Identification of 64 Novel Genetic Loci Provides an Expanded View on the Genetic Architecture of Coronary Artery Disease

van der Harst, Pim; Verweij, Niek

Published in:
Circulation research

DOI:
10.1161/CIRCRESAHA.117.312086

The Identification of 64 Novel Genetic Loci Provides an Expanded View on the Genetic Architecture of Coronary Artery Disease

Pim van der Harst[1,2,3], Niek Verweij[1]

[1]University of Groningen, University Medical Center Groningen, Department of Cardiology, Groningen, The Netherlands; [2]University of Groningen, University Medical Center Groningen, Department of Genetics, Groningen, The Netherlands, and; [3]Durrer Center for Cardiogenetic Research, Netherlands Heart Institute, Utrecht, The Netherlands.

Running title: 64 Novel Genetic Loci for Coronary Artery Disease

Subject Terms:
Coronary Artery Disease
Genetic, Association Studies
Genetics
Gene Expression and Regulation
Translational Studies

Address correspondence to:
Dr. Pim van der Harst
University of Groningen
University Medical Center Groningen
Department of Cardiology
Hanzeplein 1
9700RB Groningen
Tel. +31 50 3612355
p.van.der.harst@umcg.nl

ABSTRACT

Rationale: Coronary artery disease (CAD) is a complex phenotype driven by genetic and environmental factors. 97 genetic risk loci have been identified so far, but the identification of additional susceptibility loci might be important to enhance our understanding of the genetic architecture of CAD.

Objective: To expand the number of genome-wide significant loci, catalog functional insights, and enhance our understanding of the genetic architecture of CAD.

Methods and Results: We performed a genome-wide association study (GWAS) in 34,541 CAD cases and 261,984 controls of UK biobank Resource followed by replication in 88,192 cases and 162,544 controls from CARDIoGRAMplusC4D. We identified 75 loci that replicated and were genome-wide significant ($P<5\text{x}10^{-8}$) in meta-analysis, 13 of which had not been reported previously. Next, to further identify novel loci we identified all promising ($P<0.0001$) loci in the CARDIoGRAMplusC4D data and performed reciprocal replication and meta-analyses with UK biobank. This led to the identification of 21 additional novel loci reaching genome-wide significance ($P<5\text{x}10^{-8}$) in meta-analysis. Finally, we performed a genome wide meta-analysis of all available data revealing 30 additional novel loci ($P<5\text{x}10^{-8}$) without further replication. The increase in sample size by UK Biobank raised the number of reconstituted gene-sets from 4.2% to 13.9% of all gene-sets to be involved in CAD. For the 64 novel loci, 155 candidate causal genes were prioritized, many without an obvious connection to CAD. Fine-mapping of the 161 CAD loci generated lists of credible sets of single causal variants and genes for functional follow-up. Genetic risk variants of CAD were linked to development of atrial fibrillation, heart failure and death.

Conclusions: We identified 64 novel genetic risk loci for CAD and performed fine-mapping of all 161 risk loci to obtain a credible set of causal variants. The large expansion of reconstituted gene-sets argues in favor of an expanded "omnigenic model" view on the genetic architecture of CAD.

Keywords:
Coronary artery disease, genetics, fine-mapping, genome wide association study, computational biology.

Nonstandard Abbreviations and Acronyms:

1000Genomes	a deep catalog of variation in the human genome based on DNA sequencing
CAD	coronary artery disease
CARDIoGRAM	the Coronary Artery Disease Genome-wide Replication and
DEPICT	Data-driven Expression-Prioritized Integration for Complex
eQTL	expression quantitative trait locus
GTEx	The Genotype-Tissue Expression
GWAS	genome-wide association study
IPA	Ingenuity pathway analysis
LD	linkage disequilibrium
meQTL	methylation quantitative trait locus
SNP	Single-nucleotide polymorphism

INTRODUCTION

Coronary artery disease (CAD) is the predominant cause of ischemic heart disease often leading to myocardial infarction and a leading cause of death. Globally, deaths due to ischemic heart disease increased by 16.6% from 2005 to 2015 to 8.9 million deaths. However, the age-standardized mortality rates are decreasing (fell by 12.8%)[1] due to preventive and treatment strategies established on evolving knowledge of the underlying pathophysiology of CAD.

CAD is a complex disease, resulting from numerous additive and interacting contributions in an individual's environment and lifestyle in combination with their underlying genetic architecture. Since the first genome wide association studies for CAD in 2007[2–4], multiple additional studies with progressively larger sample sizes identified 97 genome-wide significant genetic loci associated with CAD[5–10] at the time of analysis. The continuous effort to identify additional loci associated with CAD and share these early with the scientific community is important, especially to enhance our understanding of the biological underpinnings of CAD and to catalyze the development of drugs. A comprehensive understanding of the genetic architecture of CAD is also essential to enable precision medicine approaches by identifying subgroups of patients at increased risk of CAD or its complications and might identify those with a specific driving pathophysiology in whom a particular therapeutic or preventive approach would be most useful[11].

To further our knowledge of the genetic architecture of CAD, we performed a *de novo* genome-wide association study (GWAS) of the UK Biobank Resource and meta-analyses with CARDIoGRAMplusC4D data. Our approach led to the identification of 64 novel loci associated with CAD, expanding the grand total to 161. These loci were interrogated using bioinformatic approaches to catalog and interpret the potential biological relevance of our findings. We also performed network and gene-set analyses and propose the omnigenic model to explain our findings. This expanding resource is now available for other investigators to help to further elucidate the underlying biology and relevance.

METHODS

The data that support the findings of this study are available from the corresponding author upon reasonable request. The *de novo* GWAS analysis and meta-analysis have been posted on Mendeley (doi:10.17632/2zdd47c94h.1; doi:10.17632/gbbsrpx6bs.1). A summary of the methods is provided below, and a more detailed description of the experimental procedures is provided in the Online Data Supplement.

Study design and samples.

The study design consisted of a reciprocal two-stage sequential discovery and replication approach (Online Figure I) providing the most robust statistical evidence followed by an overall meta-analysis of all available data for which currently no replication data were available in this study. First, using the UK Biobank Resource we conducted a GWAS to discover SNPs associated with CAD. In stage 2, we took forward all promising SNPs reaching nominal significance (P<0.0001) for replication in CARDIoGRAMplusC4D data. Replicating SNPs (P<0.05 after Bonferroni adjustment) were meta-analyzed and considered true when surpassing the genome-wide significance threshold ($P<5\times10^{-8}$). The reciprocal stage 1 entailed the identification for all promising SNPs (P<0.0001) in CARDIoGRAMplusC4D and replication in UK Biobank (P<0.05 after Bonferroni adjustment) followed by meta-analysis. Again, SNPs replicating and surpassing the genome-wide significance threshold were considered true. A sentinel SNP in a locus was defined as the most significant variant in a 1MB region that was independent from other sentinel SNPs (r^2<0.1). A locus was defined as a region of 1MB at either side of the sentinel SNP. A locus was considered novel if the sentinel SNP was not within a 1MB window (at either side) of earlier reported genome-wide significant SNPs (Online Table I). Finally, we performed a genome-wide meta-analysis of the UK Biobank Resource and CARDIoGRAMplusC4D to identify additional CAD associated loci ($P<5\times10^{-8}$ in meta-analysis). A potential sample overlap between the UK Biobank and cohorts of CARDIoGRAMplusC4D was estimated to be smaller than 0.1%, no evidence was found that this biased the test-statistics (Online Data Supplement).

Candidate genes and insights in biology.

Candidate causal genes at each of the loci were prioritized based on proximity, eQTL data, Data-driven expression-prioritized integration for complex traits (DEPICT)[12] analyses and long-range chromatin interactions of variants with gene-promoters (see Online Data Supplement).[8,13] Summary information of genes was obtained via queries in GeneCards, EntrezGene, UniProt, and Tocris. The Mouse Genomic Informatic (MGI) database was used for obtaining insights into mammalian phenotypes associated with disruption of candidate genes. DEPICT was also used to test for enrichment of gene sets and identify relevant tissues and cell types. Ingenuity Pathway Analysis (IPA, June 2017 release) was performed to strengthen the biological relevancy of the novel loci.

Insights in loci by associations with other phenotypes.

The GWAS catalog was queried and a phenome-scan was carried out by intersecting the identified loci with the GWAS-catalog and by testing the association of the newly identified SNPs with a wide range of phenotypes using linear or logistic regression analysis in UK Biobank (see Online Data Supplement). Genetic risk scores (GRS) were constructed using effect estimates obtained CARDIoGRAMplusC4D data as described previously.[8] Multivariable Cox proportional hazards models were fitted for quintiles of the GRS in the UK Biobank Resource, to assess the extent to which the GRS could predict new onset atrial fibrillation/flutter and heart failure.

Regulatory DNA and fine-mapping of probable causal variants.

To systematically characterize the functional, cellular and regulatory contribution of genetic variation we employed GARFIELD[14], analyzing the enrichment of genome-wide association summary statistics in tissue specific functional elements at given significance thresholds. Probabilistic Annotation INtegraTOR (PAINTOR) was used to fine-map loci by integrating genetic association signal strength with genomic functional annotation data[15]. We explored the potential target genes of these candidate causal variants by determining their direct effects on protein function (missense variants) and evidence connecting the causal variant in a Utr-3' region to gene expression (eQTL) or physical interactions (Hi-C) with the promotor of an eQTL gene. Determination of potential causal mechanisms of the potential causal variants based on a) missense variation, b) chromatin-interaction between the causal variant and the promotor of a gene for which the causal variant was also significantly associated with gene-expression by eQTL analyses, or c) Utr3' overlapping variants that were also significantly associated with gene expression of the same gene corresponding to the Utr3' position. In addition, for genes/mechanisms to be prioritized by eQTL analyses and chromatin-interactions or Utr'3, the respective causal variant was required to be in an enhancer region.

RESULTS

Genome wide analyses of 34,541 cases and 261,984 controls.

The stage 1 GWAS analysis in UK biobank (34,541 cases and 261,984 controls, Online Table II) of 7,947,838 SNPs revealed 630 suggestive SNPs (P<0.0001) in 442 loci (Online Table III). 86 independent SNPs in 75 loci both replicated (P<0.05 Bonferroni adjusted) in stage 2 in up to 88,192 cases and 162,544 controls of CARDIoGRAMplusC4D, and achieved genome wide significance ($P<5\times10^{-8}$) with no evidence of heterogeneity of effects ($P_{\text{het}} \geq 0.10$). 13 of the 75 loci are not established CAD associated loci (Table 1).

Table 1. 64 novel genome-wide significant CAD loci.

Cytoband	Position	Lead SNP	A1	A2	Freq	Variant function	Candidate Genes	Resource	OR (95%CI)	P-value
1p36.33	2252205	rs36096196	T	C	0.15	downstream	MORN1[%],SKI[#]	MA	1.05(1.03-1.06)	1.3x10^{-8}
1p36.32	3325912	rs2493298	A	C	0.14	intronic	PRDM16[%*#],PEX10[^],PLCH2[^],RER1[^]	UK	1.06 (1.04-1.08)	1.9x10^{-9}
1p34.3	38461319	rs61776719	A	C	0.53	intergenic	FHL3[%$#], UTP11[$], SF3A3[$], MANEAL[$], INPP5B[$]	UK	1.04 (1.03-1.06)	1.1x10^{-9}
1p13.2	115753482	rs11806316	A	G	0.37	intergenic	NGF[%^#,] CASQ2[^]	CA	0.96 (0.95-0.97)	4.9x10^{-10}
1q32.2	210468999	rs60154123	T	C	0.15	intergenic	HHAT[%$],SERTAD4[#^],DIEXF[^]	MA	1.05(1.03-1.06)	2.5x10^{-8}
1q42.2	230845794	rs699	A	G	0.58	missense	AGT[%*$], CAPN9[^], GNPAT[^]	CA	0.96 (0.95-0.98)	2.1x10^{-8}
2p21	45896437	rs582384	A	C	0.53	intronic	PRKCE[%],TMEM247[^]	MA	1.03(1.02-1.05)	7.6x10^{-9}
2q24.3	164957251	rs12999907	A	G	0.82	intergenic	FIGN[%]	UK	1.06 (1.04-1.07)	2.4x10^{-11}
2q32.1	188196469	rs840616	T	C	0.35	intergenic	CALCRL[%$#^], TFPI[$#]	CA	0.96 (0.95-0.97)	3.0x10^{-9}
2q37.3	238223955	rs11677932	A	G	0.32	intergenic	COL6A3[%#^]	MA	0.97(0.96-0.98)	2.6x10^{-8}
3p21.31	46688562	rs7633770	A	G	0.41	intergenic	ALS2CL[%$#],RTP3[^]	MA	1.03(1.02-1.04)	1.1x10^{-8}
3p21.31	48193515	rs7617773	T	C	0.67	intergenic	CDC25A[%], SPINK8[$], MAP4[$],ZNF589[$]	UK	1.04 (1.03-1.05)	2.3x10^{-11}
3q22.1	132257961	rs10512861	T	G	0.14	downstream	DNAJC13[%#], NPHP3[#], ACAD11[^], UBA5[^]	CA	0.96 (0.94-0.97)	1.5x10^{-8}
3q22.3	136069472	rs667920	T	G	0.78	intronic	STAG1[%#^],MSL2[#],NCK1[#],PPP2R3A[#]	MA	1.05(1.04-1.06)	6.0x10^{-15}
3q25.31	156852592	rs4266144	C	G	0.68	intergenic	CCNL1[%],TIPARP[$^]	MA	0.97(0.95-0.98)	1.4x10^{-8}
3q26.31	172115902	rs12897	A	G	0.59	UTR3	FNDC3B[%$#]	CA	0.96 (0.95-0.97)	1.9x10^{-10}
4p16.3	3449652	rs16844401	A	G	0.07	missense	HGFAC[%*#], RGS12[%], MSANTD1[^]	CA	1.07 (1.04-1.10)	4.0x10^{-8}
4q21.1	77416627	rs12500824	A	G	0.36	intronic	SHROOM3[%$#^], SEPT11[^], FAM47E[^],STBD1[^]	UK	1.04 (1.03-1.05)	4.1x10^{-10}
4q21.22	82587050	rs11099493	A	G	0.69	intergenic	HNRNPD[%], RASGEF1B[#]	UK	1.04 (1.03-1.06)	5.1x10^{-10}
4q22.3	96117371	rs3775058	A	T	0.23	intronic	UNC5C[%#]	MA	1.04(1.03-1.05)	7.6x10^{-9}
4q32.3	169687725	rs7696431	T	G	0.51	intronic	PALLD[%#^], DDX60L[^]	UK	1.04 (1.02-1.05)	2.7x10^{-8}
5p15.31	9556694	rs1508798	T	C	0.81	intergenic	SEMA5A[%$#], TAS2R1[^]	CA	1.05 (1.04-1.07)	4.8x10^{-13}
5q11.2	55860781	rs3936511	A	G	0.82	intronic	MAP3K1[%#^],MIER3[#]	MA	0.96(0.95-0.98)	3.7x10^{-8}
6p25.3	1617143	rs9501744	T	C	0.13	intergenic	FOXC1[%]	CA	0.95 (0.93-0.96)	2.2x10^{-8}
6p21.2	36638636	rs1321309	A	G	0.49	intergenic	CDKN1A[%$#],PI16[#]	MA	1.03(1.02-1.04)	3.4x10^{-8}
6p21.1	43758873	rs6905288	A	G	0.57	intergenic	VEGFA[%#],MRPL14[^],TMEM63B[^]	UK	1.05 (1.03-1.06)	1.9x10^{-12}
6p11.2	57160572	rs9367716	T	G	0.32	intergenic	PRIM2[%], RAB23[$], DST[^], BEND6[^]	CA	0.96 (0.95-0.97)	9.6x10^{-10}
6q14.1	82612271	rs4613862	A	C	0.53	intergenic	FAM46A[%#^]	MA	1.03(1.02-1.04)	6.5x10^{-10}
6q22.32	126717064	rs1591805	A	G	0.49	intergenic	CENPW[%$]	UK	1.04 (1.03-1.06)	2.1x10^{-10}
6q25.1	150997401	rs17080091	T	C	0.08	intronic	PLEKHG1[%#],IYD[^]	MA	0.95(0.93-0.96)	6.0x10^{-9}
7p22.3	1937261	rs10267593	A	G	0.20	intronic	MAD1L1[%$]	MA	0.96(0.95-0.98)	1.8x10^{-8}
7p22.1	6486067	rs7797644	T	C	0.23	intronic	DAGLB[%*$],RAC1[$#^],FAM220A[$],KDELR2[#]	MA	0.96(0.95-0.98)	2.1x10^{-8}
7p21.3	12261911	rs11509880	A	G	0.36	intronic	TMEM106B[%$], THSD7A[^]	CA	1.04 (1.02-1.05)	2.8x10^{-8}
7p13	45077978	rs2107732	A	G	0.09	missense	CCM2[%*$],MYO1G[^]	MA	0.94(0.93-0.96)	3.6x10^{-8}
7q31.2	117332914	rs975722	A	G	0.60	intergenic	CTTNBP2[%],CFTR[#],ASZ1[^]	MA	0.97(0.96-0.98)	4.1x10^{-8}
8p22	18286997	rs6997340	T	C	0.31	intergenic	NAT2[%^]	CA	1.04 (1.02-1.05)	4.9x10^{-9}
8p21.3	22033615	rs6984210	C	G	0.94	intronic	BMP1[%$#], SFTPC[#], DMTN[$],PHYHIP[$],DOK2[^], XPO7[^]	UK	0.92 (0.90-0.94)	2.1x10^{-11}
8q23.1	106565414	rs10093110	A	G	0.42	intronic	ZFPM2[%#^]	MA	0.97(0.96-0.98)	1.8x10^{-8}
9q31.2	110517794	rs944172	T	C	0.72	intergenic	KLF4[%#]	UK	0.96 (0.95-0.97)	1.1x10^{-11}

