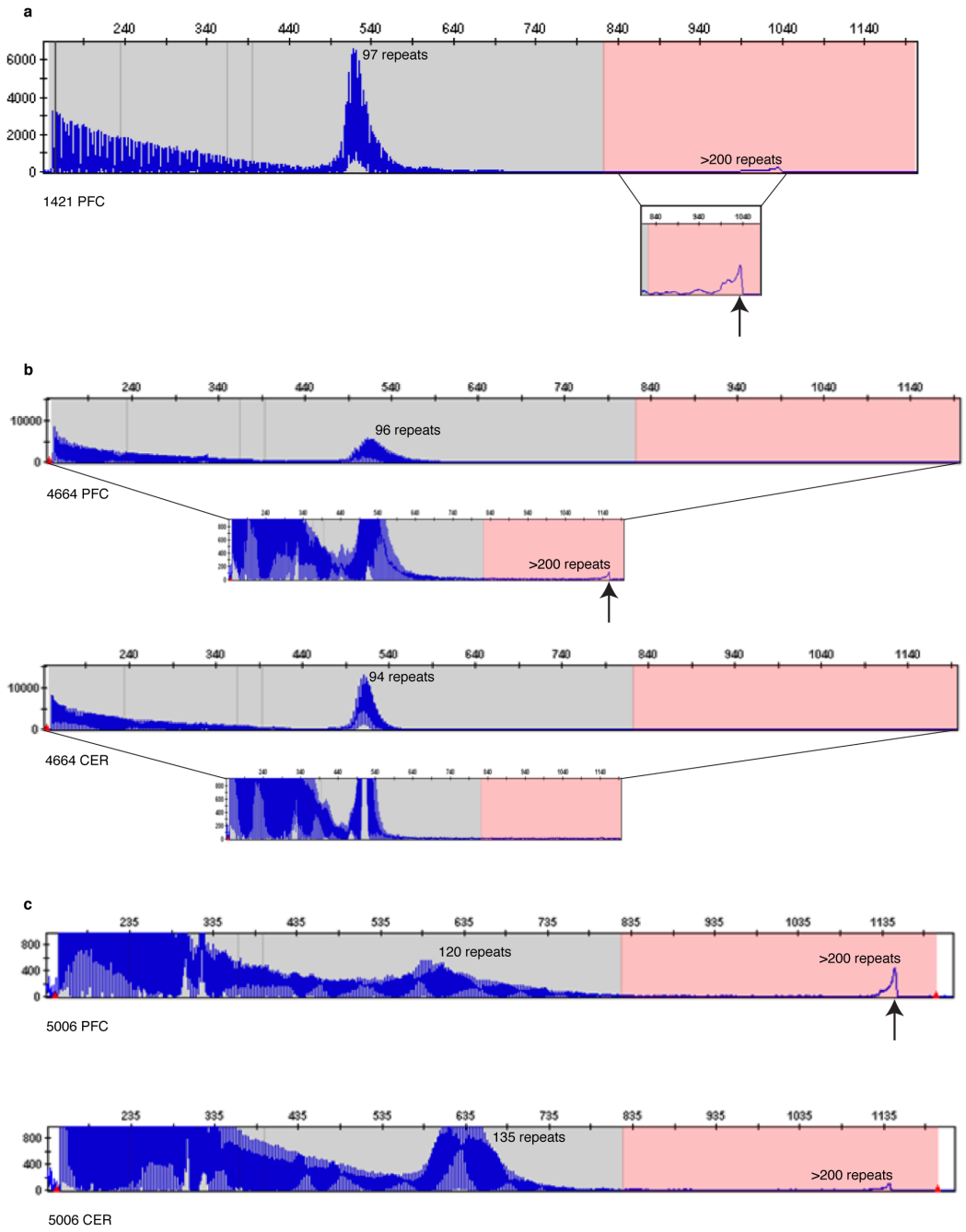
Supplemental Data:

Figure S1, Related to Table 1: Detection of somatic Fragile X CGG expansions in Cases 1421, 4664, and 5006



a. Case 1421 has both a premutation and full mutation peak in the PFC. b. Case 4664 has a premutation and a full mutation peak in the PFC, but only a premutation peak in the CER. Note that the test is not quantitative, since the smaller premutation peaks amplify more efficiently in PCR, so the relative proportion of premutation and full mutation cannot be determined by this method. c. Case 5006 has both a premutation and full mutation peak in the PFC and CER. Measured repeat number is noted above peaks. CER: cerebellum, PFC: prefrontal cortex.

