Supplementary Information

Supplementary Tables

Supplementary Table 1: Families in this study (.xlsx)

All families included in the study are listed. For each family, we show: the genders of the probands and the unaffected sibling; the centers that sequenced the family; the collection the family is from (Methods); the SFARI database ids of the sequenced probands and siblings (for example, 13416.p1 and 13416.s1 are the proband and the sibling of family 13416); the race of the parents; the verbal and non-verbal IQ of the proband; and the ages of the parents at the birth dates of their children. Also included are the number of identified de novo LGDs, missense and synonymous mutations in the sequenced children.

Supplementary Table 2: All de novo events (.xlsx)

All de novo variants identified in the study are listed. For each observed de novo event, we list:

- family (*familyId:* SFARI family identifier) and child affected status and gender (*inChild*: pM – male proband; pF – female proband; sM – male sibling; and sF – female sibling);
- genomic location (*location*) and the type of the variant (*variant*, like substitutions (sub), insertions (ins) or deletions (del); see the supplement of Iossifov *et al*. *Neuron* **74**, 285–99 (2012) for details of the nomenclature);
- representation of the variant in the VCF standard format (*vcfVariant*);
- parental genome from which the mutation originates when we could so determine (*fromParent*);
- gene target (*effectGene*) and the effect type (*effectType*);
- genotype confidence scores from the common pipeline (strong, weak or "not called" when the common pipeline missed the variant) over the data from the three sequencing centers and the validation status: "valid" for the successfully validated variants and missing for the variants that have not been validated (*CSHL*, *YALE*, *UW*).

CSHL

Strength	Effect	verified	rejected	failed validation	un-tested
Substitutions					
strong	LGDs	50	1	8	0
strong	non LGDs	64	0	1	926
strong	non coding	7	0	0	200
weak	LGDs	2	0	0	0
weak	non LGDs	0	0	0	44
weak	non coding	0	0	0	12
Indels					
strong	LGDs	51	2	6	1
strong	non LGDs	13	0	1	0
strong	non coding	11	1	1	13
weak	LGDs	0	0	2	0
weak	non LGDs	2	0	0	0
weak	non coding	0	0	0	4

UW

Strength	Effect	verified	rejected	failed validation	un-tested
Substitutions					
strong	LGDs	33	0	0	0
strong	non LGDs	413	0	0	31
strong	non coding	2	0	0	104
weak	LGDs	0	2	0	1
weak	non LGDs	45	0	0	36
weak	non coding	0	0	0	21
Indels					
strong	LGDs	46	1	0	0
strong	non LGDs	5	0	1	6
strong	non coding	2	0	0	15
weak	LGDs	1	1	0	0
weak	non LGDs	0	0	0	1
weak	non coding	0	0	0	5

YALE

Strength	Effect	verified	rejected	failed validation	un-tested
Substitutions					
strong	LGDs	49	0	0	2
strong	non LGDs	542	2	11	279
strong	non coding	7	0	0	200
weak	LGDs	3	10	0	8
weak	non LGDs	25	2	1	59
weak	non coding	0	0	0	16
Indels					
strong	LGDs	47	3	4	6
strong	non LGDs	6	0	0	9
strong	non coding	1	0	2	33
weak	LGDs	9	2	0	1
weak	non LGDs	0	0	0	2
weak	non coding	0	0	0	6

Supplementary Table 3: Experimental validation in the 40X target

The results of the experimental validation of de novo variants in the well-covered 40X target are tabulated for the each of the sequencing centers. Variants are classified by their type (*substitutions* or *indels*) and by the confidence assigned by the common pipeline (*strong* or *weak*). Each attempted validation either verifies the variant (*verified*), rejects the variant (*rejected*), or fails (*failed validation*).

		Groups of X events			
Role	Group Distance	1	2	3	4
prb	gene	3,291	32	1	1
sib	gene	2,200	24	2	0
prb	100,000	3,328	34	1	1
sib	100,000	2,232	25	2	0
prb	10,000	3,332	32	1	1
sib	10,000	2,234	24	2	0
prb	1000	3,344	28	1	0
sib	1000	2,236	23	2	0
prb	100	3,346	27	1	0
sib	100	2,239	23	1	0
prb	10	3,350	25	1	0
sib	10	2,241	22	1	0
prb	5	3,366	17	1	0
sib	5	2,251	17	1	0
prb	1	3,376	12	1	0
sib	1	2,272	8	0	0

Supplementary Table 4: Multiple de novo events

The de novo events are grouped if found in the same child and appear ‘near’ to each other. The table lists the number of groups of various sizes in probands and in siblings under different definitions of near. In the first two lines, the events are grouped if they are in the same gene; in the following lines, the events are grouped if they are closer than the specified (*Group Distance*) genomic distance. The number of groups of given size is shown in columns *1*, *2*, *3* and *4*.

