nature portfolio

https://doi.org/10.1038/s41588-021-00899-8

Supplementary information

Recent ultra-rare inherited variants implicate new autism candidate risk genes

In the format provided by the authors and unedited

Supplementary Note

Patterns of private, transmitted variants in protein-coding regions (X chromosome)

If we partition our discovery cohort by sex to include the X chromosome, we find the increased burden of private LGD variants relative to increasing gene constraint becomes both stronger and more significant in males (**Supplementary Figures 8 and 9, and Supplementary Table 9**, pLI ≥ 0.90, male vs. female OR = 1.36 vs. 1.10, permutation p-value = 0.004). Since we focused on only the X chromosome, higher pLI thresholds were not considered due to the limited number of genes at these constraint cutoffs. We also find one ASD/NDD gene set, Coe et al., comprised of genes enriched for DNMs in autism cases reaches nominal significance in females (**Supplementary Figure 14, and Supplementary Table 11**, OR = 1.71, nominal p = 0.0316) if we assess the burden in males and females separately to include genes on the X chromosome.

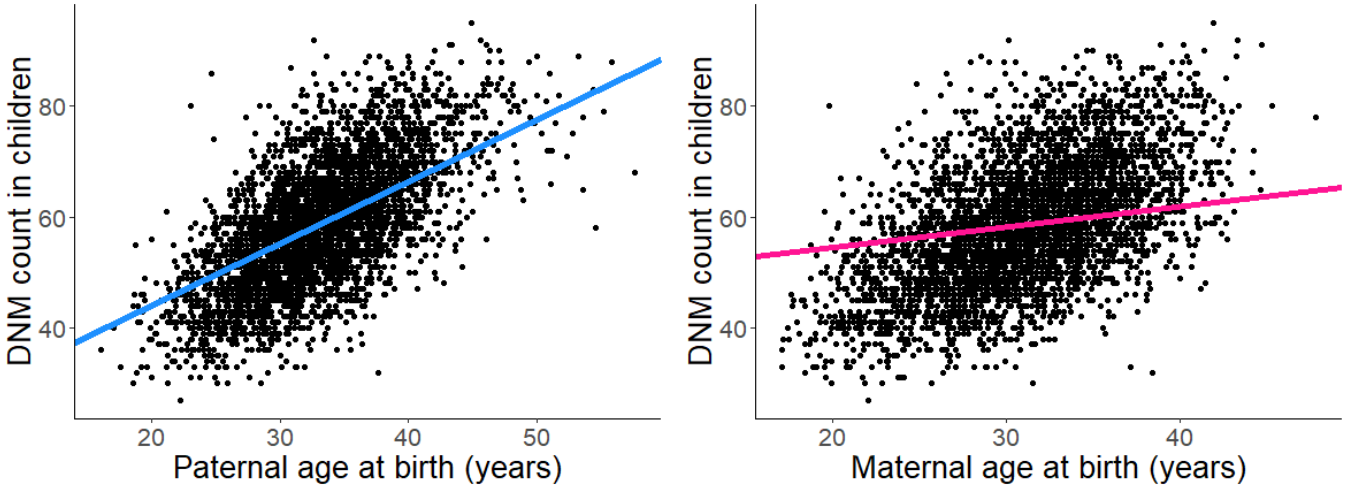
Patterns of private, transmitted variants in protein-coding regions (CNVs)

Based on SNV findings, we considered CNVs for potential to similarly disrupt genes and defined "private CNVs" as those transmitted and observed only once in the parent population. We observe a trend of probands carrying more private CNVs than siblings as pLI scores increase (**Supplementary Table 13**), but the observation does not reach statistical significance, likely due to insufficient sample size and rarity of such CNVs (**Supplementary Figure 16**). Combining genome and exome data where CNV sensitivity is reduced fails to achieve significance, although a trend of increased burden over genes enriched for DNMs in NDD genes is observed (**Supplementary Table 14**).

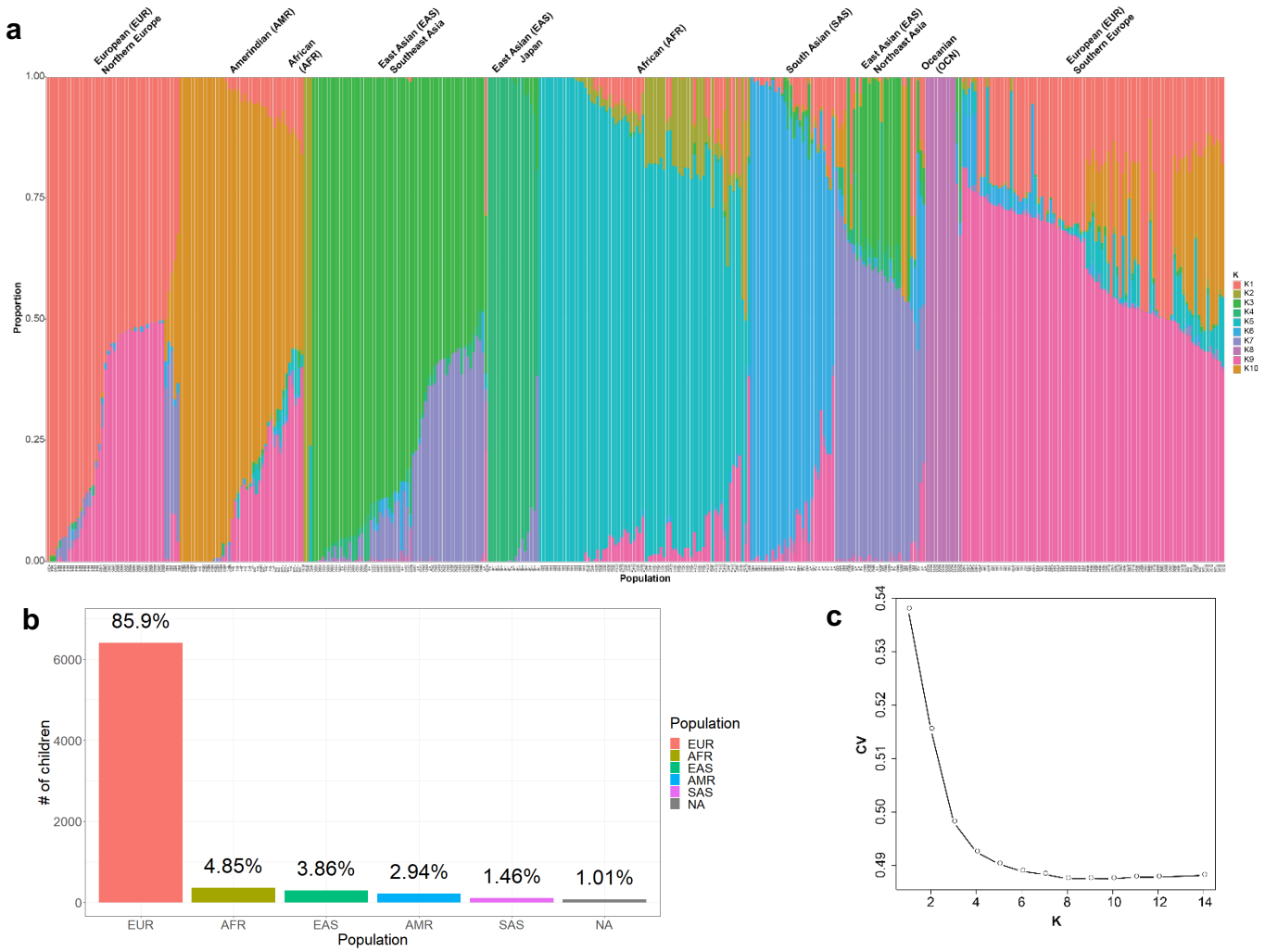
A multi-hit model for ASD

If we consider variants with pLI ≥ 0.9, we find 53.8% of the 158 probands with two hits identified in this study inherit each variant from a different parent and such transmissions may explain why neither parent is affected. To build on this finding, we consider 172 families where both parents carry exactly one LGD variant in a conserved gene. Under Mendel's laws, we would expect 25% of probands to

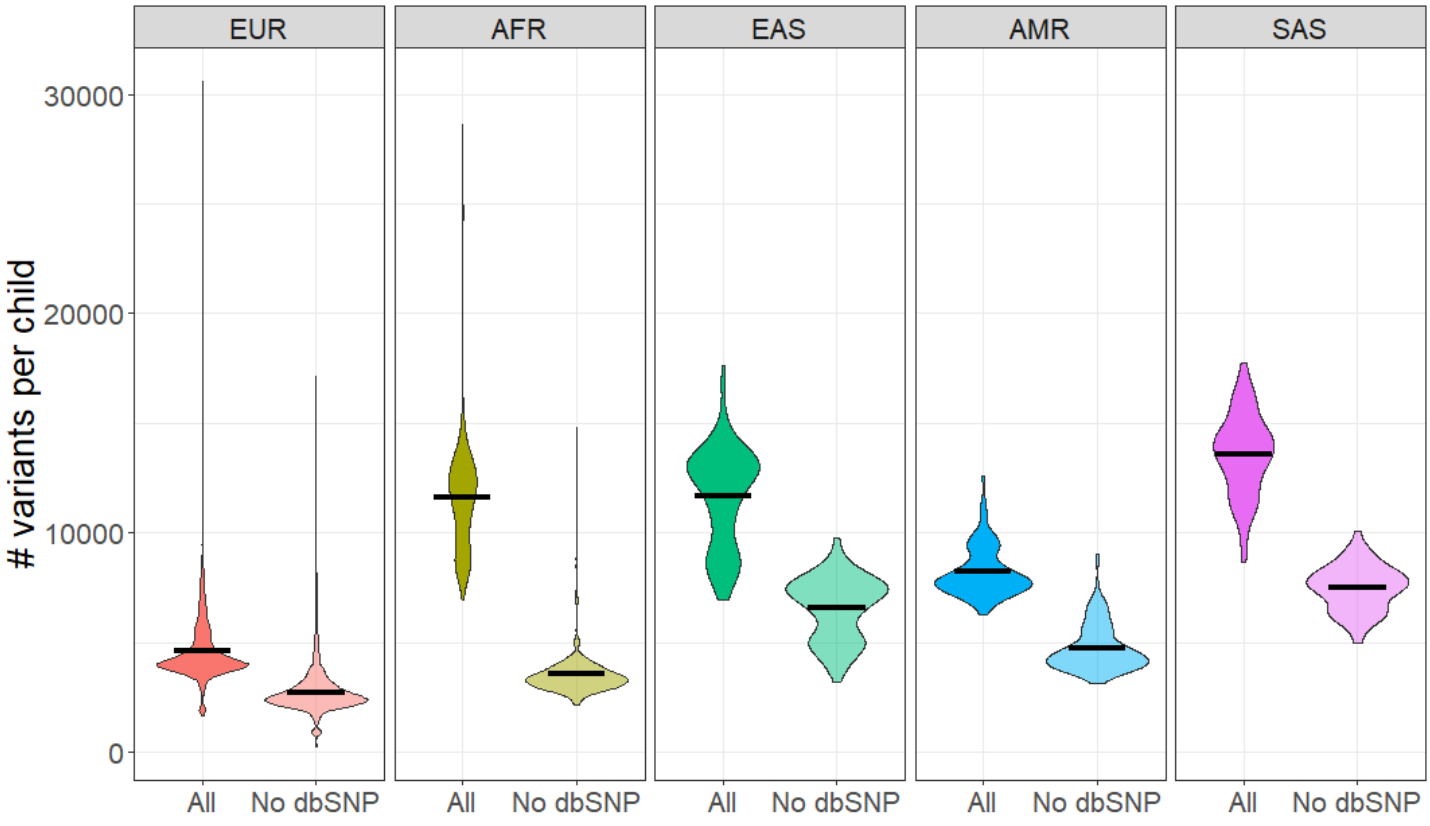
inherit two hits. We find 28% of probands (n=63 out of 225) carry two hits out of all probands with both parents carrying an inherited LGD variant, which is not significantly more than the expected 25% (binomial test, one-sided p-value = 0.32) and 22.5% of siblings (n=23 out of 102) carry two hits out of all siblings with both parents carrying an inherited LGD variant (binomial test, p-value = 0.64).



