Cohort-level analysis of human *de novo* mutations points to drivers of clonal expansion 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 Author: Shamil Sunyaev, ssunyaev@hms.harvard.edu

Abstract

In renewing tissues, mutations conferring selective advantage may result in clonal expansions[1–3]. 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[4–7]. CES was originally discovered through extreme recurrence of *de novo* mutations causing Apert syndrome[4]. 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[8–12], 6,065 control trios[11,13–18], and population variation from 807,162 mostly healthy individuals[19] identifies genes manifesting rates of *de novo* mutations inconsistent with plausible models of disease ascertainment. We propose 15 genes hypermutable at loss-of-function (LoF) sites as candidate CES drivers. Additional 17 genes feature hypermutable missense mutations at individual positions, suggesting CES acting through gain-of-function (GoF). Among candidates are 5 of 13 known CES drivers[6,7], 11 cancer drivers, and 16 members of major signaling pathways; notably, 12 genes show CES evidence in direct sperm sequencing[20]. CES increases the average mutation rate ~16-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 15 LoF CES driver candidates, only 5 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[13,22]. *De novo* mutation is a common cause of sporadic monogenic disease[8–11]. The interest in characterizing *de novo* mutation in disease contexts has led to multiple international efforts collecting large cohorts of parent-child trios[8–12,18]. These trios include unaffected parents and their children who, depending on the cohort, may be affected by conditions ranging from neurodevelopmental disorders to congenital heart diseases[8–11]. The simple yet effective methodology used in these studies to implicate genes in disease biology is based on comparing the observed count of *de novo* mutations in a gene with the count expected from its baseline mutation rate[8,23,24]. Due to the initial ascertainment on disease phenotype, excess of *de novo* mutations in a gene provides evidence for its causal role. The analysis is frequently assisted by the application of a selective constraint measure that reflects the depletion of population genetic variation relative to the expectation given by mutation rate[8,23]. The rationale for this approach is that genes with heterozygous functional variants posing high risk of a severe disease are expected to be under strong negative selection, resulting in their depletion at the population level.

Studies of *de novo* variation, most prominently studies of sporadic neurodevelopmental disorders (NDD)[8] and autism spectrum disorders (ASD)[11], have identified multiple gene-disease associations. These discoveries have provided valuable biological insights into disease etiology. However, these studies also produced unexpected observations. For instance, many identified genes have also been implicated in carcinogenesis. Additionally, some of these genes exhibit low selective constraint, which appears inconsistent with a significant impact of *de novo* mutations on early-onset conditions[8,11].

Is it possible that a basic premise of random mutagenesis underlying the observed statistics is occasionally fallacious? Over two decades ago, hypermutable sites in *FGFR3* and in *FGFR2* causal for achondroplasia and Apert syndrome, respectively, were explained by positive selection for the variants in spermatogonia due to gain-of-function (GoF) mutations[4,5]. The resulting *de novo* rate of these driver mutations dramatically exceeds the rate of baseline mutagenesis. Subsequent studies identified additional driver mutations of clonal expansions in spermatogonia (CES) in 13 genes[6,7].

Recent research demonstrated that clonal expansions in healthy tissues, commonly caused by mutations in known cancer driver genes, are ubiquitous and widely distributed across organs[1,2]. For example, 23% of 40–70-year-old individuals in the UK Biobank have large clonal expansions in blood[3]. In gametogenic tissues, specifically spermatogonia, such clonal expansions can uniquely increase the likelihood of transmitting driver variants, elevating the probability of *de novo* mutations being passed onto offspring[6,7] (Fig. 1a). If clonal expansions in spermatogonia are common, similar to somatic tissues, CES drivers would be expected to have rates of *de novo* mutations far exceeding the baseline mutation rate[4–7]. In disease cohorts, this

could lead to an increased prevalence of disease caused by mutations in genes that drive both CES and disease[4,5]. Conversely genes that drive CES without directly causing disease, may lead to false-positive findings in disease-focused studies. They could appear to be disease-causing due to the elevated number of *de novo* mutations observed in probands.

To investigate the possible effect of CES on the human germline mutation and identify potential CES drivers, we developed a statistical approach leveraging human *de novo* mutation data[8–18] along with population genetic variation resources[19]. Building on the observation that CES drivers may act with or without causing disease, our method is designed to capture both categories of CES drivers. By applying this approach to a joint dataset of 54,715 rare disease trios[8–11] and gnomAD-v4 dataset[19] of 807,162 individuals, we uncovered 15 CES driver gene candidates acting through loss-of-function (LoF) mechanism and 20 candidate driver mutations across other 17 genes acting through GoF. Most of these genes follow biological expectation for drivers of clonal expansions such as involvement in MAPK[11] and other major signalling pathways. We validated our findings using the already discovered CES genes[6,7], cohorts of unaffected trios[11,13–18] and comparison with the results of the direct sperm sequencing[20]. We observe an average 16-fold increase in mutation rate for LoF CES genes and a 524-fold increase in mutation rate at GoF positions.

Results

Detecting excess of *de novo* mutation to identify clonal expansions in spermatogonia

Rates of DNA damage, together with imperfections of DNA repair and replication determine the baseline *de novo* mutation rate in the male germline[25]. Selective advantage conveyed by CES in spermatogonia should elevate observed mutation rates of the drivers relative to the baseline. For mutations of identical selective advantages, CES will linearly amplify the baseline mutation rate: relative mutation rates between sites will still be determined by the baseline mutation rate. The approaches to identify CES genes taken here will generally compare the observed mutation counts to the ones expected under a model of baseline mutation rate.

Under the assumption of the accurate mutation rate predictions, observed counts of *de novo* mutations are Poisson distributed. To estimate the baseline mutation rate at the single base-pair resolution, we rely on the Roulette model[26]. We first evaluated the accuracy of Roulette for predicting mutation rates in the trio datasets as uncertainty of Roulette estimates may yield false positive findings and power loss. At the level of individual sites, inaccuracy of mutation rate predictions is absorbed by a Negative Binomial fit. The fit is informed by *de novo* synonymous mutations in the largest trio dataset (31,058 probands)[8]. We show that Roulette estimates are unbiased across more than two orders of magnitude of mutation rate (Fig. 1d, Suppl. Tab. S1) and the variance in synonymous mutation rate unaccounted for by Roulette is 0.4% (see Methods).

