Genome-wide haplotype association study identifies the SLC22A3-LPAL2-LPA gene cluster as a risk locus for coronary artery disease

nature genetics

Genome-wide haplotype association study identifies the *SLC22A3-LPAL2-LPA* gene cluster as a risk locus for coronary artery disease

David-Alexandre Trégouët[1], Inke R König[2], Jeanette Erdmann[3], Alexandru Munteanu[1], Peter S Braund[4], Alistair S Hall[5], Anika Großhennig[2,3], Patrick Linsel-Nitschke[3], Claire Perret[1], Maylis DeSuremain[1], Thomas Meitinger[6], Ben J Wright[7], Michael Preuss[2], Anthony J Balmforth[5], Stephen G Ball[5], Christa Meisinger[6], Cécile Germain[8], Alun Evans[9], Dominique Arveiler[10], Gérald Luc[11], Jean-Bernard Ruidavets[12], Caroline Morrison[13], Pim van der Harst[4], Stefan Schreiber[14], Katharina Neureuther[15], Arne Schäfer[14], Peter Bugert[16], Nour E El Mokhtari[14], Jürgen Schrezenmeir[17], Klaus Stark[15], Diana Rubin[17], H-Erich Wichmann[6], Christian Hengstenberg[15], Willem Ouwehand[18], Wellcome Trust Case Control Consortium[19], Cardiogenics Consortium[19], Andreas Ziegler[2], Laurence Tiret[1], John R Thompson[7], Francois Cambien[1], Heribert Schunkert[3] & Nilesh J Samani[4]

We identify the *SLC22A3-LPAL2-LPA* gene cluster as a strong susceptibility locus for coronary artery disease (CAD) through a genome-wide haplotype association (GWHA) study. This locus was not identified from previous genome-wide association (GWA) studies focused on univariate analyses of SNPs. The proposed approach may have wide utility for analyzing GWA data for other complex traits.

Our analysis comprised three stages. In the first stage, the Wellcome Trust Case Control Consortium (WTCCC) CAD study, composed of 1,926 CAD cases and 2,938 controls genotyped with the Affymetrix 500K chip[1,2], was used to identify regions with suggestive evidence of haplotype association using a sliding-windows approach (**Supplementary Methods** online). The analysis was carried out using the European EGEE grid[3] of computers. We searched for regions fulfilling two a priori defined criteria: (i) a haplotype-based *P* value lower than 10^{-5} and (ii) a haplotype-based *P* value at least 100 times smaller than the smallest single-SNP *P* value observed in the corresponding region. We fixed the first criterion at a threshold that was not too stringent, in order to increase the sensitivity of the analysis given the higher number of degrees of freedom of haplotype tests as compared to single-SNP tests. The second criterion was applied to identify new susceptibility regions where haplotypes are more informative than single SNPs to predict disease status. Using these criteria, we identified 29 regions (involving 94 SNPs) in the WTCCC data (**Supplementary Table 1** online).

In the second stage, we investigated these regions in the GWA data of the German Myocardial Infarction Family Study I (GerMIFS I)[1], composed of 875 CAD cases and 1,644 controls typed with the same Affymetrix array. Because of the lower performance of the algorithm used for calling GerMIFS I genotypes[1], 24 of the 94 SNPs did not pass the quality control and the 21 regions involving these SNPs were not further analyzed. We confirmed the high validity of the remaining genotypes that were called in the GerMIFS I study by comparing the genotypes obtained on the GWA platform for a proportion of the SNPs with those obtained with an alternate platform (**Supplementary Methods**). Of the eight regions where exact haplotype replication analysis could be undertaken, one showed strong evidence for association ($P = 3.70 \times 10^{-4}$ after Bonferroni correction for the number of tested regions). This region includes four SNPs overlapping the *SLC22A3-LPAL2-LPA* gene cluster on chromosome 6q26–q27 (see **Supplementary Fig. 1** online for their pattern of linkage disequilibrium). The frequency distribution of the haplotypes derived from these four SNPs was consistently and significantly different between cases and controls of the WTCCC and the GerMIFS I studies (**Table 1**).

