Hotspots of human mutation point to clonal expansions in spermatogonia

Vladimir Seplyarskiy[1,2*†], Mikhail A Moldovan[1*†], Evan Koch[1], Prathitha Kar[1], Matthew DC Neville[3], Raheleh Rahbari[3] & Shamil Sunyaev[1,2†]

[1] Department of Biomedical Informatics, Harvard Medical School, Boston, MA, USA
[2] Brigham and Women's Hospital, Division of Genetics, Harvard Medical School, Boston, MA, USA
[3] Cancer, Ageing and Somatic Mutation, Wellcome Sanger Institute, Hinxton, United Kingdom

***Equal contribution**
[†]Corresponding Authors:
Shamil Sunyaev, ssunyaev@hms.harvard.edu
Mikhail A Moldovan, Mikhail_Moldovan@hms.harvard.edu
Vladimir Seplyarskiy, alicodendrochit@gmail.com

Abstract

In renewing tissues, mutations conferring selective advantage may result in clonal expansions[1–4]. In contrast to somatic tissues, mutations driving clonal expansions in spermatogonia (CES) are also transmitted to the next generation. This results in an effective increase of *de novo* mutation rate for CES drivers[5–8]. CES was originally discovered through extreme recurrence of *de novo* mutations causing Apert syndrome[5]. Here, we develop a systematic approach to discover CES drivers as hotspots of human *de novo* mutation. Our analysis of 54,715 trios ascertained for rare conditions[9–13], 6,065 control trios12,14–19, and population variation from 807,162 mostly healthy individuals[20] identifies genes manifesting rates of *de novo* mutations inconsistent with plausible models of disease ascertainment. We propose 23 genes hypermutable at loss-of-function (LoF) sites as candidate CES drivers. An additional 17 genes feature hypermutable missense mutations at individual positions, suggesting CES acting through gain-of-function (GoF). CES increases the average mutation rate ~17-fold for LoF genes in both control trios and sperm and ~500-fold for pooled GoF sites in sperm. Positive selection in the male germline elevates the prevalence of genetic disorders and increases polymorphism levels, masking the effect of negative selection in human populations. Despite the excess of mutations in disease cohorts for 19 LoF CES driver candidates, only 9 show clear evidence of disease causality[21], suggesting that CES may lead to false-positive disease associations.

Introduction

The genome of the average newborn harbors ~70 *de novo* point mutations, of which ~80% arise in paternal germline[14,22]. *De novo* mutation is a common cause of sporadic monogenic disease[9–13,15], which has led to multiple international efforts collecting large cohorts of parent-child trios[9–13,15]. These trios typically include unaffected parents and their children who, depending on the cohort, may be affected by conditions ranging from neurodevelopmental disorders to congenital heart diseases[9–13,15]. To discover the genes associated with disease, these studies compare the observed counts of *de novo* mutations in a gene with the ones expected from a baseline mutation rate model[9,12]. Due to the initial ascertainment on disease phenotype, this signal provides evidence for the gene's causal role.

Studies of *de novo* variation, most prominently studies of sporadic neurodevelopmental disorders (NDD)[9] and autism spectrum disorders (ASD), have identified hundreds of disease genes. As expected, the vast majority were found to be under strong negative selection consistent with a significant impact of *de novo* mutations on severe early-onset conditions[9]. However, we observe that a small subset of the identified genes defy a simple dependence of disease effect and selection and exhibit substantial levels of population polymorphism. Another observed discrepancy is that counts of specific high-impact *de novo* mutations in disease trio cohorts exceed the maximal levels explainable by disease ascertainment. A potential explanation is that forces beyond random mutagenesis increase the observed incidence of mutations.

Many genes within this subset are involved in carcinogenesis or clonal expansions in somatic tissues suggesting that a plausible explanation may be offered by the positive selection leading to clonal expansions in spermatogonia (CES). Clonal expansions have been recently found to be a feature of all renewing tissues[1–4], including spermatogonia. Indeed, gain-of-function (GoF) mutations in 13 genes[7,8,] have already been demonstrated to drive CES. In contrast to somatic tissues, CES increases the likelihood of transmitting driver variants to the next generation, elevating the observed mutation rate in the offspring (Fig. 1a). For this reason, the first CES drivers were discovered through their recurrence in patients with monogenic disease[5,6]. By increasing mutation rate, CES can lead to an increased prevalence of disease. Conversely, genes that drive CES without directly causing disease could appear as disease-causing due to the elevated number of *de novo* mutations observed in affected children.

To systematically investigate the possible effect of CES on human germline mutation and identify potential CES driver candidates, we developed a statistical approach leveraging human *de novo* mutation data[9–15,22] along with population genetic variation resources[20]. Most of the identified genes follow biological expectation for drivers of clonal expansions such as involvement in MAPK[7] and other major signaling pathways. We validated our findings using the already discovered CES genes[7,8], cohorts of unaffected trios[12,14–19] and comparison with the results of the direct sperm sequencing[23].

Results

Rates of DNA damage, together with imperfections of DNA repair and replication determine the baseline *de novo* mutation rate in the male germline[24]. Positive selection in spermatogonia elevates observed mutation rates of the drivers relative to the baseline (Fig. 1a). The approaches to identify CES genes taken here will generally compare observed mutation counts to those expected under a model of baseline mutation rate.

Historically, mutation rates at functional sites have been estimated from the prevalence of phenotypes caused by mutations at those sites. We generalized this logic to identify CES genes in a large trio sequencing cohort ascertained by any phenotype. Specifically, we counted *de novo* mutations in the largest assembled neurodevelopmental disorder (NDD) trio cohort[9] (31,058 affected probands). The effect of ascertainment has a strict upper bound given by the inverse prevalence of ascertained phenotype. Denoting V the presence of a *de novo* variant and D – presence of the disease, we have:

$$\frac{\#\,obs}{\#\,exp} \approx \frac{P(V\,|D)}{P(V)} = \frac{P(D\,|\,V)}{P(D)} \leq \frac{1}{P(D)}, \qquad (1)$$

where $P(V)$ is the probability of *de novo* variants defined by a mutation rate model (Supplementary Text 1) and $P(D)$ is the overall disease prevalence (Fig. 1b). Unlike disease-causing genes with no CES effect, CES drivers could violate this upper bound, offering an approach for finding CES drivers. We develop a statistical approach based on Eq. 1 and demonstrate that it is robust to non-monogenic inheritance (Supplementary Text 4).

From the biological standpoint, because many known CES driver mutations cause NDD[7,8], we may expect the uncharacterized CES drivers to be similarly prone to causing NDD. The count of CES-driving mutations in the NDD cohort should be greater than in an unascertained set of trios, maximizing the chance of identifying CES drivers involved in NDD. Genes that drive CES but do not cause NDD, can be found by this approach only if the effect of CES significantly exceeds the maximal effect of disease ascertainment given by the inverse prevalence (Supplementary Text 4).

This approach is sensitive to misspecifications of the mutation rate model. For the mutation rate model Roulette[25], we demonstrated the accuracy of predictions for the synonymous variants in the NDD cohort (Fig 1c, Methods-5).

Precision of the prevalence estimates, $P(D)$, might also affect the applicability of Eq. (1). To address this issue we adopted a conservative lower bound on prevalence[26] of NDD of 1% (see Methods-6.2). We also show that the phenotypic sampling is uniform across the NDD subcohorts ruling out false positive CES findings due to heterogeneous ascertainment by phenotype (Methods-6.1).

Gain-of-function mutations causing CES

All CES genes discovered to date act through gain-of-function (GoF) mechanisms, and all known CES mutations are individual missense variants[7,8]. We therefore first apply Eq. (1) to all

possible (5×10^7) individual missense *de novo* mutations in the NDD cohort and identify 21 variants in 18 unique genes that pass the 20% FDR threshold (Extended Data Figure 1a, Suppl. Tab. S11). Among the 18 identified genes, 5 have been previously experimentally established to cause CES[7,8]. From the functional perspective, the identified genes are enriched in RAF activation (adjusted $p=1.7\times10^{-7}$) and MAPK1/MAPK3 (Fig 1d, adjusted $p=2.2\times10^{-7}$) pathways[27] in agreement with previous studies[7] (Suppl. Tab. S12, see Methods-7). All but one of these genes are expressed in spermatogonia ($p=6.5\times10^{-7}$). The exception, *GRIN2B,* was thus excluded from the GoF set (Extended Data Figure 2) as a potential false positive.

As explained above, the test based on Eq. 1 has greater power to identify CES drivers that also cause NDD. Indeed, 16 out of 17 genes have independent evidence of association with disease phenotypes through gain-of-function variants (Fig. 1e).

Loss-of-function mutations causing CES

Although existing experimental work has focused solely on GoF mutations driving CES, LoF mutations are known to also drive clonal expansions in cancers and healthy renewing tissues[1]. This motivated us to investigate LoF mutations as plausible CES drivers (Fig. 2a). Because every gene harbors multiple possible LoF mutations, we may extend Eq (1) to test sets of variants within genes rather than individual variants. We validated that the aggregation of LoF sites by gene does not bias our mutation rate expectation (Fig. 2b, Methods-5).

Five genes display numbers of de novo LoF mutations in the NDD cohort significantly exceeding any plausible ascertainment by disease (by Eq. 1): *PURA*, *ARID1B*, *CTNNB1*, *DYRK1A* and *FOXG1* (Fig. 2cd, Extended Data Figure 1bc). We call this list the LoF-1 set. Three of them, *ARID1B*, *CTNNB1*, *DYRK1A*, are involved in carcinogenesis or other clonal expansions[28,29] (Fig 2c). Although *PURA* and *FOXG1* have not been previously identified as drivers of somatic clonal expansions, their functions fit the profile of clonal expansion drivers. *PURA* is involved in replication and transcription control and *FOXG1* is a transcription factor regulating early development (Fig. 2c).

The independent evidence of association with disease phenotypes supports the effect of LoF mutations on NDD for all five of these genes. Also, these genes are highly selectively constrained in the human population (LOEUF < 0.3), consistent with their role in severe pediatric conditions.

CES and loss-of-function polymorphism

We now bring our attention to CES drivers with weak or no effect on NDD. The approach outlined above is not suited well to identify such genes, because it requires the counts of *de novo* mutations in the NDD cohort to exceed the expectation by at least 100 fold (inverse prevalence of NDD). Therefore, we rely on additional considerations. In particular, we note that CES should increase the rate of functionally consequential mutations and thus polymorphism levels in the general population. The rate of *de novo* mutations in a disease cohort should be elevated for both CES genes and disease genes. However, unlike CES driver with no effect on NDD,

disease-causing genes should be deprived of functional polymorphism in the predominantly healthy population. This discrepancy between amounts of genetic variation in the population among genes significant for the excess of *de novo* variants in the NDD cohort over the mutation rate expectation (not necessarily exceeding the boundary of Eq. 1) yields a procedure for identifying CES genes. Specifically, a gene with both large amounts of polymorphism in the general population and the excess of de novo variants in a disease cohort may be deemed a CES driver candidate. Note that this procedure may be applied to cohorts with mixed disease phenotypes, because the expectation is defined just by the mutation rate model and is independent of ascertainment, motivating us to increase power by merging *de novo* variation in cohorts ascertained by NDD (31,058 affected probands)[9] and autism spectrum disorders (ASD, 16,877 affected probands)[12]. We quantify LoF polymorphism in the general population using LOEUF[20], an upper bound estimate of the observed-to-expected ratio of LoF SNV counts.

122 genes show a significant excess of LoF *de novo* mutations in the ASD/NDD cohort at FDR < 0.1 (see Methods-6.5, Suppl. Tab S7). From these, we selected 19 genes (Fig 3a, Extended Data Figures 3,4) with LOEUF > 0.5 (concordant with cutoffs usually chosen for relaxed negative selection[12,30–32]). High values of LOEUF may be indicative of both relaxed selection and misspecifications of the mutation rate model, as should be the case with CES-driven mutation rate elevation. Genes with misannotated LoF variants[33,34] or involved in clonal expansions in blood[3], may also have spuriously high LOEUF values for reasons unrelated to CES. Due to these reasons, four genes were flagged as potential false findings (Supplementary Text 2,3). The remaining 15 genes we call the LoF-2 set.

Four additional considerations support this set of genes as CES driver candidates. First, we show that LoF-2 genes have a significant excess of *de novo* LoF mutations not just in the joint ASD/NDD dataset, but also in a healthy trio cohort[12,14–19] (Fig 3b). The excess of mutations in the healthy trio cohort could not be affected by disease ascertainment, allowing us to estimate the average CES effect for LoF-2 genes. Surprisingly, the magnitude of the effect in LoF-2 genes is comparable between ASD/NDD and control cohorts (Fig 3b, 18.7 vs 26.9, p=0.2), implying that the effect of disease ascertainment is most likely small and the majority of LoF-2 genes do not cause ASD/NDD. Indeed, only four genes from this set have independent evidence of the causal role in ASD/NDD[21] (Fig 3c). Just one of these genes, *PTEN,* demonstrates an excess of *de novo* LoF mutations in the NDD cohort comparable to the excesses in the LoF-1 set (Fig 3a, c). In line with small effects of NDD ascertainment in LoF-2 genes, we observe similar magnitudes of the enrichment of *de novo* LoF mutations observed in other trio cohorts ascertained by different phenotypes: encephalopathic epilepsy and congenital heart disease. Finally, using the cohort of healthy controls, we additionally estimated that just 15 LoF-2 genes explain 3.8% of the entire *de novo* LoF variation genome-wide.

To investigate the robustness of the LOEUF threshold used to select genes in the LoF-2 set, we compared the excesses of LoF mutations in the control cohort for different thresholds. The magnitude of the effect does not change substantially in the range of LOEUF threshold

values from 0.3 to 2.0 (Extended Data Figure 4). The effect in the healthy cohort dissipates for ASD/NDD genes with LOEUF<0.3, consistent with causal disease effect of these genes.

Second, all LoF-2 genes barring *KCNA1* are expressed in spermatogonia (expression enrichment p = 0.01, Mann-Whitney U test). *KCNA1* was flagged as a potential false-positive finding (Extended Data Figure 2). Four LoF-2 genes are also known to cause cancers[28].

Third, with the help of population variation data, we can find additional evidence that the high level of LoF polymorphism in LoF-2 genes likely reflects elevated *de novo* mutation rate rather than relaxed selection. The large sample size of gnomAD-v4 (1.6M haploid genomes) enables analysis of frequencies of LoF variants generated by recurrent mutations[25,35–37]. LoF mutations within the same gene are expected to have identical functional effects, including on CES. All individual LoF sites within the same gene therefore share a common linear inflation factor relative to baseline mutation rate expectations. In the presence of CES, allele frequencies remain linearly proportional to the baseline mutation rate attesting to CES being a linear effect, as illustrated by the example of *MIB1* in Fig 3d.

