No large-effect low-frequency coding variation found for myocardial infarction

Oddgeir L. Holmen1,2,†, He Zhang3,‡, Wei Zhou3,‡, Ellen Schmidt3,5, Daniel H. Hovelson3,5, Arnulf Langhammer1, Maja-Lisa Løchen6, Santhi K. Ganesh3, Ellis B. Mathiesen7,8, Lars Vatten9, Carl Platou1, Tom Wilsgaard6, Jin Chen3, Frank Skorpen10, Håvard Dalen11,12, Michael Boehnke13, Goncalo R. Abecasis13, Inger Njølstad6, Kristian Hveem1,12 and Cristen J. Willer3,4,5,∗

1HUNT Research Centre, Department of Public Health and General Practice, Norwegian University of Science and Technology, 7600 Levanger, Norway, 2St. Olav Hospital, Trondheim University Hospital, Trondheim, Norway, 3Department of Internal Medicine, Division of Cardiology, 4Department of Human Genetics, University of Michigan Medical School, Ann Arbor, MI 48109, USA, 5Department of Computational Medicine and Bioinformatics, University of Michigan, Ann Arbor, MI 48109, USA, 6Epidemiology of Chronic Diseases Research Group, Department of Community Medicine, Faculty of Health Sciences, 7Brain and Circulation Research Group, Department of Clinical Medicine, Faculty of Health Sciences, UiT The Arctic University of Norway, 9037 Tromsø, Norway, 8Department of Neurology and Neurophysiology, University Hospital of North Norway, 9037 Tromso, Norway, 9Department of Public Health, 10Department of Laboratory Medicine, Children’s and Women’s Health Faculty of Medicine, 11Medical Imaging Laboratory for Innovative Future Healthcare, Lab and Department of Circulation and Medical Imaging, Norwegian University of Science and Technology, 7030 Trondheim, Norway, 12Department of Medicine, Levanger Hospital, Nord-Trøndelag Health Trust, 7600 Levanger, Norway and 13Center for Statistical Genetics, Department of Biostatistics, University of Michigan School of Public Health, Ann Arbor, MI 48109, USA

Received December 10, 2013; Revised April 7, 2014; Accepted April 9, 2014

Genome-wide association studies have identified variants, primarily common, that are associated with coronary artery disease or myocardial infarction (MI), but have not tested the majority of the low frequency and rare variation in the genome. We explored the hypothesis that previously untested low frequency (1–5% minor allele frequency) and rare (<1% minor allele frequency) coding variants are associated with MI. We genotyped 2906 MI cases and 6738 non-MI controls from Norway using the Illumina HumanExome Beadchip, allowing for direct genotyping of 85 972 polymorphic coding variants as well as 48 known GWAS SNPs. We followed-up 34 coding variants in an additional 2350 MI cases and 2318 controls from Norway. We evaluated exome array coverage in a subset of these samples using whole exome sequencing (N = 151). The exome array provided successful genotyping for an estimated 72.5% of Norwegian loss-of-function or missense variants with frequency >1% and 66.2% of variants <1% frequency observed more than once. Despite 80% power in the two-stage study (N = 14 312) to detect association with low-frequency variants with high effect sizes [odds ratio (OR) >1.86 and >1.36 for 1 and 5% frequency, respectively], we did not identify any novel genes or single variants that reached significance. This suggests that low-frequency coding variants with large effect sizes (OR >2) may not exist for MI. Larger sample sizes may identify coding variants with more moderate effects.

†The authors wish it to be known that, in their opinion, the first three authors should be regarded as joint First Authors.

∗To whom correspondence should be addressed at: Division of Cardiovascular Medicine and Department of Human Genetics, University of Michigan Medical School, Ann Arbor, MI, USA. Tel: +1 734 6476018; Fax: +1 734 7644142; Email: cristen@umich.edu

© The Author 2014. Published by Oxford University Press. All rights reserved.
For Permissions, please email: journals.permissions@oup.com
INTRODUCTION

Recent genome-wide association studies (GWAS) have discovered 47 loci associated with coronary artery disease (CAD), most of which are also associated with myocardial infarction (MI) (1–4). While the CAD-associated variants typically have minor allele frequency (MAF) > 5%, they have small individual effect sizes and collectively explain <10% of the heritability in MI (1) which is estimated to be 40–60% (5,6).

