Published in final edited form as:

J Med Genet. 2015 January ; 52(1): 10–16. doi:10.1136/jmedgenet-2014-102697.

Case-Only Exome Sequencing and Complex Disease Susceptibility Gene Discovery: Study Design Considerations

Lang Wu^{1,2}, Daniel J. Schaid¹, Hugues Sicotte¹, Eric D. Wieben³, Hu Li⁴, and Gloria M. Petersen¹

¹Department of Health Sciences Research, Mayo Clinic, Rochester, Minnesota, USA 55905

²Center for Clinical and Translational Science, Mayo Clinic, Rochester, Minnesota, USA 55905

³Department of Biochemistry and Molecular Biology, Mayo Clinic, Rochester, Minnesota, USA 55905

⁴Department of Molecular Pharmacology and Experimental Therapeutics, Mayo Clinic, Rochester, Minnesota, USA 55905

Abstract

Whole exome sequencing (WES) provides an unprecedented opportunity to identify the potential etiologic role of rare functional variants in human complex diseases. Large-scale collaborations have generated germline WES data on patients with a number of diseases, especially cancer, but less often on healthy controls under the same sequencing procedures. These data can be a valuable resource for identifying new disease susceptibility loci, if study designs are appropriately applied. This review describes suggested strategies and technical considerations when focusing on case-only study designs that use WES data in complex disease scenarios. These include variant filtering based on frequency and functionality, gene prioritization, interrogation of different data types, and targeted sequencing validation. We propose that if case-only WES designs were applied in an appropriate manner, new susceptibility genes containing rare variants for human complex diseases can be detected.

Introduction

Human genetics has always attempted to incorporate emerging technologies into studies of disease etiology. Various epidemiologic study designs have also been employed to identify genetic predisposition to human diseases. These designs include case series, case-control, cohort, and family studies. For decades, the candidate gene approach was widely applied, comparing prevalence of genetic variants in biologically plausible genes or pathways between controls and cases with disease [1]. This approach begins with assembling known gene(s) of interest, in which polymorphisms that may have functional or population

Corresponding Author: Gloria M. Petersen, Ph.D., Department of Health Sciences Research, Mayo Clinic College of Medicine, 200 First Street SW, Rochester, MN 55905. Phone: 507-538-1563; Fax: 507-266-2478; Petersen.Gloria@mayo.edu.

Contributors

LW and GMP contributed to the conceptualization of this manuscript. LW wrote the first draft of this manuscript. All authors revised and approved this manuscript.

relevance are selected for genotyping. Although demonstrated to be relatively cost and time effective, this design generally tests hypothesized associations, and does not typically discover novel genetic loci beyond these candidates [2]. Moreover, due to variation in sample composition and assays, candidate genes are infrequently replicated.

In the late 1990's and early 2000's, the combination of extensive variation uncovered by genome sequencing, high throughput genotyping technology, and availability of very large samples of cases and controls with DNA biospecimens spurred genome-wide association study (GWAS) design [3]. These studies were based on the hypothesis that common diseases could be explained by common variants [4]. The GWAS studies combine case-control design with high throughput genotyping of pre-selected common single nucleotide polymorphisms (SNPs) from across the genome, with the goal of agnostically characterizing genetic susceptibility to complex diseases [5 6]. For nearly a decade, through GWAS many SNPs associated with disease susceptibility were identified [7 8]. However, in contrast to early expectations of this methodology, findings from GWAS analyses ultimately explained only a modest proportion of disease heritability [9 10]. Signals for susceptibility detected from GWAS remain difficult to meaningfully interpret or facilely translate, since most SNPs utilized in GWAS were selected because their frequencies were more common and represented regions by their linkage disequilibrium structure [11 12]; the vast majority of GWAS SNPs have no known functional implication.

In contrast, the family-based genetic linkage study design [13–15] has been successful in localizing genomic regions that harbor rare disease mutations typically with high penetrance, pointing to genes for Mendelian diseases or Mendelian subsets of complex diseases [16 17]. However, performing linkage analysis of common diseases has been challenging, with study power weakened due to genetic heterogeneity, extensive phenocopies, and potentially low penetrance of causal mutations.

Since the majority of complex disease heritability has not been explained by either very rare mutations or common variants, it has been proposed that there is a catalog of “missing heritability” due to rare variants in the moderate portion of the penetrance spectrum [9 18–20]. Granted that non-coding regulatory regions in the genome may play a role [21 22], we will focus here on gene discovery in exome regions. The whole exome sequencing (WES) study design has been proposed to provide a new strategy to identify missing heritability. It has been demonstrated that as next-generation sequencing (NGS) technologies improve, there is improved economy of scale [23–25]. Analysis of WES data can be a cost-effective means to explore etiologic roles of rare, functional variants in complex diseases.

Researchers have begun to address NGS studies, including proposals for appropriate study design and statistical and bioinformatics analysis in the conduct of WES in family-based or case-control settings [26–31]. However, a typical WES study using a case-control design has been argued to require a very large sample size, which even today creates an unaffordable cost burden. Recently, with very large-scale collaborations in human diseases, especially cancer, extensive datasets from germline WES have become available, though not of the magnitude posited to be needed in a case-control design [32 33]. Furthermore, it can be expected that many more such datasets will be available rapidly[34 35]. Since sequencing data are generated on affected patients only, these resources do not lend themselves to

conventional study designs. Yet, if carefully executed approaches are used, they may be valuable for identifying new disease susceptibility genes. This is a very active area of research at present, in that many investigators in large-scale collaborations are applying these data to gene discovery in complex diseases. Researchers are also enhancing statistical methods and tools for appropriate analysis when using publicly available controls [36]. We compare the main advantages and disadvantages in Table 1 and provide examples of four major study types, including (I) Family-based whole genome sequencing (WGS) or WES, (II) Case-control candidate gene deep sequencing, (III) Case-control WGS or WES, and (IV) Case-only WES. Each study type has its specific advantages and disadvantages; as we have described, the first three study types have been extensively deployed to detect predisposition genes in many diseases. The case-only WES study design is the least explored, but can be promising for identifying new susceptibility rare variants/genes for complex diseases. We review lessons learned and discuss study design considerations when case-only WES data are utilized. Figure 1 illustrates proposed strategies to filter variants according to frequency, functionality, disease phenotype spectrum, gene prioritization that combines information from other data, and validation through targeted sequencing. Each step of the proposed process is discussed. Some of these proposed strategies are suitable for other designs beyond the case-only WES,

Considerations in variant filtering and prioritization

Variant filtering and prioritization is a process to determine a limited number of high-priority variants from a vast number obtained in a sequencing study [37]. Variant filtering is necessary because the initial variants identified are so numerous that effective analysis is impossible, and high false discovery rates will occur if variants are not prioritized. After variants are called in the WES analysis, approaches to variant filtering include assessment of variant frequencies and variant functionality (Figure 1):

A: Variant frequency filtering

Using allele frequency as a means to filter is suggested as a first step to prioritize variants because it is relatively robust. Researchers can consult published frequencies of variants in the population of interest. Examples of publicly available data on control subjects include the 1000 Genomes Project and the Exome Sequencing Project (ESP) [38–40]. Researchers need to be aware that participants in ESP are not all completely healthy individuals, so caution should be exercised when this dataset is used as a reference for evaluating diseases which are seen in some participants of ESP. Researchers also have to consider several factors by which to filter when choosing the lower boundary or cutoff of the minor allele frequency (MAF) (i.e., 1% or less). This can include basing the cutoff on the current understanding of the disease paradigm and the prevalence of the disease. Generally, a threshold of 1% is appropriate for MAF filtering. There are several reasons for using this threshold. First, the more common variants are well captured by the GWAS design. Through the 1000 Genomes Project, the majority (over 99%) of variants with a MAF above 1% in the general population were captured, and through imputing with the 1000 Genomes data, more common variants (>1% MAF) can be captured in combination with the GWAS design, which has been widely utilized for discovery of genetic etiology of complex diseases [37].

Second, the 1000 Genomes Project showed that relatively rare variants ($MAF < 1\%$ or even $< 0.5\%$) contain the majority of more “deleterious” variants, and thus may represent an important source of disease etiology. This is particularly true for the most highly conserved coding sites, in which 85% of non-synonymous variants and over 90% of stop-gain and splice-disrupting variants are within the category of rare variants [37].

