Human Molecular Genetics, 2013, Vol. 22, Review Issue 1 **R16–R21**
doi:10.1093/hmg/ddt376
Advance Access published on August 6, 2013

In search of low-frequency and rare variants affecting complex traits

Kalliope Panoutsopoulou, Ioanna Tachmazidou and Eleftheria Zeggini*

Wellcome Trust Sanger Institute, Hinxton, UK

Received July 3, 2013; Revised July 9, 2013; Accepted July 31, 2013

The allelic architecture of complex traits is likely to be underpinned by a combination of multiple common frequency and rare variants. Targeted genotyping arrays and next-generation sequencing technologies at the whole-genome sequencing (WGS) and whole-exome scales (WES) are increasingly employed to access sequence variation across the full minor allele frequency (MAF) spectrum. Different study design strategies that make use of diverse technologies, imputation and sample selection approaches are an active target of development and evaluation efforts. Initial insights into the contribution of rare variants in common diseases and medically relevant quantitative traits point to low-frequency and rare alleles acting either independently or in aggregate and in several cases alongside common variants. Studies conducted in population isolates have been successful in detecting rare variant associations with complex phenotypes. Statistical methodologies that enable the joint analysis of rare variants across regions of the genome continue to evolve with current efforts focusing on incorporating information such as functional annotation, and on the meta-analysis of these burden tests. In addition, population stratification, defining genome-wide statistical significance thresholds and the design of appropriate replication experiments constitute important considerations for the powerful analysis and interpretation of rare variant association studies. Progress in addressing these emerging challenges and the accrual of sufficiently large data sets are poised to help the field of complex trait genetics enter a promising era of discovery.

The genetic architecture of complex traits has not been fully elucidated yet. Following the advent of genome-wide association studies (GWASs) and large-scale consortial meta-analyses of GWASs, several thousands of variants have been robustly associated with complex phenotypes of medical relevance (http://www.genome.gov/gwastudies), giving valuable insights into underlying biological processes. GWASs are designed to provide a survey of common variation [minor allele frequency (MAF) > 0.05], therefore examining only a portion of the genomic landscape of complex traits. Low-frequency (MAF 0.01–0.05) and rare (MAF < 0.01) variation has thus far been more challenging to access. Early studies on data from deep sequencing of small numbers of loci and more recently larger-scale studies (e.g. 1000 Genomes Project) demonstrate that rare variants constitute the majority of polymorphic sites in human populations (1–3).

ACCESSING RARE VARIANTS

Current approaches to investigate the effect of rare variants in complex traits involve direct genotyping—for example, through targeted arrays like the exome chip (http://genome.sph.umich.edu/wiki/Exome_Chip_Design), metabochip (4) or immunochip (5), using the GWAS as a scaffold to impute low-frequency variants based on a sequenced reference panel (e.g. 1000 Genomes Project) (3), or resequencing of specific regions and increasingly the whole exome (WES) or the whole genome sequencing (WGS) (schematic overview in Fig. 1). The most commonly used WGS platforms generate millions of short sequence reads that are then aligned to a reference genome through read mapping. Variant calling algorithms are subsequently employed to identify candidate sites at which one or more samples differ from the reference sequence and to

*To whom correspondence should be addressed at: Wellcome Trust Sanger Institute, The Morgan Building, Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1HH, UK. Tel: +44-1223496868; Fax: +44-1223496826; Email: eleftheria@sanger.ac.uk

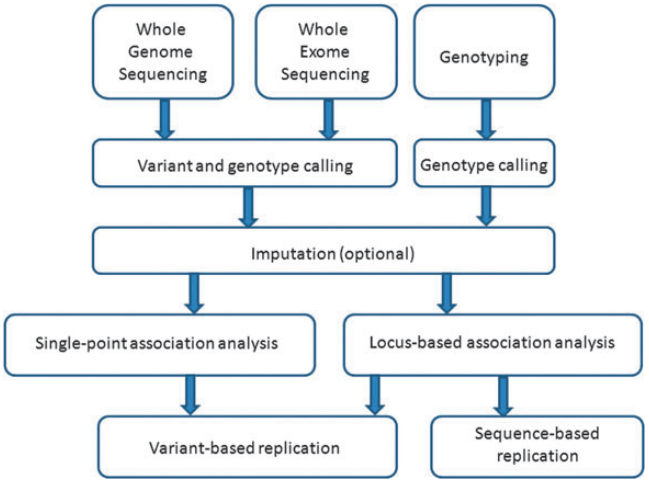

Figure 1. An overview of steps taken in the search for low-frequency and rare variants affecting complex traits.

