Hum Genet (2010) 128:627–633
DOI 10.1007/s00439-010-0889-1

ORIGINAL INVESTIGATION

Rare variation at the *TNFAIP3* locus and susceptibility to rheumatoid arthritis

John Bowes · Robert Lawrence · Stephen Eyre · Kalliope Panoutsopoulou · Gisela Orozco · Katherine S. Elliott · Xiayi Ke · Andrew P. Morris · UKRAG · Wendy Thomson · Jane Worthington · Anne Barton · Eleftheria Zeggini

Received: 23 July 2010 / Accepted: 7 September 2010 / Published online: 18 September 2010

Abstract Genome-wide association studies (GWAS) conducted using commercial single nucleotide polymorphisms (SNP) arrays have proven to be a powerful tool for the detection of common disease susceptibility variants. However, their utility for the detection of lower frequency variants is yet to be practically investigated. Here we describe the application of a rare variant collapsing method to a large genome-wide SNP dataset, the Wellcome Trust Case Control Consortium rheumatoid arthritis (RA) GWAS. We partitioned the data into gene-centric bins and collapsed genotypes of low frequency variants (defined here as MAF $\leq$0.05) into a single count coupled with univariate analysis. We then prioritised gene regions for further investigation in an independent cohort of 3,355 cases and 2,427 controls based on rare variant signal *p* value and prior evidence to support involvement in RA. A total of 14,536 gene bins were investigated in the primary analysis and signals mapping to the *TNFAIP3* and chr17q24 loci were selected for further investigation. We detected replicating association to low frequency variants in the *TNFAIP3* gene (combined $p = 6.6 \times 10^{-6}$). Even though rare variants are not well-represented and can be difficult to genotype in GWAS, our study supports the application of low frequency variant collapsing methods to genome-wide SNP datasets as a means of exploiting data that are routinely ignored.

Introduction

Genetic epidemiology of common disease has been transformed by the recent advent of the genome-wide association study (GWAS) that allows the unbiased testing of hundreds of thousands of single nucleotide polymorphisms (SNPs) in large sample collections. This approach has greatly accelerated the identification of genetic markers associated with disease susceptibility and continuous traits. The Catalogue of Published Genome-Wide Association Studies contains details from 378 published GWAS reporting 1,700 SNP associations to diseases or quantitative traits (with $p < 10^{-5}$) to date (Hindorff et al. 2009).

While the GWAS strategy has undoubtedly been successful in identifying a multitude of disease-associated makers, these have been shown to explain very little of the estimated heritable component of complex disease (Maher 2008). For example, all of the currently known rheumatoid arthritis (RA) risk alleles, including the major histocompatibility complex (MHC) locus, have been estimated to explain less than half of the total genetic component (Raychaudhuri et al. 2008), indicating that a substantial proportion of susceptibility variants remain undiscovered.

The deficit between estimated heritability of a particular trait and its known genetic variation may be attributed to a

Electronic supplementary material The online version of this article (doi:10.1007/s00439-010-0889-1) contains supplementary material, which is available to authorized users.

The members of the UKRAG are given in Appendix.

J. Bowes · S. Eyre · G. Orozco · X. Ke · W. Thomson · J. Worthington · A. Barton
Arthritis Research UK, Epidemiology Unit, University of Manchester, Manchester, UK

R. Lawrence · K. S. Elliott · A. P. Morris
Wellcome Trust Centre for Human Genetics, University of Oxford, Oxford, UK

R. Lawrence · K. Panoutsopoulou · E. Zeggini (✉)
Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1SA, UK
e-mail: Eleftheria@sanger.ac.uk

