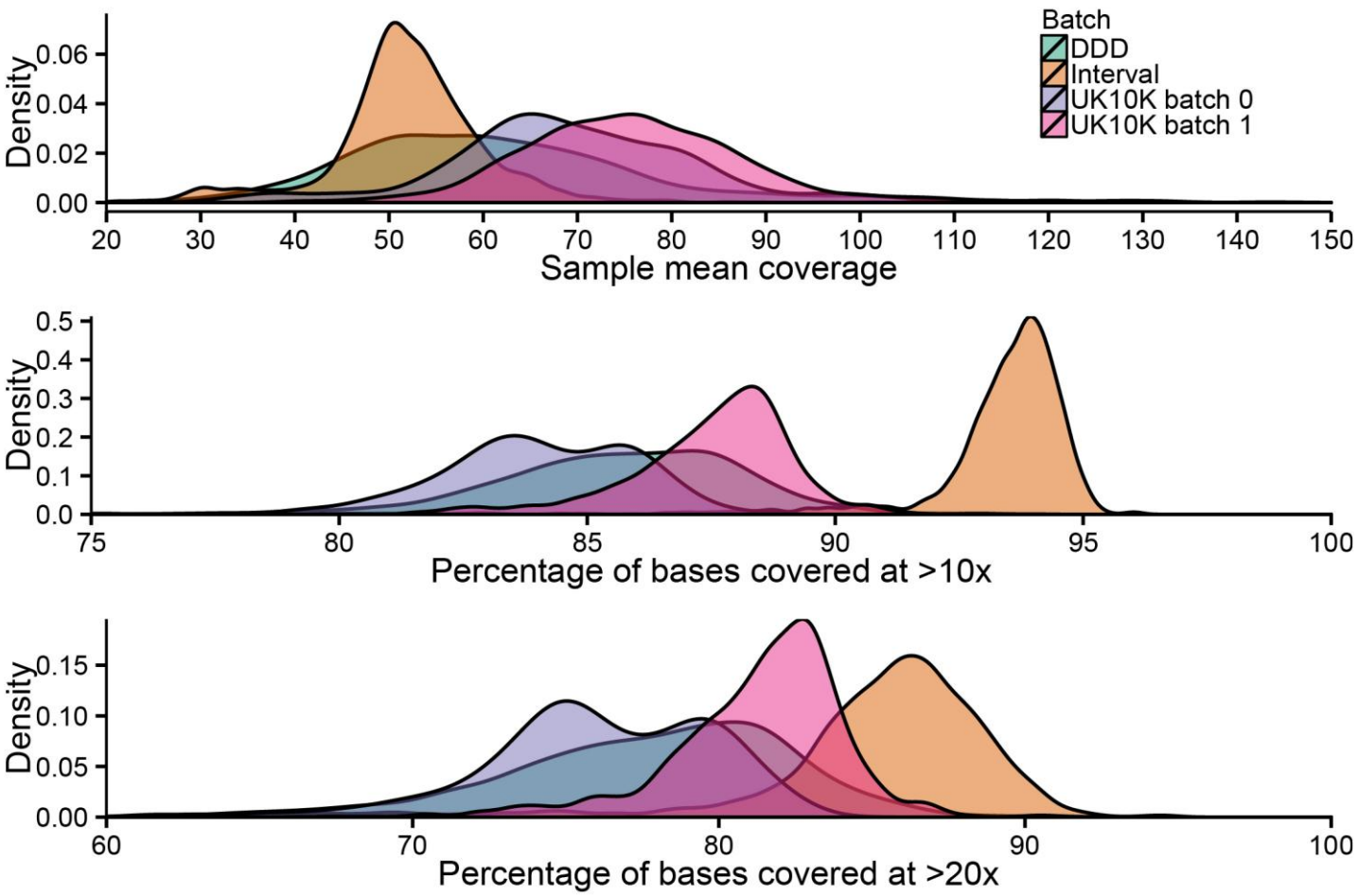
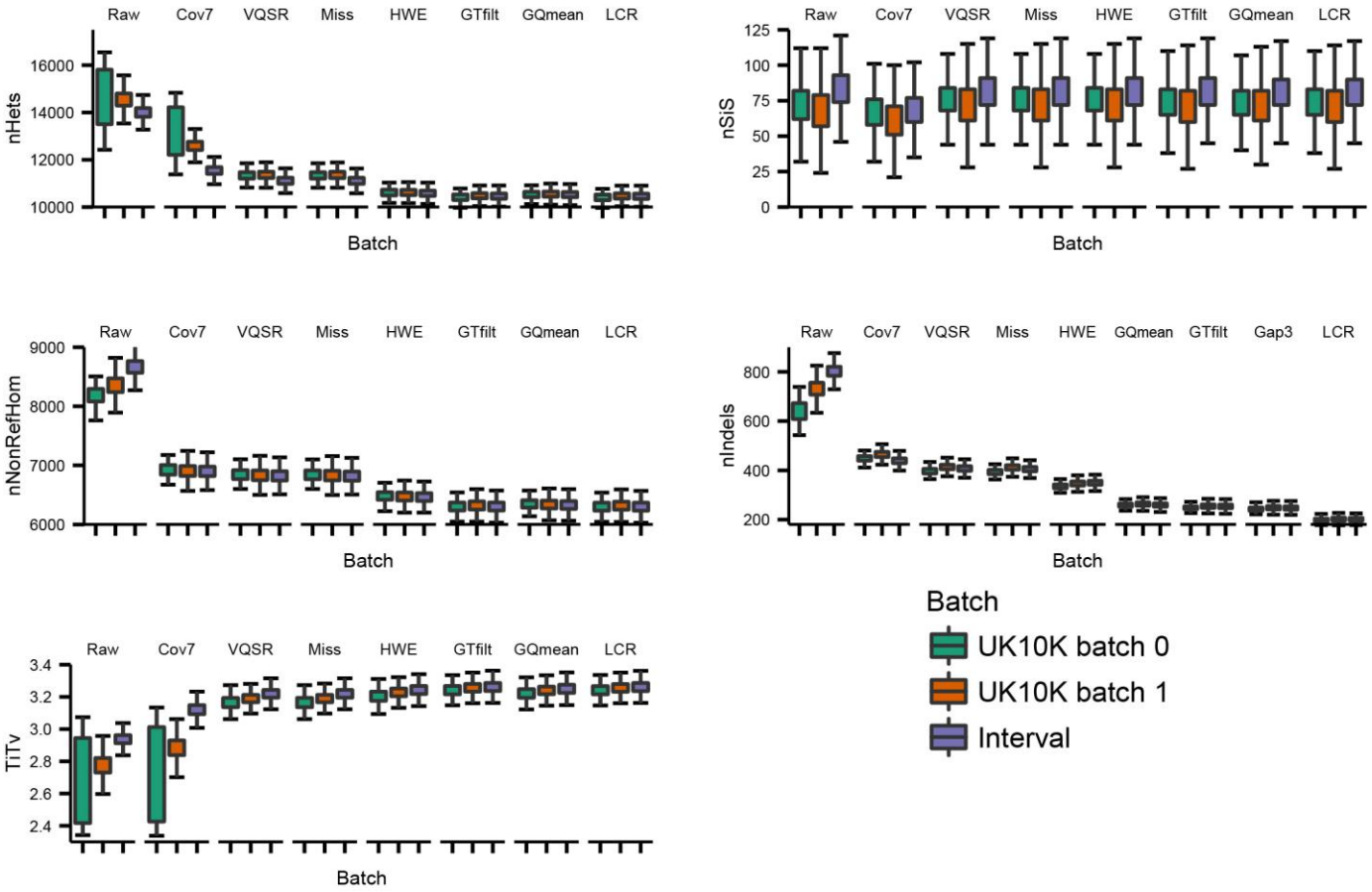


Supplementary Figure 1

Density plots of sequence coverage in the UK10K, INTERVAL and DDD data sets.