9q33.2	124420173	rs885150	T	C	0.73	intronic	*DAB2IP*$^{\%\$\#\wedge}$	MA	0.97(0.95-0.98)	7.8x10^{-10}
10p13	12303813	rs61848342	T	C	0.64	intergenic	*CDC123*$^{\%}$,*NUDT5*$^{\$}$,*OPTN*$^{\wedge}$	MA	0.96(0.95-0.98)	6.3x10^{-10}
10q23.1	82251514	rs17680741	T	C	0.72	intronic	*TSPAN14*$^{\%\$\#}$, MAT1A$^{\#}$,*FAM213A*$^{\wedge}$	UK	1.05 (1.03-1.06)	2.3x10^{-11}
10q24.33	105693644	rs4918072	A	G	0.27	intergenic	*STN1*$^{\%\$}$, *SH3PXD2A*$^{\#}$	UK	1.04 (1.03-1.06)	2.6x10^{-9}
10q26.13	124237612	rs4752700	A	G	0.55	intronic	*HTRA1*$^{\%\#}$,*PLEKHA1*$^{\wedge\#}$	MA	0.97(0.96-0.98)	8.0x10^{-11}
11p15.4	5701074	rs11601507	A	C	0.07	missense	*TRIM5*$^{\%*}$, *TRIM22*$^{\%}$, *TRIM6*$^{\wedge}$, *OR52N1*$^{\wedge}$, *OR52B6*$^{\wedge}$	UK	1.09 (1.06-1.11)	2.1x10^{-12}
11p11.2	43696917	rs7116641	T	G	0.69	intergenic	*HSD17B12*$^{\%}$	MA	0.97(0.96-0.98)	1.0x10^{-8}
11q22.1	100624599	rs7947761	A	G	0.72	intronic	*ARHGAP42*$^{\%\#}$	CA	0.96 (0.95-0.97)	3.0x10^{-9}
12p13.31	7175872	rs11838267	T	C	0.87	intronic	*C1S*$^{\%*\#}$	MA	1.05(1.04-1.07)	6.1x10^{-10}
12q22	95355541	rs7306455	A	G	0.10	intergenic	*NDUFA12*$^{\%}$, *FGD6*$^{\#}$	CA	0.95 (0.93-0.96)	1.0x10^{-8}
13q13.1	33058333	rs9591012	A	G	0.34	intronic	*N4BP2L2*$^{\$\wedge}$,*PDS5B*$^{\$\#}$	CA	0.96 (0.94-0.97)	7.0x10^{-11}
13q34	113631780	rs1317507	A	C	0.26	intronic	*MCF2L*$^{\%\$}$, *PCID2*$^{\wedge}$, *CUL4A*$^{\wedge}$	CA	1.04 (1.03-1.06)	8.4x10^{-12}
14q23.1	58794001	rs2145598	A	G	0.58	intronic	*ARID4A*$^{\%\$\#}$,*PSMA3*$^{\wedge}$	MA	0.97(0.96-0.98)	4.2x10^{-8}
14q32.13	94838142	rs112635299	T	G	0.02	intergenic	*SERPINA2*$^{\%\#}$,*SERPINA1*$^{\%*}$	MA	0.87(0.84-0.91)	8.4x10^{-10}
15q26.2	96146414	rs17581137	A	C	0.75	intergenic		MA	1.04(1.02-1.05)	1.2x10^{-8}
16q23.3	81906423	rs7199941	A	G	0.40	intronic	*PLCG2*$^{\%\#\wedge}$, *CENPN*$^{\$}$	CA	1.04 (1.03-1.05)	9.2x10^{-13}
17q11.2	27941886	rs13723	A	G	0.51	UTR3	*CORO6*$^{\%\$}$, *ANKRD13B*$^{\%\$}$, *GIT1*$^{\$}$, *SSH2*$^{\#\$\wedge}$, *EFCAB5*$^{\wedge}$	CA	0.96 (0.95-0.97)	5.6x10^{-10}
17q11.2	30033514	rs76954792	T	C	0.22	intergenic	*COPRS*$^{\%\$}$,*RAB11FIP4*$^{\wedge}$	MA	1.04(1.03-1.05)	1.1x10^{-8}
17q21.2	40257163	rs2074158	T	C	0.82	missense	*DHX58*$^{\%*\$}$, *KAT2A*$^{\%\#}$,*RAB5C*$^{\$}$#, *NKIRAS2*$^{\wedge}$, *DNAJC7*$^{\wedge}$, *KCNH4*$^{\wedge}$, *HCRT*$^{\wedge}$, *GHDC*$^{\wedge}$	CA	0.95 (0.93-0.96)	2.2x10^{-10}
18q21.1	47229717	rs9964304	A	C	0.72	intergenic	*ACAA2*$^{\%\#}$,*RPL17*$^{\wedge}$	MA	0.96(0.95-0.97)	1.1x10^{-9}
19p13.11	17855763	rs73015714	C	G	0.80	intergenic	*MAP1S*$^{\%\#}$, *FCHO1*$^{\%\$}$, *COLGALT1*$^{\wedge}$	CA	0.94 (0.93-0.96)	8.3x10^{-14}
20q12	39924279	rs6102343	A	G	0.25	intronic	*ZHX3*$^{\%}$,*PLCG1*$^{\$\#\wedge}$,*TOP1*$^{\#\wedge}$	MA	1.04(1.02-1.05)	1.1x10^{-8}
20q13.12	44586023	rs3827066	T	C	0.14	intronic	*PCIF1*$^{\%\#}$,*ZNF335*$^{\%\#}$,*NEURL2*$^{\$}$,*PLTP*$^{\$\#}$	MA	1.04(1.03-1.06)	4.4x10^{-9}
20q13.32	57714025	rs260020	T	C	0.13	intergenic	*ZNF831*$^{\%}$	MA	1.05(1.04-1.07)	7.9x10^{-10}
21q21.3	30533076	rs2832227	A	G	0.82	intronic	*MAP3K7CL*$^{\%\$}$,*BACH1*$^{\#}$	UK	0.96 (0.94-0.97)	1.7x10^{-9}

List of novel CAD associated replicating (P<0.05 Bonferroni adjusted, direction of effect consistent) and surpassing the genome-wide significance threshold in meta-analysis of UK Biobank and CARDIoGRAMplusC4D. Full details are shown in Online Tables I, III-XII. Results are shown for the discovery, replication, and combined meta-analysis. OR; odds ratio. Resource; UK; UK Biobank as discovery and CARDIoGRAMplusC4D as replication, CA; CARDIoGRAMplusC4D as discovery and UK biobank as replication, MA; Genome wide significant in the GWAS meta-analysis. % nearest; * coding variants; $ eQTL; # Depict; ^ Hi-C.

Next, we re-analyzed the data from the MetaboChip meta-analysis of CARDIoGRAMplusC4D[9], the CARDIoGRAMplusC4D 1000 genomes meta-analysis[7], and the CARDIoGRAM Exome array data[16] to identify the promising SNPs (P<0.0001). We identified 568 promising SNPs located in 375 loci (Online Table IV). 113 independent SNPs in 96 loci both replicated (P<0.05 Bonferroni adjusted) in stage 2, UK biobank, and achieved genome wide significance in meta-analysis ($P<5\times10^{-8}$), including 21 additional novel loci (Table 1 and Online Table V).

Finally, we performed a meta-analysis of CARDIoGRAMplusC4D[9], the CARDIoGRAMplusC4D 1000 genomes meta-analysis[7] with UK biobank and identified 30 additional loci for which no replication test was available (Table 1, Online Table VI) increasing the total number of genome wide significant CAD loci to 161 (Online Figure II). The novel variants were common (>5%, except for 1, rs112635299 near *SERPINA1*). Online Figure III shows the regional associations plot of each novel locus. For some variants, a dominant or recessive linkage model appears to be a better fit compared to an additive model (Online Table VII). Complete summary statistics of all SNPs in UK Biobank and the UK Biobank-CARDIoGRAMplusC4D meta-analysis are available as download on www.cardiomics.net.

Candidate genes and deeper insights into biology.

To disentangle whether associations were driven more by acute myocardial infarction as opposed to stable CAD we performed multinomial logistic regression analyses for all genome wide significant ($P<5\times10^{-8}$) loci in UK Biobank. 16,875 individuals were only diagnosed with CAD and 17,666 also with myocardial infarction. None of the novel loci and only two previously identified variant (rs9349379, rs10947789) appear to be mainly driven by its association with myocardial infarction rather than stable CAD (FDR P<0.05; Online Table VIII).

We further explored the potential biology of the 64 novel CAD associated loci by prioritizing 155 candidate causal genes in these loci: 69 genes were in proximity (the nearest gene and any additional gene within 10kb) of the lead variant, 9 genes contained coding genetic variation in LD ($r^2>0.8$) with the lead variant (Online Table IX), 50 genes were selected based on expression quantitative trait loci (eQTL) analyses (Online Table X), 64 genes showed significant chromatin-interactions (Hi-C) between the genetic variant and promoter of the gene (Online Table XI), and 60 genes were prioritized based on DEPICT analyses (Online Table XII). Of the 155 candidate genes, 63 were prioritized by multiple methods of identification, which may be used to prioritize candidate causal genes. A summary of the current function annotation of each novel candidate gene is provided in Online Table XIII and knowledge on pharmacologic compounds and nutrients influencing these genes is provided in Online Table XIV. Next, we performed a systematic search in the Mouse Genome Informatics (MGI) database to identify the effect of mutations in orthologous genes for these candidate causal genes (details in Online Table XV). In brief, we identified 34 genes that expressed at least one cardiovascular system phenotype (*AGT, ARHGAP42, BACH1, CALCRL, CASQ2, CCM2, CDC123, CDKN1A, FIGN, FOXC1, GIT1, GNPAT, HCRT, HSD17B12, MAP1S, MAP3K1, MSANTD1, NGF, NPHP3, PCIF1, PDS5B, PLCG1, PLEKHA1, PPP2R3A, PRDM16, PRKCE, RAC1, SEMA5A, SH3PXD2A, TFPI, TIPARP, TMEM106B, VEGFA, ZFPM2*) and 34 genes that affected other potentially plausible traits linked to CAD, including metabolic/lipid/adipose/weight abnormalities (*AGT, CORO6, FIGN, GIT1, KAT2A, NGF, PPP2R3A, NPHH3, SH3PXD2A, TMEM106B, VEGFA, ZHX3, OPTN, FAM213A, DNAJC7, COPRS*), abnormalities in inflammation or white blood cells (*DHX58, FHL3, HNRNPD, PLCG2, PRDM16, TFPI, VEGFA, ZNF335, PRKCE, MYO1G, RAC1, ARID4A*), and abnormalities in platelets or coagulation (*FHL3, PLCG2, TFPI, VEGFA, DST, KLF4*).

Novel insights from pathway analyses.

Ingenuity pathway analysis (IPA) restricted to the 155 candidate causal genes confirmed that these are enriched for effects on the cardiovascular system and cell cycle functions (Online Table XVI). Pathway insights provided by the DEPICT framework identified 1,525 reconstituted genes sets that could be captured in 156 meta gene sets (Online Table XVII). The 4 most significant meta-sets were "complete embryonic lethality during organogenesis", "blood vessel development", "Anemia" and "SRC PPI subnetwork". The "platelet alpha granule lumen", "SRC PPI subnetwork", "blood vessel development" and "hemostasis" had the largest betweenness centrality, an indicator of a node's centrality in the network. The tissue enrichment analyses by DEPICT indicated blood vessels as the most relevant tissue ($P= 4\times10^{-7}$), 41 additional tissues or cell types were significantly enriched at FDR<0.05 (Online Table XVIII). We compared the contribution of novel information to previous work. The previous CardiogramPlusC4D analysis led to 457 reconstituted gene-sets (at FDR<0.05), the addition of the intermediate dataset UK Biobank of 150k individuals identified a total of 889 significant gene sets, substantially less than the current 1,525 gene sets (Figure 1; Online Table XVII). Considering all 10,968 possible gene sets, this study represents an increase from 4.16% to 13.90% of all gene sets involved in CAD since the 1000 Genomes analysis of CardiogramPlusC4D in 2015. Genes implicated by DEPICT on the FDR<0.05 level are 94 in the previous data which has increased to 540 genes.

Insights in loci by associations with other phenotypes.

To increase our understanding of potentially mediating mechanisms at the genetic variant level we searched the GWAS-catalog for previously reported variants. Of the 64 novel loci, 23 loci were in LD (r^2>0.6) with genetic variants previously reported to be associated with other traits surpassing the genome wide significant ($P<5\times10^{-8}$) threshold (Online Table XIX). We found associations with anthropometric measurements (rs6905288, rs1591805, rs3936511, rs840616), anti-neutrophil antibody associated vasculitis (rs112635299), angiotensinogen measurements (rs699), coffee consumption (rs13723), C-reactive protein (rs667920), pulmonary function (rs61848342, rs13723, rs112635299), fibrinogen levels (rs67920, rs16844401, rs2074158), glomerular filtration rate (rs12500824), HDL cholesterol (rs667920, rs10512861, rs6905288), LDL cholesterol (rs10512861), total cholesterol (rs6997340), triglycerides (rs667920, rs3936511, rs6905288, rs6997340), diabetes (rs1591805, rs3936511), blood pressure indices (rs260020, rs17080091, rs61776719, rs7696431, rs1317507), transferrin levels (rs6997340), QRS amplitude (rs13723), abdominal aortic aneurysm (rs885150, rs3827066), adiponectin measurements (Rs6905288), and age at menarche (rs1591805); full details can be found in Online Table XIX. We also explored the association of the 64 lead SNPs with a range of traits in UKbiobank Resource. Consistent with the GWAS-catalog search and in keeping with earlier observations in established CAD loci, several of our novel loci were associated with hyperlipidemia, blood pressure traits, diabetes and anthropometric traits (Figure 2). For example, rs6905288 (*VEGFA*) was also associated with waist-to-hip ratio and hyperlipidemia and rs61776719 (*FHL3, UTP11L*) was also closely associated with pulse pressure in UK biobank. Interestingly, we observed that 15 of 64 loci were associated with platelet counts.

Genetic risk for CAD and association with CAD risk factors and outcome.

To explore potential clinical relevance, we constructed a genetic risk score (GRS), weighted for their effects in CardiogramPlusC4D by multiplying the effect sizes with the number of effect variants of each variant in each individual and divided this GRS into quintiles. The associations with many different traits and diseases from the UK Biobank are visualized in Figure 2. The risk of a future diagnosis of atrial fibrillation and heart failure in UK Biobank participants was higher in quantile 5 individuals as compared to quantile 1 (HR 1.18 [95CI 1.10-1.27, P=1.2x10^{-6}] and HR 1.59 [95%CI 1.43-1.77, P=3.3x10^{-18}], respectively - Online Figure IV). In addition, all-cause mortality and especially cardiovascular mortality was higher in individuals of quantile 5 compared to quantile 1 (HR 1.12 [95%CI 1.06-1.19, P=4x10^{-4}] and HR 1.94 [95%CI 1.70-2.21, P=2x10^{-23}], respectively - Online Figure IV).

Role of regulatory DNA and fine-mapping of candidate causal variants.

Across the genome, virtually all tissues showed significant enrichment of DNase I hypersensitivity sites providing limited indications for involved biology (**Figure 3a** and **b**). Minimal differential enrichment of functional elements for the identified genetic loci was observed in blood vessels and liver. To facilitate future functional studies directed at causal variants and molecular mechanisms, we prioritized variants via the probabilistic framework of Probabilistic Annotation INTegratOR (PAINTOR). As no clear differential enrichment was observed for tissue specific functional elements, we focused on DNA annotations from the study of Finucane *et al*[17] that are not specific for tissue or cell types. PAINTOR determined the significance of each annotation to be causal (Figure 3c and d) and a model was constructed using LD information, *P*-value distribution, and information on coding variation, conservation and H3K4me1 sites to prioritize potential causal SNPs of all 161 (known and novel) loci. This analysis yielded 28 variants at or above the 95% confidence level for which we prioritized candidate genes (Online Table XX, Table 2).

Table 2. For 28 loci, the 95% credible set of causal variants consisted of a single CAD variant.