For sites aggregated by individual genes, we elected a conservative approach by censoring gene outliers according to synonymous variation in gnomAD-v4[19]. This procedure flagged 2.5% of autosomal genes (439) with the observed deviation from the synonymous

expectation at Z-score > 3; the threshold corresponding to ~20% increase in mutation rate in those genes (Fig. 1c, Extended Data Figure 1, Suppl. Tab. S2). Further, we flagged known drivers of clonal hematopoiesis[3] (Suppl. Tab. S3) as most of the sequencing is performed on DNA samples from blood and such drivers could be misclassified as germline variants[3,19]. Last, we flagged genes where only NMD-escaping protein-truncating variants cause NDD[27,28], as they violate our assumption of the same functional effect of all predicted LoFs. We clearly mark the flagged genes in all figures, and they were not included in any of the aggregated statistics presented further.

CES explains hypermutability of LoFs in a disease cohort

The discrepancy between baseline and observed mutation rates indicative of CES would likely be best investigated in cohorts of trios randomly sampled from the population. However, such datasets are not currently large enough for CES gene discovery, even in aggregate[11,13–18] (Supl. Tab. S4). Instead, the largest cohort published to date is that of NDD cases[8], where the CES signal is obscured by ascertainment on phenotype. Thus, the excess of *de novo* mutations over baseline expectation will also be observed for genes causing monogenic forms of NDD. For any disease-causing gene *G* not driving CES, there exists a limit to the relative excess of mutations due to ascertainment given by complete penetrance:

$$P(G \mid D) = \frac{P(D \mid G)\mu}{P(D)} \leq \frac{\mu}{P(D)}, \quad (1)$$

where μ is the probability of a functional *de novo* mutation occurring in *G* and *P(D)* is the overall disease prevalence (Fig. 1b). CES drivers could theoretically violate this upper bound. This procedure is sensitive to the estimates of *P(D)* and the uniformity of phenotypic sampling of the sequencing cohort from the disease pool.

The uniformity of phenotypic sampling underlies the key assumption of the same value of *P(D)* across genes implicitly made in eq. (1). For monogenic diseases driven by *de novo* mutations, non-uniform ascertainment by phenotype would manifest as variable genetic contributions of disease-causing genes across cohorts. We compared the contributions of individual genes to *de novo* variation in three cohorts comprising the aggregated NDD dataset used here (see Methods). Our analysis supported the assumptions of uniformity for the three cohorts assembled in different countries.

Potential overestimation of prevalence renders our approach non-conservative. Based on epidemiological evidence and the genetic observations in the NDD cohort suggesting that the prevalence of individual conditions in such cohorts is larger than 1%. We consider *P(D)*=1% as a conservative estimate[29] (see Methods).

We applied Poisson test with the expectation given by μ*/P(D)*, to the counts of *de novo* LoF variants aggregated by gene (see Methods, Suppl. Tab. S5). We identified two genes significant at the level of $p < 3 \times 10^{-6}$ (0.05 after Bonferroni for 17731 genes): *ARID1B*, *CTNNB1*. Two more genes: *DYRK1A and FOXG1* were significant at FDR 10%. *PURA* also passed the Bonferroni significance threshold, but has been flagged as a potential Roulette outlier due to an excess of synonymous variants in the population (Fig. 1c). We call these four genes “set 1” of putative CES genes (Fig. 1e). Regarding the source of the observed effect, two possibilities exist. First, mutations in these genes do not cause disease, and the increase in the effective mutation rate above baseline is driven purely by CES. In contrast, the second possibility is that

LoF mutations in these genes drive both CES and disease phenotype. In this scenario, the elevated mutation rate is influenced by both the expansion in CES and ascertainment bias, as these mutations are more likely to be observed in disease cohorts.

It is possible to distinguish between these two hypotheses using the observed levels of LoF polymorphism in the general population. If CES drivers of set 1 do not cause NDD, the mutation rate should be increased by more than two orders of magnitude to explain the excess of LoFs in the NDD cohort. This would make at least 80% of LoF sites in gnomAD-v4 having at least 1 variant representing a *de novo* mutation in the cohort (10^{-8} average mutation rate × 1.6×10^{6} haplotypes × 10^{2} CES effect ≈ mean number of *de novo* per site 1.6)[13,19,22]. 80% represents the lower bound corresponding to extreme negative selection effectively prohibiting segregation. In contrast, the observed rate of LoF variation in gnomAD at any frequency (not just *de novo*) in these genes is in the range of 1.2-2.4%. This suggests that set 1 genes both cause NDD and drive CES. Moreover, the association with NDD for all four genes and *PURA* is backed by independent clinical evidence[21] (Suppl. Tab S6), and genes involvement in CES is indirectly supported by involvement for some of these genes in clonal expansions in other tissues[30,31].