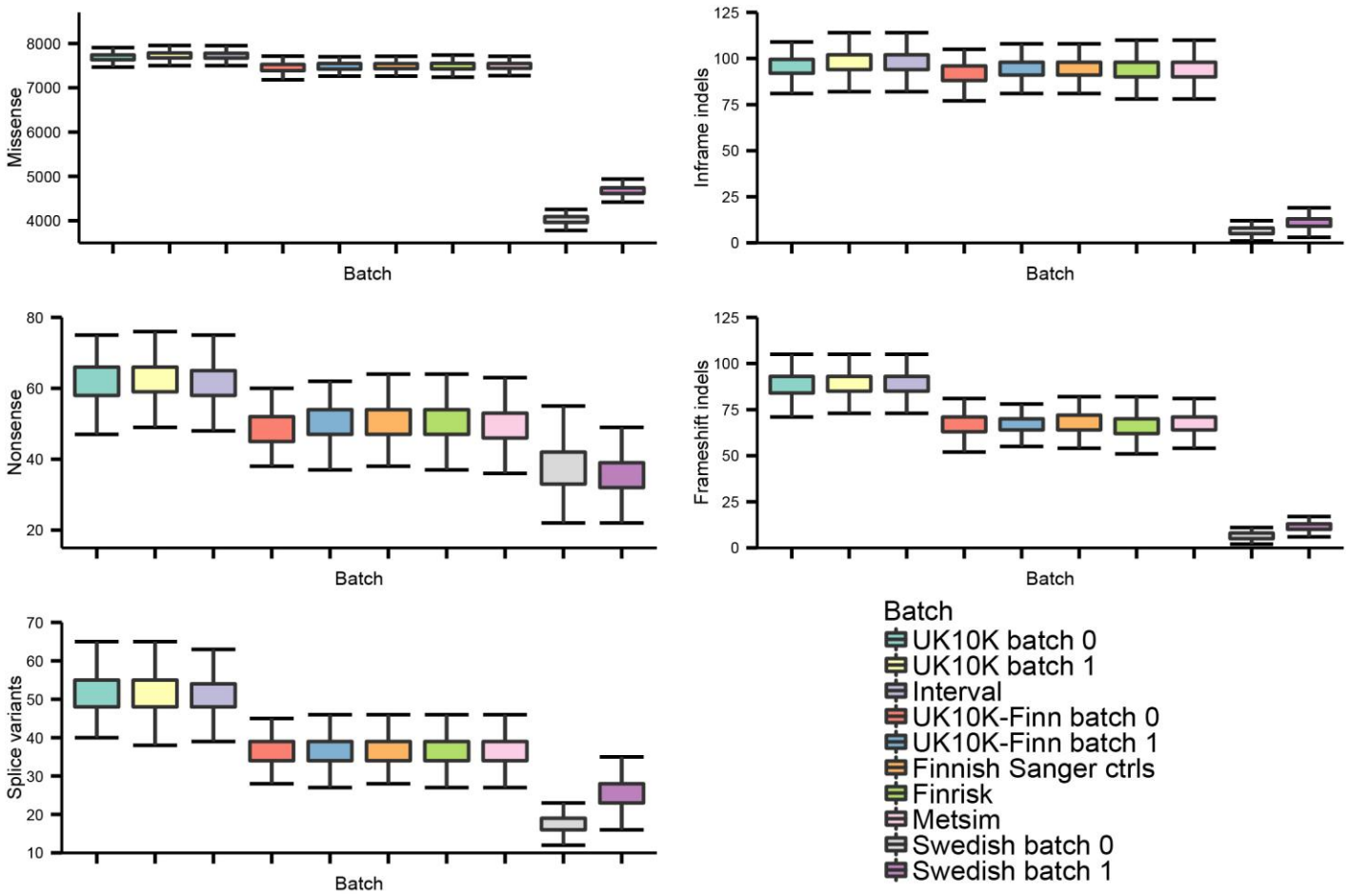
Per-sample sequence coverage was calculated and summarised from exome sequencing data generated in the UK10K (N = 4,734 in batch 0, and N = 562 in batch 1), INTERVAL (N = 4,502), and DDD (N = 1,972) data sets. The UK10K data set was separated into two sequencing batches. Top: sample mean coverage; Middle: percentage of Gencode v19 coding bases covered at 10x or more in each sample; Bottom: percentage of Gencode v19 coding bases covered at 20x or more in each sample.



Supplementary Figure 2

Variant metrics in the UK10K and INTERVAL data sets after each variant filtering step.

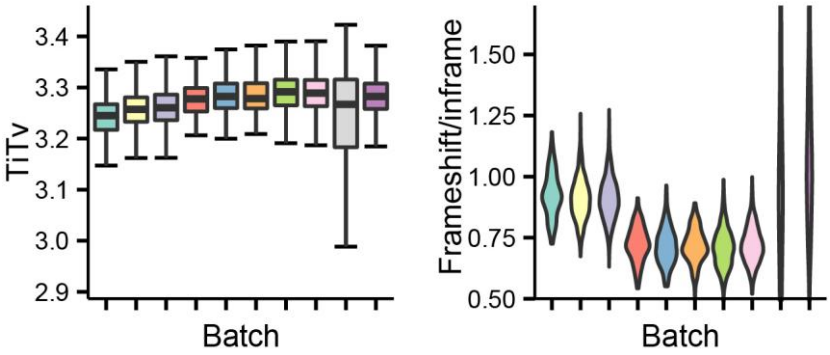
Box plots of per-sample heterozygote count (nHets), non-reference homozygote count (nNonRefHom), transition/transversion rate (TiTv), number of singletons (nSiS), and number of indels (nIndels) following each variant QC step. Variant metrics were summarised across all samples in the UK10K and INTERVAL data sets. Raw: no variant QC steps applied; Cov7: restricting to variants with at least 7x mean coverage; VQSR: GATK variant calibration using default parameters; Miss: filter for excess missingness; HWE: filter for deviation from Hardy-Weinberg equilibrium; GTfilt: filter for low alternate allele read depth, and abnormal allelic balance; GQmean: filter for low genotype quality; LCR: exclude variants in low-complexity regions. See Online Methods on more information on each step of variant QC.



Supplementary Figure 3

Variant counts summarized according to variant class and sequencing batch in the UK10K, INTERVAL, Finnish and Swedish data sets.