B: Variant functionality filtering

Another important criterion researchers may use to filter variants is the predicted functionality of the variant. Variants that result in short insertion/deletions (indels), stoploss/stopgain, and splice site alterations are widely accepted as evidence for how deleterious (disease-causing) they are. However, the sensitivity and specificity of currently available bioinformatically-based function annotation tools are not satisfactory, especially in the case of insertion-deletion variants (indels) [41]. Researchers obtain high false-positive and false-negative rates when calling this type of variant. Another functional category of variants, the nonsynonymous variant, is also of interest because of its potentially deleterious effect on protein function. A variety of *in silico* tools, such as SIFT and PolyPhen-2 [42 43], have been developed and widely used to predict the functionality of these variants. To date, the overall performance of these available tools is less than optimal, with a relatively high sensitivity (~70%) but low specificity (~15%), which thus limits our confidence in variant filtering if using only these tools [44]. It is an active area of research to develop new tools/methods, including those building a classifier based on multiple prediction tools to address these problems [22 45–56]. For example, several new methods beyond the scope of SIFT or PolyPhen-2 have been developed, based on amino acid contact energy change [56], integrating scores on information about known disease genes, with haploinsufficiency prediction, and variant impact prediction [51], or employing sequence, structure, and residue-contact network features [49]. Others improve the effects of SIFT and PolyPhen-2 by more comprehensive considerations, including evolutionary balancing [46]. In addition, combining multiple prediction tools and generating a consensus probability score have been extensively investigated and found to demonstrate superior outcomes [45 52 53 55]. However, one concern raised by researchers in this arena is that these different tools or consensus classifications were often applied to unique datasets for evaluation. More unbiased “gold-standard” datasets are required to provide a comprehensive evaluation about their sensitivities and specificities. Until the final, most optimal tool is validated, current approaches should be utilized with caution in the variant filtering process.

C: Disease phenotype spectrum

Knowledge of the spectrum of the disease phenotype and its characteristics brings useful information to help prioritize variants. For certain diseases, especially cancer, there is a possibility that heterozygous variants at the germline level can contribute to disease etiology if further somatic change affects the target tissue, i.e., through loss of heterozygosity. Variant filtering can be focused if further somatic information (if applicable) is available, as discussed below.

D: Gene prioritization

In addition to variant filtering, prioritization at the gene level can be another informative procedure in identifying genetic susceptibility factors. Several strategies can be potentially useful to prioritize potential susceptibility genes.

(1): Established genes known to play roles in etiology of other biologically relevant diseases—Numerous studies have provided evidence for correlation between different human diseases, including correlation between Mendelian and complex diseases [57–61] as well as between complex diseases [62–67]. Thus, it may be valid to hypothesize that genes causing related conditions would be good candidates for explaining susceptibility to the disease of interest. It also may be warranted to conduct pathway or network analysis [68–72] to determine whether specific pathways or networks involving those established genes are frequently mutated in patients. This could provide clues for prioritizing genes from the candidate list. For example, protein-protein interaction networks can be generated from genes carrying potentially deleterious mutations, and simulation studies could be conducted to determine whether the experimental network has more edges and a higher clustering coefficient than expected by chance. This method has been successfully demonstrated for autism [73], and could be useful in WES studies as well.

(2): Publicly available data on controls—The use of publicly available sequence data on healthy population controls can be considered in designing studies when researchers with limited resources perform WES only on disease cases. It is tempting to employ publicly available data (1000 Genomes Project or ESP) to perform comparisons by association analysis, either at the variant level or gene level. The goal would be to discover statistical evidence of differential variant/mutation burden in cases versus the data from publicly available controls. We would argue, however, that employing this type of direct comparison mandates extreme caution. Different sequencing initiatives may use very different sequencing kits/reagents, bioinformatics programs/pipelines for variant calling and annotation, and quality control filters and model parameters when generating the data. There is also the concern of variation from study to study in sequencing depth of coverage [74]. All these procedure differences may cause systematic deviations between different sequencing results. Thus, a direct statistical comparison to these publicly available data may not be sufficient to address these potential complications. A recent study [36] used external control datasets in association testing, and found problematic consequences, including residual inflation from sequencing differences from cases and controls [36]. The study concluded that variant filters are necessary for correcting for inflation in the tails of the distribution while it does not adjust for inflation in the median. More research is needed to resolve this problem. Caution is mandated in concluding that there is a greater preponderance of rare variants in sequenced cases when compared to publicly available control data. Nevertheless, we are not calling for complete avoidance of publicly available control datasets. Certain considerations could at least partially resolve the problem. For example, when concerned whether cases and publicly available controls have the same genetic ancestry, principal component analysis using common variants of these individuals could be conducted to address the potential population stratification issue. Improved

analysis methods are warranted to better employ publicly available data on controls in such scenarios.

E: Combining information from other data

Additional data types, including epigenetic variation, mRNA expression, and somatic genetic variation (if applicable), can be helpful in prioritizing promising gene candidates. In the case of cancers, large-scale sequencing and other analyses have been conducted and made available through The Cancer Genome Atlas (TCGA) [75] and the International Cancer Genome Consortium (ICGC) [76]. It is useful to interrogate all of these information sources when identifying candidate susceptibility genes/variants because they can provide further understanding as to how germline genetic variation can translate to the genesis of the disease process. When additional evidence can be added to germline variant discovery, it brings plausibility to the biological relevance of the susceptibility gene. This can better guide us in determining the most likely susceptibility genes.

F: Validation through targeted sequencing

An important design element, if resources are available, is to perform a validation step of targeted deep sequencing of the identified promising susceptibility genes in an independent sample of cases and controls. Cases and controls should be sequenced contemporaneously and preferably by the same laboratory or using the same platform and parameters.

Appropriate statistical tests comparing the burden of rare variants in cases versus controls can then be conducted to determine disease susceptibility genes. Relevant methods including burden tests, adaptive burden tests, variance-component tests, combined tests and collapsing methods, etc. have been extensively summarized in literature and can be applied in this setting [77–78].

Applying lessons to future study designs

Were case-only WES designs applied in an appropriate manner, we would argue that new susceptibility genes containing rare variants for human complex diseases can be detected. It is expected that larger reference datasets which are clean enough, more trustworthy, and of better quality, will increase usability for the case-only exome sequencing design. Efforts, including the Exome chip [79], are being implemented to provide a better reference for rare variants with functionality annotation. With further large-scale collaborations, standard WES with a case-control design may be more accessible. Detection of new predisposition genes with rare variants will require advances toward improved statistical and bioinformatics tools. After clarifying the role of rare coding variants in the etiology of complex traits, it will be necessary to investigate the role of non-coding variants using the WGS design. Suitable analysis tools are being actively developed [80–82] and will be critical for adding to our knowledge of genetic basis of complex diseases. Finally, other designs that explore functionality and improved strategies for efficient functional assays will more rapidly bring a better understanding about the biological significance of susceptibility variants and mutations [28].

Acknowledgments

This publication was made possible by CTSA Grant Number UL1 TR000135 from the National Center for Advancing Translational Sciences (NCATS), a component of the National Institutes of Health (NIH) as well as Mayo Graduate School fellowship. Its contents are solely the responsibility of the authors and do not necessarily represent the official view of NIH.

