New mutations, old statistical challenges

Based on targeted sequencing of 208 genes in 11,730 neurodevelopmental disorder cases, Stessman et al. report the identification of 91 genes associated (at a False Discovery Rate [FDR] of 0.1) with autism spectrum disorders (ASD), intellectual disability (ID), and developmental delay (DD)—including what they characterize as 38 novel genes, not previously reported as connected with these diseases[1].

If true, this would represent a substantial step forward. Unfortunately, each of the two discovery analyses (1. *De novo* mutation analysis and, 2. a comparison of private mutations with public control data) contain critical statistical flaws. When one accounts for these problems, fewer than half of the genes--and very few, if any, of the novel findings--survive. These errors have implications for how future analyses should be conducted, for understanding the genetic basis of these disorders, and for genomic medicine.
We discuss the two main analyses in turn and provide more detailed treatment of the issues in a supplementary technical note.

1. Two-stage analysis of *de novo* mutations. The authors selected 208 genes, consisting of 130 with one or more *de novo* truncating mutations in prior published studies, along with 78 others belonging to related pathways or having related Mendelian disease association. Of these genes, 19 have an already documented 'genome-wide significant' excess of *de novo* mutations from 5-6,000 neurodevelopmental disorder patients used as the "discovery sample" by the authors. Moreover, a more recently published exome dataset[2] has convincingly elevated the number of formally genome-wide significant genes to 93.

The authors resequenced this collection of 208 genes in 11,730 neurodevelopmental disorder cases ("replication sample") and looked for an excess of *de novo* mutations. They report that analysis of *de novo* truncating mutations identifies 68 disease-associated genes, while analysis of *de novo* damaging missense mutations adds an additional 10 genes (with 32 of these 78 genes described as novel). These claims, however, are made based on a flawed statistical analysis, and we estimate fewer than half this number of genes achieve an FDR of 0.1, nearly all of which are known.

The study belongs to the traditional 'two-stage' design, which was first popularised 10-25 years ago, in early days of genome-wide association and, before that, linkage studies. There are two valid ways to analyze such a design: (1) count events only in the replication sample, in which case a significance threshold based on the number of genes tested in the replication sample (here, 208) is applied or (2) count events in both the initial and replication sample, in which case a significance threshold based on the number of genes in the genome (~20,000) must be applied.

The authors, however, used an invalid approach: they counted events in both the initial and replication sample, but they applied the weaker significance threshold appropriate for testing only 208 genes. This approach inflates the significance of genes: because the 208 genes were chosen due to a higher-than-expected number of events in the initial sample, the number of events in the combined initial and replication sample will be artificially high; it will not follow the null distribution even if there are no associated genes. While seemingly subtle, this pernicious pitfall applies to any genomewide scanning technique: examining the first half of the data and then completing the study for only the most promising sites is still performing a full genome scan when both halves of the data are analyzed together.

The problem is underscored by some simple observations regarding the *de novo* likely gene disruptive (LGD) analysis:

(1) The rate of *de novo* truncating mutations in the 208 genes is dramatically lower in the replication sample than in the initial sample (0.7% vs. 5.4%)

(2) For 28 of the 68 genes claimed to be significant based on the presence of truncating *de novo* mutations, *no events at all* were observed in the replication sample. Despite the fact that the replication study (twice the size of the discovery sample) therefore provided evidence *against* these genes, the analysis declared the genes significant due only to the use of an inappropriately liberal significance level. These 28 genes include 11 of the 'novel' genes. When one applies a correct significance threshold to the data for truncating *de novo* mutations, the number associated at FDR=0.1 falls from 68 genes to 30 genes (23 of which reach the much stricter genome-wide significance threshold) (Technical Note 1).

Only *one* of the 30 genes, *SETBP1*, has not previously been reported in the exome literature (summarized in recent meta-analyses[2-6]) as associated with ASD, ID and/or DD, but this gene is a well-established cause of severe autosomal dominant intellectual disability, as published by some of the same authors in 2014[7]. Of the 29 others, 28 were documented in exome studies in 2015 or earlier and the 29th (*TCF4*) was noted in the most recent 2017 DDD paper but has been widely known as a cause of ID due to Pitt-Hopkins syndrome since 2007[8-10].