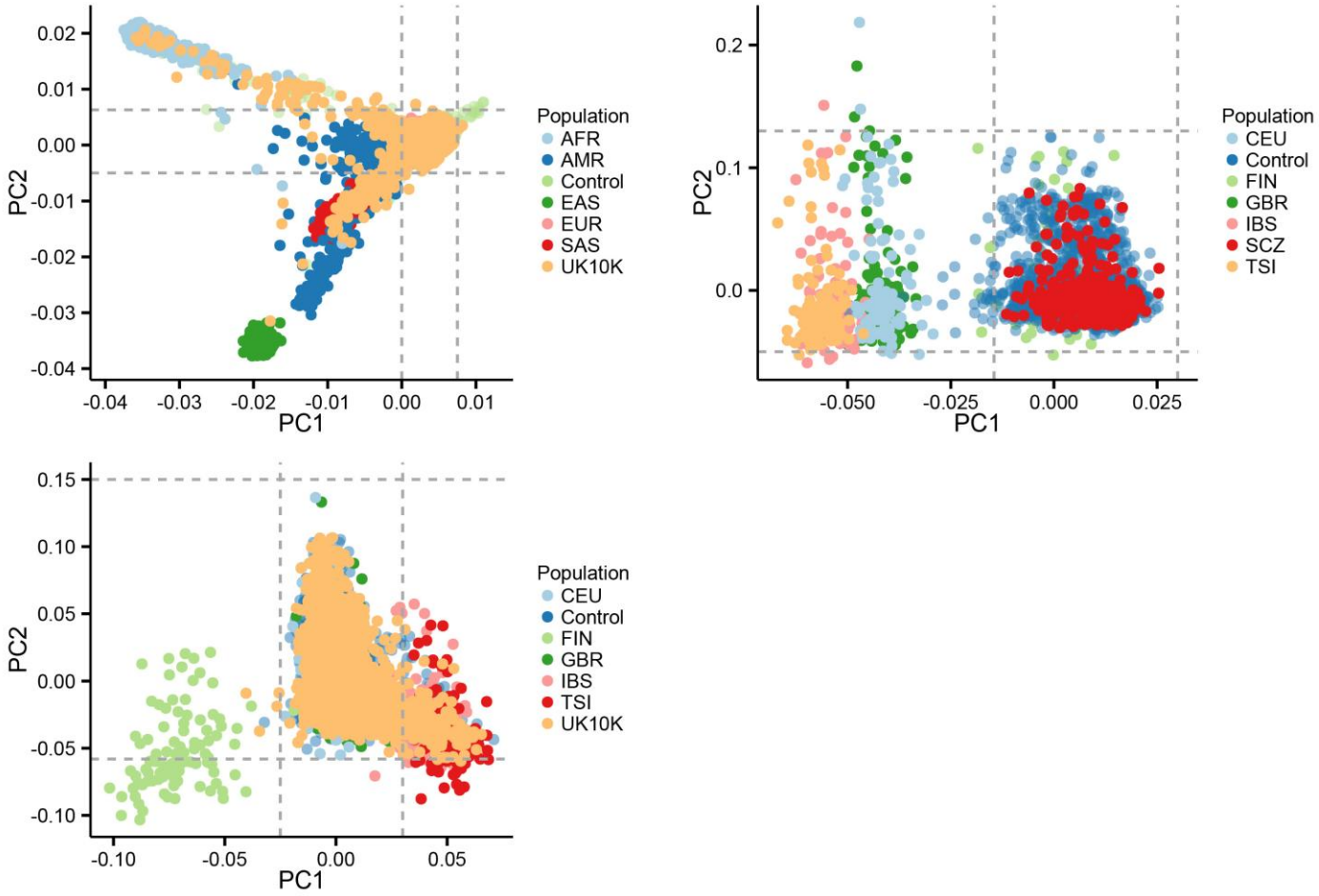
Box plots of per-sample variant counts in the UK10K, INTERVAL, Finnish, and Swedish data sets. All samples included in our meta-analysis are represented in the figure. The UK10K data sets was sub-divided according to sequencing batches (batch 0 and batch 1), and sample ancestry (UK and Finnish). The Finnish control data sets was separated by study of origin (Metsim, Finrisk, and Sanger controls). The Swedish case-control data set was separated into two sequencing batches. Differences exist in total variant counts between the UK, Finnish, and Swedish collections, likely reflecting differences in sequencing depth, capture reagents, sequencing protocol, read alignment, and variant calling. However, variant counts and population genetics metrics were consistent between cases and controls within each population group.



Supplementary Figure 4

Distributions of TiTv and frameshift/in-frame ratios in the UK10K, INTERVAL, Finnish and Swedish data sets.

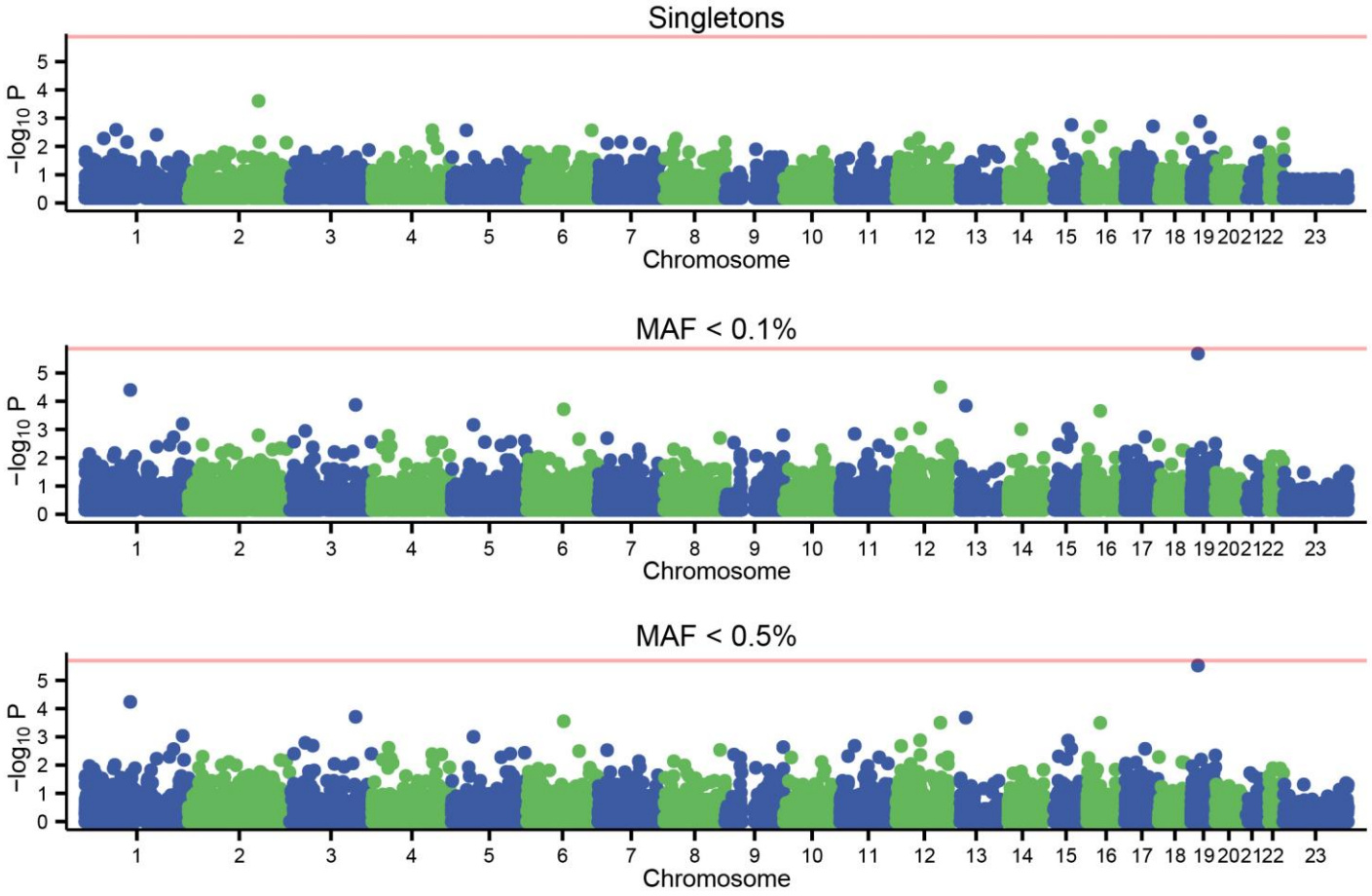
Box plot of sample TiTv (left) and violin plot of sample frameshift-to-inframe ratio (right) in the UK10K, INTERVAL, Finnish, and Swedish data sets. All samples included in our meta-analysis are represented in the figure. See Supplementary Figure 3 for the legend, and a description of each batch and sub-study. Following sample and variant QC, the per-sample transition/transversion rate was comparable between all populations (mean ~3.25).



Supplementary Figure 5

Principal component analysis of UK and Finnish samples in our UK10K schizophrenia data set.

