Chapter 2

A protocol for the quality control of whole-exome sequencing data sets

2.1 Challenges behind the production and analysis of sequencing data

Whole-exome sequencing has emerged as the technology of choice in investigating the contribution of rare variation in the genetic basis of complex disorders. It has been most successful in identifying genes underlying rare Mendelian disorders, in which only a small number of samples are needed to reveal causal variants of large effect [110, 111]. Early results from complex diseases have demonstrated that a genome-wide burden of disruptive variants exists in cases compared to controls. However, the identification of individual risk genes remains elusive because a large number of genes appear to underlie many complex traits and our ability to differentiate pathogenic variants from neutral polymorphisms remains limited [78, 103, 88]. Much larger sample sizes, possibly in the tens of thousands, are required to identify sufficient numbers of rare variants to implicate individual risk genes [88, 105]. While studies have individually analysed a small number of exomes, in aggregate tens of thousands of whole-exome sequences have been generated to date [112]. Meta-analyses leveraging published data sets are beginning to have sufficient power for gene discovery.

Standardized protocols currently exist for performing variant discovery on whole-exome sequence data [113–117]. Raw reads in a FASTQ files are first mapped to a genome reference, duplicated reads are marked in the resulting BAM file to reduce amplification bias, and base quality scores are empirically adjusted for systematic errors. A variant caller such as Samtools mpileup or GATK HaplotypeCaller identifies sites at which a potential variant exists relative to the reference, and calculates the probabilities of each possible genotype

at that site [113, 114]. For very large data sets, samples are called individually and merged before variant calling occurs in aggregate. This enables the incorporation of variant-level information across samples when determining the appropriate genotype. Subsequently, a variant classifier, such as GATK Variant Quality Score Recalibration (VQSR), filters out mapping and sequencing artefacts. The remaining variants are annotated with predicted biological consequences and analysed. These best practices were successfully applied in Mendelian disorder studies and parent-proband trio studies analysing *de novo* mutations [110, 118, 98].

As we begin to jointly analyse thousands of samples aggregated from published studies, additional complexities in the preparation and production of whole-exome sequencing data begin to emerge. First, sequencing technologies have a higher genotyping error rate than array-based calls, and unlike common variant association studies, genotype refinement using a reference panel is unlikely to improve the quality of variant calls at the lowest end of the allele frequency spectrum [116]. To partially address this, each sample is sequenced to sufficiently high depth to ensure reasonable coverage ($40\times$ or greater) over the entire exome [116]. However, the enrichment of coding sequences using DNA hybridization inherently leads to uneven coverage: certain regions are captured to much greater affinity due to sequence context (high GC content), while other baits fail when overlapping polymorphisms modify its annealment affinity. Baits targeting low complexity regions capture reads from other repetitive sequences, leading to a even greater disparity of coverage across the exome. These limitations are further exacerbated by the substantial batch effects that appear from combining data from different exome sequencing studies. Depending on study design, researchers sequence samples to different mean coverage, which result in higher quality calls in some samples over others. Furthermore, a number of commercial captures are available for target enrichment, and each have systematic biases in its regional coverage of the exome. Finally, sequencing centres have different protocols for sample preparation, sequencing, and data production that are subject to change as technology progresses, all of which introduces additional variability between groups of samples. To aggregate and meta-analyse published sequencing data sets, we must first address these sources of systematic bias which often confound the results of rare variant association tests.

In this Chapter, I first describe the whole-exome sequencing data generated in the UK10K project, the Deciphering Developmental Disorders (DDD) study, INTERVAL study, Swedish Schizophrenia study, and the Sequencing Initiative Suomi (SiSU) project, all of which are analysed in Chapters 3 and 4. I then highlight the steps taken to prepare these data for analysis, and detail best practices to harmonize sequence production, variant calling, and variant- and sample-level QC across many thousands of whole-exome sequences. Useful

metrics for comparing variant quality between data sets are shown and discussed. Using diagnostic *de novo* mutations from the DDD study, I determine which *in silico* annotation tool best differentiated pathogenic from benign variants, which I then use to classify missense variants in subsequent analyses. Finally, I describe published whole-exome sequencing data sets of schizophrenia parent-proband trios, and extend a method of modelling the recurrence of *de novo* mutations for gene discovery.