2. Comparison of 'private mutations' in cases vs. public control data. The second type of analysis performed by the authors involves looking for an excess of "private mutations" (that is, singletons) in cases vs. controls. To assess significance, they use a permutation test in which they permute the labels of the cases and controls. Based on this analysis, the authors report 13 additional disease-associated genes (on top of the earlier 78).

This study design is statistically valid—*provided* that the cases and controls are chosen from the sample population and have been sequenced in the same manner. However, the results are *not* valid if the controls (i) did not come from the same populations (population history affects the allelic spectrum, including the frequency of singletons), (ii) were not sequenced and analysed in identical ways (differences in the average sequence depth or the coverage of specific exons can affect singleton detection) or (iii) sites called singleton in one study have unknown frequency in the other (i.e. might be absent, or might be present). The use of a permutation test to assess significance does nothing to eliminate these problems: the systematic differences are present in the case vs. control contrast, but absent in the permutations which distribute the technical differences randomly.

Unfortunately, the authors compared their case samples to a public control sample from the Exome Aggregation Consortium (ExAC)[11] in a fashion that violated all three of the critical criteria above. The ExAC database is chosen from a different mix of populations and, being drawn from heterogeneous exome capture experiments, has differences in local and average coverage and is not at all matched to the targeted MIP study the authors use. For example, *UNC80* is highlighted as having an excess of private mutations in cases. However, examination of ExAC shows it is particularly poorly covered in most exome sequencing studies (in 58 of 64 *UNC80* exons fewer than half of ExAC samples achieve 20x depth)--which will dramatically decrease the observed rate of singleton mutations.

To compare private mutations in cases vs. controls, it is essential that the samples be taken from the same populations, that the technical aspects of sequencing are matched across cases

and controls, and that the authors are careful to remove sites that are singleton in one study, but also observed in the other study at any frequency, or are absent from the other study for technical failures. Of note, the ExAC resource does provide summaries broken down by major continental ancestries and also provides coverage-depth information such that the data could be used much more accurately, albeit still imperfectly, in this context.

As presented in Stessman et al., the results from this analysis are not readily salvageable – even though many true positives may rise to the top of such an analysis, it is almost certain that false positives will also be present owing to the considerations outlined here. In the companion technical note, both the theory and an empirical example are provided in support of these points. This analysis is conceptually sound and the field should be motivated to create shared genomic resources for which technical and population matching can be performed in a way to make these results reliable.

Conclusion. Re-analysis suggests far fewer significant genes, and little or no truly novel genes relative to the existing literature, including the recent mega-analyses in ASD and DD/ID as well as well-established Mendelian genes. As the assignment of disease association to each gene is a foundational step on which years of molecular and clinical research will be built, it is absolutely imperative that studies of rare and *de novo* variation take heed of the hard-earned lessons and strict principles of the past decades of statistical genetics.

We urge the authors to update the analyses in this manuscript and generate corrected tables and figures reflecting an appropriate FDR correction in order to provide a more accurate view of per-gene association probabilities.

Signed (alphabetically)

Jeffrey C. Barrett[1], Joseph D. Buxbaum[2], David J. Cutler[3], Mark J. Daly[4,5], Bernie Devlin[6], Jacob Gratten[7,8], Matthew E. Hurles[1], Jack A. Kosmicki[4,5], Eric S. Lander[5,9,10], Daniel G. MacArthur[4,5], Benjamin M. Neale[4,5], Kathryn Roeder[11], Peter M. Visscher[8,12], Naomi R. Wray[8]