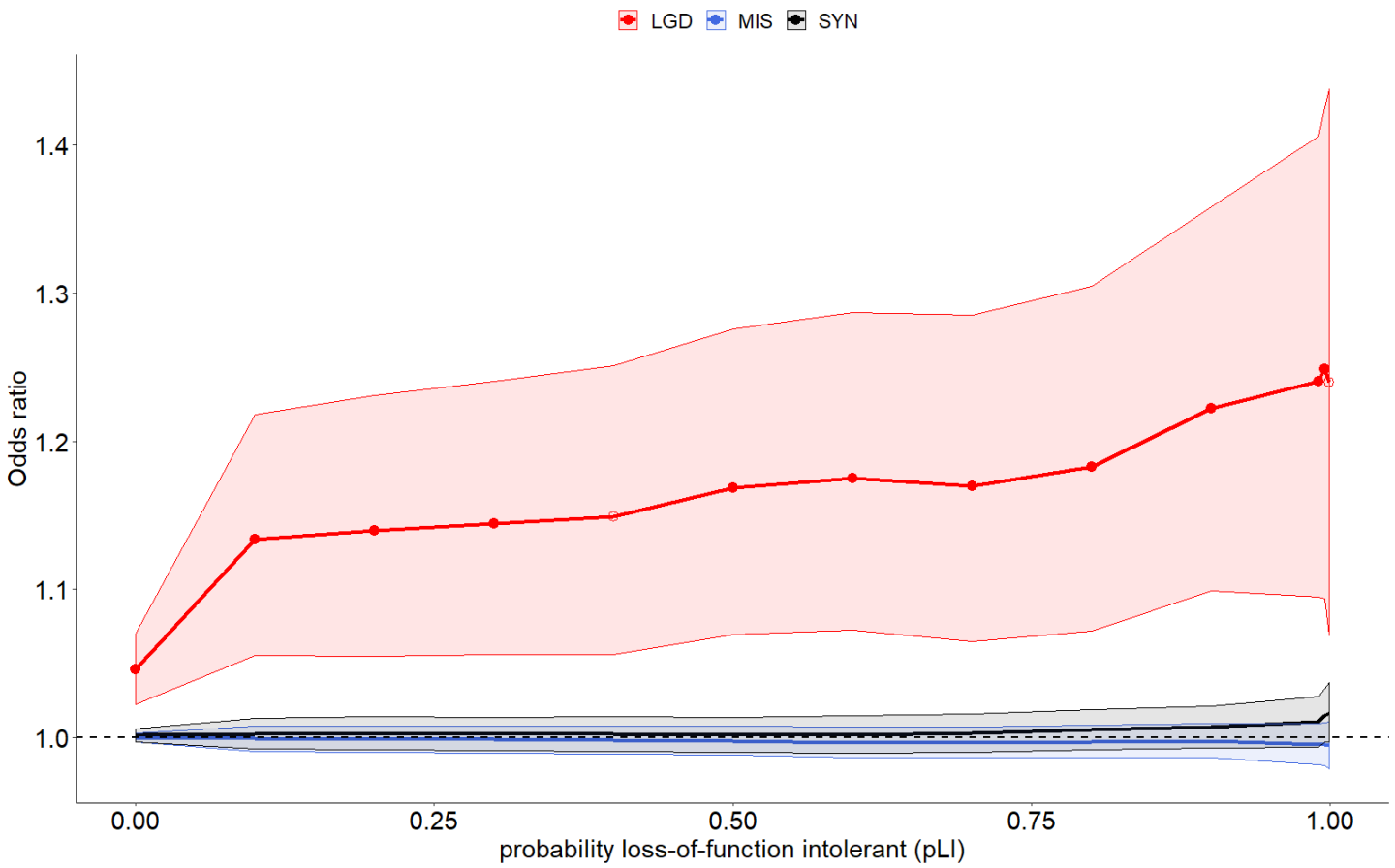
Supplementary Figure 1: Relationship between parental age and DNM counts in children. We observe a significant correlation between parental age and DNM counts in children. As expected, fathers contribute, on average, more DNMs per year of age than mothers (1.11 vs. 0.37).



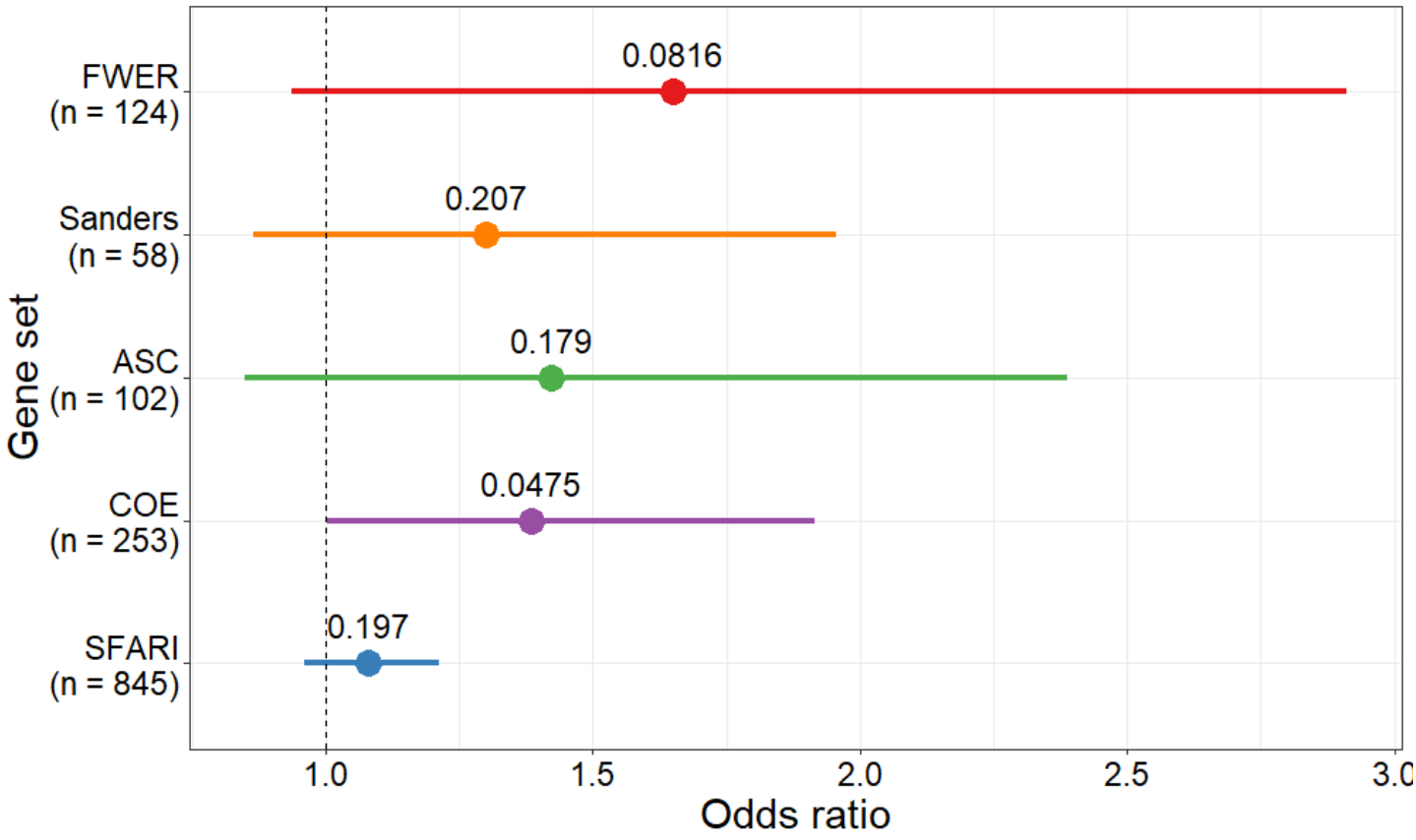
Supplementary Figure 2: Ancestry inference for diversity panel and autism family samples.
a) Plot of ancestry proportions as estimated by ADMIXTURE for 15 randomly sampled individuals from the SGDP and 1KG populations from each reported population. Known population for each individual is labeled on the x-axis and likely geographic origin is noted above each cluster.
b) Ancestry assignments for our discovery population. Ancestry for each individual was assigned according to the population with the largest proportion in that individual. c) Cross-validation (CV) error across different K for ADMIXTURE.



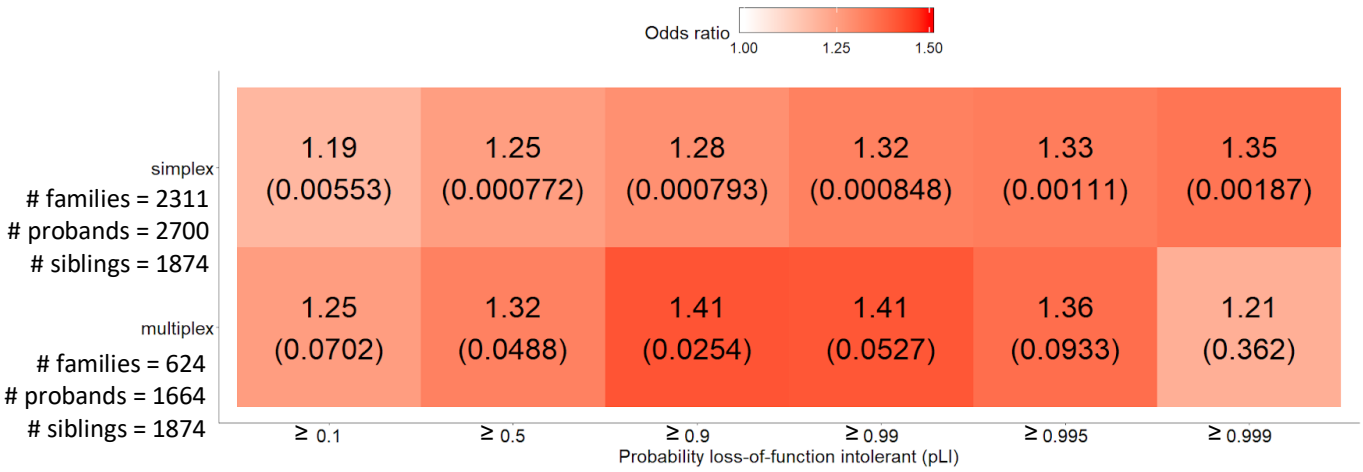
Supplementary Figure 3: Private, transmitted variant counts. Variant counts per child grouped by ancestry (EUR = European (n = 5,685), AFR = African (n = 290), EAS = East Asian descent (n = 252), AMR = Amerindian (n = 193), SAS = South Asian (n = 103)) before (All) and after (No dbSNP) filtering with dbSNPv150. Excess of private variants is partially but not fully resolved after excluding sites observed in dbSNP. We were unable to assign ancestry to one of these five population groups for 74 of the children in this study.



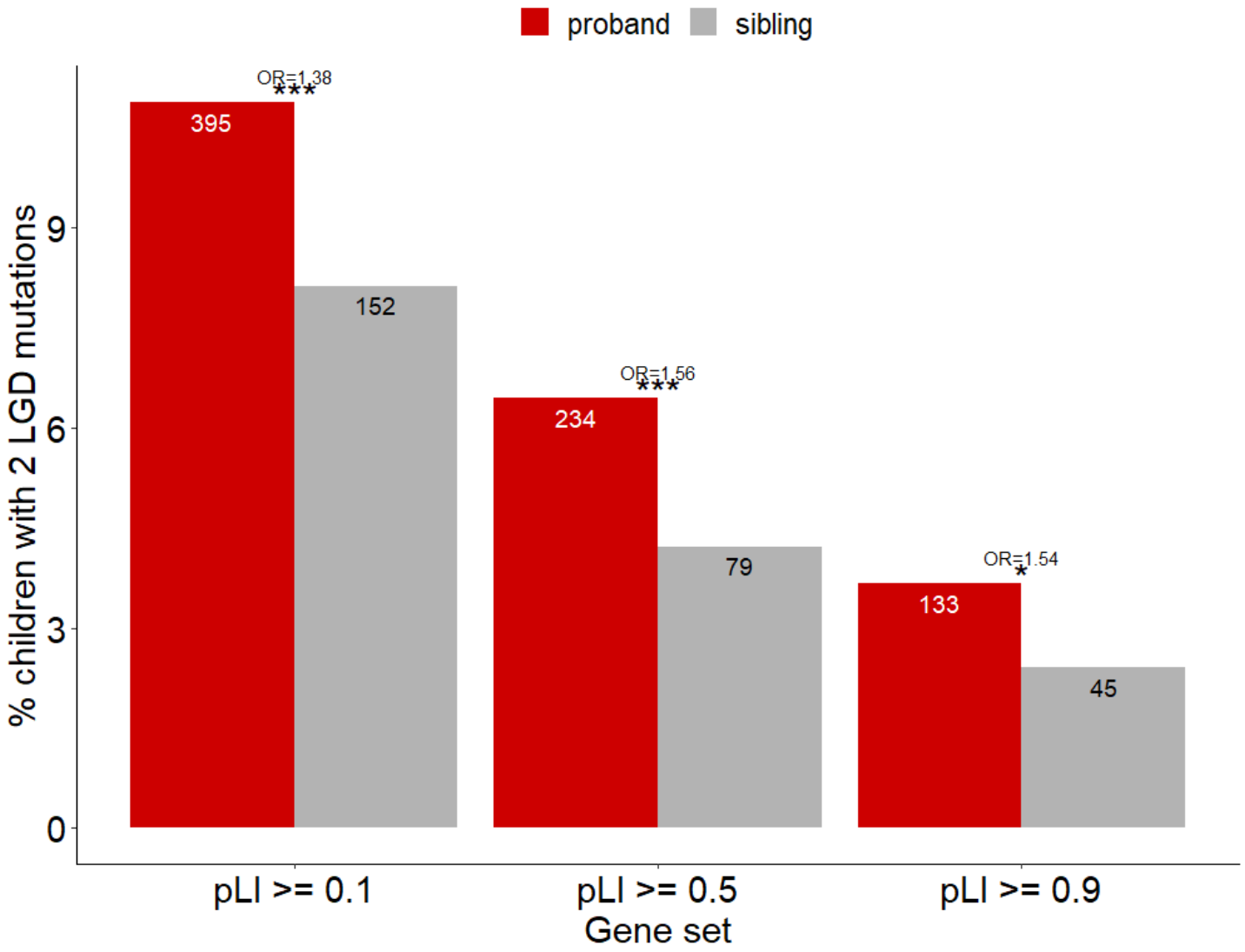
Supplementary Figure 4: Patterns of private variant burden with gene constraint in EUR subset. Burden of private LGD variants in probands increases with gene constraint in the EUR subset of the discovery cohort (n = 3,636 probands and 1,873 siblings). Excludes families with monozygotic twins. Filled circles indicate Bonferroni-corrected p-values < 0.05 (42 tests). Unfilled circles indicated uncorrected p-value < 0.05. Shaded areas indicate the 95% confidence interval for the odds ratio estimates. Odds ratios and confidence intervals calculated by logistic regression.