Since these common variants explain only a fraction of the heritability, other factors must contribute. Multiple possible scenarios have been proposed for this ‘missing heritability’ such as: common variants with small effects (7); additional variants at known loci (8); overestimates in the statistical estimation of heritability (9); and low-frequency (1–5% MAF) or rare (<1% MAF) variants with larger effect sizes (10,11).

The majority of low-frequency (1–5%) and rare (<1%) variants have not been included in modern GWAS arrays and theoretically cannot be imputed from modern reference panels based on HapMap or 1000 Genome Project samples. While exome sequencing studies have successfully uncovered many low-frequency and rare coding variants (12), few studies have systematically evaluated their associations with complex traits (13), often due to lack of power stemming from small sample sizes. Several groups have now shown that a more cost-effective genotyping array-based approach can be a valuable tool for interrogating the impact of coding variation, much of which is low-frequency, on complex traits (14–17).

To systematically evaluate the potential contribution of low-frequency and rare coding variants to the risk of MI, we carried out exome array-based genotyping on a MI case and control sample taken from a homogenous population-based cohort in Norway. Specifically, we addressed the question of whether untested low-frequency and rare coding variants exist with moderate to high effect sizes as proposed (10,11). Exome array-based genotyping was carried out on 205 318 predicted coding variants in 9644 samples, including 2906 MI cases, from the Nord-Trøndelag Health Study (HUNT). We performed follow-up genotyping of the most promising signals in an additional 4668 Norwegian MI cases and controls from a separate Norwegian population-based cohort: the Tromsø study. Our combined sample of 14 312 individuals was well-powered to detect low-frequency variants with moderate to high effect on risk of MI (odds ratio (OR) > 1.86 for variants with MAF > 1%).

RESULTS

Our first goal was to evaluate the coverage of the exome array. To do this, we estimated the number of coding variants present in 151 Norwegian individuals using exome sequencing (75 MI cases and 76 controls, 45x average depth). We identified 46 864 coding variants genome-wide, of which 33 387 had <5% minor allele frequency (Table 1). Using the Illumina HumanExome Beadchip array, designed to capture a large fraction of coding variants observed in European individuals, we genotyped the 151 exome sequenced samples and examined array coverage of variants detected by sequencing. The array captured a large fraction of loss-of-function (LoF) and missense variation observed in the sequenced samples (72.5% and 66.2% for variants with MAF > 5 and >1%, respectively). Of 8706 rare LoF or missense variants (<1% frequency) identified in sequenced individuals and with two or more copies, suggesting they are unlikely to be sequencing artifacts, we directly genotyped 5767 (66.2%) with the array (Table 1).

Having confirmed that a large fraction of overall and low-frequency coding variation was captured by the exome array, we next tested the power of our sample to detect association with low-frequency variants with moderate to large effect sizes. We declared an alpha level for significance of 7 × 10⁻⁷.

Table 1. Coverage of coding variation on exome array

<table>
<thead>
<tr>
<th>Variant type</th>
<th>Polyorphic variants on exome array</th>
<th>Comparison to exome sequenced samples</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Frequency in N = 9644</td>
<td>Number of variants genotyped by exome array (N = 9644)</td>
</tr>
<tr>
<td>LoF (splice, nonsense, read-through)</td>
<td>&gt;5%</td>
<td>174</td>
</tr>
<tr>
<td>1–5%</td>
<td>115</td>
<td>1–5%</td>
</tr>
<tr>
<td>6 Copies—1%</td>
<td>1012</td>
<td>2 Copies—1%</td>
</tr>
<tr>
<td>1–5 Copies</td>
<td>992</td>
<td>1 Copy</td>
</tr>
<tr>
<td>Missense</td>
<td>&gt;5%</td>
<td>11 174</td>
</tr>
<tr>
<td>1–5%</td>
<td>7841</td>
<td>1–5%</td>
</tr>
<tr>
<td>6 Copies—1%</td>
<td>35 792</td>
<td>2 Copies—1%</td>
</tr>
<tr>
<td>1–5 Copies</td>
<td>28 872</td>
<td>1 Copy</td>
</tr>
<tr>
<td>LoF</td>
<td>≥6 Copies</td>
<td>1301</td>
</tr>
<tr>
<td>Missense</td>
<td>≥6 Copies</td>
<td>54 807</td>
</tr>
<tr>
<td>LoF + missense</td>
<td>≥6 Copies</td>
<td>56 108</td>
</tr>
<tr>
<td>CAD GWAS (non-coding)</td>
<td>All frequencies</td>
<td>45 (48)</td>
</tr>
</tbody>
</table>