number of issues. Firstly, the SNPs identified in the primary scans are likely to serve only as proxies for the true causal variant, therefore estimates of their contribution to genetic risk may be greatly underestimated (McCarthy and Hirschhorn 2008). Secondly, further variants at an associated locus may confer additional effects independent of the primary signal (Plenge et al. 2007). In addition, structural variation, such as insertions, deletions, and duplications, is now recognised as a well-established feature of the genome (Iafrate et al. 2004; Redon et al. 2006; Sebat et al. 2004), with a number of emerging studies reporting associations to disease susceptibility (Fanciulli et al. 2007; Fellermann et al. 2006; Gonzalez et al. 2005; Hollox et al. 2008; McKinney et al. 2008). Importantly, a significant proportion of inherited susceptibility to common disease may be attributable to the effects of multiple, low frequency variants with moderate effects in a number of genes (Bodmer and Bonilla 2008). The majority of variants in the human genome are predicted to have minor allele frequency (MAF) below 0.05 and are correlated with increased functional significance (Gorlov et al. 2008). Evidence to support the involvement of rare variants in the susceptibility to common disease comes from both theoretical (Pritchard 2001; Pritchard and Cox 2002) and empirical data from studies such as colorectal cancer (Fearnhead et al. 2004), HDL cholesterol levels (Cohen et al. 2004), hypertension (Ji et al. 2008), obesity (Ahituv et al. 2007) and type 1 diabetes (Nejentsev et al. 2009).

The design of currently available commercial SNP platforms has been heavily biased towards common SNPs. GWAS rely on indirect association methods exploiting local linkage disequilibrium (LD) patterns. Studies performed on large empirical and simulated datasets have shown that LD-tagging based on HapMap data is robust for common SNPs, but that performance declines rapidly for lower frequency variants when common tags are used (Ahmadi et al. 2005; Zeggini et al. 2005). In addition, rare variants are notoriously difficult to genotype accurately, and single-point association analysis of rare variants has inherently low power given currently available sample sizes. Indeed, the majority of GWAS exclude low frequency SNPs from further analysis, thus ignoring any information afforded by the rare variants included in the arrays. The issue of power to detect association is being addressed by the development of statistical methods for the combined analysis of rare variants (Li and Leal 2008; Morris and Zeggini 2010).

Here we demonstrate, with a proof of principle experiment, that the direct analysis of rare variants (defined here as MAF $\leq$0.05) present on commercially available arrays can indeed lead to the identification of real complex disease susceptibility signals. We have analysed all rare variants on the Affymetrix 500K array for the RA branch of the Wellcome Trust Case Control Consortium (2007) and have followed up four signals with strong priors for association with disease. Following extensive checks on genotype quality and after excluding poorly clustering variants, we find replicating association of rare variation at the *TNFAIP3* locus with RA susceptibility.

Materials and methods

Ethics statement

This study was approved by the North West Multicentre Research Ethics Committee (MREC 99/8/84), and all subjects provided informed consent.

Primary scan

Samples, genotyping and quality control

We used data from the RA branch of the WTCCC study (The Wellcome Trust Case Control Consortium 2007), genotyped on the Affymetrix GeneChip Human Mapping 500K Array Set. The dataset consisted of 1,860 RA cases and 2,938 controls, all from the UK, after the exclusion of samples failing quality control (QC) (The Wellcome Trust Case Control Consortium 2007). We excluded from further analysis SNPs failing QC thresholds as described in the WTCCC study (The Wellcome Trust Case Control Consortium 2007). Briefly, SNPs were excluded based on deviation from Hardy–Weinberg equilibrium ($p < 5.7 \times 10^{-7}$) or if the study-wise missing data proportion was >0.05. Additionally, SNPs with a MAF <0.05 were excluded if the study-wise missing data proportion exceeded 0.01. Genotype cluster plot inspection was not possible at the genome-wide scale, so we inspected the clustering properties of SNPs within signals of interest after the rare variant analysis had been carried out.