We expand this logic to statistically detect elevation of mutation rate for LoF-2 set in the human population data. Variants with the same mutation rate and selection coefficient have different allele frequencies due to the effect of genetic drift. Strong selection overpowers the effect of drift and makes allele frequencies approach the expectation under the balance between mutation and selection. CES is expected to shift this balance and increase the average allele frequency, but is not expected to affect deviations from this balance. Slightly more formally, the distribution of sampling allele frequencies in the presence of recurrent mutation and relatively strong selection, is given by the Nei approximation[38] (see Methods-9). Relying on the Nei approximation, we isolate variance inflation due to genetic drift. It is inversely proportional to selection coefficient. The mean allele frequency is similarly inversely proportional to selection coefficient, but is also directly proportional to the effect of CES. Thus, CES effects may be identified from the comparison of the mean and variance terms for a single gene. Even with the current sample sizes, this procedure has low power, as most of the information is contained in frequencies of highly mutable yet rare CpG sites, and so we restricted our analysis to 952 genes with more than 10 CpGs. Fig 3e shows the evidence of CES effects for LoF-2 genes included in the analysis.

Our final analysis relies on GeneBayes[39], which evaluates per-gene selective constraint relying on a protein function-informed prior and a likelihood based on LoF variation within the gene. The prior, unlike the likelihood, is not influenced by mutation rate and the effects of CES. We observe LoF-2 genes have the most inconsistent prior and likelihood estimates, concordant with the effects of CES on the likelihood (Extended Data Figure 5).

Comparison with sperm sequencing

Direct sequencing of mature sperm provides important data for the evaluation of CES effects without any ascertainment by phenotype. In parallel with this study, NanoSeq, an accurate single-molecule resolution DNA sequencing technology, was applied to sequencing sperm cells

of 63 donors[23]. A panel of 263 genes implicated in cancer was sequenced with a mean coverage of 8×10^4 and whole exomes were sequenced with a mean coverage of 2.1×10^4. Neville et al. identified 40 genes under positive selection in spermatogonia based on enrichment of non-synonymous mutations in sperm[23]. Direct sequencing of sperm is effectively equivalent to assaying paternal *de novo* mutations across all genes in a large cohort of randomly selected trios barring those mutations not compatible with live birth, a case that is discussed below. We used these data to verify some of our predictions.

To evaluate the level of concordance, we tested if CES driver candidates identified in human genetics datasets achieved significance in the test of Neville et al. aggregating all types of non-synonymous mutations[23]. Of all 40 genes identified here, 17 are nominally- (20 at the p-value threshold of 0.1 used by Neville et al.) - and 11 study-wide significant in Neville et al. We compiled the joint list of all putative CES genes between studies (Supplementary Table S20).

Next, we compared genes of LoF-1 and LoF-2 sets with 16 genes significant in the NanoSeq dataset exclusively due to LoFs (FDR 20%). As shown in Fig. 4a, out of 16 NanoSeq genes, 7 overlap with our LoF sets.

Partial overlap between the lists could be attributed to multiple factors. First of all, datasets obtained by NanoSeq, have variable coverage among the genes, and indeed our CES candidates that are not significant in Neville et al, have lower coverage (Extended Data Figure 6). Second, for GoF-1 and LoF-1 the two approaches have differential statistical power. For the human genetics approach, the power is enhanced by the effect of CES drivers on NDD, whereas the power of the NanoSeq experiment is not boosted by the ascertainment. Finally, the power of both approaches is affected by sampling variance in the low mutation count data.

To demonstrate that the overlap is limited by power, we show that a residual CES effect exists in genes outside of the overlap. We calculated the observed-to-expected ratio for LoF mutations in the NanoSeq dataset for genes that did not attain genome-wide significance, but are included in our LoF sets (Fig 4b). We also measured the observed-to-expected ratio for genes outside the overlap using *de novo* variants in control trios (Fig 4b). In both cases, we observe significant enrichment of LoF mutations ($p < 0.01$, Poisson test).

In the NanoSeq dataset, the combined LoF-1 and LoF-2 gene sets demonstrate a 16.6-fold (Poisson 95% CI 13.5-20.2) excess of LoF mutations. For the GoF set, we observe a remarkable 524-fold excess of mutation counts in sperm sequencing (Poisson 95% CI 311-828).

Discussion

The unique position of gametogenic tissues as a bridge between somatic and germline evolution allows CES to be viewed from two different perspectives and in terms of two different fields. From the biological perspective, CES can be related to cancer with similar genes being involved in clonal expansions in normal tissues and oncogenic transformation. From the human genetics perspective, CES acts as a factor inflating mutation rate at certain positions. Inflation of mutation rate leads to increase in population frequency of CES drivers. However, because the putative

CES drivers reported here are kept at low frequency in the population, we expect that they are subject to strong negative selection even in cases with no obvious disease association.

Biological function of CES drivers

Biological functions of many of our CES driver candidates are consistent with the role in clonal expansions (Suppl. Tab S6). We note that the method used here biases the resulting gene sets towards NDD and ASD causal genes, but the functional roles of CES driver candidates are distinct from the bulk of known ASD/NDD genes.

Specifically, out of 40 identified genes, 20 play a role in major signaling pathways (Suppl. Tabs. S6, S15-S18). The pathways include MAPK, WNT and TGF-β (Suppl. Tab. S13). As expected, the MAPK pathway has the highest enrichment (8 genes, Fig. 4c). We noticed that the putative CES drivers are concentrated on the pathway structure and the LoF and GoF annotations are mirrored in the activator/repressor roles (Fig. 4c).

Serine-threonine kinase CSNK2A1 and genes activated by it (PACS1 and PACS2)[40] form an intriguing group of functionally related CES driver candidates within the GoF set. Another related kinase, CSNK2B[40], has been reported by Neville *et al.*. R203T mutation in PACS1 is the most recurrent variant in the NDD cohort[41] and explains 0.1% of all cases and E209K mutation in PACS2 explains another 0.04%. The high recurrence of these variants in the NDD cohort may be explained by their role in both NDD[41,42] and CES.

As expected for genes involved in clonal expansions, 12 out of 40 genes identified here are COSMIC census cancer drivers[28]. Four of them (*FGFR3*, *PTEN*, *MTOR* and *BRAF*) have been specifically associated with testicular tumors[43–46].

There is a discrepancy in the average elevations of mutation rate due to CES between GoF and LoF mutations (524-fold vs 17 -fold). We propose two possible explanations. First, spermatogonia cells are diploid, and the variants discussed here are likely heterozygous. LoF mutations are often partially recessive with moderate effects in heterozygotes[47]. In contrast, GoF mutations are usually dominant. The second explanation comes from population genetics. We observe very limited numbers of GoF CES drivers in a single gene (Fig 1e). On the other hand, since all LoF mutations within a gene have identical functional consequences, the number of potential CES driver mutations is large. As CES drivers are frequently involved in severe diseases, the GoF-like 500-fold elevation in mutation rate might generate a substantial rate of *de novo* pathogenic mutations, e.g. is the same inflation was present for LoF mutations in *ARID1B*, about 0.5% of all individuals would be born with NDD if the CES effect was so high. Such pathogenic burden created by LoF mutations in a single gene will trigger efficient negative selection on the CES effect.

Impact of CES drivers on NDD

Here, we develop a series of tests to identify CES drivers using the counts of *de novo* mutations in cohorts ascertained by disease, primarily NDD. CES effect and disease ascertainment independently increase the counts of *de novo* mutations in these cohorts. This gives us an upper

hand to find CES genes with involvement in NDD as compared to non-ascertained cohorts like trio controls or sperm sequencing. Indeed, out of 40 recovered genes, 26 have orthogonal functional or genetic evidence of involvement in NDD. As expected based on the properties of the tests discussed above, GoF and LoF-1 gene sets are more enriched in genes causal for NDD. This falls in line with the notion of CES determining the prevalence of multiple types of NDD. A well-known example is given by the Noonan syndrome where clonal expansions underlie both etiology and prevalence of the disease. For Noonan syndrome mutations, the effect observed in NanoSeq is 440-fold (CE 91-1285), which resolves the controversy between the net baseline mutation rate of $\sim 10^{-7}$ and the prevalence of the syndrome of $\sim 10^{\wedge -4}$. Remarkably, prevalence of CES-related diseases such as Apert syndrome and achondroplasia scales exponentially with paternal age[8,48,49]. Therefore, it is expected that prevalence of any CES-related condition would similarly show the exponential dependency on the age of the father.

On the other hand, many of the recovered CES driver candidates, especially the ones from the LoF-2 set, e.g. *MIB1* or *TCF7L2*, have no strong additional supporting evidence for NDD meaning that they could be false-positives in studies of gene-disease association that rely solely on *de novo* enrichment. However, these studies may find evidence of the causal role in NDD not confounded by CES such as: i. transmission distortion and other types of familial segregation (Extended Data Figure 7), ii. case-control analyses, iii. phenotypic similarity of mutation carriers and iv. functional assays *in vitro* and *in vivo*.

Other modulators of mutation counts

Aside from CES and disease ascertainment, one modulator of the observed *de novo* counts is embryonic deleteriousness (partial lethality) (Table 1). It is expected to decrease the observed mutation rates and therefore renders our approaches for finding CES genes conservative. The comparison of *de novo* counts between sperm sequencing and trio sequencing has a potential to highlight lethal mutations. Following this logic, we find that LoF mutations in *RASA2* are consistent with embryonic deleteriousness.

Although here we interpret the elevation of mutation rate relative to the baseline as the effect of CES, the mutation rate increase may be due to clonal expansions in early development or in oocyte progenitors. Formally, our approach does not pinpoint the source of the clonal expansions. Still, substantial overlap (given statistical limitations) with sperm sequencing data supports the CES hypothesis. No mutations reported by Neville et al. have appreciable variant allele frequency in sperm[23], ruling out early developmental origin. Historically, CES is the only reported clonal mechanism responsible for the elevation of germline mutation rate.

In this study we have shown that the phenomenon of CES might be of unexpectedly great importance for future studies in the fields of genetics of rare disease, population genetics, cancer biology and in the emerging field of clonal evolution in somatic tissues.

Table 1 | Different data modalities offer possibilities for distinguishing types of modulators of observed *de novo* mutation rates

	NDD	Clonal expansions			Embryonic deleteriousness
		Spermatogonia	Early development	Oocyte progenitors	
***de novo* effect in cases**	↑	↑	↑	↑	↓
***de novo* effect in controls**	0	↑	↑	↑	↓
Effect in sperm sequencing	0	↑	↑	0	0
Segregation in population	↓	↑	↑	↑	↓
Other considerations					
Disease phenotype	yes	no	no	no	no
Somatic mosaicism	no	no	yes	no	no
Parental bias	baseline	paternal	50/50	maternal	baseline

References

1. Lawson, A. R. J. *et al.* Somatic mutation and selection at epidemiological scale. 2024.10.30.24316422 Preprint at https://doi.org/10.1101/2024.10.30.24316422 (2024).
2. Maeda, H. & Kakiuchi, N. Clonal expansion in normal tissues. *Cancer Sci.* **115**, 2117–2124 (2024).
3. Bernstein, N. *et al.* Analysis of somatic mutations in whole blood from 200,618 individuals identifies pervasive positive selection and novel drivers of clonal hematopoiesis. *Nat. Genet.* **56**, 1147–1155 (2024).
4. Fowler, J. C. & Jones, P. H. Somatic Mutation: What Shapes the Mutational Landscape of

Normal Epithelia? *Cancer Discov.* **12**, 1642–1655 (2022).

5. Goriely, A., McVean, G. A. T., Röjmyr, M., Ingemarsson, B. & Wilkie, A. O. M. Evidence for selective advantage of pathogenic FGFR2 mutations in the male germ line. *Science* **301**, 643–646 (2003).
6. Tiemann-Boege, I. *et al.* The observed human sperm mutation frequency cannot explain the achondroplasia paternal age effect. *Proc. Natl. Acad. Sci. U. S. A.* **99**, 14952–14957 (2002).
7. Maher, G. J. *et al.* Selfish mutations dysregulating RAS-MAPK signaling are pervasive in aged human testes. *Genome Res.* **28**, 1779–1790 (2018).
8. Wood, K. A. *et al.* SMAD4 mutations causing Myhre syndrome are under positive selection in the male germline. *Am. J. Hum. Genet.* **111**, 1953–1969 (2024).
9. Kaplanis, J. *et al.* Evidence for 28 genetic disorders discovered by combining healthcare and research data. *Nature* **586**, 757–762 (2020).
10. Hamdan, F. F. *et al.* High Rate of Recurrent De Novo Mutations in Developmental and Epileptic Encephalopathies. *Am. J. Hum. Genet.* **101**, 664–685 (2017).
11. Jin, S. C. *et al.* Contribution of rare inherited and de novo variants in 2,871 congenital heart disease probands. *Nat. Genet.* **49**, 1593–1601 (2017).
12. Zhou, X. *et al.* Integrating de novo and inherited variants in 42,607 autism cases identifies mutations in new moderate-risk genes. *Nat. Genet.* **54**, 1305–1319 (2022).
13. Richter, F. *et al.* Genomic analyses implicate noncoding de novo variants in congenital heart disease. *Nat. Genet.* **52**, 769–777 (2020).
14. Francioli, L. C. *et al.* Genome-wide patterns and properties of de novo mutations in humans. *Nat. Genet.* **47**, 822–826 (2015).
15. Iossifov, I. *et al.* The contribution of de novo coding mutations to autism spectrum disorder. *Nature* **515**, 216–221 (2014).
16. Genome of the Netherlands Consortium. Whole-genome sequence variation, population structure and demographic history of the Dutch population. *Nat. Genet.* **46**, 818–825 (2014).

17. An, J.-Y. *et al.* Genome-wide de novo risk score implicates promoter variation in autism spectrum disorder. *Science* **362**, eaat6576 (2018).
18. Gulsuner, S. *et al.* Spatial and temporal mapping of de novo mutations in schizophrenia to a fetal prefrontal cortical network. *Cell* **154**, 518–529 (2013).
19. Zhao, G. *et al.* Gene4Denovo: an integrated database and analytic platform for de novo mutations in humans. *Nucleic Acids Res.* **48**, D913–D926 (2020).
20. Karczewski, K. J. *et al.* The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature* **581**, 434–443 (2020).
21. Amberger, J. S., Bocchini, C. A., Scott, A. F. & Hamosh, A. OMIM.org: leveraging knowledge across phenotype–gene relationships. *Nucleic Acids Res.* **47**, D1038–D1043 (2019).
22. Halldorsson, B. V. *et al.* Characterizing mutagenic effects of recombination through a sequence-level genetic map. *Science* **363**, eaau1043 (2019).
23. Neville, M. D. *et al.* Sperm sequencing reveals extensive positive selection in the male germline. 2024.10.30.24316414 Preprint at https://doi.org/10.1101/2024.10.30.24316414 (2024).
24. Seplyarskiy, V. B. & Sunyaev, S. The origin of human mutation in light of genomic data. *Nat. Rev. Genet.* **22**, 672–686 (2021).
25. Seplyarskiy, V. *et al.* A mutation rate model at the basepair resolution identifies the mutagenic effect of polymerase III transcription. *Nat. Genet.* **55**, 2235–2242 (2023).
26. Totsika, V., Liew, A., Absoud, M., Adnams, C. & Emerson, E. Mental health problems in children with intellectual disability. *Lancet Child Adolesc. Health* **6**, 432–444 (2022).
27. Raudvere, U. *et al.* g:Profiler: a web server for functional enrichment analysis and conversions of gene lists (2019 update). *Nucleic Acids Res.* **47**, W191–W198 (2019).
28. Tate, J. G. *et al.* COSMIC: the Catalogue Of Somatic Mutations In Cancer. *Nucleic Acids Res.* **47**, D941–D947 (2019).
29. Boßelmann, C. M. *et al.* Analysis of 1386 epileptogenic brain lesions reveals association

with DYRK1A and EGFR. *Nat. Commun.* **15**, 10429 (2024).