Cytoband	Causal variant	#SNPs in locus	MAF (EUR)	GWAS P	Posterior P	Annotation	Candidate gene/mechanism
1p32.3	rs11591147	4	0.02	1.9x10^{-22}	1.00	missense (T)	***PCSK9****
1p13.3	rs602633	67	0.21	3.6x10^{-58}	1.00	downstream	***SORT1***$^{\$(130)\wedge(51.1)}$,*SARS*$^{\$(11.6)}$,*PSRC1*$^{(152)}$,*CELSR2*$^{\$(108)}$, *ATXN7L2*$^{\$(11.7)}$
1q32.1	rs6700559	83	0.49	1.8x10^{-08}	0.97	intronic	***CAMSAP2***$^{\$(9.5)\wedge(23.9)}$, *DDX59*$^{\$(42.0)}$
2p24.1	rs16986953	66	0.07	1.1x10^{-16}	1.00	intergenic	-
2q35	rs2571445	50	0.40	1.6x10^{-12}	0.97	missense (T)	***TNS1***$^{*\$(121.5)}$,*DIRC3*$^{\$(7.9)}$
3p21.31	rs7633770†	49	0.44	1.1x10^{-08}	0.97	intergenic	*ALS2CL*$^{\$(8.6)}$,*RTP3*$^{\wedge(49.6)}$,*LTF*$^{\wedge(49.6)}$
3q26.31	rs12897†	67	0.41	1.2x10^{-09}	1.00	UTR3	***FNDC3B***$^{\&\$(8.8)}$
4q21.22	rs11099493†	54	0.37	2.5x10^{-10}	1.00	intergenic	-
5q31.3	rs246600	15	0.50	6.5x10^{-17}	1.00	intronic	*HMHB1*$^{\wedge(159.2)}$
6p24.1	rs9349379	268	0.41	2.7x10^{-76}	1.00	intronic	***EDN1***$^{\$(2,2)\wedge(23.9)}$,*TBC1D7*$^{\$(15.9),}$*PHACTR1*$^{\$(55.7)}$,*GFOD1*$^{\$(8.1)}$
7p21.1	rs2107595	71	0.18	1.3x10^{-24}	1.00	intergenic	TWIST1$^{\$(36.8)}$
7p13	rs2107732†	11	0.10	3.6x10^{-08}	0.98	missense (T)	***CCM2***$^{*\wedge(9.6)}$,*MYO1G*$^{\wedge(22.9)}$
7q32.2	rs11556924	95	0.38	1.4x10^{-23}	1.00	missense (D)	***ZC3HC1****,*NRF1*$^{\wedge(38)}$,*KLF14*$^{\wedge(216.9)}$
7q36.1	rs3918226	25	0.09	1.4x10^{-20}	1.00	intronic	*NOS3*$^{\$(6.0)}$
9p21.3	rs4977574	161	0.49	8.8x10^{-223}	1.00	intronic	***CDKN2B***$^{\$(4.7),\wedge(133)}$, *MTAP*$^{\wedge(168)}$
11p15.4	rs11601507†	3	0.06	5.6x10^{-13}	1.00	missense (D)	***TRIM5****,OR52N1$^{\wedge(45)}$,*TRIM6*$^{\wedge(49)}$,*OR52B6*$^{\wedge(49)}$
11q13.1	rs3741380	207	0.48	2.8x10^{-11}	0.95	missense (T)	***EHBP1L1***$^{*\$(51.5)}$
11q13.5	rs590121	122	0.28	1.5x10^{-10}	0.98	intronic	*SERPINH1*$^{\$(5.5)}$,*KLHL35*$^{\wedge(23.7)}$
11q22.3	rs974819	428	0.24	1.1x10^{-28}	0.99	intergenic	***PDGFD***$^{\$(20.6),\wedge(87.0)}$
13q34	rs11617955	19	0.11	6.9x10^{-18}	1.00	intronic	-
13q34	rs1317507†	94	0.25	8.2x10^{-12}	1.00	intronic	*PCID2*$^{\wedge(22.5),}$ *CUL4A*$^{\wedge(22.5)}$
15q25.1	rs7173743	367	0.46	5.5x10^{-36}	0.96	intergenic	*RASGRF1*$^{\$(4.2),}$*ADAMTS7*$^{\$(33.3)}$
16q23.3	rs7500448	257	0.22	1.6x10^{-16}	1.00	intronic	***CDH13***$^{\$(70.2)\wedge(64.8)}$
17q21.32	rs17608766	178	0.17	8.2x10^{-10}	1.00	UTR3	*GOSR2*$^{\&}$
19p13.2	rs116843064	7	0.03	3.6x10^{-10}	1.00	missense (D)	***ANGPTL4****
19q13.32	rs7412	39	0.07	2.1x10^{-35}	1.00	missense (D)	***APOE****,*APOC2*$^{\wedge(64.1)}$,*CLPTM1*$^{\wedge(64.1)}$,*APOC4*$^{\wedge(64.1)}$
20q11.22	rs867186	>500	0.10	6.8x10^{-12}	0.97	missense (T)	***PROCR***$^{*\$(20.3)}$,***TRPC4AP***$^{\$(42.9)\wedge(116)}$,*GGT7*$^{\$(4.6)}$,*EDEM2*$^{\$(7.9)}$,*NCOA6*$^{\wedge(75.1)}$,*HMGB3P1*$^{\wedge(75.1)}$
21q22.11	rs28451064	104	0.13	2.6x10^{-33}	1.00	intergenic	***MRPS6***$^{\$(17.4)\wedge(238.6)}$, ***SLC5A3***$^{\$(32.1)\wedge(238.6)}$

*= Gene with a missense causal variant; ^=chromatin-interaction between the causal variant and the promotor of the gene; &=Gene of which the three prime untranslated region overlaps with the causal variant; $=eQTL gene; ()=the number between brackets indicates the significance –log(P), of the eQTL or chromatin interaction. †=SNPs of novel loci. *D=deleterious (SIFT), T=Tolerated(SIFT)*. The column '#SNPs in locus' corresponds

to the number of SNPs with a *P*<0.01 that are in low LD (r^2>0.1) with the sentinel SNP. Genes in bold indicates converging evidence of a potential functional SNP-gene mechanism, further described in the methods. Online Table XX contains full details of the loci on the variant level.

For example, rs974819 was prioritized as causal variant and could be linked to PDGFD by hi-C evidence and eQTL data in relevant tissues (Online Figure V). In total 15 of the 28 fine-mapped loci could be pinpointed to one single potential causal mechanism implicating a single variant. For two loci, there were 2 potential causal mechanisms (*TRPC4AP*/*PROCR* and *MRPS6/SLC5A3*) with equal evidence.

DISCUSSION

The present study is the largest genetic association study of CAD performed to date. We report on the primary results and downstream bioinformatic analyses of the meta-analysis of *de novo* GWAS data derived from the UK Biobank combined with existing data from CARDIoGRAMplusC4D, leading to the inclusion of up to 122,733 cases and 424,528 controls. This study contributes to the existing literature by reporting 64 *novel* genetic loci representing 38% of all 161 GWAS identified CAD loci to date[18]. For the novel loci, a detailed catalogue of 155 candidate genes (based on proximity, gene-expression data, coding variation and physical chromatin interaction) is provided. We demonstrate that the increase in significantly associated CAD loci results in a large expansion of implicated reconstituted gene-networks, from 4% to almost 14%. Finally, by integrating genetic association strength, LD and functional annotation data, we performed fine-mapping of all 161 CAD loci, providing a novel credible list of causal variants and plausible genes to be prioritized for functional validation.

The 64 novel genetic loci reported in this single manuscript is exceptionally large compared to previous manuscripts, including those of CARDIoGRAMplusC4D and others reporting on 10-15 novel loci each[2–10]. 34 of the 64 loci are significant in a robust reciprocal replication strategy between CARDIoGRAMplusC4D and the UK Biobank but another 30 are genome-wide significant in the overall meta-analysis as is commonly considered sufficient evidence[7,10]. The obvious reason for the large number of novel loci is the considerable number of novel CAD cases and non-CAD controls compared to these earlier efforts combined with less heterogeneity in samples, collection and definitions used. By increasing the sample size, more loci can be identified, more genes can be implicated and more gene-networks or pathways can be constructed. Not only is the increase of associated loci in the past decade rapidly outpacing functional validation, even understanding biological networks appears to insufficiently accommodate the increased amount of GWAS hits under the conceptual "*polygenetic model*". This can be illustrated by the large increase of reconstituted gene-networks observed in our study. For the first time, we show that almost 14% of all existing gene-networks are involved in the complex CAD trait (Figure 1), and this will only increase when further samples are added to the GWAS study making it increasingly more difficult to consider these all to be "key pathways". In our data, we also observed genetic associations signals to be spread across most of the genome, and many of the novel 155 candidate genes do not have an obvious connection to CAD. In addition, virtually all cell types showed significant enrichment of DNase I hypersensitivity and other functional elements. These notions are all supportive of the "*omnigenic model*" which has recently been proposed by the Pritchard's team suggesting that prevailing conceptual models for complex diseases are incomplete. The Omnigenic model hypothesizes that all gene regulatory networks are sufficiently interconnected such that all genes expressed in disease-relevant cells can influence the function of core disease-related genes and a major proportion of heritability can be explained by effects of genes outside key pathways[19]. To further our knowledge, it is questionable whether further increasing the GWAS sample size will resolve the outstanding issues concerning our incomplete understanding of cellular regulatory networks and our ability to differentiate core genes from peripheral genes. If the omnigenic model is indeed correct, detailed mapping of cell-specific regulatory networks will be essential to understand CAD.

To facilitate functional research based on our findings, we not only provided extensive bioinformatic analyses of coding variation, gene-expression and chromatin interactions for the 64 novel loci, we also performed novel fine-mapping and presented statistically convincing arguments for causal genetic variants at 28 loci, linking 19 genes in the 161 CAD loci. In the known loci, these genes included *APOE*, *PCSK9, ANGPTL4,* and *SORT1*, all implicated as core-genes in lipid-metabolism. Recently, *PCSK9* has been validated in clinical trials[20], and functional studies are also supporting a key role for *SORT1*[21]. More recently, *EDN1* has indeed been identified as the likely causal gene in the pathogenesis of CAD instead of the nearby *PHACTR*[22]. In the novel loci, we found evidence for causal variants linked to *FNDC3B*

(Fibronectin Type III Domain Containing 3B), *CCM2* (CCM2 Scaffolding Protein), and *TRIM5* (Tripartite Motif Containing 5). Indeed, the functional link between these genes and CAD is not obvious and remains to be determined. *FNDC3B* has been suggested to function as a positive regulator of adipogenesis.[23] *CCM2* has been implicated in abnormal vascular morphogenesis in the brain, leading to cerebral cavernous malformations[24] but is also expressed in the heart. Although its effect in the coronary arteries has not been investigated but *Ccm2* knockdown in the mouse brain endothelial cells leads to increased monolayer permeability, decreased tubule formation, and reduced cell migration following wound healing[25]. *TRIM5* has been suggested to promote innate immune signaling and its activity is amplified by retroviral infections[26]. All SNP-gene mechanisms proposed in this manuscript should be experimentally sought out. Also, the analyses were restricted to variants available in the HRC imputation panel. Although this is the largest imputation panel to date it is only comprised of SNPs; future fine-mapping efforts are necessary that include non-SNPs as well, such as indels, to cover the additional aspects of the human variation landscape. However, a 95% credible set that contains just 1 potential causal variant per locus provides a first starting point for generating new hypotheses and scientific explorations.

In our current work, we validated our previous finding that these genetic variants of CAD also predict the risk of atrial fibrillation, heart failure[8] and extended it to all cause death. We also aimed to differentiate between stable CAD and acute myocardial infarction by performing multinomial logistic regression analyses. Most loci were not driven by one clinical presentation specifically. However, for two previously identified loci (rs9349379 (*EDN1*) and rs10947789 (*KCNK5*)) we found statistical evidence that these loci may be driven by acute myocardial infarction and not stable CAD. Also for this observation, functional hypotheses are to be developed and tested. Our variants might be driven mainly by non-fatal CAD and different variants might exist for fatal heart disease.

Some limitations of the current work are to be acknowledged. This work is based on statistical evidence and does not provide functional experimental validation. The genetic variants identified and the genes prioritized require further direct investigations in future studies to elucidate their role, and function, in the development and progression of CAD. However, in the short term, these data open up new possibilities to improve quantitative measures of genetic risk prediction. Recent data suggests that instead of operating in a deterministic fashion, high genetic risk is indeed modifiable by lifestyle[27], pharmacotherapy[28], and also by incorporation of genetic risk into shared decision making sessions with patients[29].

In conclusion, our GWAS, meta-analyses, and bioinformatic analyses provide several novel insights into the biology of CAD. We report 64 novel loci, link 155 candidate genes and performed fine-mapping of all old and novel loci, providing a credible list of causal genetic variants. However, with the ever-increasing sample size, our work is the first to indicate that an omnigenic model may be more appropriate to accommodate the complex genetic architecture of CAD, compared to a polygenic model. In addition to an expanded view, it also suggests new methods and tools are required to further our understanding of CAD biology through genetics.

ACKNOWLEDGEMENT

This research has been conducted using the UK Biobank Resource under Application Number 12006. We thank the CARDIoGRAMplusC4D investigators for making publicly available their data. We would like to thank the Center for Information Technology of the University of Groningen for their support and for providing access to the Peregrine high-performance computing cluster.

SOURCES OF FUNDING

N.V. is supported by Marie Sklodowska-Curie GF (call: H2020-MSCA-IF-2014, Project ID: 661395) and an NWO VENI grant (016.186.125). We acknowledge the support from the Netherlands Cardiovascular Research Initiative: an initiative with support of the Dutch Heart Foundation, CVON2015-17 EARLY-SYNERGY.

DISCLOSURES

Nothing to disclose.

REFERENCES

1. Wang H, Naghavi M, Allen C, Barber R, Bhutta ZA, Carter C, Casey C, Charlson F, Chen C, Coates M, Dandona H. Global, regional, and national life expectancy, all-cause mortality, and cause-specific mortality for 249 causes of death, 1980-2015: a systematic analysis for the Global Burden of Disease Study 2015. *Lancet*. 2016;388:1459–1544.
2. Samani NJ, Erdmann J, Hall AS, et al. Genomewide association analysis of coronary artery disease. *N Engl J Med*. 2007;357:443–453.
3. Helgadottir A, Thorleifsson G, Manolescu A, et al. A common variant on chromosome 9p21 affects the risk of myocardial infarction. *Science*. 2007;316:1491–1493.
4. McPherson R, Pertsemlidis A, Kavaslar N, Stewart A, Roberts R, Cox DR, Hinds DA, Pennacchio LA, Tybjaerg-Hansen A, Folsom AR, Boerwinkle E, Hobbs HH, Cohen JC. A common allele on chromosome 9 associated with coronary heart disease. *Science*. 2007;316:1488–1491.
5. Schunkert H, König IR, Kathiresan S, et al. Large-scale association analysis identifies 13 new susceptibility loci for coronary artery disease. *Nat Genet*. 2011;43:333–8.
6. CARDIoGRAMplusC4D Consortium, Deloukas P, Kanoni S, et al. Large-scale association analysis identifies new risk loci for coronary artery disease. *Nat Genet*. 2013;45:25–33.
7. Nikpay M, Goel A, Won H-H, et al. A comprehensive 1,000 genomes-based genome-wide association meta-analysis of coronary artery disease. *Nat Genet*. 2015;47:1121–30.
8. Verweij N, Eppinga RN, Hagemeijer Y, van der Harst P. Identification of 15 novel risk loci for coronary artery disease and genetic risk of recurrent events, atrial fibrillation and heart failure. *Sci Rep*. 2017;7:2761.
9. Howson JMM, Zhao W, Barnes DR, et al. Fifteen new risk loci for coronary artery disease highlight arterial-wall-specific mechanisms. *Nat Genet*. 2017;49:1113–1119.
10. Nelson CP, Goel A, Butterworth AS, et al. Association analyses based on false discovery rate implicate new loci for coronary artery disease. *Nat Genet*. 2017;49:1385-1391.
11. Khera A V., Kathiresan S. Genetics of coronary artery disease: discovery, biology and clinical translation. *Nat Rev Genet*. 2017;18:331–344.
12. Pers TH, Karjalainen JM, Chan Y, et al. Biological interpretation of genome-wide association studies using predicted gene functions. *Nat Commun*. 2015;6:5890.
13. van der Harst P, van Setten J, Verweij N, et al. 52 Genetic loci influencing myocardial mass. *J Am Coll Cardiol*. 2016;68:1435-1448.
14. Iotchkova V, Huang J, Morris JA, et al. Discovery and refinement of genetic loci associated with cardiometabolic risk using dense imputation maps. *Nat Genet*. 2016;48:1303–1312.
15. Kichaev G, Yang W-Y, Lindstrom S, Hormozdiari F, Eskin E, Price AL, Kraft P, Pasaniuc B.

Integrating functional data to prioritize causal variants in statistical fine-mapping studies. *PLoS Genet*. 2014;10:e1004722.

16. Myocardial Infarction Genetics and CARDIoGRAM Exome Consortia Investigators, Stitziel NO, Stirrups KE, et al. Coding variation in ANGPTL4, LPL, and SVEP1 and the risk of coronary disease. *N Engl J Med*. 2016;374:1134–1144.
17. Finucane HK, Bulik-Sullivan B, Gusev A, et al. Partitioning heritability by functional annotation using genome-wide association summary statistics. *Nat Genet*. 2015;47:1228–1235.
18. Klarin D, Martin Zhu Q, Emdin CA, et al. Genetic analysis in UK Biobank links insulin resistance and transendothelial migration pathways to coronary artery disease. *Nat Genet.* 2017;49:1392-1397.
19. Boyle EA, Li YI, Pritchard JK. An expanded view of complex traits: from polygenic to omnigenic. *Cell*. 2017;169:1177–1186.
20. Ridker PM, Revkin J, Amarenco P, et al. Cardiovascular efficacy and safety of bococizumab in high-risk patients. *N Engl J Med*. 2017;376:1527–1539.
21. Musunuru K, Strong A, Frank-Kamenetsky M, et al. From noncoding variant to phenotype via SORT1 at the 1p13 cholesterol locus. *Nature*. 2010;466:714–719.
22. Gupta RM, Hadaya J, Trehan A, et al. A genetic variant associated with five vascular diseases is a distal regulator of endothelin-1 gene expression. *Cell*. 2017;170:522–533.
23. Kishimoto K, Kato A, Osada S, Nishizuka M, Imagawa M. Fad104, a positive regulator of adipogenesis, negatively regulates osteoblast differentiation. *Biochem Biophys Res Commun*. 2010;397:187–191.
24. Liquori CL, Berg MJ, Siegel AM, et al. Mutations in a gene encoding a novel protein containing a phosphotyrosine-binding domain cause type 2 cerebral cavernous malformations. *Am J Hum Genet*. 2003;73:1459–1464.
25. Crose LES, Hilder TL, Sciaky N, Johnson GL. Cerebral cavernous malformation 2 protein promotes smad ubiquitin regulatory factor 1-mediated RhoA degradation in endothelial cells. *J Biol Chem*. 2009;284:13301–13305.
26. Pertel T, Hausmann S, Morger D, et al. TRIM5 is an innate immune sensor for the retrovirus capsid lattice. *Nature*. 2011;472:361–365.
27. Khera A V., Emdin CA, Drake I, et al. Genetic risk, adherence to a healthy lifestyle, and coronary disease. *N Engl J Med*. 2016;375:2349–2358.
28. Mega JL, Stitziel NO, Smith JG, et al. Genetic risk, coronary heart disease events, and the clinical benefit of statin therapy: an analysis of primary and secondary prevention trials. *Lancet*. 2015;385:2264–2271.
29. Kullo IJ, Jouni H, Austin EE, et al. Incorporating a genetic risk score into coronary heart disease risk estimates: effect on low-density lipoprotein cholesterol levels (The MI-GENES Clinical Trial). *Circulation*. 2016;133:1181–1188.