[1] Wellcome Trust Sanger Institute, Wellcome Genome Campus, Hinxton, UK
[2] Seaver Autism Center for Research and Treatment, Department of Psychiatry, Icahn School of Medicine at Mount Sinai, New York, NY 10029, USA
[3] Department of Human Genetics, Emory University School of Medicine, Atlanta, GA 30322, USA
[4] Analytic and Translational Genetics Unit, Department of Medicine, Massachusetts General Hospital and Harvard Medical School, Boston, MA 02114, USA
[5] Broad Institute of Harvard and MIT, 7 Cambridge Center, Cambridge, MA 02142, USA
[6] Department of Psychiatry, University of Pittsburgh School of Medicine, Pittsburgh, PA 15213, USA
[7] Institute for Molecular Bioscience, The University of Queensland, Brisbane, QLD 4072, Australia
[8] Queensland Brain Institute, The University of Queensland, Brisbane, QLD 4072, Australia
[9] Department of Biology, Massachusetts Institute of Technology, Cambridge, MA 02139, USA
[10] Department of Systems Biology, Harvard Medical School, Boston, MA 02115, USA
[11] Department of Statistics, Carnegie Mellon University, Pittsburgh, PA 15213, USA
[12] The University of Queensland Diamantina Institute, The Translation Research Institute, Brisbane, QLD 4102, Australia

Citations

1. Stessman, H.A.F. *et al.* Targeted sequencing identifies 91 neurodevelopmental-disorder risk genes with autism and developmental-disability biases. *Nat Genet* **advance online publication**(2017).
2. Deciphering Developmental Disorders, S. Prevalence and architecture of de novo mutations in developmental disorders. *Nature* **542**, 433-438 (2017).
3. De Rubeis, S. *et al.* Synaptic, transcriptional and chromatin genes disrupted in autism. *Nature* **515**, 209-215 (2014).
4. Deciphering Developmental Disorders, S. Large-scale discovery of novel genetic causes of developmental disorders. *Nature* **519**, 223-8 (2015).
5. Iossifov, I. *et al.* The contribution of de novo coding mutations to autism spectrum disorder. *Nature* **515**, 216-221 (2014).
6. Sanders, S.J. *et al.* Insights into Autism Spectrum Disorder Genomic Architecture and Biology from 71 Risk Loci. *Neuron* **87**, 1215-33 (2015).
7. Coe, B.P. *et al.* Refining analyses of copy number variation identifies specific genes associated with developmental delay. *Nat Genet* **46**, 1063-71 (2014).
8. Amiel, J. *et al.* Mutations in TCF4, encoding a class I basic helix-loop-helix transcription factor, are responsible for Pitt-Hopkins syndrome, a severe epileptic encephalopathy associated with autonomic dysfunction. *Am J Hum Genet* **80**, 988-93 (2007).
9. Zweier, C. *et al.* Haploinsufficiency of TCF4 causes syndromal mental retardation with intermittent hyperventilation (Pitt-Hopkins syndrome). *Am J Hum Genet* **80**, 994-1001 (2007).
10. Brockschmidt, A. *et al.* Severe mental retardation with breathing abnormalities (Pitt-Hopkins syndrome) is caused by haploinsufficiency of the neuronal bHLH transcription factor TCF4. *Hum Mol Genet* **16**, 1488-94 (2007).
11. Lek, M. *et al.* Analysis of protein-coding genetic variation in 60,706 humans. *Nature* **536**, 285-291 (2016).

Technical Note 1:

This study[1] builds from several years of successful publications in this field demonstrating an excess of spontaneously arising protein-truncating variants (referred to in Stessman et al as Likely Gene Disrupting [LGD]). The authors select a set of 208 candidates - largely based on those genes having 1 or more *de novo* truncating mutations in prior published studies (74.52%). Thus, the experimental strategy falls squarely into a traditional and well-established 'two-stage' genetic study design, initially popular 10-25 years ago in early genome-wide association and, before that, linkage studies. Such a study design, in which an initial discovery sample has a full genome scan and is followed by a second replication study of independent samples which examines only the most promising regions or genes to emerge from the first, was widely-used when cost of genotyping was the primary limiting factor in genetic studies[12].

As outlined in Skol et al (2006)[13], among others, when candidate genes are selected from a genome-wide screen (Study1, discovery) and a second, targeted study (Study2, replication) is conducted on regions selected based on top results from Study1, there are two ways of analyzing the resulting data. Analysis option 1 is to consider the results of Study2 as a stand-alone entity and interpret it against a null background determined by the testing burden of only the targeted set of markers. Option 2 is to combine the results from Study1 and Study2, but in this case the analysis must be considered to have been derived from the entire genome, and results compared against a genome-wide testing burden. Failure to do this is self-evidently invalid: if one were to peek at half of a genome-wide study, pick the top 100 genes, and then complete the study, those 100 genes will obviously not be null distributed in the full data set even if no true associations are present.