30. Gogate, A. *et al.* The genetic landscape of autism spectrum disorder in an ancestrally diverse cohort. *NPJ Genomic Med.* **9**, 62 (2024).
31. Gallego-Martinez, A. *et al.* Using coding and non-coding rare variants to target candidate genes in patients with severe tinnitus. *NPJ Genomic Med.* **7**, 70 (2022).
32. Wainberg, M. *et al.* Deletion of Loss-of-Function-Intolerant Genes and Risk of 5 Psychiatric Disorders. *JAMA Psychiatry* **79**, 78–81 (2022).
33. Rodan, L. H. *et al.* Gain-of-function variants in the ODC1 gene cause a syndromic neurodevelopmental disorder associated with macrocephaly, alopecia, dysmorphic features, and neuroimaging abnormalities. *Am. J. Med. Genet. A.* **176**, 2554–2560 (2018).
34. Jansen, S. *et al.* De Novo Truncating Mutations in the Last and Penultimate Exons of PPM1D Cause an Intellectual Disability Syndrome. *Am. J. Hum. Genet.* **100**, 650–658 (2017).
35. Harpak, A., Bhaskar, A. & Pritchard, J. K. Mutation Rate Variation is a Primary Determinant of the Distribution of Allele Frequencies in Humans. *PLoS Genet.* **12**, e1006489 (2016).
36. Schraiber, J. G., Spence, J. P. & Edge, M. D. Estimation of demography and mutation rates from one million haploid genomes. 2024.09.18.613708 Preprint at https://doi.org/10.1101/2024.09.18.613708 (2024).
37. Wakeley, J., Fan, W.-T. L., Koch, E. & Sunyaev, S. Recurrent mutation in the ancestry of a rare variant. *Genetics* **224**, iyad049 (2023).
38. Nei, M. The frequency distribution of lethal chromosomes in finite populations. *Proc. Natl. Acad. Sci.* **60**, 517–524 (1968).
39. Zeng, T., Spence, J. P., Mostafavi, H. & Pritchard, J. K. Bayesian estimation of gene constraint from an evolutionary model with gene features. *Nat. Genet.* **56**, 1632–1643 (2024).
40. Szklarczyk, D. *et al.* The STRING database in 2021: customizable protein-protein networks,

and functional characterization of user-uploaded gene/measurement sets. *Nucleic Acids Res.* **49**, D605–D612 (2021).

41. Schuurs-Hoeijmakers, J. H. M. *et al.* Recurrent de novo mutations in PACS1 cause defective cranial-neural-crest migration and define a recognizable intellectual-disability syndrome. *Am. J. Hum. Genet.* **91**, 1122–1127 (2012).
42. Olson, H. E. *et al.* A Recurrent De Novo PACS2 Heterozygous Missense Variant Causes Neonatal-Onset Developmental Epileptic Encephalopathy, Facial Dysmorphism, and Cerebellar Dysgenesis. *Am. J. Hum. Genet.* **102**, 995–1007 (2018).
43. Goriely, A. *et al.* Activating mutations in FGFR3 and HRAS reveal a shared genetic origin for congenital disorders and testicular tumors. *Nat. Genet.* **41**, 1247–1252 (2009).
44. Ichimura, K. *et al.* Recurrent neomorphic mutations of MTOR in central nervous system and testicular germ cell tumors may be targeted for therapy. *Acta Neuropathol. (Berl.)* **131**, 889–901 (2016).
45. Kimura, T. *et al.* Conditional loss of PTEN leads to testicular teratoma and enhances embryonic germ cell production. *Dev. Camb. Engl.* **130**, 1691–1700 (2003).
46. Sommerer, F. *et al.* Mutations of BRAF and RAS are rare events in germ cell tumours. *Int. J. Cancer* **113**, 329–335 (2005).
47. Knudson Hypothesis - an overview | ScienceDirect Topics. https://www.sciencedirect.com/topics/medicine-and-dentistry/knudson-hypothesis.
48. Goriely, A. & Wilkie, A. O. M. Paternal age effect mutations and selfish spermatogonial selection: causes and consequences for human disease. *Am. J. Hum. Genet.* **90**, 175–200 (2012).
49. Penrose, L. S. Parental age and mutation. *Lancet Lond. Engl.* **269**, 312–313 (1955).

Acknowledgements

The authors would like to thank Drs. Carles Biox, Colby Chang, and Remus Stana as well as Prof. Aaron Quinlan and Jason Kunisaki for valuable discussions and helpful suggestions that improved the analyses presented in this study. This work was supported by the National Institutes of Health through grants R35GM12713, R01MH101244, and U01HG012009.

Author contributions

V.S., M.A.M., and S.S. jointly conceived the study and developed the methodological framework. V.S., M.A.M., E.K., P.K., and M.D.C.N. performed data analysis and interpreted results. V.S., M.A.M., E.K., P.K., M.D.C.N., R.R., and S.S. collaboratively drafted and revised the manuscript. S.S., R.R., V.S., and M.A.M. jointly supervised the project. All authors reviewed and approved the final version of the manuscript.

Data availability

All data used in this study, including intermediate processed datasets, are provided in the Supplementary Tables. Associated summary files have also been deposited on Zenodo at https://zenodo.org/records/15660433.

Code availability

The full analysis pipeline, including code to reproduce figures and statistical analyses, is available on GitHub at https://github.com/mikemoldovan/CES_Discovery.

Conflict of interest statement

The authors declare no competing interests.

Captions for the main figures

Figure 1 | Clonal expansions in spermatogonia and disease ascertainment. (a) Effect of clonal expansions in sperm on the observed mutation rate **(b)** In a randomly sampled cohort of individuals affected by a specific disease, the frequency of any variant is elevated due to ascertainment by at most the inverse of disease prevalence. *P(D)* is disease prevalence and μ is the mutation rate **(c)** Observed synonymous *de novo* variant counts in the neurodevelopmental disorder (NDD) cohort published[9] by Kaplanis et al., 2020 stratified by Roulette mutation rate

bins vs. Roulette predictions. 95% Poisson confidence intervals are shown. **(d)** Fractions of genes included in Reactome RAF activation and MAPK1/MAPK3 pathways. P-values of Fisher's exact test for comparison of fractions below 0.001 are shown as triple asterisks (***). **(e)** CES driver candidates due to GoF.

Figure 2 | LoF-1 set of putative CES drivers. (a) Observed vs. expected LoF counts in the NDD cohort. LoF-1 genes are shown in red. **(b)** Observed synonymous *de novo* mutation counts in the NDD cohort vs. the ones expected under the Poisson counts around Roulette estimates **(c)** Properties of LoF-1 genes **(d)** Ratio of observed to expected *de novo* LoF variant counts (o/e) for genes at FDR < 0.1 in the neurodevelopmental disorder (NDD) cohort published[9] by Kaplanis et al., 2020. The outliers with the ratio above the maximal ascertainment (LoF-1 genes) are highlighted in red.

Figure 3 | Effect of CES on *de novo* mutations in disease and on the levels of LoF polymorphism in population. (a) Observed-to-expected variant count ratio (o/e) for *de novo* LoFs in genes with FDR < 0.1 in the neurodevelopmental disorder cohort (NDD)9 merged with the autism spectrum disorder (ASD)[12] cohort plotted against LOUEF scores. The dashed violet line indicates the 0.5 threshold for LOEUF used to construct the LoF-2 set, and the histogram to the right shows LOEUF values. **(b)** Observed-to-expected LoF variant count ratio in four different trio cohorts. Violet bars indicate LoF-2 genes and empty grey bars indicate other genes significant at FDR < 0.1 in the ASD/NDD cohort. 95% Poisson confidence intervals are shown. Counts of variants are shown below the bars. **(c)** Properties of LoF-2 genes. The single gene not expressed in spermatogonia (*KCNA1*) is shown in grey **(b)** Frequency of LoF in gnomAD-v4 against mean scaled mutation rate predicted by Roulette in *MIB1*. LoF sites were aggregated by mutation rate and each point represents the mean taken for at least 10 sites. **(e)** x-axis: inflation of the variance of the LoF allele frequencies in gnomAD-v4 due to random genetic drift proportional to the inverse selection coefficient (1/s). y-axis: mean allele frequency scaled by mutation rate equal to CES effect multiplied by 1/s. Each point corresponds to an individual gene with at least 10 LoF CpG sites.

Figure 4 | Comparison with direct sequencing data (a) Observed-to-expected ratio for LoFs counts in NanoSeq sequencing of sperm for genes with significant excess of LoFs. Poisson-conjugate (Gamma) 95% confidence intervals are shown. Grey bar indicates that *PPM1D* has been flagged in our previous analysis. Venn diagram shows the numbers of genes in sets 1 and 2 of genes significant due to LoFs in our study, number of genes significant due to LoFs in NanoSeq study[23] and the number of common genes. Numbers in parentheses include *PPM1D*. Below, overlaps with set 1 and set 2 genes are shown. **(b)** Enrichments *de novo* mutations in controls for genes overlapping between NanoSeq and LoF-1 + LoF-2, genes private to NanoSeq, and genes private for set 1 + set 2. Number of asterisks indicates the Binomial significance threshold: (*): 0.05, (**): 0.01, (***): 0.001, NS: non-significant. Counts of variants

are shown below the bars. **(c)** Core pathway of RAF activation. Gray dashed frame shows protein complexes, activator and repressor activities are denoted with pointed and blunt arrows, respectively.

Captions for extended data figures

Extended Data Figure 1 | Counts of *de novo* variants in the NDD cohort. (a) Numbers of *de novo* missense variants stratified by recurrence. Genes harboring variants occurring >10 times in the cohort are shown in blue. **(b)** Numbers of *de novo* loss-of-function variants aggregated by gene stratified by recurrence. **(c)** Scatter plot of observed vs. expected *de novo* loss-of-function variant counts in the NDD cohort. LoF-1 set genes are shown in red; LoF-2 set genes are shown in purple. Upper bound of disease ascertainment in Eq (1) given by the lower bound of prevalence of 1% is shown as a dashed line.

Extended Data Figure 2 | Expression of the identified CES drivers in germline tissues. (a) nTPM values for spermatogonia reported in The Human Protein Atlas single-cell dataset. **(b)** nTPM values normalized by the maximal expression across all tissues for each gene. **(c)** nTPM values in oocytes. **(d)** Normalized nTPM values in oocytes.

Extended Data Figure 3 | Stability of LoF-2 set with respect to the metric of loss-of-function constraint. (a) Observed-to-expected variant count ratio (o/e) for *de novo* LoFs in genes with FDR < 0.1 in the neurodevelopmental disorder cohort (NDD) merged with the autism spectrum disorder (ASD) cohort plotted against the Loss-of-function Observed/Expected Lower-bound Fraction (LOELF) scores. The dashed violet line indicates the minimal LOELF value across LoF-2 genes of 0.23. LoF-2 genes are shown in violet, LoF-1 genes are shown in red, genes above the chosen LOELF threshold but not included in the LoF-2 set (*SHANK3* and *ZMYM2*) are shown in black. **(b)** Same as in (a), but for the Loss-of-function Observed/Expected (LOE) metric. The upper bound for LOE (shown as violet dashed line) is 0.355.

Extended Data Figure 4 | Ratio of observed-to-expected counts of LoF de novo mutations in a cohort of control trios for LoF-2 set genes. The ratio is shown as a function of the LOEUF threshold: we aggregate all genes with LOEUF values lower than the value indicated on the x-axis and calculate the cumulative observed-to-expected ratio. The shaded grey area represents the 95% confidence interval obtained by permuting the LOEUF labels.

Extended Data Figure 5 | Validation of LoF-2 genes with a non-LOE metric. Prior of the GeneBayes s_{het} calculated using biological features of genes (x-axis) and the s_{het} values updated with LoF polymorphism data from gnomAD-v4 (y-axis). See Methods-8 for details.

Extended Data Figure 6 | Expected LoF rate in NanoSeq data for LoF-1 and LoF-2 genes.

Rates are shown separately for genes overlapping with those significant in NanoSeq and for private LoF-1/2 genes. An asterisk (*) indicates $p < 0.05$ from the Mann–Whitney U test.

Extended Data Figure 7 | Power analysis of paternal transmission. Statistical power (i.e., the probability of correctly detecting a signal when it exists calculated as the complement of type-2 error rate ß) of the Binomial test for paternal overtransmission relative to the baseline of 0.75 is shown across the range of CES-related mutation rate inflations κ and counts of observed variants. Results are presented for three significance levels: 0.05, 0.01, and 0.001.

Methods

1. Data

We used the dataset of *de novo* variants in probands affected by neurodevelopmental disorders (NDD) published by Kaplanis et al.[9]. For autism spectrum disorders (ASD), we used the *de novo* portion of the SFARI dataset published by Zhou et al.[12]. For congenital heart diseases (CHD), encephalopathic epilepsy (EE) and trios with healthy offspring used as controls, we used aggregated datasets of *de novo* variants in cohorts with probands affected by the respective conditions[9–19,50–54,]. For any non-NDD and non-ASD cohorts, we filtered out probands with duplicate IDs and samples where *de novo* variants were not obtained by whole genome or whole exome sequencing. Next, for the cohorts with *de novo* variants reported in hg19 coordinates, we mapped the coordinates to hg38 using LiftOver[55].

A recently published NanoSeq dataset of sperm cell sequencing was downloaded from the supplementary materials of Neville et. al., 2024 and candidates due to LoF were taken as genes with the LoF enrichment at the FDR < 0.2 level given in the original table (*qtrunc_cv* column). COSMIC cancer tiers of the putative CES genes and observed recurrence of variants in these genes with respect to cancer association, were obtained from COSMIC database on October 25th, 2024[28]. The list of genes causing clonal expansions in hematopoietic tissues was obtained from Bernstein et al., 2024[3].

Data on single-cell gene expression in sperm progenitors and oocytes were downloaded from The Human Protein Atlas[56] and data on single-cell RNAseq of early human embryos was obtained from the Supplementary information of the study by Xu et al., 2023[57].

The structure of the MAPK pathway shown on Fig. 4d was obtained from WikiPathways/WP382[58]. For the per-site mutation rate estimates and quality control tracks, we used Roulette-MR and Roulette-QUAL features of Roulette, respectively[25].