Rare variant analysis

In order to maximise power to detect association, we followed a "super-locus" approach to analysing rare variation in the GWAS RA data (Li and Leal 2008; Morris and Zeggini 2010). SNPs with a study-wise MAF <0.05 were included in the analysis. We defined genic regions genome-wide based on the coordinates of known genes and (arbitrarily) included 50 kilobases (kb) flanking either side of each gene's transcriptional start and stop site in an attempt to include SNPs affecting localised regulatory elements. For each such region, we collapsed the rare variant allele counts by assigning each individual a carrier/non-carrier status based on the presence or absence of rare variant

minor alleles (Supplementary Figure 1) (Li and Leal 2008; Morris and Zeggini 2010). This classification system is not affected by LD, as the number of rare alleles carried by each individual (the rare allele load) was not taken into account. This allowed the construction of a simple 2 × 2 contingency table, which we used to test differences in rare variant minor allele carriage between cases and controls using the Chi-squared test (or Fisher's exact test where necessary). For regions attaining $p < 10^{-4}$, we also permuted case/control status 100,000 times in order to assess significance by means of empirical p values. We used CCRaVAT software to carry out these analyses (http://www.sanger.ac.uk/resources/software/ccravat-qutie/). For the purposes of this study, variants with a MAF ≤0.05 were considered low frequency/rare and included in the analysis. Although this value is arbitrary, it represents a typical threshold at which SNPs are excluded from further consideration in many GWAS.

Signal prioritisation and quality control

Signals were selected for replication based on the primary rare variant association scan p value and the existence of prior evidence to suggest involvement in the susceptibility to RA. Bins containing only a single rare variant were excluded from further investigation, as were bins mapping to the extended MHC region. For the purposes of this study, the extended MHC was considered to be a 7.6 Mb region consisting of the extended class I, classical class I, extended class II, classical class II, and classical class III subregions. The region is defined by the co-ordinates Chr6:25,809,997-33,486,772 (NCBI build 36). As low frequency/rare variants are notoriously difficult to accurately genotype, we examined SNP clustering properties for all variants contributing to the prioritised signals. To investigate the influence of poor quality genotyping on the performance of the super-locus method we excluded any SNPs considered to be of low quality, based on clustering, from the gene region under study and re-analysed.

Replication study

Follow-up patient samples

DNA samples for 3,838 RA cases and 2,719 healthy unrelated controls were made available from the UK Rheumatoid Arthritis Genetics (UKRAG) consortium, a collaboration of six UK rheumatology research groups.

Genotyping and quality control

Low frequency/rare variants contributing to prioritised association signals in the primary scan were selected for follow-up. Genotyping was performed using the Sequenom MassARRAY platform with 10 ng of genomic DNA in concordance with the manufacturer's specifications for the iPLEX protocol. SNPs rs5029939 and rs7749323 had already been typed in a subset of the follow-up patient samples as part of a related study (Orozco et al. 2009). All Sequenom genotype cluster plots were manually evaluated to determine satisfactory resolution of clusters. Samples with a call rate <0.95 and SNPs with a call rate <0.90 were excluded prior to analysis. Allele counts for all SNPs mapping to a gene were combined and then compared between cases and controls using a Fisher's exact test.

Meta-analysis

We combined individual summary statistics across the WTCCC RA and UK replication studies using a fixed effects inverse-variance meta-analysis approach for each genic region. We also investigated the presence of significant heterogeneity across studies, by means of the I^2 and Q statistics.

Results

Primary scan

Following QC procedures, the GWAS consisted of 1,860 RA cases and 2,938 healthy controls genotyped for 459,446 SNPs. This dataset included 40,482 SNPs with MAF ≤0.05 suitable for inclusion in the current analysis. Application of the super-locus method yielded a total of 22,344 genomic bins, of which 14,536 contained at least two rare variants and mapped outside of the defined MHC region ($\lambda GC = 1.09$, Supplementary Figure 2). We identified 25 gene regions with a $p < 10^{-4}$. For the purposes of this study, we prioritised four signals for follow-up genotyping that emerged from two distinct chromosomal regions. This selection was based on their rare variant analysis p value and prior evidence to support disease association. Firstly, we selected SNPs from three gene regions mapping to chromosome 17q24; *ACE*, *CYB561*, and *PRKCA* ($p = 6.0 \times 10^{-5}$, $p = 2.7 \times 10^{-5}$ and $p = 0.0012$ respectively). This region is implicated in susceptibility to RA by previous evidence from human and animal linkage studies and is of particular interest to our research group (Backdahl et al. 2008; Barton et al. 2001; Jawaheer et al. 2001). To date, no common susceptibility variants that could explain the observed linkage signal have been discovered in this region, and this could be attributable to the presence of associated rare variants. The second candidate locus was the *TNFAIP3* gene ($p = 4.2 \times 10^{-5}$), a known RA-associated locus with evidence for allelic heterogeneity (Orozco et al. 2009; Plenge et al. 2007; Thomson