FIGURE LEGENDS

Figure 1. Network analyses of reconstituted gene sets. The total number of significant gene sets involved in CAD increased to 13.90% since the 1000 genome GWAS of CardiogramPlusC4D, considering all possible gene sets. Clustering by modularity using Gephi software indicated that pathways specific for cardiovascular/heart development, inflammation, lipids, kidney and coagulation clustered together. 'PPI networks & Others' indicates a remaining bin predominantly populated by Protein-protein interaction networks.

Figure 2. Heatmap of associations in UK Biobank with novel loci. Heatmap of z-scores for different diseases and phenotypes in UK Biobank, aligned to increased risk of CAD. Only significant associations (FDR<0.01) are shown. The genetic risk score constructed with the known and novel loci, weighted using coefficients of CardiogramPlusC4D, is highlighted by the red rectangle.

Figure 3. The role of regulatory DNA underlying CAD associated SNPs. Enrichment of genome-wide association analysis p-values in DNaseI hypersensitive sites (DHS). CAD SNPs at different GWAS threshold were significantly enriched in DHS footprints **(a)** and hotspots **(b)** across many different tissues and cell types. The fold enrichment was highly significant for most tissues and cell types ($P<1\times10^{-8}$) as indicated by the 4 colored circles next to the labels, 3 colored circles indicates $P<1\times10^{-7}$. Label sizes of tissue types were down sized due to space limitations; tissues-types may be represented by multiple samples, indicated by hash marks of the same color. **(c)** Subsequent prioritization of potential causal annotations underlying the 161 CAD loci also suggested that regions of DHS may be underlying the associations, but coding variants, conservation, 5-prime UTR and H3K4me1 annotations were more likely to be causal. **(d)** Posterior probabilities for causality for each variant in the 164 CAD loci were calculated by an empirical Bayes approach implemented in the Probabilistic Annotation INtegraTOR Framework (PAINTOR), taking into account LD, association statistics and the potentially causal annotations, and summarized in Table 2 and Online Table XX.

NOVELTY AND SIGNIFICANCE

What Is Known?

- Coronary Artery Disease (CAD) is a multifactorial disease with a substantial heritable component.

- Genome wide association studies (GWAS) in the past decade have identified 96 loci associated with CAD and are believed to provide biological insights into "key pathways" under the presumption of a "*polygenetic*" model.

What New Information Does This Article Contribute?

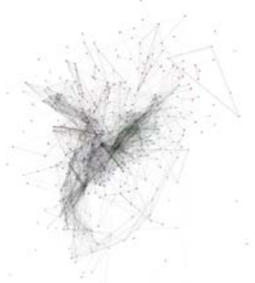
- We have identified 64 additional loci, which were associated with CAD. We fine-mapped all new and known loci to provide evidence in support of the causal role of the genetic variants or genes in CAD.

- Network analyses suggest a complex genetic architecture of CAD, which might not be fully captured by the prevailing "*polygenetic*" model of CAD.

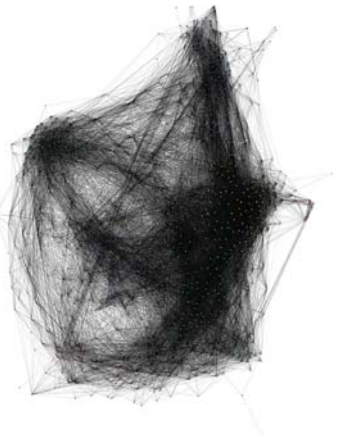
- This work lends supports to the "*omnigenetic*" model, proposing that associated genetic variants might not necessarily lay in key disease pathways. Instead, all gene regulatory networks maybe sufficiently interrelated such that all genes expressed, including those outside key disease pathways, may influence key disease related genes.

CAD, a leading cause of death, is a complex multifactorial disease. GWAS of CAD have offered new biological insights and added to risk prediction and identification of drugable targets. We performed a large systematic meta-analysis of GWAS, involving 122,733 cases and 424,528 controls and identified 64 new genetic loci that were associated with CAD. Fine-mapping of all known and novel CAD loci highlighted potential causal SNP-gene mechanisms. A large proportion of all biological pathways and a plethora of human tissues were found to be associated with CAD for no obvious reason. This finding could indicate that the "*polygenic*" model may not uphold with ever increasing sample sizes for CAD genetics and the "*omnigenic*" model may be more appropriate to accommodate the increasing complexity. This study underscores the importance of tissue-specific dedicated mechanistic studies. New methods and tools are required to advance our understanding of genetic mechanisms influencing the development and progression of CAD.

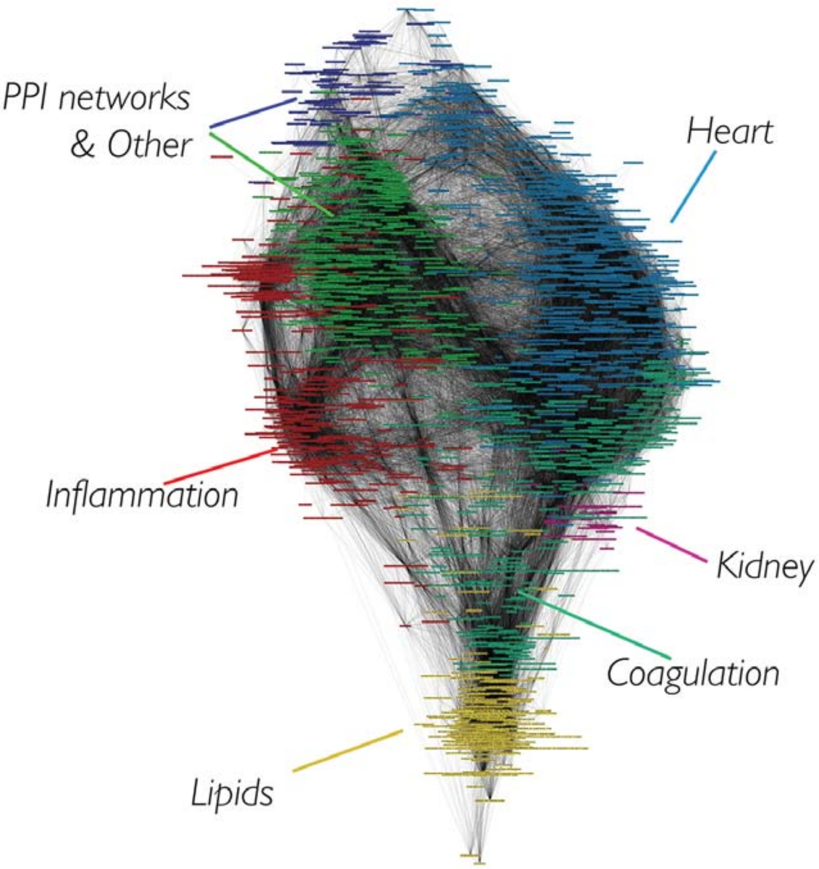
FIGURE 1



CardiogramplusC4D only
458 genesets - 21,836 edges



CardiogramplusC4D & UK Biobank 150k
889 genesets - 66,554 edges



CardiogramplusC4D & UK Biobank 500k (this study)
1,525 genesets - 140,375 edges

FIGURE 2

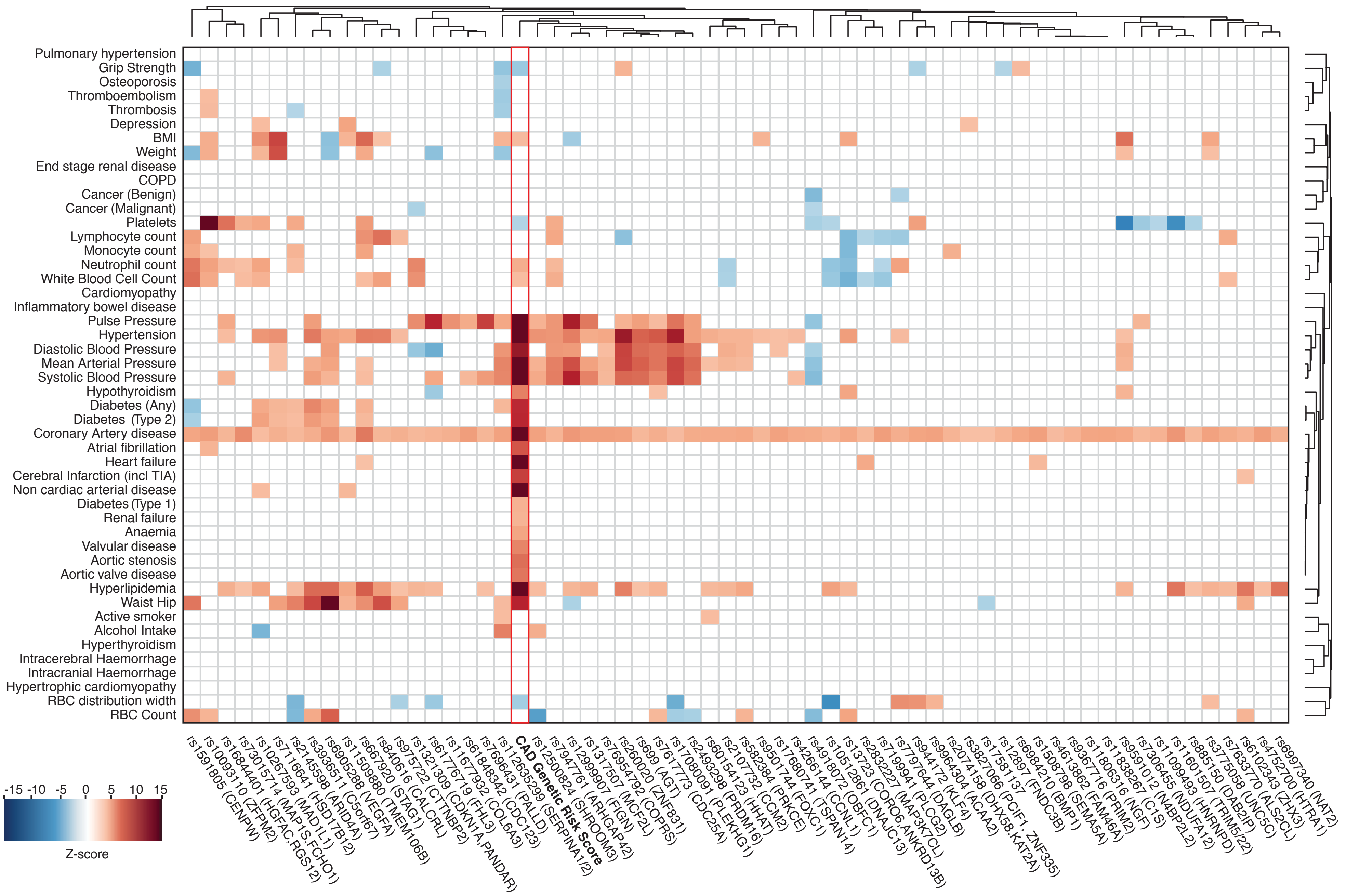
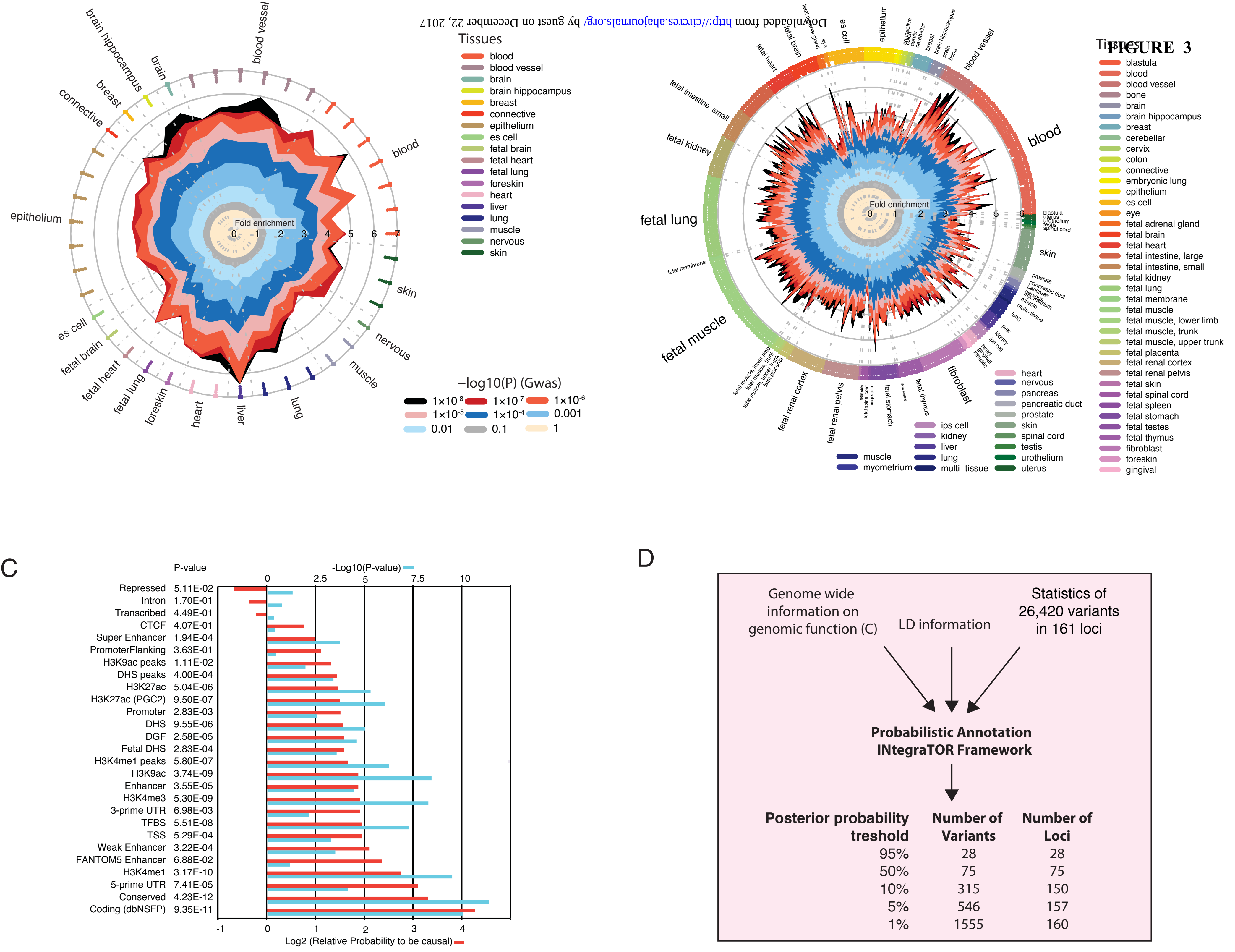

Pulmonary hypertension
Grip Strength
Osteoporosis
Thromboembolism
Thrombosis
Depression
BMI
Weight
End stage renal disease
COPD
Cancer (Benign)
Cancer (Malignant)
Platelets
Lymphocyte count
Monocyte count
Neutrophil count
White Blood Cell Count
Cardiomyopathy
Inflammatory bowel disease
Pulse Pressure
Hypertension
Diastolic Blood Pressure
Mean Arterial Pressure
Systolic Blood Pressure
Hypothyroidism
Diabetes (Any)
Diabetes (Type 2)
Coronary Artery disease
Atrial fibrillation
Heart failure
Cerebral Infarction (incl TIA)
Non cardiac arterial disease
Diabetes (Type 1)
Renal failure
Anaemia
Valvular disease
Aortic stenosis
Aortic valve disease
Hyperlipidemia
Waist Hip
Active smoker
Alcohol Intake
Hyperthyroidism
Intracerebral Haemorrhage
Intracranial Haemorrhage
Hypertrophic cardiomyopathy
RBC distribution width
RBC Count
rs1591805 (CENPW)
rs10093110 (ZFPM2)
rs16844401 (HGFAC,RGS12)
rs73015714 (MAP1S,FCHO1)
rs10267593 (MAD1L1)
rs7116641 (HSD17B12)
rs2145598 (ARID4A)
rs3936511 (C5orf67)
rs6905288 (VEGFA)
rs11509880 (TMEM106B)
rs667920 (STAG1)
rs840616 (CALCRL)
rs975722 (CTTNBP2)
rs1321309 (CDKN1A,PANDAR)
rs61776719 (FHL3)
rs11677932 (COL6A3)
rs61848342 (CDC123)
rs7696431 (PALLD)
rs112635299 (SERPINA1/2)
CAD Genetic Risk Score
rs12500824 (SHROOM3)
rs7947761 (ARHGAP42)
rs12999907 (FIGN)
rs1317507 (MCF2L)
rs76954792 (COPRS)
rs260020 (ZNF831)
rs699 (AGT)
rs7617773 (CDC25A)
rs17080091 (PLEKHG1)
rs2493298 (PRDM16)
rs60154123 (HHAT)
rs2107732 (CCM2)
rs582384 (PRKCE)
rs9501744 (FOXC1)
rs17680741 (TSPAN14)
rs4266144 (CCNL1)
rs4918072 (OBFC1)
rs10512861 (DNAJC13)
rs13723 (CORO6,ANKRD13B)
rs2832227 (MAP3K7CL)
rs7199941 (PLCG2)
rs7797644 (DAGLB)
rs944172 (KLF4)
rs9964304 (ACAA2)
rs2074158 (DHX58,KAT2A)
rs3827066 (PCIF1, ZNF335)
rs17581137
rs12897 (FNDC3B)
rs6984210 (BMP1)
rs1508798 (SEMA5A)
rs4613862 (FAM46A)
rs9367716 (PRIM2)
rs11806316 (NGF)
rs11838267 (C1S)
rs9591012 (N4BP2L2)
rs7306455 (NDUFA12)
rs11099493 (HNRNPD)
rs11601507 (TRIM5/22)
rs885150 (DAB2IP)
rs3775058 (UNC5C)
rs7633770 (ALS2CL)
rs6102343 (ZHX3)
rs4752700 (HTRA1)
rs6997340 (NAT2)
-15 -10 -5 0 5 10 15
Z-score

FIGURE 3



Genome wide information on genomic function (C)

LD information

Statistics of 26,420 variants in 161 loci

Probabilistic Annotation INtegraTOR Framework

Posterior probability treshold	Number of Variants	Number of Loci
95%	28	28
50%	75	75
10%	315	150
5%	546	157
1%	1555	160

Circulation Research
JOURNAL OF THE AMERICAN HEART ASSOCIATION

American Heart Association®

The Identification of 64 Novel Genetic Loci Provides an Expanded View on the Genetic Architecture of Coronary Artery Disease

Pim van der Harst and Niek Verweij

Circ Res. published online December 6, 2017;
Circulation Research is published by the American Heart Association, 7272 Greenville Avenue, Dallas, TX 75231
Copyright © 2017 American Heart Association, Inc. All rights reserved.
Print ISSN: 0009-7330. Online ISSN: 1524-4571

Supplemental Material

The identification of 64 novel genetic loci provides an expanded view on the genetic architecture of coronary artery disease

Pim van der Harst[1,2,3], Niek Verweij[1]

(1.) University of Groningen, University Medical Center Groningen, Department of Cardiology, Groningen, The Netherlands. (2.) University of Groningen, University Medical Center Groningen, Department of Genetics, Groningen, The Netherlands. (3.) Durrer Center for Cardiogenetic Research, Netherlands Heart Institute, Utrecht, The Netherlands.