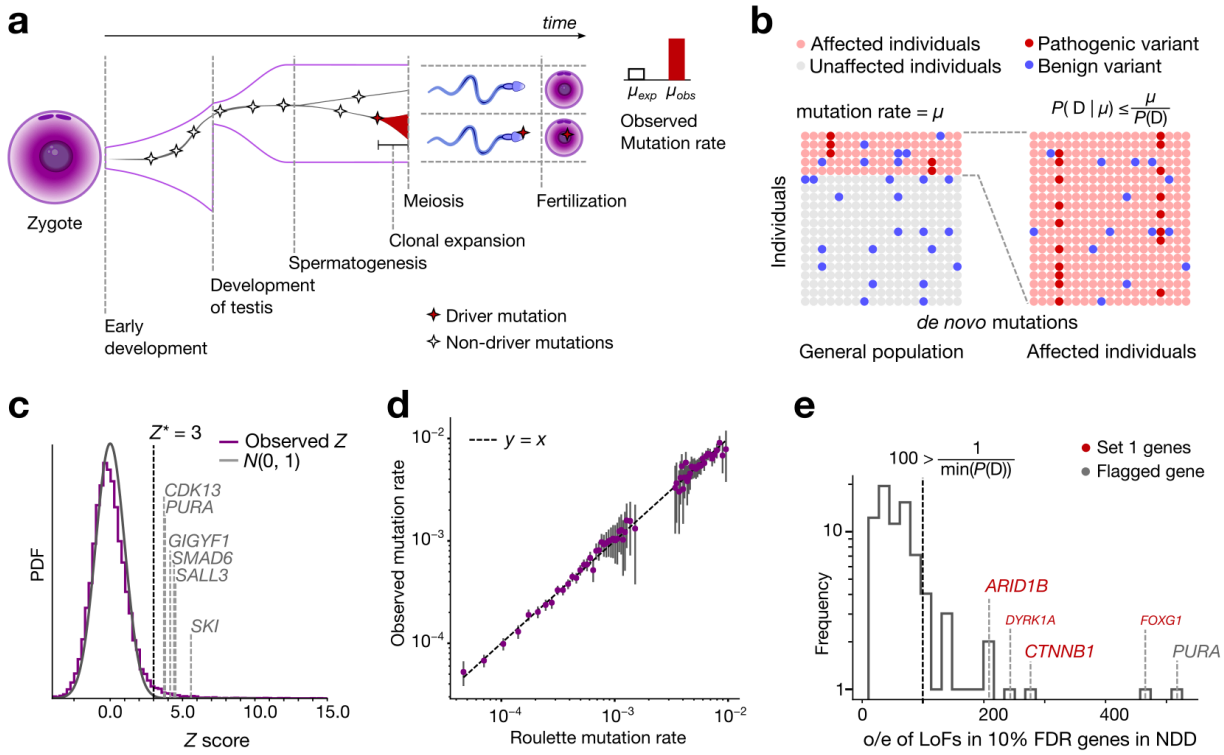


Figure 1 | Clonal expansions in spermatogonia and disease ascertainment. (a) Effect of clonal expansions in sperm on the observed mutation rate **(b)** In a randomly sampled cohort of individuals affected by a specific disease, the frequency of any variant is elevated due to ascertainment by at most the inverse of disease prevalence. *P(D)* is disease prevalence and μ is mutation rate **(c)** Per-gene *Z*-scores of comparison between the observed synonymous counts in gnomAD-v4[19] and expectations given by Roulette. Threshold *Z*=3 is shown as a dashed black line. *Z*-scores of genes with *Z* > 3 significant in any of our downstream analyses are shown as grey dashed lines. **(d)** Observed synonymous *de novo* variant counts in NDD

cohort stratified by Roulette mutation rate bins vs. Roulette predictions. 95% Poisson confidence intervals are shown. **(e)** 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[8] by Kaplanis et al., 2020. The outliers with the ratio above the maximal ascertainment (set 1 genes) are highlighted in red. Gene with biased baseline mutation rate estimates is shown in grey.

Highly polymorphic CES drivers

Genes of set 1 simultaneously drive CES and cause NDD, which explains the large excess of *de novo* LoF mutations in the NDD cohort. For CES drivers with no effect on disease, the test that produced set 1 is severely underpowered, as it requires the increase of mutation rate above two orders of magnitude. To identify such genes, we can rely on additional considerations. In particular, CES causes the increase of *de novo* mutation rate in driver genes in any cohort irrespective of phenotypic ascertainment. Thus, we merged the NDD dataset with the largest trio dataset[8] of Autism Spectrum Disorders (ASD)[11]. Furthermore, conditionally on the fitness effect, the increase in mutation rate is naturally expected to elevate the level of LoF polymorphism in the general population. This motivates a joint analysis of large trio sequencing datasets and population variation resources.

We first identified genes with significant excess of LoF *de novo* mutations in the joint ASD/NDD cohort. Then, we analyzed segregating LoF SNVs in these genes in gnomAD-v4. To quantify LoF polymorphism, we used LOEUF[19], an upper bound estimate of the observed-to-expected ratio of LoF SNVs. By construction, high values of LOEUF indicate relaxed negative selection or underestimation of mutation rate. Genes under relaxed negative selection and not causing CES may have high LOEUF values, but are not expected to show the excess of *de novo* mutations in the ASD/NDD trio cohort. Genes causing ASD/NDD but not CES may have elevated mutation counts in trio cohorts, but are expected to ultimately have low LOEUF values. In contrast, CES drivers are expected to have both the excess of *de novo* mutations in the ASD/NDD cohort and could have high LOEUF values.

122 genes show a significant excess of LoF *de novo* mutations in the ASD/NDD cohort at FDR < 0.1 (see Methods, Suppl. Tab S7). From these we selected 19 genes with LOEUF > 0.5 (concordant with cutoffs usually chosen for relaxed negative selection[11,32–34]). Eight out of the 19 genes were flagged as possible Roulette outliers, genes involved in clonal expansions in blood[3], or genes where only NMD-escaping protein-truncating variants cause NDD[27,28]. The remaining 11 genes we call set 2 of putative CES genes (Fig 2a). To further validate our procedure we used GeneBayes[35], which is an orthogonal approach that estimates gene constraint using biological features (see Methods). These estimates were then compared with a population-based estimate of constraint, which would be deflated if the effective mutation rate exceeds the baseline. Indeed, set 2 genes appear to be markedly more constrained according to biological features than according to population data, implying inflated mutation rate (Extended Data Figure 2, Suppl. Tab. S8).

With the help of population variation data, we can find additional evidence that the high level of LoF polymorphism (e.g. as indicated by LOEUF) in set 2 genes likely reflects elevated *de novo* mutation rate rather than relaxed selection. Large sample size of gnomAD-v4 (1.6M chromosomes) enables analysis of frequencies of LoF variants generated by recurrent

mutations[26,36–38]. Impact of mutation rate on polymorphism is clearly visible from the proportional increase of LoF allele frequencies with baseline mutation rate. Figure 2b shows this effect for *MIB1*, where mutation recurrence determines allele frequencies. Note that relative LoF mutation rates within the same gene are proportional to the baseline rate irrespective of CES.

Negative selection does not only decrease the average allele frequencies in the population, but also reduces variance among frequencies of individual alleles. Variants with the same mutation rate and selection coefficient have different allele frequencies due to the effect of genetic drift. Stronger 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[39] (see Methods). 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. As we expect a linear relationship between the mean allele frequency and variance inflation due to drift, CES genes will deviate from this relationship by exactly the factor increasing mutation rate. Figure 2c shows this relationship for 952 genes with a sufficient number of highly mutable LoF sites (ten or more CpG LoF sites). Set 2 genes satisfying this criterion systematically deviate in the direction consistent with the CES effect.

Set 2 genes have a significant excess of *de novo* mutations not just in the joint ASD/NDD dataset, but also in the joint healthy trio cohort[11,13–18] and cohorts of encephalopathic epilepsy trios[9] and congenital heart disease trios[10,12] (Suppl. Tabs S4,S9,S10). Moreover, the 20-fold excess of LoF mutations in these genes is strikingly similar between healthy controls and all other cohorts (Fig. 2d). This suggests that the set 2 genes could include false positives for disease association. Indeed, the existing literature strongly supports only one gene (*PTEN*) as associated with either ASD or NDD. Compared to other set 2 genes, *PTEN* has much stronger enrichment of LoF *de novo* mutations in the ASD/NDD cohort. In contrast, seven of eleven genes show inconsistent or no evidence, with the remaining three genes showing relatively weak evidence of involvement in ASD/NDD (Supplementary Table S6). Their enrichment in ASD/NDD could therefore be driven primarily by CES. Using the cohort of healthy controls, we additionally estimated that set 2 genes explain 3% of *de novo* LoF variation. This means that 11 genes contribute the same amount of *de novo* variation as the average of 554 genes. In addition, all of the set 1 and set 2 genes, barring *KCNA1*, are expressed in spermatogonia according to single-cell RNAseq data[40] (Fig. 2e). Noticeably, the discovery of set 2 genes is robust to the LOEUF threshold as judged by the excess of LoF mutations in the control cohort (Extended Data Figure 3).