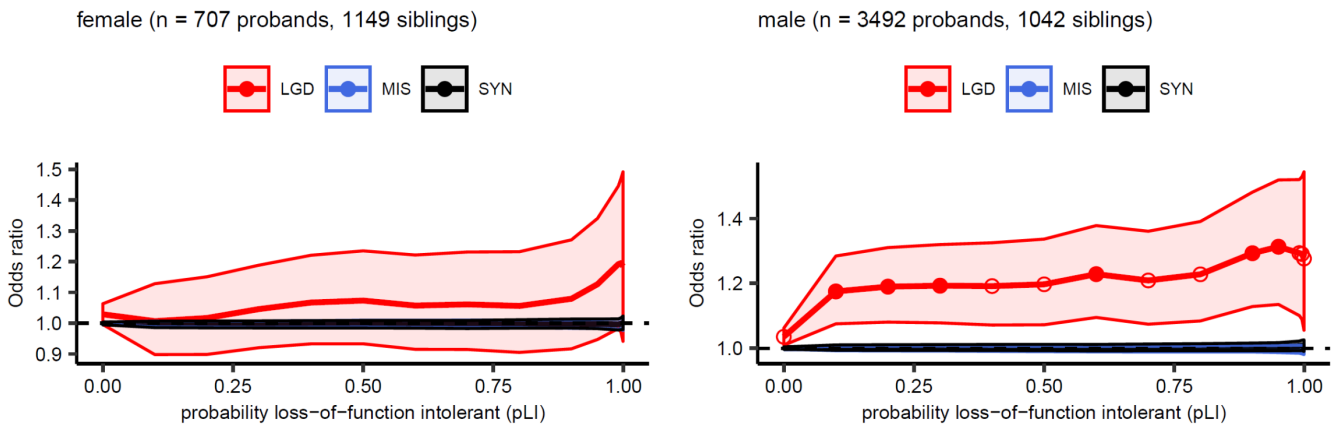
Supplementary Figure 5: Private LGD variant burden across five DNM-enriched gene sets in EUR subset. The EUR subset of the discovery cohort is comprised of 3,636 probands and 1,873 siblings. Analysis excludes families with monozygotic twins. Reported p-values are nominal, and odds ratios and confidence intervals are estimated by logistic regression.



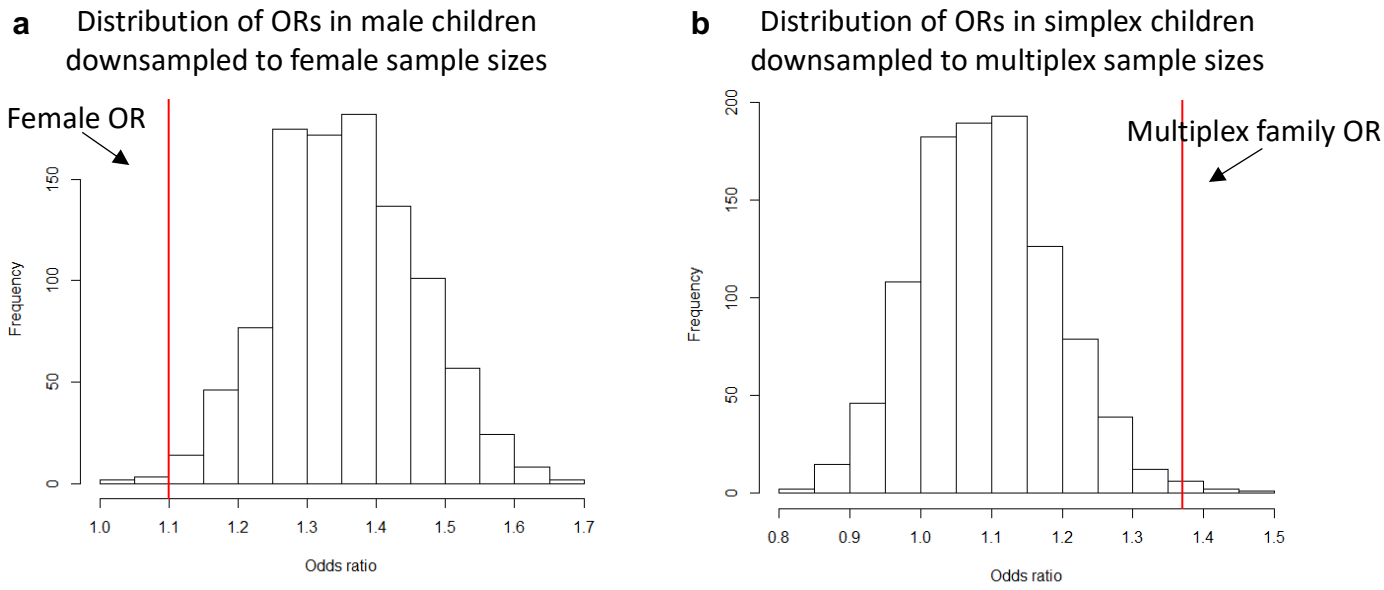
Supplementary Figure 6: Burden of private LGD variants across gene constraint in EUR subset of the discovery cohort. Excludes families with monozygotic twins. Reported p-values are Bonferroni corrected for 12 tests. Odds ratios are calculated by two-sided Fisher's exact test.



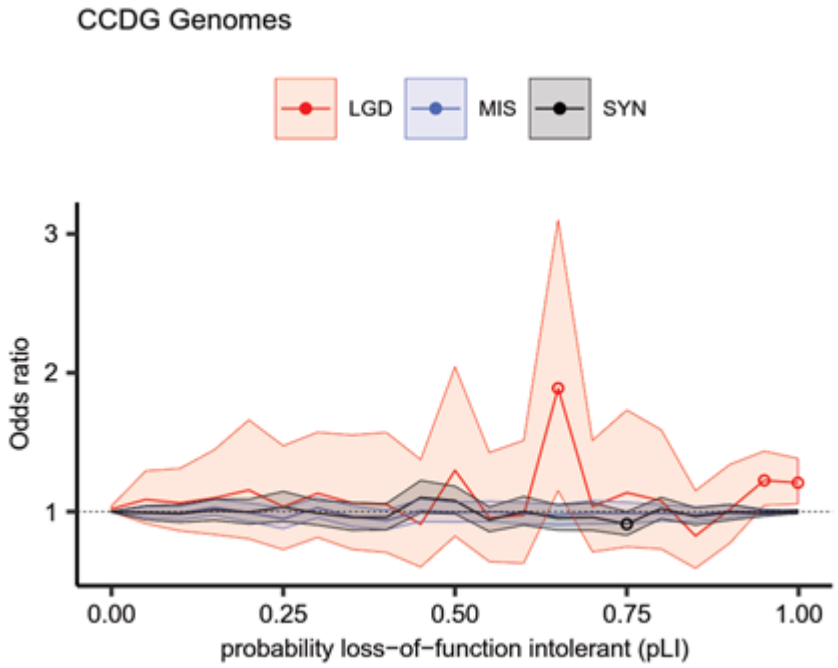
Supplementary Figure 7: Burden of two LGD variants in EUR subset. The EUR subset of the discovery cohort is comprised of 3,636 probands and 1,873 siblings. Excludes families with monozygotic twins. Odds ratios are calculated by two-sided Fisher's exact test. Reported p-values are nominal. *= p<0.05, **= p<0.01, ***= p< 0.005.

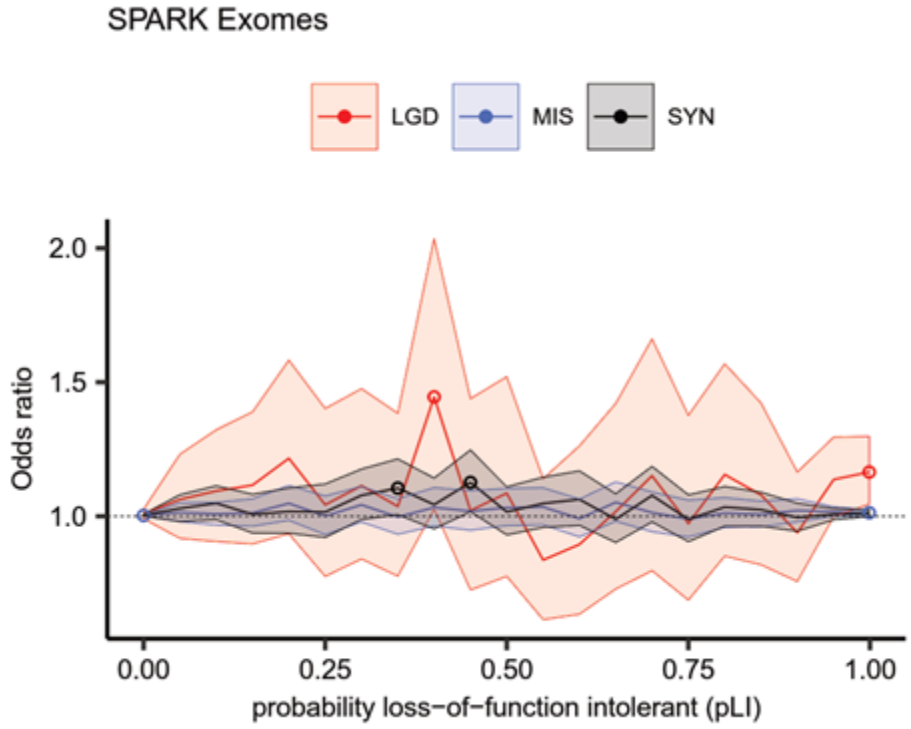


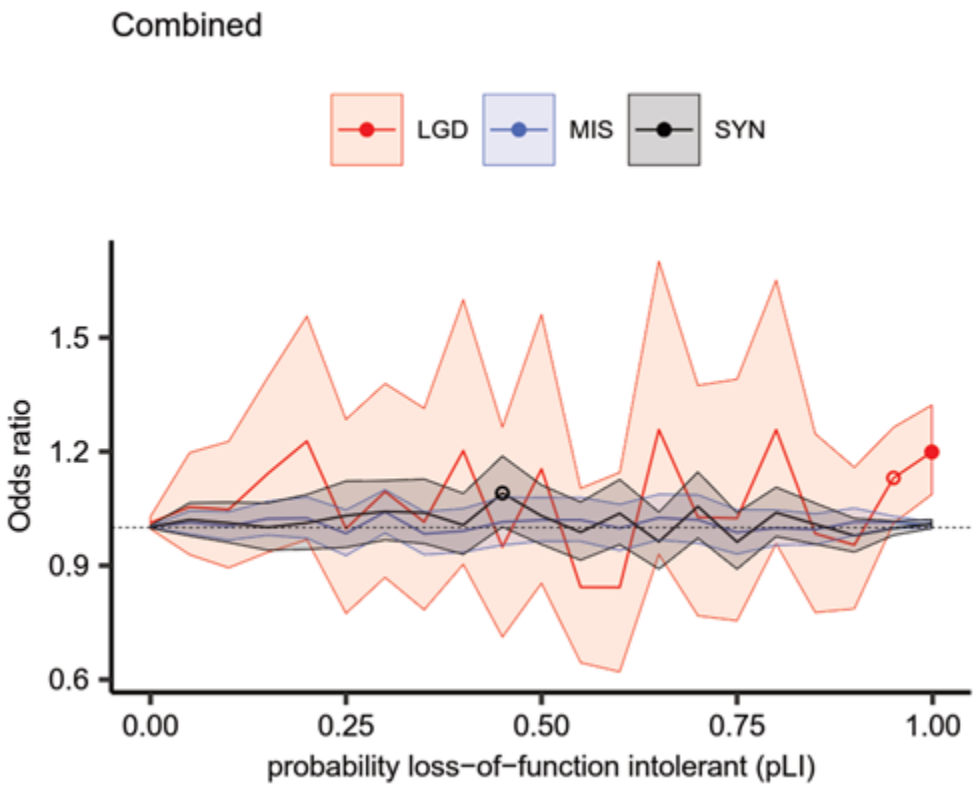
Supplementary Figure 8: Patterns of private variant burden with gene constraint by sex. Burden of private LGD variants in probands increases with gene constraint in males and females. Filled circles indicate Bonferroni-corrected p-values < 0.05 (84 tests). Unfilled circles indicated uncorrected p-value < 0.05. Shaded area indicates the 95% confidence interval around the odds ratio estimates. Odds ratios are calculated by two-sided Fisher's exact test