Online Data Supplement Contents

Online Material & Methods

UK Biobank individuals

Participants were recruited with an age range of 40-69 years of age that registered with a general practitioner of the UK National Health Service (NHS). Between 2006–2010, in total 503,325 individuals were included. All study participants provided informed consent and the study was approved by the North West Multi-centre Research Ethics Committee. Detailed methods used by UK Biobank have been described elsewhere.

Ascertainment of coronary artery disease (CAD)

Prevalence and incidence data on CAD was obtained at the Assessment Centre in-patient Health Episode Statistics (HES) and at any of the visits as described previously[1]. The prevalence and incidence of coronary artery disease conditions and events were captured by data collected at the Assessment Centre in-patient Health Episode Statistics (HES) download on September 10, 2016. CAD was defined using the following ICD 10 codes: I21-I25 covering ischemic heart diseases and the following Office of Population Censuses and Surveys Classification of Interventions and Procedures, version 4 (OPCS-4) codes: K40-K46, K49, K50 and K75 which includes replacement, transluminal balloon angioplasty, and other therapeutic transluminal operations on coronary artery and percutaneous transluminal balloon angioplasty and insertion of stent into coronary artery. Self reported CAD was also used in the definition (heart attack/myocardial infarction, coronary angioplasty +/- stent, cabg and triple heart bypass).

Non-CAD individuals defined the control population but to reduce biological misclassification and to improve power we excluded individuals from the control population if their mother, father or sibling was reported to suffer from 'heart disease'. This approach has been validated previously[1] and lead to the inclusion of 34,541 CAD cases and 261,984 non-CAD controls of the UK Biobank.

The exact phenotype definitions of UK Biobank used in the baseline table and phenome scan are described in detail elsewhere[1].

Genotyping and imputation

The Wellcome Trust Centre for Human Genetics performed quality control before imputation and imputed to HRC v1.1 panel. Analyses have been restricted to variants that are in the HRC v1.1. reference panel because UK10K imputation was unreliable at time of analyses. Quality control of samples and variants, and imputation was performed by the Wellcome Trust Centre for Human Genetics, as described in more detail elsewhere[2]. Sample outliers based on heterozygosity and missingness were excluded by the Welcome Trust Centre for Human Genetics, 373 additional participants were excluded based on gender discrepancies between the reported and inferred gender (using X-chromosome heterozygosity).

Genetic analyses

All genetic analyses in UK Biobank that are reported in this manuscript were adjusted for age, gender, the first 30 principal components (PCAs) to account for population stratification and genotyping array (Affymetrix UK Biobank Axiom® array or Affymetrix UK BiLEVE Axiom array). Genome wide association analysis in UK Biobank was performed using BOLT-LMM v2.3beta2, employing a mixed linear model that corrects for population structure and cryptic relatedness in a time efficient manner[3]. For the mixed model, we used directly genotyped variants that passed quality control, which were extracted from the imputed dataset, to ensure 100% call rate, and pruned on linkage disequilibrium (first $r^2 < 0.05$ and a second round of $r^2<0.045$) to obtain roughly 400k variants across the genome, as recommended by BOLT. The GWAS in UK Biobank was performed on a confined set of 7,947,838 SNPs that were common in UK Biobank and available in the CARDIoGRAMplusC4D 1000Genomes[4], Metabochip[5,6] and/or exome-chip[7] studies (downloaded from http://www.cardiogramplusc4d.org/ downloads and http://www.phenoscanner.medschl.cam.ac.uk). The genomic control, lambda was 1.25 but the intercept (1.0287(se=0.0093)) of the LDscore regression suggested no inflation due to non-polygenic signals (**Online Figure VI**), the residual inflation of 2.8% is most likely due to the sample/reference LD Score mismatch between UKBiobank and 1000 Genomes[8].

The Beta estimates of the top-SNPs in UK Biobank were re-estimated using a logistic regression with sandwich robust standard errors that were clustered by family to account for relatedness among UK Biobank participants, which were used in the meta-analyses[9]. Families were inferred from the kinship matrix, clustering all 3rd degree relatives or higher together (kinship coëfficiënt > 0.0442).

The GWAS dataset of CARDIoGRAMplusC4D's 1000Genomes[4] was complemented with the 1000Genomes[4]+Metabochip study[5,6], which was used for the discovery of SNPs associated with CAD P<0.0001 and used to replicate independent signals found in the UK Biobank GWAS.

To obtain a set of independent SNPs, SNPs *P*<0.0001 were clumped together based on LD $r^2>0.1$ and 5,000kb distance, using plink's clumping procedure. A locus was defined as a 1MB region at either side of the highest associated SNP in a locus.

We used a reciprocal two-stage sequential discovery and replication design with independent SNPs that were suggestive for their association with CAD (P<0.0001). First, we determined independent SNPs (P<0.0001) in UK Biobank and used the exome-chip[7] or 1000Genomes[4]+Metabochip study[5,6] for replication, second we determined independent SNPs in the exome-chip[7] or 1000Genomes[4]+Metabochip study[6] and used UK Biobank as replication. To account for multiple testing in the replication phase we applied *P*<0.05 after Bonferroni adjustment for the number of tests (649 tests in stage 1, 568 tests in stage 2), considering the direction of effect in the discovery phase (1-sided). To minimize false positive findings, a SNP was only considered to be true if significant if both replicated and surpassing the genome-wide significance threshold (*P*<5x10-8) in the inverse-variance meta-analysis. Finally, we performed a meta-analysis of the GWAS of UK Biobank and the 1000Genomes[4]+ study[6], which was used to plot the regional associations of the novel loci and Manhattan. Over-dispersion of association statistics in the 1000Genomes[4]+Metabochip study[5,6] was adjusted using genomic control by the consortium; in UK Biobank we adjusted the over-dispersion by the LDscore's intercept 1.0287.

The additive model was compared to a dominant and recessive model for genome wide significant SNPs using SNPTEST v2. Since SNPTEST does not account for relatedness, the analysis was restricted to independent individuals (more details under "Details on regression analyses and accounting for relatedness"), explaining why some SNP-associations are less significant compared to the discovery analysis.

Heritability

SNP-heritability was estimated in the UK Biobank using BOLT-REML, results were transformed from the observed to the liability scale using linear transformation. The proportion of variance explained by the identified variants was calculated by taking the difference between the total-SNP heritability and the SNP-heritability after including the identified SNPs in the model as covariates. BOLT-REML estimated the pseudo (SNP) heritability at 0.104 (0.001), which is 27.8% on the liability scale, slightly higher than the previous estimate[1]; 15.1% of the SNP-heritability could be explained by the 161 SNPs, 4.1% by the 64 novel SNPs.

Assessment of potential bias due to overlapping samples

The exome-chip[7] or 1000Genomes[4]+Metabochip study[6] recruited individuals living in the UK (see table below) that might also have been invited to participate in the UK Biobank cohort.

We estimated the influence of potential duplicate samples on test statistics. The CARDIoGRAM cohorts are heterogeneous for their age-ranges, inclusion-dates, and often of mixed-ancestry reducing the risk of duplicates, as the age range of recruitment for the UK Biobank was 40-69yrs (in 2006-2010).

Table CARDIOGRAM+C4D cohorts that may contain UK individuals overlapping with UK Biobank.

Cohort	*PMID*	*% recruited in UK*	*% age range with UK Biobank (40-69yrs)*	*Total in analysis*	*Total potentially overlapping UK Biobank*
EPIC CAD [7]	10466767	100	50	8423	4212
GoDARTS CAD [7]	16710446	100	100	4340	4340
PROCARDIS[7]	16710446	61.61	100	4710	2902
EPIC- CVD [6]	17295097	5.85	71	18642	1091
GODARTS[4]	9329309	100	100	3064	3064
LOLIPOP[4]	18454146	100	100	6548	6548
WTCCC[4]	17554300	100	100	4864	4864
HPS[4]	12114036	100	24	5458	1305
CARDIOGENICS[4]	22144904	67.5	100	802	541
ITH_2[4]	16271645	0.02	100	850	0
PROSPER[4]	12097148	43.42	0	5244	0
PROCARDIS[4]	16710446	61.61	70	12264	5289
				Total meta-gwas	22702
				total exome	11454

The 503,325 participants of UK Biobank represents a 5.5% response rate[10], of the approximately 9.150.000 invited individuals. Considering around 23.800.000 UK inhabitants within the age-range (40-69 years) and 22,702 Cardiogram UK participants, the chance for random sample overlap selection is 0.021, which equals 480 individuals; or less than 0.1% of this study's sample size, at most.

The influence of 0.01% duplicated sample on test-statistics is negligible (**Online Figure VII**). To estimate the effect of overlapping samples, a random set of 1000 case/control samples and SNPs associated at $P=1x10^{-8}$ was simulated; then a random overlap of X% samples was introduced for 5000 times and the mean –log(P-value) and 95% confidence interval of the SNP-CAD associations was determined. This process was repeated for 0.1 to 100% sample overlap.

Separately, we estimated the influence of potential sample overlap between UK Biobank and 1000Genomes[4]+Metabochip study[6] using mtag ("Multi-Trait Analysis of GWAS" , https://github.com/omeed-maghzian/mtag). Mtag (with the options '--equal_h2 --perfect_gencov') performs a meta-analysis taking into potential sample overlap using estimates of LD Score regression and modeling it snp-by-snp. The LD score regression intercept of the genetic covariance was estimated at 0.022 (0.008), which could suggest that there may be some degree of sample overlap, but meta-analysis results by mtag are virtually unchanged (**Online Table VI**).

Genetic risk score

A weighted genetic risk score was constructed using effect estimates of the CARDIoGRAMplusC4D data as previously described[1] using the 97 previously identified variants and the 64 novel ones. For this, the number of CAD increasing risk alleles were summed after multiplying the alleles with the corresponding β (based on the CARDIoGRAMplusC4D 1000Genomes[4]+Metabochip study[5,6]) to avoid any potential reverse causation and standardized to a mean of zero and a standard deviation of 1.

eQTL and meQTL analyses

To search for evidence of functional effects of SNPs at CAD loci multiple eQTL databases were examined; GTEX version 6[11], Stockholm-Tartu Atherosclerosis Reverse Network Engineering Task (STARNET)[12], cis-eQTL datasets of Blood[13–15] and cis-meQTLs[16]. Only eQTLS/meQTLs that achieved $P<1x10^{-6}$ and were in LD ($r^2>0.8$) with the queried GWAS variant were considered significant. Several eQTLs were observed in different tissues/studies, adding to the evidence of being a true eQTL.

Identification of Candidate Genes

We prioritized candidate genes in each of the 64 loci based on the following criteria: (1) The nearest gene or any gene located within 10 kb of the sentinel genetic variant, (2) Any gene containing protein coding variants in linkage disequilibrium (r2 > 0.8, UK Biobank) with the sentinel genetic variant. (3) Expression QTL (eQTL) analyses in cis; we search for eQTLs (sentinel genetic variants or genetic variants in linkage disequilibrium, $r^2 > 0.8$, UK Biobank) see above. We only considered eQTLs for which the top-eQTL was in linkage disequilibrium (r 2 > 0.8, UK Biobank) with the sentinel genetic variant and for which the eQTL $P<1x10^{-6}$. (4) DEPICT-genes (see section below for more details). (5) Long range chromatin interaction genes, Hi-C data was queried using http://yunliweb.its.unc.edu/HUGIn for the variant location using the 'Association' tab. A gene was considered significant if the interaction between the variant's location and promoter of a gene was significant by -log(P)<20.

Fine-mapping for causal variants and regulatory elements

To obtain insights into cell type specific functional annotations and enrichment we applied the nonparametric GARFIELD[17] approach on the genome-wide summary statistics of the complete meta-complete using default settings. Candidate causal variant were identified by the probabilistic framework of Probabilistic Annotation INTegratOR (PAINTOR) incorporating LD information, the P-value distribution of significance across GWAS loci and genetic annotations[18]. As input we used the 64 newly identified leadSNPs and 97 previously identified variants (the highest associated variant was selected in CardiogramPlusC4D 1000 genomes analysis), all SNPs in LD of $r^2>0.1$ and a P-value <0.01 in the complete meta-analysis of UKBiobank and CardiogramPlusC4D was considered as a locus. Because GARFIELD and DEPICT indicated that many different types of tissues and cell types may be underlying the CAD loci, and the number of annotations in complete PAINTOR model is limited to 3-5, we used genetic annotations that were not cell-type specific as described previously in Finucane *et al*[19], and coding variants identified by dbNSFP[20]. PAINTOR determined the significance of each annotation (**Figure 3c**) after which we performed a forward selection process based on significance to select the most relevant genetic annotations for the loci. The model for prioritization included coding variants, conservation scores and H3K4me1 regions; any additional annotations did not increase the model-fit and/or were highly correlated with one of the annotations already in the model. Potential SNP-Gene mechanisms were highlighted by missense variants (highly enriched among causal variants), utr-3 variants together with evidence of an eQTL of the same gene and significant long-range interactions of the SNP location with a gene's promoter that is also a significant eQTL.

In silico functional annotations of regulatory elements of the leads SNPs were performed with annotations described in Finucane *et al*[19], but also cell-type specific chromatin state, protein binding annotation from the Roadmap Epigenomics[21] and ENCODE[22] projects, sequence conservation across mammals and the effect on regulatory motifs by the HaploReg tool (v4.1)[23].

Mouse Genome Informatics (MGI) analyses

We systematically searched the international database resource for the laboratory mouse (MGI-Mouse Genome Informatics) for all candidate genes and manually curated Mammalian Phenotypes (MP) identifiers related to the cardiovascular system and others potentially relevant to the pathogenesis of CAD.

Data-driven Expression-Prioritized Integration for Complex (DEPICT) analyses

DEPICT systematically identifies the most likely causal gene at a given associated locus, tests gene sets for enrichment in associated SNPs, and identifies tissues and cell types in which genes from associated loci are highly expressed (see Pers et al.[24] for a detailed description of the method). DEPICT.v1.beta version rel194 for 1KG imputed GWAS (8.2G) obtained from https://data.broadinstitute.org/mpg/depict/ was used to perform an integrated gene function analyses. DEPICT was ran with default settings using all SNPs that achieved $P<1x10^{-5}$ as input, as suggested[24]. For comparison with previous datasets DEPICT was applied to the CardiogramPlusC4D 1000 genomes analysis[4], and CardiogramPlusC4D + intermediate dataset of UK Biobank[1], using the same settings. Both nominal P-values and false discovery rates (FDRs) were calculated. Networks were visualized using Gephi software (www.gephi.org).

Ingenuity Pathway Analysis (IPA) analyses

Ingenuity Pathway Analysis® (IPA, Qiagen's Ingenuity Systems, Redwood City, CA, USA; www.ingenuity.com) June 2017 release was used and focused on the potentially relevance of the novel 47 candidate genes. The default parameters set were: 1. The Ingenuity Knowledge Base was set as the reference set; 2. Both direct and indirect relationships were considered; and 3. Only relationships that were experimentally observed were considered. 4. Networks were generated with a maximum size of 35 genes.

GWAS catalog analyses

The GWAS catalog database was downloaded from https://www.ebi.ac.uk/gwas/ and queried by searching for SNPs in a 1MB region of the novel SNPs found in this study. Next, LD was determined by calculating the r^2 and D' in UK Biobank between the GWAS catalog SNPs and the SNPs of this study.

Details on regression analyses, accounting for relatedness and phenome wide analysis

All regression analyses (Linear, logistic or multinomial logistic) were carried out using STATA-SE between phenotypes and SNPs or the genetic risk score, and were performed with sandwich robust standard errors that were clustered on family to account for relatedness (unless stated otherwise). Families were inferred from the kinship matrix and based on 3th degree relatives or higher (kinship coëfficiënt > 0.0442). All analyses in this manuscript were adjusted for age, gender, the first 30 principal components and genotyping array to account for population stratification. Multinomial logistic regression analyses in STATA (mlogit) were performed to assess the extent to which genome wide significant SNPs were or were not driven by myocardial infarction. A Wald test was performed to determine significance in beta–estimates between CAD(non-myocardial infarction) vs controls and CAD (myocardial infaction) vs controls, by STATA-SE's 'test' command.

Cox regression analyses adjusted for age, gender, 30 principal components and the genotyping array were used to evaluate the predictive power of the genetic risk score on new onset disease and mortality. To account for relatedness in the Cox regression analyses we pruned families (>3th degree, kinship coefficient <0.0442) using an iterative approach to keep a maximum independent set of participants, based on the genotype missingness rate; as a result, 74,477 related individuals were removed for the cox regression analyses. For disease phenotypes, individuals were excluded when they were reported to have the particular disease at baseline or in history to study new onset disease only.