Supplementary Table 5: De novo substitution mutations (.xlsx)

Each de novo substitution is characterized by the parental nucleotide (*From*) and the novel nucleotide (column headers). The “Observed substitution counts” section shows the number of observed de novo substitution split by ‘from’ and ‘to’ nucleotides. The “Opportunity” sections shows the number of Mendelian genotypes in children for the given “From” allele at nearly fixed loci (at most one alternative allele allowed). The “Rates” section shows per-base-per-generation rate of substitutions.

Supplementary Table 6: Properties of mutational classes in SSC child types (.xlsx)

We show various properties of the mutation-child-types. We define mutation-child-type to refer to a set of events of a certain mutational type (e.g. missense or LGD) in children of a certain type (e.g. male affecteds or unaffected siblings). We grouped children into seven types: unaffected siblings, probands, affected females, affected males, affected males with low nvIQ (<90), affected girls plus low nvIQ boys, and affected males with high nvIQ. We partitioned mutations into three types: LGDs, missense, and synonymous.

For each mutation-child-type we show:

1. Details of 40X-based rate and ascertainment-differential computation, including: the number of children included in the 40X analysis (quads only); total length of 40X target; number of de novo mutation-child-type events in the 40X target; estimated rate and its standard deviation; estimated ascertainment differential, and its standard deviation and a p-value of the ascertainment differential.
2. Total number of children, total number of events, and the number of gene targets.
3. The number of recurrent genes and the number of recurrent events.
4. The expected number of recurrent genes under gene length-based null model and p value of the observer number of recurrent genes.
5. Details of the target size estimation including: proportion of events in the target and the approximate number of contributory events; target size estimated with a 95% CI.
6. Observed and expected overlaps with the eight gene classes defined in Table 1 and p-values computed by comparing the observed and expected number. Expectations are based on the gene length-based null model.
7. Observed and expected overlaps with all the other mutation-child-types. The expectation and the p-value are based on the gene length null model. De novo variants shared between siblings and de novo mutations occurring in the same gene in the same child are removed prior to testing for overlaps. We only tested for overlap between two mutation-child-types with distinct sets of children.

Supplementary Table 7: Numbers of mutations by gene and type (.xlsx)

All genes included within SeqCap EZ Human Exome Library v2.0 (Roche NimbleGen) are listed within the table. Total coding region length in nucleotides (predicted by RefSeq) is displayed in column “B,” whereas the length represented within the actual capture reagent is shown in column “C.” Gene sets (columns D–K) are as defined in Table 1; the presence of a “1” in these columns indicates that the gene falls within the particular class. Columns L–AC tabulate de novo (dnv) events of specific classes (LGD, missense or synonymous) that fall within the specified child cohorts: all probands (prb); male probands (prbM); male probands with nvIQ<90 (prbML); male probands with nvIQ≥90 (prbMH); all female probands (prbF); and unaffected siblings (sib).

					Prb. vs.
	Proband	Sibling	Proband		Sib. p-
Gene Group	only	only	& Sibling	Neither	Value
FMRP-associated	40	41	11	108	1.00
targets of de novo LGDs in probands	22	23	7	71	1.00
targets of de novo missense in probands	130	115	36	377	0.37
all genes	373	348	115	1,025	0.37

Supplementary Table 8: Compound non-synonymous hits in targets from quad families

The Supplementary Table hows the numbers of genes hit by two rare nonsynonymous variants, one inherited from the mother and one inherited from the father (compound hit). Since each parent carries an affected allele, the compound hit can be in proband only, in the unaffected sibling only, in both proband and sibling, or in neither of the children. We report either all compound hits or only those in three sets of genes: FMRP-associated genes, genes affected by de novo LGDs in probands, and genes affected by de novo missense variants in probands. The p-values are calculated under the null model assuming that the probability of a compound hit in a proband only is the same as the probability of a compound hit in a sibling only.