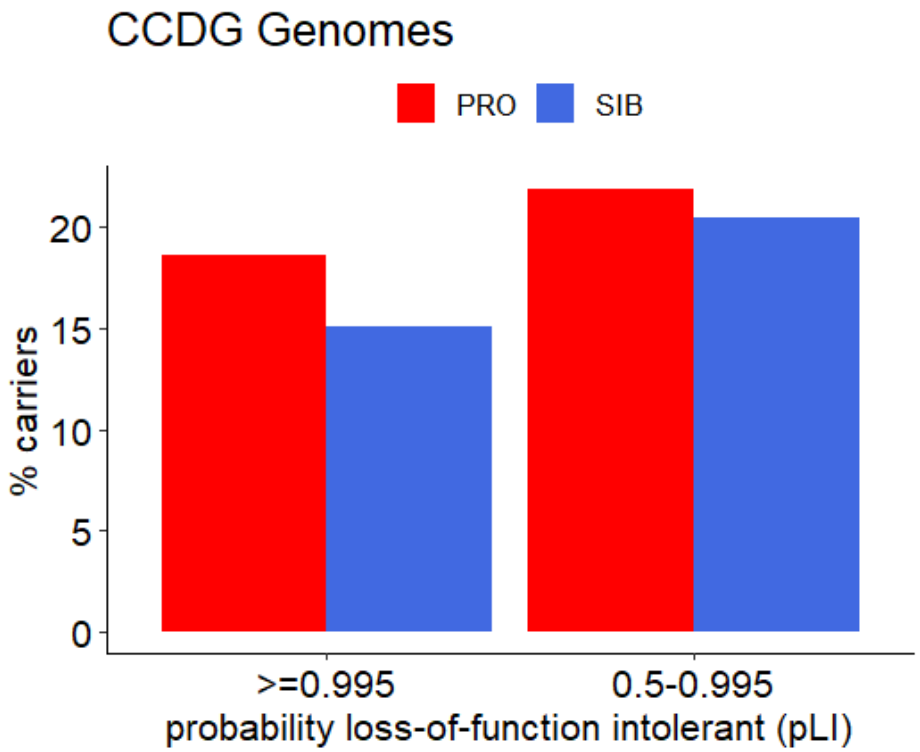
Supplementary Figure 9: Permutation tests for estimated burden in probands. a) Male permutation tests were performed by randomly sampling 707 male probands to match the sample size of the female probands. Since there are ~100 fewer male siblings as compared to female siblings, all male siblings were used (i.e., no downsampling for male siblings occurred) and then the burden of private LGD variants in probands vs. siblings in genes with pLI ≥ 0.9 was calculated using a Fisher's exact test. b) Simplex tests were performed by randomly sampling 2,691 probands and 533 siblings to match the sample sizes observed for the multiplex families and then calculated the burden (OR) of private LGDs variants in probands vs. siblings in genes with pLI ≥ 0.1 using a Fisher's exact test. Red lines indicate the observed OR for females and multiplex as calculated by Fisher's exact test, respectively. We performed 1,000 permutations of each analysis.


CCDG Genomes
LGD
MIS
SYN
Odds ratio
3
2
1
0.00
0.25
0.50
0.75
1.00
probability loss−of−function intolerant (pLI)

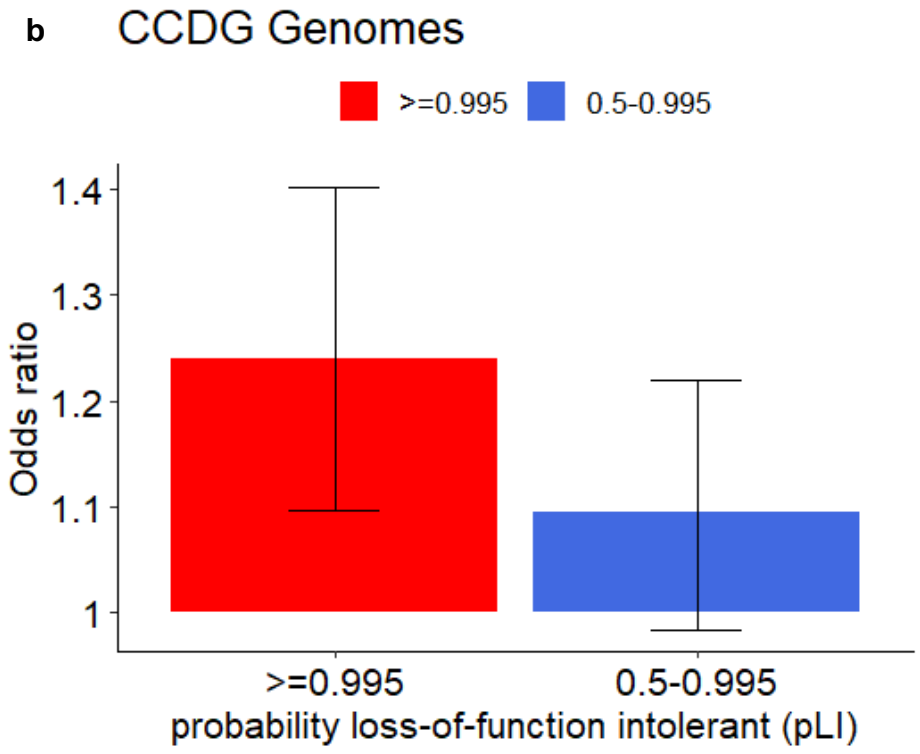

SPARK Exomes
LGD
MIS
SYN
Odds ratio
2.0
1.5
1.0
0.00
0.25
0.50
0.75
1.00
probability loss−of−function intolerant (pLI)



Supplementary Figure 10: Odds ratios for discrete pLI bins for the discovery, replication and combined cohorts. Open circles represent nominal significance and filled circles represent Bonferroni-corrected significance. The shading around the line indicates the 95% confidence interval around each odds ratio estimate. Odds ratios were calculated by two-sided Fisher's exact test.

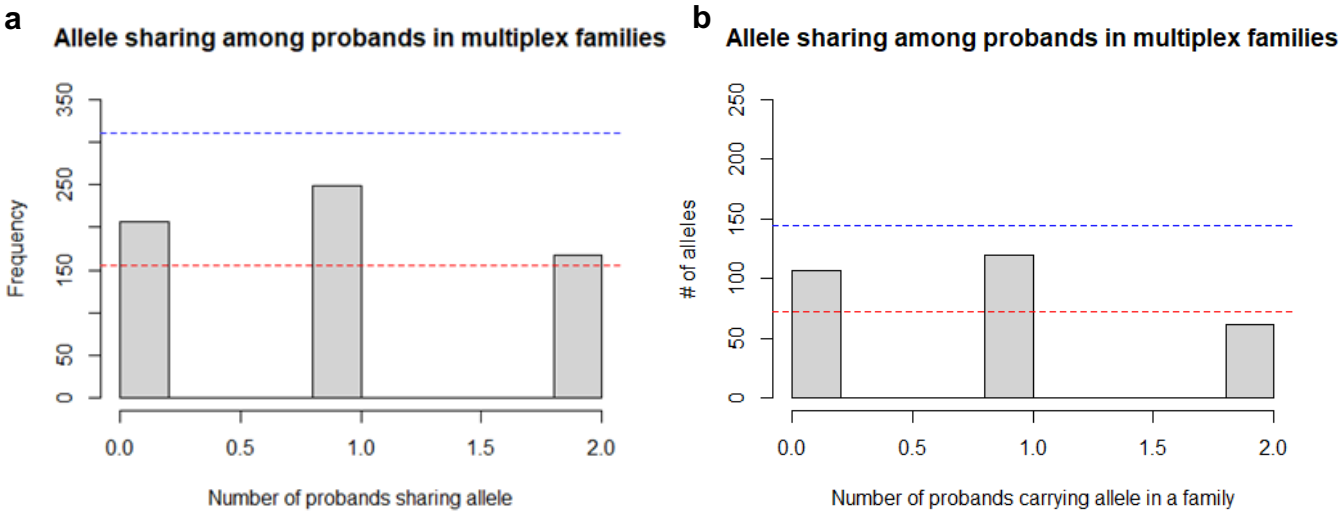


b



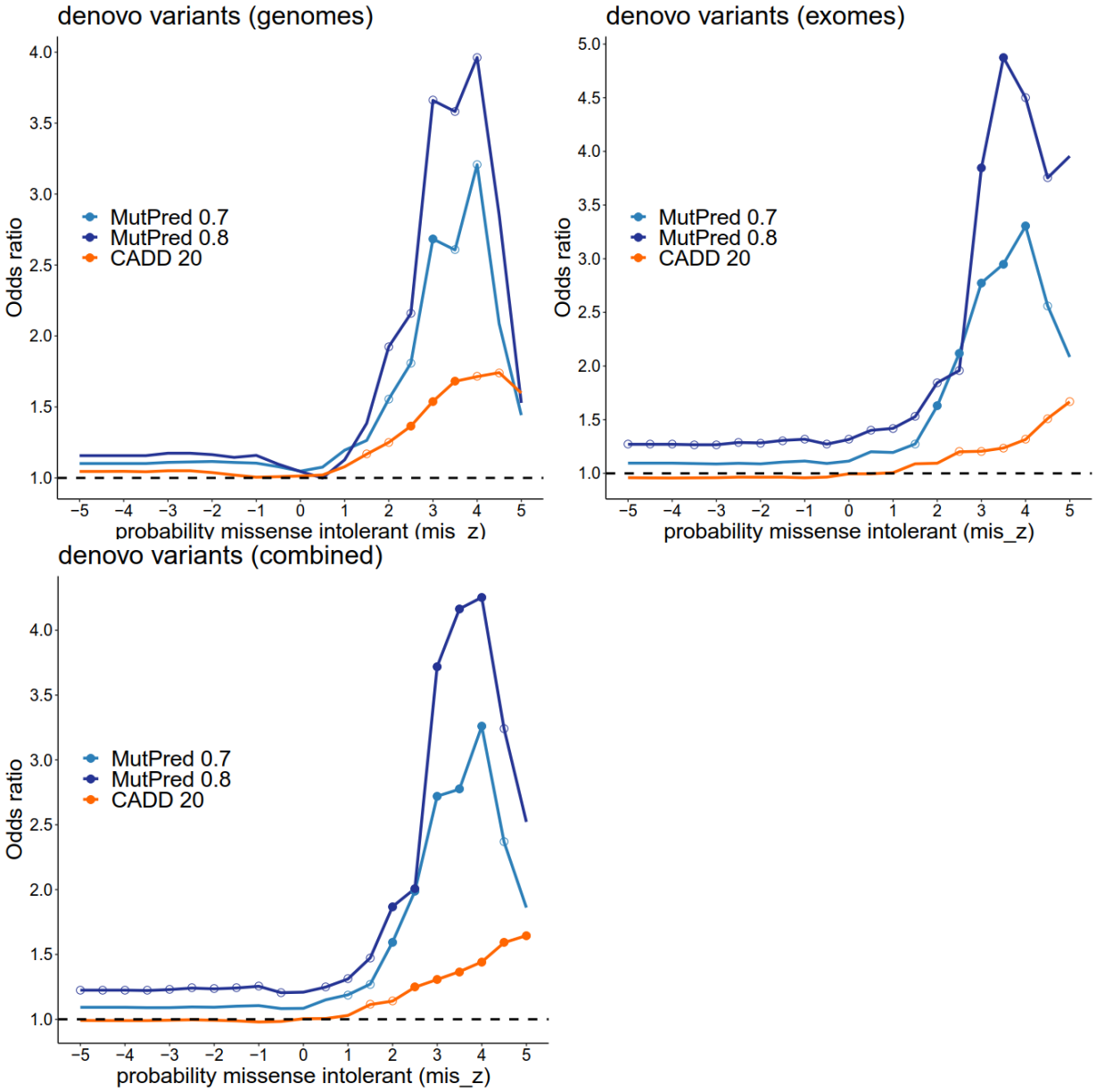
Supplementary Figure 11: Replication of ASC signal with private, transmitted LGD variants. a) Percentage of carriers by case–control status for the two discrete pLI bins used in the Satterstrom et al. For both bins we observe an increase in carriers among probands as compared to siblings (this difference is significant by two-sided Fisher's exact test for pLI ≥ 0.995). b) Odds ratio as calculated by two-sided Fisher's exact test comparing the proportion of affected and unaffected children for the

