The molecular basis of early onset cardiovascular disease
Sivapalaratnam, S.

Citation for published version (APA):

General rights
It is not permitted to download or to forward/distribute the text or part of it without the consent of the author(s) and/or copyright holder(s), other than for strictly personal, individual use, unless the work is under an open content license (like Creative Commons).

Disclaimer/Complaints regulations
If you believe that digital publication of certain material infringes any of your rights or (privacy) interests, please let the Library know, stating your reasons. In case of a legitimate complaint, the Library will make the material inaccessible and/or remove it from the website. Please Ask the Library: http://uba.uva.nl/en/contact, or a letter to: Library of the University of Amsterdam, Secretariat, Singel 425, 1012 WP Amsterdam, The Netherlands. You will be contacted as soon as possible.
RARE MUTATIONS IN MCF2L AND ZC3HC1 ARE ASSOCIATED WITH PREMATURE ATHEROSCLEROSIS


*Equally contributed

Department of Vascular Medicine (SS, SM, MM, GKH, ML, JJP and MDT), Academic Medical Center, Amsterdam, The Netherlands; Department of Human Genetics (CA, WHO), Wellcome Trust Sanger Institute, Hinxton, United Kingdom; Department of Hematology (JSam, WHO), University of Cambridge, Cambridge, United Kingdom; National Health Service Blood and Transplant (WHO), Cambridge, United Kingdom; Department of Cardiology (MDT), Academic Medical Center, Amsterdam, The Netherlands; Department of Cardiovascular Sciences (CN, NS), Leicester University, Leicester, United Kingdom.

Preliminary report
ABSTRACT

Heritable factors account for 30%–60% of the inter-individual variation in the risk of cardiovascular disease (CVD), but to date, a mere 10% of the heritability is explained by common genetic variants. Our aim was therefore to identify novel genes involved in premature coronary artery disease (CAD) in families suffering from autosomal dominantly inherited CAD. We performed exclusion linkage analysis and identified rare mutations in ZC3HC1 (NM_016478.3, c.913A>G, p.Ile305Val) and MCF2L (NM_001112732.1, c.2066A>G, p.Asp689Gly). It is noteworthy that common variation in ZC3HC1 has been associated to CAD risk in recent large genome wide association studies (GWAS). For MCF2L, SNP analysis in the Cardiogram Consortium resulted in an association that did not reach GWAS significance. Further studies are warranted to unravel the putative role of these genes in the complex phenotype of premature atherosclerosis.
REPORT

Cardiovascular disease (CVD) is the major cause of morbidity and mortality in most societies. In combination with the strong influence of environmental factors, genetic determinants are pivotal in the etiology of atherosclerosis, the pathophysiological process ultimately leading to clinical cardiovascular events.

So far, twenty-five loci comprising common genetic low-risk variants associated with Coronary Artery Disease (CAD) were identified by Genome Wide Association (GWA) studies. These variants combined explain up to 10% of the heritability of CAD. The inability to unravel the full heritability of CAD is merely a reflection of its genetic heterogeneity and of our limited ability to classify CAD cases into subgroups that are more homogeneous in their molecular pathology.

Mendelian disorders are considered the most ultimate form of homogeneity in human disease, and as such, hold great promise to identify monogenetic causes of disease. We therefore set out to identify the molecular defect in a family with an autosomal dominant form of premature CAD.

The index case in this family is a Caucasian female patient of Dutch descent who was admitted to the hospital with an acute myocardial infarction (AMI) at the age of 39 years (figure 1 for pedigree and supplement for CAD definitions). Her past medical history was unremarkable and no CVD risk factors were present except for smoking and a strong familial predisposition for premature CAD. Upon family collection, a total of four siblings with premature CAD were identified. Only one individual (II-4) was classified as unaffected based on the absence of any clinical CVD event and in this subject, no signs of subclinical atherosclerosis were identified by coronary Computer Tomography (CT) angiography; her calcium-score was 0 and no vascular abnormalities were present. Lipid levels were within the normal range in all cases and Familial Hypercholesterolemia was ruled out by sequencing of LDL-C related genes in the index case.