Skol et al establish that the combined analysis has clear power advantages over the analysis of Study2 alone – despite the testing burden being vastly different between the two designs[13]. Wang et al start with the assumption that the combined analysis is superior and carefully explore optimal two-stage study designs via simulation[14]. The design of a permissive first stage (as Stessman et al employ - advancing genes even with 1 de novo LGD) is optimal, however Wang et al prove that the second (combined) stage analysis must use a significance threshold only imperceptibly higher (1.1-1.2x) than the genome-wide threshold to properly control type I error. Earlier papers addressing an analogous issue pertaining to the follow-up of genome-scans for linkage drew the same conclusions regarding two-stage designs[15,16].

Thus, the analysis option selected (employing a combined analysis of the original and replication data sets) and the use of a permissive first stage including 208 genes are both excellent choices. The comparison to a null distribution and FDR correction based on only 208 genes, however, is disastrous. The primary statistical analyses in Table 1 and Figure 2 from Stessman et al. combine the new targeted sequencing results from 11-13,000 samples with published exome-wide data from 5-6,000 neurodevelopmental disorder patients – including the same exome sequencing studies which generated the earlier *de novo* mutation lists[1]. While the combined analysis is clearly the most powerful approach[13] (per Skol et al), the misstep is that the FDR results and definitions of significance thresholds used in Figure 2 correct only for 208

genes targeted, and not for the entire exome. The effects of this oversight are substantial – 28 of the 68 genes claimed as significant in the principal analysis of *de novo* LGD mutations actually have no new *de novo* mutations discovered in the targeted analysis – they are elevated into apparent significance solely by the existence of a handful (in 19 cases only 2) *de novo* LGD mutations in previously published data and a new FDR correction assuming only 208 genes are considered. This includes 11 genes described as novel findings of this study based on LGD excess. For these genes, the evidence presented is in fact negative – no new mutations have been discovered in twice as large a sample as previously published, and thus these genes must be considered less likely candidates for genuine association than before.

Assessing how negative the results are for the 28 genes with no new mutations involves many considerations. One obvious factor is the 'winner's curse' analog – the top results from a moderately powered Study1 will be demonstrated to have overestimated effects when re-examined in Study2 (whether the results are true or false positives[17-19]). Evidence of this phenomenon is clear - the rates of de novo LGD mutations in the published data (5.4%) and new data (0.74%) are dramatically different. We recomputed the QQ plots (using a comparable mutational expectation model[20]) for the original, new and combined data. The original data (**Figure 1a**) show strong signal, as expected given that the genes were selected largely based on the existence of *de novo* LGDs – as such this is not a random set of 208 genes but rather the ~1% extreme tail of an exome-wide distribution drawn from published studies.

Assuming the molecular assays detected mutation equivalently to previous studies, and diagnoses were similarly equivalent, the new data by itself constitutes an independent, bias-free test of 208 genes (Analysis Option 1 from Skol; **Figure 1b**) and is the only one of these three that can be legitimately compared to a null expectation of a 208 gene test. Here rather than 78 genes, a much smaller number show substantial deviation from the null QQ plot (the most extreme of which correspond to well-known genes). The combined analysis (**Figure 1c** presented here - nearly identical to Figure 2 in Stessman et al) is only legitimate when considered as the tail of an entire exome-wide study – following the well-established principle that combined two-stage analyses have a testing burden that is much closer to an entire full genome scan of both stages than it is to that of a randomly chosen set of genes or variants. In both Figure 1a and 1c, we have added a heavy dashed line representing where the 1% tail of a null distribution drawn from the entire exome should appear to clarify how different the correct and incorrect interpretation of the observed data are. With an appropriate correction, we estimate that only 23 genes meet strict genomewide significance and 30 genes meet the more liberal Stessman et al. FDR threshold of 0.1 (**Supplementary Table 1**)[1]. Of the remaining 38 of 68 LGD 'significant' genes, 30 are genes in which only 2 LGD mutations in ~18,000 cases have been observed in total (along with several others for which observing 3 mutations does not provide statistically compelling evidence against the background of the entire genome).