Table S3, Related to Figure 1: Comparison of variants identified in ASD cases and neurotypical controls

	Total number of mutations		Mutations per subject		Number of subjects with mutation			
	Cases	Controls	Cases, N=55	Controls, N=50	Cases (%)	Controls (%)	p Value[a]	Odds Ratio[c] (95% CI)
Silent	13	13	0.24	0.26	10 (18%)	12 (24%)	0.483	0.70 (0.27 - 1.81)
Protein-altering	34	15	0.62	0.3	26 (47%)	12 (24%)	0.015	2.84 (1.23 - 6.56)
Deleterious	20	5	0.36	0.10	16 (29%)	5 (10%)	0.016	3.69 (1.24 - 11.00)
LOF	6	0[b]	0.11	0	6 (11%)	0 (0%)	0.028	N/A

a: p Value was calculated using a two-sided Fisher's exact test
b: One control harbors a frameshift mutation that is not an exonic frameshift in all isoforms and is in an alternatively spliced exon considered noncritical for protein function. This mutation was not included in the LOF statistics.
c: The Odds Ratio is the ratio of cases with a type of mutation to cases without that type of mutation divided by the corresponding ratio in controls
CI: Confidence Interval

Table S4, Related to Figure 1: NGS and subcloning details for potential somatic variants that did not validate

Case	Gene	Region	Chr	Position	Ref	Alt	Total Reads	Alt Reads	NGS AAF(%)	Total Colonies	Alt Colonies	TOPO AAF (%)	p Value[a] (NGS vs TOPO)
1474	*ADNP*	PFC	20	49508945	C	A	91	12	13.2	34	0	0	0.0353
4899	*AFF2*	CER	X	147743569	C	A	60	10	16.7	40	0	0	0.0054
AN00764	*AGTR2*	CER	X	115303498	G	T	54	5	9.3	60	0	0	0.0215
4231	*AP1S2*	PFC	X	15845352	G	T	98	6	6.1	72	0	0	0.0394
5144	*AUTS2*	PFC	7	69064653	C	A	51	3	5.9	93	0	0	0.0428
4334	*CACNA1H*	PFC	16	1252038	C	A	66	8	12.1	39	0	0	0.0245
5144	*CHD8*	PFC	14	21870426	C	A	194	10	5.2	89	0	0	0.0338
5470	*CHD8*	PFC	14	21864011	G	T	158	20	12.7	27	0	0	0.0486
AN09714	*FMR1*	BA19	X	147014095	C	A	45	6	13.3	46	0	0	0.0122
818	*FMR1*	PFC	X	147026517	C	A	86	10	11.6	40	0	0	0.0299
5470	*GABRB3*	PFC	15	26793187	G	T	59	10	16.9	39	0	0	0.0054
1712	*GRIA3*	PFC	X	122599588	C	A	46	10	21.7	24	0	0	0.0124
4672	*GRIK2*	PFC	6	102266296	G	T	103	12	11.7	39	0	0	0.0366
5470	*IL1RAPL1*	PFC	X	29417297	C	A	79	16	20.3	24	0	0	0.0201
5452	*MDM2*	PFC	12	69203068	C	A	74	12	16.2	30	0	0	0.0172
UK25363	*NLGN3*	PFC	X	70375140	G	T	75	10	13.3	38	0	0	0.0157
1499	*PTCHD1*	PFC	X	23411323	C	A	80	11	13.8	34	0	0	0.0321
UK20244	*SBF1*	PFC	22	50906799	A	G	723	68	9.4	95	0	0	0.0002
5408	*SCN1A*	PFC	2	166908288	C	A	87	10	11.5	39	0	0	0.0305
5176	*SCN2A*	CER	2	166245181	C	A	46	6	13.0	80	0	0	0.0019
5470	*SCN2A*	PFC	2	166153564	G	A	63	12	19.0	26	0	0	0.0157
5027	*SCN2A*	CER	2	166231195	G	T	92	12	13.0	33	0	0	0.0350
M3746M	*SETD2*	PFC	3	47161989	G	T	65	8	12.3	56	0	0	0.0072
5027	*SLC9A6*	CER	X	135080718	G	T	77	16	20.8	23	0	0	0.0195
967	*SYN1*	PFC	X	47464767	G	A	220	10	4.5	91	0	0	0.0376
5297	*TSC2*	CER	16	2135991	C	T	526	15	2.9	154	0	0	0.0290