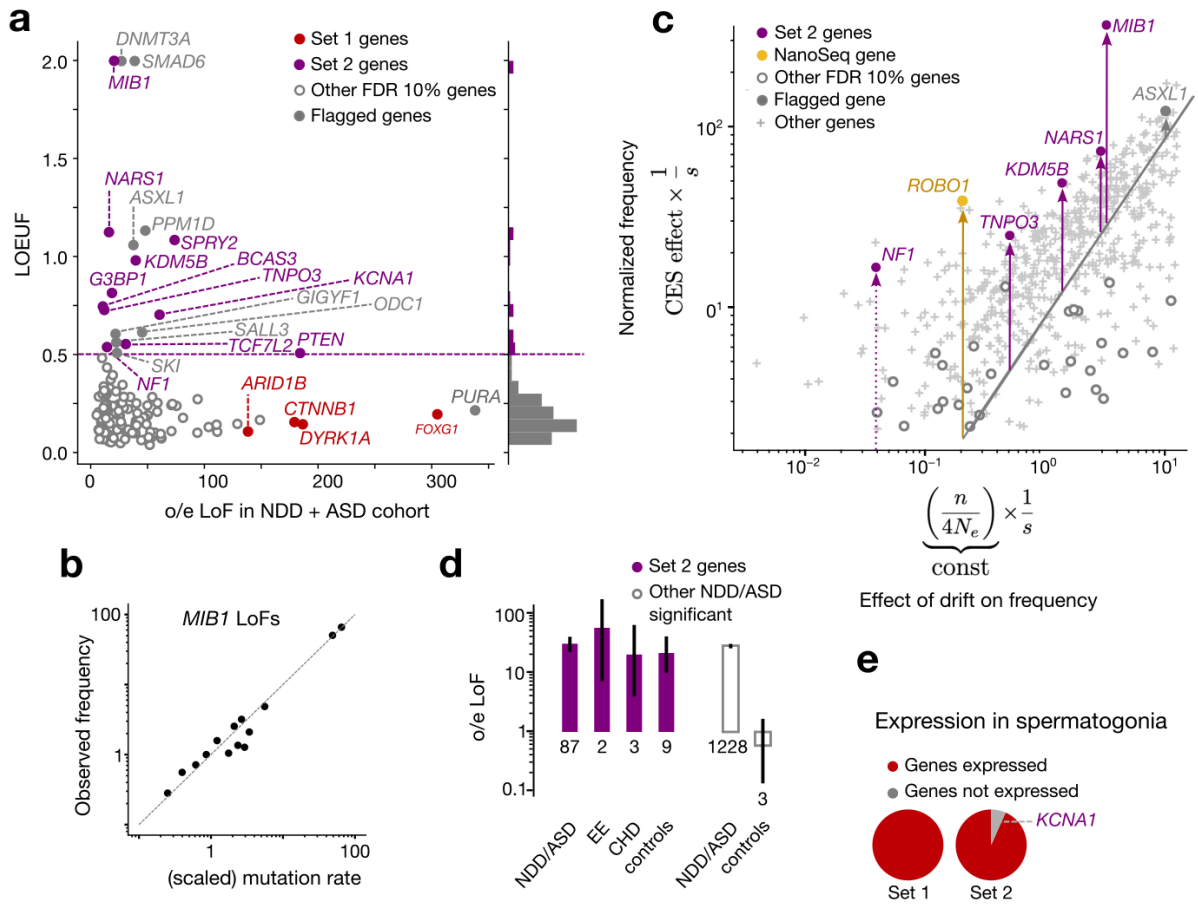


Figure 2 | Effect of CES on *de novo* mutations in disease and on the levels of LoF polymorphism in population. (a) Observed-to-expected variant count ratio (o/e) for *de novo* LoFs in genes with FDR < 0.1 in neurodevelopmental disorder (NDD)[8] merged with autism spectrum disorder (ASD)[11] against LOUEF scores. Dashed violet line indicates the 0.5 threshold for LOEUF used to construct set 2 and the histogram to the right shows LOEUF values. **(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 **(c)** 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. **(d)** Observed-to-expected LoF variant count ratio in four different trio cohorts. Violet bars indicate set 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. **(e)** Fraction of set 2 genes expressed in spermatogonia.

Gain-of-function CES drivers

Many CES genes discovered to date act through a gain-of-function (GoF) mechanism[4–7]. This motivated us to investigate *de novo* missense variation in the NDD cohort alongside LoF

variation. Similar to LoFs, missense variants causing CES through GoF should manifest as the increase of effective mutation rate relative to the baseline expectation in individual positions. This prompted us to employ an approach similar to that as in the case of set 1, which is to test for the excess of variants for each individual site above the upper bound given by disease prevalence (as described above).

Even though the number of effective tests is equal to all possible autosomal missense mutations (5×10^7), this procedure yielded 21 variants significant at 20% FDR landing in 18 unique genes (Suppl. Tab. S11). One of the 18 genes was flagged before as deviating from Roulette mutation rate predictions, leaving the total of 17 genes which we call set 3 (Fig. 3a). Five set 3 genes have been previously experimentally established to cause CES[6,7]. In addition, all set 3 genes, except for *GRIN2B* and *PP2R5D*, are expressed in spermatogonia[40] (Fig 3b). In agreement with the previous studies of CES acting through GoF mechanism[6,7] set 3 genes are enriched in RAF activation (adjusted $p=1.7\times10^{-7}$) and MAPK1/MAPK3 (adjusted $p=2.2\times10^{-7}$) pathways[41] (Suppl. Tab. S12, see Methods).

Somatic clonal expansions have been linked to genes driving oncogenic transformation of cells. To investigate this connection, we intersected set 3 genes with the curated COSMIC tier-1 list of genes for which ample evidence of oncogenic activity exists[30] and found that six genes in set 3 are COSMIC tier-1 genes (Fig. 3a). Because set 3 genes were discovered purely through an excess of *de novo* mutations, agnostic to any mechanistic hypothesis, they provide an opportunity to speculate about the biology underlying CES. First, we compared positions of putative GoF mutations within the six tier 1 cancer genes in our list with mutations in the same gene in COSMIC[30]. In the context of cancer, these genes also work predominantly through GoF mechanisms. However, the specific GoF mutations underlying CES are barely observed in COSMIC, except for *MTOR* (Fig. 3d). Remarkably, the most recurrent *de novo* mutations in NDD cohort yielding N540K variant in *FGFR3* and Q257R in *BRAF* do not occur in the COSMIC database at all (Fig. 3d).