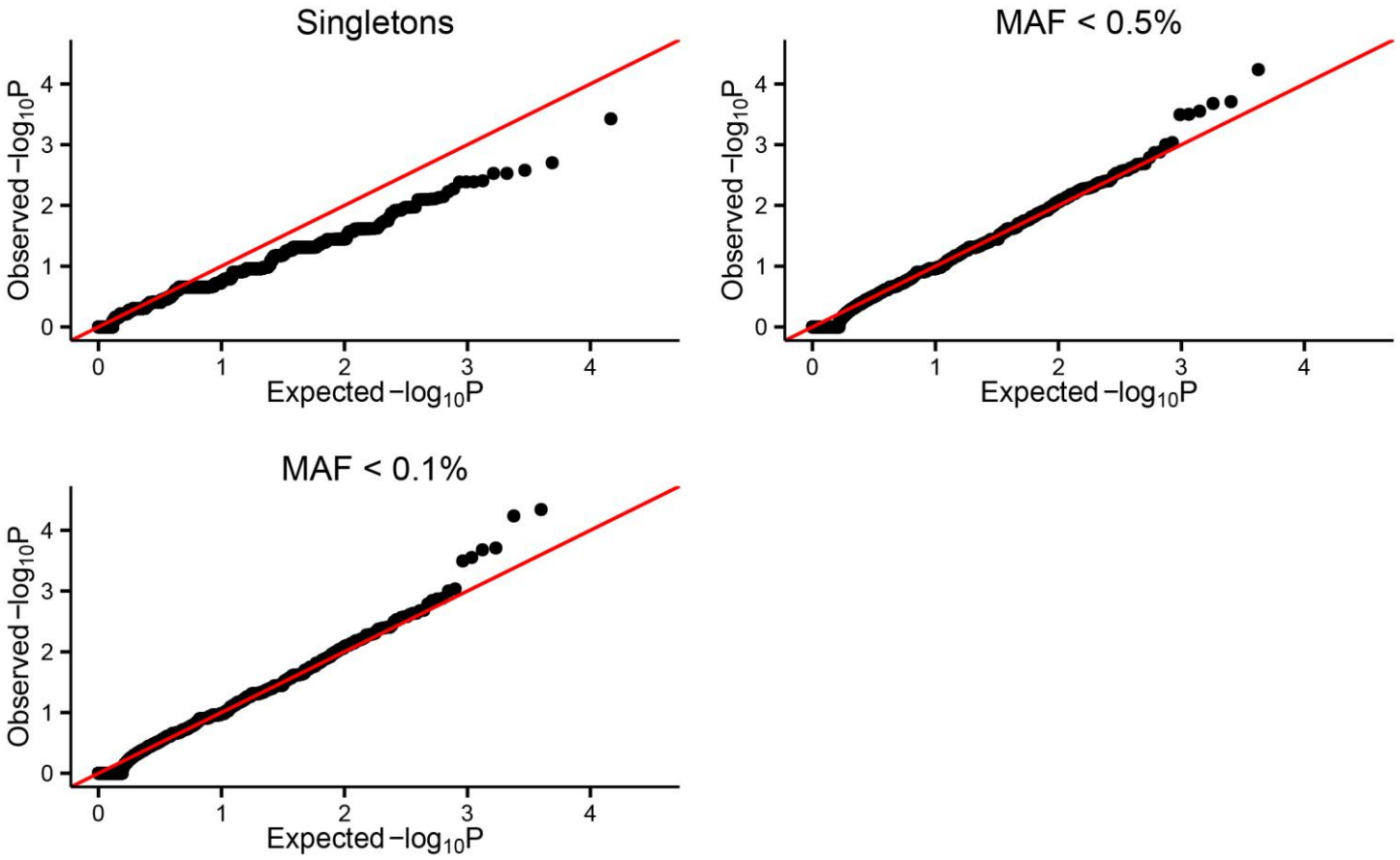
Principal components were estimated using 1000 Genomes samples, onto which we projected our cases and controls. We verified if samples had the same population ancestry (UK, Finnish or Swedish) as reported in the sample manifests, and excluded individuals who were of non-European ancestry. Thresholds for sample inclusion and exclusion are shown as dashed lines in each plot. Our analyses incorporated this information to perform permutations within each population (UK, Finnish, and Swedish) to control for ancestry and batch-specific differences. Top left: Population structure of all UK10K samples, with 1000 Genomes populations used as bases. We restricted our analyses to individuals of European ancestry; Bottom left: PCA plot of individuals of non-Finnish European ancestry in the UK10K data set with 1000 Genomes European populations used as bases. Samples not within the UK cluster (center of the plot) were excluded from analysis; Top right: PCA plot of individuals of Finnish ancestry in the UK10K data set. Samples not in the Finnish cluster (right of plot) were excluded from analysis. The three-letter symbols describing each population originate from nomenclature in the 1000 Genomes Project.



Supplementary Figure 6

Manhattan plot of the rare variant association analysis of LoF variants in 4,264 cases and 9,343 controls.

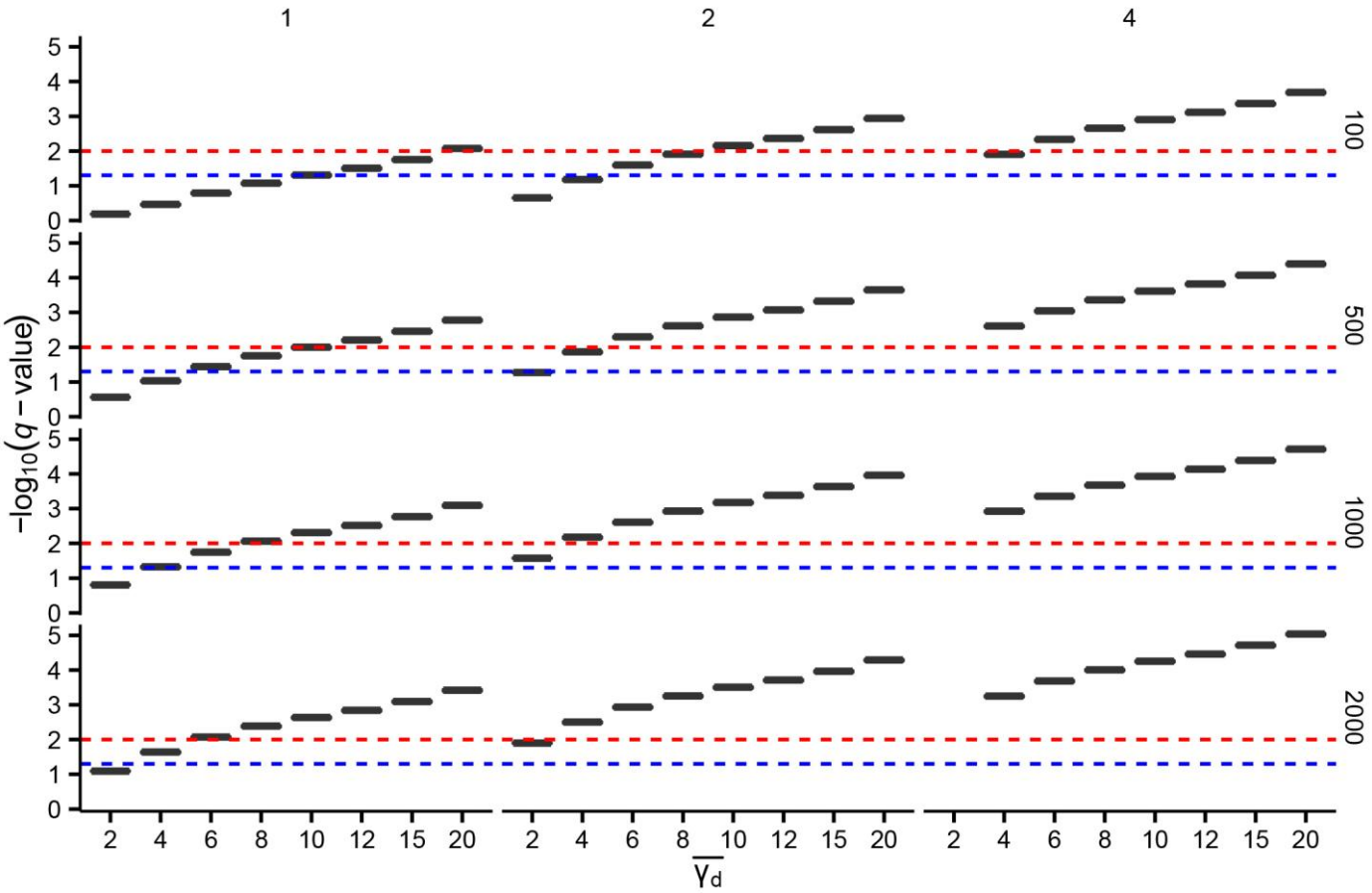
We tested for an excess of LoF variants within 18,271 genes using Fisher's exact test. – $\log_{10}$ *P*-values were plotted against the chromosomal location (mid-point) of each gene. We showed results from three allele frequency thresholds (singletons, < 0.1% and < 0.5%) for aggregating rare variants. No gene exceeded the exome-wide significant threshold of $P = 1.25 \times 10^{-6}$ (red line).



Supplementary Figure 7

Q-Q plots of the rare variant association analysis of LoF variants in 4,264 cases and 9,343 controls.