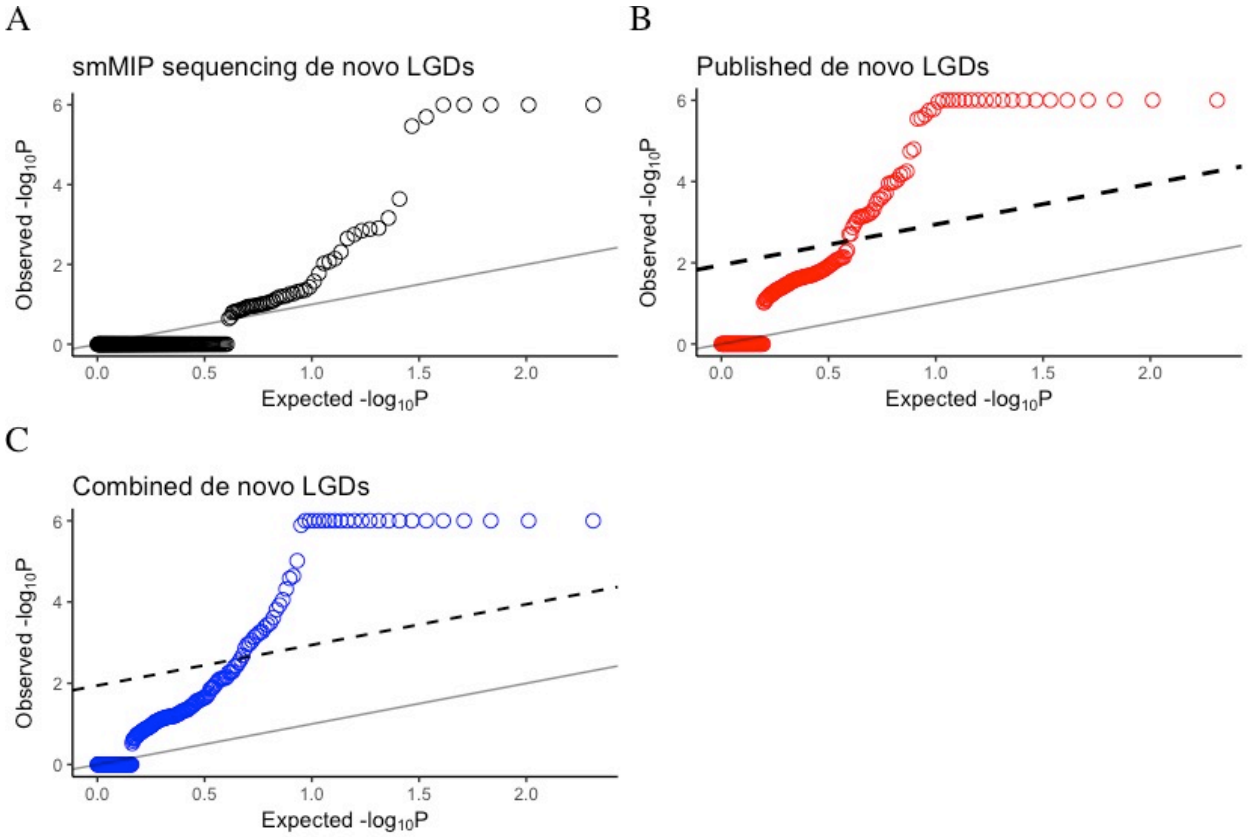


Figure 1: Quantile-quantile plots for gene-based *P*-values based on the number of *de novo* LGD variants found from a) smMIP sequencing[1], b) published data, and c) combined smMIP sequencing plus published data. Inverse, log transformed *P*-values represent the Poisson probability of observing more than the expected number of *de novo* LGD variants based on a similar mutation model[20] and the number of samples used. *P*-values were capped at 1 x 10^{-6} to display the bulk of the distribution. Light grey line represents the standard null distribution for the number of genes tested. The heavy dashed black line represents where the 1% tail of a null distribution drawn from the entire exome would appear in (**b**) and (**c**). Underlying data located in **Supplementary Table 1**.