2. Processing of genes

Since Roulette mutation rates are available only for autosomes, we restricted all analyses to autosomal genes. To avoid ambiguity in assigning variants, we further limited our dataset to genes with non-overlapping coding sequences (CDS). For any pair of genes with overlapping CDS regions, we retained the longer gene and excluded the shorter one. In cases where more

than two genes overlapped, we selected the largest subset of non-overlapping genes that maximized the total CDS length. This filtering yielded a final set of 17,791 autosomal protein-coding genes, listed in Supplementary Table S14.

3. Processing of variants

We included only coding variants with a "high" or "TFBS" quality score in the Roulette-QUAL track. For each gene, variants were assigned to one of three functional categories. First, loss-of-function (LoF) variants were defined as those with both LOFTEE[20] high-confidence annotations and a VEP[59] consequence of *stop_gained*, *splice_donor_variant*, or *splice_acceptor_variant*. Second, synonymous variants were defined as those with a VEP annotation[59] containing only the term "synonymous," excluding any variants [possibly] involved in splicing. Third, we included missense variants only if they were annotated as "missense" in VEP and had AlphaMissense[60] scores greater than 0.1.

4. Mutational expectation

To scale mutational expectations in an unbiased way, we normalized by the observed number of synonymous *de novo* mutations. This approach accounts for cohort-specific factors such as average parental age and sequencing coverage. For a given variant v, we define the expected number of occurrences in a *de novo* dataset as:

$$\lambda_v := N_S \frac{\mu_v}{\sum_{v \in S} \mu_v},$$

where N_S is the total number of synonymous *de novo* mutations observed in the cohort, S is the set of all admissible synonymous variants, and μ_v is the unscaled Roulette mutation rate assigned to v. This formulation reflects the assumption that *de novo* mutation events occur independently, allowing expected counts to be proportional to the summed rates.

For any annotation $A = \{v\}$, the expected number of *de novo* mutations is:

$$\lambda_A := \sum_{v \in A} \lambda_v = N_S \frac{\sum_{v \in A} \mu_v}{\sum_{v \in S} \mu_v}. \tag{1}$$

This enables comparison of observed and expected mutation counts across functional categories while controlling for mutation rate variation and technical biases.

5. Control for Biases in Roulette

To test for overdispersion in Roulette estimates at the single-variant level, we compared two models of counts of *de novo* synonymous mutations in the NDD cohort: a Poisson model assuming no overdispersion, and a Negative Binomial model assuming overdispersion by a

factor of $\gamma + 1$. Using expected mutation counts defined as in Eq. (1), the likelihood under the Poisson model is:

$$L_0 = \prod_{v \in S} e^{-\lambda_v} \frac{\lambda_v^{n_v}}{n_v!}, \tag{2}$$

where n_v is the observed number of recurrent *de novo* mutations at site v. The likelihood under the Negative Binomial model, with variance inflated by a factor of $\gamma + 1$, is:

$$L_\gamma = \prod_{v \in S} \binom{n_v + \frac{\lambda_v}{\gamma} - 1}{n_v} \left(\frac{\gamma}{\gamma+1}\right)^{n_v} \left(\frac{1}{\gamma+1}\right)^{\frac{\lambda_v}{\gamma}}, \tag{3}$$

using the parametrization of the Negative Binomial distribution by mean λ_v and overdispersion factor γ, such that the variance is $\lambda_v(1 + \gamma)$. Since Eq. (3) can be rewritten as a Poisson–Gamma mixture, λ_v/γ need not be an integer.

After maximizing the likelihood in Eq. (3) over γ, we obtain $\hat{\gamma} = 0.004$. Comparing the log-likelihoods of Eq. (2) and Eq. (3), we find a difference of $logL_\gamma - logL_0 = 31.1$. Applying the Akaike Information Criterion (AIC), which penalizes additional parameters by 2 log-likelihood units, this strongly supports the presence of modest but statistically significant overdispersion. As a result, we use the Negative Binomial distribution with $\gamma = 0.004$ rather than the Poisson model when modeling individual mutations in the GoF gene set.

Beyond single-site effects, Roulette estimates might have regionally correlated errors resulting in some regions having generally over- or underpredicted mutation rates. In particular, correlated errors in residuals could systematically distort the expected counts for specific regions of interest, such as protein-coding genes and may be undetectable at the per-site level. To assess this, we tested for overdispersion in the number of synonymous *de novo* mutations aggregated by gene. Let n_g be the observed and λ_g the expected count of synonymous mutations in gene g, with λ_g obtained by summing site-specific rates as in Eq. (1). We then substitute $n_v \to n_g$ and $\lambda_v \to \lambda_g$ in Eq. (2) and Eq. (3), and re-estimate the log-likelihoods.

In this gene-level analysis, the log-likelihood difference is $logL_\gamma - logL_0 = 0.69$, which falls well below the AIC threshold for including an extra parameter. Thus, we find no evidence of significant variance inflation at the gene level.

To further test for overdispersion in the gene-level expectations λ_g, we used a regression-based approach. Applying the law of total variance:

$$Var(n_g) = E[Var(n_g|\lambda_g)] + Var[E(n_g|\lambda_g)],$$

where the first term captures residual (Poisson) noise and the second term reflects underlying variation in λ_g. Since $E(n_g|\lambda_g) = \lambda_g$ and $Var(n_g|\lambda_g) = \lambda_g$ under the Poisson model, we expect:

$$Var(n_g) = E(\lambda_g) + Var(\lambda_g).$$

To test this, we computed sample estimates of each term and tested the Poisson assumption with the chi-squared goodness-of-fit test on statistic

$$\chi^2 = \sum_g \frac{(n_g - \lambda_g)^2}{\lambda_g},$$

with degrees of freedom equal to the number of genes minus 1 (17,790). The resulting p-value of 0.6 confirms consistency with the Poisson model, with no evidence of excess dispersion. We thus conclude that gene-level Roulette residuals are effectively uncorrelated, and that variance inflation in gene-level mutational expectations is negligible.

Consequently, for gene-level aggregation of variant counts in LoF-1 and LoF-2 sets, we used Poisson tests to assess enrichments.

6. Statistics

6.1. Phenotypic homogeneity in the NDD cohort

Our procedures yielding LoF-1 and GoF sets of putative CES drivers assume uniform phenotypic sampling across the NDD cohort (see Main Text and below). Deviations from uniformity—e.g., overrepresentation of certain syndromes—could result in differential ascertainment of associated mutations, potentially inflating effect sizes for certain genes and introducing false positives to LoF-1 and LoF-2 sets.

To assess homogeneity of recruitment in the composite NDD cohort used here, we compared contributions of the three sub-cohorts (DDD, GeneDx and RUMC) comprising it to *de novo* LoF variation in individual genes. If the cohorts were sampled in a uniform way, the fractions of *de novo* variants coming from each cohort across all genes should correspond to the per-cohort mutation rate, which is proportional to the number of synonymous mutations per individual cohort (Eq. 1). The numbers of variants in each gene should therefore come from the Multinomial distribution with probabilities given by per-cohort synonymous variant counts and the number of events given by the number of observed *de novo* variants.

We tested this hypothesis by comparing four models:

1. A Multinomial model with probabilities fixed by observed synonymous variant counts (0 free parameters).
2. A Multinomial model with unconstrained probabilities to test for shifts in representation (2 free parameters due to 3 probabilities which need to sum up to 1).
3. A Dirichlet-Multinomial model centered on the observed synonymous proportions, allowing for overdispersion (1 free parameter: amplitude of Dirichlet).
4. A fully flexible Dirichlet-Multinomial model accounting for both overdispersion and shifted means (3 free parameters: 2 for unconstrained Multinomial and 1 additional parameter for the amplitude of Dirichlet).

For each model, we computed the total likelihood across all genes and compared models using the Akaike Information Criterion (AIC). The unbiased Multinomial model (Model 1) was the one

preferred by AIC, indicating that the observed gene-level LoF variant distributions are consistent with uniform cohort sampling.
To further validate that our LoF-1 and LoF-2 sets are not confounded by non-uniform sampling, we repeated the analysis on specifically LoF-1 and LoF-2 putative CES drivers. In both cases, the AIC again favored the null Multinomial model. Together, these results suggest that recruitment across the DDD, GeneDx, and RUMC cohorts was phenotypically homogeneous with respect to LoF variant ascertainment. Full likelihood and AIC values for all model comparisons are reported in Supplementary Table S2.

6.2. Prevalence of NDD

The identification of putative CES drivers in the LoF-1 and GoF sets relies on the assumption that the disease prevalence estimate $P(D)$ used in Eq. 1 (Main Text) is not inflated. Overestimating $P(D)$ would lead to underestimating the maximum plausible effect size attributable to ascertainment, thereby increasing the risk of false positive CES calls in highly penetrant genes.
Epidemiological estimates of neurodevelopmental disorder (NDD) prevalence—based on diagnoses of intellectual disability or global developmental delay—typically fall within the range of 2–3%[26]. To ensure a conservative interpretation, we imposed a lower bound of 1% on $P(D)$, which corresponds to an upper bound of 100 on the effect of ascertainment.
One additional consideration suggests that 1% is an appropriate lower boundary for NDD prevalence. Almost all significantly mutated genes have the effects below 100 with very few genes greatly exceeding this threshold (Fig. 3d). The possibility that the prevalence is even lower than 1% and the observed distribution of gene effects is fully explained by disease ascertainment is inconsistent with the existence of genes with substantial penetrance for NDD.

6.3. Construction of the GoF set

To define the GoF set of putative CES driver genes, we applied a Negative Binomial test. The expected count of *de novo* variants at position v was set to $\mu_v/P^*(D)$, where μ_v is the Roulette mutation rate at v, and $P^*(D)$ is the lower bound on disease prevalence, fixed at 0.01 (see Main Text). To account for overdispersion at the single-site level, we used a variance inflation parameter $\gamma = 0.004$, as defined in Eq. (3). For multiple testing correction, we applied a Bonferroni adjustment using a correction factor equal to the total number of autosomal missense variants annotated by VEP and filtered to have an AlphaMissense score above 0.1 (49,686,008 tests).

6.4. Construction of the LoF-1 set

To define the LoF-1 set of putative CES driver genes, we applied a Poisson test to LoF *de novo* variants observed in the NDD cohort. The expected count of variants in gene g was set to $\lambda_g/P^*(D)$, where λ_g is the LoF mutation rate for gene g as defined in Eq. (1), and $P^*(D)$ is the

lower bound on disease prevalence, fixed at 0.01 (see Main Text). As in the GoF analysis, this expectation corresponds to the number of events expected under maximal disease ascertainment (i.e., full penetrance). The choice of the Poisson model is justified in the "Control for Biases in Roulette" section above. For multiple testing correction, we used a Bonferroni factor equal to the number of autosomal genes with at least one filtered LoF site (17,791 tests). This yielded five Bonferroni-significant genes at the 0.05 level, and one additional gene (*FOXG1*) significant at FDR < 0.1, using the Benjamini–Hochberg procedure (Fig. 2ad).

6.5. Construction of the LoF-2 set

The LoF-2 set of putative CES driver genes was constructed using the combined ASD + NDD cohort in two steps. First, we identified genes with a significant excess of *de novo* LoF variants relative to expectation ($n_g > \lambda_g$), where n_g is the observed LoF count and λ_g is the Roulette-based expectation for gene g. As in the LoF-1 analysis, we used a Poisson test with correction for 17,791 tested autosomal genes (Supplementary Table S14). On the second step, we defined the LoF-2 set as those genes passing a false discovery rate threshold of FDR < 0.1 and having gnomAD v4 LOEUF scores greater than 0.5 (Fig. 2a).

6.6. Excess of LoF Variants in NDD Compared to NanoSeq

To test whether the excess of *de novo* LoF variants observed in the NDD cohort exceeded that in the NanoSeq sperm dataset, we used a Binomial test. For each gene g, we set the probability of success to the expected fraction of LoF mutations arising from NanoSeq, defined as $\frac{\lambda_{m,g}}{\lambda_{m,g}+\lambda_n}$, where $\lambda_{m,g}$ is the expected LoF count in NanoSeq and λ_n is the expected LoF count in NDD. The number of trials was set to $n_g + m_g$, where n_g and m_g are the observed LoF counts in gene g in the NDD and NanoSeq datasets, respectively. The test compared the observed fraction $\frac{m_g}{m_g+n_g}$ to the expected null fraction. Only genes with at least one LoF variant in both datasets were tested. A gene was deemed significant if its P-value, Bonferroni-corrected for the number of tests ($N = 5$), was below 0.05. This test identified only *CTNNB1* and *PTEN* as significantly enriched for LoF variants in the NDD cohort relative to NanoSeq. Since the NanoSeq donors were older on average than the parents of individuals in the NDD cohort, this comparison is conservative. The test for the excess of variants in NanoSeq compared to excess in the NDD cohort highlights *CUL3* and *ARID1A* as genes with a larger excess in NanoSeq. Although this effect may be confounded by the age differences between the parents in the NDD cohort and the donors of samples used in the NanoSeq experiment, an alternative explanation may be embryonic deleteriousness of LoF variants in these genes.

7. Expression and Gene Set Enrichment Analyses

Expression analysis of candidate genes was performed using single-cell data from The Human Protein Atlas. For each gene, the Atlas reports TPM expression values across a wide range of tissues. To account for gene-specific differences in baseline expression, we normalized the TPM values in spermatogonia and oocytes by the maximum TPM observed for each gene across all tissues. To statistically assess differences in expression patterns between CES driver candidates and other genes, we applied the Mann–Whitney U test to the normalized TPM values.
Gene set enrichment analysis was performed using the g:Profiler[27] web service with default parameters. We tested five gene sets: the GoF, LoF-1, and LoF-2 sets individually; the union of LoF findings (LoF-1 and LoF-2); and the union of all three sets. Full g:Profiler results are provided in Supplementary Tables S15-S18.

8. GeneBayes update and CES

To obtain estimates of LoF constraint in the LoF-2 set of genes unimpacted by mutation rate misspecification we used prior estimates of the selection coefficient against heterozygotes, produced by the GeneBayes method. These estimates result from a prior distribution fit to all LoF variation using a set of global gene features and should be minimally impacted by variation in any one gene[39]. To assess the robustness of the resulting list of genes to the constraint estimation, we additionally performed an orthogonal analysis of differences between prior and posterior GeneBayes s_{het} estimates, which were obtained on gnomAD-v3. As expected under moderate selection against LoF mutations and CES, prior estimates of s_{het} were substantially higher than those estimated using polymorphism alone. The rationale here is that, given downwardly biased mutation rate estimates of CES genes, the data should shift the prior down by factors larger than those expected for other genes. This analysis recapitulates our results obtained with LOEUF.
Due to the demographic complexity of the gnomAD v4 sample, we used a data-driven approach to estimate s_{het} for each gene. We first fit how both the proportion of polymorphic sites as well as the shape of the site frequency spectrum depend on s_{het}, conditional on μ. For this, we fit a multinomial model to the observed allele counts using prior estimates of as a covariate. To ensure that the SFS was not too sparse at higher allele counts, we binned sites by allele count with boundaries $[0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 16, 24, 36, 100]$. The observed distribution of alleles counts in each Roulette mutation rate bin was then transformed to multinomial coefficients such that

$$log\frac{P(K=i|\mu)}{P(K=0|\mu)} = \beta_i^{\mu} \tag{4}$$

This model for the neutral SFS in each bin was first estimated using synonymous SNV counts. Next, we fit a model to all LoF polymorphism such that

$$log\frac{P(K=i|\mu,)}{P(K=0|\mu,)} = \beta_i^{\mu} - \beta_i^{s}\sqrt{s_{het}}. \tag{5}$$

The β_i^{μ} coefficients were determined by the synonymous SFS and β_i^{s} coefficients were fit with maximum likelihood using the *L-BFGS-B* algorithm as implemented in scipy. The square root

dependency on was determined by manual inspection of the SFS, binned by s_{het} and μ. The values used to fit this global multinomial model were prior mean values output by GeneBayes. This procedure yielded a model for the SFS in each mutation rate bin, given a value of s_{het}. We then re-estimated for LoF mutations in each gene using this model using maximum likelihood and bounding to $[0, 1]$.