This table shows HumanExome Beadchip exome array coverage (QC + polymorphic variants) of variants identified by exome sequencing a subset of the samples. Sequencing-based variant calls are derived from moderate-depth exome sequencing (average depth of 45 × in coding regions) in a sample of 151 Norwegian individuals (HUNT study) samples. Variants were annotated relative to ANNOVAR (version 2012-05-25 with hg19 as reference genome) (18). The exome array contains 45 directly genotyped CAD GWAS markers (and three high-r² proxies) as published by Deloukas et al. (2013) (1), Kathiresan et al. (2008) (3) and Schunkert et al. (2011) (4). CAD, coronary artery disease. LoF, loss-of-function.
reflecting 73 653 single-variant tests with minor allele count (MAC) > 5. For gene-based burden tests we set the significance threshold to $1 \times 10^{-6}$ to account for 37 607 gene models tested. Modeling our two-stage design (19), the jointly analyzed sample set of 14 312 individuals provided 80% power to detect variants, or a gene burden, of 1% frequency ($1/100$) and an OR of at least 1.86, or with 5% frequency and OR $> 1.36$ (Supplementary Material, Fig. S1).

In Stage 1 (see Materials and Methods), we genotyped 85 972 polymorphic coding variants and 48 GWAS variants in 9644 individuals taken from the HUNT study (20) using the exome array: 2906 medical-record confirmed MI cases and 6738 controls. We first selected 2906 birth year and sex-matched healthy controls without cardiovascular disease, diabetes, hypertension and no reported first or second degree relatives with cardiovascular disease (see Materials and Methods for details) to maximize our ability to detect association with genetic variants that impact MI. In addition, we included 3833 non-MI controls from the same cohort available from a separate study of high-density lipoprotein (HDL) cholesterol extremes. Analysis performed either with or without the additional controls was found to be highly consistent, thus we report results from the combined control set. Clinical characteristics for the genotyped individuals are summarized in Supplementary Material, Table S1.

To test if our approach and MI phenotype definition were sufficient to detect association with MI, we examined evidence for association with previously identified GWAS variants present on the array (Fig. 1). Of 51 variants in 47 loci previously reported to be associated with CAD or MI by large-scale GWAS (1,3,4), 48 variants from 44 loci were tested in our sample (45 directly tested, 3 with high-$r^2$ proxies). We observed a direction of effect consistent with published reports for 45 of the 48 CAD variants which is significantly different from the null expectation of 50% ($P_{\text{binomial}} = 1 \times 10^{-10}$, Supplementary Material, Table S2). We also observed an excess of suggestive $P$-values (10 with $P$-value $< 0.01$) for known variants compared with expectation under the null hypothesis of no association ($P_{\text{binomial}} = 4 \times 10^{-11}$). We further confirmed association with MI using a combined weighted allele score for the 48 previously reported variants ($P = 1 \times 10^{-28}$) and demonstrated that the risk of MI is higher for individuals with higher genetic risk scores (Fig. 2).

Examining the 48 CAD GWAS variants individually, the strongest evidence for association with MI in the 9644 individuals was observed in the CDKN2B cluster at the 9p21 locus for an intergenic variant 3 kb downstream of CAD GWAS lead identified CAD SNPs. This figure shows an increased risk of MI ($OR = 1.36$, Supplementary Fig. S1). We further confirmed association with MI using a combined weighted allele score for the 48 previously reported variants ($P = 1 \times 10^{-28}$) and demonstrated that the risk of MI is higher for individuals with higher genetic risk scores (Fig. 2).

Figure 1. Quantile–quantile (QQ) plots of single-variant analysis results of myocardial infarction. This figure shows quantile–quantile plots for single-variant association with myocardial infarction status in Stage 1 samples ($N = 9644$) for: 48 GWAS variants as previously published (1,3,4) represented on the exome array (red), all coding variants within 500 kb of CAD GWAS lead SNPs (blue) (1,3,4), and all other coding variants tested (black). The expectation under the null hypothesis is a thin black line at $x = y$, and the 95% confidence interval (CI) is shown in gray.