call genotypes across samples. Currently, studies tend to focus on single nucleotide variants, as accurate calling of copy number variation is less straightforward. High-depth WGS is currently the preferred approach to exhaustively study variation across the full allelic spectrum genome-wide but for complex trait studies, where a large number of individuals need to be sampled costs remain prohibitively expensive. For mapping complex trait variants, study designs that increase the number of sequenced samples by decreasing sequencing depth are more powerful and cost-effective than sequencing fewer individuals at high depth (4,6–8), but detection and calling accuracy at rare variant sites can be compromised. WES interrogates only the coding regions of genes. The lower cost of WES compared with WGS means that higher depths are feasible, leading to higher accuracy in rare variant calls.

EXAMPLES OF RARE VARIANTS CONTRIBUTING TO COMPLEX TRAITS

There is growing evidence that rare variants can play a role in complex disease aetiology. One of the earliest examples came from the field of breast and ovarian cancers with the discovery of multiple rare mutations in the *BRAC1* and *BRAC2* genes (reviewed in 9–11). A more recent example is provided by a study which identified four rare variants acting independently on type 1 diabetes (T1D) risk through targeted resequencing of *IFIH1*, a gene located in a region previously associated with T1D by GWAS (12). A further report of rare variants exerting individual effects showed that five distinct rare variants in *NOD2* are associated with the risk of Crohn's disease and appear to act independently from each other and from the previously implicated low-frequency causal variants (13–15). In a different allelic architecture paradigm, sequencing the exons of the GWAS-implicated type 2 diabetes (T2D) gene *MTNR1B* identified several rare variants impairing melatonin receptor 1B function which collectively contribute to T2D risk (16). Several examples of rare and low-frequency variants with individually large effect sizes have also been reported. For example, a rare missense variant in *MYH6* was found to be associated with high (~12-fold) risk of sick sinus syndrome (17) in a pioneering study from Iceland, which combined whole-genome sequencing, genome-wide genotyping and imputation-based approaches. Notably, over the past few years, WGS of affected trios has led to the identification of several *de novo* mutations implicated in the aetiology of autism (18–21), schizophrenia (22,23) and intellectual disability (24).

Perhaps, the most abundant examples of rare variants acting collectively have emerged from the study of medically relevant quantitative traits and in particular circulating lipid levels. Screening for variants in genes implicated in Mendelian forms of low high-density lipoprotein cholesterol (HDL-C) levels revealed an aggregation of rare alleles in individuals with low HDL-C compared with those with high HDL-C (25). Resequencing of *ANGPTL4* uncovered both rare and common variants that reduce triglycerides and increase HDL (26). Recently, a region near *PARM1* was implicated in HDL-C level variation through a sliding-window burden testing approach performed on variants with MAF < 0.01 in one of the first WGS-based complex trait studies to date (27). An aggregation of rare alleles has also been associated with low-density lipoprotein cholesterol (LDL-C) (28) and with blood pressure reduction and protection against hypertension (29). The first reported application of exome array genotyping identified five independently acting, low-frequency variants associated with fasting proinsulin concentrations (30).

POPULATION ISOLATES

The study of rare variation can be empowered by focusing on population isolates (31,32). Isolated populations are characterized by increased phenotypic, genetic and environmental homogeneity. In these populations, rare variants may have drifted up in frequency and linkage disequilibrium (LD) tends to be extended. Founder populations carry a subset of the genetic variation present in the original population from which they have diverged. The effect of random genetic drift on increasing allele frequencies is higher for rare compared with common variants. Population bottlenecks can also affect the genetic architecture of isolates by reducing the population size and hence heterogeneity, increasing endogamy levels and subsequently increasing levels of homozygosity and LD. Population isolates tend to demonstrate geographical and/or cultural isolation frequently commensurate with a homogeneous set of environmental exposures, e.g. diet and lifestyle.

For example, the Iceland-based deCODE study has been successful in identifying rare variants contributing to complex traits by leveraging these characteristics of population isolates in conjunction with extended genealogical information for a variety of complex traits (including prostate cancer, Alzheimer's disease, gout and serum uric acid levels) (33–36). Association of a rare functional variant (R19X) in the *APOC3* gene with HDL-C and triglycerides levels was first detected in the Amish founder population and later confirmed in a Greek population isolate from Crete. R19X appears to have drifted up in frequency independently in the two population isolates (37,38).

