI, Wright FA, Sullivan PF, Nikkola E, Alvarez M, Civelek M, Lusis AJ, Lehtimäki T, Raitoharju E, Kähönen M, Seppälä I, Raitakari OT, Kuusisto J, Laakso M, Price AL, Pajukanta P, Pasaniuc B.** Integrative approaches for large-scale transcriptome-wide association studies. *Nat Genet* 48: 245–252, 2016. doi:10.1038/ng.3506.
41. **Gusev A, Lee SH, Trynka G, Finucane H, Vilhjálmsson BJ, Xu H, Zang C, Ripke S, Bulik-Sullivan B, Stahl E, Kähler AK, Hultman CM, Purcell SM, McCarroll SA, Daly M, Pasaniuc B, Sullivan PF, Neale BM, Wray NR, Raychaudhuri S, Price AL; Schizophrenia Working Group of the Psychiatric Genomics Consortium; SWE-SCZ Consortium.** Partitioning heritability of regulatory and cell-type-specific variants across 11 common diseases. *Am J Hum Genet* 95: 535–552, 2014. doi:10.1016/j.ajhg.2014.10.004.
42. **Hao K, Bossé Y, Nickle DC, Paré PD, Postma DS, Laviolette M, Sandford A, Hackett TL, Daley D, Hogg JC, Elliott WM, Couture C, Lamontagne M, Brandsma CA, van den Berge M, Koppelman G, Reicin AS, Nicholson DW, Malkov V, Derry JM, Suver C, Tsou JA, Kulkarni A, Zhang C, Vessey R, Opiteck GJ, Curtis SP, Timens W, Sin DD.** Lung eQTLs to help reveal the molecular underpinnings of

asthma. *PLoS Genet* 8: e1003029, 2012. [Erratum in *PLoS Genet* 8: 2012]. doi10.1371/journal.pgen.1003029.

43. **Harber MP, Kaminsky LA, Arena R, Blair SN, Franklin BA, Myers J, Ross R.** Impact of cardiorespiratory fitness on all-cause and disease-specific mortality: advances since 2009. *Prog Cardiovasc Dis* 60: 11–20, 2017. doi:10.1016/j.pcad.2017.03.001.
44. **Hawkins MN, Raven PB, Snell PG, Stray-Gundersen J, Levine BD.** Maximal oxygen uptake as a parametric measure of cardiorespiratory capacity. *Med Sci Sports Exerc* 39: 103–107, 2007. [Erratum in *Med Sci Sports Exerc* 39: 574, 2007]. doi:10.1249/01.mss.0000241641.75101.64.
45. **Heger A, Webber C, Goodson M, Ponting CP, Lunter G.** GAT: a simulation framework for testing the association of genomic intervals. *Bioinformatics* 29: 2046–2048, 2013. doi:10.1093/bioinformatics/btt343.
46. **Henderson KK, Wagner H, Favret F, Britton SL, Koch LG, Wagner PD, Gonzalez NC.** Determinants of maximal $O_2$ uptake in rats selectively bred for endurance running capacity. *J Appl Physiol (1985)* 93: 1265–1274, 2002. doi:10.1152/japplphysiol.00809.2001.
47. **Hon CC, Ramilowski JA, Harshbarger J, Bertin N, Rackham OJ, Gough J, Denisenko E, Schmeier S, Poulsen TM, Severin J, Lizio M, Kawaji H, Kasukawa T, Itoh M, Burroughs AM, Noma S, Djebali S, Alam T, Medvedeva YA, Testa AC, Lipovich L, Yip CW, Abugessaisa I, Mendez M, Hasegawa A, Tang D, Lassmann T, Heutink P, Babina M, Wells CA, Kojima S, Nakamura Y, Suzuki H, Daub CO, de Hoon MJ, Arner E, Hayashizaki Y, Carninci P, Forrest AR.** An atlas of human long non-coding RNAs with accurate 5′ ends. *Nature* 543: 199–204, 2017. doi:10.1038/nature21374.
48. **Huang W, Sherman BT, Lempicki RA; W.** Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat Protoc* 4: 44–57, 2009. doi:10.1038/nprot.2008.211.
49. **Hunter GR, Weinsier RL, McCarthy JP, Enette Larson-Meyer D, Newcomer BR.** Hemoglobin, muscle oxidative capacity, and $\dot{V}O_{2max}$ in African-American and Caucasian women. *Med Sci Sports Exerc* 33: 1739–1743, 2001. doi:10.1097/00005768-200110000-00019.
