



# Unravelling the genetic architecture of human complex traits through whole genome sequencing

Ozvan Bocher, Cristen J. Willer & Eleftheria Zeggini



Whole genome sequencing has enabled new insights into the genetic architecture of complex traits, especially through access to low-frequency and rare variation. This *Comment* highlights the key contributions from this technology and discusses considerations for its use and future perspectives.

The field of human complex trait genetics has been enriched by high-throughput whole genome sequencing (WGS) technologies. WGS complements array-based genotyping by offering the opportunity to access most sequence in the genome and not only a set of known genetic variants. As sequencing costs gradually drop, an increasing number of study designs involve WGS approaches. Large sequencing projects have been undertaken in the general population that can be used as reference panels for genotype imputation in association studies, such as the Haplotype Reference Consortium (HRC<sup>1</sup>) project. Further initiatives, such as UKBioBank<sup>2</sup> and TOPMed<sup>3</sup>, also make use of WGS technologies to sequence thousands of phenotypically-diverse individuals, providing resources of unprecedented scale to study the genetic architecture underlying complex diseases and traits<sup>4</sup>. Here, we comment on the progress and successes in the field heralded through WGS, as well as on future perspectives of this technology.

## Advantages of WGS

**Access to rare variation.** The main advantage of using WGS approaches is direct access to genetic variation across the whole frequency spectrum, without first knowing where such variation occurs, as is required for array-based genotyping. Low frequency and rare variants are often not imputed accurately from reference panels (which are also frequently not a perfect population match), but can be detected by WGS. WGS offers more accurate and complete information capture of rare variation observed in sequenced individuals, their family members, and others with shared ancestry<sup>5</sup>. WGS-based studies have reported associations with rare variants of large effect size. These associations have been described in case-control studies, for example, in the TOPMed project where Zhao et al. identified a rare variant with a large effect on reduced lung function<sup>6</sup>, as well as in quantitative trait studies, for example, in the study by Benonisdottir et al., which identified nine rare variants associated with urinary biomarkers<sup>7</sup>. Detecting single-point rare variant associations requires very large sample sizes, especially if effect sizes are not large. To maximise the chance to detect rare variants associated with complex diseases and to consider genetic heterogeneity between individuals, rare variant association tests (RVAT) have been developed. These methods have enabled the

detection of associations between medically relevant traits and an accumulation of rare variants in a chromosomal region, typically a single gene. For example, Gilly et al. described a cardioprotective rare variant burden in the *APOC3* gene, composed of exonic and splice variants, which could not be detected using imputation of genotype data<sup>8</sup>. In a larger study, by applying gene-based burden tests in over 17,000 binary phenotypes, Wang et al. identified over 1,700 significant associations, highlighting the importance of rare genetic variation in complex diseases<sup>9</sup>. Followed by functional investigation, these findings are bringing new insights into the biological mechanisms behind complex diseases. Nevertheless, biological interpretation can be more readily reached for the exome, on which the majority of RVAT have currently been applied.

Despite the description of several associations with rare variants, it is still unclear how much they contribute to the heritability of human complex traits. This proportion is especially hard to estimate for rare variants as they correspond to observations only in a few individuals resulting in high standard errors<sup>10</sup>. As common variants are present in more individuals, it is expected that they will contribute more to the phenotypic variance than rare variants. Apart from a few examples such as height<sup>11</sup> or type 2 diabetes<sup>12</sup>, several studies have indeed shown that complex trait heritability due to rare variants is expected to be rather low. For example, by looking at 22 common traits, Weiner et al. showed that rare coding variants explain on average only 1.3% of the overall phenotypic variance, ranging from 0.4% for asthma to 3.6% for height<sup>13</sup>. While rare variants will unlikely explain all the remaining phenotypic variability of complex traits, they can also be useful for prediction<sup>14</sup>. For instance, a study to predict haemoglobin A1C levels showed that the integration of many rare variants into prediction scores could lead to the identification of a substantial number of undiagnosed type 2 diabetes cases<sup>15</sup>.