et al. 2007). These four bins comprise a total of 27 unique SNPs (due to overlap the *ACE* and *CYB561* gene regions share two SNPs). Individual SNP genotype data and association results are reported in Supplementary Table 1. For the purposes of evaluating data quality we examined the clustering properties of rare variants residing in the associated gene regions and identified six SNPs with unsatisfactory clustering. Re-analysis of the data after excluding the badly clustering variants did not alter the results for the *TNFAIP3* and *PRKCA* loci qualitatively, but did affect significance at the other two genes (Table 1; Supplementary Table 2).

Replication study

We attempted to genotype all 27 SNPs from the four bins, regardless of clustering results, in an independent dataset consisting of 3,355 cases and 2,427 controls after QC filtering. Three SNPs failed assay multiplexing and 24 SNPs were successfully genotyped. SNPs were partitioned into appropriate gene regions and proportions of carriers and non-carriers were compared between the case and control groups. Evidence to support rare variant association with RA in the *TNFAIP3* gene was further substantiated ($p = 0.03$) (Table 1). No evidence was found to support association in the three genes mapping to the chr17 candidate region.

Meta-analysis

Meta-analysis across the discovery and replication datasets revealed stronger evidence for association of low frequency variants in *TNFAIP3* with RA (combined $p = 6.6 \times 10^{-6}$; OR 1.27[1.14:1.40]) (Table 1). The direction of association indicated an over-representation of rare variant minor alleles in cases compared to controls.

Discussion

To date GWAS have concentrated on the investigation of common variation therefore discarding a significant proportion of data already collected, since a common QC criterion is the exclusion of SNPs with low MAF. The evidence presented here highlights the potential importance of investigating low frequency variant data from commercial SNP arrays. It should be noted that the primary purpose of these analyses is signal detection and not accurate effect estimation. The application of a rare variant analysis method (Li and Leal 2008; Morris and Zeggini 2010), which accumulates information from multiple rare SNP alleles within a genic region, identified an association between RA and the *TNFAIP3* gene.

Multiple independent RA associations have previously been reported at the *TNFAIP3* locus. The most convincing evidence for association to rare variants in this region comes from the meta-analysis described here ($p = 6.6 \times 10^{-6}$). Association to the 6q23 region was originally identified for common variants by two independent GWAS (Plenge et al. 2007; The Wellcome Trust Case Control Consortium 2007). The WTCCC RA GWAS identified a common RA risk variant >180 kb away from the *TNFAIP3* gene, which was replicated in a UK study (Thomson et al. 2007). This finding prompted a fine mapping experiment covering the intergenic region containing the index association and the *TNFAIP3* gene across ~7,000 individuals, identifying an independent association at rs5029937 (MAF ~0.04) in intron 2 of *TNFAIP3* (Orozco et al. 2009). The evidence for multiple independent effects in the *TNFAIP3* gene region make this a compelling candidate for re-sequencing.

Application of the super-locus approach using the original WTCCC data alone (2,000 cases, 3,000 controls) prior to the fine-mapping project would have directed attention to this signal obviating the need for time-consuming and costly region-wide fine-mapping. LD between the low frequency variants in the *TNFAIP3* gene region reported here and the intergenic SNPs, previously shown to be associated with RA, is minimal ($r^2 < 0.04$). However, there is strong LD between two SNPs from the super-locus analysed here and rs5029937 (rs7749323 $r^2 = 0.91$ and rs5029939 $r^2 = 0.97$). Our results highlight the potential strength of analysing lower frequency/rare variants in a