9. Variants in LoF-2 Genes Segregate in the Population Due to Elevated Mutation Rate

LoF-2 genes, despite showing a significant excess of LoF mutations in the NDD cohort, harbor many LoF variants segregating in the general population. This is reflected, for instance, in their relatively high LOEUF values. The number of segregating sites in a gene is governed by both mutation rate and selection. If these genes are genuine CES drivers, we expect an elevated mutation rate to contribute to the number and frequency of LoF SNVs. In such cases, both the number of segregating sites and their allele frequencies would appear higher than expected under a fixed selection coefficient.

In a cohort with over a million chromosomes, this hypothesis can be evaluated by comparing the empirical expectation and variance of LoF allele frequency distributions across genes. While both mutation rate and selection influence the mean and variance of allele frequencies, recurrent mutation produces alleles that tend to be more common but exhibit lower variance in frequency, due to reduced genetic drift. In contrast, relaxed selection increases both mean and variance via enhanced drift. Thus, we expect genes with elevated mutation rate to show lower variance in allele frequency relative to their mean than genes under relaxed selection.

To formalize this, we applied the Gamma-Poisson approximation for allele frequency spectra under strong selection and recurrent mutation[38].

Gene Filtering. For this analysis, we retained only genes with at least 10 LoF sites passing Roulette-QUAL QC and having unscaled Roulette mutation rate $\mu > 1$ (e.g., CpG>TpG transitions). This filtering ensures reliable estimates of the mean and variance of allele frequencies for each gene.

Model. We assume a three-step generative model for LoF allele counts in each gene:

1. Mutations occur randomly at any LoF site in the gene.
2. Each derived variant segregates in a diploid population of effective size N_e, experiencing negative selection of intensity s in heterozygotes. We assume strong selection and no lineage interference.
3. A sample of n alleles is drawn from the population.

Let k_v denote the number of observed derived LoF variants v in the population sample. In what follows, we will drop the index v. We assume the variance of k to be the sum of contributions of the three assumed steps without any additional terms:

$$Var(k) = Var_{sampling}(k) + Var_{seg}(k) + Var_{\mu}(k), \quad (6)$$

where the components correspond to mutation heterogeneity across sites, drift-induced segregation, and Poisson sampling noise, respectively. Under the Gamma approximation introduced by Nei (1968), the distribution of allele frequencies at LoF sites is modeled with $4N_e\mu$ as the shape parameter and $1/(4N_e s)$ as the scale parameter. The expectation of k is then:

$$E(k) = \frac{n\,E(\mu)}{s}.$$

And the first two terms of (6) are given by the variance of the Poisson and Gamma components, respectively:

$$Var_{sampling}(k) \equiv E(k) = \frac{n\,E(\mu)}{s},$$

$$Var_{seg}(k) = \frac{n\,E(\mu)}{s}\frac{n}{4N_e s}.$$

The final term of (6) may be obtained by noticing that $Var_{sampling}(k) + Var_{seg}(k) = E_\mu[Var(k|\mu)]$. Applying the law of total variance,

$$\begin{aligned} Var(k) &= [Var_{sampling}(k) + Var_{seg}(k)] + Var_\mu(k) \\ &= E_\mu[Var(k|\mu)] + Var_\mu[E(k|\mu)] \\ &= \frac{n\,E(\mu)}{s}\left(1 + \frac{n}{4N_e s}\right) + \left(\frac{n}{s}\right)^2 Var(\mu) \\ &= \frac{n\,E(\mu)}{s} + \frac{n\,E(\mu)}{s}\frac{n}{4N_e s} + \left(\frac{n}{s}\right)^2 Var(\mu), \end{aligned}$$

$$Var_\mu(k) = \left(\frac{n}{s}\right)^2 Var(\mu).$$

We now model the effect of CES as a multiplicative factor $\kappa \geq 1$ on the mutation rate so that:

$$E(k) = \kappa \times \frac{n\,E(\mu)}{s}, \tag{7}$$

$$Var(k) = \kappa \times \frac{n\,E(\mu)}{s}\left(1 + \frac{n}{4N_e s}\right) + \kappa^2 \times \left(\frac{n}{s}\right)^2 Var(\mu). \tag{8}$$

The final term in Equation (8) containing dependency on κ^2 can be estimated from the regression of k on μ with Poisson error. Subtracting this term from $Var(k)$ leaves only the sample estimate of $E_\mu[Var(k|\mu)]$.

Estimating Selection. This framework yields two estimators of $1/s$:

1. From the mean allele count (Eq. 7):

$$\frac{\kappa}{\hat{s}} = \frac{\bar{k}}{n\bar{\mu}}, \tag{9}$$

   where $\bar{\mu}$ and $\bar{k}$ are the mean mutation rate and mean LoF allele count per gene.
2. From the conditional variance (Eq. 8), subtracting sampling noise:

$$\frac{Var_{seg}(k)}{Var_{sampling}(k)} = \frac{E_{\mu}[Var(k|\mu)] - Var_{sampling}(k)}{Var_{sampling}(k)} = \frac{n}{4N_e s}. \quad (10)$$

Note that only the first estimator (Eq. 9) depends on κ and is therefore sensitive to CES effects.

Figure 3e plots the log estimate of $\frac{n}{4N_e s}$ (excess variance due to drift) against the log estimate of $\frac{\kappa}{s}$ (elevated mean allele count due to CES). Because $log(n/4N_e) - log(s)$ is a constant offset from $log(\kappa) - log(s)$, genes with elevated mutation due to CES appear as positive residuals from the regression line.

This approach is conservative: the Gamma-Poisson approximation is invalid for genes under weak or no selection, where allele frequency variance is even higher. Thus, such genes would not mimic the CES signal.

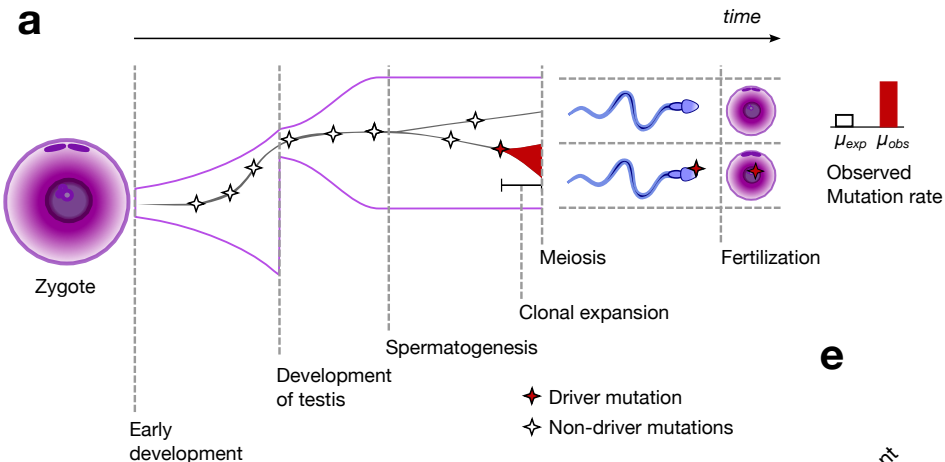
True NDD genes deviate negatively from the regression line, having lower mean allele frequencies than predicted under a fixed s. This is consistent with either sampling bias in gnomAD v4 resulting from difficulties in recruitment of individuals affected with NDD or from selection acting against LoF variants in these genes even prenatally, leading to inherently reduced representation in population datasets.

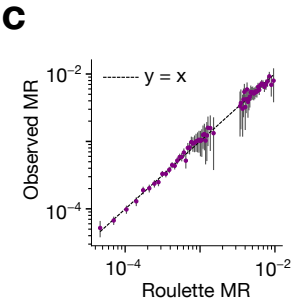
MIB1 Example. For *MIB1*, we used Poisson regression to quantify the dependence of LoF allele frequency on mutation rate. To visualize this relationship (Figure 3d), we binned sites by scaled mutation rate, merging bins with fewer than 10 sites with the next-highest bin.

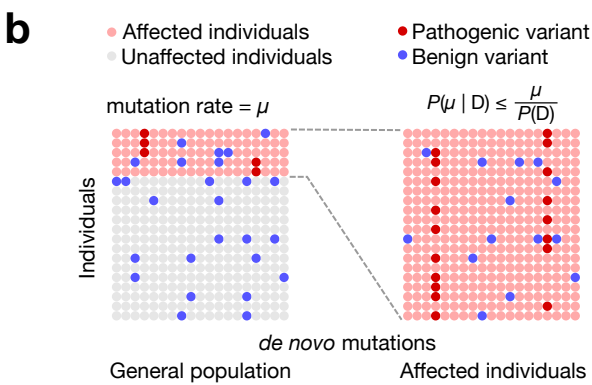
Methods References

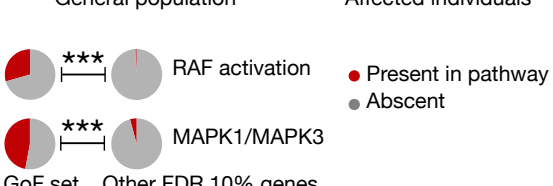
50. EuroEPINOMICS-RES Consortium, Epilepsy Phenome/Genome Project, & Epi4K Consortium. De novo mutations in synaptic transmission genes including DNM1 cause epileptic encephalopathies. *Am. J. Hum. Genet.* **95**, 360–370 (2014).
51. Helbig, K. L. *et al.* Diagnostic exome sequencing provides a molecular diagnosis for a significant proportion of patients with epilepsy. *Genet. Med.* **18**, 898–905 (2016).
52. Heyne, H. O. *et al.* De novo variants in neurodevelopmental disorders with epilepsy. *Nat. Genet.* **50**, 1048–1053 (2018).
53. Klöckner, C. *et al.* De novo variants in SNAP25 cause an early-onset developmental and epileptic encephalopathy. *Genet. Med. Off. J. Am. Coll. Med. Genet.* **23**, 653–660 (2021).
54. Allen, A. S. *et al.* De novo mutations in epileptic encephalopathies. *Nature* **501**, 217–221 (2013).

55. Perez, G. *et al.* The UCSC Genome Browser database: 2025 update. *Nucleic Acids Res.* gkae974 (2024) doi:10.1093/nar/gkae974.
56. Karlsson, M. *et al.* A single–cell type transcriptomics map of human tissues. *Sci. Adv.* **7**, eabh2169 (2021).
57. Single-cell and spatial transcriptomics during human organogenesis. *Nat. Cell Biol.* **25**, 522–523 (2023).
58. Agrawal, A. *et al.* WikiPathways 2024: next generation pathway database. *Nucleic Acids Res.* **52**, D679–D689 (2023).
59. McLaren, W. *et al.* The Ensembl Variant Effect Predictor. *Genome Biol.* **17**, 122 (2016).
60. Cheng, J. *et al.* Accurate proteome-wide missense variant effect prediction with AlphaMissense. *Science* **381**, eadg7492 (2023).