The maximum theoretical LOD score in this core family was 1.5 and we therefore followed an approach where we combined the data derived from exome sequencing and exclusion linkage analysis. A detailed description of the genotyping and sequencing protocols are provided in supplementary methods. In short, the Agilent SureSelect protocol was exploited to enrich for 39.3 Mb of coding sequence and the Illumina GAII platform was used for exome sequencing of the index case and the cousin (II-1). A total of five family cases (II.2; II.3; II.5-II.7) and two controls (I.4 and II.4) of the core family were genotyped using a human CytoSNP-12 BeadChip array (Town Centre Drive, San Diego, CA, USA) for linkage analysis. A total of 45310 variants met the commonly described quality criteria of which 4510 were non-synonymous, indels or splice site variations. 961 of these variants were not present in the large public databases (dbSNP131, SNPs...
from 8 HapMap exomes\textsuperscript{5}, the 1000 Genomes pilot project SNPs and indels \textsuperscript{7} (http://www.ncbi.nlm.nih.gov/projects/SNP/)(see supplement table 1). The vast majority of these variants (899 = 93\%) were located within the regions with LOD scores below -2, and were therefore excluded from further analysis. Fifteen variants were predicted to have deleterious effects by SIFT and Polyphen predictions\textsuperscript{8,9} and of these, 3 dideoxynucleotide sequencing confirmed mutations were shown to co-segregate within the core family. The heterozygous mutations were identified in the genes encoding for calmodulin regulated spectrin-associated protein 1 (\textit{CAMSAP1}) on chromosome 9q34, MCF.2 cell line derived transforming sequence-like (\textit{MCF2L}) on chromosome 13q34 and zinc finger C3HC-type containing 1 (\textit{ZC3HC1}) on chromosome 7q32.

\textbf{Figure 1.} Pedigree of the family with early onset CAD

Arrow for the index case. The right half of a symbol denotes CVD status. If fully colored this indicates CVD. On the left half, the presence of mutation in \textit{ZC3HC1} and \textit{MCF2L} is indicated with green (top corner) and yellow (bottom corner) colors respectively. The first number is an identifier within the generation. The number between brackets is the current age. AMI denotes Acute Myocardial Infarction; AP denotes Angina Pectoris, PTCA denotes Percutaneous Transluminal Coronary Angiography; CABG denotes Coronary Artery Bypass Graft; this line also contains the age of event in years. RF denotes Risk Factors at time of event; py denotes cigarette pack years; HT denotes Hypertension, None of the subjects suffered from diabetes mellitus or was obese. FH was excluded in the index case (arrow). Subsequently, the lipid profiles are displayed; TC, HDL-c, LDL-c and TG in mmol/L. On the final line Ator denotes atorvastatin; Simva denotes simvastatin and Ez denotes ezetimibe treatment.
To our surprise, none of these variants were identified in the ID II-1, a cousin who was included in this study after performing the linkage and sequencing. He suffered from an AMI. The cousin and the index case shared 7 genetic variants in KIAA1875, C12orf53, NOP2, USO1, USP34, GSPT1 and OPLAH. However these variants did not co-segregate with the phenotype in the core family and we therefore consider the cousin to be a phenocopy.