Table 1 Results for the four rare variant gene bins

Locus	WTCCC			Replication			Meta-analysis		
	n SNPs	OR	*p* value	*n* SNPs	OR	*p* value	OR	*p* value	I^2
ACE	2	1.27 (0.59:2.71)	0.54	2	2.00 (0.88:4.52)	0.09	1.59 (0.92:2.75)	0.098	0
PRKCA	11	1.27 (1.09:1.48)	1.54E−03	10	1.01 (0.88:1.16)	0.85	1.12 (1.02:1.24)	0.023	79.1
CYB561	2	1.59 (0.69:3.67)	0.28	2	1.65 (0.83:3.28)	0.15	1.63 (0.96:2.76)	0.072	0
TNFAIP3	7	1.39 (1.20:1.62)	1.24E−05	6	1.16 (1.01:1.34)	0.03	1.27 (1.14:1.40)	6.60E−06	66

OR odds ratio (95% confidence intervals in parentheses); *n* number

Results presented here are following exclusion of the six SNPs identified as having poor clustering properties

complementary approach to the primary GWAS of common variants.

The study of low frequency/rare variants from commercial SNP arrays is limited by a number of factors. Firstly, the arrays contain only a small proportion of the low MAF variants that actually exist; this is due to the array design being motivated by the detection of common susceptibility variants. The gold standard strategy for investigating lower frequency/rare variants will be the use of sequencing data, capturing a much higher proportion of genetic variation (Morris and Zeggini 2010). However, in spite of the rapid advances in sequencing technology, this approach is often limited by economic constraints when considering large sample sizes. The method applied here utilises existing data that are generally discarded in GWAS. It therefore increases the amount of available data that are analysed and comes at no additional cost to the researcher. Collapsing methods are also very flexible in their application. In the study presented here we have focused on gene-centric bins, but this could be expanded to include bins based on coordinates of known conserved non-coding regions or functional pathways as the functional unit for analysis.

Secondly, the genotype quality of rare variants typed on GWAS platforms tends to be low, mainly driven by poor automated clustering and genotype calling. This limitation affects both single-point and collapsing methods and necessitates stringent and exhaustive inspection of cluster plots in association signals of interest. The potential for confounding is demonstrated in our study by the effect of excluding poorly clustering SNPs in the *ACE* and *CYB561* gene regions. These two bins were no longer statistically significant upon removal of poor quality SNPs, thus highlighting the potential for inflated false positive rates in the absence of stringent quality assurance when dealing with low frequency SNP data. In addition, it is highly likely that bins containing only a few SNPs have an increased sensitivity to such artefacts. Genotype calling algorithms fine-tuned for rare variant clustering have recently been developed and hold the promise of helping to overcome many of these issues.

Thirdly, collapsing methods can be limited by the properties of the component variants. For example, they have been shown to be sensitive to the inclusion of neutral alleles, whose presence reduces power. Ideally, each of the bins would only contain putatively functional variants, which could be achieved by the use of bioinformatics tools. However, the accuracy of such tools is currently limited and the reduction of power is greater when true functional SNPs are excluded compared to the inclusion of functionally neutral SNPs (Li and Leal 2008). Additionally, it is conceivable that bins could contain variants with diverse effects, where the presence of both protective and risk alleles would confound the analysis. Rare variant collapsing methods are being further developed and extended to account for direction of effect, probability that a variant be functional, and genotype or sequence call uncertainty.

Finally, the phenotype under investigation could strongly dictate the success of identifying novel, rare disease-related variants. It could be argued that late-onset diseases, such as RA, are less likely to be influenced by a contribution from rare variants. However, the underlying genetic architecture of the disease is yet to be fully elucidated; collapsing methods represent one strategy that can help address this question.

Despite these limitations, our results illustrate that collapsing methods represent a strategy with the potential for identifying novel variants at no additional cost to the researcher, thus maximising return from costly primary GWAS. Applying this approach to existing SNP array data would have expedited the identification of the *TNFAIP3* intron 2 SNP association with RA. We advocate the application of this method in conjunction with primary GWAS for complex diseases, but also suggest extreme caution with the interpretation of promising signals, which need to be scrutinised for genotype quality. Additionally, collapsing methods hold great promise for the future with the application to sequencing data.