RARE VARIANT REFERENCE PANELS

Large-scale collaborative efforts aiming to help understand the full spectrum of sequence variation serve as valuable resources for the scientific community. The 1000 Genomes Project (3) has helped enhance our understanding of low-frequency and rare variants by providing a catalogue of common and uncommon variation through WGS and exon sequencing across several global populations. Importantly, it has also enabled the first-line use of imputation approaches (39) to infer genotypes at untyped low-frequency variants in large-scale GWAS meta-analysis efforts. This approach has started contributing to the identification of novel disease-associated variants (e.g. 40). The National Heart Lung and Blood Institute Exome Sequencing Project (NHLBI-ESP) (https://esp.gs.washington.edu) has provided insights into rare coding variants through WES of 6500 samples in phenotyped sets from the USA. The UK10K Project (www.uk10k.org) has undertaken high-depth WES of 6000 and low-depth WGS of 4000 well-phenotyped individuals primarily from the UK. The imputation of low-frequency and rare variants can be challenging compared with common alleles, and the availability of very large-scale reference panels can improve imputation performance. It is envisaged that large-scale WGS efforts will join forces to generate an overarching reference panel to enable efficient and widespread use of the generated data.

RARE VARIANT ASSOCIATION ANALYSIS

Statistical genetics considerations of rare variant association analysis have been the focus of intensive method development over the last few years. The single-point analysis of rare variants is under-powered, because not enough copies of the rare variant allele are observed in sample sizes typically available to date. Instead of examining the association of each rare variant in isolation, multivariate methods that combine information across multiple variant sites within a gene or other functional genomic region are a viable alternative strategy (Fig. 1). A plethora of such locus-specific statistical approaches have been developed and fall broadly into a few categories: collapsing methods based on summary statistics (Cohort Allelic Sum Test (41); Combined Multivariate and Collapsing Test (42); Weighted Sum Test (43); Variable-Threshold Approach (44)); methods based on similarities among individual sequences (Kernel Based Association Test (45); Sequence Kernel Association Test (46)); and regression models that use collapsed sets of variants and other factors as predictors (collapsing test using proportion of rare variants (47); Adaptive Sum Test (48); LASSO and Ridge Regression (49)) (50).

Collapsing methods aggregate information across multiple variants within a region of interest into a single quantity, which is then used to test for trait association with an accumulation of rare minor alleles. Collapsing methods vary in the way they collapse the variants and in the chosen statistical test (42,47), for example options include using a regression approach that models the phenotype as a function of the proportion of rare variants at which an individual carries a minor allele, or as a function of the presence or absence of a minor allele at any rare variant within an individual. Collapsing approaches assume that all collapsed variants are associated with the disease, and that they can be either deleterious or protective. Alternative approaches that model similarities among individual sequences using various kernel functions, such as KBAT (45) and SKAT (46), are multivariate tests that combine single-variant test statistics. They make no assumptions about the probability or direction of effect of each rare variant, and are therefore more flexible, given that the allelic architecture of complex traits is unknown. A unified approach between collapsing methods and SKAT (SKAT-O (51)) adapts to the data to give more weight to the test that makes the most realistic assumptions for the specific region and trait of interest.

Genotype uncertainty metrics for imputed genetic variants or for sequencing-derived variants could be incorporated as weights in different statistical tests, instead of filtering out variants with low imputation or quality scores, to increase association power as shown by (52). Moreover, variants in rare association tests can be down or up weighted according to their probability of being functional. Such weights can be based on the MAF under the assumption that rarer variants are more likely to be deleterious according to the natural selection theory. Alternatively, weights can be based on functional annotation predictions. Coding variants that are predicted to have severe functional consequences may be hypothesized to confer larger phenotypic effects, have higher translational potential and may be more amenable to designing downstream functional experiments. Functional annotation of non-coding variation is more challenging and an active area of current research.

An open question for rare variant analysis in WGS studies is how to define the region of interest. In WES studies, such a decision is more straightforward, as a gene unit is an intuitive option. In WGS studies, a potential approach is to divide the genome into windows of certain physical size. However, it is not clear what the size of the windows should be and whether they should be overlapping or by how much. Genome-wide significance levels for the GWAS era were estimated to be at $P = 5 \times 10^{-8}$ based on the number of independent common-frequency variants across the genome calculated based on the European population data from the HapMap Project (53). This threshold has served the scientific community well, representing a standard to be attained before declaring significance. The analysis of rare variants across the genome requires a more stringent significance threshold that takes into account single-point common and rare variant tests as well as burden tests. This threshold is likely to vary depending on the study design parameters like sample size and sequencing depth, and is expected to be lower for African-descent populations.