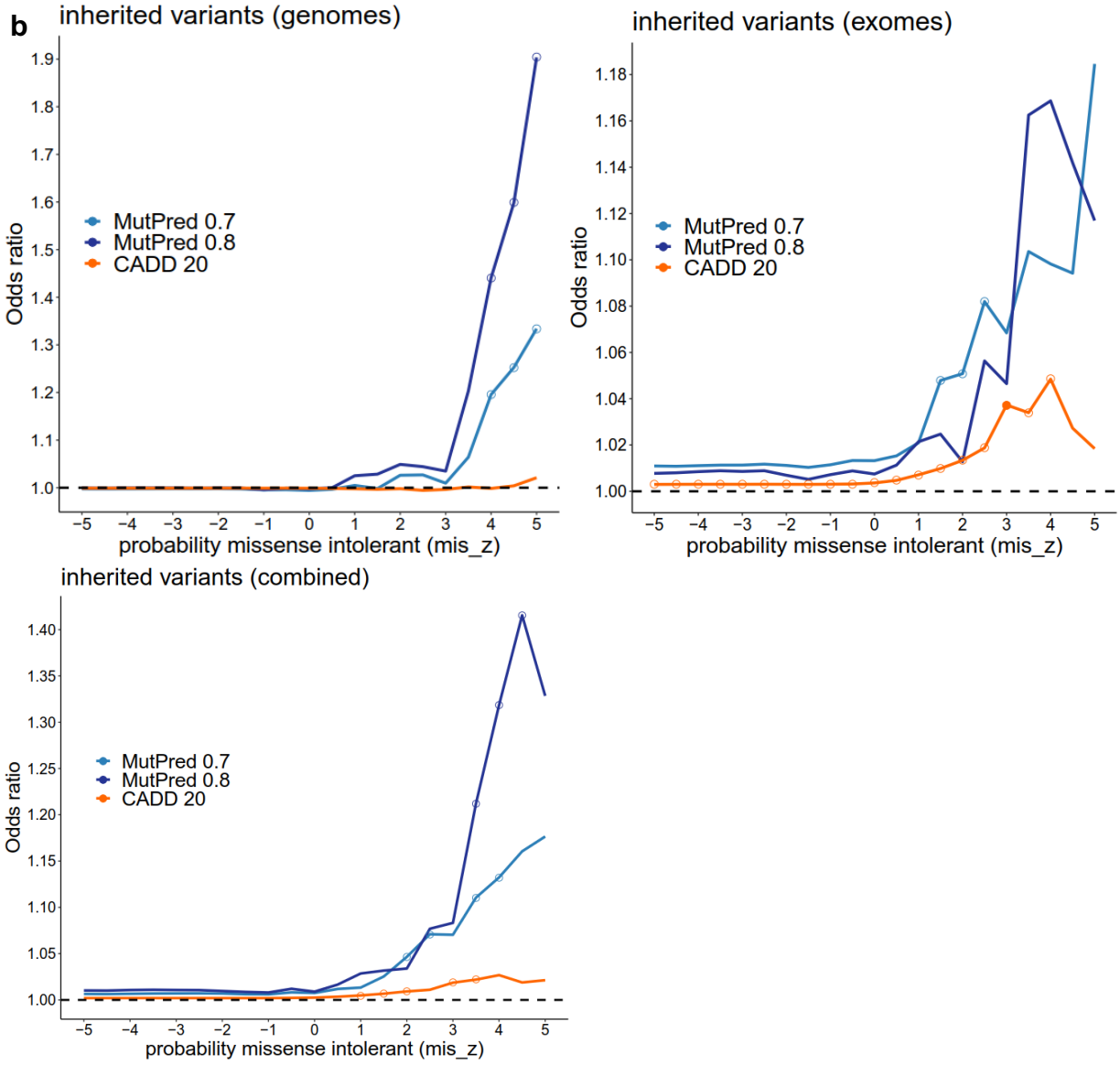
two pLI bins used in Satterstrom et al. (n = 4,201 affected and 2,191 unaffected children). Error bars indicate the 95% confidence interval around the odds ratio estimate.



Supplementary Figure 12: Allele sharing in families with two probands. Under the null hypothesis, we would expect each child to have a 50% chance of inheriting a transmitted LGD variant from their parent. The red line indicates the expected number of families with zero- and two-proband transmissions (expect 25% of events), and the dark blue line indicates the number of single proband transmissions (expect 50% of events). a) At pLI ≥ 0.1, we observe a significant depletion of transmissions to only one proband (two-sided binomial test, p = 5.0e-4) and a significant increase in transmissions to neither proband (two-sided binomial test, p = 1.3e-4). b) We assessed 238 multiplex families with exactly two probands (and no unaffected siblings) and where at least one parent was carrying a private LGD variant in a gene with pLI ≥ 0.99. We observe a significant depletion of transmissions to only one proband (two-sided binomial test, p = 0.04) and a significant increase in transmissions to neither proband (two-sided binomial test, p = 4.3e-5).

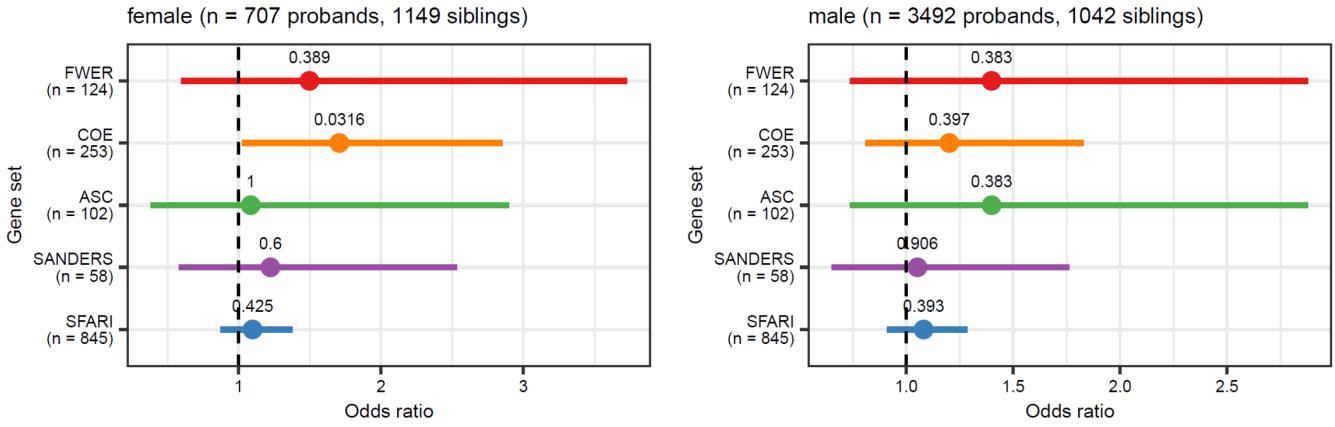
a


denovo variants (genomes)
Odds ratio
MutPred 0.7
MutPred 0.8
CADD 20
probability missense intolerant (mis_z)
denovo variants (exomes)
Odds ratio
MutPred 0.7
MutPred 0.8
CADD 20
probability missense intolerant (mis_z)
denovo variants (combined)
Odds ratio
MutPred 0.7
MutPred 0.8
CADD 20
probability missense intolerant (mis_z)