We tested the association of the newly identified SNPs and the genetic risk score of CAD with a wide range of phenotypes using linear or logistic regression analysis in UK Biobank to create a heatmap of the Z-scores that are aligned with CAD increasing risk, for this only FDR<0.01 significant associations are depicted in colored squares (hierarchical clustered).

A phenome-scan for each SNP was carried out by intersecting the identified loci with the GWAS-catalog.

External resources and bioinformatics tools

Name	Description	URL
BOLT-LMM/REML	Tool for Genome wide association using mixed model approach	https://data.broadinstitute.org/alkesgroup/BOLT-LMM/
PLINK 1.9/2.0	used to handle genetic data	https://www.cog-genomics.org/plink2
QCTOOL	used to handle genetic data	http://www.well.ox.ac.uk/~gav/qctool/#overview
Bgenix	used to handle genetic data	https://bitbucket.org/gavinband/bgen/wiki/bgenix
DEPICT	Pathway analyses	https://data.broadinstitute.org/mpg/depict
HUGIN	Analysis of Hi-C data	http://yunliweb.its.unc.edu/HUGIn
GWAS catalog	Data based of gwas hits	https://www.ebi.ac.uk/gwas/
PAINTOR	Fine-mapping software	https://github.com/gkichaev/PAINTOR_V3.0
Annotations of figure 3	-	https://data.broadinstitute.org/alkesgroup/LDSCORE/baseline_bedfiles.tgz
GARFIELD	Tissue specific enrichment of functional DNA elements	https://www.ebi.ac.uk/birney-srv/GARFIELD/
Haploreg	Annotation of fine-mapped variants	http://archive.broadinstitute.org/mammals/haploreg
Ingenuity IPA	Pathway analysis	www.ingenuity.com
Cardiogram GWAS		http://www.cardiogramplusc4d.org; http://www.phenoscanner.medschl.cam.ac.uk
UK Biobank		http://www.ukbiobank.ac.uk
LD score	LD score regression	https://github.com/bulik/ldsc
mtag	Multi-Trait Analysis of GWAS	https://github.com/omeed-maghzian/mtag
ukpheno	R package used to generate phenotypes of UK Biobank	https://github.com/niekverw/ukpheno
Locuszoom	Generation of regional plots	http://locuszoom.sph.umich.edu
GTEx	eQTL database	http://GTExportal.org
NESDA NTR	eQTL database	https://eqtl.onderzoek.io
Blood eQTL browser	eQTL database	http://genenetwork.nl/bloodeqtlbrowser/
BIOS QTL browser	eQTL, trans-meQTL and eQTM databases	http://genenetwork.nl/biosqtlbrowser/

Online Figure Legends

Online Figure I. Flow scheme of analysis strategy.

Flow scheme of the analysis strategy. 2-stage reciprocal design was adopted, (A) stage 1 was UK Biobank as discovery cohort, and stage 2 data of the CARDIoGRAMplusC4D as replication identifying 13 novel loci, (B) stage 1 was CARDIoGRAMplusC4D consortium as discovery and stage 2 UK Biobank as replication, identifying 21 additional novel loci. (C) Upon genome wide meta-analysis, 30 additional new loci were identified (**Table 1**).

Online Figure II. Manhattan plot

Manhattan plot showing meta-analysis results for CAD under an additive model. *P* values are truncated at $-\log_{10}$ (P) = 40. Markers shown are from the meta-analysis of UK Biobank with CARDIoGRAMplusC4D (see also **Online Table III-VI**). The gray dashed line denotes the GWAS ($P < 5\times10^{-8}$) significance threshold. Known CAD risk markers are shown in grey. The 64 novel CAD-associated are represented by the red peaks (**Table 1**).

Online Figure III. Regional association plots

Regional plots of the 64-novel genome-wide associated loci with CAD. LD (r^2) was based on the UK Biobank cohort. *P*-values were based on the genome wide meta-analysis to provide an accurate overview of the *P*-value distribution among variants at each locus; this is also the reason that for some regional plots the highest associated SNP is not the highlighted variant that was taken for replication.

Online Figure IV. Clinical outcome for GRS Quintiles

Online Figure V. Example of Fine-Mapping result

The location of the fine-mapped causal variant rs974819 can be functionally linked with the *PDGFD* gene via long-range interactions in aorta in Hi-C data (a) and the association of rs974819 with gene expression of *PDGFD* in aorta (b). The black vertical line in the middle of the Hi-C plot indicates the location of rs974819 (the 'anchor' position); the black horizontal line indicates observed interaction counts, the red horizontal line shows the expected interaction counts and the blue horizontal line indicates bonferonni significant interactions (threshold shown in purple dotted line). The yellow area indicates that the significant interaction is overlapping a promoter region of a candidate gene. The long-range interaction was especially significant in mesenchymal stem cells, which may be explained by the role of very similar platelet-derived growth factors such as PDGFD-AA and BB in mesenchymal stem cells proliferation and angiogenesis (c). The variant overlapped with an enhancer mark specific for mesenchymal cells (d).

Online Figure VI. Quantile-Quantile (QQ) plot

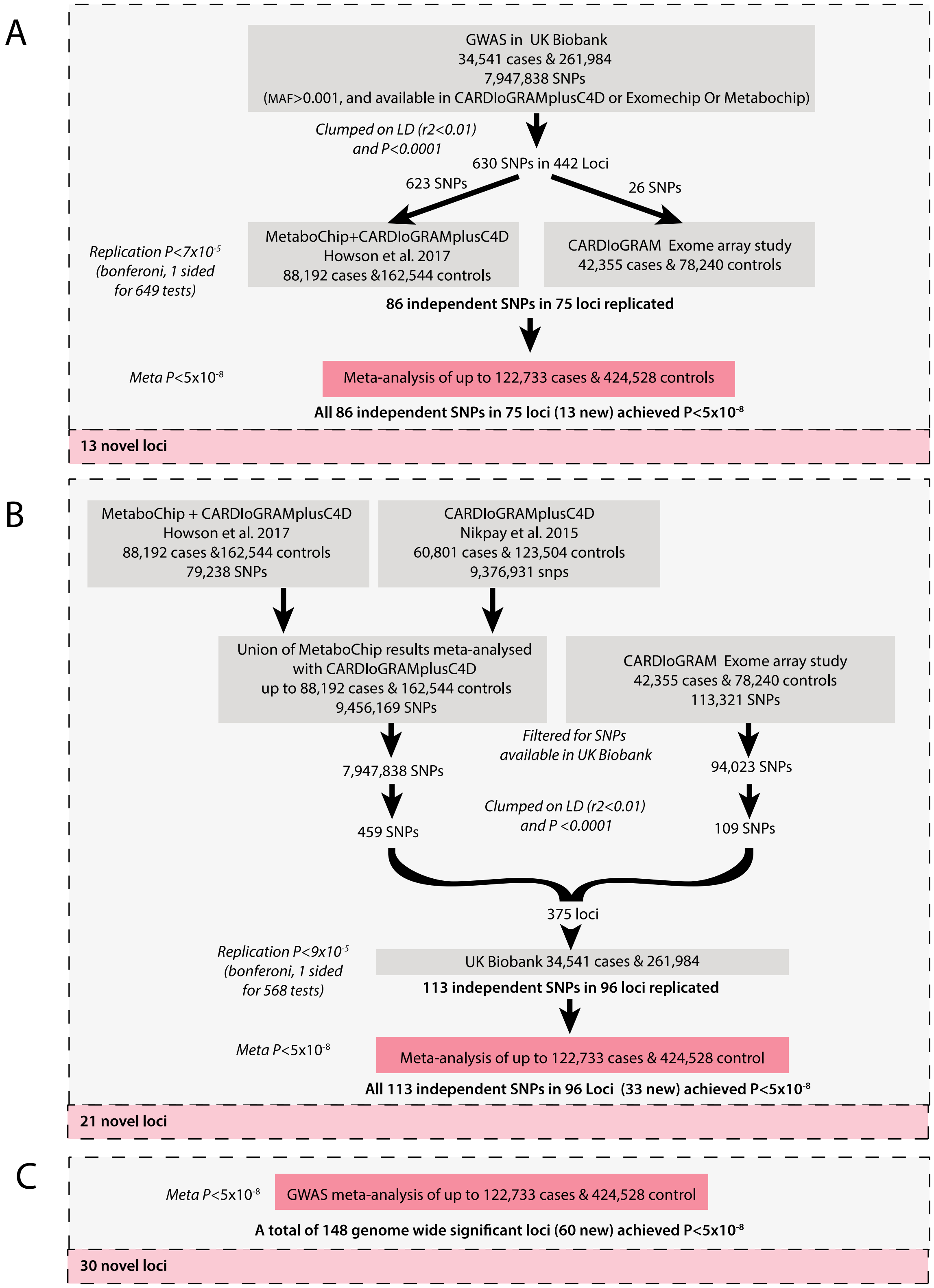
The Quantile-quantile (Q-Q) plot of the GWAS in UK Biobank. The intercept of LD score suggests no inflation due to non-polygenic causes.

Online Figure VII. Influence of potential sample overlap

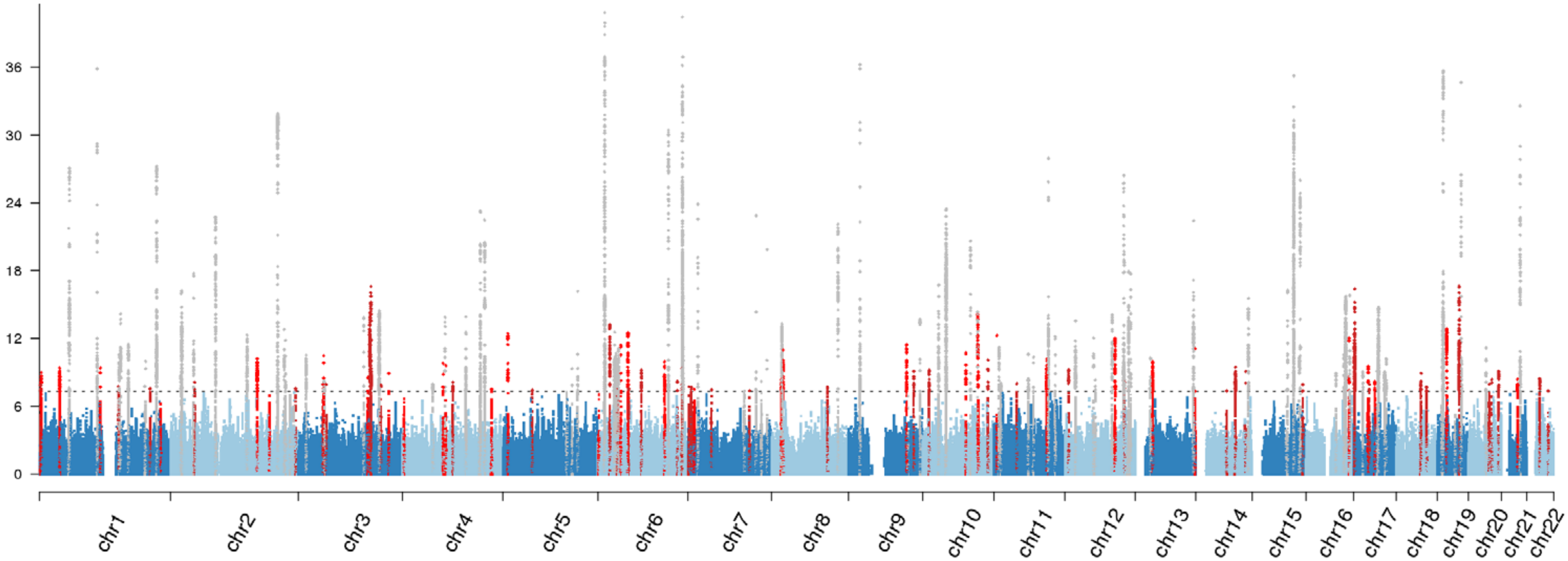
The influence of sample overlap was investigated by simulating random CAD observations and an associated SNP at a predefined P value (red horizontal line). Then the mean P value (y-axis) with the 95% confidence interval was determined of the P value distribution that was created by randomly introducing a sample overlap of X% for 5000 times. This process was repeated for each X between 0.1 and 100% sample overlap (x-axis, black dots). The predefined P value was set at P=1e-8. The SNP simulated had a minor allele frequency of 0.4, but results should be independent of this considering common alleles, like the ones identified in this manuscript.

Online Figures

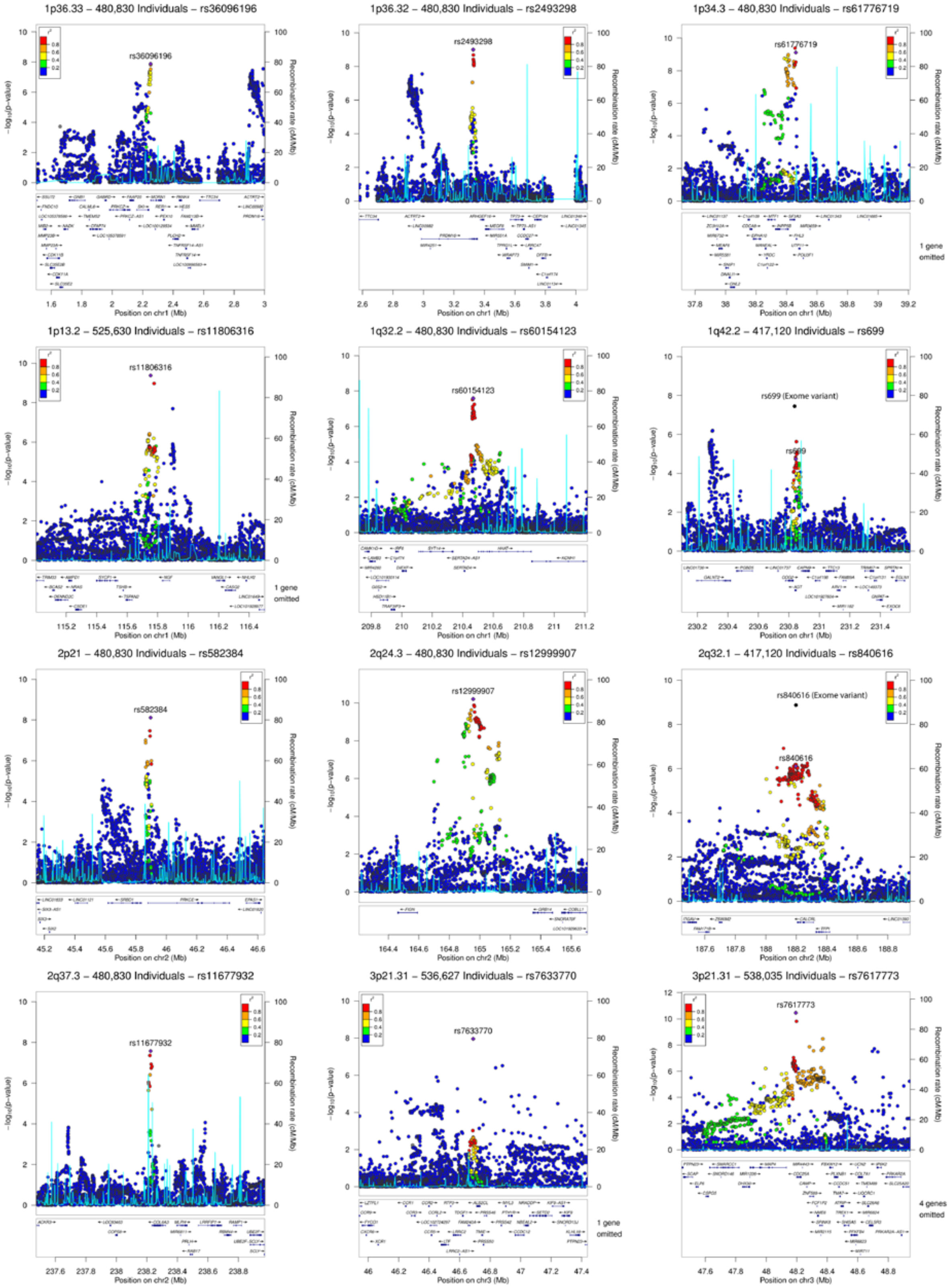
Online Figure I


A
GWAS in UK Biobank
34,541 cases & 261,984
7,947,838 SNPs
(MAF>0.001, and available in CARDIoGRAMplusC4D or Exomechip Or Metabochip)
Clumped on LD (r2<0.01)
and P<0.0001
630 SNPs in 442 Loci
623 SNPs
26 SNPs
MetaboChip+CARDIoGRAMplusC4D
Howson et al. 2017
88,192 cases &162,544 controls
CARDIoGRAM Exome array study
42,355 cases & 78,240 controls
Replication $P<7x10^{-5}$
(bonferoni, 1 sided
for 649 tests)
86 independent SNPs in 75 loci replicated
Meta $P<5x10^{-8}$
Meta-analysis of up to 122,733 cases & 424,528 controls
All 86 independent SNPs in 75 loci (13 new) achieved $P<5x10^{-8}$
13 novel loci
B
MetaboChip + CARDIoGRAMplusC4D
Howson et al. 2017
88,192 cases &162,544 controls
79,238 SNPs
CARDIoGRAMplusC4D
Nikpay et al. 2015
60,801 cases & 123,504 controls
9,376,931 snps
Union of MetaboChip results meta-analysed
with CARDIoGRAMplusC4D
up to 88,192 cases & 162,544 controls
9,456,169 SNPs
CARDIoGRAM Exome array study
42,355 cases & 78,240 controls
113,321 SNPs
Filtered for SNPs
available in UK Biobank
7,947,838 SNPs
94,023 SNPs
Clumped on LD (r2<0.01)
and P <0.0001
459 SNPs
109 SNPs
375 loci
Replication $P<9x10^{-5}$
(bonferoni, 1 sided
for 568 tests)
UK Biobank 34,541 cases & 261,984
113 independent SNPs in 96 loci replicated
Meta $P<5x10^{-8}$
Meta-analysis of up to 122,733 cases & 424,528 control
All 113 independent SNPs in 96 Loci (33 new) achieved $P<5x10^{-8}$
21 novel loci
C
Meta $P<5x10^{-8}$
GWAS meta-analysis of up to 122,733 cases & 424,528 control
A total of 148 genome wide significant loci (60 new) achieved $P<5x10^{-8}$
30 novel loci