Set 3 genes were discovered in the NDD cohort with a procedure analogous to the one used to construct set 1. This means that both CES and NDD ascertainment effects likely contribute to the signal. Thus, the differences between the mutations observed here and in COSMIC might be due to different effects of variants within a single gene on cancer and NDD. Indeed, set 3 is enriched in genes causal for NDD, e.g. associated with Noonan syndrome (Suppl. Tab. S6).

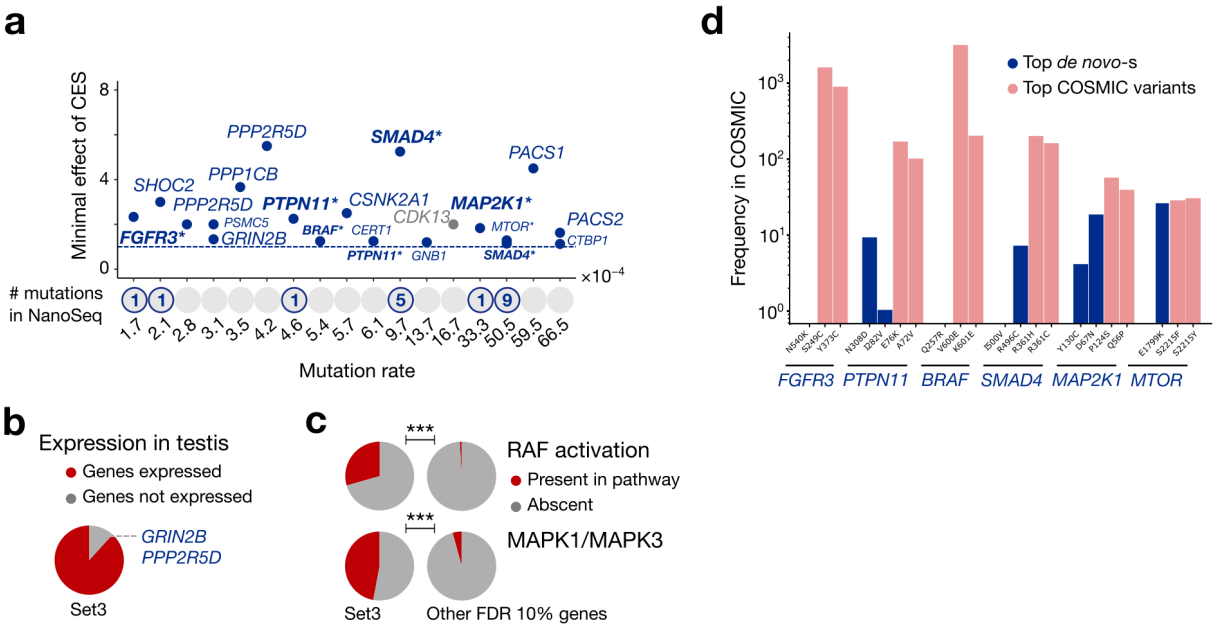


Figure 3 | Missense GoF mutations driving CES. (a) Minimal effects of CES estimated for individual significant 20 missense variants across 17 genes (See Methods). Genes identified in refs.[6,7] are shown in bold font. COSMIC tier-1 cancer driver genes are marked with asterisks (*). Mutations are stratified by relative Roulette mutation rate (x axis). Counts of mutations observed in the dataset of *Neville et al.*[20] in respective positions are shown as numbers. **(b)** Fraction of genes from set 3 expressed in spermatogonia[40]. **(c)** 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 (***). **(d)** Frequencies in COSMIC[30] of putative CES variants (blue bars) in the NDD cohort vs. top cancer driver variants (red bars).

Comparison with direct NanoSeq 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[20]. 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[20]. In terms presented here, direct sequencing of sperm is effectively equivalent to assaying paternal *de novo* mutations in a large cohort of randomly selected trios in all genes barring those which are not compatible with live births. We used these data here to verify some of our hypotheses and make quantitative statements about a number of parameters of interest.

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[20]. Out of 32 genes included in sets 1-3 reported here, 12 are significant in Neville et al. This includes 7 out of 15 LoF genes from sets 1 and 2.

Next, we compared genes of sets 1 and 2 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 set 1 or 2. Out of those 7, 2 genes have been flagged. One additional NanoSeq-significant gene (*ROBO1*) has features suggesting CES effect in the analysis of LoF polymorphism in gnomAD-v4 (Fig. 2b), although it did not have enough variants in the joint ASD/NDD cohort to be included in set 2.

When aggregated, the union of set 1 and set 2 genes demonstrates a 15.6-fold (Poisson 95% CI 13.5-19.1) excess of LoF mutations in the NanoSeq dataset. Notably, set 1 and set 2 genes not achieving significance in NanoSeq, partially due to lower power (Extended Data Figure 4), feature a significant 3-fold LoF enrichment in aggregate (Poisson P = 0.0098). The excess of LoF mutations in the control trio cohort for LoF-significant genes overlapping between NanoSeq and the union of set 1 and set 2 was estimated to be 23.1 (Poisson 95% CI 8.5-50.2). In the rest of genes from set 1 and 2, the cumulative effect in controls is 10.0 (Poisson 95% CI 2.1-29.3). Finally, we observe a remarkable 524-fold increase in mutation rate in NanoSeq data[20] for aggregated missense mutations in set 3 (Poisson 95% CI 311-828), which is almost 10-fold more than enrichment observed for all sites mutated 5 or more times in NDD[20].

Biological and clinical significance of putative 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. However, the functional roles of CES driver candidates are not representative of the broader set of known ASD/NDD genes.

Out of 32 identified genes, 16 play a role in major signalling pathways. 8 of these 16 are nominally significant in Neville et al. (additionally, SPRY2 has a nominal p-value of 0.07). The pathways include MAPK, WNT, TGF-β and RTK (Suppl. Tab. S13). As expected, the MAPK pathway has the highest enrichment (8/32 genes). 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. 4d).

Three genes are involved in chromatin regulation (*ARID1B*, *KDM5B* and *CTBP1*). *KDM5B* is highly significant in the direct sperm sequencing experiment, and so is *ARID1A*, a member of the chromatin remodelling complex SWI/SNF highly related to *ARID1B*. Set 2 genes *TNPO3* and *NARS1* do not have obvious biological functions relevant to clonal expansion, although both are nominally significant in Neville *et al*.[20].