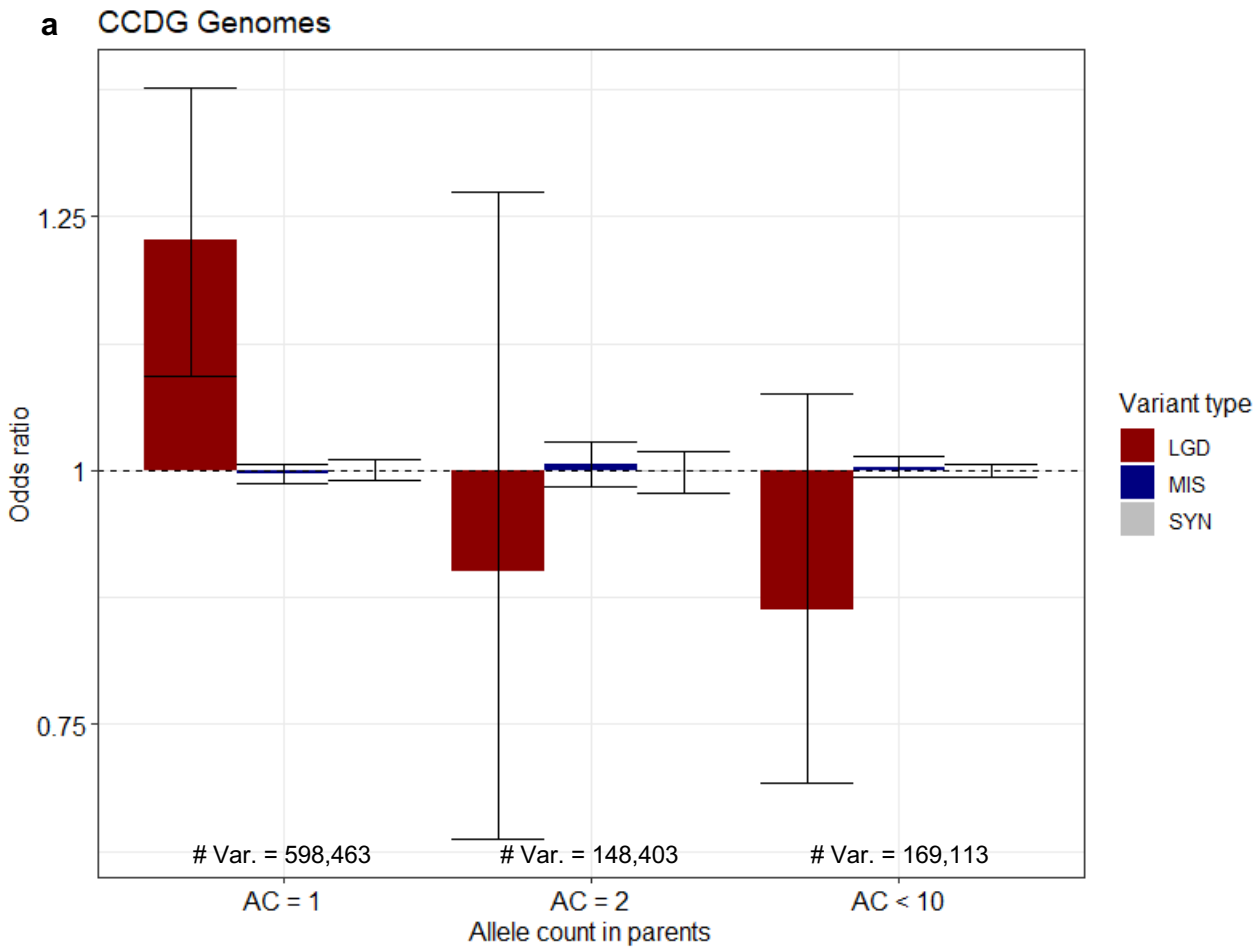


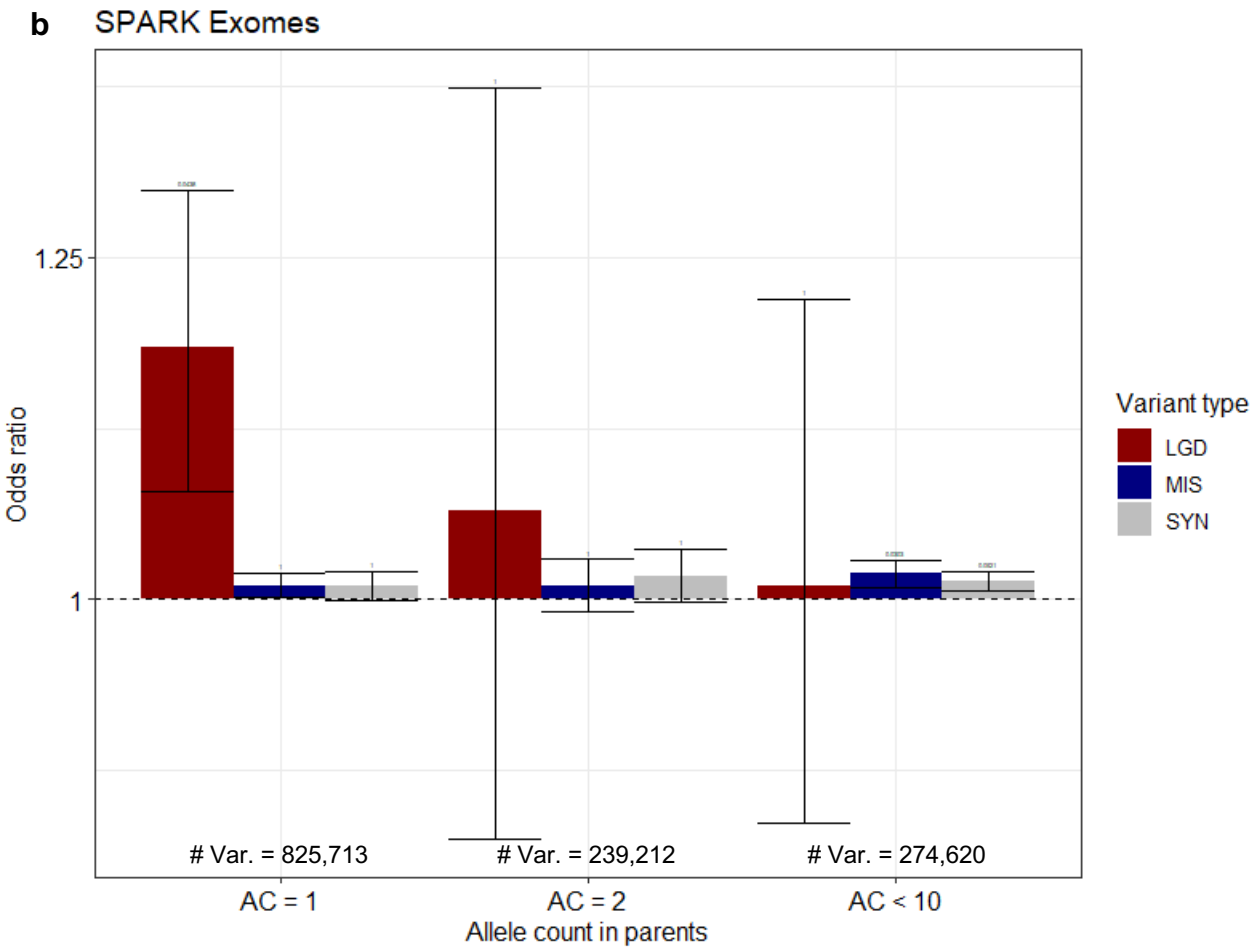
Supplementary Figure 13: Excess of intolerant missense variants in autism probands. Odds ratios are plotted against missense Z-score for two pathogenicity predictors, CADD and MutPred, at two different thresholds for a) *de novo* mutations and b) inherited variants. The genome dataset is combined from the SSC, SAGE, TASC and AGRE studies for a total of 3,587 families (5,062 probands and 2,411 siblings). The exome dataset is from SPARK and includes 5,906 families (6,556 probands and 3,037 siblings). Open circles represent nominal significant differences ($p < 0.05$) while filled circles represent Bonferroni-corrected significance for 33 tests. Odds ratios estimated by logistic regression.

a

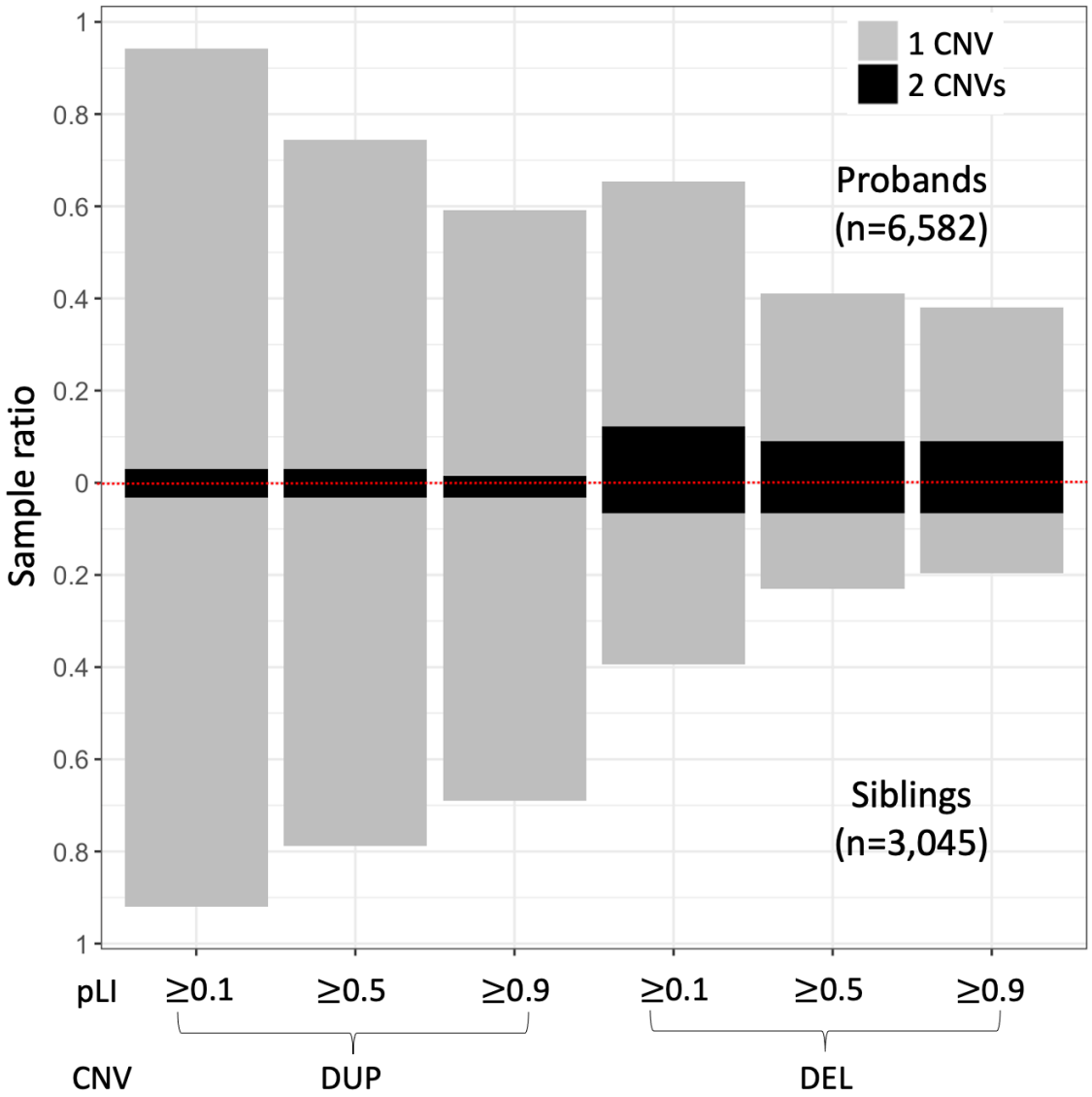


Supplementary Figure 14: Private LGD variant burden across five DNM-enriched gene sets in males and females. Points indicate the odds ratio for each gene set as calculated by Fisher's exact test, and error lines indicate the 95% confidence interval around the odds ratio estimate. P-values have been Bonferroni corrected for 10 tests. No gene sets reach significance after Bonferroni-correction.

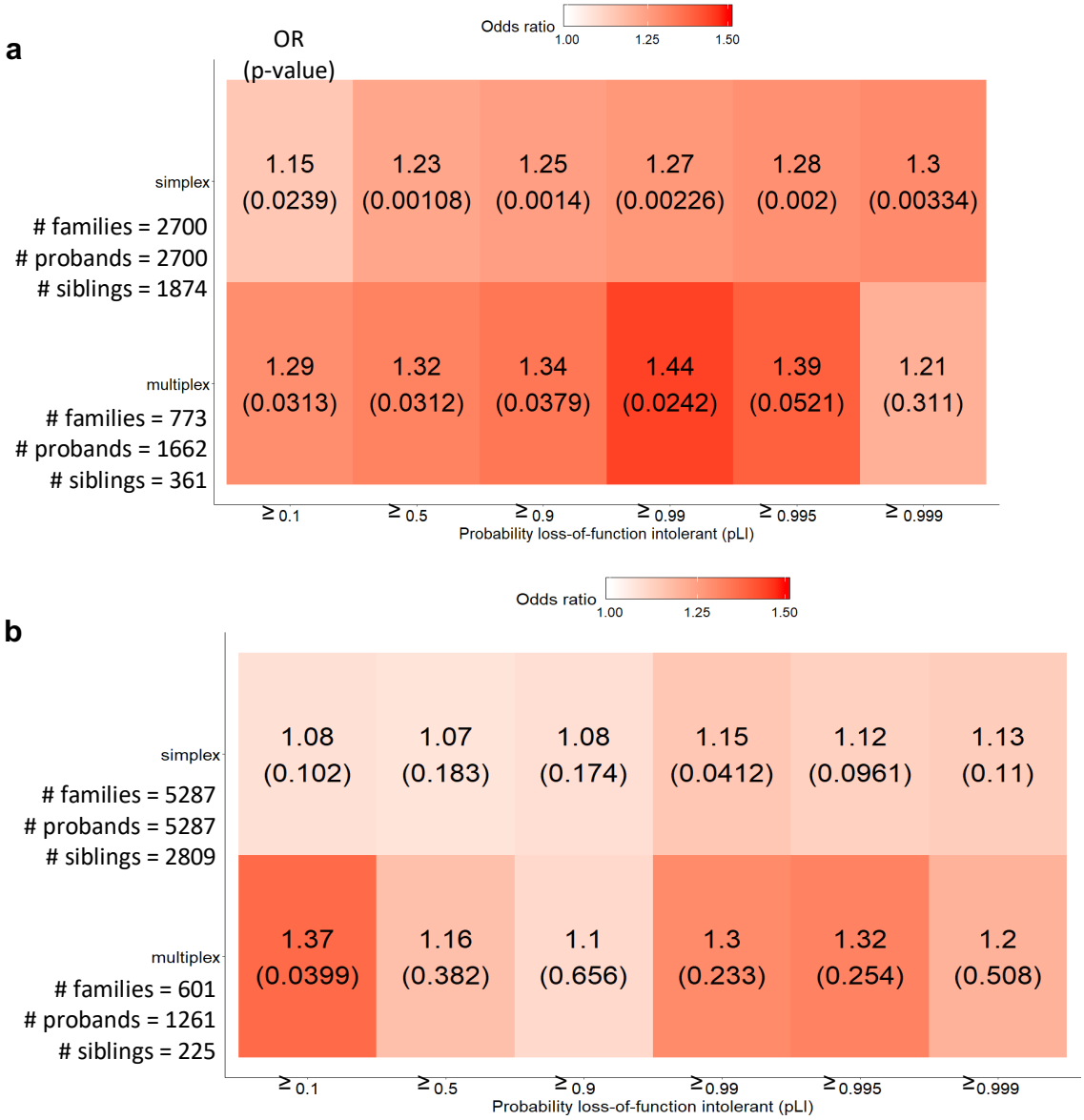

a
CCDG Genomes
1.25
1
0.75
Odds ratio
# Var. = 598,463
# Var. = 148,403
# Var. = 169,113
AC = 1
AC = 2
AC < 10
Allele count in parents
Variant type
LGD
MIS
SYN



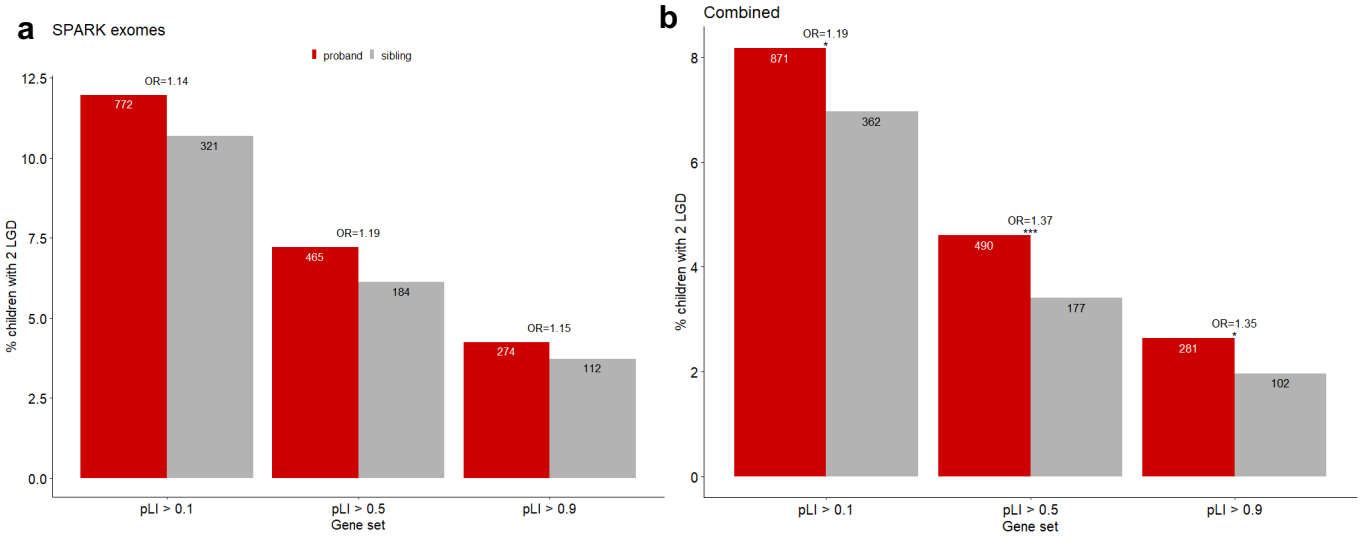
Supplementary Figure 15: LGD burden in probands isolated to private variants. Burden of LGD, MIS and SYN variants were compared in probands vs. siblings at three allele count bins using a Fisher's exact test for pLI ≥ 0.99 in the a) discovery (n = 4,201 affected and 2,191 unaffected children) and b) replication cohorts (n = 6,453 affected and 3,007 unaffected children). Error bars indicate the 95% confidence interval for the odds ratio estimates. Analyses exclude families with monozygotic twins.