2.1.1 Publication note and contributions

The results described in this chapter was peer-reviewed and published earlier this year [119]. I briefly summarise the various contributions to this project. The neuro group within the UK10K study recruited and whole-exome sequenced schizophrenia cases. This initiative was led by Aarno Palotie, Michael J. Owen, Jeffrey C. Barrett, and Daniel Geschwind. The sequencing team at the Wellcome Trust Sanger Institute performed exome capture, sequencing, and alignment for the UK10K and INTERVAL studies. I received the raw VCF for the Finnish case-control data set from Mitja I. Kurki and Aarno Palotie. I performed all subsequent production, and QC steps for these data under the supervision of Jeffrey C. Barrett. Unless explicitly stated, the parts of the peer-reviewed publication reproduced in this Chapter are my original work.

2.2 Materials and methods

2.2.1 Sample collections

Individuals clinically diagnosed with schizophrenia were recruited and exome sequenced as part of eight neurodevelopmental collections (Aberdeen, Collier, Edinburgh, Gurling, Muir, UK-SCZ, Finnish-SCZ, and Kuusamo) in the UK10K sequencing project. Matched population controls were selected from non-psychiatric arms of the UK10K project, healthy blood donors from the INTERVAL project, and five Finnish population studies (ENGAGE, Familial dyslipidemia, FINRISK, Health 2000, and METSIM). Additional details on the UK10K dataset are described in Table 2.1 and 2.2, and the sequence data have been deposited into the European Genome-phenome Archive (EGA) under study accession EGAO00000000079. The Swedish schizophrenia case-control study had been described in an earlier publication [103], and I acquired processed VCFs for this data set via dbGaP authorized access (Accession: phs000473.v1.p1). A total of 2,536 schizophrenia cases and 2,543 controls were available for analysis. The DDD study was designed to further our understanding of broader developmental disorders while advancing clinical genetics practice in the UK. 4,281 children

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 2.1 Description of samples collections included as cases in the UK10K schizophrenia analysis. 1,353 cases remained after sample quality control.

with diverse, severe undiagnosed developmental disorders and their parents were exome sequenced to identify novel risk genes carrying variants of large effect. Patient recruitment, sample collection, sequencing production, and initial analysis of the dataset were described in detail in a previous publication [118]. The sequence data had been deposited into the EGA under study accession EGAS00001000775.

The SiSU project is an international collaboration generating whole genome and whole-exome sequence data from Finnish samples, and consists of a number of prospective and case-control cohorts, including the ENGAGE, FINRISK, Health 2000, and METSIM studies (http://www.sisuproject.fi/content/cohorts). The Northern Finnish 1966 Birth Cohort (NFBC) is a geographically based representative birth cohort including 96% (N = 12,068) of all live births in the two most northern provinces of Finland in 1966. The Northern Finnish Intellectual Disability Cohort (NFID) is an ongoing sample collection of individuals who have been diagnosed with ICD-10 diagnosis of intellectual disability or specific developmental disorder of speech and language of unknown etiology (ICD-10 codes: F70-F79 and F80-F89). The current sample includes 324 patients and their first-degree family members ($N = 631$, 92 full trios) with GWAS and WES data available. Combined, 5,720 Finnish exomes from the SiSU project were available for 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 2.2 Description of samples collections included as controls in the UK10K schizophrenia analysis. 4,769 controls remained after sample quality control.

Informed consent was obtained for all samples. Further information is available at http://uk10k.org/, http://www.ddduk.org, http://www.intervalstudy.org.uk/, and http://www.sisuproject.fi/.

2.3 Sequence data production

2.3.1 Sample preparation

DNA samples in the UK10K, DDD, and INTERVAL studies were sequenced at the Wellcome Trust Sanger Institute (Hinxton, Cambridge). One to three micrograms of DNA was sheared to $\sim$100 to 400 base pairs using either a Covaris E210 or LE220 machine (Covaris, Woburn, MA, USA), and processed using Illumina paired-end DNA library preparation. Three different captures were used to capture targeted coding regions: an expanded custom Agilent SureSelect Human All Exon v.3 capture with custom ELID C0338371 in the UK10K project, the Agilent SureSelect Human All Exon v.3 Kit (ELID S02972011) in the DDD study, and the Agilent SureSelect Human All Exon v.5 kit in INTERVAL study. All libraries were subsequently sequenced on Illumina HiSeq 2000 with 75 base paired-end reads in multiple batches according manufacturer's protocol over the duration of each project.