Technical Note 2:

It is a surprising fact that in two independent samples drawn from a constant size neutral population, the expected number of apparently "private mutations" observed is independent of sample size. This is to say, if one sequences 1,000 individuals and observes X variants with exactly one copy of the minor allele (singletons) and then conducts different studies of 100, 1,000, 10,000 or 1,000,000 samples, the expected number of singleton sites observed is X in all of those other studies[21]. This fact is relatively simple to prove. Since a singleton variant in a sample of 100 alleles has expected frequency 1/100 and a singleton variant in a sample of 1,000,000 alleles has expected frequency 1/1,000,000, singletons in the larger study are fundamentally "rarer" than singletons in the smaller study. Thus, a variant at frequency 1 in 100 in the general population might appear as a singleton (and be counted as such) in the small

study, but most will not be singleton in the larger study (most will rightly be observed to be much more frequent, although their expected relative frequency is the same).

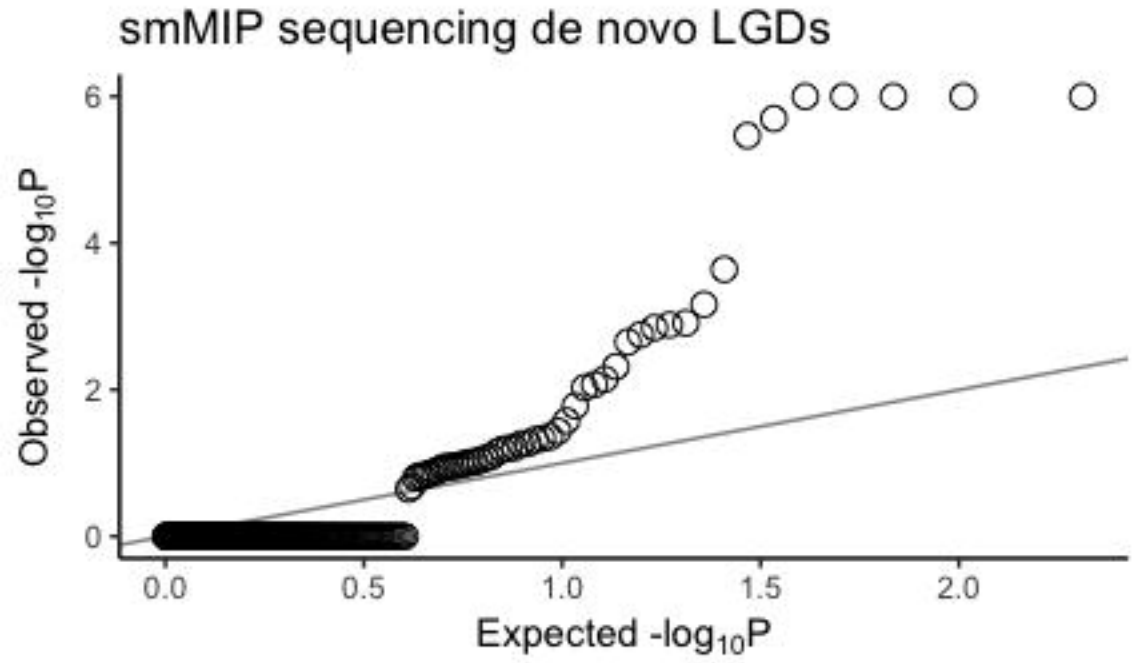
Intuition aside, in a constant size neutral population, a valid statistical way to compare the number of singletons in the present study to the number found in ExAC[11] is to ask whether the number in the present study is larger than ExAC, independent of sample size (i.e., a McNemar chi^2 test, or binomial exact test with $p = 0.5$). A quick glance at Table 2 suggests perhaps 2 of these results are experiment wide significant, after correcting for 208 tests[1]. Readers might sensibly object to the assumption of a constant size human population as the proper null, but in this case an even greater challenge develops. If we admit that population sizes are growing, then we must account in a rigorous way for population structure in this analysis. If some individuals are from populations growing slightly faster (or slower) than others, or have undergone unique demographic events that influence the allele frequency distribution, the entire statistical testing framework for rare alleles is uncertain, but it cannot even begin without a careful analysis of population structure, and accounting for that structure in the statistical testing framework. This is not possible using ExAC summary count data as presented on the public website.