In a third stage, we investigated the association of the haplotypes derived from these four SNPs (rs2048327 in the *SLC22A3* gene, rs3127599 in the *LPAL2* gene and rs7767084 and rs10755578 in the *LPA* gene) with CAD in four additional studies composed of 1,222

[1]Institut National de la Santé Et de la Recherche Médicale (INSERM) Unité Mixte de Recherche (UMR_S) 525, Université Pierre et Marie Curie (UPMC). Paris 06, Paris 75013, France. [2]Institut für Medizinische Biometrie und Statistik, Universität zu Lübeck, D-23538 Lübeck, Germany. [3]Medizinische Klinik II, Universität zu Lübeck, Ratzeburger Allee 160, D-23538 Lübeck, Germany. [4]Department of Cardiovascular Sciences, University of Leicester, Clinical Sciences Wing, Glenfield Hospital, Groby Road, Leicester, LE3 9QP, UK. [5]Leeds Institute of Genetics, Health and Therapeutics (LIGHT), University of Leeds, Leeds, LS1 3EX, UK. [6]Institut für Humangenetik, Helmholtz Zentrum München, Deutsches Forschungszentrum für Umwelt und Gesundheit, D-85764 Neuherberg, Germany. [7]Department of Health Sciences and Genetics, University of Leicester, LE1 7RH Leicester, UK. [8]Centre National de la Recherche Scientifique (CNRS), UMR_S 8623, Université Paris-Sud, Laboratoire de Recherche en Informatique, 91403 Orsay, France. [9]Department of Epidemiology and Public Health, Queen's University of Belfast, BT7 1NN Belfast, UK. [10]EA1801-Laboratoire d'Epidémiologie et de Santé Publique, Faculté de Médecine, Multinational MONItoring of Trends and Determinants in CArdiovascular Disease (MONICA), Strasbourg, 67085 Strasbourg, France. [11]INSERM U545, Pasteur Institute of Lille, University of Lille 2, 59019 Lille, France. [12]INSERM U518, Faculté de Médecine, MONICA Toulouse, 31073 Toulouse, France. [13]The MONICA Project, Glasgow Royal Infirmary, Scotland, G4 0SF, UK. [14]Institut für Klinische Molekularbiologie, Christian-Albrechts Universität, D-24105 Kiel, Germany. [15]Klinik und Poliklinik für Innere Medizin II, Universität Regensburg, D-93042 Regensburg, Germany. [16]Molecular Biology Laboratory, Institute of Transfusion Medicine and Immunology, University of Heidelberg, Medical Faculty of Mannheim Mannheim, D-69117 Germany. [17]Bundesforschungsanstalt für Ernährung und Lebensmittel, Institut für Physiologie und Biochemie der Ernährung, D-24121 Kiel, Germany. [18]Department of Haematology, University of Cambridge, Long Road, Cambridge, CB2 2PT, UK. [19]A full list of members is provided in the **Supplementary Note** online. Correspondence should be addressed to D.-A.T. (david.tregouet@upmc.fr).

Received 25 July 2008; accepted 22 December 2008; published online 8 February 2009; doi:10.1038/ng.314

Table 1 Association of the haplotypes derived from the rs2048327, rs3127599, rs7767084 and rs10755578 with CAD in six independent studies

Polymorphisms				Haplotype frequencies					
				WTCCC		GerMIFS I		GerMIFS II	
rs2048327 (*SLC22A3*)	rs3127599 (*LPAL2*)	rs7767084 (*LPA*)	rs10755578 (*LPA*)	Controls (n = 2,938)	Cases (n = 1,926)	Controls (n = 1,644)	Cases (n = 875)	Controls (n = 1,298)	Cases (n = 1,222)
T	C	T	C	0.479	0.453	0.496	0.469	0.487	0.474
T	T	T	C	0.021	0.016	0.012	0.018	0.015	0.010
T	T	T	G	0.107	0.102	0.137	0.100	0.122	0.118
C	C	T	C	0.021	0.041	0.018	0.034	0.017	0.025
C	C	T	G	0.031	0.023	0.025	0.025	0.043	0.033
C	C	C	G	0.148	0.149	0.151	0.167	0.144	0.163
C	T	T	G	0.178	0.203	0.142	0.172	0.152	0.163
Test of haplotypic association				χ^2 = 45.16 with 6 d.f. $P = 4.34 \times 10^{-8}$		χ^2 = 29.63 with 6 d.f. $P = 4.62 \times 10^{-5}$		χ^2 = 13.31 with 6 d.f. P = 0.038	