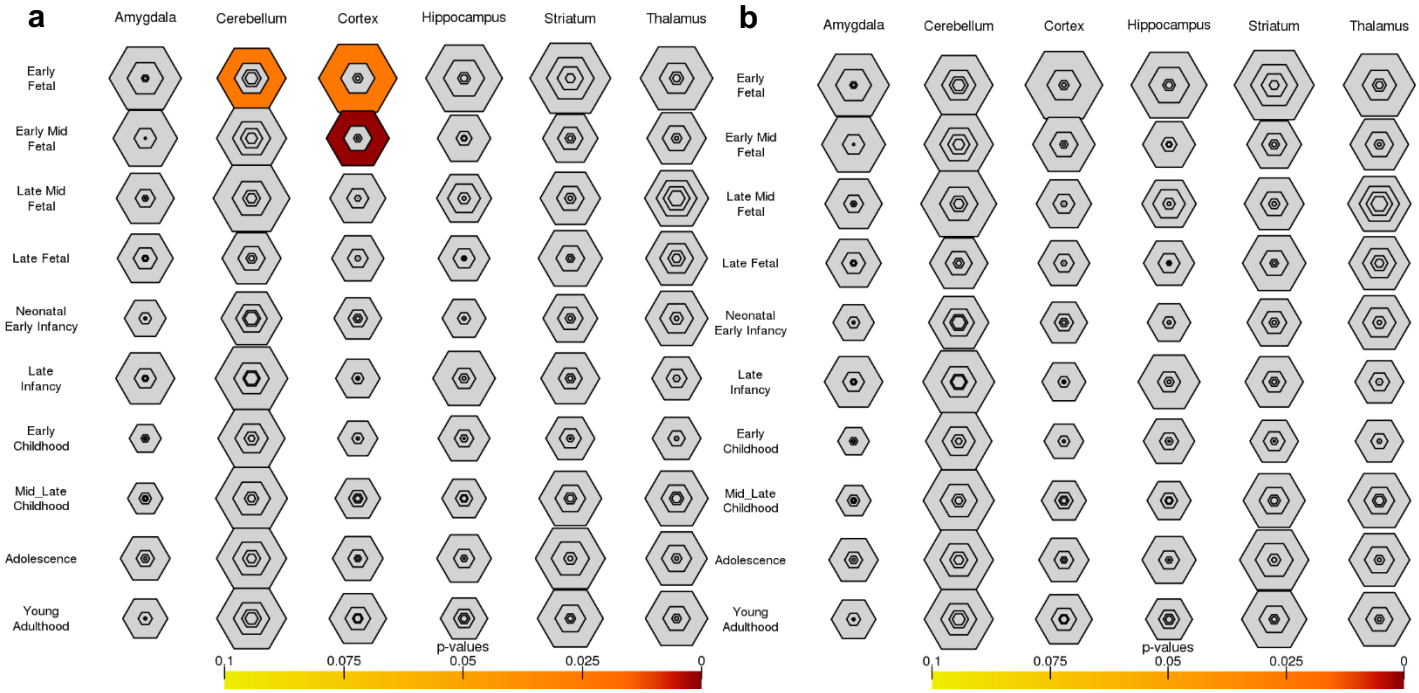
Supplementary Figure 16: Distribution of "gene-killing" private CNVs in SPARK probands and siblings by pLI score. Y-axis indicates the proportion of probands (bars facing up) and siblings (bars facing down) carrying either one CNV (gray) or two CNVs (black), broken down by duplications and deletions and by pLI score (cutoffs: pLI ≥ 0.1, ≥ 0.5, and ≥ 0.9) as shown on x-axis.



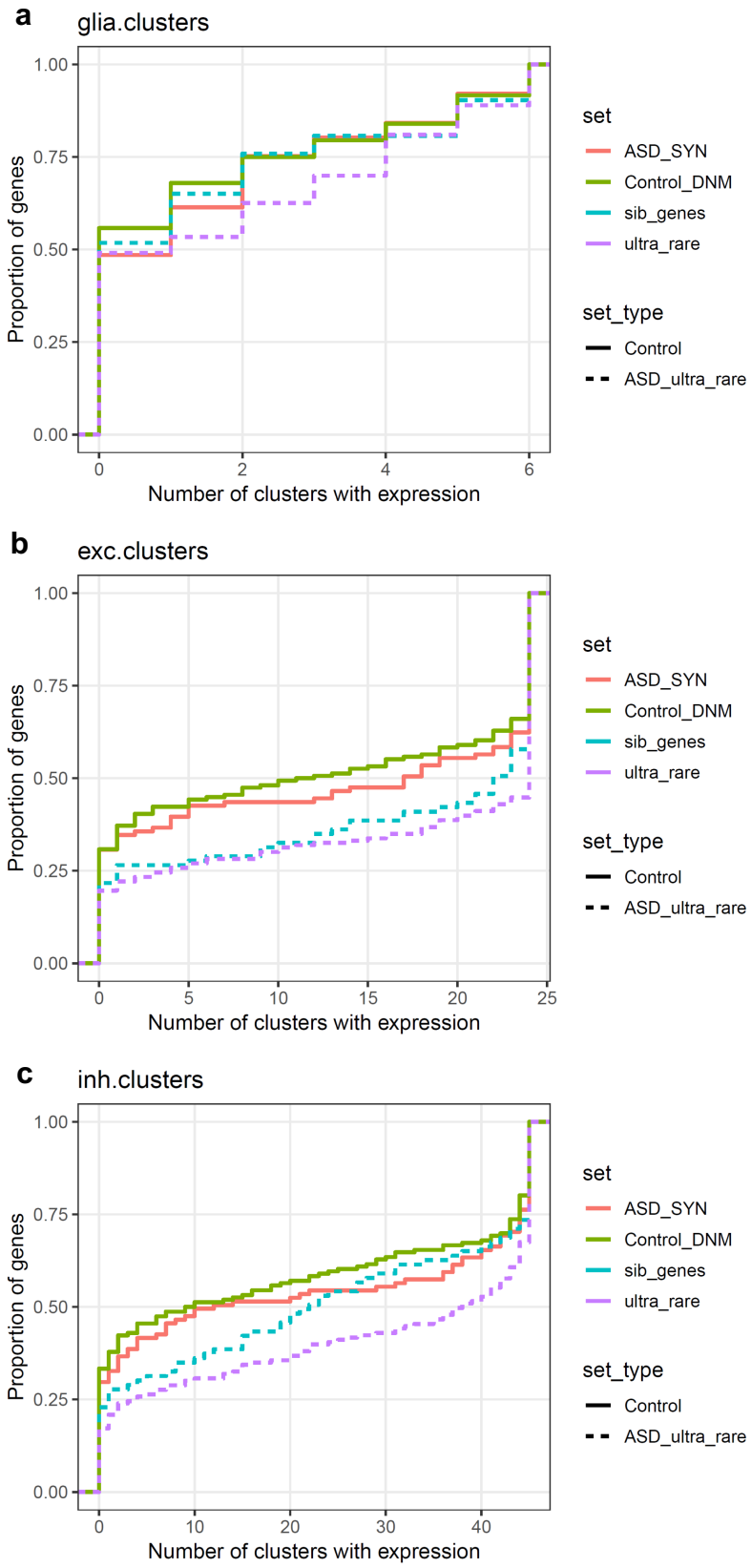
Supplementary Figure 17: Burden of private LGD variants in simplex and multiplex families. a) Our discovery cohort and b) our replication cohort. Excludes families with monozygotic twins. Reported p-values are Bonferroni corrected for 12 tests. Odds ratios calculated by two-sided Fisher's exact test.



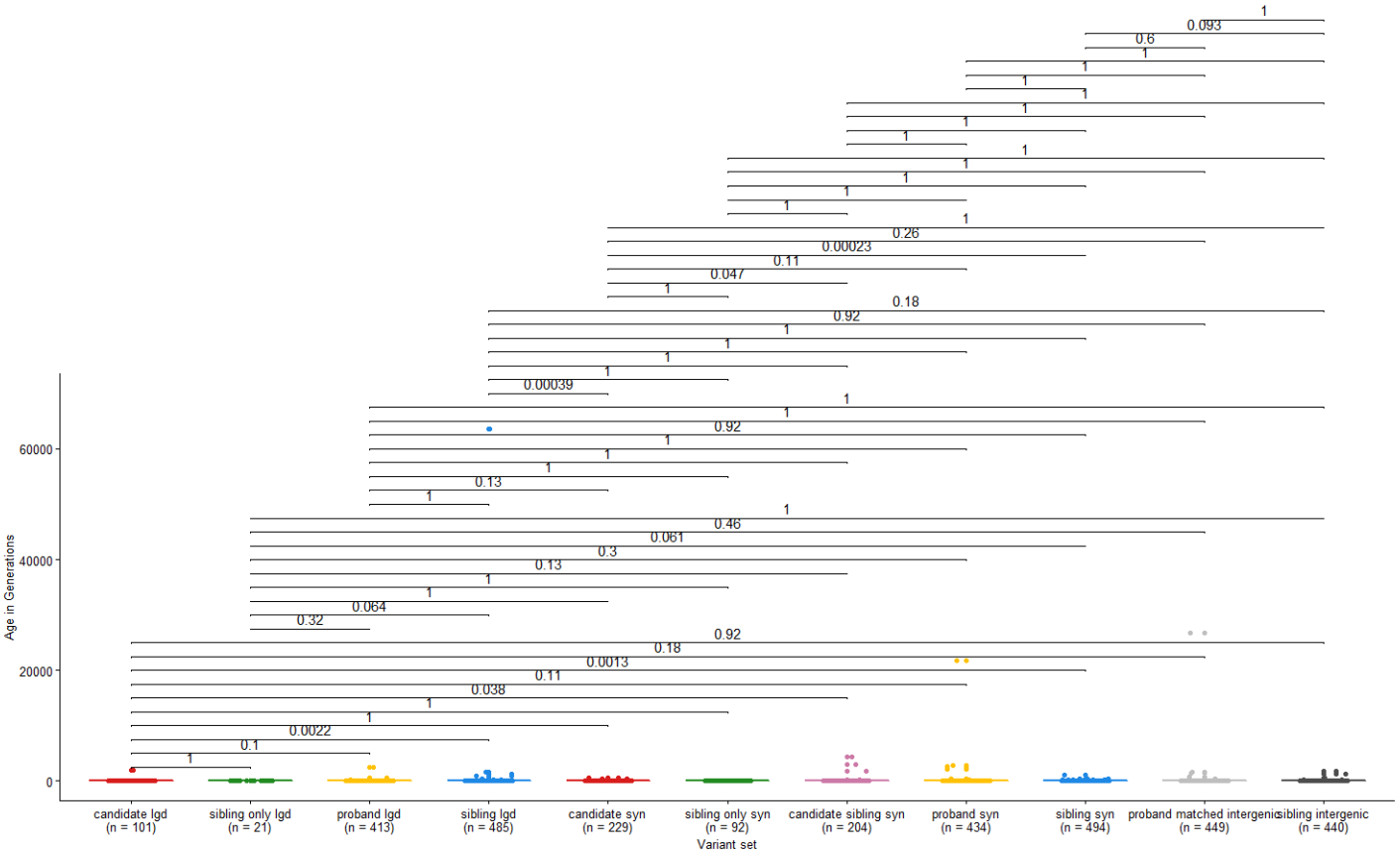
Supplementary Figure 18: Burden of two private LGD variants. a) Replication cohort (n = 6,539 probands, 3,034 siblings) and b) combined discovery and replication cohorts (n = 10,638 probands, 5,188 siblings). Probands are enriched for multiple private, transmitted LGD variants as compared to siblings across increasing thresholds of pLI. Families with monozygotic twins (n = 75 in discovery, n = 63 in replication, and n = 138 in combined) were removed from analysis. For the combined set, variants were restricted to regions with at least 20x average coverage in the exomes. Significance stars reflect Bonferroni-corrected p-values for three tests. Odds ratios calculated by two-sided Fisher's exact test.