Furthermore, this analysis is not statistically valid unless all the same exons are analysed with the same calling accuracies and efficiencies in the new targeted sequencing as they were in ExAC. Any difference in sensitivity or false positive rate between the two studies will lead to an imbalance in singleton counts. Because we know the technologies used for target enrichment are wholly different, the depth of coverage at individual sites and exons is almost surely different, and the observed rate of *de novo* variation appears to differ by a factor of seven or more (5.4 versus 0.74), concerns about the underlying differences in technology (or perhaps clinical ascertainment of individuals) must be considered. Importantly, that the authors assessed statistical significance via a permutation does nothing to relieve these technical and population mismatch concerns. Had the same number of cases (this study) and controls (ExAC) been sequenced with the same technology capturing the same exons to the same average depth, in individuals coming from well-matched populations, the permutation assessment of significance would be an excellent choice. In the present case in which there are fundamental mismatches between the case and control sets, permuted-label data sets are unhelpful as they have randomized the confounding structure and cannot therefore speak to the impact of those confounders in the observed data.

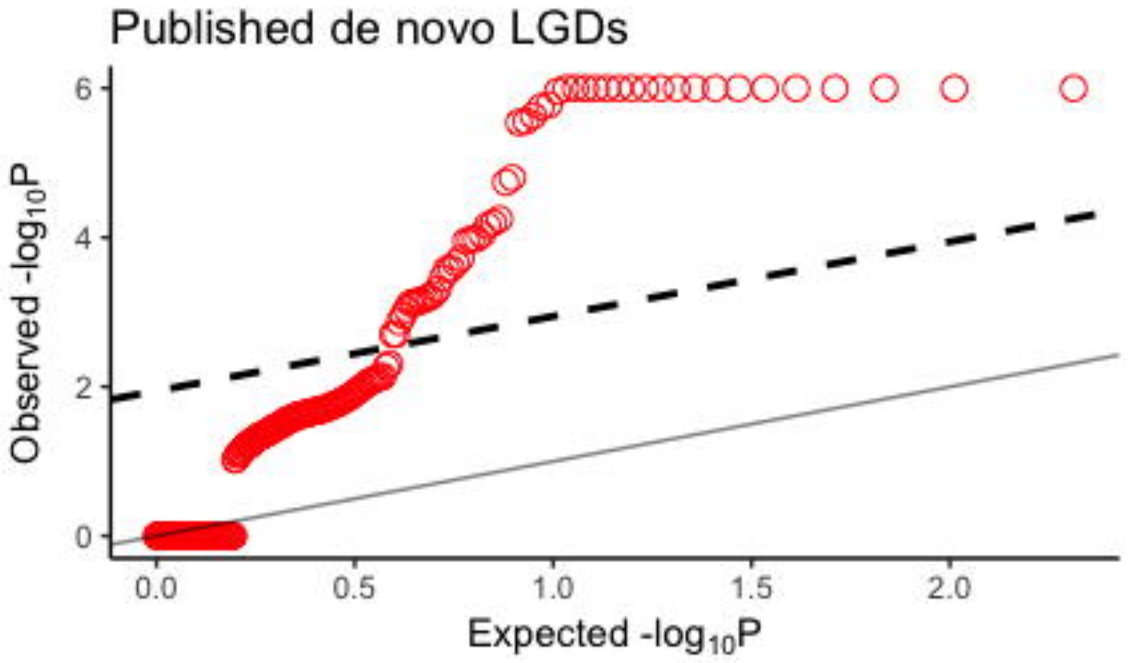
Citations

12. Satagopan, J.M. & Elston, R.C. Optimal two-stage genotyping in population-based association studies. *Genet Epidemiol* **25**, 149-57 (2003).
13. Skol, A.D., Scott, L.J., Abecasis, G.R. & Boehnke, M. Joint analysis is more efficient than replication-based analysis for two-stage genome-wide association studies. *Nat Genet* **38**, 209-13 (2006).