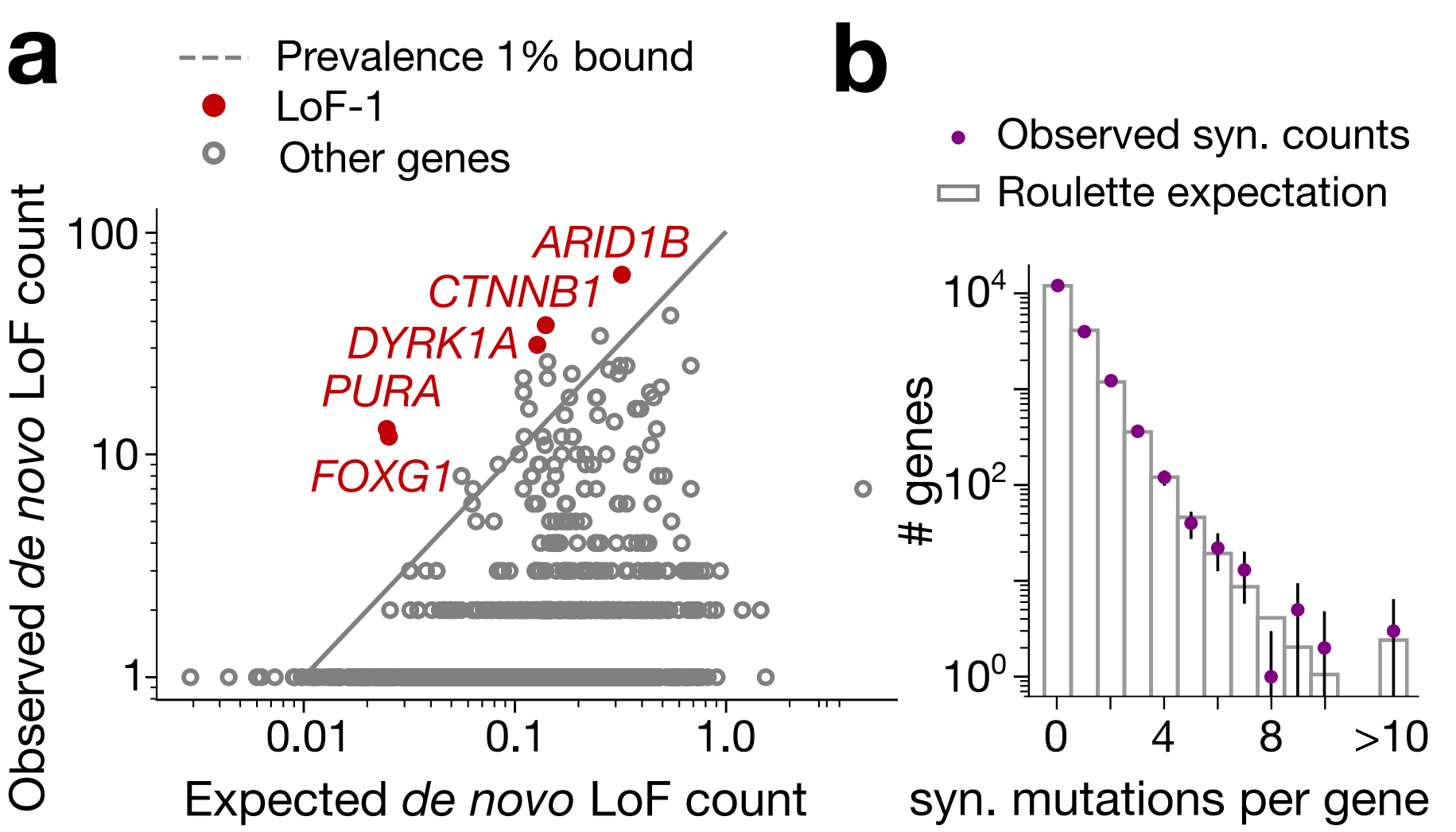






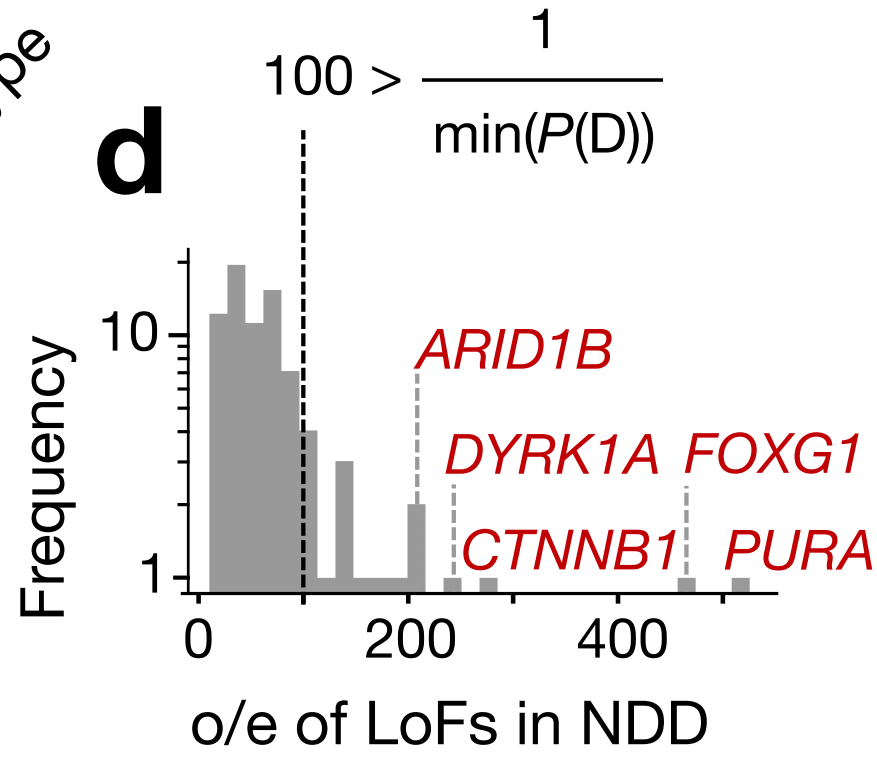
e

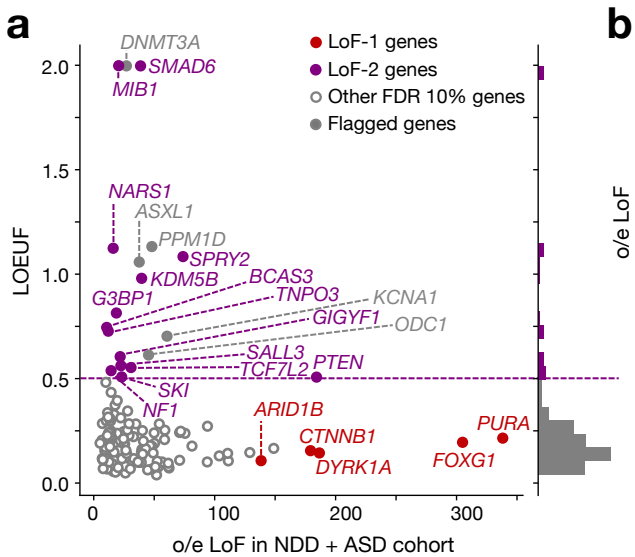
	Variant	# mutations observed	o/e ratio	Noonan [-like] variant	NDD variant	Cancer driver gene	known CES gene	FDR
PPP2R5D	E198K	22	51900		✓			$< 2.2\times10^{-16}$
SHOC2	S2G	9	42800	✓	✓			$< 2.2\times10^{-16}$
PPP1CB	P49R	11	31500	✓	✓			$< 2.2\times10^{-16}$
SMAD4	I500V	21	21700		✓	✓	✓	$< 2.2\times10^{-16}$
PTPN11	N308D	9	19600	✓	✓	✓	✓	$< 2.2\times10^{-16}$
CSNK2A1	K198R	10	17500		✓			$< 2.2\times10^{-16}$
PACS1	R203W	36	6050		✓			$< 2.2\times10^{-16}$
FGFR3	N542K	7	40200			✓	✓	2.2×10^{-8}
CDK13	N842S	10	6000		✓			6.4×10^{-8}
MAP2K1	Y130C	11	3300	✓	✓	✓	✓	9.9×10^{-7}
PACS2	E209K	13	1950		✓			3.1×10^{-6}
PPP2R5D	E420K	6	21500					1.7×10^{-5}
GRIN2B	G820A	6	19100		✓			2.7×10^{-5}
SMAD4	R496C	9	1780		✓	✓	✓	0.018
BRAF	Q257R	5	9340	✓	✓	✓	✓	0.024
PTPN11	I282V	5	8220	✓	✓	✓	✓	0.035
CERT1	S132L	5	8220		✓			0.035
GNB1	I80T	6	4380		✓			0.035
CTBP1	R342W	9	1350		✓			0.124
PSMC5	P320R	4	12700		✓			0.197
MTOR	E1799K	8	1580		✓	✓		0.197

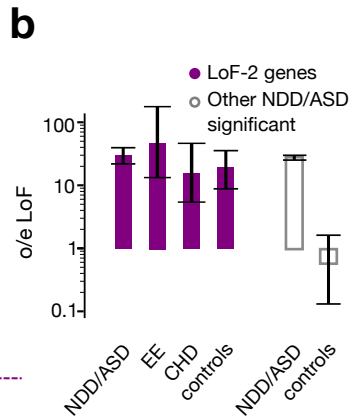


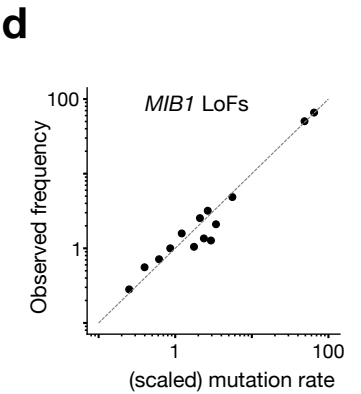
c

	# mutations observed	o/e ratio	Somatic driver gene	NDD phenotype	FDR
ARID1B	66	205	✓	✓	0.001
CTNNB1	38	270	✓	✓	0.001
PURA	14	563		✓	0.002
DYRK1A	32	249	✓	✓	0.022
FOXG1	12	471		✓	0.051