Serine-threonine kinase *CSNK2A1* and genes activated by it (*PACS1* and *PACS2*)[42] form an intriguing group of functionally related CES driver candidates within set 3. Another related kinase, *CSNK2B*[42], has been reported by Neville *et al*.. R203T mutation in *PACS1* is the most recurrent variant in the NDD cohort 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[43,44] and CES.

The NanoSeq dataset allows us to verify our assumptions about set 1 and set 2 genes regarding their clinical significance. This can be only done for 5 genes overlapping between our sets 1 and 2 and the 16 NanoSeq LoF significant genes (Fig. 4a). We test whether the level of

enrichment of LoF *de novo* mutations in the NDD cohort can be fully explained by CES (as measured in NanoSeq) without additional disease ascertainment (see Methods). As expected for any gene of set 1, *CTNNB1* is significant for the excess effect of disease ascertainment and the excess of LoF mutations in the NDD cohort cannot be fully explained by CES. In addition, a set 2 gene, *PTEN*, shows significant ascertainment on NDD as well. On the contrary, for *MIB1*, *NF1* and *KDM5B* the significant effect of disease ascertainment could not be established. Concordantly, OMIM database includes clear evidence of *PTEN* and *CTNNB1* as dominantly acting NDD genes and markedly less evidence for the other three (Suppl. Tab. S6).

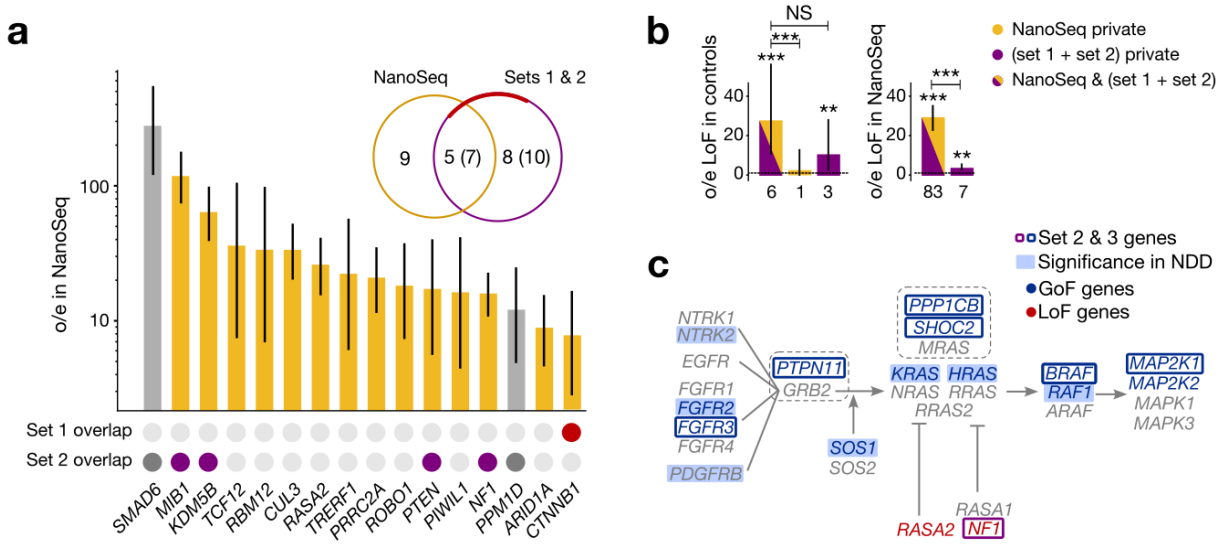


Figure 4 | Comparison with direct sequencing data (a) Observed-to-expected ratio for LoFs counts in NanoSeq sequencing of sperm for genes with significant excess of LoFs. Poisson 95% confidence intervals are shown. Grey bar indicates that *SMAD6* and *PPM1D* have been flagged in our previous analysis. 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 and the number of common genes. Numbers in parentheses include *SMAD6* and *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 set 1 + set 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. Genes associated with CES in any of the CES sets are highlighted with blue font and genes with missenses significant in NDD are highlighted with blue shading. Gray dashed frame shows protein complexes.

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 human genetics perspective, CES acts as a factor inflating mutation rate at certain positions. From the biological perspective, CES can be related to cancer with similar genes being involved in clonal expansions in normal tissues and oncogenic malformations.

Here, we studied two groups of effects of mutation rate biases introduced by CES. First, CES may result in false positives in the gene-disease associations (e.g. *MIB1*, *TCF7L2, TNPO3* and others). Second, CES might coincide with the causal involvement of certain genes in diseases. Genes like *FOXG1, DYRK1A, ARID1B*, *CTNNB1* or *PTEN* and possibly few others may contribute disproportionately to disease precisely due to simultaneous effects of CES drivers on clonal dynamics in spermatogonia and NDD phenotypes. Thus, the assessment of the effects of such genes on disease as well as risk conveyed by their mutations, must necessarily take the effects of CES into account.

In the context of population genetics, the substantial elevation of mutation rates due to CES can lead to mutational saturation and increased recurrence, even at negatively selected sites (Fig. 2b). As population cohorts grow larger, the sheer number of mutations will rise, making recurrence a dominant force[38,45]. This shift could render traditional absence/presence metrics, such as LOEUF, less informative in these contexts. CES-driven mutation rate increase in selected sites give a glimpse of opportunities and challenges for the analysis of future population datasets.

As noted above, CES may be studied both with experimental and statistical approaches. Remarkably, despite the difference in technology and even tissue that was sequenced to identify mutations, genes in set 2 and variants in set 3 are highly concordant with results in *Neville et. al.* In contrast, set 1 genes are more dissimilar between our studies. Set 1 genes should have substantial disease effect and moderate CES effect. Agnostic sperm sequencing, employed in *Neville et.*[20] *al.* could be underpowered to identify CES for such genes. Indeed out of 4 genes in set 1 only *CTNNB1* has significant excess of LoFs in NanoSeq study, with the smallest effect among significant genes. Ascertainment on phenotype effectively increases the number affected carriers and, correspondingly, informative data in the form of mutation counts. This directly increases power to detect CES for disease-associated genes.This is reminiscent of the first genetics-driven discovery of CES activity[4,5] of *FGFR2* and *FGFR3*.

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.