References

1. Kwon JM, Goate AM. The candidate gene approach. *Alcohol Res Health*. 2000; 24(3):164–8. [PubMed: 11199286]
2. Lunetta KL. Genetic association studies. *Circulation*. 2008; 118(1):96–101. [published Online First: Epub Date]. 10.1161/CIRCULATIONAHA.107.700401 [PubMed: 18591452]
3. Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, Devon K, Dewar K, Doyle M, FitzHugh W, Funke R, Gage D, Harris K, Heaford A, Howland J, Kann L, Lehoczky J, LeVine R, McEwan P, McKernan K, Meldrim J, Mesirov JP, Miranda C, Morris W, Naylor J, Raymond C, Rosetti M, Santos R, Sheridan A, Sougnez C, Stange-Thomann N, Stojanovic N, Subramanian A, Wyman D, Rogers J, Sulston J, Ainscough R, Beck S, Bentley D, Burton J, Clee C, Carter N, Coulson A, Deadman R, Deloukas P, Dunham A, Dunham I, Durbin R, French L, Grafham D, Gregory S, Hubbard T, Humphray S, Hunt A, Jones M, Lloyd C, McMurray A, Matthews L, Mercer S, Milne S, Mullikin JC, Mungall A, Plumb R, Ross M, Shownkeen R, Sims S, Waterston RH, Wilson RK, Hillier LW, McPherson JD, Marra MA, Mardis ER, Fulton LA, Chinwalla AT, Pepin KH, Gish WR, Chissoe SL, Wendl MC, Delehaunty KD, Miner TL, Delehaunty A, Kramer JB, Cook LL, Fulton RS, Johnson DL, Minx PJ, Clifton SW, Hawkins T, Branscomb E, Predki P, Richardson P, Wenning S, Slezak T, Doggett N, Cheng JF, Olsen A, Lucas S, Elkin C, Uberbacher E, Frazier M, Gibbs RA, Muzny DM, Scherer SE, Bouck JB, Sodergren EJ, Worley KC, Rives CM, Gorrell JH, Metzker ML, Naylor SL, Kucherlapati RS, Nelson DL, Weinstock GM, Sakaki Y, Fujiyama A, Hattori M, Yada T, Toyoda A, Itoh T, Kawagoe C, Watanabe H, Totoki Y, Taylor T, Weissenbach J, Heilig R, Saurin W, Artiguenave F, Brottier P, Bruls T, Pelletier E, Robert C, Wincker P, Smith DR, Doucette-Stamm L, Rubenfield M, Weinstock K, Lee HM, Dubois J, Rosenthal A, Platzer M, Nyakatura G, Taudien S, Rump A, Yang H, Yu J, Wang J, Huang G, Gu J, Hood L, Rowen L, Madan A, Qin S, Davis RW, Federspiel NA, Abola AP, Proctor MJ, Myers RM, Schmutz J, Dickson M, Grimwood J, Cox DR, Olson MV, Kaul R, Raymond C, Shimizu N, Kawasaki K, Minoshima S, Evans GA, Athanasiou M, Schultz R, Roe BA, Chen F, Pan H, Ramser J, Lehrach H, Reinhardt R, McCombie WR, de la Bastide M, Dedhia N, Blöcker H, Hornischer K, Nordsiek G, Agarwala R, Aravind L, Bailey JA, Bateman A, Batzoglu S, Birney E, Bork P, Brown DG, Burge CB, Cerutti L, Chen HC, Church D, Clamp M, Copley RR, Doerks T, Eddy SR, Eichler EE, Furey TS, Galagan J, Gilbert JG, Harmon C, Hayashizaki Y, Haussler D, Hermjakob H, Hokamp K, Jang W, Johnson LS, Jones TA, Kasif S, Kasprzyk A, Kennedy S, Kent WJ, Kitts P, Koonin EV, Korf I, Kulp D, Lancet D, Lowe TM, McLysaght A, Mikkelsen T, Moran JV, Mulder N, Pollara VJ, Ponting CP, Schuler G, Schultz J, Slater G, Smit AF, Stupka E, Szustakowski J, Thierry-Mieg D, Thierry-Mieg J, Wagner L, Wallis J, Wheeler R, Williams A, Wolf YI, Wolfe KH, Yang SP, Yeh RF, Collins F, Guyer MS, Peterson J, Felsenfeld A, Wetterstrand KA, Patrino A, Morgan MJ, de Jong P, Catanese JJ, Osoegawa K, Shizuya H, Choi S, Chen YJ, International Human Genome Sequencing Consortium. Initial sequencing and analysis of the human genome. *Nature*. 2001; 409(6822):860–921. [published Online First: Epub Date]. 10.1038/35057062 [PubMed: 11237011]
4. Cargill M, Altshuler D, Ireland J, Sklar P, Ardlie K, Patil N, Shaw N, Lane CR, Lim EP, Kalyanaraman N, Nemesh J, Ziaugra L, Friedland L, Rolfe A, Warrington J, Lipshutz R, Daley GQ, Lander ES. Characterization of single-nucleotide polymorphisms in coding regions of human genes. *Nat Genet*. 1999; 22(3):231–8. [published Online First: Epub Date]. 10.1038/10290 [PubMed: 10391209]
5. Bush WS, Moore JH. Chapter 11: Genome-wide association studies. *PLoS Comput Biol*. 2012; 8(12):e1002822. [published Online First: Epub Date]. 10.1371/journal.pcbi.1002822 [PubMed: 23300413]

6. Marian AJ. Molecular genetic studies of complex phenotypes. *Transl Res.* 2012; 159(2):64–79. [published Online First: Epub Date]. 10.1016/j.trsl.2011.08.001 [PubMed: 22243791]
7. Witte JS. Genome-wide association studies and beyond. *Annu Rev Public Health.* 2010; 31:9–20. 4 p following 20. [published Online First: Epub Date]. 10.1146/annurev.publhealth.012809.103723 [PubMed: 20235850]
8. Visscher PM, Brown MA, McCarthy MI, Yang J. Five years of GWAS discovery. *Am J Hum Genet.* 2012; 90(1):7–24. [published Online First: Epub Date]. 10.1016/j.ajhg.2011.11.029 [PubMed: 22243964]
9. Manolio TA, Collins FS, Cox NJ, Goldstein DB, Hindorff LA, Hunter DJ, McCarthy MI, Ramos EM, Cardon LR, Chakravarti A, Cho JH, Guttmacher AE, Kong A, Kruglyak L, Mardis E, Rotimi CN, Slatkin M, Valle D, Whittemore AS, Boehnke M, Clark AG, Eichler EE, Gibson G, Haines JL, Mackay TF, McCarroll SA, Visscher PM. Finding the missing heritability of complex diseases. *Nature.* 2009; 461(7265):747–53. [published Online First: Epub Date]. 10.1038/nature08494 [PubMed: 19812666]
10. Stranger BE, Stahl EA, Raj T. Progress and promise of genome-wide association studies for human complex trait genetics. *Genetics.* 2011; 187(2):367–83. [published Online First: Epub Date]. 10.1534/genetics.110.120907 [PubMed: 21115973]
11. Gabriel SB, Schaffner SF, Nguyen H, Moore JM, Roy J, Blumenstiel B, Higgins J, DeFelice M, Lochner A, Faggart M, Liu-Cordero SN, Rotimi C, Adeyemo A, Cooper R, Ward R, Lander ES, Daly MJ, Altshuler D. The structure of haplotype blocks in the human genome. *Science.* 2002; 296(5576):2225–9. [published Online First: Epub Date]. 10.1126/science.1069424 [PubMed: 12029063]
12. A haplotype map of the human genome. *Nature.* 2005; 437(7063):1299–320. [published Online First: Epub Date]. 10.1038/nature04226 [PubMed: 16255080]
13. Dawn Teare M, Barrett JH. Genetic linkage studies. *Lancet.* 2005; 366(9490):1036–44. [published Online First: Epub Date]. 10.1016/S0140-6736(05)67382-5 [PubMed: 16168786]
14. Teare MD. Approaches to genetic linkage analysis. *Methods Mol Biol.* 2011; 713:55–67. [published Online First: Epub Date]. 10.1007/978-1-60327-416-6-5 [PubMed: 21153611]
15. Pericak-Vance MA. Analysis of genetic linkage data for Mendelian traits. *Curr Protoc Hum Genet.* 2001 Chapter 1: Unit 14. [published Online First: Epub Date]. 10.1002/0471142905.hg0104s09
16. Jorde LB. Linkage disequilibrium and the search for complex disease genes. *Genome Res.* 2000; 10(10):1435–44. [PubMed: 11042143]
17. Botstein D, Risch N. Discovering genotypes underlying human phenotypes: past successes for mendelian disease, future approaches for complex disease. *Nat Genet.* 2003; 33(Suppl):228–37. [published Online First: Epub Date]. 10.1038/ng1090 [PubMed: 12610532]
18. Kiezun A, Garimella K, Do R, Stitzel NO, Neale BM, McLaren PJ, Gupta N, Sklar P, Sullivan PF, Moran JL, Hultman CM, Lichtenstein P, Magnusson P, Lehner T, Shugart YY, Price AL, de Bakker PI, Purcell SM, Sunyaev SR. Exome sequencing and the genetic basis of complex traits. *Nat Genet.* 2012; 44(6):623–30. [published Online First: Epub Date]. 10.1038/ng.2303 [PubMed: 22641211]
19. Schork NJ, Murray SS, Frazer KA, Topol EJ. Common vs rare allele hypotheses for complex diseases. *Curr Opin Genet Dev.* 2009; 19(3):212–9. [published Online First: Epub Date]. 10.1016/j.gde.2009.04.010 [PubMed: 19481926]
20. McCarthy MI, Abecasis GR, Cardon LR, Goldstein DB, Little J, Ioannidis JP, Hirschhorn JN. Genome-wide association studies for complex traits: consensus, uncertainty and challenges. *Nat Rev Genet.* 2008; 9(5):356–69. [published Online First: Epub Date]. 10.1038/nrg2344 [PubMed: 18398418]
21. Zhang X, Bailey SD, Lupien M. Laying a solid foundation for Manhattan—‘setting the functional basis for the post-GWAS era’. *Trends Genet.* 2014; 30(4):140–9. [published Online First: Epub Date]. 10.1016/j.tig.2014.02.006 [PubMed: 24661571]
22. Ward LD, Kellis M. Interpreting noncoding genetic variation in complex traits and human disease. *Nat Biotechnol.* 2012; 30(11):1095–106. [published Online First: Epub Date]. 10.1038/nbt.2422 [PubMed: 23138309]