Figure 2. CAD genetic risk score categories and risk for MI based on previously identified CAD SNPs. This figure shows an increased risk of MI (OR $> 1$) for individuals with higher weighted genetic risk scores based on 48 CAD GWAS variants. The black line shows ORs and 95% CIs for association with MI for individuals grouped into risk score quintiles compared with the first quintile (labels on right vertical axis). Compared with the first quintile, the second to fifth quintiles have ORs 1.04, 1.32, 1.56 and 1.78, respectively, and 95% CIs [0.90, 1.20], [1.15, 1.52], [1.36, 1.80] and [1.55, 2.04], respectively. The risk scores were calculated for individuals in the Stage 1 sample ($N = 9644$) for 45 previously identified CAD variants that were directly typed on the exome array and three proxy markers in high linkage disequilibrium (LD) with the previously identified CAD variants (1,3,4). By testing all individuals in a regression framework, the genetic risk score was strongly associated with MI ($P = 1 \times 10^{-28}$). Alleles were weighted by their published effect sizes. The bars represent the number of MI cases (red bars) and controls (blue bars) in each risk score quintile (labels on left vertical axis).
Table 2. Genome-wide significant single-variant association with myocardial infarction

<table>
<thead>
<tr>
<th>Gene</th>
<th>Position</th>
<th>SNP</th>
<th>Stage 1 (N = 9644)</th>
<th>Stage 2 (N = 4668)</th>
<th>Combined (N = 14312)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>P-value MAF (%) OR (95% CI)</td>
<td>P-value MAF (%) OR (95% CI)</td>
<td>P-value MAF (%) OR (95% CI)</td>
</tr>
<tr>
<td>LPA</td>
<td>6:160961137</td>
<td>rs3798220</td>
<td>Ile1891Met</td>
<td>1.9</td>
<td>1.67 (1.33–2.09)</td>
</tr>
<tr>
<td></td>
<td>9p21</td>
<td>rs10757278</td>
<td>Non-coding</td>
<td>4.77</td>
<td>3.34 (2.09–5.38)</td>
</tr>
</tbody>
</table>

Positions are given for hg19 (chromosome:base pair position). The ‘Variant’ column lists predicted changes in amino acid sequence relative to ANNOVAR (version 2012-05-25 with hg19 as reference genome) (18). OR are reported for the minor allele. MAF, minor allele frequency; CI, confidence interval; nt, not tested.

We next examined all assayed coding variation across the genome for association with MI using single-variant association tests with age, sex and 10 ancestry informative principal components as covariates. Focusing on low-frequency variation, we selected variants for follow-up with \( P < 1 \times 10^{-3} \) in the preliminary analysis using matched controls, MAF < 10%, and MAC > 5. Follow-up genotyping was performed in an additional 4668 Norwegian individuals from the Tromsø study (22): 2350 MI cases and 2318 non-MI controls (Stage 2; see Materials and Methods). Thus, the total sample size available for these 24 variants was 14312.

In the combined analysis, one low-frequency coding variant reached the threshold for study-wide significance; the variant encoding p.Ile1891Met in the Lipoprotein A (LPA) gene (rs3798220, MAF = 1.9%, OR = 1.67, \( P = 5 \times 10^{-8} \), Table 2). This variant, which was previously known as p.Ile4399Met, is part of the SLC22A3-LPAL2-LPA gene cluster located at 6q26-27 that has previously been associated with CAD (1,23,24) and apolipoprotein(a) levels (25). No other variants reached exome-wide significance (\( P < 7 \times 10^{-8} \)) in the combined sample; top results for single-variant tests of the combined analysis are given in Table 2 and Supplementary Material, Table S4. To test for potential dominant or recessive effects of coding variants, we also performed single-variant association tests using these models. No additional variants reached exome-wide significance (data not shown).

We searched previously annotated biological pathways to test for over-representation of variants associated with MI using gene set enrichment analysis as implemented in MAGENTA (26). We identified one enriched pathway (B-cell proliferation) across five loci using a relaxed single-variant threshold in our combined Stage 1 sample (\( P < 10^{-4} \), \( N = 9644 \); Supplementary Material, Table S5). This pathway does not include known CAD genes.