50. **Knollmann BC, Chopra N, Hlaing T, Akin B, Yang T, Ettensohn K, Knollmann BE, Horton KD, Weissman NJ, Holinstat I, Zhang W, Roden DM, Jones LR, Franzini-Armstrong C, Pfeifer K.** Casq2 deletion causes sarcoplasmic reticulum volume increase, premature $Ca^{2+}$ release, and catecholaminergic polymorphic ventricular tachycardia. *J Clin Invest* 116: 2510–2520, 2006. doi:10.1172/JCI29128.
51. **Kokkinos P, Myers J, Faselis C, Panagiotakos DB, Doumas M, Pittaras A, Manolis A, Kokkinos JP, Karasik P, Greenberg M, Papademetriou V, Fletcher R.** Exercise capacity and mortality in older men: a 20-year follow-up study. *Circulation* 122: 790–797, 2010. doi:10.1161/CIRCULATIONAHA.110.938852.
52. **Kokkinos P, Myers J, Kokkinos JP, Pittaras A, Narayan P, Manolis A, Karasik P, Greenberg M, Papademetriou V, Singh S.** Exercise capacity and mortality in black and white men. *Circulation* 117: 614–622, 2008. doi:10.1161/CIRCULATIONAHA.107.734764.
53. **Kokkinos P, Myers J, Nylen E, Panagiotakos DB, Manolis A, Pittaras A, Blackman MR, Jacob-Issac R, Faselis C, Abella J, Singh S.** Exercise capacity and all-cause mortality in African American and Caucasian men with type 2 diabetes. *Diabetes Care* 32: 623–628, 2009. doi:10.2337/dc08-1876.
54. **Lahat H, Pras E, Olender T, Avidan N, Ben-Asher E, Man O, Levy-Nissenbaum E, Khoury A, Lorber A, Goldman B, Lancet D, Eldar M.** A missense mutation in a highly conserved region of CASQ2 is associated with autosomal recessive catecholamine-induced polymorphic ventricular tachycardia in Bedouin families from Israel. *Am J Hum Genet* 69: 1378–1384, 2001. doi:10.1086/324565.
55. **Lamparter D, Marbach D, Rueedi R, Kutalik Z, Bergmann S.** Fast and rigorous computation of gene and pathway scores from SNP-based summary statistics. *PLOS Comput Biol* 12: e1004714, 2016. doi:10.1371/journal.pcbi.1004714.
56. **Levine BD.** VO2max: what do we know, and what do we still need to know? *J Physiol* 586: 25–34, 2008. doi:10.1113/jphysiol.2007.147629.
57. **Li Y, Willer CJ, Ding J, Scheet P, Abecasis GR.** MaCH: using sequence and genotype data to estimate haplotypes and unobserved genotypes. *Genet Epidemiol* 34: 816–834, 2010. doi:10.1002/gepi.20533.
58. **Lundby C, Montero D, Joyner M.** Biology of $VO_2$ max: looking under the physiology lamp. *Acta Physiol (Oxf)* 220: 218–228, 2017. doi:10.1111/apha.12827.
59. **Machiela MJ, Chanock SJ.** LDlink: a web-based application for exploring population-specific haplotype structure and linking correlated alleles of possible functional variants. *Bioinformatics* 31: 3555–3557, 2015. doi:10.1093/bioinformatics/btv402.
60. **Mäkinen VP, Civelek M, Meng Q, Zhang B, Zhu J, Levian C, Huan T, Segrè AV, Ghosh S, Vivar J, Nikpay M, Stewart AF, Nelson CP, Willenborg C, Erdmann J, Blakenberg S, O'Donnell CJ, März W, Laaksonen R, Epstein SE, Kathiresan S, Shah SH, Hazen SL, Reilly MP, Lusis AJ, Samani NJ, Schunkert H, Quertermous T, McPherson R, Yang X, Assimes TL; Coronary ARtery DIsease Genome-Wide Replication And Meta-Analysis (CARDIoGRAM) Consortium.** Integrative genomics reveals novel molecular pathways and gene networks for coronary artery disease. *PLoS Genet* 10: e1004502, 2014. doi:10.1371/journal.pgen.1004502.
61. **McNally EM, Duggan D, Gorospe JR, Bönnemann CG, Fanin M, Pegoraro E, Lidov HG, Noguchi S, Ozawa E, Finkel RS, Cruse RP, Angelini C, Kunkel LM, Hoffman EP.** Mutations that disrupt the carboxyl-terminus of gamma-sarcoglycan cause muscular dystrophy. *Hum Mol Genet* 5: 1841–1847, 1996. doi:10.1093/hmg/5.11.1841.