2.3.2 Alignment and BAM processing

Sequencing reads that failed quality control (QC) were first removed using the Illumina GA pipeline. Remaining raw reads were mapped to the reference genome (GRCh37 in UK10K; GRCh37_hs37d5 in DDD and INTERVAL studies) using BWA (v0.5.9-r16 in UK10K; v0.5.10 in DDD and INTERVAL) [113], and duplicate fragments were marked using Picard (v1.36 in UK10K; v1.98 in DDD; v1.114 in INTERVAL) [120]. GATK (version 1.1-5-g6f43284 in UK10K; version 3.1-1-g07a4bf8 in DDD; version 3.2-2-gec30cee in INTERVAL) was used to perform local realignment around indels [115], and recalibrate base qualities in each sample BAM. I applied VerifyBamID (v1.0) to estimate the Freemix value, which is representative of the contamination fraction in our sequence data [121]. I used the recommended thresholds for contamination, and removed samples if they had Freemix score ≥ 0.03. 31 samples or 2% of the UK10K data set were excluded, while 201 samples or 4.5% of the INTERVAL data set were excluded. We were unsure if the excess contamination in the INTERVAL study occurred during sample extraction, preparation, or sequencing.

2.3.3 Variant calling

I first called variants in individual samples using GATK Haplotype Caller (version 3.2-2-gec30cee). All samples were merged into random batches of 200 using CombineGVCFs, and then joint-called using GenotypeGVCFs at default settings [115, 122]. Because three different exome captures had been used, variant calling was performed on the union of Agilent v.3 and v.5 baits with 100 base pairs of flanking sequence. I subsequently ran the GATK VQSR on all GENCODE coding variants using default settings. This joint calling protocol was suggested by the GATK development team for the production of large sequencing data sets.

2.4 Variant calling and quality control across capture and batch

2.4.1 Adjusting for differences between capture and batch

The sequence data for individuals of UK ancestry was generated at the Wellcome Trust Sanger Institute using the same Illumina sequencing platform and some version of the SureSelect Human All Exon v.3 or v.5 captures. However, substantial differences exist between the exome captures, and this must be carefully adjusted for if samples were to be jointly analysed in a case-control framework. The v.5 capture improved coverage across the entire exome by shifting problematic coding baits into the intronic region and excluding a small percentage of repetitive and problematic genes. Because of this, the v.3 and v.5 captures shared only 77% of their targeted regions, and a simple intersection could not be used to prioritise genomic regions for a joint analysis. To best harmonize calls across projects, I first re-called samples together using a common calling pipeline at the union of both Agilent captures with 100 bp of flanking sequence. Instead of calculating coverage at v.3 and v.5 captures, I calculated per-sample read depth at all coding exons defined by GENCODE version 19 to evaluate differences in coverage and sequence quality [123]. From these data, I identified a set of well-behaved coding regions with sufficient coverage across batches and captures for subsequent QC and analysis.

In Figure 2.1, the v.5-captured samples (INTERVAL) had lower read depth across the entire exome, but covered a larger percentage of coding regions than in earlier v.3 captures (DDD and UK10K). The samples in the UK10K study were divided into two batches, reflecting a known chemistry change that occurred early in the project. DDD exomes more closely resembled the UK10K v.3 samples in regional coverage but clear differences still