Supplementary Table 9: Gene targets of recurrent mutation in coding regions (.xlsx)

Genes with recurrent mutations (defined here as two independent missense and/or severe de novo events) are ranked based on a per-gene model (O'Roak *et al.*, Science **338**, 1619–1622, 2012). The term "severe" as used here refers to LGDs plus all coding mutations that are neither synonymous nor missense, such as frame-preserving small indels and mutations that remove start or stop codons. Genes with recurrent hits that were rejected by this model are indicated by asterisks (*).

Supplementary Table 10: UW sequencing protocols (.xlsx)

SSC families processed by the UW group are listed by family ID and the version of capture and sequencing protocols used. If previously published, the corresponding study is referenced: O'Roak, B. J. *et al.*, *Nat Genet* **43**, 585–589 (2011), or O'Roak, B. J. *et al.*, *Nature* **485**, 246–250(2012). Families new to this study are indicated accordingly.

			CSHL		
Strength	**Effect**	**verified**	**rejected**	**failed validation**	**un-tested**
			Substitutions		
strong	LGDs	75	1	14	0
strong	non LGDs	101	0	1	1561
strong	non coding	11	0	0	394
weak	LGDs	9	1	5	0
weak	non LGDs	12	0	1	269
weak	non coding	2	0	1	92
			Indels		
strong	LGDs	100	9	15	1
strong	non LGDs	27	5	4	0
strong	non coding	16	5	2	43
weak	LGDs	2	22	17	0
weak	non LGDs	4	5	4	0
weak	non coding	0	3	0	38

			UW		
Strength	**Effect**	**verified**	**rejected**	**failed validation**	**un-tested**
			Substitutions		
strong	LGDs	53	0	0	0
strong	non LGDs	625	0	0	86
strong	non coding	3	0	0	202
weak	LGDs	6	2	0	1
weak	non LGDs	91	0	0	100
weak	non coding	0	0	0	82
			Indels		
strong	LGDs	68	6	0	6
strong	non LGDs	7	0	1	14
strong	non coding	3	0	0	37
weak	LGDs	2	4	0	30
weak	non LGDs	0	0	0	8
weak	non coding	0	0	0	35

			YALE		
Strength	**Effect**	**verified**	**rejected**	**failed validation**	**un-tested**
			Substitutions		
strong	LGDs	65	0	0	5
strong	non LGDs	700	3	16	395
strong	non coding	7	0	0	288
weak	LGDs	4	12	0	9
weak	non LGDs	63	3	1	169
weak	non coding	1	0	0	90
			Indels		
strong	LGDs	60	20	5	9
strong	non LGDs	7	0	0	15
strong	non coding	2	0	3	57
weak	LGDs	9	6	0	120
weak	non LGDs	0	0	0	7
weak	non coding	0	0	0	79

Supplementary Table 11: Validation summary by center (.docx)

The results of the experimental validation of de novo variants in the well-covered are tabulated for the each of the sequencing centers. Variants are classified by their type (*Substitutions* or *Indels*) and by the confidence assigned by the common pipeline (*strong* or *weak*). Each attempted validation either verifies the variant (*verified*), rejects the variant (*rejected*), or fails (*failed validation*).

Supplementary Table 12: Data agreement (.docx)

We intentionally had a number of families sequenced independently in more than one sequencing center (Extended Data Fig. 1). For the three centers, we show the agreement of the results of the common pipeline of the families sequenced independently at the two centers. Variants are classified by their type (*Substitutions* of *Indels*) and by the confidence assigned by the common pipeline (*strong* or *weak*). "Not called" labels are used for variants identified in one dataset but not in the other.

Description of genes	Median length	p-val
Selected based on length	2,175	
Selected uniformly	1,335	
Hit by LGDs in probands	2,670	0.001
Hit by LGDs in siblings	2,013	0.940
Hit by missense in proband	2,424	0.001
Hit by missense in sibling	2,274	0.432
Hit by synonymous (sib + prb)	2,206	0.822

Supplementary Table 13: Median gene lengths

Median gene lengths of simulated and genes sets and genes hit by different types of mutations. The first line shows the median gene lengths if genes are selected with probability proportional to their length. The second line shows the median gene length if genes are randomly selected independent of their length. The lines following show the median gene lengths of genes hit by de novo LGDs in probands, LGDs in siblings, missense in probands, missense in siblings, and synonymous mutations in probands or siblings. In addition, we show the empirical p-values of the observed median lengths under the gene-length based null model. The p-values were computed by selecting the same number of random genes (proportional to their length) and recording the number of times (out of 10,000 attempts) that the median lengths of selected genes were larger than those observed.