23. Niedringhaus TP, Milanova D, Kerby MB, Snyder MP, Barron AE. Landscape of next-generation sequencing technologies. *Anal Chem.* 2011; 83(12):4327–41. [published Online First: Epub Date]. 10.1021/ac2010857 [PubMed: 21612267]
24. Sboner A, Mu XJ, Greenbaum D, Auerbach RK, Gerstein MB. The real cost of sequencing: higher than you think! *Genome Biol.* 2011; 12(8):125 d. [published Online First: Epub Date]. 10.1186/gb-2011-12-8-125 [PubMed: 21867570]
25. Hall N. After the gold rush. *Genome Biol.* 2013; 14(5):115 d. [published Online First: Epub Date]. 10.1186/gb-2013-14-5-115 [PubMed: 23657273]
26. Li B, Leal SM. Methods for detecting associations with rare variants for common diseases: application to analysis of sequence data. *Am J Hum Genet.* 2008; 83(3):311–21. [published Online First: Epub Date]. 10.1016/j.ajhg.2008.06.024 [PubMed: 18691683]
27. Morris AP, Zeggini E. An evaluation of statistical approaches to rare variant analysis in genetic association studies. *Genet Epidemiol.* 2010; 34(2):188–93. [published Online First: Epub Date]. 10.1002/gepi.20450 [PubMed: 19810025]
28. Do R, Kathiresan S, Abecasis GR. Exome sequencing and complex disease: practical aspects of rare variant association studies. *Hum Mol Genet.* 2012; 21(R1):R1–9. [published Online First: Epub Date]. 10.1093/hmg/dds387 [PubMed: 22983955]
29. Stitzel NO, Kiezun A, Sunyaev S. Computational and statistical approaches to analyzing variants identified by exome sequencing. *Genome Biol.* 2011; 12(9):227. [published Online First: Epub Date]. 10.1186/gb-2011-12-9-227 [PubMed: 21920052]
30. Larson NB, Schaid DJ. Regularized rare variant enrichment analysis for case-control exome sequencing data. *Genet Epidemiol.* 2014; 38(2):104–13. [published Online First: Epub Date]. 10.1002/gepi.21783 [PubMed: 24382715]
31. Kilpinen H, Barrett JC. How next-generation sequencing is transforming complex disease genetics. *Trends Genet.* 2013; 29(1):23–30. [published Online First: Epub Date]. 10.1016/j.tig.2012.10.001 [PubMed: 23103023]
32. ICGC Data Portal. Secondary ICGC Data Portal. <http://dcc.icgc.org/>.
33. The Cancer Genome Atlas Data Portal. Secondary The Cancer Genome Atlas Data Portal. <https://tcga-data.nci.nih.gov/tcga/>.
34. Genomic Data Sharing (GDS). Secondary Genomic Data Sharing (GDS). <http://gds.nih.gov/>.
35. Data use under the NIH GWAS Data Sharing Policy and future directions. *Nat Genet.* 2014; 46(9):934–8. [Published Online First: Epub Date]. 10.1038/ng.3062 [PubMed: 25162809]
36. Hendricks, AE. Whole Exome Sequencing Cases: Finding and Testing with External Controls; 62nd Annual Meeting of The American Society of Human Genetics; San Francisco, CA. 2012.
37. Isakov O, Perrone M, Shomron N. Exome sequencing analysis: a guide to disease variant detection. *Methods Mol Biol.* 2013; 1038:137–58. [published Online First: Epub Date]. 10.1007/978-1-62703-514-9_8 [PubMed: 23872973]
38. A map of human genome variation from population-scale sequencing. *Nature.* 2010; 467(7319):1061–73. [published Online First: Epub Date]. 10.1038/nature09534 [PubMed: 20981092]
39. 1000 Genomes Project Consortium. Abecasis GR, Auton A, Brooks LD, DePristo MA, Durbin RM, Handsaker RE, Kang HM, Marth GT, McVean GA. An integrated map of genetic variation from 1,092 human genomes. *Nature.* 2012; 491(7422):56–65. [published Online First: Epub Date]. 10.1038/nature11632 [PubMed: 23128226]
40. NHLBI Exome Sequencing Project (ESP) Exome Variant Server. Secondary NHLBI Exome Sequencing Project (ESP) Exome Variant Server. <http://evs.gs.washington.edu/EVS/>
41. O'Rawe J, Jiang T, Sun G, Wu Y, Wang W, Hu J, Bodily P, Tian L, Hakonarson H, Johnson WE, Wei Z, Wang K, Lyon GJ. Low concordance of multiple variant-calling pipelines: practical implications for exome and genome sequencing. *Genome Med.* 2013; 5(3):28 d. [published Online First: Epub Date]. 10.1186/gm432 [PubMed: 23537139]
42. Ng PC, Henikoff S. Predicting deleterious amino acid substitutions. *Genome Res.* 2001; 11(5):863–74. [published Online First: Epub Date]. 10.1101/gr.176601 [PubMed: 11337480]
43. Adzhubei IA, Schmidt S, Peshkin L, Ramensky VE, Gerasimova A, Bork P, Kondrashov AS, Sunyaev SR. A method and server for predicting damaging missense mutations. *Nat Methods.*