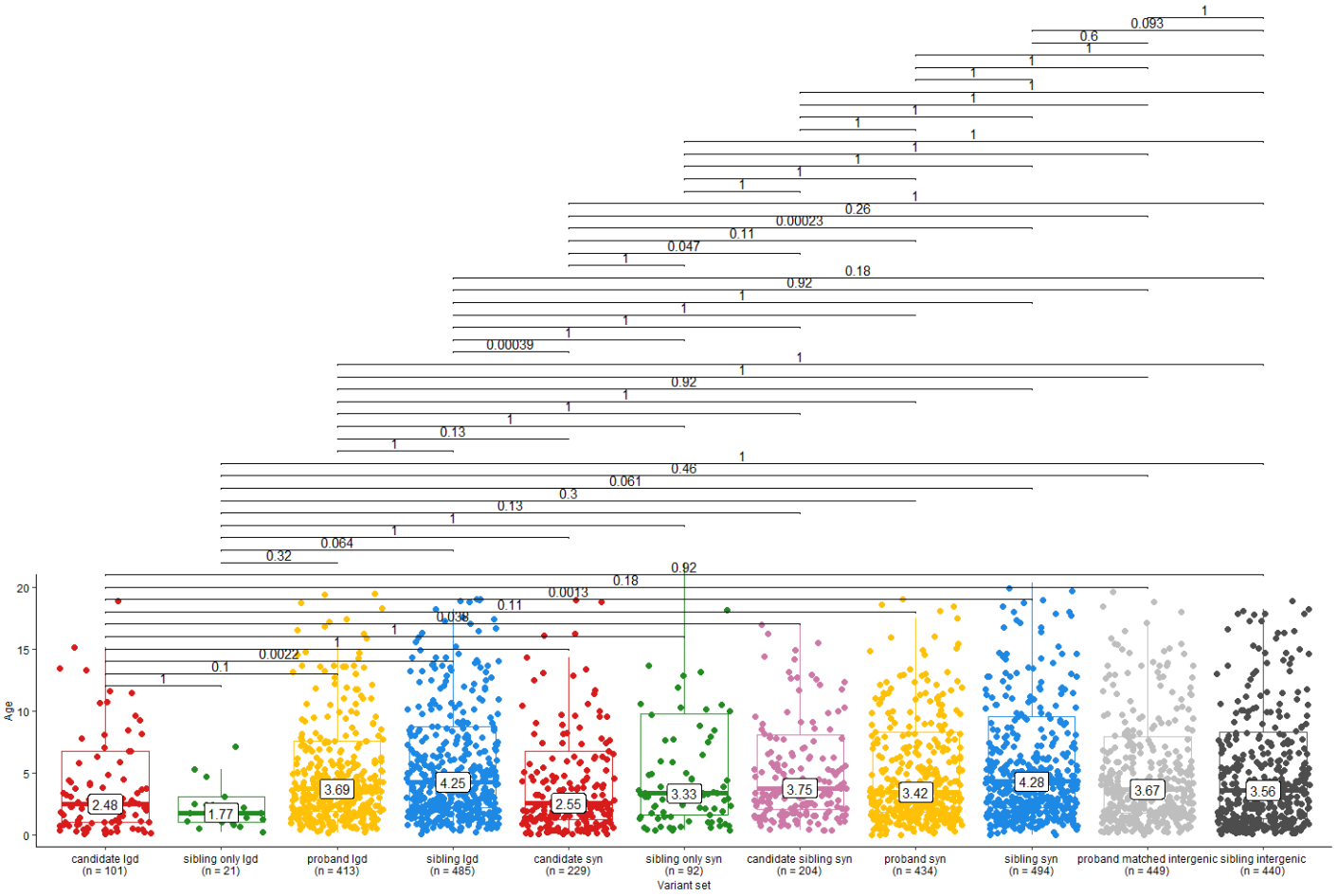
Supplementary Figure 19: Cell-type-specific expression analysis. a) 163 candidate genes with private LGD variants in probands only and b) 83 candidate genes with LGD variants in siblings only.



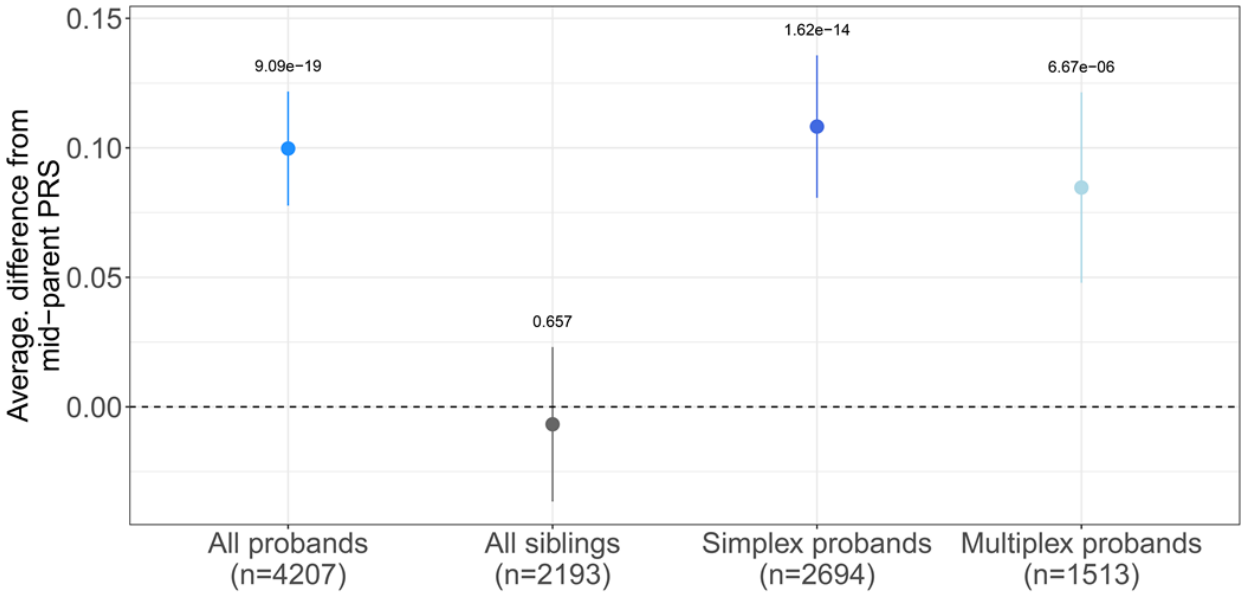
Supplementary Figure 20: Proportion of candidate genes expressed brain cells. a) Glia, b) excitatory, and c) inhibitory neurons. Excitatory and inhibitory neurons are enriched for expression of candidate genes as compared to control sets, but not in comparison to genes ascertained in siblings for the same criteria.



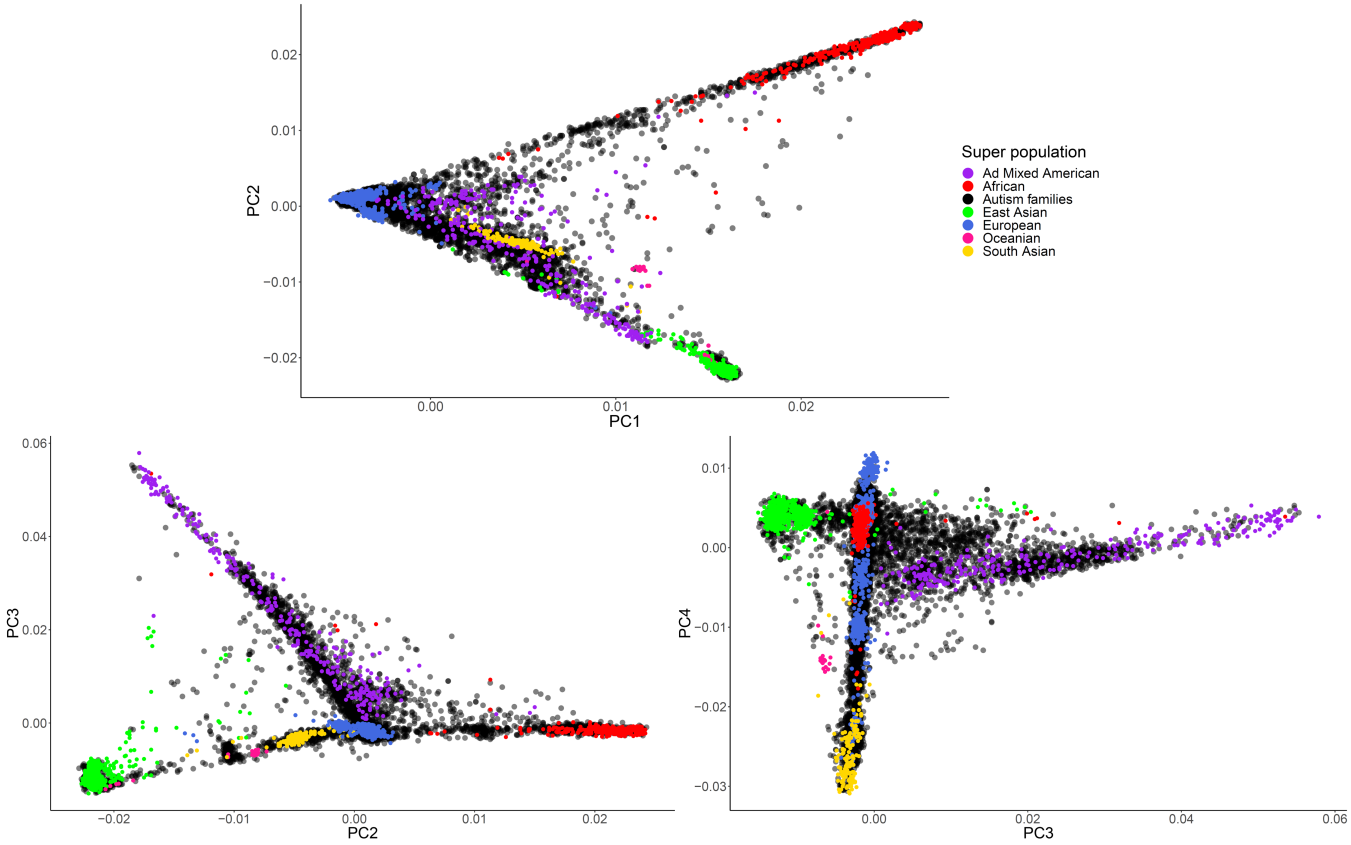
Supplementary Figure 21: Allele age estimates and comparisons for LGD and SYN variants in the EUR subset of our discovery cohort. P-values calculated by two-sided t-test and are Bonferroni-corrected for 36 tests.



Supplementary Figure 22: Allele age estimates for LGD and SYN variants within the EUR subset of our discovery cohort. Plot is zoomed in to 20 generations (data older than this is included in calculating represented statistics [e.g., boxplots, medians, p-values] but is not visualized). Medians and p-values represent the full dataset. P-values are calculated using two-sided t-test and are Bonferroni-corrected for 36 tests. Boxplot whiskers represent 1.5 times the upper and lower interquartile ranges. Upper and lower hinges correspond to the 25th and 75th percentiles, and the middle line represents the median. Mean values are noted on the plot.



Supplementary Figure 23: Polygenic transmission disequilibrium in autism. We calculated the polygenic risk score (PRS) using published GWAS summary statistics in our discovery cohort. The average PRS in both parents was compared to the PRS in probands and unaffected siblings. In addition to considering all children, we partitioned affected children by whether they were from a simplex or multiplex family. P-values were calculated using a two-sided t-test where the null hypothesis assumes the difference between the child's PRS and the average of both parent's PRS is centered at zero. Error lines indicate the 95% confidence interval around the average difference between the child and mid-parent PRS.



Supplementary Figure 24: Plots of the first four principle components from PCA comprised of two reference cohorts, our discovery cohort, and our replication cohort. Plots highlight the diversity of the cohorts.