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. 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).
5. 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).
6. Maher, G. J. *et al.* Selfish mutations dysregulating RAS-MAPK signaling are pervasive in aged human testes. *Genome Res.* **28**, 1779–1790 (2018).
7. 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).
8. Kaplanis, J. *et al.* Evidence for 28 genetic disorders discovered by combining healthcare and research data. *Nature* **586**, 757–762 (2020).
9. 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).
10. 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).
11. 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).
12. Richter, F. *et al.* Genomic analyses implicate noncoding de novo variants in congenital heart disease. *Nat. Genet.* **52**, 769–777 (2020).
13. Francioli, L. C. *et al.* Genome-wide patterns and properties of de novo mutations in humans. *Nat. Genet.* **47**, 822–826 (2015).
14. Iossifov, I. *et al.* The contribution of de novo coding mutations to autism spectrum disorder. *Nature* **515**, 216–221 (2014).
15. Genome of the Netherlands Consortium. Whole-genome sequence variation, population

structure and demographic history of the Dutch population. *Nat. Genet.* **46**, 818–825 (2014).

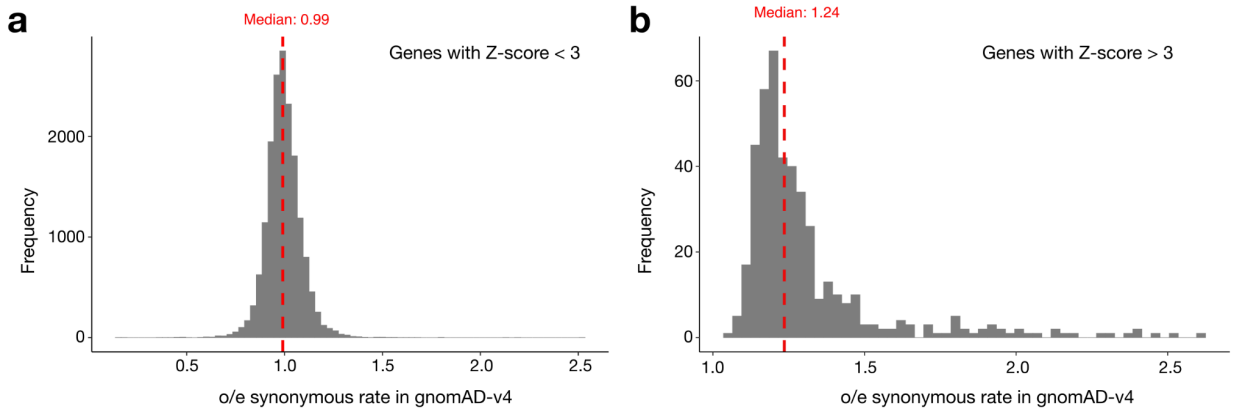
16. An, J.-Y. *et al.* Genome-wide de novo risk score implicates promoter variation in autism spectrum disorder. *Science* **362**, eaat6576 (2018).
17. Gulsuner, S. *et al.* Spatial and temporal mapping of de novo mutations in schizophrenia to a fetal prefrontal cortical network. *Cell* **154**, 518–529 (2013).
18. Zhao, G. *et al.* Gene4Denovo: an integrated database and analytic platform for de novo mutations in humans. *Nucleic Acids Res.* **48**, D913–D926 (2020).
19. Karczewski, K. J. *et al.* The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature* **581**, 434–443 (2020).
20. 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).
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. Kong, A. *et al.* Rate of de novo mutations and the importance of father's age to disease risk. *Nature* **488**, 471–475 (2012).
23. Neale, B. M. *et al.* Patterns and rates of exonic de novo mutations in autism spectrum disorders. *Nature* **485**, 242–245 (2012).
24. O'Roak, B. J. *et al.* Exome sequencing in sporadic autism spectrum disorders identifies severe de novo mutations. *Nat. Genet.* **43**, 585–589 (2011).
25. Seplyarskiy, V. B. & Sunyaev, S. The origin of human mutation in light of genomic data. *Nat. Rev. Genet.* **22**, 672–686 (2021).
26. 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).
27. 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).

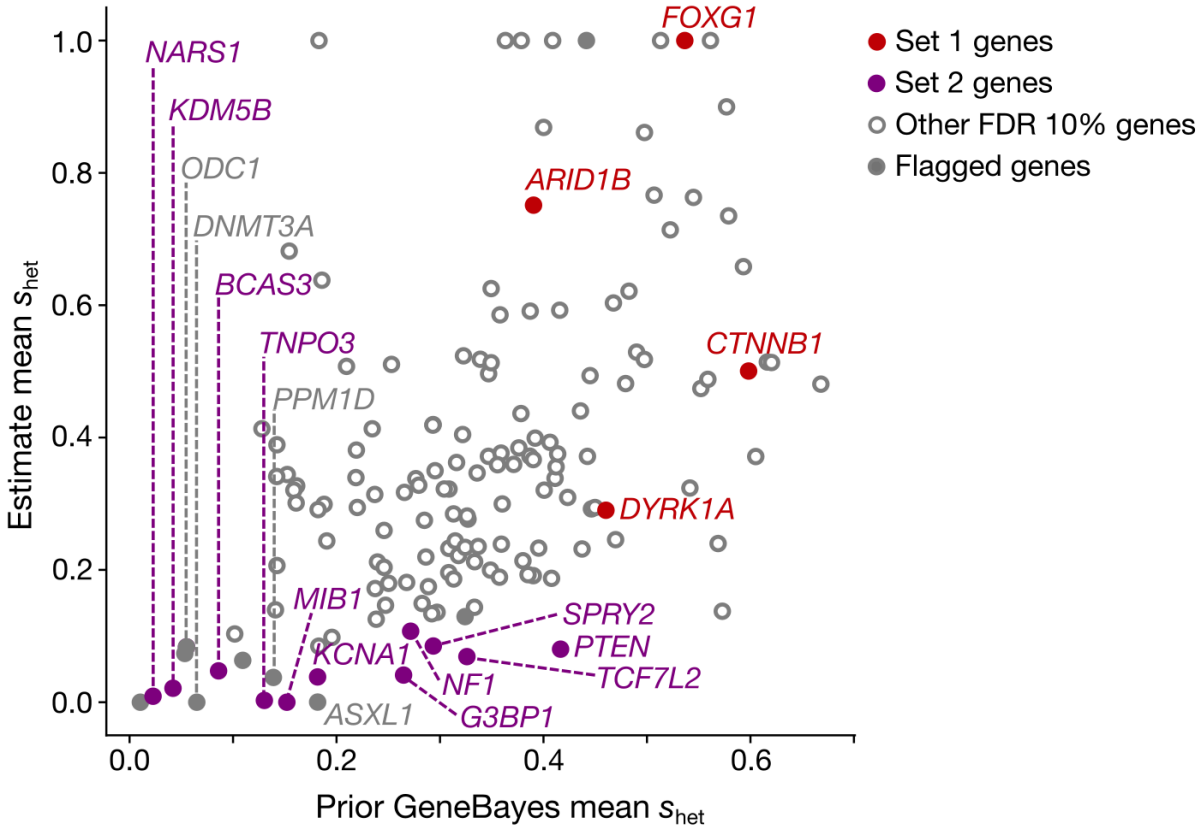
28. Jansen, S. *et al.* De Novo Truncating Mutations in the Last and Penultimate Exons of PPM1D Cause an Intellectual Disability Syndrome. *Am. J. Hum. Genet.* **100**, 650–658 (2017).
29. Totsika, V., Liew, A., Absoud, M., Adnams, C. & Emerson, E. Mental health problems in children with intellectual disability. *Lancet Child Adolesc. Health* **6**, 432–444 (2022).
30. Tate, J. G. *et al.* COSMIC: the Catalogue Of Somatic Mutations In Cancer. *Nucleic Acids Res.* **47**, D941–D947 (2019).
31. Boßelmann, C. M. *et al.* Analysis of 1386 epileptogenic brain lesions reveals association with DYRK1A and EGFR. *Nat. Commun.* **15**, 10429 (2024).
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. Gogate, A. *et al.* The genetic landscape of autism spectrum disorder in an ancestrally diverse cohort. *NPJ Genomic Med.* **9**, 62 (2024).
34. 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).
35. 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).
36. 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).
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. 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).