- 2010; 7(4):248–9. [published Online First: Epub Date]. 10.1038/nmeth0410-248 [PubMed: 20354512]
44. Leslie EJ, Standley J, Compton J, Bale S, Schutte BC, Murray JC. Comparative analysis of IRF6 variants in families with Van der Woude syndrome and popliteal pterygium syndrome using public whole-exome databases. *Genet Med*. 2013; 15(5):338–44. [published Online First: Epub Date]. 10.1038/gim.2012.141 [PubMed: 23154523]
 45. Lopes MC, Joyce C, Ritchie GR, John SL, Cunningham F, Asimit J, Zeggini E. A combined functional annotation score for non-synonymous variants. *Hum Hered*. 2012; 73(1):47–51. [published Online First: Epub Date]. 10.1159/000334984 [PubMed: 22261837]
 46. Liu L, Kumar S. Evolutionary balancing is critical for correctly forecasting disease-associated amino acid variants. *Mol Biol Evol*. 2013; 30(6):1252–7. [published Online First: Epub Date]. 10.1093/molbev/mst037 [PubMed: 23462317]
 47. Li B, Krishnan VG, Mort ME, Xin F, Kamati KK, Cooper DN, Mooney SD, Radivojac P. Automated inference of molecular mechanisms of disease from amino acid substitutions. *Bioinformatics*. 2009; 25(21):2744–50. [published Online First: Epub Date]. 10.1093/bioinformatics/btp528 [PubMed: 19734154]
 48. Cingolani P, Platts A, Wang le L, Coon M, Nguyen T, Wang L, Land SJ, Lu X, Ruden DM. A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of *Drosophila melanogaster* strain w1118; iso-2; iso-3. *Fly (Austin)*. 2012; 6(2):80–92. [published Online First: Epub Date]. 10.4161/fly.19695 [PubMed: 22728672]
 49. Wang M, Zhao XM, Takemoto K, Xu H, Li Y, Akutsu T, Song J. FunSAV: predicting the functional effect of single amino acid variants using a two-stage random forest model. *PLoS One*. 2012; 7(8):e43847. [published Online First: Epub Date]. 10.1371/journal.pone.0043847 [PubMed: 22937107]
 50. Chen X, Ishwaran H. Random forests for genomic data analysis. *Genomics*. 2012; 99(6):323–9. [published Online First: Epub Date]. 10.1016/j.ygeno.2012.04.003 [PubMed: 22546560]
 51. Sifrim A, Popovic D, Tranchevent LC, Ardeschirdavani A, Sakai R, Konings P, Vermeesch JR, Aerts J, De Moor B, Moreau Y. eXtasy: variant prioritization by genomic data fusion. *Nat Methods*. 2013; 10(11):1083–4. [published Online First: Epub Date]. 10.1038/nmeth.2656 [PubMed: 24076761]
 52. Bendl J, Stourac J, Salanda O, Pavelka A, Wieben ED, Zendulka J, Brezovsky J, Damborsky J. PredictSNP: robust and accurate consensus classifier for prediction of disease-related mutations. *PLoS Comput Biol*. 2014; 10(1):e1003440. [published Online First: Epub Date]. 10.1371/journal.pcbi.1003440 [PubMed: 24453961]
 53. Frousios K, Iliopoulos CS, Schlitt T, Simpson MA. Predicting the functional consequences of non-synonymous DNA sequence variants—evaluation of bioinformatics tools and development of a consensus strategy. *Genomics*. 2013; 102(4):223–8. [published Online First: Epub Date]. 10.1016/j.ygeno.2013.06.005 [PubMed: 23831115]
 54. Castellana S, Mazza T. Congruency in the prediction of pathogenic missense mutations: state-of-the-art web-based tools. *Brief Bioinform*. 2013; 14(4):448–59. [published Online First: Epub Date]. 10.1093/bib/bbt013 [PubMed: 23505257]
 55. Li MX, Kwan JS, Bao SY, Yang W, Ho SL, Song YQ, Sham PC. Predicting mendelian disease-causing non-synonymous single nucleotide variants in exome sequencing studies. *PLoS Genet*. 2013; 9(1):e1003143. [published Online First: Epub Date]. 10.1371/journal.pgen.1003143 [PubMed: 23341771]
 56. Yang Y, Chen B, Tan G, Vihinen M, Shen B. Structure-based prediction of the effects of a missense variant on protein stability. *Amino Acids*. 2013; 44(3):847–55. [published Online First: Epub Date]. 10.1007/s00726-012-1407-7 [PubMed: 23064876]
 57. Blair DR, Lyttle CS, Mortensen JM, Bearden CF, Jensen AB, Khiabani H, Melamed R, Rabadan R, Bernstam EV, Brunak S, Jensen LJ, Nicolae D, Shah NH, Grossman RL, Cox NJ, White KP, Rzhetsky A. A nondegenerate code of deleterious variants in mendelian Loci contributes to complex disease risk. *Cell*. 2013; 155(1):70–80. [published Online First: Epub Date]. 10.1016/j.cell.2013.08.030 [PubMed: 24074861]

58. Duggan S, Prichard D, Kirca M, Kelleher D. Inherited Syndromes Predisposing to Inflammation and GI Cancer. *Recent Results Cancer Res.* 2011; 185:35–50. [published Online First: Epub Date]. 10.1007/978-3-642-03503-6_2 [PubMed: 21822818]
59. Jorge AA, Malaquias AC, Arnhold IJ, Mendonca BB. Noonan syndrome and related disorders: a review of clinical features and mutations in genes of the RAS/MAPK pathway. *Horm Res.* 2009; 71(4):185–93. [published Online First: Epub Date]. 10.1159/000201106 [PubMed: 19258709]
60. Segel GB, Lichtman MA. Familial (inherited) leukemia, lymphoma, and myeloma: an overview. *Blood Cells Mol Dis.* 2004; 32(1):246–61. [PubMed: 14757442]
61. Novarino G, Fenstermaker AG, Zaki MS, Hofree M, Silhavy JL, Heiberg AD, Abdellateef M, Rosti B, Scott E, Mansour L, Masri A, Kayserili H, Al-Aama JY, Abdel-Salam GM, Karminejad A, Kara M, Kara B, Bozorgmehri B, Ben-Omran T, Mojahedi F, Mahmoud IG, Bouslam N, Bouhouche A, Benomar A, Hanein S, Raymond L, Forlani S, Mascaro M, Selim L, Shehata N, Al-Allawi N, Bindu PS, Azam M, Gunel M, Caglayan A, Bilguvar K, Tolun A, Issa MY, Schroth J, Spencer EG, Rosti RO, Akizu N, Vaux KK, Johansen A, Koh AA, Megahed H, Durr A, Brice A, Stevanin G, Gabriel SB, Ideker T, Gleeson JG. Exome sequencing links corticospinal motor neuron disease to common neurodegenerative disorders. *Science.* 2014; 343(6170):506–11. [published Online First: Epub Date]. 10.1126/science.1247363 [PubMed: 24482476]
62. Goldstein AM, Fraser MC, Struwing JP, Hussussian CJ, Ranade K, Zametkin DP, Fontaine LS, Organic SM, Dracopoli NC, Clark WH Jr. Increased risk of pancreatic cancer in melanoma-prone kindreds with p16INK4 mutations. *N Engl J Med.* 1995; 333(15):970–4. [published Online First: Epub Date]. 10.1056/NEJM199510123331504 [PubMed: 7666916]
63. Goldstein AM, Chan M, Harland M, Gillanders EM, Hayward NK, Avril MF, Azizi E, Bianchi-Scarra G, Bishop DT, Bressac-de Paillerets B, Bruno W, Calista D, Cannon Albright LA, Demeais F, Elder DE, Ghiorzo P, Gruis NA, Hansson J, Hogg D, Holland EA, Kanetsky PA, Kefford RF, Landi MT, Lang J, Leachman SA, Mackie RM, Magnusson V, Mann GJ, Niendorf K, Newton Bishop J, Palmer JM, Puig S, Puig-Butille JA, de Snoo FA, Stark M, Tsao H, Tucker MA, Whitaker L, Yakobson E, Melanoma Genetics Consortium (GenoMEL). High-risk melanoma susceptibility genes and pancreatic cancer, neural system tumors, and uveal melanoma across GenoMEL. *Cancer Res.* 2006; 66(20):9818–28. [published Online First: Epub Date]. 10.1158/0008-5472.CAN-06-0494 [PubMed: 17047042]
64. Huxley R, Ansary-Moghaddam A, Berrington de Gonzalez A, Barzi F, Woodward M. Type-II diabetes and pancreatic cancer: a meta-analysis of 36 studies. *Br J Cancer.* 2005; 92(11):2076–83. [published Online First: Epub Date]. 10.1038/sj.bjc.6602619 [PubMed: 15886696]
65. Su GH, Hruban RH, Bansal RK, Bova GS, Tang DJ, Shekher MC, Westerman AM, Entius MM, Goggins M, Yeo CJ, Kern SE. Germline and somatic mutations of the STK11/LKB1 Peutz-Jeghers gene in pancreatic and biliary cancers. *Am J Pathol.* 1999; 154(6):1835–40. [published Online First: Epub Date]. 10.1016/S0002-9440(10)65440-5 [PubMed: 10362809]
66. Ekblom A, McLaughlin JK, Nyren O. Pancreatitis and the risk of pancreatic cancer. *N Engl J Med.* 1993; 329(20):1502–3. [published Online First: Epub Date]. 10.1056/NEJM19931113292016 [PubMed: 8413467]
67. Wu L, Goldstein AM, Yu K, Yang XR, Rabe KG, Arslan AA, Canzian F, Wolpin BM, Stolzenberg-Solomon R, Amundadottir LT, Petersen GM. Variants associated with susceptibility to pancreatic cancer and melanoma do not reciprocally affect risk. *Cancer Epidemiol Biomarkers Prev.* 2014; 23(6):1121–4. [published Online First: Epub Date]. 10.1158/1055-9965.EPI-13-0627 [PubMed: 24642353]
68. Jia P, Zhao Z. VarWalker: personalized mutation network analysis of putative cancer genes from next-generation sequencing data. *PLoS Comput Biol.* 2014; 10(2):e1003460. [published Online First: Epub Date]. 10.1371/journal.pcbi.1003460 [PubMed: 24516372]
69. Carter H, Hofree M, Ideker T. Genotype to phenotype via network analysis. *Curr Opin Genet Dev.* 2013; 23(6):611–21. [published Online First: Epub Date]. 10.1016/j.gde.2013.10.003 [PubMed: 24238873]
70. Cukier P, Wright H, Rulfs T, Silveira LF, Teles MG, Mendonca BB, Arnhold IJ, Heger S, Latronico AC, Ojeda SR, Brito VN. Molecular and gene network analysis of thyroid transcription factor 1 (TTF1) and enhanced at puberty (EAP1) genes in patients with GnRH-dependent pubertal

disorders. *Horm Res Paediatr*. 2013; 80(4):257–66. [published Online First: Epub Date]. 10.1159/000354643 [PubMed: 24051510]