To increase power to detect MI association with low-frequency and rare variants, we combined evidence across these variants in each gene using burden tests. We used two approaches to either weight variants by the inverse of their frequency (SKAT-O) (27) or test a variety of allele frequency thresholds for exclusion of common variants (VT) (28). We examined (i) all non-synonymous variants (including splice), (ii) only those variants predicted to be LoF (premature stop, read-through or splice) and (iii) LoF and ‘probably damaging’ missense variants as defined by Polyphen-2 (29). The tests showed an excess of suggestive \( P \)-values (\( P \)-values < 0.01) when analyzing all non-synonymous variants (1.5-fold excess, \( P_{\text{binomial}} = 8 \times 10^{-8} \)) and LoF or probably damaging variants (\( P_{\text{binomial}} = 0.02 \); Supplementary Material, Fig. S2), although no genes reached our significance threshold (\( P < 1 \times 10^{-8} \)) in Stage 1 (Supplementary Material, Table S6). Preliminary analysis of MI cases and the matched control set identified seven genes with \( P < 5 \times 10^{-8} \) and at least six individuals carrying a burden of rare allele(s). Within these seven genes we successfully genotyped in the follow-up sample 11 variants that individually reached \( P < 0.10 \) in Stage 1 samples. After follow-up, no genes reached exome-wide significance (\( P < 1 \times 10^{-6} \)) in Stage 1 + 2 samples using gene-based tests (Supplementary Material, Tables S6B and S7).

We extended the burden tests to aggregate individuals with rare variants (MAF < 1%) in genes that belong to the same...
pathway (defined by Gene Ontology) (30). These pathway-based burden tests identified one significant pathway related to lipid metabolism, low-density lipoprotein-mediated lipid transport ($P < 5 \times 10^{-5}$), which included two known CAD genes ($LPA$ and $APOB$; Supplementary Material, Table S8).

To more carefully examine candidate MI/CAD and lipid related genes, we selected genes within 500 kb of known GWAS signals for CAD or lipids. We applied a reduced significance threshold to account for the multiple testing burden for any single-variant test ($N = 7635$ variants, $P < 6.5 \times 10^{-5}$) or gene-based burden test ($N = 1635$ gene-based tests, $P < 3 \times 10^{-3}$) in these regions. No additional variants or genes reached these reduced thresholds for significance (Supplementary Material, Fig. S4 and Table S9).

Having been unable to identify any novel variants, genes or pathways associated with MI, we next examined the proportion of genetic heritability that could be explained by rare coding variants (MAF < 1%), low-frequency coding variants (MAF 1–5%) and common coding variants (MAF > 5%). Using a linear mixed model as implemented in the GCTA software (31), we found that the 56 108 coding variants genotyped together explained 15.9% of the trait variance. Intriguingly, rare variants (6 copies—1% MAF) explained a slightly higher proportion of the variance (10.9%) compared with common variation (MAF > 5%, 7.4% of variance explained; Table 3).

<table>
<thead>
<tr>
<th>Frequency of variant</th>
<th>Proportion of variance explained using Stage 1 samples ($N = 9644$)</th>
<th>Proportion of variance explained using matched controls only ($N = 5785$)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>All variants (%)</td>
<td>Coding variants (%)</td>
</tr>
<tr>
<td>&gt;5%</td>
<td>11.6</td>
<td>7.4</td>
</tr>
<tr>
<td>1–5%</td>
<td>2.2</td>
<td>1.7</td>
</tr>
<tr>
<td>6 Copies — 1%</td>
<td>11.1</td>
<td>10.9</td>
</tr>
<tr>
<td>All (&gt;6 Copies)</td>
<td>18.3</td>
<td>15.9</td>
</tr>
</tbody>
</table>

This table shows proportion of variance explained by genetic variants in different frequency strata as estimated by GCTA software (31) for our Stage 1 samples as described in Supplementary Material, Table S1. The increase in variance explained by ‘All variants’ in comparison to coding variants for the frequency >5% stratum is due to the inclusion of known GWAS variants.

Lastly, to investigate whether putatively associated variants were present at reduced significant thresholds, we examined the observed and expected number of significantly associated coding variants at several thresholds ($P < 0.01$, $1 \times 10^{-3}$ and $1 \times 10^{-4}$). As expected for a polygenic phenotype where power was limited to find the underlying genetic effects, we observed an excess of coding variants for several of the frequency strata (Supplementary Material, Table S10) suggesting additional association results will be identified in larger studies.

**DISCUSSION**

To study the contribution of coding variants to susceptibility of a common complex disease, we evaluated 85 972 polymorphic coding variants across the allele frequency spectrum for association with MI, with a particular focus on rare and low-frequency variation. We identified an association with a variant encoding p.Ile1891Met in $LPA$ and with common non-coding variants at 9p21 near the $CDKN2B$ gene cluster, suggesting that our approach and phenotype definition were sufficient to detect association. Although we had 80% power to detect association with low-frequency variants (MAF 1–5%) with moderate or high effect sizes (OR > 1.86), we did not identify any novel genes or single variants that reached statistical significance in up to 14 312 individuals.