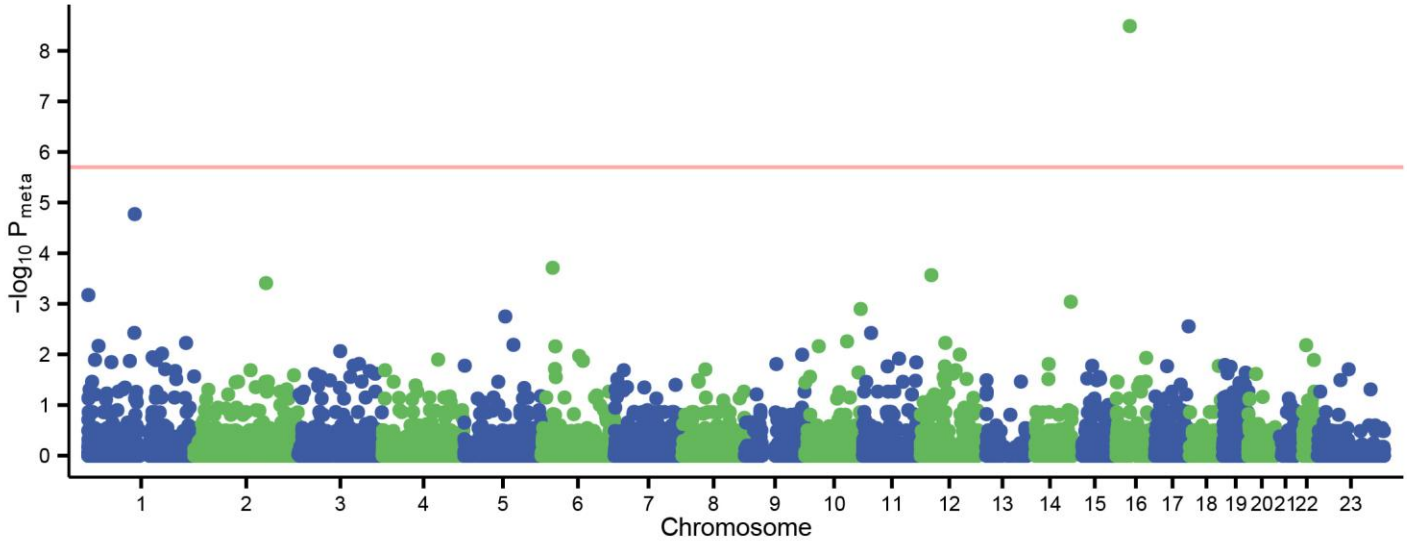
We tested for an excess of LoF variants within 18,271 genes using Fisher's exact test, and plotted the ordered – $\log_{10}$ *P*-values against transformed *P*-values sampled from the uniform distribution. The *Q-Q* plots for gene burden tests with minor allele frequency cut-offs of 0.1% and 0.5% followed an expected null distribution. The Q-Q plot for the burden test of singleton variants still showed deflation because the per-gene counts are too low and the data does not meet the asymptotic requirements of the statistical test. We included *P*-values from informative tests in which genes have at least one case LoF count.



Supplementary Figure 8

The robustness of the *SETD1A* result across reasonable parameters in the TADA model.

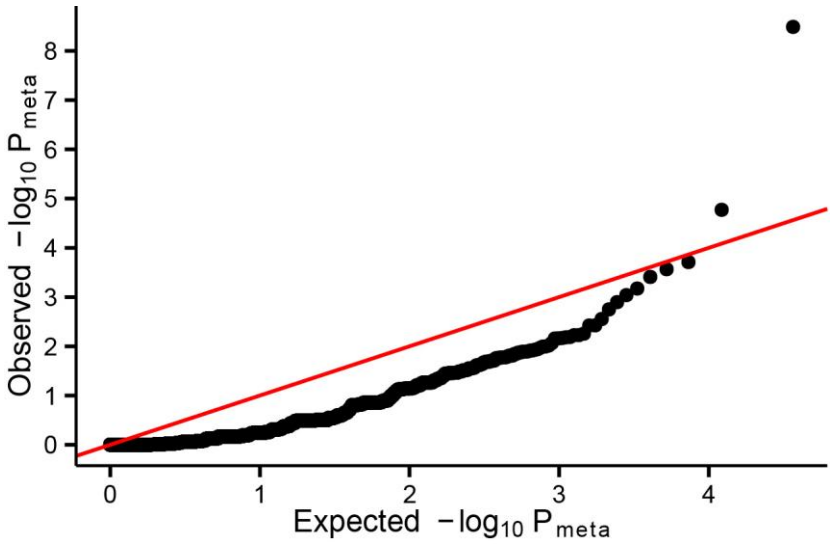
Because the TADA model depended heavily on the specification of its hyperparameters, we calculated the log *q*-value of *SETD1A* across different mean relative risk of *de novo* variants ($\bar{\gamma}_d$), mean relative risk of case-control variants ($\bar{\gamma}$), and numbers of true schizophrenia risk genes (*k*). Each vertical column is a different value for $\bar{\gamma}$, and each horizontal facet is a different value for *k*. Our signal in *SETD1A* had a *q*-value < 0.01 across all reasonable parameters. Blue line: *P* = 0.05; red line: *P* = 0.01.



Supplementary Figure 9

Manhattan plot of the meta-analysis of *de novo* mutations and case-control variants in 1,077 trios, 4,264 cases and 9,343 controls.

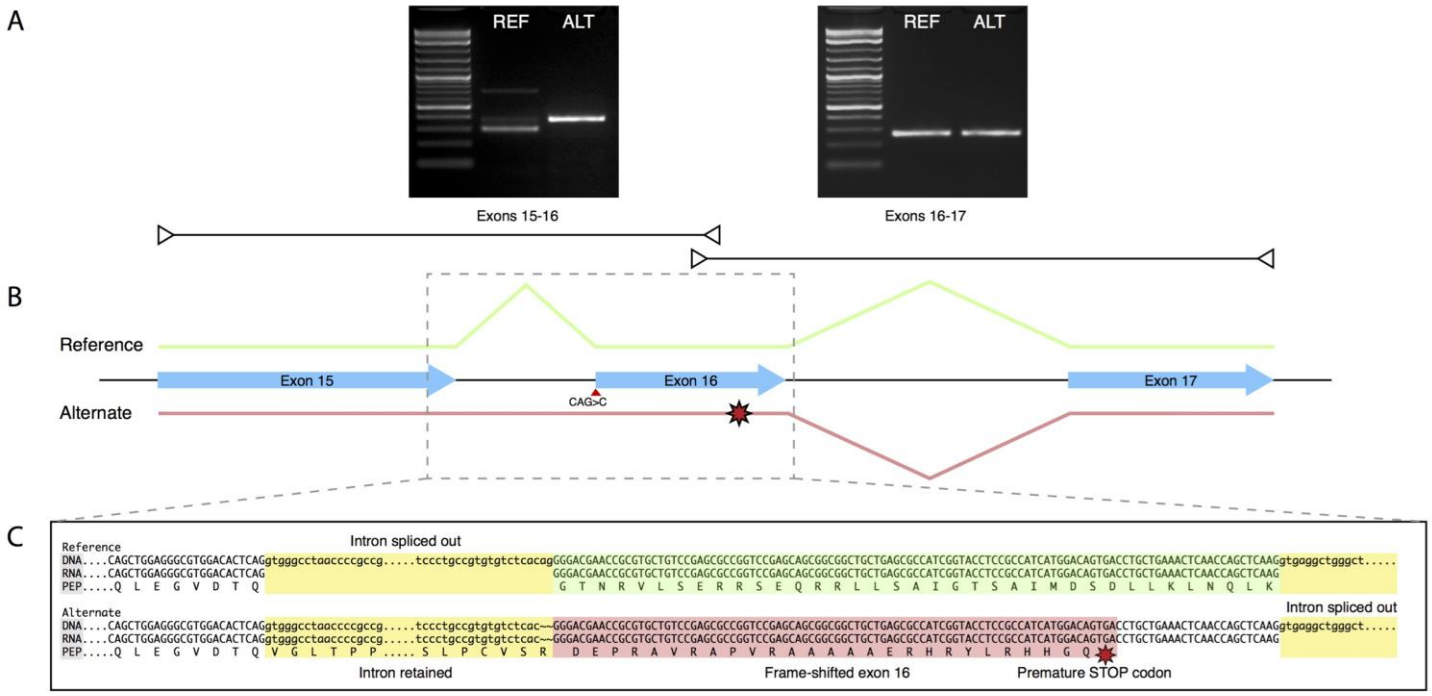
De novo and case-control burden *P*-values were meta-analysed using Fisher's combined probability method. – $\log_{10}$ *P*-values were plotted against the chromosomal location (mid-point) of each gene. A total of 18,271 genes were tested. Only *SETD1A* exceeded exome-wide significance, with $P = 3.3 \times 10^{-6}$. Red line: $P = 1.25 \times 10^{-6}$.