Polymorphisms				Haplotype frequencies					
				Angio-Lueb		PopGen		ECTIM	
rs2048327 (*SLC22A3*)	rs3127599 (*LPAL2*)	rs7767084 (*LPA*)	rs10755578 (*LPA*)	Controls (n = 1,524)	Cases (n = 1,679)	Controls (n = 1,809)	Cases (n = 2,199)	Controls (n = 1,050)	Cases (n = 1,098)
T	C	T	C	0.494	0.464	0.484	0.453	0.532	0.476
T	T	T	C	0.013	0.015	0.013	0.013	0.012	0.018
T	T	T	G	0.131	0.124	0.115	0.112	0.112	0.106
C	C	T	C	0.017	0.022	0.017	0.033	0.018	0.027
C	C	T	G	0.031	0.034	0.042	0.035	0.028	0.024
C	C	C	G	0.156	0.161	0.156	0.160	0.140	0.158
C	T	T	G	0.144	0.162	0.156	0.178	0.149	0.177
Test of haplotypic association				χ^2 = 8.41 with 6 d.f. P = 0.209		χ^2 = 30.24 with 6 d.f. $P = 3.53 \times 10^{-5}$		χ^2 = 19.44 with 6 d.f. P = 0.0035	

The WTCCC Study was the primary study where GWHA was done (stage 1). This was followed by replication in the GerMIFS I GWA data (stage 2) and follow-up replication in the other four studies (stage 3).

CAD cases and 1,298 controls (GerMIFS II)[4,5], 1,679 CAD cases and 1,524 controls (Angio-Luebeck study), 2,199 CAD cases and 1,809 controls (PopGen)[6] and 1,098 myocardial infarction cases and 1,050 controls (ECTIM)[7] (**Supplementary Methods**). A population-adjusted combined analysis of these four studies provided strong support ($P = 1.23 \times 10^{-9}$) for haplotype association at the 6q26–q27 locus.

In the six studies, the same seven common (frequency >0.01) haplotypes were inferred from the four SNPs (**Table 1**). In a combined individual-level analysis of the six datasets totaling 8,999 cases and 10,263 controls, the overall evidence for global haplotypic association adjusted for study was strong ($P = 1.34 \times 10^{-21}$). The association was mainly attributable to two haplotypes (CTTG and CCTC) whose frequency was increased in cases compared to controls (**Fig. 1**). The CTTG haplotype was associated with a combined odds ratio (OR) of 1.202 (95% CI = 1.133–1.275, $P = 1.19 \times 10^{-9}$) when compared to the most frequent TCTC haplotype, with no evidence for heterogeneity across the six studies (P = 0.826). The less frequent CCTC haplotype was associated with a combined OR of 1.822 (95% CI = 1.568–2.116, $P = 4.20 \times 10^{-15}$, P for heterogeneity = 0.532; **Fig. 1**). These two haplotypes both carry the rs2048327[C] allele, which was also strongly associated with an increased risk in the combined samples ($P = 3.98 \times 10^{-13}$, **Supplementary Table 2** online). We therefore tested whether the association with CAD in all datasets combined might be compatible with the sole effect of the rs20848327[C] allele, but this hypothesis was rejected ($P = 3.14 \times 10^{-11}$, **Supplementary Table 2**). We also examined whether the association observed could be explained by untyped SNP(s) located close to those that had been genotyped. Genotypes for up to 2.5 million SNPs have been imputed from the WTCCC data[8]. No single imputed SNP at this locus showed stronger evidence of association with CAD than the haplotypes identified by the sliding-window approach (**Supplementary Fig. 2** online).