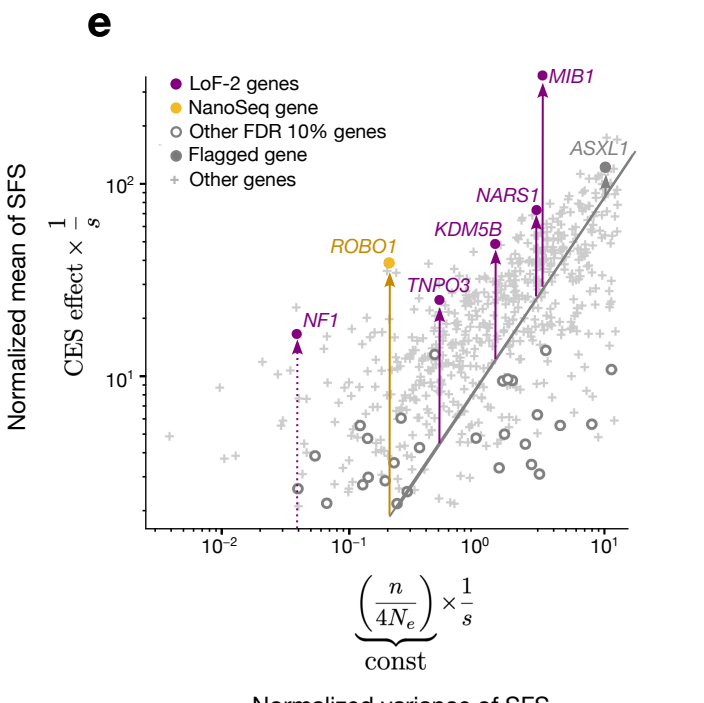


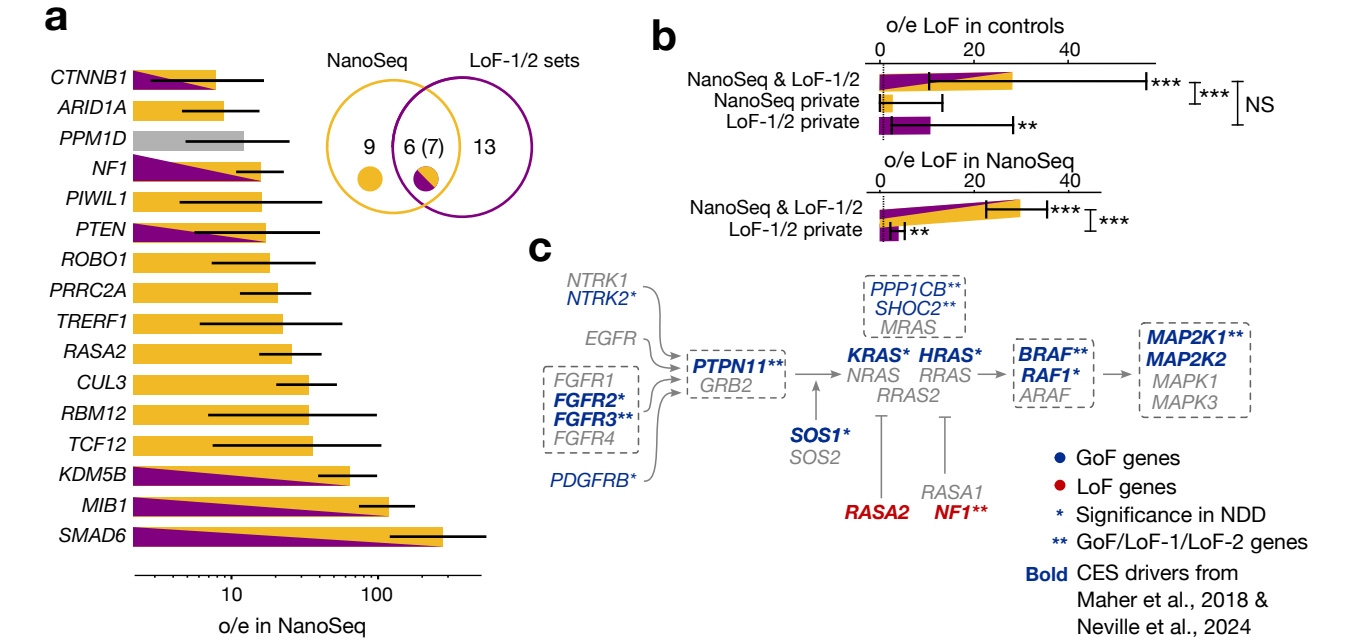


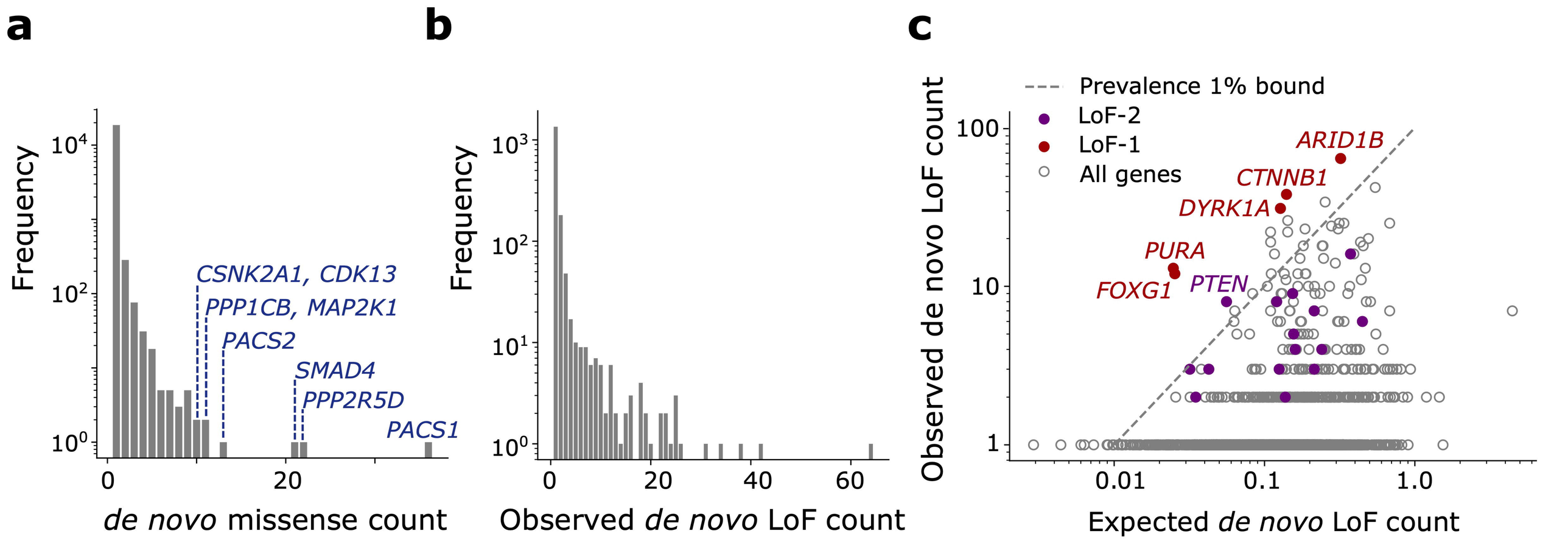


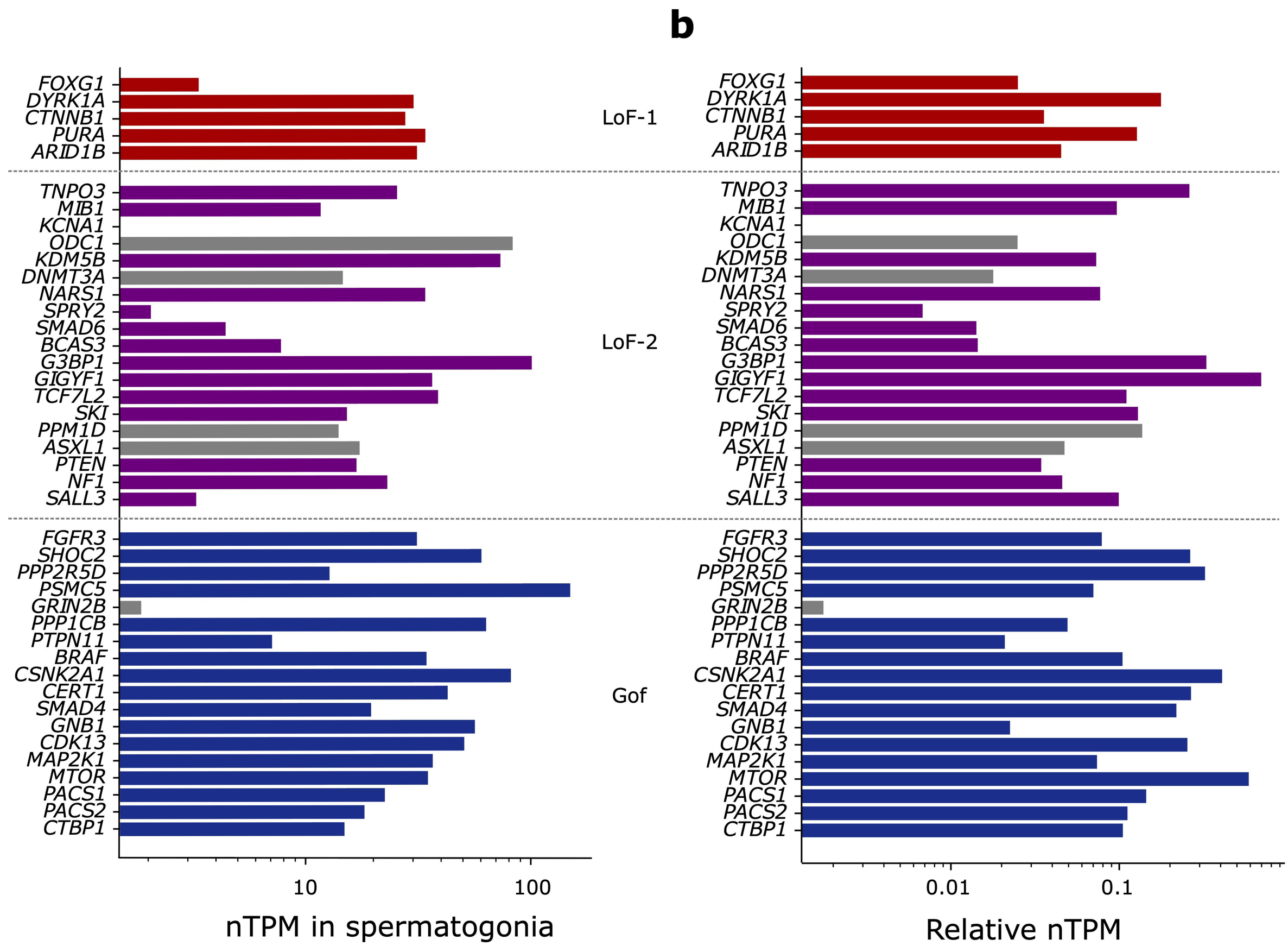
c

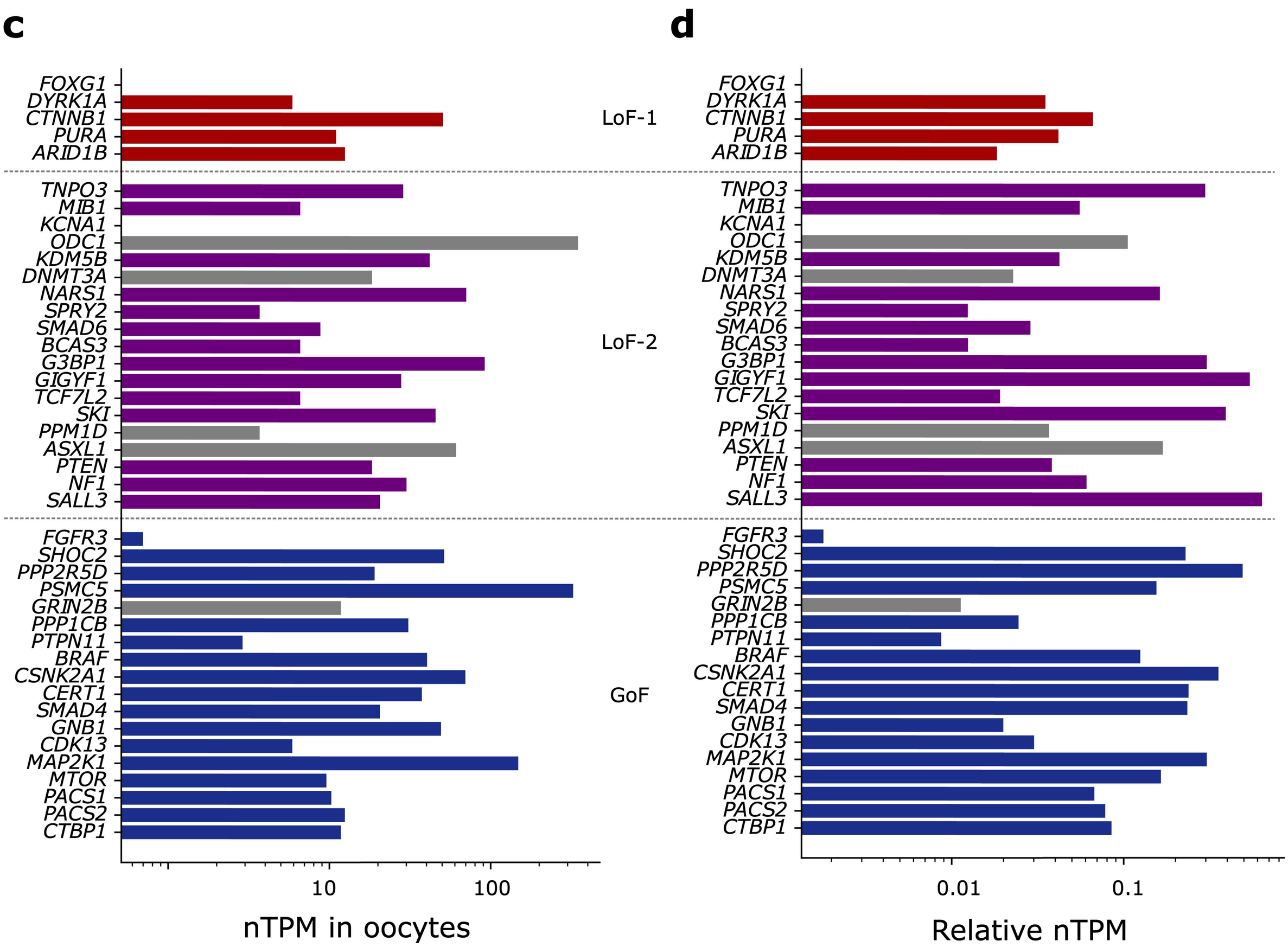
	# mutations observed	o/e ratio	LoF dominant NDD phenotype	Cancer driver	FDR
PTEN	16	184	✓	✓	$< 2.2\times10^{-16}$
KDM5B	23	39.6			$< 2.2\times10^{-16}$
MIB1	9	27			2.5×10^{-8}
NF1	10	14.4	✓	✓	1.1×10^{-6}
GIGYF1	7	21.9			1.4×10^{-5}
TCF7L2	6	30.8		✓	1.7×10^{-5}
SPRY2	4	73.8			7.7×10^{-5}
KCNA1	3	60.6	✓		0.004
SMAD6	3	38.7			0.013
G3BP1	4	18.7			0.013
NARS1	4	16.1	✓		0.022
SKI	3	23.1	✓	✓	0.05
SALL3	3	22.5			0.054
TNPO3	4	12			0.058
BCAS3	4	10.7			0.085

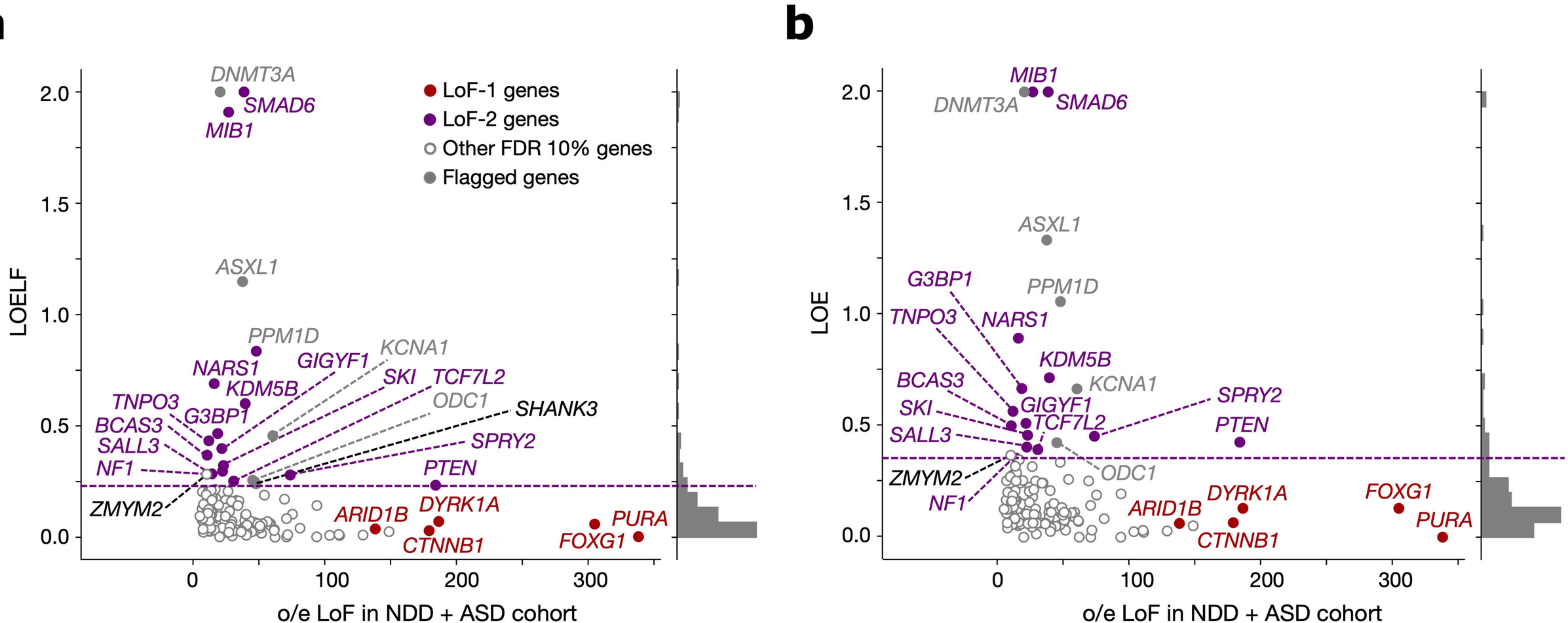


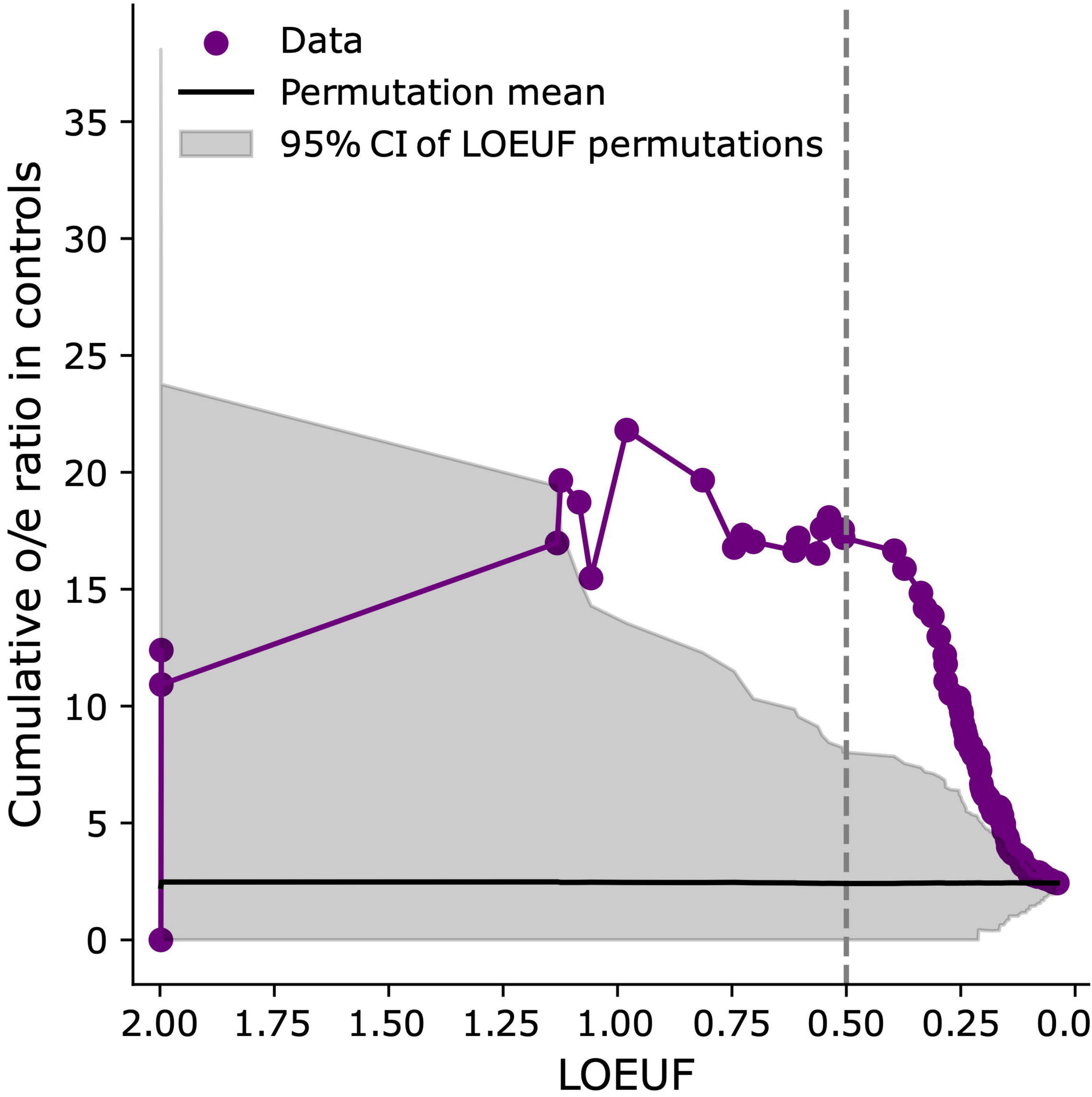

a
CTNNB1
ARID1A
PPM1D
NF1
PIWIL1
PTEN
ROBO1
PRRC2A
TRERF1
RASA2
CUL3
RBM12
TCF12
KDM5B
MIB1
SMAD6
10
100
o/e in NanoSeq
NanoSeq
LoF-1/2 sets
9
6 (7)
13
b
o/e LoF in controls
0
20
40
NanoSeq & LoF-1/2
NanoSeq private
LoF-1/2 private
***
***
NS
**
o/e LoF in NanoSeq
0
20
40
NanoSeq & LoF-1/2
LoF-1/2 private
***
***
**
c
NTRK1
NTRK2*
EGFR
FGFR1
FGFR2*
FGFR3**
FGFR4
PDGFRB*
PTPN11**
GRB2
SOS1*
SOS2
PPP1CB**
SHOC2**
MRAS
KRAS*
NRAS
HRAS*
RRAS
RRAS2
RASA2
RASA1
NF1**
BRAF**
RAF1*
ARAF
MAP2K1**
MAP2K2
MAPK1
MAPK3
GoF genes
LoF genes
* Significance in NDD
** GoF/LoF-1/LoF-2 genes
Bold CES drivers from Maher et al., 2018 & Neville et al., 2024