71. Iyer G, Al-Ahmadie H, Schultz N, Hanrahan AJ, Ostrovskaya I, Balar AV, Kim PH, Lin O, Weinhold N, Sander C, Zabor EC, Janakiraman M, Garcia-Grossman IR, Heguy A, Viale A, Bochner BH, Reuter VE, Bajorin DF, Milowsky MI, Taylor BS, Solit DB. Prevalence and co-occurrence of actionable genomic alterations in high-grade bladder cancer. *J Clin Oncol*. 2013; 31(25):3133–40. [published Online First: Epub Date]. 10.1200/JCO.2012.46.5740 [PubMed: 23897969]
72. Gu Y, Wang H, Qin Y, Zhang Y, Zhao W, Qi L, Zhang Y, Wang C, Guo Z. Network analysis of genomic alteration profiles reveals co-altered functional modules and driver genes for glioblastoma. *Mol Biosyst*. 2013; 9(3):467–77. [published Online First: Epub Date]. 10.1039/c2mb25528f [PubMed: 23344900]
73. O’Roak BJ, Vives L, Girirajan S, Karakoc E, Krumm N, Coe BP, Levy R, Ko A, Lee C, Smith JD, Turner EH, Stanaway IB, Vernot B, Malig M, Baker C, Reilly B, Akey JM, Borenstein E, Rieder MJ, Nickerson DA, Bernier R, Shendure J, Eichler EE. Sporadic autism exomes reveal a highly interconnected protein network of de novo mutations. *Nature*. 2012; 485(7397):246–50. [published Online First: Epub Date]. 10.1038/nature10989 [PubMed: 22495309]
74. Sims D, Sudbery I, Illott NE, Heger A, Ponting CP. Sequencing depth and coverage: key considerations in genomic analyses. *Nat Rev Genet*. 2014; 15(2):121–32. [published Online First: Epub Date]. 10.1038/nrg3642 [PubMed: 24434847]
75. Ciriello G, Miller ML, Aksoy BA, Senbabaoglu Y, Schultz N, Sander C. Emerging landscape of oncogenic signatures across human cancers. *Nat Genet*. 2013; 45(10):1127–33. [published Online First: Epub Date]. 10.1038/ng.2762 [PubMed: 24071851]
76. International Cancer Genome Consortium. Hudson TJ, Anderson W, Artez A, Barker AD, Bell C, Bernabé RR, Bhan MK, Calvo F, Eerola I, Gerhard DS, Guttmacher A, Guyer M, Hemsley FM, Jennings JL, Kerr D, Klatt P, Kolar P, Kusada J, Lane DP, Laplace F, Youyong L, Nettekoven G, Ozenberger B, Peterson J, Rao TS, Remacle J, Schafer AJ, Shibata T, Stratton MR, Vockley JG, Watanabe K, Yang H, Yuen MM, Knoppers BM, Bobrow M, Cambon-Thomsen A, Dressler LG, Dyke SO, Joly Y, Kato K, Kennedy KL, Nicolás P, Parker MJ, Rial-Sebbag E, Romeo-Casabona CM, Shaw KM, Wallace S, Wiesner GL, Zeps N, Lichter P, Biankin AV, Chabannon C, Chin L, Clément B, de Alava E, Degos F, Ferguson ML, Geary P, Hayes DN, Hudson TJ, Johns AL, Kasprzyk A, Nakagawa H, Penny R, Piris MA, Sarin R, Scarpa A, Shibata T, van de Vijver M, Futreal PA, Aburatani H, Bayés M, Botwell DD, Campbell PJ, Estivill X, Gerhard DS, Grimmond SM, Gut I, Hirst M, López-Otín C, Majumder P, Marra M, McPherson JD, Nakagawa H, Ning Z, Puente XS, Ruan Y, Shibata T, Stratton MR, Stunnenberg HG, Swerdlow H, Velculescu VE, Wilson RK, Xue HH, Yang L, Spellman PT, Bader GD, Boutros PC, Campbell PJ, Flicek P, Getz G, Guigó R, Guo G, Haussler D, Heath S, Hubbard TJ, Jiang T, Jones SM, Li Q, López-Bigas N, Luo R, Muthuswamy L, Ouellette BF, Pearson JV, Puente XS, Quesada V, Raphael BJ, Sander C, Shibata T, Speed TP, Stein LD, Stuart JM, Teague JW, Totoki Y, Tsunoda T, Valencia A, Wheeler DA, Wu H, Zhao S, Zhou G, Stein LD, Guigó R, Hubbard TJ, Joly Y, Jones SM, Kasprzyk A, Lathrop M, López-Bigas N, Ouellette BF, Spellman PT, Teague JW, Thomas G, Valencia A, Yoshida T, Kennedy KL, Axton M, Dyke SO, Futreal PA, Gerhard DS, Gunter C, Guyer M, Hudson TJ, McPherson JD, Miller LJ, Ozenberger B, Shaw KM, Kasprzyk A, Stein LD, Zhang J, Haider SA, Wang J, Yung CK, Cros A, Liang Y, Gnaneshan S, Guberman J, Hsu J, Bobrow M, Chalmers DR, Hasel KW, Joly Y, Kaan TS, Kennedy KL, Knoppers BM, Lowrance WW, Masui T, Nicolás P, Rial-Sebbag E, Rodriguez LL, Vergely C, Yoshida T, Grimmond SM, Biankin AV, Bowtell DD, Cloonan N, deFazio A, Eshleman JR, Etemadmoghadam D, Gardiner BB, Kench JG, Scarpa A, Sutherland RL, Tempero MA, Waddell NJ, Wilson PJ, McPherson JD, Gallinger S, Tsao MS, Shaw PA, Petersen GM, Mukhopadhyay D, Chin L, DePinho RA, Thayer S, Muthuswamy L, Shazand K, Beck T, Sam M, Timms L, Ballin V, Lu Y, Ji J, Zhang X, Chen F, Hu X, Zhou G, Yang Q, Tian G, Zhang L, Xing X, Li X, Zhu Z, Yu Y, Yu J, Yang H, Lathrop M, Tost J, Brennan P, Holcatova I, Zaridze D, Brazma A, Egevard L, Prokhorchouk E, Banks RE, Uhlén M, Cambon-Thomsen A, Viksna J, Ponten F, Skryabin K, Stratton MR, Futreal PA, Birney E, Borg A, Børresen-Dale AL, Caldas C, Foekens JA, Martin S, Reis-Filho JS, Richardson AL, Sotiropoulos C, Stunnenberg HG, Thoms G, van de Vijver M, van’t Veer L, Calvo F, Birnbaum D, Blanche H, Boucher P, Boyault S, Chabannon C, Gut I, Masson-Jacquemier JD, Lathrop M,