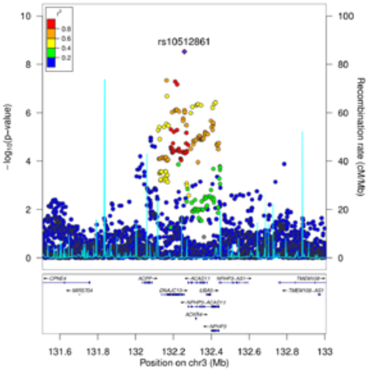
Online Figure II



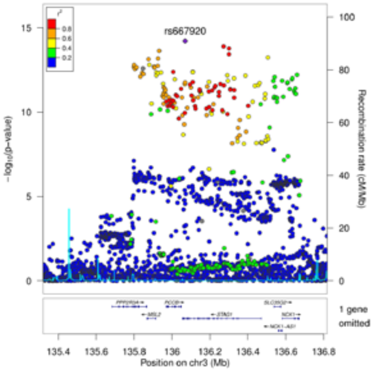
Online Figure III



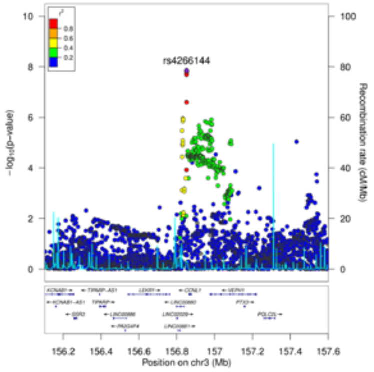
3q22.1 – 538,823 Individuals – rs10512861



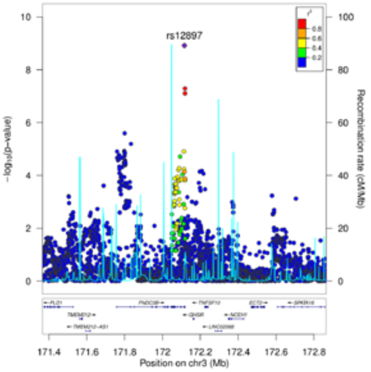
3q22.3 – 516,126 Individuals – rs667920



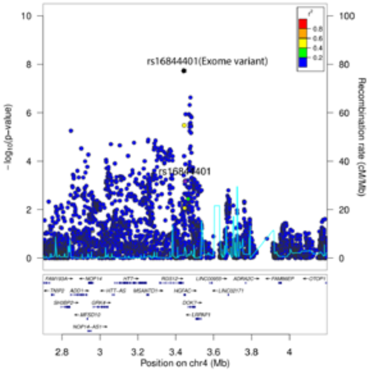
3q25.31 – 480,830 Individuals – rs4266144



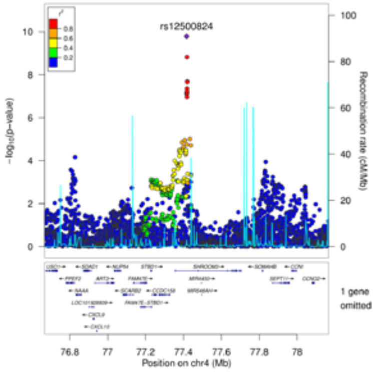
3q26.31 – 480,830 Individuals – rs12897



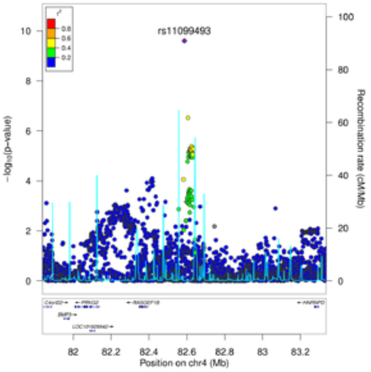
4p16.3 – 417,120 Individuals – rs16844401



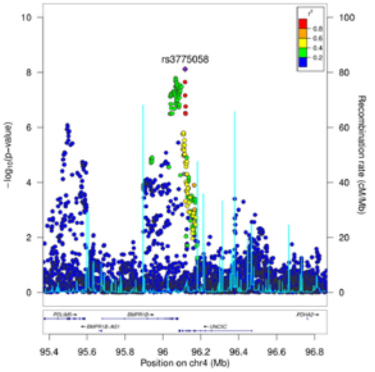
4q21.1 – 540,742 Individuals – rs12500824



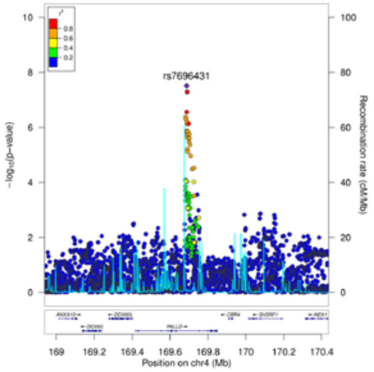
4q21.22 – 480,830 Individuals – rs11099493



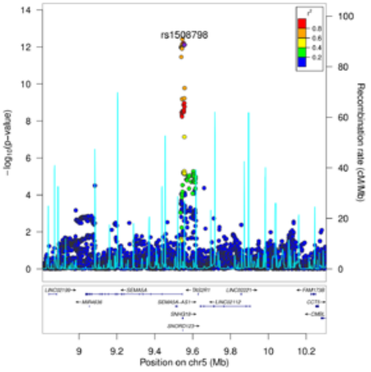
4q22.3 – 480,830 Individuals – rs3775058



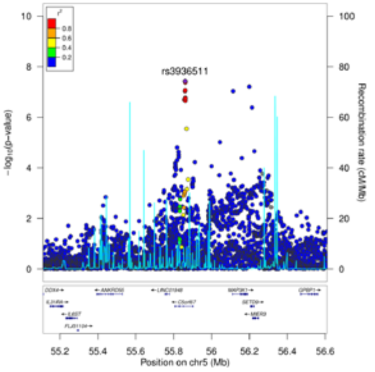
4q32.3 – 480,830 Individuals – rs7696431



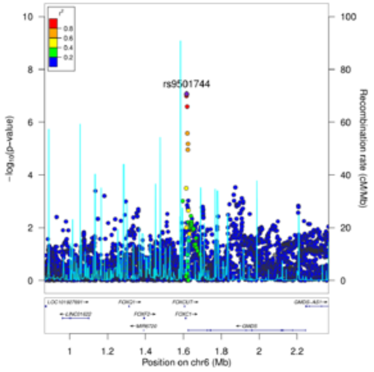
5p15.31 – 543,202 Individuals – rs1508798

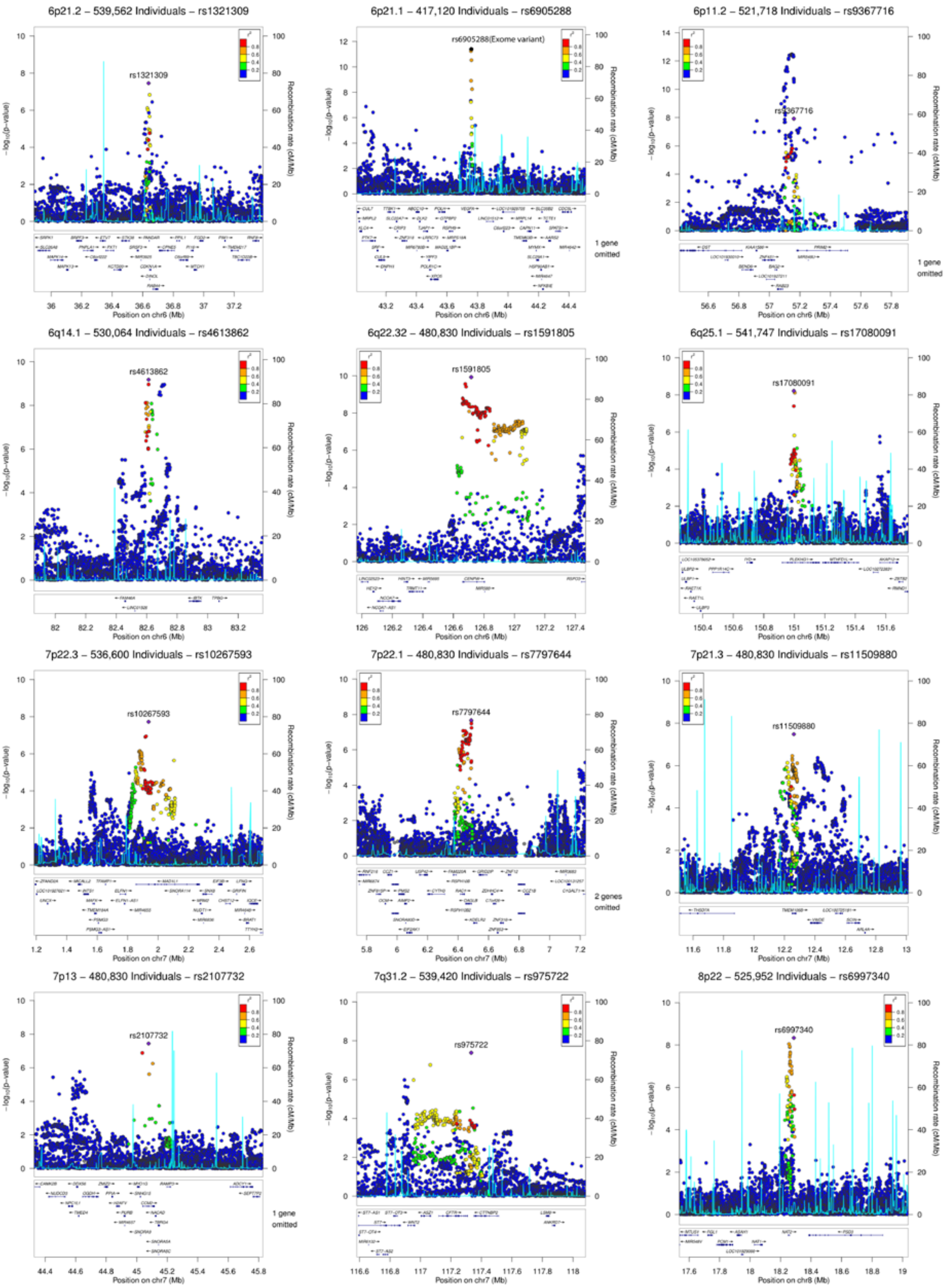


5q11.2 – 539,137 Individuals – rs3936511



6p25.3 – 480,830 Individuals – rs9501744





6p21.2 – 539,562 Individuals – rs1321309

6p21.1 – 417,120 Individuals – rs6905288

6p11.2 – 521,718 Individuals – rs9367716

6q14.1 – 530,064 Individuals – rs4613862

6q22.32 – 480,830 Individuals – rs1591805

6q25.1 – 541,747 Individuals – rs17080091

7p22.3 – 536,600 Individuals – rs10267593

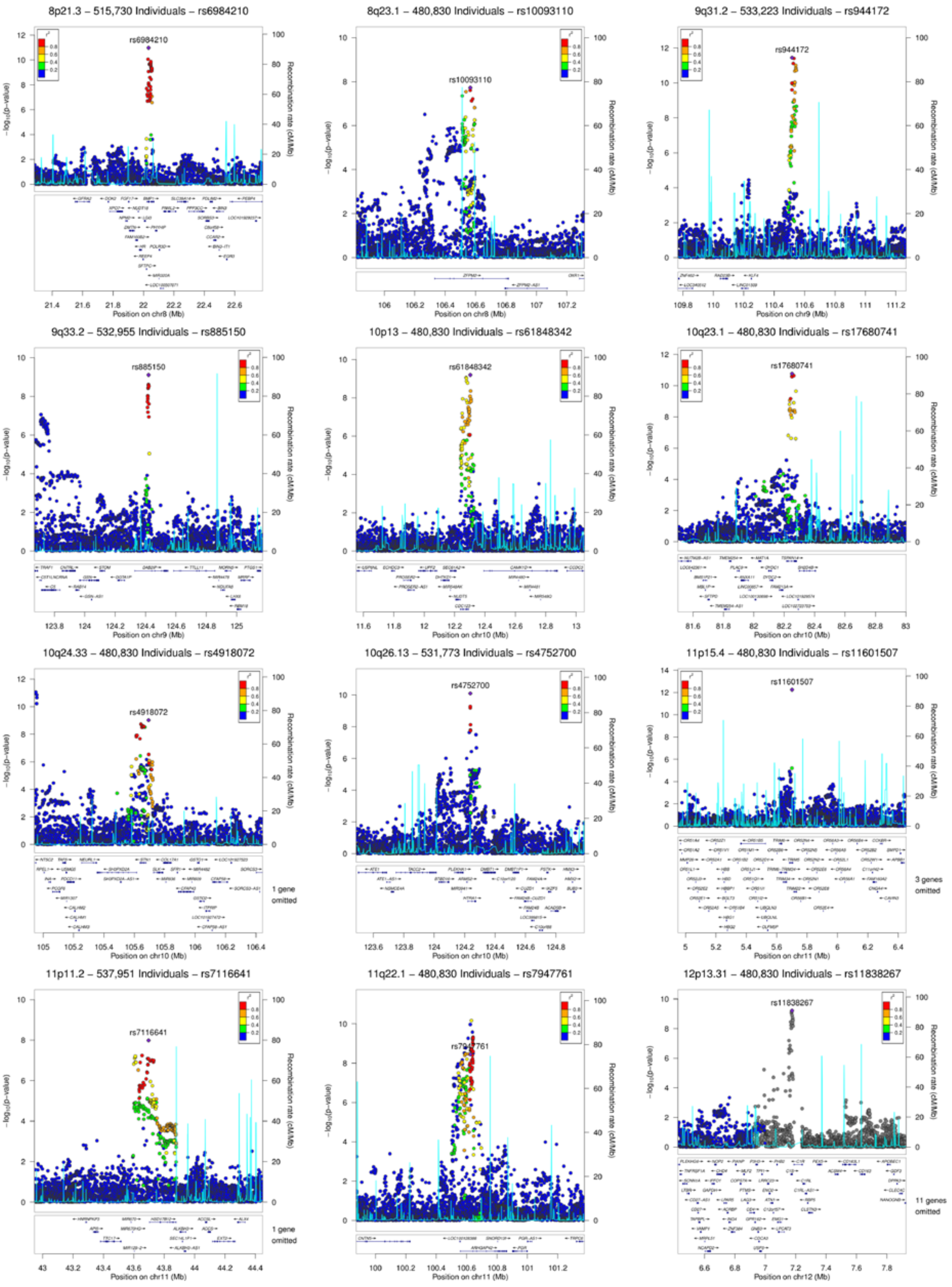
7p22.1 – 480,830 Individuals – rs7797644

7p21.3 – 480,830 Individuals – rs11509880

7p13 – 480,830 Individuals – rs2107732

7q31.2 – 539,420 Individuals – rs975722

8p22 – 525,952 Individuals – rs6997340



8p21.3 – 515,730 Individuals – rs6984210

8q23.1 – 480,830 Individuals – rs10093110

9q31.2 – 533,223 Individuals – rs944172

9q33.2 – 532,955 Individuals – rs885150

10p13 – 480,830 Individuals – rs61848342

10q23.1 – 480,830 Individuals – rs17680741

10q24.33 – 480,830 Individuals – rs4918072

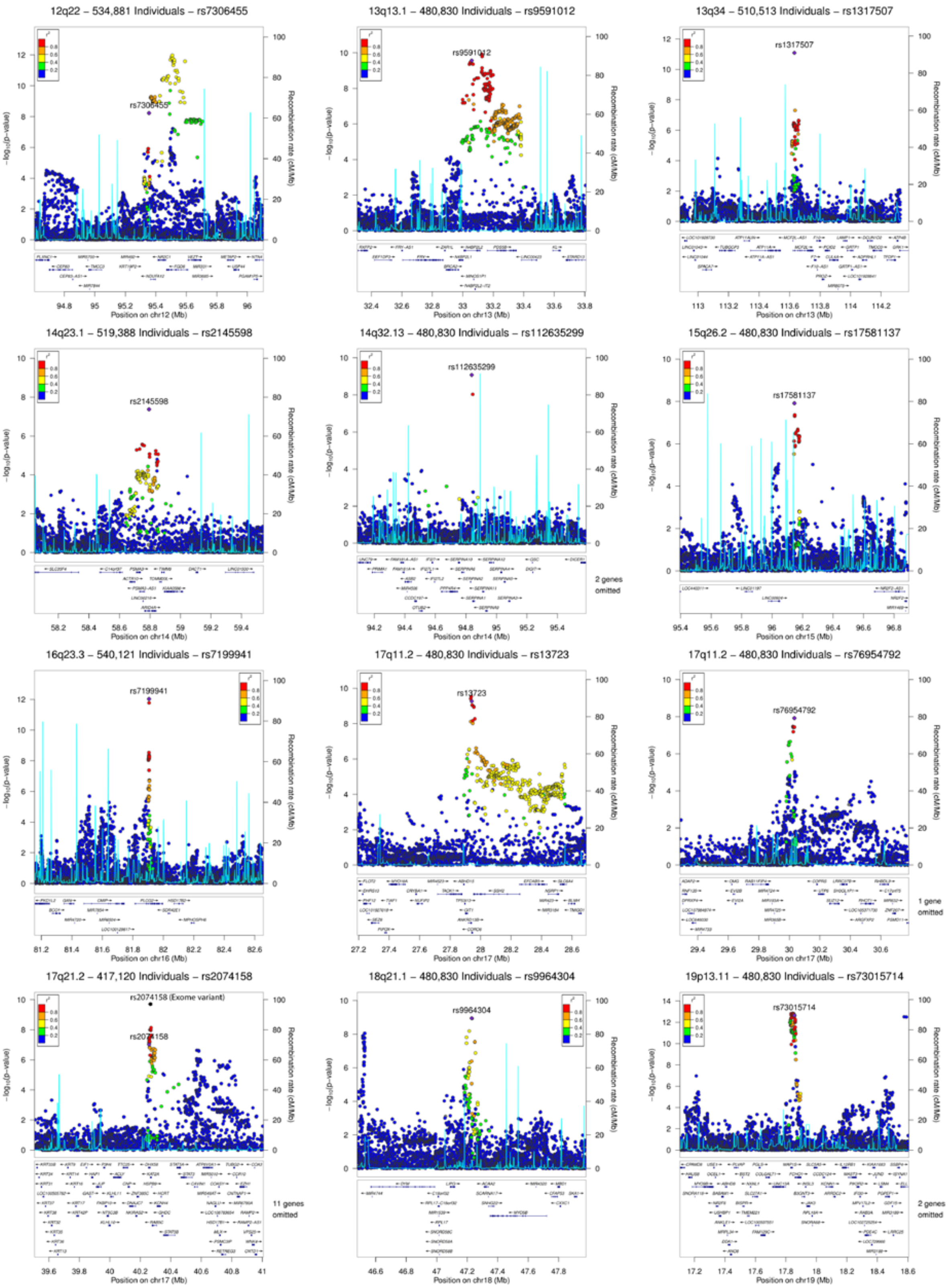
10q26.13 – 531,773 Individuals – rs4752700