**Ancestry-diverse studies.** A further important benefit of WGS is the investigation of under-represented populations that have not been well characterised by currently available sequencing data, in which rare or population-enriched variation is therefore not accurately described<sup>16</sup>. For example, using genotyping and low-depth sequencing in 6,400 individuals from the Uganda population, numerous associations were identified with complex traits, including both novel findings and associations at previously reported loci but with different allelic effects<sup>17</sup>. Similarly, sequencing followed by imputation of the Icelandic population has resulted in novel insights, including an association between a splice variant in *RPL3L* and atrial fibrillation<sup>18</sup>. Considering sub-populations within Europeans is also of interest: in Norwegians, low-depth sequencing followed by a custom genotyping array performed on 70,000 individuals resulted in new associations, e.g., between *ZNF529* p.K405X and LDL-C<sup>19</sup>. A GWAS performed in the

Finnish population on 1,932 phenotypes found 2,491 significant associations, including newly associated variants that could be identified due to their higher frequency in the Finnish population. For example, an intronic variant of *TNRC18* strongly associated with IBD but almost absent from other European populations<sup>20</sup>. WGS in diverse populations represents one of the most active areas of research, as getting an overview of the genetic architecture in diverse populations will enable better comprehension of complex diseases as well as the differences in effect direction and sizes of the associated variants that are observed across populations<sup>10</sup>.

## Considerations in WGS-based studies

**Challenges in study design.** Genotype-based GWAS is an established field where power has been shown to clearly depend on the sample size and the detectable genetic effect<sup>21</sup>, however, planning the design of a sequencing-based study can be more challenging. Li et al. showed that power indeed also depends on read depth and distribution<sup>22</sup>. In addition, the power of RVAT is less straightforward to estimate compared to single-point association analysis, because it depends on additional parameters such as the filtering strategy used to select qualifying variants and their directions of effect. Power is a major driver of the

success of a study, and multiple software packages are available to estimate the expected power of WGS-based studies, as reviewed in Li et al.<sup>22</sup>. Conducting studies in diverse populations can provide useful insights into the genetics of complex diseases. Furthermore, power to detect associations can be boosted by studying isolated populations, in which variant frequencies and effect sizes may be larger<sup>8,18,20</sup>. As the majority of WGS projects to date have been focused on European populations, conducting sequencing-based studies in under-represented populations is expected to be of benefit<sup>10</sup> and represents an important direction for future WGS applications.

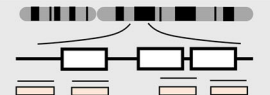

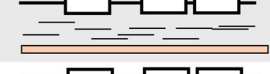




**Determination of WGS approach.** WGS-based studies can take various forms, for which the optimal choice will depend on several parameters including the population under study, the biological hypothesis investigated and the computational and financial resources available, with the cost of WGS being its largest disadvantage. These forms include WGS coupled to imputation in genome-wide association studies, cohort-wide low- or very low-depth WGS, and deep WGS (Box 1). When the interest is in related samples or under-represented populations, low and very-low depth WGS approaches may be relatively more efficient compared to when examining well-studied

## BOX 1

### Overview of the different sequencing techniques currently available

White boxes correspond to coding exons and thin black lines to sequencing reads. The sequencing depth is represented at the bottom

of each graphic by brown shades. Pros and cons of the different depths and genome coverage are highlighted.

Sequencing Strategy	Pros	Cons
 <p>Ultra low depth &lt;0.5x</p>	Can use off-target reads from exome sequencing to provide frequencies in understudied populations	Can't reliably provide individual-level genotypes
 <p>Very low depth 0.5-2x</p>	Works best in related samples	Heavy computational requirements
 <p>Low depth 2-8x</p>	Can provide haplotypes for imputation and variant-site discovery	Requires computing and imputation to estimate individual-level calls
 <p>Medium depth (8-25x)</p>	Reasonably high genotype accuracy and capture of rare variants relative to cost	Incomplete capture of indels and very rare variants (e.g., singletons)
 <p>Exome (&gt;20x at target)</p>	Highly accurate genotype calls in covered coding regions - the most directly interpretable portion of the genome	Little coverage outside of coding regions, small fraction of genes missed, harder to call indels and structural variants. Some challenges comparing across capture methods
 <p>High depth (&gt;25x)</p>	Best coverage of sequencable genome, rare variants, and high accuracy of genotype calls	More expensive, doesn't detect structural variants (as previous approaches)
 <p>Long-read</p>	Detection of structural variants, better assembly, mapping and phasing	Data processing, most expensive approach