14. Wang, H., Thomas, D.C., Pe'er, I. & Stram, D.O. Optimal two-stage genotyping designs for genome-wide association scans. *Genet Epidemiol* **30**, 356-68 (2006).
15. Guo, X. & Elston, R.C. One-stage versus two-stage strategies for genome scans. *Adv Genet* **42**, 459-71 (2001).
16. Kruglyak, L. & Daly, M.J. Linkage thresholds for two-stage genome scans. *American Journal of Human Genetics* **62**, 994-997 (1998).
17. Göring, H.H.H., Terwilliger, J.D. & Blangero, J. Large Upward Bias in Estimation of Locus-Specific Effects from Genomewide Scans. *American Journal of Human Genetics* **69**, 1357-1369 (2001).
18. Ioannidis, J.P., Ntzani, E.E., Trikalinos, T.A. & Contopoulos-Ioannidis, D.G. Replication validity of genetic association studies. *Nature genetics* **29**, 306-309 (2001).
19. Lohmueller, K.E., Pearce, C.L., Pike, M., Lander, E.S. & Hirschhorn, J.N. Meta-analysis of genetic association studies supports a contribution of common variants to susceptibility to common disease. *Nat Genet* **33**, 177-82 (2003).
20. Samocha, K.E. *et al.* A framework for the interpretation of de novo mutation in human disease. *Nat Genet* **46**, 944-50 (2014).
21. Johnston, H.R., Hu, Y. & Cutler, D.J. Population genetics identifies challenges in analyzing rare variants. *Genet Epidemiol* **39**, 145-8 (2015).

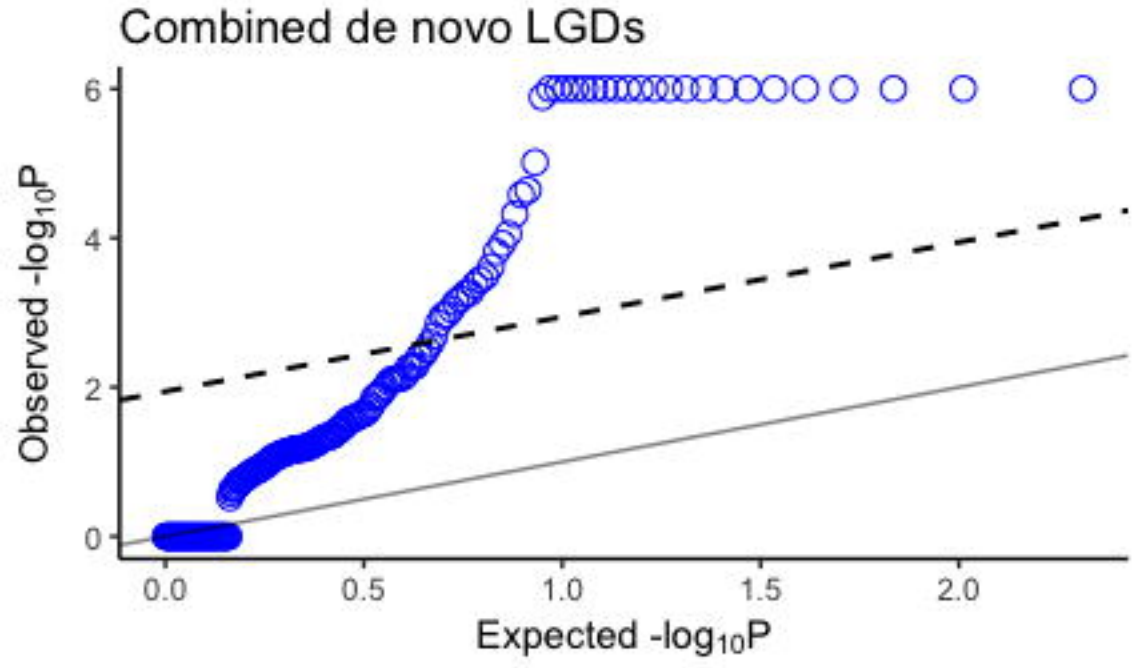
A


smMIP sequencing de novo LGDs
Observed -log10P
Expected -log10P
0
2
4
6
0.0
0.5
1.0
1.5
2.0

B


Published de novo LGDs
Observed -log10P
Expected -log10P
0
2
4
6
0.0
0.5
1.0
1.5
2.0

C


Combined de novo LGDs
Observed -log10P
Expected -log10P
0
2
4
6
0.0
0.5
1.0
1.5
2.0