Acknowledgments The authors are grateful to Edward Flynn and Paul Martin for Sequenom SNP genotyping. We are also grateful for the support of the Manchester Biomedical Research Centre and the Manchester Academy of Health Sciences.

Appendix

United Kingdom Rheumatoid Arthritis Genetics (UKRAG). School of Medicine and Biomedical Sciences, Sheffield University, Sheffield S10 2JF (Dr Anthony G Wilson); NIHR-Leeds Musculoskeletal Biomedical Research Unit, Leeds Institute of Molecular Medicine, University of Leeds, UK (Prof. Ann W Morgan, Prof. Paul Emery); Clinical and Academic Rheumatology, Kings College Hospital NHS Foundation Trust, Denmark Hill, London SE5 9RS (Dr. Sophia Steer); Musculoskeletal and Genetics Section, Division of Applied Medicine, University of Aberdeen, UK, AB25 2ZD (Dr Lynne J Hocking, Dr David M Reid); University of Oxford Institute of Musculoskeletal Sciences, Botnar Research Centre, Oxford OX3 7LD, UK (Dr Pille Harrison, Professor Paul Wordsworth).

References

Ahituv N, Kavaslar N, Schackwitz W, Ustaszewska A, Martin J, Hebert S, Doelle H, Ersoy B, Kryukov G, Schmidt S, Yosef N, Ruppin E, Sharan R, Vaisse C, Sunyaev S, Dent R, Cohen J, McPherson R, Pennacchio LA (2007) Medical sequencing at the extremes of human body mass. Am J Hum Genet 80:779–791

Ahmadi KR, Weale ME, Xue ZY, Soranzo N, Yarnall DP, Briley JD, Maruyama Y, Kobayashi M, Wood NW, Spurr NK, Burns DK, Roses AD, Saunders AM, Goldstein DB (2005) A single-nucleotide polymorphism tagging set for human drug metabolism and transport. Nat Genet 37:84–89

Backdahl L, Guo JP, Jagodic M, Becanovic K, Ding B, Olsson T, Lorentzen JC (2008) Definition of arthritis candidate risk genes by combining rat linkage-mapping results with human case control association data. Ann Rheum Dis 67:1742–1749

Barton A, Eyre S, Myerscough A, Brintnell B, Ward D, Ollier WE, Lorentzen JC, Klareskog L, Silman A, John S, Worthington J (2001) High resolution linkage and association mapping identifies a novel rheumatoid arthritis susceptibility locus homologous to one linked to two rat models of inflammatory arthritis. Hum Mol Genet 10:1901–1906

Bodmer W, Bonilla C (2008) Common and rare variants in multifactorial susceptibility to common diseases. Nat Genet 40:695–701

Cohen JC, Kiss RS, Pertsemlidis A, Marcel YL, McPherson R, Hobbs HH (2004) Multiple rare alleles contribute to low plasma levels of HDL cholesterol. Science 305:869–872

Fanciulli M, Norsworthy PJ, Petretto E, Dong R, Harper L, Kamesh L, Heward JM, Gough SC, De SA, Blakemore AI, Froguel P, Owen CJ, Pearce SH, Teixeira L, Guillevin L, Graham DS, Pusey CD, Cook HT, Vyse TJ, Aitman TJ (2007) FCGR3B copy number variation is associated with susceptibility to systemic, but not organ-specific, autoimmunity. Nat Genet 39:721–723

Fearnhead NS, Wilding JL, Winney B, Tonks S, Bartlett S, Bicknell DC, Tomlinson IP, Mortensen NJ, Bodmer WF (2004) Multiple rare variants in different genes account for multifactorial inherited susceptibility to colorectal adenomas. Proc Natl Acad Sci USA 101:15992–15997