11p15.4 – 480,830 Individuals – rs11601507

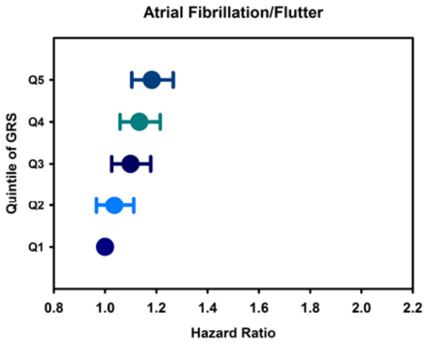
11p11.2 – 537,951 Individuals – rs7116641

11q22.1 – 480,830 Individuals – rs7947761

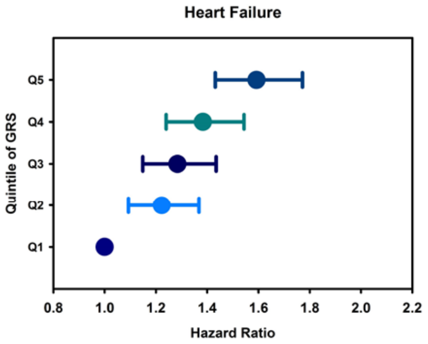
12p13.31 – 480,830 Individuals – rs11838267


12q22 – 534,881 Individuals – rs7306455
rs7306455
Position on chr12 (Mb)
13q13.1 – 480,830 Individuals – rs9591012
rs9591012
Position on chr13 (Mb)
13q34 – 510,513 Individuals – rs1317507
rs1317507
Position on chr13 (Mb)
14q23.1 – 519,388 Individuals – rs2145598
rs2145598
Position on chr14 (Mb)
14q32.13 – 480,830 Individuals – rs112635299
rs112635299
2 genes omitted
Position on chr14 (Mb)
15q26.2 – 480,830 Individuals – rs17581137
rs17581137
Position on chr15 (Mb)
16q23.3 – 540,121 Individuals – rs7199941
rs7199941
Position on chr16 (Mb)
17q11.2 – 480,830 Individuals – rs13723
rs13723
Position on chr17 (Mb)
17q11.2 – 480,830 Individuals – rs76954792
rs76954792
1 gene omitted
Position on chr17 (Mb)
17q21.2 – 417,120 Individuals – rs2074158
rs2074158 (Exome variant)
rs2074158
11 genes omitted
Position on chr17 (Mb)
18q21.1 – 480,830 Individuals – rs9964304
rs9964304
Position on chr18 (Mb)
19p13.11 – 480,830 Individuals – rs73015714
rs73015714
2 genes omitted
Position on chr19 (Mb)
−log₁₀(p−value)
Recombination rate (cM/Mb)
r²
0.8
0.6
0.4
0.2

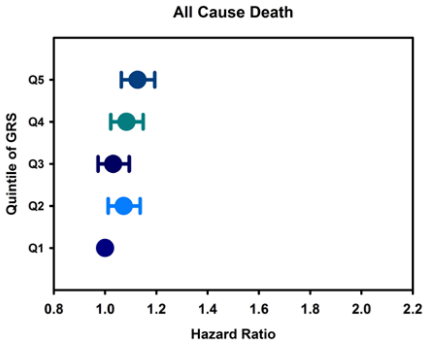
Online Figure IV



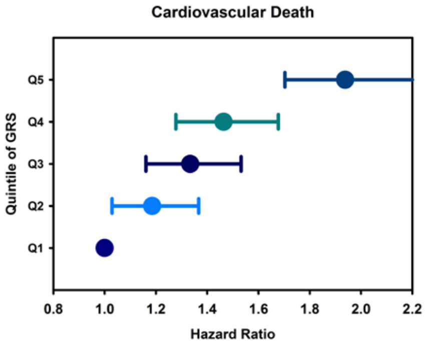
Q	HR	[95%Conf.	Interval]	P>z
2	1.04	0.97	1.11	0.32
3	1.10	1.03	1.18	0.01
4	1.13	1.06	1.21	0.00
5	1.18	1.10	1.27	0.00



Q	HR	[95%Conf.	Interval]	P>z
2	1.22	1.09	1.37	0.00
3	1.28	1.15	1.43	0.00
4	1.38	1.24	1.54	0.00
5	1.59	1.43	1.77	0.00

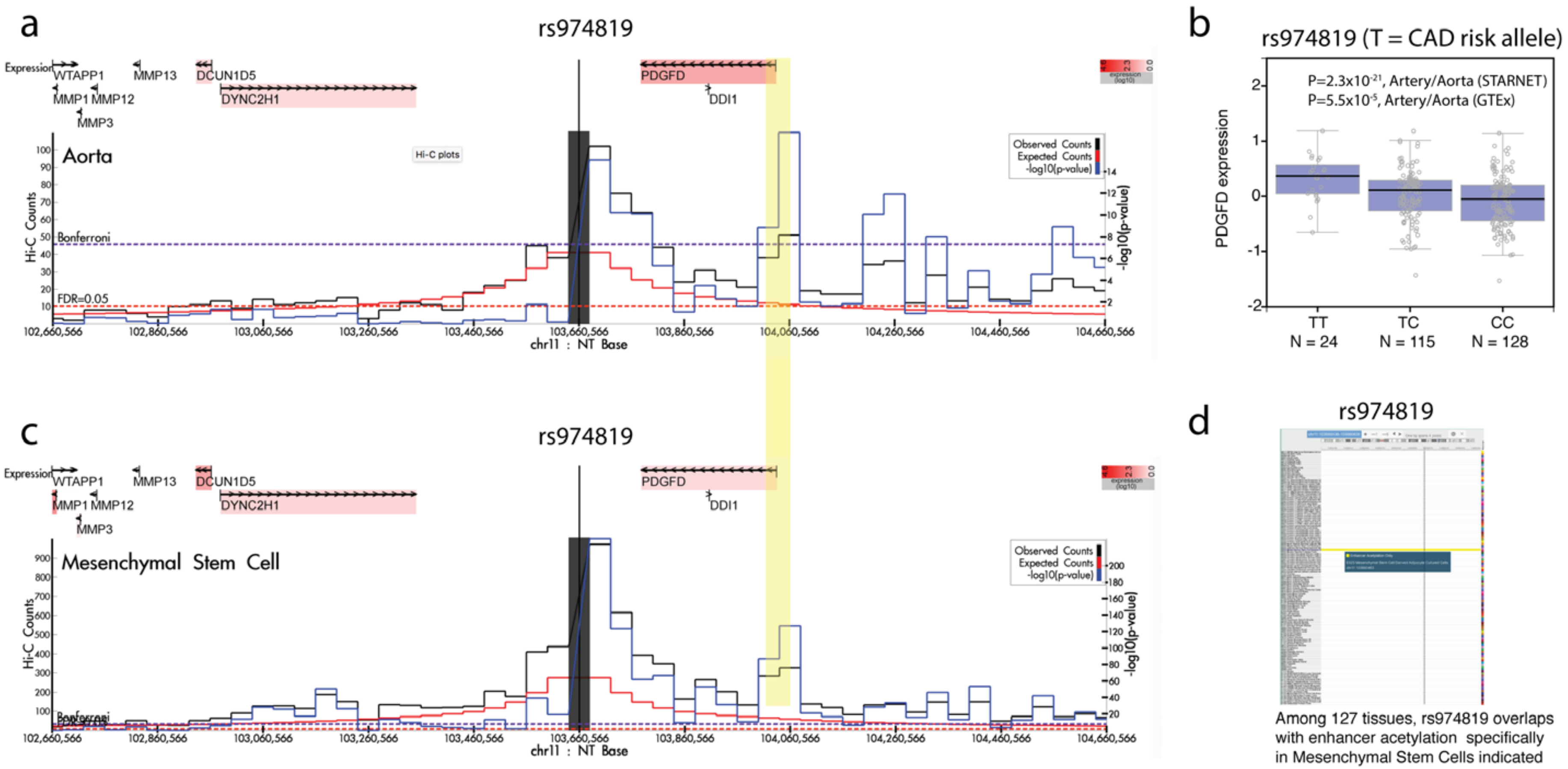


Q	HR	[95%Conf.	Interval]	P>z
2	1.07	1.01	1.14	0.02
3	1.03	0.97	1.09	0.30
4	1.08	1.02	1.15	0.01
5	1.13	1.06	1.19	0.00



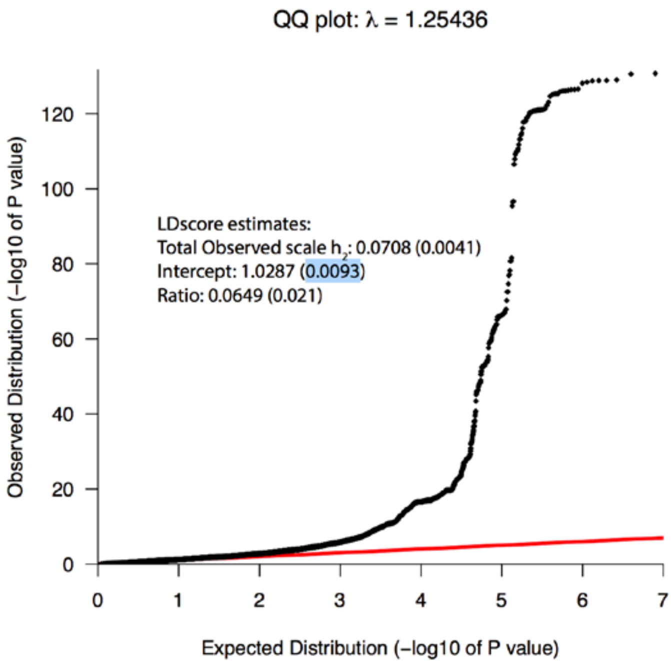
Q	HR	[95%Conf.	Interval]	P>z
2	1.19	1.03	1.37	0.02
3	1.33	1.16	1.53	0.00
4	1.46	1.28	1.68	0.00
5	1.94	1.70	2.21	0.00

Online Figure V

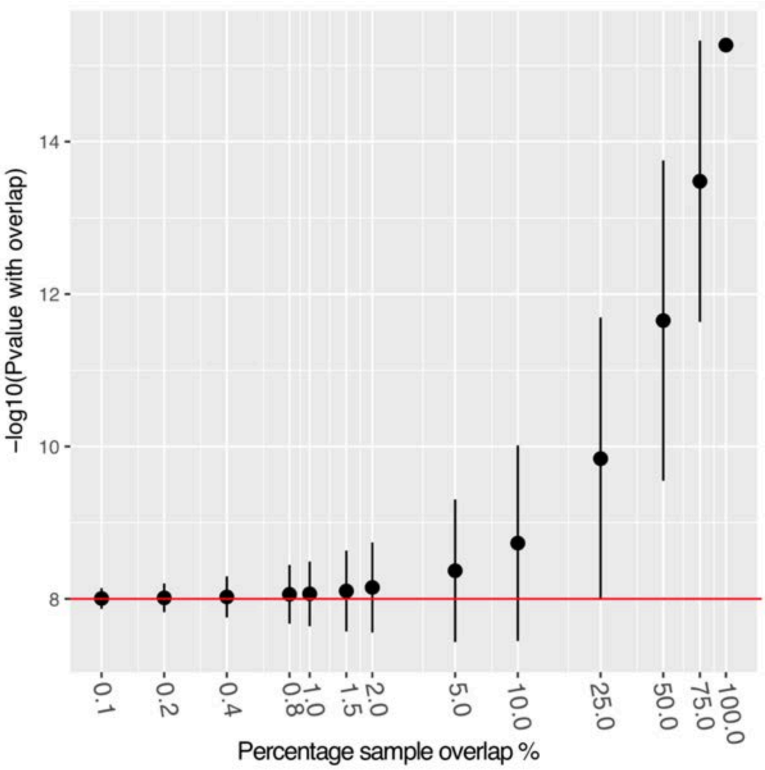


Among 127 tissues, rs974819 overlaps with enhancer acetylation specifically in Mesenchymal Stem Cells indicated by the yellow square

Online Figure VI



Online Figure VII



Online References

1 Verweij N, Eppinga RN, Hagemeijer Y, van der Harst P. Identification of 15 novel risk loci for coronary artery disease and genetic risk of recurrent events, atrial fibrillation and heart failure. *Sci Rep* 2017; **7**: 2761.

2 Bycroft C, Freeman C, Petkova D, *et al.* Genome-wide genetic data on ~500,000 UK Biobank participants. *bioRxiv* 2017. http://www.biorxiv.org/content/early/2017/07/20/166298 (accessed Aug 5, 2017).

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

4 Nikpay M, Goel A, Won H-H, *et al.* A comprehensive 1,000 Genomes-based genome-wide association meta-analysis of coronary artery disease. *Nat Genet* 2015; **47**: 1121–30.

5 CARDIoGRAMplusC4D Consortium P, Deloukas P, Kanoni S, *et al.* Large-scale association analysis identifies new risk loci for coronary artery disease. *Nat Genet* 2013; **45**: 25–33.

6 Howson JMM, Zhao W, Barnes DR, *et al.* Fifteen new risk loci for coronary artery disease highlight arterial-wall-specific mechanisms. *Nat Genet* 2017; **49**: 1113–9.

7 Investigators MIG and CardiEC. Coding Variation in ANGPTL4, LPL, and SVEP1 and the Risk of Coronary Disease. *N Engl J Med* 2016; **374**: 1134–44.

8 Bulik-Sullivan BK, Loh P-R, Finucane HK, *et al.* LD Score regression distinguishes confounding from polygenicity in genome-wide association studies. *Nat Genet* 2015; **47**: 291–5.

9 UK Biobank. Genotyping and Quality Control of UK Biobank, a Large-Scale, Extensively Phenotyped Prospective Resource: Information for Researchers. Interim Data Release. 2015; : 1–27.

10 Sudlow C, Gallacher J, Allen N, *et al.* UK Biobank: An Open Access Resource for Identifying the Causes of a Wide Range of Complex Diseases of Middle and Old Age. *PLOS Med* 2015; **12**: e1001779.

11 Lonsdale J, Thomas J, Salvatore M, *et al.* The Genotype-Tissue Expression (GTEx) project. *Nat Genet* 2013; **45**: 580–5.

12 Franzén O, Ermel R, Cohain A, *et al.* Cardiometabolic risk loci share downstream cis- and trans-gene regulation across tissues and diseases. *Science* 2016; **353**: 827–30.

13 Westra H-J, Peters MJ, Esko T, *et al.* Systematic identification of trans eQTLs as putative drivers of known disease associations. *Nat Genet* 2013; **45**: 1238–43.

14 Zhernakova D V, Deelen P, Vermaat M, *et al.* Identification of context-dependent expression quantitative trait loci in whole blood. *Nat Genet* 2016; **49**: 139–45.

15 Jansen R, Hottenga J-J, Nivard MG, *et al.* Conditional eQTL analysis reveals allelic heterogeneity of gene expression. *Hum Mol Genet* 2017; **26**: 1444–51.

16 Bonder MJ, Luijk R, Zhernakova D V, *et al.* Disease variants alter transcription factor levels and methylation of their binding sites. *Nat Genet* 2016; **49**: 131–8.

17 Iotchkova V, Huang J, Morris JA, *et al.* Discovery and refinement of genetic loci associated with cardiometabolic risk using dense imputation maps. *Nat Genet* 2016; **48**: 1303–12.

18 Kichaev G, Yang W-Y, Lindstrom S, *et al.* Integrating functional data to prioritize causal variants in statistical fine-mapping studies. *PLoS Genet* 2014; **10**: e1004722.

19 Finucane HK, Bulik-Sullivan B, Gusev A, *et al.* Partitioning heritability by functional annotation using genome-wide association summary statistics. *Nat Genet* 2015; **47**: 1228–35.

20 Liu X, Wu C, Li C, Boerwinkle E. dbNSFP v3.0: A One-Stop Database of Functional Predictions and Annotations for Human Nonsynonymous and Splice-Site SNVs. *Hum Mutat* 2016; **37**: 235–41.

21 Roadmap Epigenomics Consortium A, Kundaje A, Meuleman W, *et al.* Integrative analysis of 111 reference human epigenomes. *Nature* 2015; **518**: 317–30.

22 Dunham I, Kundaje A, Aldred SF, *et al.* An integrated encyclopedia of DNA elements in the human genome. *Nature* 2012; **489**: 57–74.

23 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* 2012; **40**: D930-4.

24 Pers TH, Karjalainen JM, Chan Y, *et al.* Biological interpretation of genome-wide association studies using predicted gene functions. *Nat Commun* 2015; **6**: 5890.