Supplementary Figure 10

Q-Q plot of the meta-analysis of *de novo* mutations and case-control variants in 1,077 trios, 4,264 cases and 9,343 controls.

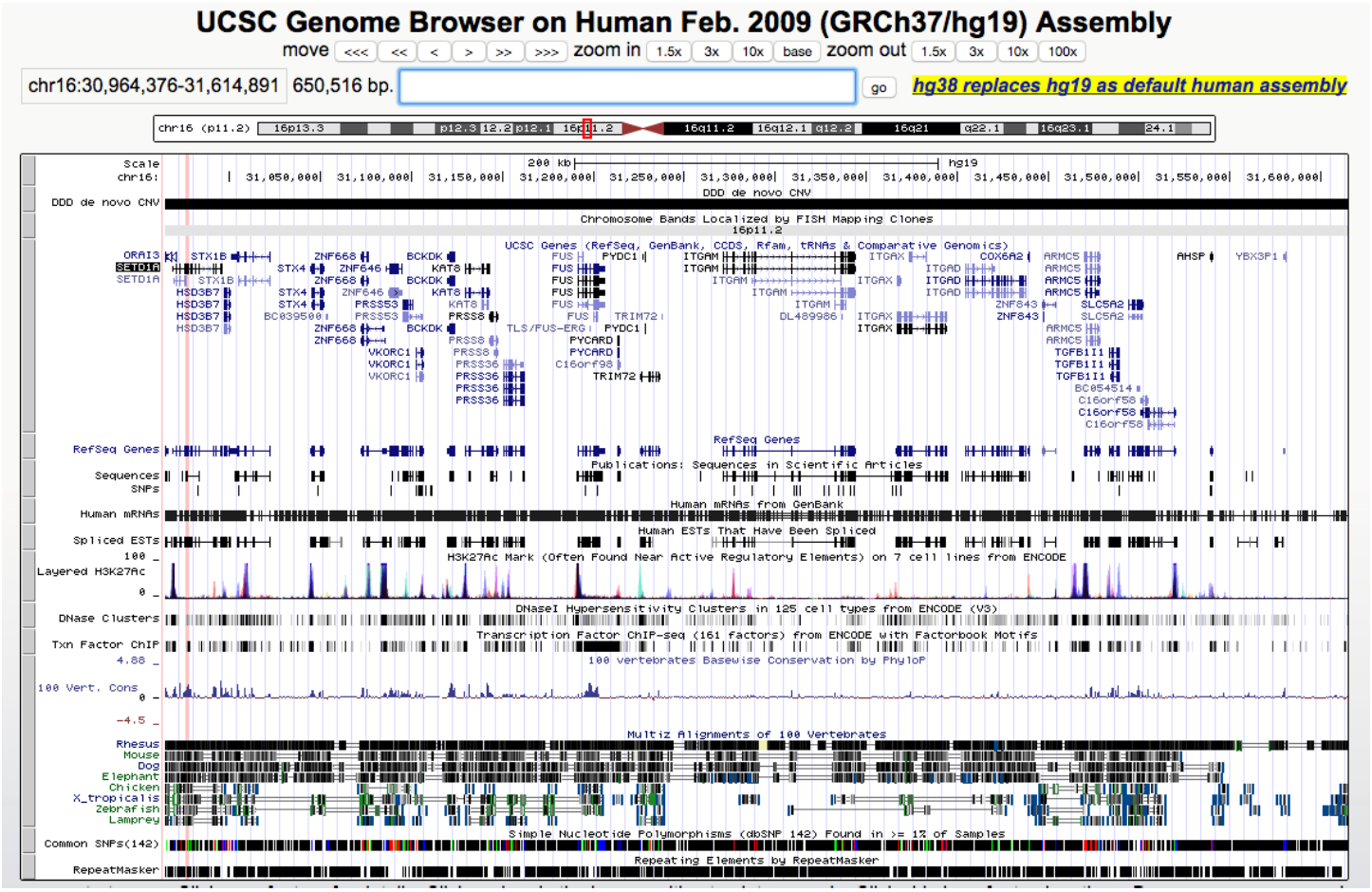
De novo and case-control burden *P*-values were meta-analysed using Fisher's combined probability method, and the $\log_{10}$ *P*-values plotted against transformed *P*-values sampled from the uniform distribution. Because only a subset of genes had *de novo* LoF variants, Fisher's method deflated the combined *P*-value of genes without any *de novo* information.



Supplementary Figure 11

Results from the minigene experiment assessing the impact of the exon 16 splice acceptor site variant.

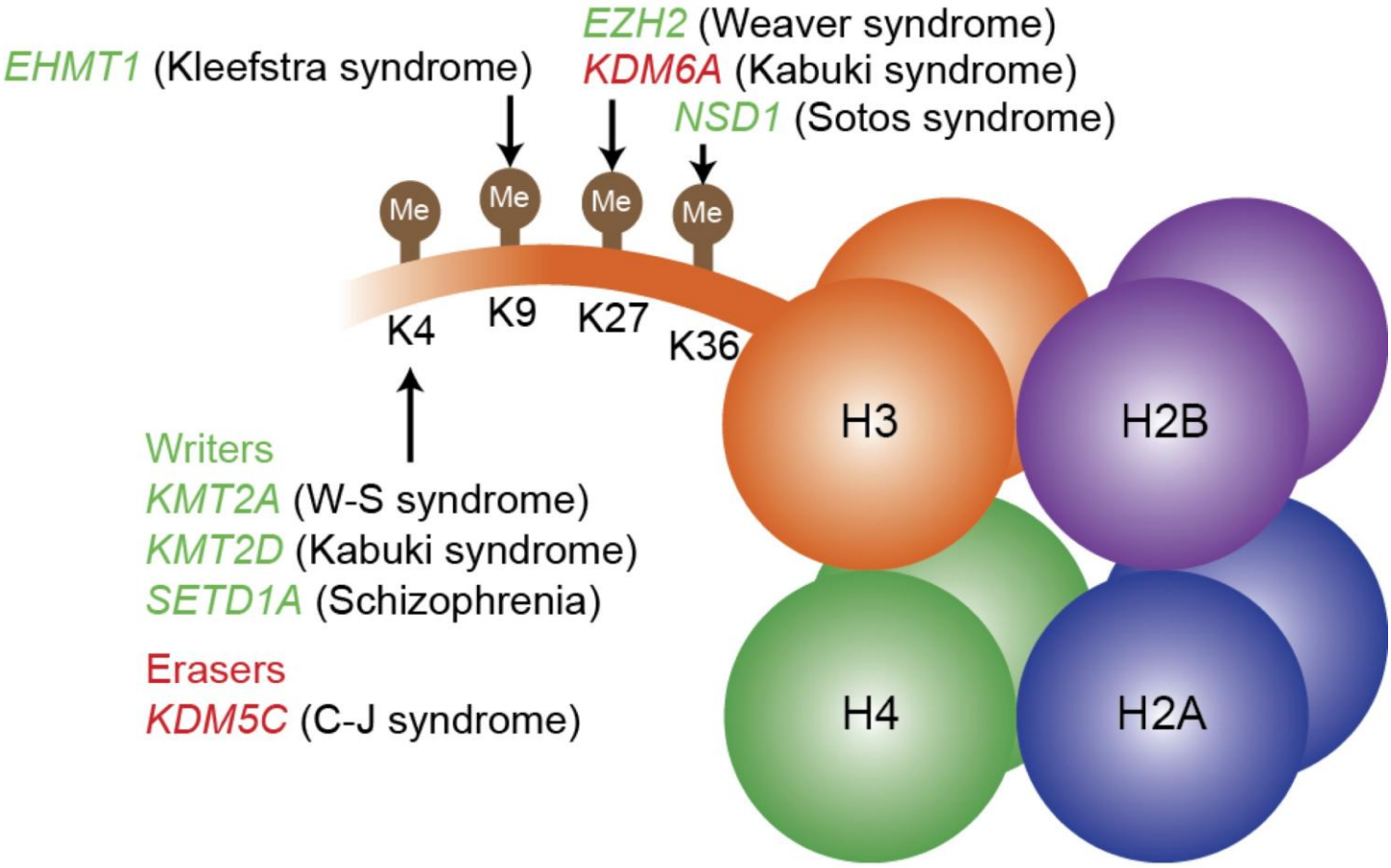
A. Minigene constructs driving expression of exons 15, 16 (ref and alt), and 17 fused to GFP were transfected into HELA cells. RT-PCR analysis of cell lysates using primer pair 2, spanning exons 15, 16, and the intervening intron reveal a change in size of PCR products suggesting retention of the intervening intron in the construct containing the splice-acceptor deletion (panel A, "Exons 15-16", REF versus ALT). PCRs with primer pair 3, spanning the intron downstream of exon 16 show no change in band sizes (panel A, "Exons 16-17", REF versus ALT), suggesting this intron is correctly spliced out in both reference and alternate forms. **B.** Cartoon of genomic locus surrounding the exon 16 splice acceptor deletion. The predicted structure of reference (green) and deletion containing (red) transcripts are shown above and below genomic map. The red star indicates a predicted premature stop codon due to intron retention and resulting frame-shifted translation. **C.** Results from capillary sequencing of PCR products from panel A confirms intron retention in the splice acceptor deletion construct (panel C, "RNA", yellow box). This will result in a predicted frame-shifted translation of exon 16 (panel C, "PEP", red box), and a premature truncation of the protein 28 amino acids into exon 16 (red star). Downstream intron splicing was confirmed by capillary sequencing to be intact in both constructs.