62. **Mewes HW, Heumann K, Kaps A, Mayer K, Pfeiffer F, Stocker S, Frishman D.** MIPS: a database for genomes and protein sequences. *Nucleic Acids Res* 27: 44–48, 1999. doi:10.1093/nar/27.1.44.
63. **Miller CL, Pjanic M, Wang T, Nguyen T, Cohain A, Lee JD, Perisic L, Hedin U, Kundu RK, Majmudar D, Kim JB, Wang O, Betsholtz C, Ruusalepp A, Franzén O, Assimes TL, Montgomery SB, Schadt EE, Björkegren JL, Quertermous T.** Integrative functional genomics identifies regulatory mechanisms at coronary artery disease loci. *Nat Commun* 7: 12092, 2016. doi:10.1038/ncomms12092.
64. **Musunuru K, Strong A, Frank-Kamenetsky M, Lee NE, Ahfeldt T, Sachs KV, Li X, Li H, Kuperwasser N, Ruda VM, Pirruccello JP, Muchmore B, Prokunina-Olsson L, Hall JL, Schadt EE, Morales CR, Lund-Katz S, Phillips MC, Wong J, Cantley W, Racie T, Ejebe KG, Orho-Melander M, Melander O, Kotelianski V, Fitzgerald K, Krauss RM, Cowan CA, Kathiresan S, Rader DJ.** From noncoding variant to phenotype via SORT1 at the 1p13 cholesterol locus. *Nature* 466: 714–719, 2010. doi:10.1038/nature09266.
65. **Orchard S, Ammari M, Aranda B, Breuza L, Briganti L, Broackes-Carter F, Campbell NH, Chavali G, Chen C, del-Toro N, Duesbury M, Dumousseau M, Galeota E, Hinz U, Iannuccelli M, Jagannathan S, Jimenez R, Khadake J, Lagreid A, Licata L, Lovering RC, Meldal B, Melidoni AN, Milagros M, Peluso D, Perfetto L, Porras P, Raghunath A, Ricard-Blum S, Roechert B, Stutz A, Tognolli M, van Roey K, Cesareni G, Hermjakob H.** The MIntAct project-IntAct as a common curation platform for 11 molecular interaction databases. *Nucleic Acids Res* 42, *D1*: D358–D363, 2014. doi:10.1093/nar/gkt1115.
66. **Paffenbarger RS Jr, Hyde RT, Wing AL, Lee IM, Jung DL, Kampert JB.** The association of changes in physical-activity level and other lifestyle characteristics with mortality among men. *N Engl J Med* 328: 538–545, 1993. doi:10.1056/NEJM199302253280804.
67. **Pasquali L, Gaulton KJ, Rodríguez-Seguí SA, Mularoni L, Miguel-Escalada I, Akerman İ, Tena JJ, Morán I, Gómez-Marín C, van de Bunt M, Ponsa-Cobas J, Castro N, Nammo T, Cebola I, García-Hurtado J, Maestro MA, Pattou F, Piemonti L, Berney T, Gloyn AL, Ravassard P, Skarmeta JL, Müller F, McCarthy MI, Ferrer J.** Pancreatic islet enhancer clusters enriched in type 2 diabetes risk-associated variants. *Nat Genet* 46: 136–143, 2014. doi:10.1038/ng.2870.
68. **Pers TH, Karjalainen JM, Chan Y, Westra HJ, Wood AR, Yang J, Lui JC, Vedantam S, Gustafsson S, Esko T, Frayling T, Speliotes EK, Boehnke M, Raychaudhuri S, Fehrmann RS, Hirschhorn JN, Franke L; Genetic Investigation of ANthropometric Traits (GIANT) Consortium.** Biological interpretation of genome-wide association studies using predicted gene functions. *Nat Commun* 6: 5890, 2015. doi:10.1038/ncomms6890.
69. **Pickrell JK.** Joint analysis of functional genomic data and genome-wide association studies of 18 human traits. *Am J Hum Genet* 94: 559–573, 2014. [Erratum in *Am J Hum Genet* 95: 126, 2014] doi:10.1016/j.ajhg.2014.03.004.
70. **Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MA, Bender D, Maller J, Sklar P, de Bakker PI, Daly MJ, Sham PC.** PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am J Hum Genet* 81: 559–575, 2007. doi:10.1086/519795.