- Pauporté I, Pivot X, Vincent-Salomon A, Tabone E, Theillet C, Thomas G, Tost J, Treilleux I, Calvo F, Bioulac-Sage P, Clément B, Decaens T, Degos F, Franco D, Gut I, Gut M, Heath S, Lathrop M, Samuel D, Thomas G, Zucman-Rossi J, Lichter P, Eils R, Brors B, Korbel JO, Korshunov A, Landgraf P, Lehrach H, Pfister S, Radlwimmer B, Reifemberger G, Taylor MD, von Kalle C, Majumder PP, Sarin R, Rao TS, Bhan MK, Scarpa A, Pederzoli P, Lawlor RA, Delledonne M, Bardelli A, Biankin AV, Grimmond SM, Gress T, Klimstra D, Zamboni G, Shibata T, Nakamura Y, Nakagawa H, Kusada J, Tsunoda T, Miyano S, Aburatani H, Kato K, Fujimoto A, Yoshida T, Campo E, López-Otín C, Estivill X, Guigó R, de Sanjosé S, Piris MA, Montserrat E, González-Díaz M, Puente XS, Jares P, Valencia A, Himmelbauer H, Quesada V, Bea S, Stratton MR, Futreal PA, Campbell PJ, Vincent-Salomon A, Richardson AL, Reis-Filho JS, van de Vijver M, Thomas G, Masson-Jacquemier JD, Aparicio S, Borg A, Børresen-Dale AL, Caldas C, Foekens JA, Stunnenberg HG, van't Veer L, Easton DF, Spellman PT, Martin S, Barker AD, Chin L, Collins FS, Compton CC, Ferguson ML, Gerhard DS, Getz G, Gunter C, Guttmacher A, Guyer M, Hayes DN, Lander ES, Ozenberger B, Penny R, Peterson J, Sander C, Shaw KM, Speed TP, Spellman PT, Vockley JG, Wheeler DA, Wilson RK, Hudson TJ, Chin L, Knoppers BM, Lander ES, Lichter P, Stein LD, Stratton MR, Anderson W, Barker AD, Bell C, Bobrow M, Burke W, Collins FS, Compton CC, DePinho RA, Easton DF, Futreal PA, Gerhard DS, Green AR, Guyer M, Hamilton SR, Hubbard TJ, Kallioniemi OP, Kennedy KL, Ley TJ, Liu ET, Lu Y, Majumder P, Marra M, Ozenberger B, Peterson J, Schafer AJ, Spellman PT, Stunnenberg HG, Wainwright BJ, Wilson RK, Yang H. International network of cancer genome projects. *Nature*. 2010; 464(7291): 993–8. [published Online First: Epub Date]. 10.1038/nature08987 [PubMed: 20393554]
77. Dering C, Hemmelmann C, Pugh E, Ziegler A. Statistical analysis of rare sequence variants: an overview of collapsing methods. *Genet Epidemiol*. 2011; 35(Suppl 1):S12–7. [published Online First: Epub Date]. 10.1002/gepi.20643 [PubMed: 22128052]
78. Lee S, Abecasis GR, Boehnke M, Lin X. Rare-variant association analysis: study designs and statistical tests. *Am J Hum Genet*. 2014; 95(1):5–23. [published Online First: Epub Date]. 10.1016/j.ajhg.2014.06.009 [PubMed: 24995866]
79. Exome Chip Design. Exome Chip Design wikipedia. 2013
80. Khurana E, Fu Y, Colonna V, Mu XJ, Kang HM, Lappalainen T, Sboner A, Lochovsky L, Chen J, Harmanci A, Das J, Abyzov A, Balasubramanian S, Beal K, Chakravarty D, Challis D, Chen Y, Clarke D, Clarke L, Cunningham F, Evani US, Flicek P, Fragoza R, Garrison E, Gibbs R, Gümüs ZH, Herrero J, Kitabayashi N, Kong Y, Lage K, Liliushvili V, Lipkin SM, MacArthur DG, Marth G, Muzny D, Pers TH, Ritchie GR, Rosenfeld JA, Sisu C, Wei X, Wilson M, Xue Y, Yu F, 1000 Genomes Project Consortium. Dermitzakis ET, Yu H, Rubin MA, Tyler-Smith C, Gerstein M. Integrative annotation of variants from 1092 humans: application to cancer genomics. *Science*. 2013; 342(6154):1235587. [published Online First: Epub Date]. 10.1126/science.1235587 [PubMed: 24092746]
81. Torkamani A, Schork NJ. Predicting functional regulatory polymorphisms. *Bioinformatics*. 2008; 24(16):1787–92. [published Online First: Epub Date]. 10.1093/bioinformatics/btn311 [PubMed: 18562267]
82. Torkamani A, Scott-Van Zeeland AA, Topol EJ, Schork NJ. Annotating individual human genomes. *Genomics*. 2011; 98(4):233–41. [published Online First: Epub Date]. 10.1016/j.ygeno.2011.07.006 [PubMed: 21839162]
83. Palles C, Cazier JB, Howarth KM, Domingo E, Jones AM, Broderick P, Kemp Z, Spain SL, Guarino E, Salguero I, Sherborne A, Chubb D, Carvajal-Carmona LG, Ma Y, Kaur K, Dobbins S, Barclay E, Gorman M, Martin L, Kovac MB, Humphray S, CORGI Consortium; WGS500 Consortium. Lucassen A, Holmes CC, Bentley D, Donnelly P, Taylor J, Petridis C, Roylance R, Sawyer EJ, Kerr DJ, Clark S, Grimes J, Kearsley SE, Thomas HJ, McVean G, Houlston RS, Tomlinson I. Germline mutations affecting the proofreading domains of POLE and POLD1 predispose to colorectal adenomas and carcinomas. *Nat Genet*. 2013; 45(2):136–44. [published Online First: Epub Date]. 10.1038/ng.2503 [PubMed: 23263490]
84. Roberts NJ, Jiao Y, Yu J, Kopelovich L, Petersen GM, Bondy ML, Gallinger S, Schwartz AG, Syngal S, Cote ML, Axilbund J, Schulick R, Ali SZ, Eshleman JR, Velculescu VE, Goggins M, Vogelstein B, Papadopoulos N, Hruban RH, Kinzler KW, Klein AP. (2012) ATM mutations in patients with hereditary pancreatic cancer. *Cancer Discov*. 2012; 2(1):41–6. [published Online First: Epub Date]. 10.1158/2159-8290.CD-11-0194 [PubMed: 22585167]