We genotyped the 3 variants co-segregating in the core family in healthy blood donors: 1000 randomly selected from the Dutch Blood Center (Sanquin Cohort/DBC) and 8000 from the Cambridge Bio Resource Cohort (CBR, see supplementary methods). In addition, we genotyped 1000 Dutch patients with early onset CAD, a selection of the youngest cases from the AMC-PAS Cohort (supplementary methods). Given the relatively high allele frequency of the CAMSAP1 variant in the 1000 DBC donors (n=13, AF 13/2000), this variant was excluded from further analysis (see supplementary methods). No carriers for the MCF2L variant were identified in the Sanquin, CBR and AMC-PAS cohort. The carrier frequency of the ZC3HC1 variant was 0.003, 0.006 and 0.005 in the Sanquin, CBR and AMC-PAS cohorts, respectively. The five carriers of the ZC3HC1 variant within the AMC-PAS cohort were contacted, and 4 were willing to participate in this family based research. Upon expansion of these four families, 16 carriers and 18 non-carriers were identified. Eleven of the heterozygous ZC3HC1 carriers suffered from CAD (mean age at event in years ± SD; 43.6 ± 9.8) and 5 did not report any CAD (mean age at genotyping in years 50.2 ± 15.8). Three of the non-carriers suffered from CAD (mean age at event 46.7 ± 11.5) and 15 non-carriers did not suffer from an event (mean age at genotyping 47.4 ± 22.7). Because of the young age of some family members, it could not be ascertained whether they suffered from atherosclerotic vascular disease. Due to the low age or unwillingness of the participants we were not able to classify them with a coronary CT scan.

Taken these data for ZC3HC1 including the original family studies, there is 1 carrier without penetration of the phenotype and 5 phenocopies in all CAD families that were identified and studied. Still, carriers of the ZC3HC1 variant have a high relative risk of 4.4 (p=0.001) of CAD. If we only include those individuals aged 30 years and above at time of examination the relative risk is still 3.4 (p=0.005).

To the best of our knowledge, no associations between common variations in the locus harboring the MCF2L gene and CAD risk have been described thus far. Analysis of the Cardiogram containing 22,000 cases and 60,000 controls showed that rs9577428 SNP, located intronically in MCF2L was associated with CAD risk. However, as expected, the associated p value of 9 x 10-4 did not meet genome wide significance.

The putative role of ZC3HC1 in CAD was substantiated by the finding in a recently published GWAS, that common variations in ZC3HC1 are associated with CAD risk. Carrier ship of the minor alleles of rs11556924 and rs3825897 were associated with a CAD odds ratio (OR) of 1.09(1.05-1.14)2, 3.
Little is known about the function of these genes in relationship to CVD.

*ZC3HC1* is a nuclear protein of 502 amino acids, with a size of 5262 Da, which is highly expressed in heart, skeletal muscle and testis. It has a putative role in apoptosis and cell cycling and interacts with ALK1, SKP1, RBx1, CCNB110.

A conditional *ZC3HC1* knockout (KO) mouse line was generated as part of the International KO Mouse Consortium program (www.knockoutmouse.org). *ZC3HC1* knockout mice fed a high fat diet (21.4% crude fat content, Western RD, 829100, Special Diets Services) for 12 weeks, starting at the age of 4 weeks, show a decrease in the total number of circulating mature B cell IgD+, which has a role in activation of other B-cells. In addition, male mice fed the high fat diet for 8 weeks showed increased carbon dioxide production, increased oxygen consumption, and increased energy expenditure. *ZC3HC1* might therefore be related with CVD both by a direct effect on inflammation and/or energy metabolism.

*MCF2L* encodes for the Guanine nucleotide exchange factor (GEF) Dbs, a plasma membrane protein, located at the cytoplasmic side which contains 1137 amino acids. It is...
highly expressed in the cerebellum and it directly interacts with Cdc42, RAC1, RhoG and RhoH\textsuperscript{11}. The GEFs together with the Rho kinases play a pivotal role in leukocyte migration\textsuperscript{12}. Rho G family GTPases act as molecular switches that are turned on in the GTP-bound state and turned off in the GDP-bound state in response to a variety of stimuli, including growth factors and cytokines and cell-cell or integrin-extra cellular matrix interactions\textsuperscript{13}. It is thus likely that activation of Rho family GTPases via \textit{MCF2L} will affect many cell systems by mediating signalling through multiple pathways. In men, little is known about the phenotypical effects of molecular defects in \textit{MCF2L}. SNPs in \textit{MCF2L} have been described to be associated with osteoarthritis\textsuperscript{14}. None of the carriers of the .. variant in the family studied by us did report clinical signs or symptoms of arthritis.