The identified locus partly overlaps the *LPA* gene, which encodes apolipoprotein(a), the main protein of lipoprotein(a) (Lp(a)), a known risk factor for CAD[9,10] whose plasma measurements were available in the ECTIM study[9]. We therefore examined the association of the identified haplotypes with Lp(a) levels in ECTIM. The haplotypes associated with CAD were those associated with the highest Lp(a) levels, and after adjustment for Lp(a) levels, these were no longer associated with myocardial infarction (P = 0.102), suggesting that their relation to risk is mediated by an effect on Lp(a). Lp(a) levels are strongly heritable and the *LPA* locus accounts for almost the whole heritability[11]. A kringle-repeat polymorphism of the *LPA* gene is known to be the key determinant of Lp(a) levels. In ECTIM, 34% of Lp(a) variability was attributable to this repeat polymorphism, but an additional fraction of 15% ($P = 6.02 \times 10^{-20}$) was explained by the haplotypes combining the four SNPs of the present analysis and a *LPA*

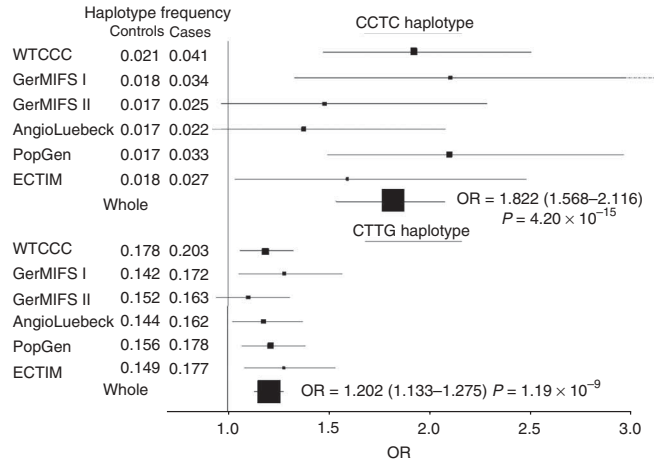

Figure 1 Haplotypic odds ratios (ORs) for CAD associated with the CCTC and CTTG haplotypes derived from rs2048327, rs3127599, rs7767084 and rs10755578 on chromosome 6q26–q27. Squares, with size proportional to the size of the study sample, represent ORs with their 95% confidence interval. These are haplotypic ORs by comparison to the most frequent TCTC haplotype, under the assumption of haplotypic additive effects. We estimated ORs from a combined analysis of all individual datasets while adjusting for study and after having checked for the homogeneity using the Mantel-Haenszel method.

promoter SNP, rs1800769, which was genotyped in ECTIM[9] but not characterized by HapMap data and not available in the studies analyzed using the DNA arrays (**Supplementary Methods** and **Supplementary Fig. 2**). Notably, the rs1800769 SNP splits the CTTG haplotype into two subhaplotypes, and only one of them is associated with increased Lp(a) levels (**Supplementary Fig. 3** online) and increased risk of CAD (**Supplementary Table 3** online).

To our knowledge, this work is the first to report results from a genome-wide haplotype analysis with replication of the findings. A recent study[12] carried out a haplotype analysis of the WTCCC data for seven diseases investigated in that project but did not detect any new CAD susceptibility locus. The statistical methodology used by these authors for detecting haplotype effects was different from ours in several aspects; in particular, they used a lower number of investigated haplotypic configurations (1.5 million versus 8.1 million in our study) and a different statistical methodology for testing association (**Supplementary Methods**). Moreover, these investigators did not include any replication study.

The genome-wide haplotype strategy identified the chromosome 6q26–q27 locus as a strong susceptibility locus for CAD. This locus encompasses the *SLC22A3*, *LPAL2* and *LPA* genes, the latter having a long history of suggestive association with CAD[13]. Further work is necessary to determine whether the detected haplotype effects are the reflection of interactions between SNPs at the haplotypic level or whether they are tagging ungenotyped functional variants. The *LPA* gene shows a highly complex pattern of variability, with several copy number variations (CNV) polymorphisms not well characterized by the DNA arrays used. This might explain why this locus was not suspected from our previous GWA studies[1] based on single-SNP analyses and was not detected through imputed data analysis. As one of the detected haplotypes was relatively uncommon and associated with strong genetic effect, our work raises the possibility that haplotypes based on common SNPs may identify relatively rare variants of clinical interest.