39. Nei, M. The frequency distribution of lethal chromosomes in finite populations. *Proc. Natl. Acad. Sci.* **60**, 517–524 (1968).
40. Bush, S. J. *et al.* Adult Human, but Not Rodent, Spermatogonial Stem Cells Retain States with a Foetal-like Signature. *Cells* **13**, 742 (2024).
41. 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).
42. 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).
43. 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).
44. 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).
45. Spence, J. P., Zeng, T., Mostafavi, H. & Pritchard, J. K. Scaling the discrete-time Wright-Fisher model to biobank-scale datasets. *Genetics* **225**, iyad168 (2023).

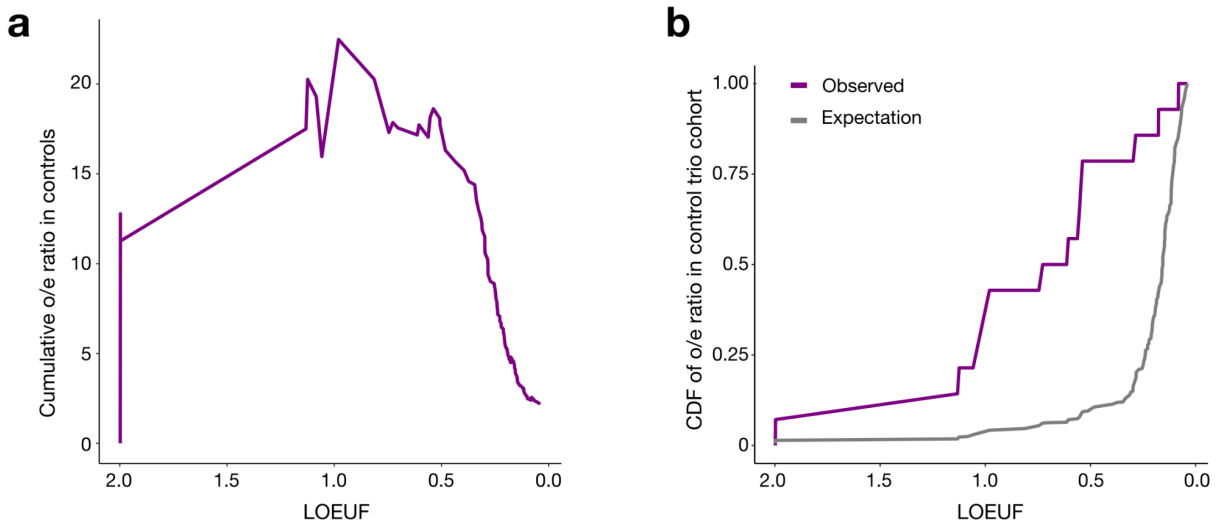
Extended Data Figures



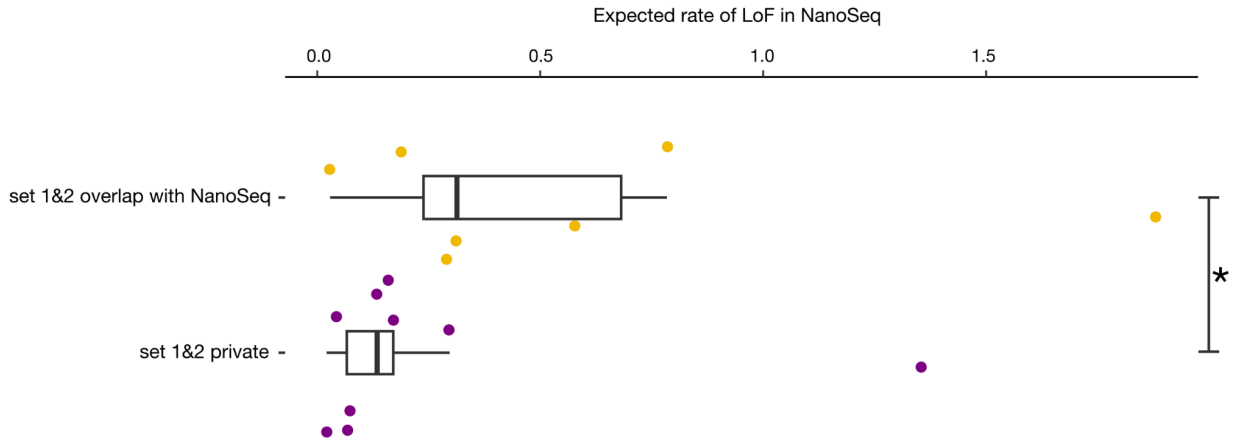
Extended Data Figure 1 | Observed the expected ratio of synonymous SNVs in gnomAD-v4 for genes **(a)** with Z-score <3 and **(b)** with Z-score >3 that were flagged.



Extended Data Figure 2 | Prior of GeneBayes s_{het} calculated using biological features of genes (x-axis) and s_{het} values updated with LoF polymorphism data from gnomAD-v4 (y-axis).



Extended Data Figure 3 | Ratio of observed-to-expected counts of LoF *de novo* mutations in cohort of control trios for genes with significant excess of LoF mutations in ASD/NDD. **(a)** We aggregate all considered genes with LOEUF values lower than the value shown on the x-axis. **(b)** CDF for LOF counts in control cohort and for expectation.



Extended Data Figure 4 | Expected LoF rate in NanoSeq data for set 1 and set 2 genes overlapping with genes significant in NanoSeq and private set 1/set 2 genes. * corresponds to p-value<0.05 Mann-Whitney U test.