Our work so far has three strengths. Firstly, the state of the art in depth phenotyping applied in this study. Because of this, we were able to confirm by coronary CT and calcium scoring that the apparently unaffected healthy sister of the index was truly free of atherosclerosis.

Secondly, the application of next generation sequencing together with exclusion linkage, which has proven to be a successful method to study genetics of rare variants in smaller families.

Finally, the identified rare variant sits in \textit{ZC3HC1}, for which common variants have shown to be associated with early onset CAD in GWAS. The fact that this gene emerges as a significant hit in two different independent approaches raises the level of confidence that this locus has indeed a role in atherosclerosis. We also carefully speculate that indirectly this validates \textit{MCF2L} as a true gene associated with CAD.

Several issues merit discussion. Firstly, we are in search of a rare variant, which should not be present in apparently healthy individuals. We used public available databases to filter the variants obtained from exome sequencing. Individuals within the public databases, however, could suffer from subclinical atherosclerosis, and we might have unrightfully filtered out variants associated with atherosclerosis. In light of the filters applied it is therefore remarkable that the variants considered to be rare based on the public databases were present in healthy controls derived from the geographic region where this study took place. This emphasizes the need to study allele frequencies of putative rare variants in well phenotyped regional control cohorts.

Secondly, there is only a small difference in carrier frequency between apparently healthy controls of the Dutch Blood Center and the AMC-PAS cohort for \textit{ZC3HC1}. Thirdly, unfortunately the family of one of the carriers of \textit{ZC3HC1} was not willing to participate. Co-segregation analysis of the variant within this pedigree could thus not be performed, which might have resulted in an overestimation of the effect size of the rare variants we identified.
Fourthly, we have now concluded that the cousin is a phenocopy. We can not rule out that there is a non-exonic genetic variant which is causative of the phenotype in this family, which we, due to the technique applied, we were not able to identify.

Finally, the major pitfall is that no biological plausible mechanisms underlying our findings have been identified yet, but studies mice models with local overexpression of the genes of interest are ongoing.

In conclusion, we have identified rare mutations in MCF2L and ZC3HC1 in a family with an autosomal trait for premature CAD. The relevance of these 2 putative genes is underscored by the finding that common variations within these genes are associated with increased CAD risk in GWAS. Further studies are warranted to unravel the pathophysiological role of these genes in the complex multifactorial phenotype of premature atherosclerosis.

FUNDING SOURCES

This project was funded by Bloodomics consortium, European Union 6th Framework Programme (LSHM-CT-2004-503485) and Ipse Movet. JK is holder of the Lifetime Achievement Award of the Dutch Heart Foundation (2010T082). WHO is supported by programme grants from the British Heart Foundation and the National Institute for Health Research, England. GKH is holder of a Veni grant (91612122) from the Netherlands Organisation for Scientific Research (NWO)

ACKNOWLEDGEMENTS

We are indebted to J. Kuiper for data management, C. Holtkamp for genetic fieldwork.
REFERENCES


229
SUPPLEMENTARY INFORMATION

CAD definitions
CAD was defined by the presence of at least one of the following: (I) myocardial infarction, proven by at least two of the following: (a) classical symptoms (>15 min), (b) specific electrocardiographic abnormalities, (c) elevated cardiac enzymes (>2× upper limit of normal); (II) percutaneous coronary intervention or other invasive procedures; (III) coronary artery bypass grafting; (IV) angina pectoris, diagnosed as classical symptoms in combination with at least one unequivocal result of one of the following: (a) exercise test, (b) nuclear scintigram (MIBI) (c) dobutamine stress ultrasound, (d) a more than 70% stenosis on a coronary angiogram or (f) requiring treatment.

