Such improvements would also enhance the compatibility of DLO Hi-C with popular enrichment technologies for 3C libraries, for example, Capture-C[12].

Predicting topological trouble

Even though generating high-quality chromatin interaction maps no longer requires millions of cells or massive sequencing efforts, performing extensive 3C studies is still challenging. This is especially true when limited cell or tissue material is available, for instance, when aiming to study genome topology in a human disease setting. Moreover, once interaction maps have been generated, it remains difficult to test the effects of specific chromatin perturbations on 3D architecture. Computational modeling of 3C data is a valuable tool for visualizing and interpreting genome topology[13]. Moreover, modeling holds promise as an approach to predict 3D structure from 1D genomic features, for example, histone modifications and chromatin factor binding[14,15]. Bianco et al.[10] have now taken an important step forward by developing an algorithm to model 3D chromatin folding and predict how structural variation of the genome alters chromatin architecture (Fig. 1b). Their approach, named polymer-based recursive statistical inference method, or 'PRISMR', uses pairwise contacts from Hi-C maps to infer the minimal factors that shape the 3D structure of a genomic region of interest. Importantly, the algorithm does not make any a priori assumptions, nor does it rely on additional or tunable parameters. Adding binding information for the architectural protein CTCF (both location and motif orientation) did not further improve the quality of the 3D models, underscoring the robustness of PRISMR. CTCF binding information alone (without Hi-C data) only partially captured chromatin folding and TAD structure at the *EPHA4* locus, indicating the involvement of other factors than CTCF in establishing proper genome topology. Leveraging wild-type Hi-C maps as input data for modeling and patient sample maps for validation purposes, the authors showed that PRISMR was able to accurately predict how large structural variants (for example, duplications or deletions) observed in individuals with limb malformations alter local enhancer–promoter contacts, leading to transcriptional misregulation and, ultimately, disease. Thus, PRISMR has the potential to greatly facilitate Hi-C data interpretation in silico to help dissect complex gene regulatory processes and explain how genomic rearrangements might cause disease phenotypes. In the future, it will be of interest to see whether PRISMR can also be applied to smaller structural variants, for example, deletions of single enhancers or CTCF sites. ❐

Ralph Stadhouders[1,2]

[1]*Department of Pulmonary Medicine, Erasmus MC, Rotterdam, The Netherlands.* [2]*Department of Cell Biology, Erasmus MC, Rotterdam, The Netherlands.*
e-mail: r.stadhouders@erasmusmc.nl

Published online: 26 April 2018
https://doi.org/10.1038/s41588-018-0112-1

References

1. Denker, A. & de Laat, W. *Genes Dev.* **30**, 1357–1382 (2016).
2. Dekker, J., Rippe, K., Dekker, M. & Kleckner, N. *Science* **295**, 1306–1311 (2002).
3. Bonev, B. & Cavalli, G. *Nat. Rev. Genet.* **17**, 661–678 (2016).
4. Stadhouders, R. et al. *Nat. Genet.* **50**, 238–249 (2018).
5. Jhunjhunwala, S., van Zelm, M. C., Peak, M. M. & Murre, C. *Cell* **138**, 435–448 (2009).
6. Gorkin, D. U., Leung, D. & Ren, B. *Cell Stem Cell* **14**, 762–775 (2014).
7. Lupiáñez, D. G. et al. *Cell* **161**, 1012–1025 (2015).
8. Gröschel, S. et al. *Cell* **157**, 369–381 (2014).
9. Lin, D. et al. *Nat. Genet.* https://doi.org/10.1038/s41588-018-0111-2 (2018).
10. Bianco, S. et al. *Nat. Genet.* https://doi.org/10.1038/s41588-018-0098-8 (2018).
11. Rao, S. S. et al. *Cell* **159**, 1665–1680 (2014).
12. Davies, J. O. et al. *Nat. Methods* **13**, 74–80 (2016).
13. Serra, F. et al. *FEBS Lett.* **589** (20 Pt. A), 2987–2995 (2015).
14. Zhu, Y. et al. *Nat. Commun.* **7**, 10812 (2016).
15. Sanborn, A. L. et al. *Proc. Natl Acad. Sci. USA* **112**, E6456–E6465 (2015).

Competing interests
The author declares no competing interests.