71. **Ramasamy A, Trabzuni D, Guelfi S, Varghese V, Smith C, Walker R, De T, Coin L, de Silva R, Cookson MR, Singleton AB, Hardy J, Ryten M, Weale ME; UK Brain Expression Consortium; North American Brain Expression Consortium.** Genetic variability in the regulation of

gene expression in ten regions of the human brain. *Nat Neurosci* 17: 1418–1428, 2014. doi:10.1038/nn.3801.

72. **Rico-Sanz J, Rankinen T, Joanisse DR, Leon AS, Skinner JS, Wilmore JH, Rao DC, Bouchard C; HERITAGE Family Study.** Familial resemblance for muscle phenotypes in the HERITAGE Family Study. *Med Sci Sports Exerc* 35: 1360–1366, 2003. doi:10.1249/01.MSS.0000079031.22755.63.
73. **Roadmap Epigenomics Consortium; Kundaje A, Meuleman W, Ernst J, Bilenky M, Yen A, Heravi-Moussavi A, Kheradpour P, Zhang Z, Wang J, Ziller MJ, Amin V, Whitaker JW, Schultz MD, Ward LD, Sarkar A, Quon G, Sandstrom RS, Eaton ML, Wu YC, Pfenning AR, Wang X, Claussnitzer M, Liu Y, Coarfa C, Harris RA, Shoresh N, Epstein CB, Gjoneska E, Leung D, Xie W, Hawkins RD, Lister R, Hong C, Gascard P, Mungall AJ, Moore R, Chuah E, Tam A, Canfield TK, Hansen RS, Kaul R, Sabo PJ, Bansal MS, Carles A, Dixon JR, Farh KH, Feizi S, Karlic R, Kim AR, Kulkarni A, Li D, Lowdon R, Elliott G, Mercer TR, Neph SJ, Onuchic V, Polak P, Rajagopal N, Ray P, Sallari RC, Siebenthall KT, Sinnott-Armstrong NA, Stevens M, Thurman RE, Wu J, Zhang B, Zhou X, Beaudet AE, Boyer LA, De Jager PL, Farnham PJ, Fisher SJ, Haussler D, Jones SJ, Li W, Marra MA, McManus MT, Sunyaev S, Thomson JA, Tlsty TD, Tsai LH, Wang W, Waterland RA, Zhang MQ, Chadwick LH, Bernstein BE, Costello JF, Ecker JR, Hirst M, Meissner A, Milosavljevic A, Ren B, Stamatoyannopoulos JA, Wang T, Kellis M.** Integrative analysis of 111 reference human epigenomes. *Nature* 518: 317–330, 2015. doi:10.1038/nature14248.
74. **Roberts AM, Wong AK, Fisk I, Troyanskaya OG.** GIANT API: an application programming interface for functional genomics. *Nucleic Acids Res* 44, *W1*: W587–W592, 2016. doi:10.1093/nar/gkw289.
75. **Rockman MV, Kruglyak L.** Genetics of global gene expression. *Nat Rev Genet* 7: 862–872, 2006. doi:10.1038/nrg1964.
76. **Roy JL, Hunter GR, Fernandez JR, McCarthy JP, Larson-Meyer DE, Blaudeau TE, Newcomer BR.** Cardiovascular factors explain genetic background differences in $VO_2$max. *Am J Hum Biol* 18: 454–460, 2006. doi:10.1002/ajhb.20509.
77. **Sarzynski MA, Ghosh S, Bouchard C.** Genomic and transcriptomic predictors of response levels to endurance exercise training. *J Physiol* 595: 2931–2939, 2017. doi:10.1113/JP272559.
78. **Skinner JS, Jaskólski A, Jaskólska A, Krasnoff J, Gagnon J, Leon AS, Rao DC, Wilmore JH, Bouchard C; HERITAGE Family Study.** Age, sex, race, initial fitness, and response to training: the HERITAGE Family Study. *J Appl Physiol (1985)* 90: 1770–1776, 2001. doi:10.1152/jappl.2001.90.5.1770.
79. **Skinner JS, Wilmore KM, Jaskolska A, Jaskolski A, Daw EW, Rice T, Gagnon J, Leon AS, Wilmore JH, Rao DC, Bouchard C.** Reproducibility of maximal exercise test data in the HERITAGE Family Study. *Med Sci Sports Exerc* 31: 1623–1628, 1999. doi:10.1097/00005768-199911000-00020.