META-ANALYSIS OF RARE VARIANTS

Meta-analysis of common variants in GWASs is a common strategy of combining studies examining the same trait to increase power to obtain statistical evidence of association. Traditional meta-analysis techniques, such as Fisher's (54) and Stouffer's (55) tests, that use region-level *P*-values from the different burden tests are not necessarily powerful in combining data across independent studies for rare variant association testing, as they do not capture all of the available information (56). Ideally, meta-analytical approaches for next-generation sequencing studies should result in little or no power loss when

compared with a joint analysis approach, in the same way as meta-analysis of single tests for common variants (57).

Lumley *et al.* (58) and Lee *et al.* (59) have independently developed meta-analytical techniques for SKAT (46) and SKAT-O (51), respectively. The latter approach can assume both homogeneous and heterogeneous genetic effects across studies, corresponding to a fixed- and random-effects meta-analysis model, respectively. Both approaches are similar in spirit and are based on study-specific summary statistics rather than individual-level data. They combine single-variant score statistics first across studies and then within a region. They also require between-variant covariance-type relationship statistics (such as LD structure) for each region, as well as MAF of variants.

Liu *et al* (60) and Tang and Lin (61) suggest approaches that encompass a number of popular gene-level association tests such as collapsing tests (42,47), variable threshold (44) and SKAT (46). Their methods also combine single-variant score statistics across studies. A unique feature of the Liu *et al* (2013) approach is that apart from calculating asymptotic *P*-values, it also evaluates significance in an empirical and numerically stable way via an adaptive Monte-Carlo simulation scheme. Another unique feature of Liu et al.'s approach is its ability to conduct conditional meta-analysis of gene-level tests. Lumley *et al.*, Lee *et al.* and Liu *et al.* (2013) show in simulation studies that their proposed approaches are as efficient as an analysis that pools individual-level data together. The evaluation of different meta-analysis approaches of rare variant tests is an active field of study.

POPULATION STRATIFICATION AT RARE VARIANTS

Population stratification at rare variants is an important consideration for next-generation association studies. Rare variants show increased population specificity (3). Based on the theoretical examination, rare variants can show a stronger pattern of population stratification than common variants, particularly in the presence of sharp spatial distributions for non-genetic risk of disease (62). Existing methods to correct for population stratification at common variants such as principal component analysis and genomic control have not been shown to effectively control stratification at rare variants with implications for both single-point and locus-based approaches (62–64). In empirical data from the UK population, rare variants were found to display different stratification patterns to common variants (65). These findings underscore the need for carefully matching samples, for example cases and controls, between geographical regions and highlight the need for replication in independent datasets.

REPLICATION OF RARE VARIANT SIGNALS

Strategies for replication of associations discovered in low-frequency and rare variants depend on the allelic architecture of the associated locus. For example, if a single low-frequency or rare variant is driving the signal, replication can be sought by genotyping the implicated variant in an independent sample set (e.g. 17). For associations uncovered via locus-based approaches, two replication strategies have been proposed: variant-based replication, where only variants found in the discovery phase are followed-up by, e.g. genotyping, and sequence-based replication, where the whole region is re-sequenced in the replication sample set and novel variants can be included in the test (Fig. 1). Under several simulation scenarios, it has been demonstrated that there are small gains in power when adopting the sequence-based replication design and more so if the discovery sample set is small. At medium- to large-scale studies and when discovery and replication sample sets are drawn from the same population, genotyping can offer a viable alternative solution (66).

The emerging generation of studies in search of low-frequency and rare variants affecting complex traits will require robust strategies to ensure high power in the context of an appropriate statistical framework. It is anticipated that sequence-based meta-analysis across diverse populations, including populations of African descent, will empower novel locus discovery, and that accruing the necessary sample sizes is likely to be a key determinant of success. Initial insights into the contribution of rare variation indicate a firm role in complex trait aetiology and suggest a combination of potential allelic architectures underpinning biological phenotypes. Their powerful detection will require tailored study design and analysis approaches.

ACKNOWLEDGEMENTS

We are grateful to Jeremy Schwartzentruber for useful comments.

Conflict of Interest statement. None declared.

FUNDING

K.P., I.T. and E.Z. are funded by the Wellcome Trust (098051). K.P. is funded by Arthritis Research UK (19542). Funding to pay the Open Access publication charges for this article was provided by The Wellcome Trust (098051).