a: p value comparing NGS read counts to TOPO counts calculated using a two-tailed Fisher's exact test
BA19: Brodmann Area 19, CER: cerebellum, PFC: prefrontal cortex

Table S5, Related to Figure 1: NGS and subcloning details for validated somatic variants

Case	Gene	Region	Total Reads	Alt Reads	NGS AAF (%)	p Value[a] (50%)	Total Colonies	Alt Colonies	TOPO AAF (%)	p Value[b] (50%)
5006	*CACNA1C*	PFC	160	67	41.88	1.78E-01	94	31	32.98	2.61E-02
		CER	157	68	43.31	2.59E-01	212	88	41.75	4.87E-02[c]
5378	*CACNA1H*	PFC	1776	89	5.02	2.13E-220	95	2	2.11	1.20E-15
5278	*SCN1A*	PFC	355	115	32.39	2.16E-06	775	367	47.35	1.61E-01[c]
		CER	623	234	37.56	1.08E-05	781	357	45.71	4.98E-02[c]
UK20244	*SETD2*	PFC	171	28	16.37	3.05E-11	132	35	26.52	1.34E-04
967	*SLC6A4*	CER	333	54	16.22	7.21E-21	24	1	4.17	6.99E-04

a: p value comparing NGS read counts to expected reads counts for a heterozygous mutation calculated using a two-tailed Fisher's exact test
b: p value comparing TOPO counts to expected counts for a heterozygous mutation calculated using a two-tailed Fisher's exact test (except where noted)
c: p value calculated using a one-tailed Fisher's exact test
CER: cerebellum, PFC: prefrontal cortex

Table S6, Related to Figure 2: Distribution of somatic variants identified

Case	Diagnosis	Gene	Mut	Chr	Pos	Ref	Alt	Brain AAF	Brain Mutant Cell Frequency	Non-brain AAF
5278	Autism	*SCN1A*	Sp	2	166911147	C	T	PFC: 32.4-47.4%	PFC: 65-95%	Liver: 46.7%
								CER: 37.6-45.7%	CER: 75-91%	Serum: 46.3%
								PAR: 42.6%	PAR: 85%	
								MED: 44.3%	MED: 89%	
UK20244	Autism	*SETD2*	Ms	3	47144882	G	C	PFC: 16.4-26.5 %	PFC: 33-53%	N/A
								CER: 0%	CER: 0%	
5006	Fragile X, premutation	*CACNA1C*	St	12	2162730	T	C	PFC: 33-42%	PFC: 66-84%	
								CER: 42-43%	CER: 84-86%	
5378	ASD/Autism Sibling, Social Anxiety Disorder	*CACNA1H*	Syn[a]	16	1268542	G	A	PFC: 2-5%	PFC: 4-10%	
								CER: 0%[b]	CER: 0%	
967	ASD/Autism, suspected	*SLC6A4*	In	17	28546044	G	A	PFC: 0%	PFC: 0%	Dura: 0%
								CER: 4-16%	CER: 8-32%	

a: rs60526088
b: 5378 CER had 2/1793 reads with A at this position (0.1%). Given that the expected base miscall rate is 0.1% (Shirley et al., 2013), that 5378 PFC has 2/1776 reads with C at this position (0.1%) and an additional 2/1776 reads with T at this position (0.1%), and that validation did not identify the alternate base in 5378 CER, we believe it is most likely a sequencing error.
CER: cerebellum, In: Intronic, MED: medulla, Ms: Missense, PAR: parietal cortex, PFC: prefrontal cortex, Sp: Splicing, St: Start Lost, Syn: Synonymous