80. **Skinner JS, Wilmore KM, Krasnoff JB, Jaskólski A, Jaskólska A, Gagnon J, Province MA, Leon AS, Rao DC, Wilmore JH, Bouchard C.** Adaptation to a standardized training program and changes in fitness in a large, heterogeneous population: the HERITAGE Family Study. *Med Sci Sports Exerc* 32: 157–161, 2000. doi:10.1097/00005768-200001000-00023.
81. **Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, Paulovich A, Pomeroy SL, Golub TR, Lander ES, Mesirov JP.** Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc Natl Acad Sci USA* 102: 15545–15550, 2005. doi:10.1073/pnas.0506580102.
82. **Taylor EM, Copsey AC, Hudson JJ, Vidot S, Lehmann AR.** Identification of the proteins, including MAGEG1, that make up the human SMC5-6 protein complex. *Mol Cell Biol* 28: 1197–1206, 2008. doi:10.1128/MCB.00767-07.
83. **Thompson JR, Gögele M, Weichenberger CX, Modenese M, Attia J, Barrett JH, Boehnke M, De Grandi A, Domingues FS, Hicks AA, Marroni F, Pattaro C, Ruggeri F, Borsani G, Casari G, Parmigiani G, Pastore A, Pfeufer A, Schwienbacher C, Taliun D; CKDGen Consortium, Fox CS, Pramstaller PP, Minelli C.** SNP prioritization using a Bayesian probability of association. *Genet Epidemiol* 37: 214–221, 2013. doi:10.1002/gepi.21704.
84. **van Ginkel S, Amami M, Dela F, Niederseer D, Narici MV, Niebauer J, Scheiber P, Müller E, Flück M.** Adjustments of muscle capillarity but not mitochondrial protein with skiing in the elderly. *Scand J Med Sci Sports* 25: e360–e367, 2015. doi:10.1111/sms.12324.
85. **Wagner PD.** Determinants of maximal oxygen transport and utilization. *Annu Rev Physiol* 58: 21–50, 1996. doi:10.1146/annurev.ph.58.030196.000321.
86. **Warburton DE, Gledhill N, Quinney HA.** Blood volume, aerobic power, and endurance performance: potential ergogenic effect of volume loading. *Clin J Sport Med* 10: 59–66, 2000. doi:10.1097/00042752-200001000-00011.
87. **Ward LD, Kellis M.** HaploReg: a resource for exploring chromatin states, conservation, and regulatory motif alterations within sets of genetically linked variants. *Nucleic Acids Res* 40, *D1*: D930–D934, 2012. doi:10.1093/nar/gkr917.
88. **Wei M, Kampert JB, Barlow CE, Nichaman MZ, Gibbons LW, Paffenbarger RS Jr, Blair SN.** Relationship between low cardiorespiratory fitness and mortality in normal-weight, overweight, and obese men. *JAMA* 282: 1547–1553, 1999. doi:10.1001/jama.282.16.1547.
89. **Williams TM, Bengtson P, Steller DL, Croll DA, Davis RW.** The healthy heart: lessons from nature's elite athletes. *Physiology (Bethesda)* 30: 349–357, 2015. doi:10.1152/physiol.00017.2015.
90. **Wilmore JH, Stanforth PR, Turley KR, Gagnon J, Daw EW, Leon AS, Rao DC, Skinner JS, Bouchard C.** Reproducibility of cardiovascular, respiratory, and metabolic responses to submaximal exercise: the HERITAGE Family Study. *Med Sci Sports Exerc* 30: 259–265, 1998. doi:10.1097/00005768-199802000-00014.
91. **Zhang K, Cui S, Chang S, Zhang L, Wang J.** i-GSEA4GWAS: a web server for identification of pathways/gene sets associated with traits by applying an improved gene set enrichment analysis to genome-wide association study. *Nucleic Acids Res* 38, *Web Server*: W90–W95, 2010. doi:10.1093/nar/gkq324.
92. **Zhou J, Troyanskaya OG.** Predicting effects of noncoding variants with deep learning-based sequence model. *Nat Methods* 12: 931–934, 2015. doi:10.1038/nmeth.3547.
93. **Zhou YB, Liu F, Zhu ZD, Zhu H, Zhang X, Wang ZQ, Liu JH, Han ZG.** N-glycosylation is required for efficient secretion of a novel human secreted glycoprotein, hPAP21. *FEBS Lett* 576: 401–407, 2004. doi:10.1016/j.febslet.2004.09.039.