REFERENCES

1. International HapMap 3 Consortium, Altshuler, D.M., Gibbs, R.A., Peltonen, L., Dermitzakis, E., Schaffner, S.F., Yu, F., Bonnen, P.E., de Bakker, P.I., Deloukas, P. *et al.* (2010) Integrating common and rare genetic variation in diverse human populations. *Nature*, **467**, 52–58.
2. Marth, G.T., Yu, F., Indap, A.R., Garimella, K., Gravel, S., Leong, W.F., Tyler-Smith, C., Bainbridge, M., Blackwell, T., Zheng-Bradley, X. *et al.*. (2011) The functional spectrum of low-frequency coding variation. *Genome Biol.*, **12**, R84.
3. 1000 Genomes Project Consortium, Abecasis, G.R., Auton, A., Brooks, L.D., DePristo, M.A., Durbin, R.M., Handsaker, R.E., Kang, H.M., Marth, G.T. and McVean, G.A. (2012) An integrated map of genetic variation from 1092 human genomes. *Nature*, **491**, 56–65.
4. Voight, B.F., Kang, H.M., Ding, J., Palmer, C.D., Sidore, C., Chines, P.S., Burtt, N.P., Fuchsberger, C., Li, Y., Erdmann, J. *et al.*. (2012) The metabochip, a custom genotyping array for genetic studies of metabolic, cardiovascular, and anthropometric traits. *PLoS Genet.*, **8**, e1002793.
5. Cortes, A. and Brown, M.A. (2011) Promise and pitfalls of the Immunochip. *Arthritis Res. Ther.*, **13**, 101.
6. Le, S.Q. and Durbin, R. (2011) SNP detection and genotyping from low-coverage sequencing data on multiple diploid samples. *Genome Res.*, **21**, 952–960.