Supplementary Figure 12

De novo microdeletion of a single copy of *SETD1A* identified in the DDD study.

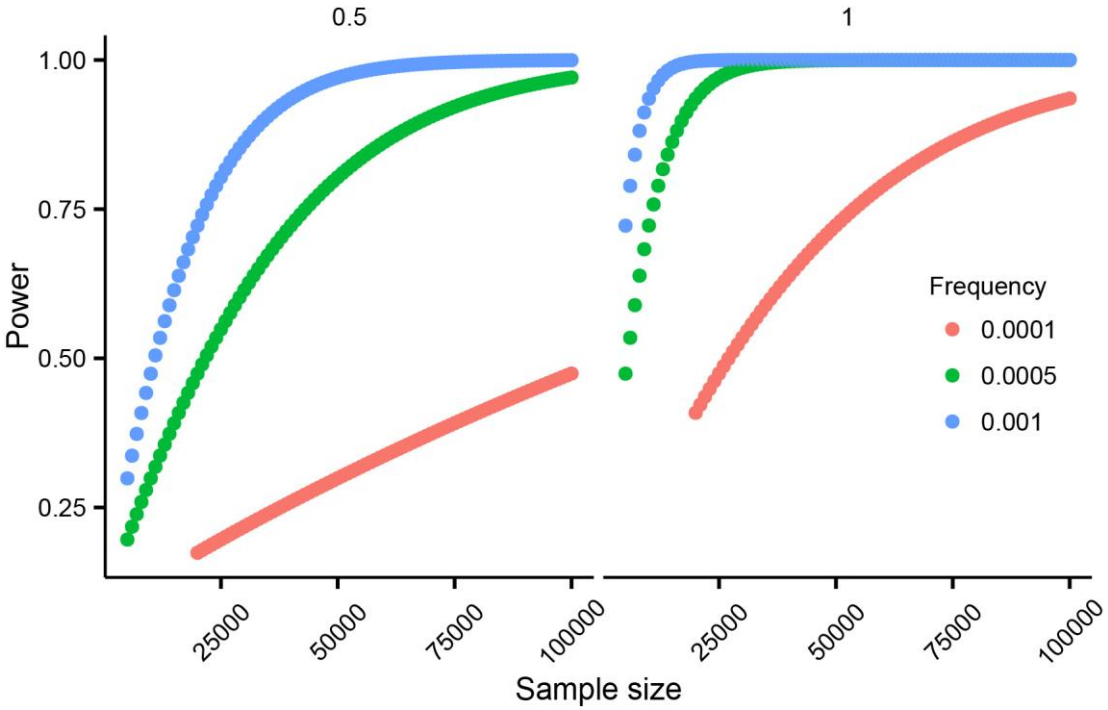
A proband was identified to have a 650 kb deletion encompassed *SETD1A* and 29 other genes. The figure showing the deletion was generated using the UCSC Genome Browser (https://genome.ucsc.edu/).



Supplementary Figure 13

Mendelian disorders of epigenetic machinery at histone H3.

Writers (in green) add methyl groups at the specified residue of the histone tail, while erasers (in red) perform targeted demethylation. Disrupting variants in writers and erasers described in the figure result in well-known examples of dominant, highly penetrant disorders characterised by developmental delay and intellectual disability. Only the tail of histone H3 and its four key lysine residues are illustrated here. Alternate nomenclature: *EHMT1* (also known as *KMT1D*), *EZH2* (*KMT6A*), *NSD1* (*KMT3B*), *SETD1A* (*KMT2F*).



Supplementary Figure 14

Sample size curves for detecting an increased risk of premorbid cognitive impairment in schizophrenia *SETD1A* LoF carriers.

We performed power calculations using a simple one-sided *t*-test to identify sample sizes required to show possible cognitive impairment in *SETD1A* schizophrenia carriers. Effect sizes *d* (0.5, 1), and allele frequencies (0.0001, 0.0005, 0.001) are varied to show their influence on statistical power. We assume a Type I error probability of 0.05. For these effect sizes and frequencies, a sample of tens of thousands of cases will be needed.

First author	Year	Journal	Sample size	Capture	Sequencer	Validation	PMID
Guipponi	14	PLOS ONE	53	Agilent SureSelect Human ALL Exon kits	HiSeq	Yes (Sanger)	25420024
Takata	14	Neuron	231	Agilent SureSelect v2 (n = 85 trios), NimbleGen SeqCap EZ v2 (n = 180 trios)	HiSeq	Yes (Sanger)	24853937
McCarthy	14	Mol Psychiatry	57	NimbleGen's SeqCap EZ Human Exome Library v2.0 probes	HiSeq (101 bp PE reads)	Yes (Sanger)	24776741
Fromer	14	Nature	617	Agilent SureSelect Human All Exon v.2, Nimblegen SeqCap EZ Human Exome Library v2.0, Agilent SureSelect Human All Exon 50MB	HiSeq (76 bp, 101 bp, 76 bp PE reads)	Yes (Sanger)	24463507
Gulsuner	13	Cell	105	NimbleGen SeqCap EZ Human Exome Library v2.0	HiSeq (101 bp PE reads)	Yes (Sanger)	23911319
Xu	12	Nature Genetics	231	Agilent SureSelect v2 (n = 85 trios), NimbleGen SeqCap EZ v2 (n = 180 trios)	HiSeq	Yes (Sanger)	23042115
Xu	11	Nature Genetics	53	Agilent SureSelect Human All Exon Target Enrichment System	HiSeq (50 bp PE reads)	Yes (Sanger)	21822266
Girard	12	Nature Genetics	14	Agilent SureSelect All Exome Kit v.1	HiSeq (76 bp PE reads)	Yes (Sanger)	21743468

Table 1: Published studies identifying *de novo* mutations in schizophrenia parent-proband trios using whole-exome sequencing.