HUMAN GENETICS

Sizing up whole-genome sequencing studies of common diseases

The triplet code underpins analyses of rare and de novo mutations in exome sequencing data, but analysis of the noncoding genome is much more challenging. A new analytical framework for common, complex diseases highlights the need for very large samples to overcome the unavoidable multiple-testing burden and hence provides preemptive warnings against underpowered studies.

Naomi R. Wray and Jacob Gratten

Whole-genome sequencing (WGS) is increasingly being promoted as a platform for investigating the full spectrum of genetic variation associated with common, complex disease, but the analytical challenges are considerable. The primary motivation for a WGS study is to understand whether structural, rare and de novo mutations in the noncoding genome contribute to disease etiology, in addition to the more well-understood contribution from coding mutations. However, until recently, analyses of noncoding variants have mostly considered a priori hypotheses regarding which noncoding variants may be relevant to disease, such as those in promoter and enhancer regions[1,2], which are only a few of many different functional annotations of the genome. These studies are analogous to candidate gene studies, but this is not widely understood or acknowledged. In this issue, Werling et al.[3] provide the first serious attempt to establish a framework for enrichment analyses of rare noncoding variation in WGS studies of common, complex diseases.

Werling et al.[3] analyzed WGS data from 519 autism spectrum disorder (ASD) quartet families (child, unaffected sibling and two parents) from the Simons Simplex Collection, selected on the basis that prior microarray and whole-exome sequencing (WES) data had failed to identify putatively

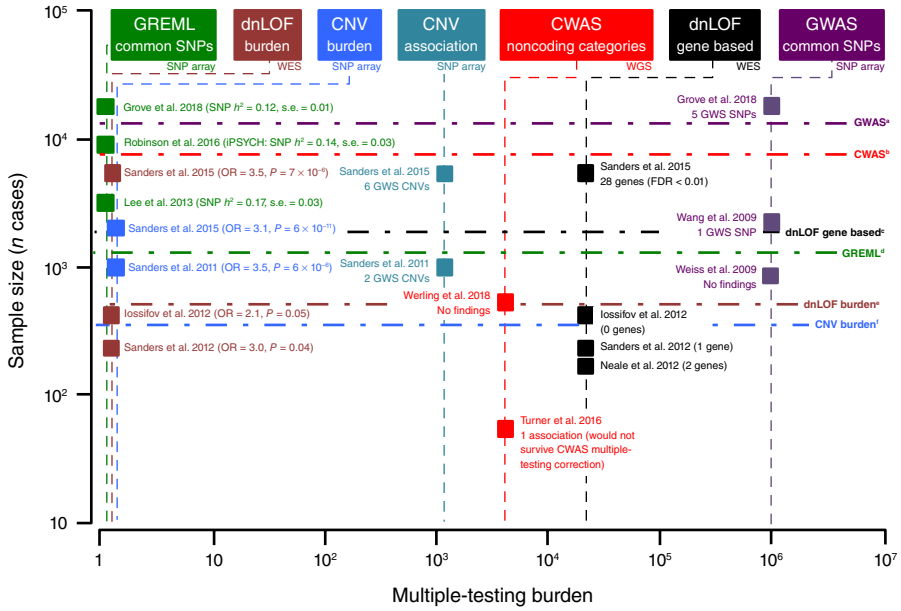

Fig. 1 | CWAS in the context of other experimental designs for investigating genetic variation in ASD. Shown are representative published studies for commonly used experimental designs in ASD genetics (for references, see the Supplementary Note) in relation to multiple-testing burden, sample size and statistical power. Studies are color-coded by experimental design and platform (SNP array, WES, WGS). Power calculation assumptions to estimate required sample size for 80% power are as follows: [a]GWAS (genome-wide association study): risk allele frequency = 0.20, odds ratio (OR) = 1.1, *P*-value threshold = 5×10^{-8}; [b]CWAS: taken from Werling et al.[3]; [c]dnLOF (de novo loss of function) gene based: maximum OR = 25, proportion of de novo loss-of-function-conferring risk = 0.05, number of de novo loss-of-function mutations per person = 0.1, *P*-value threshold = 2.5×10^{-6}; [d]GREML: SNP heritability (SNP-h^2) = 0.20 and lifetime disease risk = 0.01; [e]dnLOF burden: OR = 2, two-tailed binomial exact test; [f]CNV burden: OR = 3, two-tailed binomial exact test. Calculations assume equal sample sizes for controls and cases. Power for CNV association is not included because it is dependent on per-locus OR estimates for which there is no obvious choice. Color-coded vertical dashed lines show the multiple-testing burden for each experimental design (for CWAS, we use 4,120 tests given the sample size in Werling et al.[3]; the number of tests stabilizes to ~10,000 as sample sizes increase). Color-coded horizontal dot-dash lines show the sample size required for ≥80% power for each experimental design. CWAS for ASD will require sample sizes similar in magnitude to GWAS, and even larger sample sizes will be needed for adult-onset common diseases[13]. GWS, genome-wide significant; FDR, false discovery rate.