7. Li, Y., Sidore, C., Kang, H.M., Boehnke, M. and Abecasis, G.R. (2011) Low-coverage sequencing: implications for design of complex trait association studies. *Genome Res.*, **21**, 940–951.
8. Pasaniuc, B., Rohland, N., McLaren, P.J., Garimella, K., Zaitlen, N., Li, H., Gupta, N., Neale, B.M., Daly, M.J., Sklar, P. *et al.* (2012) Extremely low-coverage sequencing and imputation increases power for genome-wide association studies. *Nat Genet.*, **44**, 631–635.
9. Stratton, M.R. and Rahman, N. (2008) The emerging landscape of breast cancer susceptibility. *Nat Genet.*, **40**, 17–22.
10. Collins, A. and Politopoulos, I. (2011) The genetics of breast cancer: risk factors for disease. *Appl. Clin. Genet.*, **4**, 11–19.
11. Kobayashi, H., Ohno, S., Sasaki, Y. and Matsuura, M. (2013) Hereditary breast and ovarian cancer susceptibility genes (review). *Oncol Rep*. First published on 19 Jun 2013. In press.
12. Nejentsev, S., Walker, N., Riches, D., Egholm, M. and Todd, J.A. (2009) Rare variants of IFIH1, a gene implicated in antiviral responses, protect against type 1 diabetes. *Science*, **324**, 387–389.
13. Hugot, J.P., Chamaillard, M., Zouali, H., Lesage, S., Cezard, J.P., Belaiche, J., Almer, S., Tysk, C., O'Morain, C.A., Gassul, M. *et al.* (2001) Association of NOD2 leucine-rich repeat variants with susceptibility to Crohn's disease. *Nature*, **411**, 599–603.
14. Ogura, Y., Bonen, D.K., Inohara, N., Nicolae, D.L., Chen, F.F., Ramos, R., Britton, H., Moran, T., Karaliuskas, R., Duer, R.H. *et al.* (2001) A frameshift mutation in NOD2 associated with susceptibility to Crohn's disease. *Nature*, **411**, 603–606.
15. Rivas, M.A., Beaudoin, M., Gardet, A., Stevens, C., Sharma, Y., Zhang, C.K., Boucher, G., Ripke, S., Ellinghaus, D., Burtt, N. *et al.* (2011) Deep resequencing of GWAS loci identifies independent rare variants associated with inflammatory bowel disease. *Nat. Genet.*, **43**, 1066–1073.
16. Bonnefond, A., Clément, N., Fawcett, K., Yengo, L., Vaillant, E., Guillaume, J.L., Dechaume, A., Payne, F., Roussel, R., Czernichow, S. *et al.* (2012) Rare MTNR1B variants impairing melatonin receptor 1B function contribute to type 2 diabetes. *Nat. Genet.*., **44**, 297–301.
17. Holm, H., Gudbjartsson, D.F., Sulem, P., Masson, G., Helgadottir, H.T., Zanon, C., Magnusson, O.T., Helgason, A., Saemundsdottir, J., Gylfason, A. *et al.* (2011) A rare variant in MYH6 is associated with high risk of sick sinus syndrome. *Nat. Genet.*, **43**, 316–320.
18. O'Roak, B.J., Deriziotis, P., Lee, C., Vives, L., Schwartz, J.J., Girirajan, S., Karakoc, E., Mackenzie, A.P., Ng, S.B., Baker, C. *et al.* (2011) Exome sequencing in sporadic autism spectrum disorders identifies severe de novo mutations. *Nat. Genet.*, **43**, 585–589.
19. Sanders, S.J., Murtha, M.T., Gupta, A.R., Murdoch, J.D., Raubeson, M.J., Willsey, A.J., Ercan-Sencicek, A.G., DiLullo, N.M., Parikshak, N.N., Stein, J.L. *et al.* (2012) De novo mutations revealed by whole-exome sequencing are strongly associated with autism. *Nature*, **485**, 237–241.
20. Neale, B.M., Kou, Y., Liu, L., Ma'ayan, A., Samocha, K.E., Sabo, A., Lin, C.F., Stevens, C., Wang, L.S., Makarov, V. *et al.* (2012) Patterns and rates of exonic de novo mutations in autism spectrum disorders. *Nature*, **485**, 242–245.
21. O'Roak, B.J., Vives, L., Girirajan, S., Karakoc, E., Krumm, N., Coe, B.P., Levy, R., Ko, A., Lee, C., Smith, J.D. *et al.* (2012) Sporadic autism exomes reveal a highly interconnected protein network of de novo mutations. *Nature*, **485**, 246–250.
22. Girard, S.L., Gauthier, J., Noreau, A., Xiong, L., Zhou, S., Jouan, L., Dionne-Laporte, A., Spiegelman, D., Henrion, E., Diallo, O. *et al.* (2011) Increased exonic *de novo* mutation rate in individuals with schizophrenia. *Nat. Genet.*, **43**, 860–863.
23. Xu, B., Roos, J.L., Dexheimer, P., Boone, B., Plummer, B., Levy, S., Gogos, J.A. and Karayiorgou, M. (2011) Exome sequencing supports a *de novo* mutational paradigm for schizophrenia. *Nat. Genet.*, **43**, 864–868.
24. Vissers, L.E., de Ligt, J., Gilissen, C., Janssen, I., Steehouwer, M., de Vries, P., van Lier, B., Arts, P., Wieskamp, N., del Rosario, M. *et al.* (2010) A *de novo* paradigm for mental retardation. *Nat. Genet.*, **42**, 1109–1112.
25. Cohen, J.C., Kiss, R.S., Pertsemlidis, A., Marcel, Y.L., McPherson, R. and Hobbs, H.H. (2004) Multiple rare alleles contribute to low plasma levels of HDL cholesterol. *Science*, **305**, 869–872.
26. Romeo, S., Pennacchio, L.A., Fu, Y., Boerwinkle, E., Tybjaerg-Hansen, A., Hobbs, H.H. and Cohen, J.C. (2007) Source population-based resequencing of ANGPTL4 uncovers variations that reduce triglycerides and increase HDL. *Nat. Genet.*, **39**, 513–516.
27. The Cohorts for Heart and Aging Research in Genetic Epidemiology (CHARGE) Consortium, Morrison, A.C., Voorman, A., Johnson, A.D., Liu, X., Yu, J., Li, A., Muzny, D., Yu, F., Rice, K. *et al.*. (2013) Whole-genome sequence-based analysis of high-density lipoprotein cholesterol. *Nat Genet.*, **45**, 899–901.