Fellermann K, Stange DE, Schaeffeler E, Schmalzl H, Wehkamp J, Bevins CL, Reinisch W, Teml A, Schwab M, Lichter P, Radlwimmer B, Stange EF (2006) A chromosome 8 gene-cluster polymorphism with low human beta-defensin 2 gene copy number predisposes to Crohn disease of the colon. Am J Hum Genet 79:439–448

Gonzalez E, Kulkarni H, Bolivar H, Mangano A, Sanchez R, Catano G, Nibbs RJ, Freedman BI, Quinones MP, Bamshad MJ, Murthy KK, Rovin BH, Bradley W, Clark RA, Anderson SA, O'connell RJ, Agan BK, Ahuja SS, Bologna R, Sen L, Dolan MJ, Ahuja SK (2005) The influence of CCL3L1 gene-containing segmental duplications on HIV-1/AIDS susceptibility. Science 307:1434–1440

Gorlov IP, Gorlova OY, Sunyaev SR, Spitz MR, Amos CI (2008) Shifting paradigm of association studies: value of rare single-nucleotide polymorphisms. Am J Hum Genet 82:100–112

Hindorff LA, Sethupathy P, Junkins HA, Ramos EM, Mehta JP, Collins FS, Manolio TA (2009) Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. Proc Natl Acad Sci USA 106:9362–9367

Hollox EJ, Huffmeier U, Zeeuwen PL, Palla R, Lascorz J, Rodijk-Olthuis D, van de Kerkhof PC, Traupe H, De JG, den HM, Reis A, Armour JA, Schalkwijk J (2008) Psoriasis is associated with increased beta-defensin genomic copy number. Nat Genet 40:23–25

Iafrate AJ, Feuk L, Rivera MN, Listewnik ML, Donahoe PK, Qi Y, Scherer SW, Lee C (2004) Detection of large-scale variation in the human genome. Nat Genet 36:949–951

Jawaheer D, Seldin MF, Amos CI, Chen WV, Shigeta R, Monteiro J, Kern M, Criswell LA, Albani S, Nelson JL, Clegg DO, Pope R, Schroeder HW Jr, Bridges SL Jr, Pisetsky DS, Ward R, Kastner DL, Wilder RL, Pincus T, Callahan LF, Flemming D, Wener MH, Gregersen PK (2001) A genomewide screen in multiplex rheumatoid arthritis families suggests genetic overlap with other autoimmune diseases. Am J Hum Genet 68:927–936

Ji W, Foo JN, O'Roak BJ, Zhao H, Larson MG, Simon DB, Newton-Cheh C, State MW, Levy D, Lifton RP (2008) Rare independent mutations in renal salt handling genes contribute to blood pressure variation. Nat Genet 40:592–599

Li B, Leal SM (2008) Methods for detecting associations with rare variants for common diseases: application to analysis of sequence data. Am J Hum Genet 83:311–321

Maher B (2008) Personal genomes: the case of the missing heritability. Nature 456:18–21

McCarthy MI, Hirschhorn JN (2008) Genome-wide association studies: potential next steps on a genetic journey. Hum Mol Genet 17:R156–R165

McKinney C, Merriman ME, Chapman PT, Gow PJ, Harrison AA, Highton J, Jones PB, McLean L, O'Donnell JL, Pokorny V, Spellerberg M, Stamp LK, Willis J, Steer S, Merriman TR (2008) Evidence for an influence of chemokine ligand 3-like 1 (CCL3L1) gene copy number on susceptibility to rheumatoid arthritis. Ann Rheum Dis 67:409–413

Morris AP, Zeggini E (2010) An evaluation of statistical approaches to rare variant analysis in genetic association studies. Genet Epidemiol 34:188–193

Nejentsev S, Walker N, Riches D, Egholm M, Todd JA (2009) Rare variants of IFIH1, a gene implicated in antiviral responses, protect against type 1 diabetes. Science 324:387–389

Orozco G, Hinks A, Eyre S, Ke X, Gibbons LJ, Bowes J, Flynn E, Martin P, Wilson AG, Bax DE, Morgan AW, Emery P, Steer S, Hocking L, Reid DM, Wordsworth P, Harrison P, Thomson W, Barton A, Worthington J (2009) Combined effects of three independent SNPs greatly increase the risk estimate for RA at 6q23. Hum Mol Genet 18:2693–2699