populations. For example, Tran et al. performed low-depth sequencing to genetically describe the Vietnamese population and reported five disease-associated pathogenic variants with higher allelic frequencies than in other populations<sup>23</sup>. Low-depth sequencing has indeed been shown to be more efficient than classical imputation-coupled array designs in detecting GWAS signals, primarily due to a more complete assessment of genomic variation, especially in ancestries with poor coverage in existing imputation panels<sup>24</sup>. This was illustrated in a study from Gilly et al. in an isolated Greek population where twice as many variants were detected using very low-depth WGS as compared to classical imputed array genotyping data, and a vast majority of which were rare variants, leading to a twofold increase in the number of association signals<sup>25</sup>. Nevertheless, low-depth WGS shows decreased accuracy when studying rare variation (frequency lower than 1%) in the genome compared to low-frequency (frequency between 1 and 5%) and common variants. To identify such variation, medium-depth designs can be applied, or high-depth sequencing for detecting indels and ultra-rare variation, such as singletons, with high accuracy<sup>26</sup>. While high-depth WGS has proven to be useful, for example, in the study by Wessel et al., which described the contribution of rare non-coding variants to type 2 diabetes<sup>12</sup>, it remains expensive, especially for large cohorts. A solution to perform high-depth sequencing at a lower cost would be to focus on coding parts of the genome by using whole exome sequencing (WES). The lower cost associated with WES would enable the inclusion of more individuals in the study and therefore an increase in the power to detect genetic variants that reside in genes and are associated with human complex traits, as illustrated in the study by Wang et al.<sup>9</sup>. Nevertheless, using WES instead of WGS misses genetic variation in the non-coding genome (or any gene with poor coverage in whole exome studies), which has been shown to play an important role in complex diseases<sup>27</sup>. Finally, emerging technologies such as long-read sequencing offer the possibility to access genome-wide structural variants which have been found to have an impact on complex phenotypes as highlighted by Beyter et al. on LDL cholesterol levels and height<sup>28</sup>. This approach provides additional advantages, such as easier assembly and mapping of genomes, but remains the most expensive sequencing technology, preventing its use in large cohorts.

**Disadvantages compared to genotype-based studies.** Despite recent progress, the cost of WGS remains prohibitive for most large-scale studies. Genotyping coupled to imputation can retrieve most common variation in the genome<sup>29</sup>. The use of cost-efficient array-based technologies enables increasing sample sizes, which in turns results in the identification of further associations with common and low-frequency variants. One striking example is the study by Yengo et al. on over 5 million individuals, which has described all of the genetic heritability of height due to common variants<sup>30</sup>. The contribution of common variants to the genetic architecture of human complex traits is still not fully understood and array-based technologies will continue to be useful in filling this gap. In addition, while sequencing will continue to contribute to obtaining the whole picture of the genetic architecture of complex traits, it is likely that translation into the clinic to screen for polygenic risk will focus on array-genotyping approaches rather than on sequencing at first.

## Perspectives and conclusion

WGS has made an important contribution to the understanding of genetics underlying complex traits, especially in under-represented

populations, and through rare variation. Functional interpretation of association signals arising from WGS remains more challenging in the non-coding genome compared to the exome. Even if single-point associations have been described with rare and common variants in these regions, it is still arduous to biologically characterise these association signals. Combining association results from WGS with functional information at multiple levels, using, for example, other omics data such as transcriptomics, open chromatin, methylation, metabolomics or proteomics, has been shown to help in the interpretation of the associated signals<sup>31</sup>. Similarly, using RVAT in non-coding regions of the genome is not straightforward, despite affording higher power to detect genetic associations with rare variants<sup>32</sup>. Novel statistical methods are therefore needed which, for example, consider functional information across the non-coding genome<sup>33,34</sup>. WGS studies will remain useful in the future as a tool to explore the genetic underpinning of complex diseases, especially in combination with emerging functional data and their integration at multiple levels. As the cost of WGS is dropping and given the exciting prospect of long-read WGS at scale, these technologies will become increasingly accessible and will enable the description of genetic variation in hitherto understudied populations. As our understanding of the non-coding genome continues to improve, and with the further development of powerful methods to integrate functional information in rare variant association testing approaches, WGS will hopefully lead to a better and more accurate comprehension of complex diseases. In the future, it is anticipated that WGS-informed clinical decisions and interventions will accelerate personalised medicine in the wider field of complex diseases, following recent successes in cancer and rare disease, such as monogenic forms of cardiomyopathy<sup>35,36</sup>. To achieve these goals in a globally equitable fashion, WGS of diverse populations should remain a high priority going forward.