28. Cohen, J.C., Pertsemlidis, A., Fahmi, S., Esmail, S., Vega, G.L., Grundy, S.M. and Hobbs, H.H. (2006) Multiple rare variants in NPC1L1 associated with reduced sterol absorption and plasma low-density lipoprotein levels. *Proc. Natl Acad. Sci. USA*, **103**, 1810–1815.
29. Ji, W., Foo, J.N., O'Roak, B.J., Zhao, H., Larson, M.G., Simon, D.B., Newton-Cheh, C., State, M.W., Levy, D. and Lifton, R.P. (2008) Rare independent mutations in renal salt handling genes contribute to blood pressure variation. *Nat. Genet.*, **40**, 592–599.
30. Huyghe, J.R., Jackson, A.U., Fogarty, M.P., Buchkovich, M.L., Stančáková, A., Stringham, H.M., Sim, X., Yang, L., Fuchsberger, C., Cederberg, H. *et al.* (2013) Exome array analysis identifies new loci and low-frequency variants influencing insulin processing and secretion. *Nat. Genet.*, **45**, 197–201.
31. Peltonen, L., Palotie, A. and Lange, K. (2000) Use of population isolates for mapping complex traits. *Nat. Rev. Genet.*, **1**, 182–190.
32. Zeggini, E. (2011) Next-generation association studies for complex traits. *Nat. Genet.*, **243**, 287–288.
33. Gudmundsson, J., Sulem, P., Gudbjartsson, D.F., Masson, G., Agnarsson, B.A., Benediktsdottir, K.R., Sigurdsson, A., Magnusson, O.T., Gudjonsson, S.A., Magnusdottir, D.N. *et al.* (2012) A study based on whole-genome sequencing yields a rare variant at 8q24 associated with prostate cancer. *Nat. Genet.*, **44**, 1326–1329.
34. Stacey, S.N., Sulem, P., Jonasdottir, A., Masson, G., Gudmundsson, J., Gudbjartsson, D.F., Magnusson, O.T., Gudjonsson, S.A., Sigurgeirsson, B., Thorisdottir, K. *et al.* (2011) A germline variant in the TP53 polyadenylation signal confers cancer susceptibility. *Nat. Genet.*, **43**, 1098–1103.
35. Jonsson, T., Atwal, J.K., Steinberg, S., Snaedal, J., Jonsson, P.V., Bjornsson, S., Stefansson, H., Sulem, P., Gudbjartsson, D., Maloney, J. *et al.* (2012) A mutation in APP protects against alzheimer's disease and age-related cognitive decline. *Nature*, **488**, 96–99.
36. Sulem, P., Gudbjartsson, D.F., Walters, G.B., Helgadottir, H.T., Helgason, A., Gudjonsson, S.A., Zanon, C., Besenbacher, S., Bjornsdottir, G., Magnusson, O.T. *et al.* (2011) Identification of low-frequency variants associated with gout and serum uric acid levels. *Nat. Genet.*, **43**, 1127–1130.
37. Pollin, T.I., Damcott, C.M., Shen, H., Ott, S.H., Shelton, J., Horenstein, R.B., Post, W., McLenithan, J.C., Bielak, L.F., Peyser, P.A. *et al.* (2008) A null mutation in human APOC3 confers a favorable plasma lipid profile and apparent cardioprotection. *Science*, **322**, 1702–1705.
38. Tachmazidou, I., Dedoussis, G., Southam, L., Farmaki, A.E., Ritchie, G., Xifara, D., Hatzikotoulas, K., Matchan, A., Rayner, N.W., Chen, Y. *et al.*. A rare functional variant in APOC3 is associated with lipid traits and has risen in frequency in distinct population isolates. *Personal Communication*.
39. Marchini, J., Howie, B., Myers, S., McVean, G. and Donnelly, P. (2007) A new multipoint method for genome-wide association studies by imputation of genotypes. *Nat Genet.*, **39**, 906–913.
40. Day-Williams, A.G., Southam, L., Panoutsopoulou, K., Rayner, N.W., Esko, T., Estrada, K., Helgadottir, H.T., Hofman, A., Ingvarsson, T., Jonsson, H. *et al.* (2011) A variant in MCF2L is associated with osteoarthritis. *Am. J. Hum. Genet.*, **89**, 446–450.
41. Morgenthaler, S. and Thilly, W.G. (2007) A strategy to discover genes that carry multi-allelic or mono-allelic risk for common diseases: a cohort allelic sums test (CAST). *Mutat. Res.*, **615**, 28–56.
42. Li, B. and Leal, S. (2008) Methods for detecting associations with rare variants for common diseases: application to analysis of sequence data. *Am. J. Hum. Genet.*, **83**, 311–321.
43. Madsen, B.E. and Browning, S.R. (2009) A groupwise association test for rare mutations using a weighted sum statistic. *PLoS Genet.*, **5**, e1000384.
44. Price, A.L., Kryukov, G.V., de Bakker, P.I., Purcell, S.M., Staples, J., Wei, L.J. and Sunyaev, S.R. (2010) Pooled association tests for rare variants in exon-resequencing studies. *Am. J. Hum. Genet.*, **86**, 832–838.
45. Mukhopadhyay, I., Feingold, E., Weeks, D.E. and Thalamuthu, A. (2010) Association tests using kernel-based measures of multi-locus genotype similarity between individuals. *Genet. Epidemiol.*, **34**, 213–221.
46. Wu, M.C., Lee, S., Cai, T., Li, Y., Boehnke, M. and Lin, X. (2011) Rare variant association testing for sequencing data using the Sequence Kernel Association Test (SKAT). *Am. J. Hum. Genet.*, **89**, 82–93.
47. Morris, A.P. and Zeggini, E. (2010) An evaluation of statistical approaches to rare variant analysis in genetic association studies. *Genet. Epidemiol.*, **34**, 188–193.
48. Han, F. and Pan, W. (2010) A data-adaptive sum test for disease association with multiple common or rare variants. *Human Heredity*, **70**, 42–54.