Both the completeness and the fidelity of the genotyping have a more significant impact on GWHA than on single-SNP GWA analysis, especially when comparisons across studies is planned. This is highlighted by the number of loci that we were unable to compare between the WTCCC and GerMIFS I GWA data. Furthermore, the much larger number of association tests undertaken, which are not all independent, means that conventional thresholds used in single-SNP GWA analysis to indicate genome-wide significance cannot be easily translated to GWHA and a robust replication strategy is necessary, regardless of what thresholds are used to identify regions for follow-up. Because of these issues, we cannot exclude the possibility that other haplotype-based loci associated with risk of CAD were missed in the present analysis. Nonetheless, our results illustrate the utility of GWHA analysis of genotype data obtained on currently available GWA platforms to identify new susceptibility loci for complex human diseases.

Note: Supplementary information is available on the Nature Genetics website.

ACKNOWLEDGMENTS

The WTCCC Study was funded by the Wellcome Trust. Recruitment of cases for the WTCCC Study was carried out by the British Heart Foundation (BHF) Family Heart Study Research Group and supported by the BHF and the UK Medical Research Council. We also acknowledge support of the Wellcome Trust Functional Genomics Initiative in Cardiovascular Genetics. The German Study was supported by the Deutsche Forschungsgemeinschaft, the German Federal Ministry of Education and Research in the context of the German National Genome Research Network (NGFN-2 and NGFN-plus). The KORA research platform (KORA, Cooperative Research in the Region of Augsburg) was initiated and financed by the GSF-National Research Centre for Environment and Health, which is funded by the German Federal Ministry of Education and Research and of the State of Bavaria. This work has also been partially supported by the EGEE-II project funded by the European Union INFSO-RI-031688. N.J.S. is supported by a Chair funded by the BHF. The main sponsor of the current analysis is the EU funded integrated project 'Cardiogenics' (LSHM-CT-2006-037593).

AUTHOR CONTRIBUTIONS

This work was designed by D.-A.T., L.T., F.C., H.S. and N.J.S. Statistical analyses were carried out by D.-A.T., I.R.K., A.G., M.P., P.H., A.Z. and J.R.T., J.E., P.S.B., C.P., M.D., P.B. and W.O. supervised and/or participated in the genetic laboratory analyses. A.M. developed the computing grid program under the supervision of D.-A.T. and C.G. Case collections were coordinated by P.S.B., A.S.H., B.J.W., A.J.B., S.G.B., W.O., J.R.T., N.J.S. (WTCCC), J.E., T.M., H.-E.W., C. Meisinger, C.H., K.N., K.S., J.S., D.R. (GerMIFS), S.S., A.S., N.E.E.-M. (PopGene), J.E., P.L.-N. (Angio-Luebeck), A.E., D.A., G.L., J.-B.R., and C. Morrison (ECTIM). D.A.T., L.T., F.C., H.S. and N.J.S. drafted the manuscript with substantial contributions from I.R.K., J.E., A.Z.

Published online at http://www.nature.com/naturegenetics/
Reprints and permissions information is available online at http://npg.nature.com/reprintsandpermissions/

1. Samani, N.J. *et al. N. Engl. J. Med.* **357**, 443–453 (2007).
2. The Wellcome Trust Case Control Consortium. *Nature* **447**, 661–678 (2007).
3. Gagliardi, F. *et al. Phil. Transact. A. Math. Phys. Eng. Sci.* **363**, 1729–1742 (2005).
4. Fischer, M. *et al. Circulation* **111**, 855–862 (2005).
5. Wichmann, H.E., Gieger, C. & Illig, T. *Gesundheitswesen* **67** (Suppl. 1), S26–S30 (2005).
6. Krawczak, M. *et al. Community Genet.* **9**, 55–66 (2006).
7. Parra, H.J. *et al. Arterioscler. Thromb.* **12**, 701–707 (1992).
8. Li, Y. & Abecasis, G.R. *Am. J. Hum. Genet.* **S79**, 2290 (2006).
9. Brazier, L. *et al. Atherosclerosis* **144**, 323–333 (1999).
10. Holmer, S.R. *et al. Circulation* **107**, 696–701 (2003).
11. Boerwinkle, E. *et al. J. Clin. Invest.* **90**, 52–60 (1992).
12. Browning, B.L. & Browning, S.R. *Hum. Genet.* **123**, 273–280 (2008).
13. Schaefer, E.J. *et al. J. Am. Med. Assoc.* **271**, 999–1003 (1994).