**Ozvan Bocher<sup>1</sup>, Cristen J. Willer<sup>2,3,4</sup> & Eleftheria Zeggini<sup>1,5</sup>** ✉

<sup>1</sup>Institute of Translational Genomics, Helmholtz Zentrum München, German Research Center for Environmental Health, 85764 Neuherberg, Germany. <sup>2</sup>Department of Internal Medicine, Division of Cardiology, University of Michigan, Ann Arbor, MI 48109, USA. <sup>3</sup>Department of Human Genetics, University of Michigan, Ann Arbor, MI 48109, USA. <sup>4</sup>Department of Computational Medicine and Biostatistics, University of Michigan, Ann Arbor, MI 48109, USA. <sup>5</sup>Technical University of Munich (TUM) and Klinikum Rechts der Isar, TUM School of Medicine, Ismaninger Str. 22, 81675 Munich, Germany.

✉ e-mail: [eleftheria.zeggini@helmholtz-munich.de](mailto:eleftheria.zeggini@helmholtz-munich.de)

Received: 12 September 2022; Accepted: 6 June 2023;  
Published online: 14 June 2023

## References

- McCarthy, S. et al. A reference panel of 64,976 haplotypes for genotype imputation. *Nat. Genet.* **48**, 1279–1283 (2016).
- Sudlow, C. et al. UK biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age. *PLoS Med.* **12**, e1001779 (2015).
- Taliun, D. et al. Sequencing of 53,831 diverse genomes from the NHLBI TOPMed Program. *Nature* **590**, 290–299 (2021).
- Halldórsson, B. V. et al. The sequences of 150,119 genomes in the UK Biobank. *Nature* **607**, 732–740 (2022).
- Si, Y., Vanderwerff, B. & Zollner, S. Why are rare variants hard to impute? Coalescent models reveal theoretical limits in existing algorithms. *Genetics* **217** <https://doi.org/10.1093/genetics/iyab011> (2021).
- Zhao, X. et al. Whole genome sequence analysis of pulmonary function and COPD in 19,996 multi-ethnic participants. *Nat. Commun.* **11**, 5182 (2020).