49. Zhou, H., Sehl, M.E., Sinsheimer, J.S. and Lange, K. (2010) Association screening of common and rare genetic variants by penalized regression. *Bioinformatics*, **26**, 2375–2382.
50. Asimit, J. and Zeggini, E. (2010) Rare variant association analysis methods for complex traits. *Annu. Rev. Genet.*, **44**, 293–308.
51. Lee, S., Emond, M.J., Bamshad, M.J., Barnes, K.C., Rieder, M.J., Nickerson, D.A., Christiani, D.C., Wurfel, M.M. and Lin, X. (2012) Optimal unified approach for rare-variant association testing with application to small-sample case–control whole-exome sequencing studies. *Am. J. Hum. Genet.*, **91**, 224–237.
52. Asimit, J.L., Day-Williams, A.G., Morris, A.P. and Zeggini, E. (2012) ARIEL And AMELIA: testing for an accumulation of rare variants using next-generation sequencing data. *Hum Hered* **73**, 84–R94.
53. International HapMap Consortium. (2005) A haplotype map of the human genome. *Nature*, **437**, 1299–1320.
54. Fisher, R.A. (1932) *Statistical Methods for Research Workers*. Oliver and Boyd, Edinburgh.
55. Stouffer, S.A., Suchman, E.A., DeVinney, L.C. and Williams, J.R.M. (1949) *The American Soldier, Volume I: Adjustment During Army Life*. Princeton University Press, Princeton, NJ.
56. Liu, L., Sabo, A., Neale, B.M., Nagaswamy, U., Stevens, C., Lim, E., Bodea, C.A., Muzny, D., Reid, J.G., Banks, E. *et al.* (2013) Analysis of rare, exonic variation amongst subjects with autism spectrum disorders and population controls. *PLoS Genet.*, **9**, e1003443.
57. Lin, D. and Zeng, D. (2010) Meta-analysis of genome-wide association studies: no efficiency gain in using individual participant data. *Genet. Epidemiol.*, **34**, 60–66.
58. Lumley, T., Brody, J., Dupuis, J. and Cupples, A. (2013) Meta-analysis of a rare variant association test. http://stattech.wordpress.fos.auckland.ac.nz/files/2012/11/skat-meta-paper.pdf.
59. Lee, S., Teslovich, T.M., Boehnke, M. and Lin, X. (2013) General framework for meta-analysis of rare variants in sequencing association studies. *Am. J. Hum. Genet.*, **93**, 1–12.
60. Liu, D.J., Peloso, G.M., Zhan, X., Holmen, O., Zawistowski, M., Feng, S., Nikpay, M., Auer, P.L., Goel, A., Zhang, H. *et al.*. (2013) Meta-analysis of gene level association tests. http://arxiv.org/abs/1305.1318.
61. Tang, Z.Z. and Lin, D.Y. (2013) MASS: meta-analysis of score statistics for sequencing studies. *Bioinformatics*, **29**, 1803–1805.
62. Mathieson, I. and McVean, G. (2012) Differential confounding of rare and common variants in spatially structured populations. *Nat. Genet*., **44**, 243–246.
63. Liu, Q., Nicolae, D.L. and Chen, L.S. (2013) Marbled inflation from population structure in gene-based association studies with rare variants. *Genet. Epidemiol.*, **37**, 286–292.
64. He, H., Zhang, X., Ding, L., Baye, T.M., Kurowski, B.G. and Martin, L.J. (2011) Effect of population stratification analysis on false-positive rates for common and rare variants. *BMC Proc.*, **5**(Suppl 9), S116.
65. Babron, M.C., de Tayrac, M., Rutledge, D.N., Zeggini, E. and Génin, E. (2012) Rare and low frequency variant stratification in the UK population: description and impact on association tests. *PLoS One*, **7**, e46519.
66. Liu, D.J. and Leal, S.M. (2010) Replication strategies for rare variant complex trait association studies via next-generation sequencing. *Am. J. Hum. Genet.*, **87**, 790–801.