85. Neale BM, Kou Y, Liu L, Ma'ayan A, Samocha KE, Sabo A, Lin CF, Stevens C, Wang LS, Makarov V, Polak P, Yoon S, Maguire J, Crawford EL, Campbell NG, Geller ET, Valladares O, Schafer C, Liu H, Zhao T, Cai G, Lihm J, Dannenfelser R, Jabado O, Peralta Z, Nagaswamy U, Muzny D, Reid JG, Newsham I, Wu Y, Lewis L, Han Y, Voight BF, Lim E, Rossin E, Kirby A, Flannick J, Fromer M, Shakir K, Fennell T, Garimella K, Banks E, Poplin R, Gabriel S, DePristo M, Wimbish JR, Boone BE, Levy SE, Betancur C, Sunyaev S, Boerwinkle E, Buxbaum JD, Cook EH Jr, Devlin B, Gibbs RA, Roeder K, Schellenberg GD, Sutcliffe JS, Daly MJ. Patterns and rates of exonic de novo mutations in autism spectrum disorders. *Nature*. 2012; 485(7397):242–5. [published Online First: Epub Date]. 10.1038/nature11011 [PubMed: 22495311]
86. Park DJ, Lesueur F, Nguyen-Dumont T, Pertesi M, Odefrey F, Hammet F, Neuhausen SL, John EM, Andrulis IL, Terry MB, Daly M, Buys S, Le Calvez-Kelm F, Lonie A, Pope BJ, Tsimiklis H, Voegelé C, Hilbers FM, Hoogerbrugge N, Barroso A, Osorio A, Breast Cancer Family Registry; Kathleen Cuninghame Foundation Consortium for Research into Familial Breast Cancer. Giles GG, Devilee P, Benitez J, Hopper JL, Tavtigian SV, Goldgar DE, Southey MC. Rare mutations in XRCC2 increase the risk of breast cancer. *Am J Hum Genet*. 2012; 90(4):734–9. [published Online First: Epub Date]. 10.1016/j.ajhg.2012.02.027 [PubMed: 22464251]
87. Thompson ER, Doyle MA, Ryland GL, Rowley SM, Choong DY, Tothill RW, Thorne H, kConFab, Barnes DR, Li J, Ellul J, Philip GK, Antill YC, James PA, Trainer AH, Mitchell G, Campbell IG. Exome sequencing identifies rare deleterious mutations in DNA repair genes FANCC and BLM as potential breast cancer susceptibility alleles. *PLoS Genet*. 2012; 8(9):e1002894. [published Online First: Epub Date]. 10.1371/journal.pgen.1002894 [PubMed: 23028338]
88. Hoehe MR, Kopke K, Wendel B, Rohde K, Flachmeier C, Kidd KK, Berrettini WH, Church GM. Sequence variability and candidate gene analysis in complex disease: association of mu opioid receptor gene variation with substance dependence. *Hum Mol Genet*. 2000; 9(19):2895–908. [PubMed: 11092766]
89. Chien WH, Gau SS, Liao HM, Chiu YN, Wu YY, Huang YS, Tsai WC, Tsai HM, Chen CH. Deep exon resequencing of DLGAP2 as a candidate gene of autism spectrum disorders. *Mol Autism*. 2013; 4(1):26. [published Online First: Epub Date]. 10.1186/2040-2392-4-26 [PubMed: 23915500]
90. Beaudoin M, Lo KS, N'Diaye A, Rivas MA, Dubé MP, Laplante N, Phillips MS, Rioux JD, Tardif JC, Lettre G. Pooled DNA resequencing of 68 myocardial infarction candidate genes in French Canadians. *Circ Cardiovasc Genet*. 2012; 5(5):547–54. [published Online First: Epub Date]. 10.1161/CIRCGENETICS.112.963165 [PubMed: 22923420]
91. Lin H, Sinner MF, Brody JA, Arking DE, Lunetta KL, Rienstra M, Lubitz SA, Magnani JW, Sotoodehnia N, McKnight B, McManus DD, Boerwinkle E, Psaty BM, Rotter JJ, Bis JC, Gibbs RA, Muzny D, Kovar CL, Morrison AC, Gupta M, Folsom AR, Kääb S, Heckbert SR, Alonso A, Ellinor PT, Benjamin EJ, CHARGE Atrial Fibrillation Working Group. Targeted sequencing in candidate genes for atrial fibrillation: the Cohorts for Heart and Aging Research in Genomic Epidemiology (CHARGE) Targeted Sequencing Study. *Heart Rhythm*. 2013 [published Online First: Epub Date]. 10.1016/j.hrthm.2013.11.012
92. Li JM, Lu CL, Cheng MC, Luu SU, Hsu SH, Chen CH. Exonic resequencing of the DLGAP3 gene as a candidate gene for schizophrenia. *Psychiatry Res*. 2013; 208(1):84–7. [published Online First: Epub Date]. 10.1016/j.psychres.2012.12.015 [PubMed: 23414653]
93. Tang H, Jin X, Li Y, Jiang H, Tang X, Yang X, Cheng H, Qiu Y, Chen G, Mei J, Zhou F, Wu R, Zuo X, Zhang Y, Zheng X, Cai Q, Yin X, Quan C, Shao H, Cui Y, Tian F, Zhao X, Liu H, Xiao F, Xu F, Han J, Shi D, Zhang A, Zhou C, Li Q, Fan X, Lin L, Tian H, Wang Z, Fu H, Wang F, Yang B, Huang S, Liang B, Xie X, Ren Y, Gu Q, Wen G, Sun Y, Wu X, Dang L, Xia M, Shan J, Li T, Yang L, Zhang X, Li Y, He C, Xu A, Wei L, Zhao X, Gao X, Xu J, Zhang F, Zhang J, Li Y, Sun L, Liu J, Chen R, Yang S, Wang J, Zhang X. A large-scale screen for coding variants predisposing to psoriasis. *Nat Genet*. 2014; 46(1):45–50. [published Online First: Epub Date]. 10.1038/ng.2827 [PubMed: 24212883]
94. Ellinghaus D, Zhang H, Zeissig S, Lipinski S, Till A, Jiang T, Stade B, Bromberg Y, Ellinghaus E, Keller A, Rivas MA, Skieceviciene J, Doncheva NT, Liu X, Liu Q, Jiang F, Forster M, Mayr G, Albrecht M, Häslér R, Boehm BO, Goodall J, Berzuini CR, Lee J, Andersen V, Vogel U, Kupcinskis L, Kayser M, Krawczak M, Nikolaus S, Weersma RK, Ponsioen CY, Sans M,

Wijmenga C, Strachan DP, McArdle WL, Vermeire S, Rutgeerts P, Sanderson JD, Mathew CG, Vatn MH, Wang J, Nöthen MM, Duerr RH, Büning C, Brand S, Glas J, Winkelmann J, Illig T, Latiano A, Annese V, Halfvarson J, D'Amato M, Daly MJ, Nothnagel M, Karlsen TH, Subramani S, Rosenstiel P, Schreiber S, Parkes M, Franke A. Association between variants of PRDM1 and NDP52 and Crohn's disease, based on exome sequencing and functional studies. *Gastroenterology*. 2013; 145(2):339–47. [published Online First: Epub Date]. 10.1053/j.gastro.2013.04.040 [PubMed: 23624108]

95. Liu L, Sabo A, Neale BM, Nagaswamy U, Stevens C, Lim E, Bodea CA, Muzny D, Reid JG, Banks E, Coon H, Depristo M, Dinh H, Fennel T, Flannick J, Gabriel S, Garimella K, Gross S, Hawes A, Lewis L, Makarov V, Maguire J, Newsham I, Poplin R, Ripke S, Shakir K, Samocha KE, Wu Y, Boerwinkle E, Buxbaum JD, Cook EH Jr, Devlin B, Schellenberg GD, Sutcliffe JS, Daly MJ, Gibbs RA, Roeder K. Analysis of rare, exonic variation amongst subjects with autism spectrum disorders and population controls. *PLoS Genet*. 2013; 9(4):e1003443. [published Online First: Epub Date]. 10.1371/journal.pgen.1003443 [PubMed: 23593035]
96. Siemiatkowska AM, van den Born LI, van Hagen PM, Stoffels M, Neveling K, Henkes A, Kipping-Geertsema M, Hoefsloot LH, Hoyng CB, Simon A, den Hollander AI, Cremers FP, Collin RW. Mutations in the Mevalonate Kinase (MVK) Gene Cause Nonsyndromic Retinitis Pigmentosa. *Ophthalmology*. 2013; 120(12):2697–705. [published Online First: Epub Date]. 10.1016/j.ophtha.2013.07.052 [PubMed: 24084495]

Figure 1.
Proposed strategies to filter and prioritize candidate disease susceptibility variant/gene utilizing exome sequencing data

Table 1

Study designs, advantages, and disadvantages for discovery of rare variants in predisposition genes. Study design IV is the focus of this review.

Type of Study (approximate sample size)	Main Study Design	Advantages	Disadvantages	Examples
I. Family-based WGS/WES (<200 pedigrees)	WGS/WES of affected and healthy members within pedigrees (may be combined with linkage), utilizing individual phenotype information for prioritizing candidate variants/mutations	The phenotype aggregation within pedigrees greatly increases the overall power of detecting predisposition genes	Predisposition genes found in one or more pedigrees may not be present in other pedigrees. Identifying predisposition regions for sporadic disease is limited. The success rate for this approach in complex disease is much lower than for Mendelian diseases	Palles et al.[83], Roberts et al.[84], Neale et al.[85], Park et al. [86], Thompson et al. [87]
II. Case-control candidate gene deep sequencing (125–500 pairs)	Sequence candidate genes in a large sample of cases and controls (candidate genes based on prior biological or clinical knowledge)	Relatively mature statistical methods available, large sample sizes are feasible, cost typically is affordable, power is more than that of case-control WGS/WES	Potential predisposing loci other than hypothesized genes are not evaluated, limiting novel gene discovery	Hoehe et al.[88], Chien et al.[89], Beaudoin et al.[90], Lin et al.[91], Li et al.[92]
III. Case-control WGS/WES (500–1000 pairs)	WGS/WES of cases and controls	The agnostic nature of WGS/WES makes new, previously unknown loci potentially identifiable	Cost remains very high, sample size may be limited. The statistical power of this design is poor	Tang et al.[93], Ellinghaus et al.[94], Liu et al.[95], Siemiatkowska et al. [96]
IV. Case-only exome sequencing (100–1000)	WES of only affected cases (familial or sporadic)	Cost can be decreased compared to sequencing both cases and controls	More difficult to identify new disease susceptibility loci, due in part to unavailable data for controls by same sequencing procedures	