7. Benonisdottir, S. et al. Sequence variants associating with urinary biomarkers. *Hum. Mol. Genet.* **28**, 1199–1211 (2019).
8. Gilly, A. et al. Very low-depth sequencing in a founder population identifies a cardioprotective APOC3 signal missed by genome-wide imputation. *Hum. Mol. Genet.* **25**, 2360–2365 (2016).
9. Wang, Q. et al. Rare variant contribution to human disease in 281,104 UK Biobank exomes. *Nature* **597**, 527–532 (2021).
10. Selvaraj, M. S. et al. Whole genome sequence analysis of blood lipid levels in >66,000 individuals. *Nat. Commun.* **13**, 5995 (2022).
11. Wainschtein, P. et al. Assessing the contribution of rare variants to complex trait heritability from whole-genome sequence data. *Nat. Genet.* <https://doi.org/10.1038/s41588-021-00997-7> (2022).
12. Wessel, J. et al. Rare non-coding variation identified by large scale whole genome sequencing reveals unexplained heritability of Type 2 diabetes. *medRxiv.* <https://doi.org/10.1101/2020.11.13.20221812> (2020).
13. Weiner, D. J. et al. Polygenic architecture of rare coding variation across 394,783 exomes. *Nature* **614**, 492–499 (2023).
14. Kierczak, M. et al. Contribution of rare whole-genome sequencing variants to plasma protein levels and the missing heritability. *Nat. Commun.* **13**, 2532 (2022).
15. Dornbos, P. et al. A combined polygenic score of 21,293 rare and 22 common variants improves diabetes diagnosis based on hemoglobin A1C levels. *Nat. Genet.* **54**, 1609–1614 (2022).
16. Martin, A. R. et al. Low-coverage sequencing cost-effectively detects known and novel variation in underrepresented populations. *Am. J. Hum. Genet.* **108**, 656–668 (2021).
17. Gurdasani, D. et al. Uganda genome resource enables insights into population history and genomic discovery in Africa. *Cell* **179**, 984–1002.e1036 (2019).
18. Thorolfsdottir, R. B. et al. Coding variants in RPL3L and MYZAP increase risk of atrial fibrillation. *Commun. Biol.* **1**, 68 (2018).
19. Nielsen, J. B. et al. Loss-of-function genomic variants highlight potential therapeutic targets for cardiovascular disease. *Nat. Commun.* **11**, 6417 (2020).
20. Kurki, M. I. et al. FinnGen provides genetic insights from a well-phenotyped isolated population. *Nature* **613**, 508–518 (2023).
21. Purcell, S., Cherny, S. S. & Sham, P. C. Genetic Power Calculator: design of linkage and association genetic mapping studies of complex traits. *Bioinformatics* **19**, 149–150 (2003).
22. Li, C. I., Samuels, D. C., Zhao, Y. Y., Shyr, Y. & Guo, Y. Power and sample size calculations for high-throughput sequencing-based experiments. *Brief. Bioinform.* **19**, 1247–1255 (2018).
23. Tran, N. H. et al. Genetic profiling of Vietnamese population from large-scale genomic analysis of non-invasive prenatal testing data. *Sci. Rep.* **10**, 19142 (2020).
24. Li, J. H., Mazur, C. A., Berisa, T. & Pickrell, J. K. Low-pass sequencing increases the power of GWAS and decreases measurement error of polygenic risk scores compared to genotyping arrays. *Genome Res.* **31**, 529–537 (2021).
25. Gilly, A. et al. Very low-depth whole-genome sequencing in complex trait association studies. *Bioinformatics* **35**, 2555–2561 (2019).
26. Kishikawa, T. et al. Empirical evaluation of variant calling accuracy using ultra-deep whole-genome sequencing data. *Sci. Rep.* **9**, 1784 (2019).
27. French, J. D. & Edwards, S. L. The Role of Noncoding Variants in Heritable Disease. *Trends Genet.* **36**, 880–891 (2020).
28. Beyter, D. et al. Long-read sequencing of 3,622 Icelanders provides insight into the role of structural variants in human diseases and other traits. *Nat. Genet.* **53**, 779–786 (2021).
29. Hanks, S. C. et al. Extent to which array genotyping and imputation with large reference panels approximate deep whole-genome sequencing. *Am. J. Hum. Genet.* **109**, 1653–1666 (2022).
30. Yengo, L. et al. A saturated map of common genetic variants associated with human height. *Nature* **610**, 704–712 (2022).
31. Steinberg, J. et al. A molecular quantitative trait locus map for osteoarthritis. *Nat. Commun.* **12**, 1309 (2021).
32. Bocher, O. & Genin, E. Rare variant association testing in the non-coding genome. *Hum. Genet.* **139**, 1345–1362 (2020).
33. Li, X. et al. Dynamic incorporation of multiple in silico functional annotations empowers rare variant association analysis of large whole-genome sequencing studies at scale. *Nat. Genet.* **52**, 969–983 (2020).
34. Bocher, O. et al. Testing for association with rare variants in the coding and non-coding genome: RAVA-FIRST, a new approach based on CADD deleteriousness score. *PLoS Genet.* **18**, e1009923 (2022).
35. Bagnall, R. D. et al. Whole Genome Sequencing Improves Outcomes of Genetic Testing in Patients With Hypertrophic Cardiomyopathy. *J. Am. Coll. Cardiol.* **72**, 419–429 (2018).
36. Priestley, P. et al. Pan-cancer whole-genome analyses of metastatic solid tumours. *Nature* **575**, 210–216 (2019).

## Acknowledgements

The authors would like to acknowledge Kuan-Han Wu (Michigan State University) for his help with the Box 1 graphics. This project has received funding from the European Union's Horizon 2020 research and innovation programme under Grant Agreement No 101017802 (OPTOMICS).

## Author contributions

O.B., C.J.W. and E.Z. wrote and reviewed the manuscript.

## Competing interests

C.J.W. currently works at Regeneron Pharmaceuticals. The remaining authors declare no competing interests.

## Additional information

**Correspondence** and requests for materials should be addressed to Eleftheria Zeggini.

**Peer review information** *Nature Communications* thanks the anonymous reviewers for their contribution to the peer review of this work.

**Reprints and permissions information** is available at <http://www.nature.com/reprints>

**Publisher's note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

**Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2023