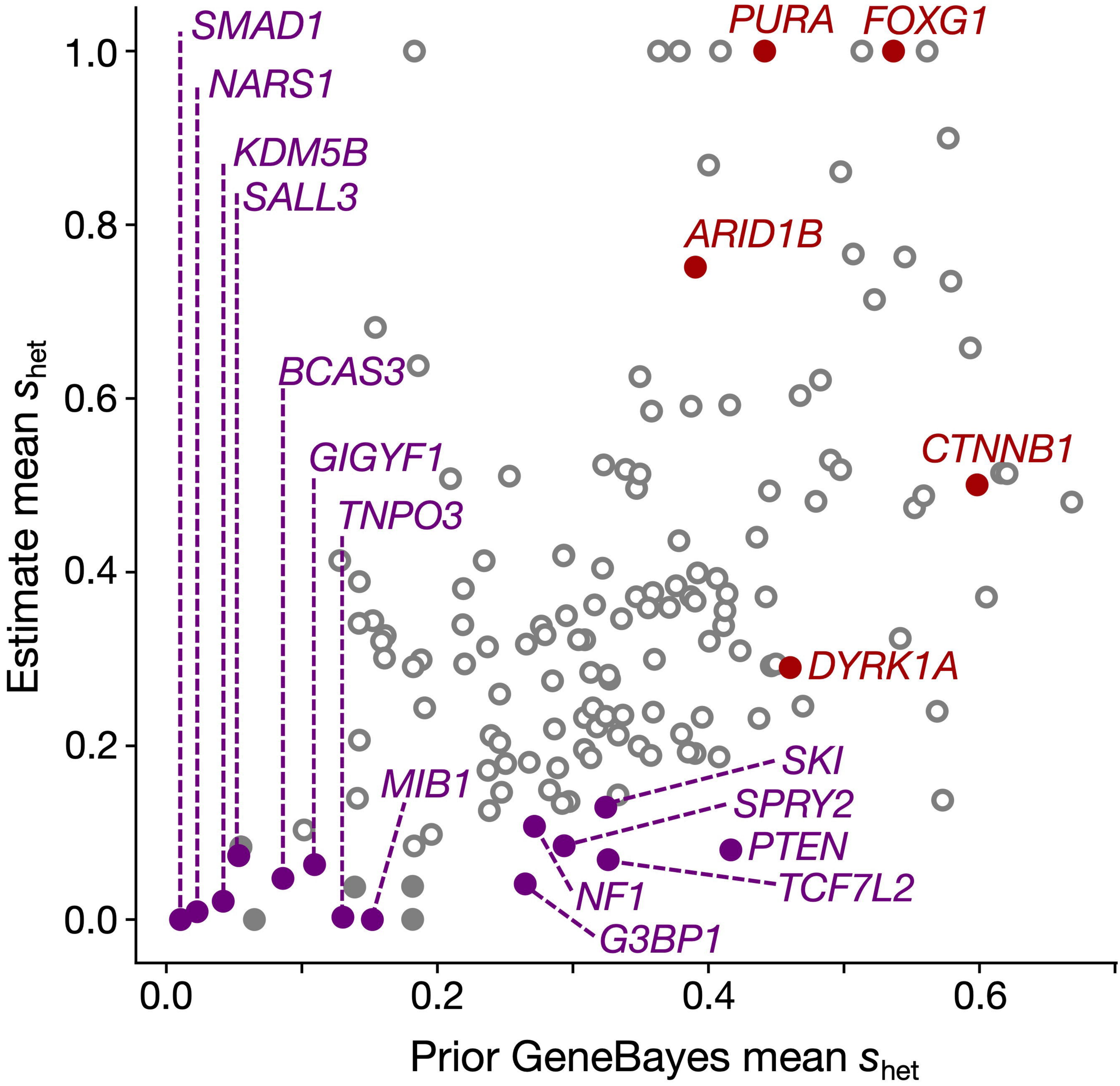

a
Frequency
10^4
10^2
10^0
CSNK2A1, CDK13
PPP1CB, MAP2K1
PACS2
SMAD4
PPP2R5D
PACS1
0
20
de novo missense count
b
Frequency
10^3
10^2
10^1
10^0
0
20
40
60
Observed de novo LoF count
c
Prevalence 1% bound
LoF-2
LoF-1
All genes
Observed de novo LoF count
100
10
1
ARID1B
CTNNB1
DYRK1A
PURA
FOXG1
PTEN
0.01
0.1
1.0
Expected de novo LoF count

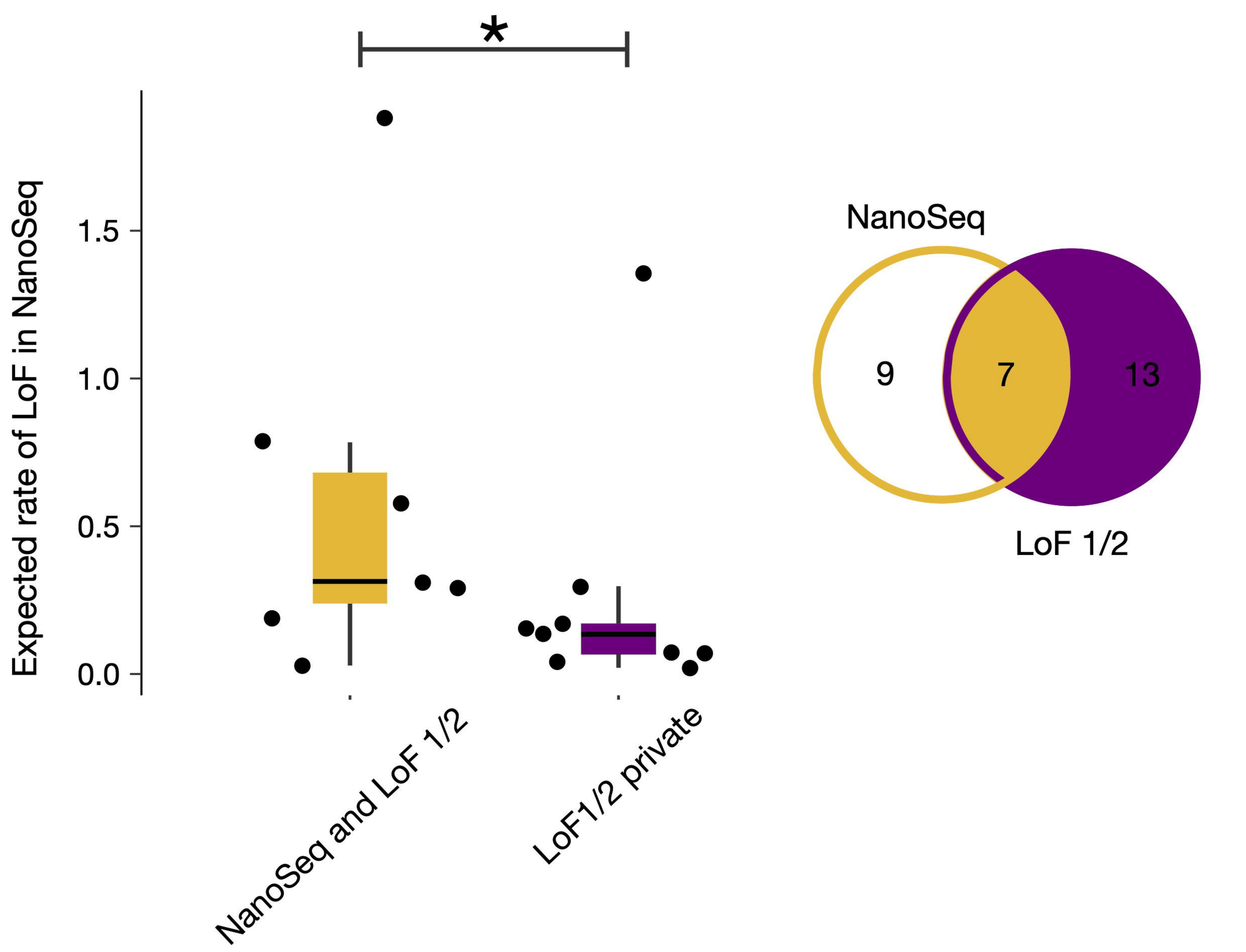

b
FOXG1
DYRK1A
CTNNB1
PURA
ARID1B
LoF-1
TNPO3
MIB1
KCNA1
ODC1
KDM5B
DNMT3A
NARS1
SPRY2
SMAD6
BCAS3
G3BP1
GIGYF1
TCF7L2
SKI
PPM1D
ASXL1
PTEN
NF1
SALL3
LoF-2
FGFR3
SHOC2
PPP2R5D
PSMC5
GRIN2B
PPP1CB
PTPN11
BRAF
CSNK2A1
CERT1
SMAD4
GNB1
CDK13
MAP2K1
MTOR
PACS1
PACS2
CTBP1
Gof
10
100
nTPM in spermatogonia
0.01
0.1
Relative nTPM

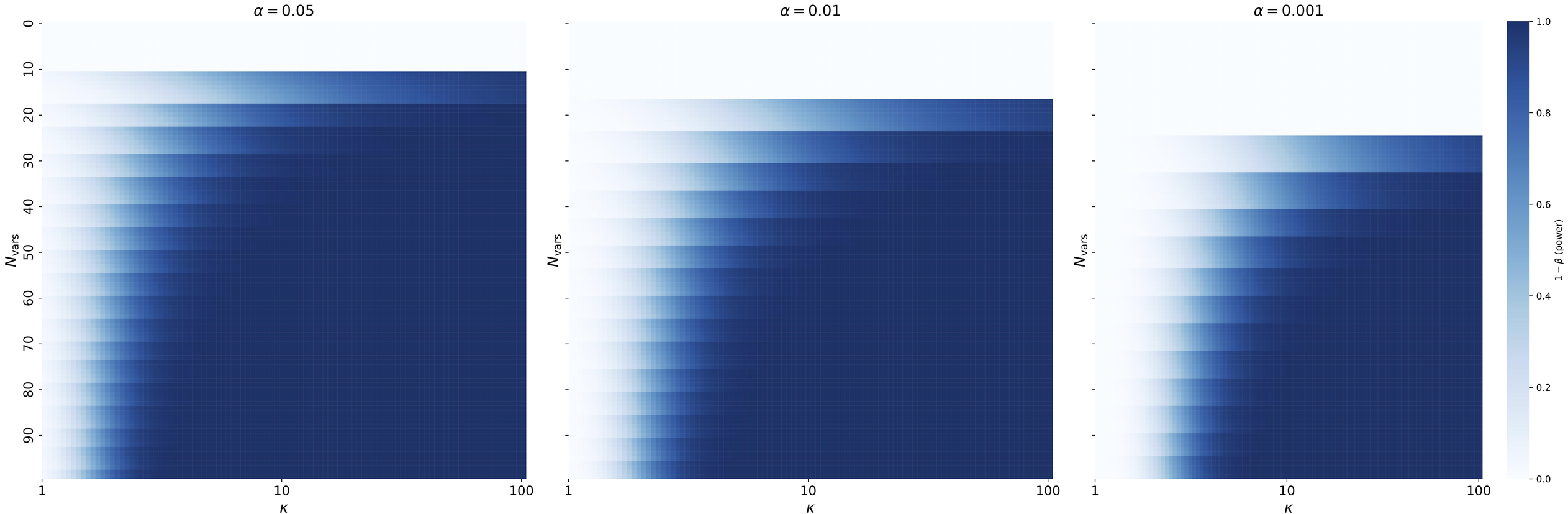

c
d
FOXG1
DYRK1A
CTNNB1
PURA
ARID1B
LoF-1
TNPO3
MIB1
KCNA1
ODC1
KDM5B
DNMT3A
NARS1
SPRY2
SMAD6
BCAS3
G3BP1
GIGYF1
TCF7L2
SKI
PPM1D
ASXL1
PTEN
NF1
SALL3
LoF-2
FGFR3
SHOC2
PPP2R5D
PSMC5
GRIN2B
PPP1CB
PTPN11
BRAF
CSNK2A1
CERT1
SMAD4
GNB1
CDK13
MAP2K1
MTOR
PACS1
PACS2
CTBP1
GoF
10
100
nTPM in oocytes
0.01
0.1
Relative nTPM


a
b
LoF-1 genes
LoF-2 genes
Other FDR 10% genes
Flagged genes
LOELF
LOE
o/e LoF in NDD + ASD cohort
o/e LoF in NDD + ASD cohort
DNMT3A
SMAD6
MIB1
ASXL1
PPM1D
KCNA1
NARS1
GIGYF1
KDM5B
SKI
TCF7L2
ODC1
SHANK3
TNPO3
G3BP1
BCAS3
SALL3
SPRY2
NF1
PTEN
ZMYM2
ARID1B
DYRK1A
CTNNB1
PURA
FOXG1
0.0
0.5
1.0
1.5
2.0
0
100
200
300


Data
Permutation mean
95% CI of LOEUF permutations
Cumulative o/e ratio in controls
35
30
25
20
15
10
5
0
2.00
1.75
1.50
1.25
1.00
0.75
0.50
0.25
0.0
LOEUF


SMAD1
NARS1
KDM5B
SALL3
BCAS3
GIGYF1
TNPO3
MIB1
PURA
FOXG1
ARID1B
CTNNB1
DYRK1A
SKI
SPRY2
PTEN
TCF7L2
NF1
G3BP1
1.0
0.8
0.6
0.4
0.2
0.0
Estimate mean s_{het}
0.0
0.2
0.4
0.6
Prior GeneBayes mean s_{het}


*
Expected rate of LoF in NanoSeq
1.5
1.0
0.5
0.0
NanoSeq and LoF 1/2
LoF1/2 private
NanoSeq
9
7
13
LoF 1/2


$\alpha = 0.05$
$\alpha = 0.01$
$\alpha = 0.001$
N_{vars}
0
10
20
30
40
50
60
70
80
90
1
10
100
κ
$1 - \beta$ (power)
0.0
0.2
0.4
0.6
0.8
1.0