Our inability to find novel rare or low-frequency variants associated with MI, despite being reasonably well-powered, suggest that low-frequency variants with very large effect (OR > 2) may not account for the missing heritability of MI. However, identifying: (i) an excess of results at thresholds below study-wide significance levels, and (ii) evidence that rare coding variants accounting for a higher proportion of the trait variance than common coding variants, suggests that rare disease-associated coding variants may be found in larger samples. We also cannot rule out the possibility that low-frequency variants with large effects exist in other ethnic groups or severely affected families.

Very rare LoF variants (<0.1% frequency), potentially with large impacts on disease risk, were not well captured by this exome array-based approach. In addition, we estimated that 20–30% of the rare and low-frequency coding variation in Norwegians was not directly evaluated by this approach. To fully evaluate the impact of rare variation on MI susceptibility, we will likely need to perform sequencing on reasonably large numbers of samples, possibly tens of thousands of cases and similar numbers of controls, and/or extend this analysis of coding variants into larger sample sizes.

**MATERIALS AND METHODS**

**Stage 1 study population**

We selected 2969 MI cases and 6738 non-MI controls from the HUNT study. The HUNT study is a comprehensive population-based health study (www.ntnu.edu/hunt) with personal and family medical histories on 106 436 people from Nord-Trøndelag County, Norway, collected in three surveys (HUNT 1, 2 and 3) (20). Inviting all residents ≥20 years of age, self-reported questionnaires, clinical examinations and non-fasting venous blood samples were available on 65 237 individuals from HUNT 2 (1995–1997; 71% of invited) and 48 290 individuals from HUNT 3 (2006–2008; 52% of invited). Information on first and second degree relationships was made available from the National Population Register and generated based on 106 436 individuals registered in one of the three surveys.

Cases were retrospectively identified as HUNT 2 and HUNT 3 participants diagnosed with acute MI (ICD-10 I21 or ICD-9 410) in the medical departments at the two local hospitals in Nord-Trøndelag County from December 1987 to June 2011. Linkage between the electronic diagnosis registry and HUNT was made possible through an 11-digit national identity number unique to every citizen of Norway. Age at earliest MI incident was registered for subjects with more than one MI admission. In the case of affected relative pairs, we selected one per family by including the individual with the earliest age of MI.

Initially, 2968 non-CAD controls were selected among HUNT 2 and HUNT 3 participants with available DNA
(N = 70 300) after excluding individuals with the following hospital diagnosed or self-reported conditions in themselves or known first and/or second degree family members: MI, angina, heart failure, stroke, aortic aneurysm, atherosclerosis, intermittent claudication and registered percutaneous coronary angioplasty procedures or bypass surgery. In addition, for individuals <80 years of age, we excluded individuals with diabetes, self-reported use of antihypertensive medication and hospital diagnosed hypertension or hyperlipidemia. Of 15 841 remaining individuals <80 years of age (22 389 individuals for all age groups), we selected one control for each case by matching on sex, birth year (±1 year), and municipality or geographical region to minimize population stratification (Supplementary Material, Table S11). In addition, 3833 non-MI controls were included from the same cohort from a separate study of HDL extremes. Clinical characteristics are described in Supplementary Material, Table S1.

Stage 1 sample genotyping
DNA was extracted from peripheral blood leukocytes from whole blood or blood clots stored in the HUNT Biobank, using the Puregene kit (Genta Systems, Minneapolis, MN) or with an AutoPure LS (Genta Systems) (20).

Genotyping of 2969 cases and 2968 controls were carried out at the Norwegian University of Science and Technology, Norway, using the iSelect Exomechip V1.0 (Illumina, San Diego, CA) and the Infinium HD ultra protocol. Each 96-well plate included both case and control individuals in random order and one sample of reference DNA that was present on every plate. Genotype calling was done using GenTrain version 2.0 in GenomeStudio V2011.1 (Illumina, San Diego, CA) followed by zCall version 2.2 (32). Genotyping of 3919 additional controls and 43 additional cases was performed by the University of Washington Resequencing and Genotyping Laboratory Center of NHLBI. Genotype clusters were generated separately for the samples genotyped at the Norwegian University of Science and Technology (N = 5937) and the University of Washington (N = 3919). Genotype calls from the two groups were combined prior to quality control procedures.