Plenge RM, Cotsapas C, Davies L, Price AL, de Bakker PI, Maller J, Pe'er I, Burtt NP, Blumenstiel B, DeFelice M, Parkin M, Barry R, Winslow W, Healy C, Graham RR, Neale BM, Izmailova E, Roubenoff R, Parker AN, Glass R, Karlson EW, Maher N, Hafler DA, Lee DM, Seldin MF, Remmers EF, Lee AT, Padyukov L, Alfredsson L, Coblyn J, Weinblatt ME, Gabriel SB, Purcell S, Klareskog L, Gregersen PK, Shadick NA, Daly MJ, Altshuler D (2007) Two independent alleles at 6q23 associated with risk of rheumatoid arthritis. Nat Genet 39:1477–1482

Pritchard JK (2001) Are rare variants responsible for susceptibility to complex diseases? Am J Hum Genet 69:124–137

Pritchard JK, Cox NJ (2002) The allelic architecture of human disease genes: common disease-common variant or not? Hum Mol Genet 11:2417–2423

Raychaudhuri S, Remmers EF, Lee AT, Hackett R, Guiducci C, Burtt NP, Gianniny L, Korman BD, Padyukov L, Kurreeman FA, Chang M, Catanese JJ, Ding B, Wong S, Van der Helm-van Mil AH, Neale BM, Coblyn J, Cui J, Tak PP, Wolbink GJ, Crusius JB, van der Horst-Bruinsma IE, Criswell LA, Amos CI, Seldin MF, Kastner DL, Ardlie KG, Alfredsson L, Costenbader KH, Altshuler D, Huizinga TW, Shadick NA, Weinblatt ME, de Vries N, Worthington J, Seielstad M, Toes RE, Karlson EW, Begovich AB, Klareskog L, Gregersen PK, Daly MJ, Plenge RM (2008) Common variants at CD40 and other loci confer risk of rheumatoid arthritis. Nat Genet 40:1216–1223

Redon R, Ishikawa S, Fitch KR, Feuk L, Perry GH, Andrews TD, Fiegler H, Shapero MH, Carson AR, Chen W, Cho EK, Dallaire S, Freeman JL, Gonzalez JR, Gratacos M, Huang J, Kalaitzopoulos

D, Komura D, MacDonald JR, Marshall CR, Mei R, Montgomery L, Nishimura K, Okamura K, Shen F, Somerville MJ, Tchinda J, Valsesia A, Woodwark C, Yang F, Zhang J, Zerjal T, Zhang J, Armengol L, Conrad DF, Estivill X, Tyler-Smith C, Carter NP, Aburatani H, Lee C, Jones KW, Scherer SW, Hurles ME (2006) Global variation in copy number in the human genome. Nature 444:444–454

Sebat J, Lakshmi B, Troge J, Alexander J, Young J, Lundin P, Maner S, Massa H, Walker M, Chi M, Navin N, Lucito R, Healy J, Hicks J, Ye K, Reiner A, Gilliam TC, Trask B, Patterson N, Zetterberg A, Wigler M (2004) Large-scale copy number polymorphism in the human genome. Science 305:525–528

The Wellcome Trust Case Control Consortium (2007) Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. Nature 447:661–678

Thomson W, Barton A, Ke X, Eyre S, Hinks A, Bowes J, Donn R, Symmons D, Hider S, Bruce IN, Wilson AG, Marinou I, Morgan A, Emery P, Carter A, Steer S, Hocking L, Reid DM, Wordsworth P, Harrison P, Strachan D, Worthington J (2007) Rheumatoid arthritis association at 6q23. Nat Genet 39:1431–1433

Zeggini E, Rayner W, Morris AP, Hattersley AT, Walker M, Hitman GA, Deloukas P, Cardon LR, McCarthy MI (2005) An evaluation of HapMap sample size and tagging SNP performance in large-scale empirical and simulated data sets. Nat Genet 37:1320–1322