Supplemental Experimental Procedures

Gene selection and panel design

The ASD candidate gene panel was designed to balance including a sufficient number of genes with achieving the depth necessary to detect low-frequency somatic mutations. Given the hundreds of candidate genes reported, we included genes with strongest evidence of association with ASD. We included genes from three sources, focusing on genes whose disease mechanisms involve dominant or X-linked modes of inheritance, as these are associated with higher *de novo* mutation rates. First, a recent study performed targeted sequencing on DNA from 2,446 individuals with ASD to identify recurrently mutated genes; we included all 44 genes used in their targeted panel (O'Roak et al., 2012). Second, we included X-linked genes associated with ASD as reviewed by Betancur (Betancur, 2011). Third, we included dominant and X-linked genes with strong evidence of association with ASD curated by the Simons Foundation Autism Research Initiative (SFARI) database (Basu et al., 2009). Overall, the panel comprises 78 genes, generating a target region of 279kb that includes all exons, exon-intron boundaries, and 10bp of flanking sequence for each gene. The design, created using Suredesign (Agilent), is predicted to cover 99.7% of the target region.

Genes included in targeted panel:

ACSL4	EN2	PTCHD1
ADCY5	FGD1	PTEN
ADNP	FMR1	RAB39B
AFF2	FOXP2	RAI1
AGTR2	FTSJ1	RBFOX1
AP1S2	GABRB3	SBF1
ARHGEF6	GRIA3	SCN1A
ARID1B	GRIK2	SCN2A
ARX	GRIN2A	SEMA5A
ASTN2	GRIN2B	SETD2
ATP10A	HOXA1	SGSM3
ATRX	IL1RAPL1	SHANK3
AUTS2	IQSEC2	SLC6A4
CACNA1C	KDM5C	SLC6A8
CACNA1H	LAMB1	SLC9A6
CASK	MDM2	SYN1
CDKL5	MECP2	SYNGAP1
CHD8	MET	TBL1XR1
CNOT4	NLGN1	TBR1
CNTN4	NLGN3	TSC1
CNTNAP2	NLGN4X	TSC2
CTNNB1	NTNG1	UBE3A
DISC1	OXTR	UBE3C
DLX2	PON1	UPF3B
DPP6	PQBP1	ZNF674
DYRK1A	PSEN1	ZNF81

DNA library preparation and next generation sequencing

Paired-end, barcoded libraries were prepared per the manufacturer's protocol with 225ng of DNA from each sample using a custom Haloplex Target Enrichment Kit (Agilent). Paired-end sequencing (250bp x 2 or 150bp x 2) was performed on MiSeq sequencers (Illumina) at the DNA Diagnostic Laboratory (now

Claritas Genomics) at Boston Children's Hospital or the Harvard BioPolymers Facility. Sequencing was performed in batches to achieve a higher read depth for each sample to optimize detection of low-frequency somatic variants.

DNA sequencing data analysis

Raw read data was processed and mapped using BWA (Li and Durbin, 2009) and SNV and insertion and/or deletion (indel) calling was performed using SAMtools (Li et al., 2009) and SNPPET (Agilent), using the Surecall software (Agilent). For cases where two brain regions were sequenced, MuTect (Cibulskis et al., 2013) was used to compare the regions for mutations that were present in one region but not the other, and vice versa. Variants were quality filtered to exclude false positives according to standard thresholds (quality<30, coverage<10X and clustered variants (window size of 10)). For somatic variants, after initial validation experiments resulted in many false positives, filtering was adjusted to alternate allele frequency <40%, coverage ≥60X and alternate allele read depth ≥ 10X. The first and last five base pairs of every read were removed from read count calculations due to bias resulting from the restriction enzyme step in library preparation. Data from the Exome Sequencing Project (http://evs.gs.washington.edy/EVS/), dbSNP 137 and 142 (Sherry et al., 2001), and 1000 Genomes Project (Abecasis et al., 2012) were used to assess variant frequencies in control populations. We excluded variants present in dbSNP or with MAF≥1% in EVS or 1000 Genomes. Previously reported mutations were identified using the Human Gene Mutation Database (Stenson et al., 2014). We used Provean (Choi et al., 2012), Sift (Ng and Henikoff, 2003), Polyphen 2 (Adzhubei et al., 2010), and CADD (Kircher et al., 2014) to assess for deleteriousness. We considered a variant to be loss-of-function if it was a nonsense, frameshift, or splicing variant, and we considered a variant to be deleterious if it was predicted as such by at least three of Provean (deleterious), Sift (damaging), Polyphen-2 (probably damaging or possibly damaging), and CADD (phred score ≥20).