Samples with <99% genotype completion rate (N = 11), evidence of gender discrepancy (N = 9), duplicates (N = 47) and/or heterozygosity rate ± 3 standard deviations from the mean (N = 184) as well as potential outliers identified by plotting the first 10 genotype-driven principle components (N = 8) were excluded from further analysis (N = 255, 2.58%). In addition, variants that did not meet a 99.9% genotyping threshold or showed deviation from Hardy–Weinberg equilibrium (P < 1 × 10−8) were removed (7904 SNPs). To avoid batch effects, we excluded genotype data from the additional controls for variants that demonstrated allele frequency differences between the non-MI control samples from the two genotyping centers, or between all samples at the two genotype centers (N = 48 variants, P < 1 × 10−4). After quality assessment, 234 997 variants (96.7%) genotyped in up to 9644 individuals (96.5% of cases and 97.8% of controls) remained for further analysis.

Of the 234 997 variants passing quality control, 105 112 were found to be polymorphic. Of the polymorphic variants, 85 972 were coding variants (LoF or missense) and 74 624 (86.8%) of these had a frequency <5%. The number of coding variants available for single-variant tests was 56 108 (MAC ≥6). The number of variants genotyped with different frequencies is given in Table 1.

Exome sequencing
We performed exome sequencing on 75 cases and 76 controls at the University of Michigan DNA Sequencing Core using Illumina Hi-Seq machines. Exome enrichment was carried out using the SeqCap EZ Human Exome Library v3.0 (Roche NimbleGen, Madison, WI), with 45 × average coverage for target regions. Non-reference genotype concordance for sequence-based genotypes and exome array genotypes was 99.4% for variants >5% frequency, 99.7% for variants with 1–5% frequency, and 99.5% for variants with frequency <1%.

Stage 2 study population and sample genotyping
We selected 2364 MI cases and 2330 controls from the Tromsø study (Tromsø 4) for follow up. Tromsø 4 was a population-based study conducted in 1994–1995 with questionnaires, clinical examination and non-fasting venous blood samples (stored at HUNT biobank) from 27 158 individuals of 25 years of age and older from the municipality of Tromsø, Norway (22). Cases of incident MI were identified by linkage to the diagnosis registry at the University Hospital of North Norway (the only hospital in the region) and to the National Causes of Death Registry from the date of enrollment to the end of the study period, 31 December 2010. Adjudication of hospitalized and out-of-hospital events was performed by an independent endpoint committee using data from hospital and out-of-hospital journals, autopsy records and death certificates. Slightly modified World Health Organization MONICA/MORGAM criteria for MI were used and included clinical symptoms and signs, findings in electrocardiograms, values of cardiac biomarkers and autopsy reports, when applicable. As in HUNT, age at earliest MI incident was registered for subjects with more than one MI admission. As controls, we selected individuals with no hospital diagnosed history of MI or self-reported history of cardiovascular disease among themselves or first and second degree relatives. Cases and controls were matched for sex and birth year (±2.5 years). Both the populations of Nord-Trøndelag County and Tromsø are ethnically homogeneous (<3% non-Caucasian ethnicity), making it especially suitable for epidemiological genetic research (33).

We selected for replication single variants with P < 1 × 10−3, MAF < 10% and MAC > 5 in a preliminary dataset to capture variants less likely to be covered by GWAS but with enough variant alleles so that replication was still a possibility. We also selected SNPs contributing to burden tests if they had gene-wise -P < 5 × 10−4, single SNP P < 0.05, and MAC > 5. We verified that gene-wise results had similar evidence of association using the subset of variants selected for follow-up (difference in log P < 0.5).

Overall, 37 SNPs were selected for replication: 25 single-variant SNPs and 12 gene-based SNPs passing our replication thres-holds. Of the SNPs selected for replication, two single-variant SNPs and one gene-based SNP failed array design and were excluded prior to genotyping. We genotyped the remaining 34 SNPs in 4694 individuals at the Centre for Integrative Genetics.
Statistical analyses

We determined statistical power using the power calculator CaTS (19) assuming ~5% MI prevalence in the HUNT study (equivalent to 3510 MI cases among 62,306 individuals with available DNA at the time of our study inclusion. We evaluated the impact of cryptic relatedness using a linear mixed model (EMMAX) (34) and found the results to be highly correlated with logistic regression (Pearson $r^2 = 0.77, P < 2.2 \times 10^{-16}$). The linkage disequilibrium (LD) metric $r^2$ was estimated using 5803 individuals from Stage 1 MI cases and matched controls that passed quality control.