causal de novo copy number variation (CNV) or loss-of-function mutations. Their challenge was to make sense of the 98% of de novo mutations that fall in noncoding regions of the genome (on average, 64 de novo single-nucleotide variants (SNVs) and 5 de novo indel events per person). Without a regulatory equivalent of the triplet code, knowledge about noncoding regions intolerant to mutation is limited. To address this, Werling et al.[3] propose a framework for a category-wide association study (CWAS), with the aim of identifying categories of genomic annotations that harbor significantly more de novo variants in cases than in sibling controls in an unbiased scan of the genome. They defined 51,801 categories, and, given the multiple-testing burden, they could not detect any proband–sibling differences. Although this is a null finding, by benchmarking against coding mutations, they demonstrate that it is unlikely that any class of noncoding variation is equivalent to coding loss-of-function variants in terms of mutation frequency and effect size for ASD. Power calculations suggest that >8,000 families will be needed to detect differences in the burden of de novo mutations in noncoding categories between cases and controls. They conclude with a strong plea to the research community to hold high standards with regard to appropriate correction for multiple testing in WGS studies. As others have noted[2], multiple-testing correction needs to be applied to both explicit and implicit testing. We must heed these warnings and recognize that very large samples will be needed to address de novo and rare variant hypotheses in WGS studies (Fig. 1). As a community, we should be skeptical of cherry-picking strategies applied to underpowered studies that are primed for false discovery and winner's curse estimates.

Category definition

The analytical framework of Werling et al.[3] can be used with the categories they provided or can be applied with user-defined categories. We summarize their category definitions, divided into five sets, as these are key to understanding the multiple-testing burden imposed. The 'variant types' set has categories of SNVs and indels (in addition to all variants), which represent 92% and 8% of the count of de novo variants, respectively. The 'conservation' set has two categories based on different conservation metrics, each representing 2% of variants. Enrichment of associated variants in conserved regions is consistently found in both common[4] and rare[5] variant analyses. The third set includes 17 categories of 'GENCODE annotations' applied to both the coding (for example, loss-of-function) and noncoding (for example, promoter or intronic) genome. The 'gene lists' set has 14 categories that represent GENCODE gene lists based on transcription annotation and ASD-specific genes, and hence this subset would need to be updated for application to non-brain-related disorders. The final 'functional annotation' set has 31 categories derived from different technologies (for example, ATAC–seq) and particularly from the Roadmap Epigenomics Project. This last category set is the most likely to be updated over time as advances in technology further improve annotation. Any variant can have multiple annotations both between and within sets, and $3 \times 3 \times 17 \times 14 \times 31$ category combinations are constructed for CWAS analysis, giving a total of 51,801 tests after redundant categories are removed. However, categories are correlated and so represent ~10,000 independent tests.

Relevance to common disease

Improved rates of diagnosis are fast-tracking WES as a first-tier test in children with suspected monogenic disorders[6]. The falling costs of WGS, combined with improved calling of the coding genome as compared to WES[7], means that in these applications WGS is likely to replace WES as the technology of choice. The high rates of molecular diagnosis for ASD based on coding de novo mutations and large de novo CNVs (currently up to 15%) mean that WGS data will grow organically. With time, datasets of tens to hundreds of thousands of families[8] will be powered for CWAS, which in turn will help prioritize genomic annotations for follow-up in ASD and other disorders. A recent study of exome-negative families of children from the Deciphering

Developmental Disorders (DDD) study[2] (that is, families in which the proband does not carry a diagnostic coding variant), which conducted hypothesis-driven targeted sequencing of noncoding regions (4.2 Mb, 0.1% of the genome, including enhancers and conserved regions), also concluded that many tens of thousands of family trios will be needed to identify pathogenic noncoding de novo variants. In line with other severe childhood-onset disorders[6], 42% of the DDD children achieved a molecular diagnosis based on de novo mutations of the exome[9], but only an additional 1–3% were found to carry pathogenic de novo mutations in regulatory elements active in fetal brain.