RNA extraction and quality assessment

Total RNA was extracted using mirVana kit (Ambion) with some modifications to the manufacturer's protocol. Each tissue sample was pulverized with liquid nitrogen in a prechilled mortar and pestle and transferred to a chilled safe-lock microcentrifuge tube (Eppendorf). Per tissue mass, equal mass of chilled stainless steel beads (Next Advance, catalog # SSB14B) along with one volume of lysis/binding buffer were added. Tissue was homogenized for 1 min in Bullet Blender (Next Advance) and incubated at 37°C for 1 min. Another nine volumes of the lysis/binding buffer were added, homogenized for 1 min, and incubated at 37°C for 2 min. One-tenth volume of miRNA Homogenate Additive was added and extraction was carried out according to the manufacturer's protocol. RNA was treated with DNase using TURBO DNA-free Kit (Ambion/ Life Technologies) and RNA integrity was measured using Agilent 2200 TapeStation System.

RNA library preparation and next generation sequencing

Barcoded libraries for RNA-seq were prepared with 5ng of RNA using TruSeq Stranded Total RNA with Ribo-Zero Gold kit (Illumina) per manufacturer's protocol. Paired-end sequencing (76bp x 2) was performed on HiSeq 2000 sequencers (Illumina) at Yale Center for Genome Analysis.

RNA-seq data analysis

The sequenced reads were processed and filtered for quality prior to alignment. First, the first base from both ends was trimmed to remove potential primer contamination. Filtered reads were aligned to hg19 (GRCh37) genome using Tophat (version 2.0.12) (Trapnell et al., 2009) Reads that were not uniquely mapped were excluded from further analysis. Gene expression levels were measured in RPKM (reads per kilobase of exon model per million mapped reads (Mortazavi et al., 2008)) using HTSeq (Anders et al., 2015) and SAMtools (Li et al., 2009). Briefly, the BAM format alignment was first converted into SAM format alignment by using the "view" function in SAMtools. Then, the "htseq-count" function in HTSeq was used to count reads mapped to genes annotated in GENCODE (version19) (http://www.gencodegenes.org/) (Harrow et al., 2012). We ran "htseq-count" function twice with different –t parameters, *i.e.*, exon and gene, so as to infer reads mapped to exon and gene, and reads different between them were mapped to introns. For each gene, a composite model of the gene (union of all exons across all transcripts of gene) was created from GENCODE (version 19) annotation, all reads overlapping this model were counted and

normalized per million mapped nucleotides and the length of the annotation item per kb to get RPKM values.

To identify differences in gene expression, we compared RNA-seq data from ASD specimens with RNA-seq data from matched neurotypical postmortem human brain specimens, which we generated as part of the BrainSpan consortium (www.brainspan.org). Each ASD sample was compared with the same region from two control samples matched closely for age and sex. Due to differences in sample and library preparations between BrainSpan controls (polyA enriched RNA was single end sequenced) and ASD samples (total RNA depleted of ribosomal RNA was paired end sequenced), several processing steps were carried out. Genes from sex chromosomes and non-coding genes were also excluded from differential expression analysis to avoid sex-bias and because BrainSpan controls do not have complete coverage of non-coding genes due to poly A enriched library preparation. The RPKM values of autosomal protein-coding genes in 6 ASD samples and 12 matched controls were pooled together to construct expression matrix, from which genes with Q3 (upper quartile) RPKM values less than 1 were filtered out, leaving 19431 protein-coding genes. Then using "normalizeBetweenArrays" function in "limma" R Bioconductor (Smyth, 2005), the log2 transformed gene (RPKM+1e-5) values were quantile normalized. Combat (Johnson et al., 2007) was then used to correct the batch effect between samples, as well as using region and ethnicity as covariance factors. The differential gene expression analyses were based on these corrected and normalized gene expression values.