Gene name	μ_{LoF}	$N_{de\ novo}$	N_{case}	N_{control}	$P_{de\ novo}$	P_{burden}	P_{meta}
SETD1A	6.6e-06	3	7	0	4.6e-07	0.0003	3.3e-09
TAF13	1.3e-06	2	1	0	3.7e-06	0.31	1.7e-05
HIST1H1E	2.4e-07	1	3	0	0.00053	0.031	0.00019
BCAT1	1.9e-06	1	8	3	0.004	0.0058	0.00027
XIRP2	3.3e-06	0	41	35	1	3.5e-05	0.00039
KLHL17	3e-06	1	4	0	0.0065	0.0096	0.00067
HSP90AA1	3.1e-06	1	5	1	0.0066	0.013	0.00091
MKI67	1e-05	2	5	10	0.00024	0.53	0.0013
CAST	3.1e-06	0	15	6	1	0.00019	0.0018
ENDOV	2.2e-06	0	10	2	1	0.00031	0.0028

Table 2: Meta-analysis results for 1,077 trios, 4,264 cases and 9,343 controls. Only *SETD1A* reached exome-wide significance.

Collection	Sample size	Population	Description
ABERDEEN	391	UK	Schizophrenia cases with cognitive measurements recruited in Aberdeen, Scotland.
COLLIER	172	UK	Subjects recruited from three different studies: the Genetics and Psychosis (GAP) study (early-onset schizophrenia), the Maudsley twin series, and the Maudsley family study (families with a history of schizophrenia or bipolar disorder).
EDINBURGH	234	UK	Subjects recruited from psychiatric facilities in Scotland with IQ > 70. 138 are familial cases, and 100 have deep neuroimaging information.
GURLING	45	UK	Subjects from multiply affected families all of which are unilineal for transmission of schizophrenia.
MUIR	103	UK	Subjects with autism, schizophrenia or some sort of psychoses with diagnoses of mental retardation. Only individuals with schizophrenia were included in our analysis.
UKSCZ	542	UK	UK and Irish subjects selected for a positive family history of schizophrenia (collected as sib-pairs or from multiplex kindreds), or are systematically recruited from South Wales and have undergone detailed cognitive testing.
KUUSAMO	120	Finland	Subjects from the Finnish Kuusamo internal isolate where there is a three-fold lifetime risk for schizophrenia.
Finnish SCZ	281	Finland	Subjects from a population cohort recruited from national registers and have two affected siblings.

Table 3: Description of sample collections included as cases in the UK10K schizophrenia analysis.

Collection	Sample size	Population	Description
UK10K Obesity TwinsUK	67	UK	Consists of individuals from the TwinsUK study with a BMI > 40.
UK10K Obesity Generation Scotland	422	UK	Subjects belong to a family-based genetic study from across Scotland, and consists of individuals with extreme obesity. Only unrelated individuals are included.
UK10K Rare Severe Insulin Resistance	119	UK	Trios in which the probands have been diagnosed with severe insulin resistance. Only unaffected parents are included as controls.
UK10K Rare Neuromuscular	116	UK	Trios in which the probands have congenital muscle dystrophies or myopathies, neurogenic conditions, mitochondrial disorders, or periodic paralysis. Only unaffected parents are included as controls.
UK10K Rare Thyroid	123	UK	Trios in which the probands have congenital hypothyroidism due to either dysgenesis or dyshormonogenesis. Only unaffected parents are included as controls.
UK10K Rare Hypercholesteromia	123	UK	Trios in which the probands have a consistently high level of LDL, and do not contain common APOB and PCSK9 mutations, or detectable LDLR mutations. Only unaffected parents are included as controls.
INTERVAL Sequencing Project	4499	UK	A cohort of healthy blood donors collected from 25 donation centres across England.
ENGAGE	283	Finland	A collection of individuals selected from Health 2000 and FINRISK cohorts based on properties of their metabolic trait profiles.
Health 2000 Survey	277	Finland	A study based on a nationally representative sample of persons aged 30 and over, with a goal of obtaining general public health information on the working-aged and aged population.
Familial dyslipidemia study	84	Finland	Individuals from families with dyslipidemia and are of Finnish origins. Only unrelated individuals are included.
FINRISK study controls	769	Finland	The FINRISK study is a large population survey investigating risk factors of chronic, non-communicable diseases. We include samples that are controls in an on-going inflammatory bowel disease exome sequencing study.
METSIM study	984	Finland	The cross-sectional METSIM study investigates genetic and non-genetic risk factors related to Type II Diabetes, cardiovascular disease, and insulin resistance. The controls included were sequenced to investigate rare variation related to these phenotypes.

Table 4: Description of sample collections included as controls in the UK10K schizophrenia analysis.

Gene name	DNM LoF	Case LoF	Ctrl LoF	DNM Mis15	Case Mis	Ctrl Mis	BF	q-value
SETD1A	3	7	0	0	24	50	2e+05	7.7e-05
XIRP2	0	41	35	0	81	145	7.4e+02	0.01
TAF13	2	1	0	0	4	9	5.9e+02	0.016
HSPA8	1	0	1	2	5	12	2.7e+02	0.025
BCAT1	1	8	3	0	10	25	2.7e+02	0.031
CAST	0	15	6	0	33	50	1.6e+02	0.041
NIPAL3	1	2	1	1	8	22	1.3e+02	0.05
HSP90AA1	1	5	1	0	20	48	1.2e+02	0.059
SSBP3	1	1	0	1	4	11	1.1e+02	0.066
KLHL17	1	4	0	0	28	58	1e+02	0.073
MKI67	2	5	10	0	27	75	1e+02	0.078
SLC25A24	0	14	6	0	8	22	92	0.084
PIK3C2B	1	3	2	1	54	83	89	0.089
DPYD	1	6	7	1	34	92	88	0.093
HIST1H1E	1	3	0	0	14	22	77	0.099
IGSF22	0	13	5	0	23	73	69	0.1
RYR3	0	11	6	2	120	242	68	0.11
ENDOV	0	10	2	0	5	10	66	0.11
LPHN2	1	2	3	1	28	49	45	0.12
PHF7	1	0	0	1	4	16	43	0.13
ORC3	0	16	10	0	25	41	42	0.14
BLNK	1	2	0	0	6	19	40	0.14
URB2	1	12	13	0	20	36	37	0.15
ZEB1	1	2	0	0	21	37	36	0.15
NUP214	1	4	2	0	74	146	33	0.16
CRYBG3	1	1	4	1	20	48	31	0.17
BTNL2	1	2	1	0	3	7	31	0.17
INHBC	1	2	1	0	9	18	30	0.18
POGZ	1	2	0	0	29	44	29	0.19
STAC2	0	3	3	2	13	30	29	0.19
DLG2	1	4	3	0	22	58	28	0.2
PRRC2A	1	3	1	0	10	37	27	0.2
ST3GAL6	1	1	0	0	7	15	27	0.21
KRT15	0	4	0	1	18	28	27	0.21
RB1CC1	1	3	2	0	24	39	23	0.22
ZDHHC5	1	1	0	0	29	59	23	0.22
SMARCC2	1	3	2	0	19	38	23	0.23
OR2T2	0	12	7	0	0	1	23	0.23
ATG12	1	2	2	0	6	24	22	0.24
XPR1	1	1	0	0	5	22	22	0.24
AOX1	0	9	6	1	36	81	22	0.25
CDKL1	0	8	2	0	10	23	21	0.25
SPDYC	0	5	2	1	17	21	21	0.25
RECK	0	7	4	1	21	52	20	0.26
RTTN	1	9	10	0	42	82	19	0.26
XIRP1	0	13	8	0	27	88	18	0.27
SLC12A7	0	9	3	0	37	55	18	0.27
SYNGAP1	1	1	0	0	20	25	18	0.28
SCLT1	0	7	1	0	8	11	18	0.28
EPHA2	1	6	7	0	43	84	18	0.28
PYCARD	0	3	0	1	1	6	17	0.29
GTPBP3	1	1	1	0	16	23	17	0.29
SHANK1	1	1	0	0	10	15	17	0.29
KDM5C	1	1	0	0	3	6	17	0.3

Table 5: TADA results using the hyperparameters in the De Rubeis *et al.* autism meta-analysis. Only *SETD1A* has a q-value < 0.01.