Single-variant tests for association with MI case/control status were performed using logistic regression, with covariates for age, sex and ancestry informative principal components 1–10 derived from all common genetic variants as implemented in PLINK (35). Our study did not show evidence for inflated test statistics or population stratification ($\alpha_{GC}$ overall $= 1.03$), nor when variants were grouped according to frequency ($\alpha_{GC}$ MAF $> 5\% = 1.11$, $\alpha_{GC}$ MAF $1–5\% = 1.02$, $\alpha_{GC}$ MAF $< 1\% = 0.99$).

We also performed conditional logistic regression accounting for CAD or MI GWAS index SNPs within 500 kb. Weighted allele scores were calculated by multiplying the sum of the effective allele dosage of the previously known CAD SNPs by their estimated log ORs from the combined analysis in the previous studies (1,3,4).

We performed gene-based and pathway-based burden tests using tests of association with rare variants weighted by allele frequency (SKAT-O) (27) and by using different frequency thresholds for inclusion of variants (VT) (28). We meta-analyzed results from the HUNT and Tromso studies using standard error weighted fixed-effect methods implemented in METAL (36).

For gene-based burden tests, we declared a significance level of $7 \times 10^{-7}$ corresponding to a Bonferroni correction for the 73,887 single-variant tests with MAC $> 5\%$, and a $1.2 \times 10^{-8}$ significance level for the 12,678 genes tested using two different models (38 210 tests overall): all non-synonymous variants (12,678 genes), variants predicted to be LoF (253 genes) and LoF and ‘probably damaging’ missense variants as defined by PolyPhen-2 (6174 genes). We set the significance level at $P < 4 \times 10^{-7}$ for the 1277 pathway-based tests performed using SKAT-O (27).

We estimated the contribution of variants with MAC $> 5\%$ on the exome array to the MI heritability using a linear mixed model as implemented in the GCTA software (31). The proportion of variance of MI risk explained by rare (six copies $\sim 1\%$), low-frequency (1–5%) and common (>5%) variants was calculated adjusting for age, sex and the first 10 ancestry informative principal components.

A gene set enrichment analysis approach implemented in MAGENTA was used to identify potential predefined biological pathways that influence MI risk (26). Ten thousand nine hundred and ninety-two pathways that are annotated from the Kyoto Encyclopedia of Genes and Genomes (KEGG; $N = 186$) (37), Ingenuity ($N = 92$) (www.ingenuity.com), Protein Analysis Through Evolutionary Relationships (PANTHER; $N = 634$) (38), BioCarta ($N = 217$) (http://cgap.nci.nih.gov/Pathways/BioCarta_Pathways), Reactome ($N = 430$) (39) and Gene Ontology ($N = 9433$) (30) databases were tested. $N$ is the number of pathways from each database.

Annotation

Variants were annotated as missense, splice, premature stop, read-through, synonymous or non-coding using ANNOVAR (Version 2012-05-15) (18). Variants are listed with respect to the hg19 genome build.

Ethics

Both the HUNT and Tromso study were conducted according to the principles expressed in the Declaration of Helsinki. Attendance was voluntary, and each participant signed a written informed consent including information on genetic analyses. Our study was approved by the Regional Committees for Medical and Health Research Ethics (REC Central), Norway.

SUPPLEMENTARY MATERIAL

Supplementary Material is available at HMG online.

ACKNOWLEDGEMENTS

The Nord-Trøndelag Health Study (The HUNT Study) is a collaboration between HUNT Research Centre (Faculty of Medicine, Norwegian University of Science and Technology NTNU), Nord-Trøndelag County Council, Central Norway Health Authority, and the Norwegian Institute of Public Health. The authors wish to thank all HUNT and Tromso study participants who contributed to scientific research.

Conflict of Interest statement. None declared.

FUNDING

Support from the National Institutes of Health (grant numbers: HL094535, HL109946 and DK062370) and the Liaison Committee between the Central Norway Regional Authority and the Norwegian University of Science and Technology (grant: 2011/13806). Genotyping services for 3992 samples were provided by the Northwest Genomics Center at the University of Washington, Department of Genome Sciences, under US Federal Government contract number HHSN268201100037C from the National Heart, Lung, and Blood Institute. The funders had no role in study design, data collection and analysis, decision to publish or preparation of the manuscript.

REFERENCES