Genes with high fold differences in expression were identified by comparing ASD samples with each control sample. To be considered to have expression differences, a gene was required to have RPKM value greater than 1 in at least one sample and absolute fold-change greater than 2. To get the most confident list of genes, only those genes that were detected as potentially differentially expressed with both control samples and which had fold changes in the same direction were considered.

Supplemental References

Abecasis, G.R., Auton, A., Brooks, L.D., DePristo, M.A., Durbin, R.M., Handsaker, R.E., Kang, H.M., Marth, G.T., and McVean, G.A. (2012). An integrated map of genetic variation from 1,092 human genomes. Nature 491, 56-65.

Adzhubei, I.A., Schmidt, S., Peshkin, L., Ramensky, V.E., Gerasimova, A., Bork, P., Kondrashov, A.S., and Sunyaev, S.R. (2010). A method and server for predicting damaging missense mutations. Nature methods 7, 248-249.

Anders, S., Pyl, P.T., and Huber, W. (2015). HTSeq--a Python framework to work with high-throughput sequencing data. Bioinformatics 31, 166-169.

Choi, Y., Sims, G.E., Murphy, S., Miller, J.R., and Chan, A.P. (2012). Predicting the functional effect of amino acid substitutions and indels. PloS one 7, e46688.

Cibulskis, K., Lawrence, M.S., Carter, S.L., Sivachenko, A., Jaffe, D., Sougnez, C., Gabriel, S., Meyerson, M., Lander, E.S., and Getz, G. (2013). Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples. Nature biotechnology 31, 213-219.

Harrow, J., Frankish, A., Gonzalez, J.M., Tapanari, E., Diekhans, M., Kokocinski, F., Aken, B.L., Barrell, D., Zadissa, A., Searle, S.*, et al.* (2012). GENCODE: the reference human genome annotation for The ENCODE Project. Genome research 22, 1760-1774.

Johnson, W.E., Li, C., and Rabinovic, A. (2007). Adjusting batch effects in microarray expression data using empirical Bayes methods. Biostatistics 8, 118-127.

Kircher, M., Witten, D.M., Jain, P., O'Roak, B.J., Cooper, G.M., and Shendure, J. (2014). A general framework for estimating the relative pathogenicity of human genetic variants. Nature genetics 46, 310-315.

Li, H., and Durbin, R. (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics 25, 1754-1760.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., and Durbin, R. (2009). The Sequence Alignment/Map format and SAMtools. Bioinformatics 25, 2078-2079.

Mortazavi, A., Williams, B.A., McCue, K., Schaeffer, L., and Wold, B. (2008). Mapping and quantifying mammalian transcriptomes by RNA-Seq. Nature methods 5, 621-628.

Ng, P.C., and Henikoff, S. (2003). SIFT: Predicting amino acid changes that affect protein function. Nucleic acids research 31, 3812-3814.

Sherry, S.T., Ward, M.H., Kholodov, M., Baker, J., Phan, L., Smigielski, E.M., and Sirotkin, K. (2001). dbSNP: the NCBI database of genetic variation. Nucleic acids research 29, 308-311.

Shirley, M.D., Tang, H., Gallione, C.J., Baugher, J.D., Frelin, L.P., Cohen, B., North, P.E., Marchuk, D.A., Comi, A.M., and Pevsner, J. (2013). Sturge-Weber syndrome and port-wine stains caused by somatic mutation in GNAQ. The New England journal of medicine 368, 1971-1979.

Smyth GK (2005). limma: Linear Models for Microarray Data. In Bioinformatics and Computational Biology Solutions Using R and Bioconductor Statistics for Biology and Health, Robert Gentleman *et al*, ed. (Springer New York), pp. 397-420.

Stenson, P.D., Mort, M., Ball, E.V., Shaw, K., Phillips, A., and Cooper, D.N. (2014). The Human Gene Mutation Database: building a comprehensive mutation repository for clinical and molecular genetics, diagnostic testing and personalized genomic medicine. Human genetics 133, 1-9.

Trapnell, C., Pachter, L., and Salzberg, S.L. (2009). TopHat: discovering splice junctions with RNA-Seq. Bioinformatics 25, 1105-1111.