Cardiac Imaging
CT was performed using a 64-slice multidetector CT scanner (Philips Medical Systems, Best, the Netherlands). The scanning protocol was as follows: tube voltage, 120 kV; tube current, 55 mAs; detector collimation, 40 × 2.5 mm; gantry rotation, 420 ms. The data was transferred to a post processing workstation (Extended Brilliance Workplace, Philips Medical Systems). We recorded CCS for the main arteries, then the total score was calculated by summing lesion scores of all sections. We evaluated CCS according to Agatston and expressed further as age/sex percentiles (ref. agaston). A score >75th percentile is considered as abnormal.

Linkage analysis
DNA of 7 family members was genotyped using Human CytoSNP-12 DNA analysis BeadChip kits (Illumina) according to manufacturer’s protocols. In order to verify the relationship between individuals, the data were subjected to standard quality control routines, including Graphical Representation of Relationship errors (GRR) and PedCheck packages 1-2. Multipoint parametric linkage analysis assuming a fully penetrant autosomal dominant model and minor allele frequency of 0.001 was conducted using the Allegro program and for this purpose, >110000 high-quality SNPs throughout the genome was used 3. Genomic intervals with LOD score > -2 were considered for filtering exome sequence data.

Exome sequencing
We applied the Agilent SureSelect protocol (Agilent, South Queensferry, UK, catalogue no. G3362A) to enrich for 39.3 Mb of sequence covering 740K exons in 79K transcripts from a highly redundant set of 34,642 genes. The enriched DNA was sequenced on the Illumina GAII platform (Illumina, Little Chesterford, UK). We generated 13.3 Gb of
sequence, resulting in a 106-fold mean coverage and 91% of the targets covered at least ten-fold.

**Sequence analysis**
The sequence analysis was divided into three phases with the first one consisting of aligning and recalibrating the reads to produce a BAM file\(^4\). The second phase consisted of calling the sequence variants using various algorithms. The third phase consisted of filtering the set of variants.

**First phase:**
1. The FastQ files generated by the sequencing pipeline were aligned to reference NCBI build 37 using the Stampy read mapper\(^5\).
2. For each lane independently, the reads were realigned around known insertions and deletions (indels) (1000 Genomes pilot project\(^6\)), followed by base quality recalibration using the Genome Analysis Toolkit (GATK)\(^6\).
3. After merging the different lanes, duplicates were flagged using the Picard software package (Web-based application Picard. [http://picard.sourceforge.net](http://picard.sourceforge.net) (Accessed 12/2010)) and excluded from subsequent analyses.

**Second phase:**
1. Indels were called using Dindel\(^7\) on each sample independently.
2. The reads were realigned in the BAM files produced in phase 1 around the indels called by Dindel using the GATK, in order to reduce alignment artefacts around all indels found in a sample.
3. The GATK software package and SAMtools were used to call Single Nucleotide Polymorphisms (SNPs)\(^4\). We called SNPs both for each sample independently and jointly using all samples. Furthermore, SNPs were called both from the BAM files produced in phase 1 and in step 2 of the second phase, so that SNPs were called using six different approaches. The joint calling was performed using exome sequencing results of 32 additional DNA samples from individuals with unrelated clinical phenotypes on the realigned BAM files produced in step 2 of the second phase.
4. We required a minimum variant quality of 20, which corresponds to a 99% confidence level.

**Third phase:**
1. Variants that did not overlap the targeted regions ±25 bp were filtered and not considered in further analyses.
2. We predicted the functional consequence of SNPs and small indels using the Ensembl Variation API\(^8\).
3. We considered only non-synonymous (ns) SNPs, SNPs introducing a premature stop-codon, indels in protein coding sequence, and variants affecting 3'UTRs and essential splice sites.