Werling et al.[3] provide an important analysis framework and a clear plan of action for researchers contemplating WGS studies of ASD. However, the lessons for those contemplating WGS of population samples and adult-onset common diseases are sobering, as there is little evidence for rare variants of large effect for common adult-onset diseases (except neurodegenerative diseases)[10–12], most likely because variants of large effect lead to childhood presentation of a more severe disorder. For this reason, the sample size estimated by Werling et al.[3] for CWAS in ASD, which is known to have an important contribution to etiology from rare and de novo variation, is likely to be an underestimate for that needed for later-onset diseases with substantially different genetic architectures[13]. Other factors also weigh into the decisions for conducting WGS in population and common disease settings. For example, sequencing technology is not yet stabilized, and the price differential of WGS as compared to SNP arrays is currently at least 30-fold. Werling et al.[3] show that, with current technology, 30× sequencing depth is needed for accurate detection of de novo mutations, and hence this is the depth needed in any study where the goal is detection of rare variants and, of course, read depth is directly related to cost. Given that very cheap SNP array analysis followed by imputation is now accurate for variants with a frequency of 0.25% or greater[14], WGS studies will need high read depth to gain over cheaper technologies. The 30× coverage was particularly needed to call indels accurately, and, even then, validation rates were lower than for SNVs (96.8% versus 82.4%). Moreover, the same ASD trios have also been studied using long-read technology, which has demonstrated a complexity of structural variants not captured by standard WGS[15]. Given the rate of technical advances in the last decade, it is likely that the next decade will see more accurate and cheaper WGS technologies. The transition from SNP arrays to WGS will inevitably come, and, when it arrives, we should be ready with large, deeply phenotyped cohorts. Hence, for application of WGS to common, adult-onset diseases, we suggest first concentrating on accumulating the sample sizes that will be powered for discovery when the technology is ripe. ❐

Naomi R. Wray[1,2]* and Jacob Gratten[1]*

[1]*Institute for Molecular Bioscience, University of Queensland, Brisbane, Queensland, Australia.* [2]*Queensland Brain Institute, University of Queensland, Brisbane, Queensland, Australia.* *e-mail: naomi.wray@uq.edu.au; j.gratten@imb.uq.edu.au

Published online: 26 April 2018
https://doi.org/10.1038/s41588-018-0113-0

References

1. Turner, T. N. et al. *Cell* **171**, 710–722 (2017).
2. Short, P. J. et al. *Nature* **555**, 611–616 (2018).
3. Werling, D. M. et al. *Nat. Genet.* https://doi.org/10.1038/s41588-018-0107-y (2018).
4. Finucane, H. K. et al. *Nat. Genet.* **47**, 1228–1235 (2015).
5. Taylor, J. C. et al. *Nat. Genet.* **47**, 717–726 (2015).
6. Costain, G. et al. *Eur. J. Hum. Genet.* https://doi.org/10.1038/s41431-018-0114-6 (2018).
7. Meynert, A. M., Ansari, M., FitzPatrick, D. R. & Taylor, M. S. *BMC Bioinformatics* **15**, 247 (2014).
8. SPARK Consortium. *Neuron* **97**, 488–493 (2018).
9. Deciphering Developmental Disorders Study. *Nature* **542**, 433–438 (2017).
10. Fuchsberger, C. et al. *Nature* **536**, 41–47 (2016).
11. Genovese, G. et al. *Nat. Neurosci.* **19**, 1433–1441 (2016).
12. Luo, Y. et al. *Nat. Genet.* **49**, 186–192 (2017).
13. Sanders, S. J. et al. *Nat. Neurosci.* **20**, 1661–1668 (2017).
14. Mahajan, A. et al. Preprint at *bioRxiv* https://doi.org/10.1101/245506 (2018).
15. Collins, R. L. et al. *Genome Biol.* **18**, 36 (2017).

Competing interests

The authors declare no competing interests.

Supplementary information is available for this paper at https://doi.org/10.1038/s41588-018-0113-0.