4. Each variation was annotated for presence in databases of genetic variation. Any variant found to be present was not considered as a candidate for GPS. The following public databases were used: dbSNP131, SNPs from 8 HapMap exomes, the 1000 Genomes pilot project SNPs and indels. In addition to these public resources we also made use of in-house databases of genetic variation. We also used as a filter the variants called on 32 exomes sequenced on the same platform for unrelated studies, which allowed us to correct for systematic errors in the variant calling.

**Sanger sequencing**

For Sanger sequencing different primer sets were used. Every primer set for the specific gene had an additional M13 tail attached for the sequencing reaction with BDTs. Full protocol detail can be found elsewhere.

<table>
<thead>
<tr>
<th>Gene</th>
<th>Primer sequence 5' – 3'</th>
</tr>
</thead>
<tbody>
<tr>
<td>MCF2L</td>
<td>Forward: TGCTTTGCTTTGATGGATG</td>
</tr>
<tr>
<td></td>
<td>Reverse: CATTCCAGCCCCCTGAAG</td>
</tr>
<tr>
<td>ZC3HC1</td>
<td>Forward: GAGAAAAACTCTCTTTTTCATCC</td>
</tr>
<tr>
<td></td>
<td>Reverse: CACCCCAAATAAGCTAAGTGAATAC</td>
</tr>
<tr>
<td>M13</td>
<td>Forward: TGTTGTAAAAACGACGGCCAGT</td>
</tr>
<tr>
<td></td>
<td>Reverse: CACAGGAACACGCTATGACC</td>
</tr>
</tbody>
</table>

**Genotyping variants for Co-segregation and in Cohorts**

For MCF2L and ZC3HC1 genotyping, allelic discrimination was performed using FAM and VIC as fluorophore. PCR conditions were denaturation for 10 min at 95°C, followed by 40 cycles (30 sec 92°C, 45 sec 60°C). PCR assay mix was obtained from Applied Biosystems (Applied Biosystems by Life Technology, ABI, Nieuwekerk aan de IJssel, NL).

**Cohorts**

Sanquin cohort/DBC is a collection of DNA samples from blood donors recruited at routine Sanquin Blood Bank donation sessions. More than 95% of the controls are from the same region as the cases of the AMC-PAS cohort.

Cambridge BioResource cohort (CBR) is a collection of 8000 pseudo-anonymised DNA samples of healthy blood donors established by the Cambridge Biomedical Research Centre in collaboration with NHS Blood and Transplant for use in genotype-phenotype association studies.
AMC-PAS cohort is a prospective study cohort recruited from the Academic Medical Centre Amsterdam and consists of patients with Premature AtheroSclerosis (PAS), defined as symptomatic CAD before the age of 51 years, defined as MI, coronary revascularization, or evidence of at least 70% stenosis in a major epicardial artery.

**Supplement Table 1.** Filtering strategy for analysis of exome data

<table>
<thead>
<tr>
<th>Total variants passing QC; 5X</th>
<th>45310</th>
</tr>
</thead>
<tbody>
<tr>
<td>Non-synonymous, splice site, indel</td>
<td>4510</td>
</tr>
<tr>
<td>Not in dbSNP, 1000 genomes</td>
<td>961</td>
</tr>
<tr>
<td>Not excluded by linkage</td>
<td>62</td>
</tr>
<tr>
<td>Serious consequence</td>
<td>15</td>
</tr>
<tr>
<td>Co-segregating with phenotype</td>
<td>3</td>
</tr>
<tr>
<td>Low frequency in general population</td>
<td>2</td>
</tr>
</tbody>
</table>

Although no significant LOD score (>3.3) was present, a large fraction of the genome was determined to be excluded from linkage to the phenotype (i.e. regions with LOD score < -2). These regions were considered for filtering exome sequence data in the next steps. LOD scores below -3 are not shown in the plot.