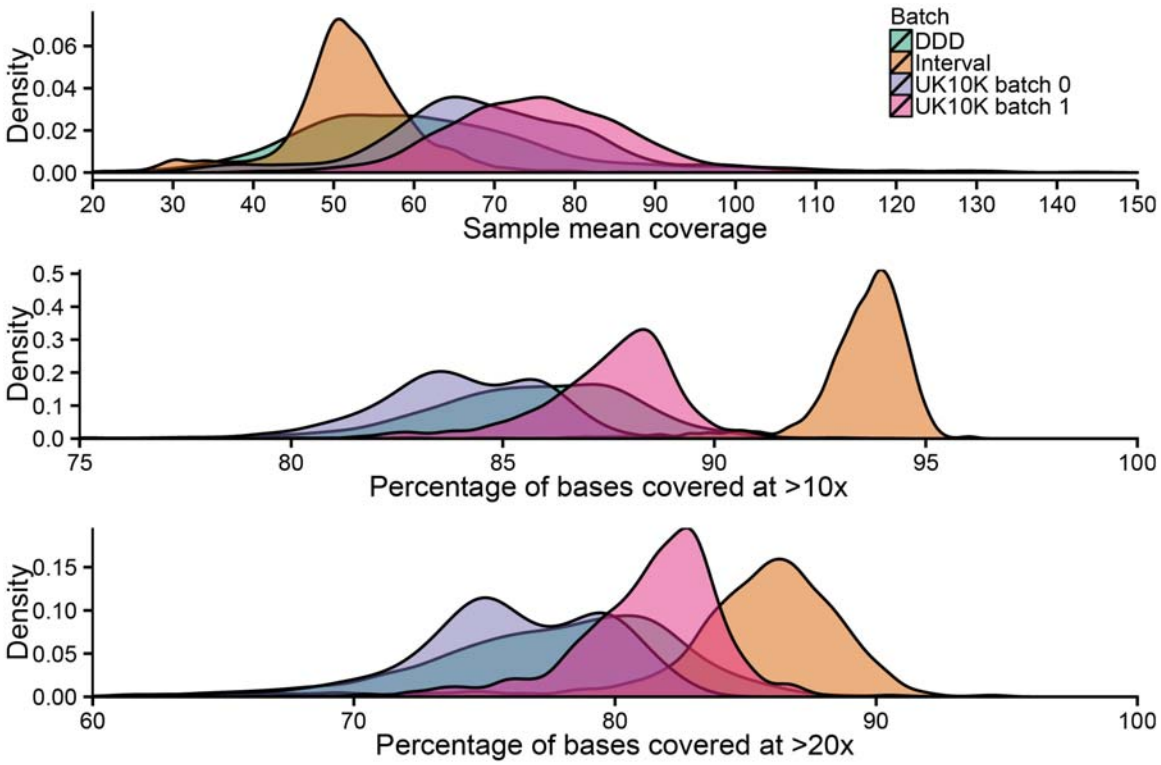

Fig. 2.1 **Density plots of sequence coverage in the UK10K, INTERVAL, and DDD datasets.** Per-sample sequence coverage was calculated and summarised from exome sequencing data generated in the UK10K (N = 4,734 in batch 0, and N = 562 in batch 1), INTERVAL (N = 4,502), and DDD (N = 1,972) datasets. The UK10K dataset was separated into two sequencing batches. Top: sample mean coverage; Middle: percentage of GENCODE v19 coding bases covered at 10× or more in each sample; Bottom: percentage of GENCODE v.19 coding bases covered at 20x or more in each sample.

existed between the v.3 and custom v.3 capture. Since all schizophrenia cases were sequenced using the v.3 capture, I have less power to detect rare variant associations in regions where this capture has limited coverage. I restricted our analysis to variants with a read depth of $7\times$ or more in at least 80% of samples in each of the four batches (UK10K batch 0, UK10K batch 1, DDD, and INTERVAL). For a more stringent filter, I identified exons that were covered at $10\times$ or more in at least 80% of samples in each batch for a total of 28.5 Mbs. By applying these filters, I excluded regions that were not covered with sufficiently high depth in our v.3-captured cases, or were not targeted in our v.5-captured controls by design.

2.5 Sample-level quality control for case-control analysis

The combined case-control data set consisted of individuals recruited from three countries: the UK, Sweden, and Finland. The UK and Finnish cases were recruited as part of the UK10K project, and the Swedish individuals were recruited in an independent study. While cases were called with nationality-matched controls, each subgroup was processed and sequenced at a different location with different reagents, and had to be analysed separately to reduce the effects of possible confounders like population stratification. Because of this, I performed sample-level and variant-level quality control steps on each nationality separately, and describe these steps in detail below.

2.5.1 Sample-level QC in the UK10K-INTERVAL case-control data set

In the UK10K data set, we sequenced the exomes of 1,488 UK individuals with schizophrenia and 5,469 matched controls without a known neuropsychiatric diagnosis. After per-sample depth analysis, I removed 22 samples with low coverage ($\leq 75\%$ of the GENCODE v.19 coding region covered at $\geq 10\times$). I next identified high-quality LD-pruned SNPs to investigate familial relatedness, non-European population ancestry, and outlying heterozygosity rates in our data set. To acquire these variants, I extracted common SNPs (MAF $> 5\%$) that passed a stringent VQSR threshold (tranche sensitivity 99.0%), had missingness $< 3\%$, and Hardy-Weinberg equilibrium χ^2 P-values $> 1 \times 10^{-3}$ in the UK10K and INTERVAL sequencing batches. I merged this subset of samples and variants with the 1000 Genomes Phase III release, and retained 43,837 SNPs with MAF $> 5\%$ and missingness $< 3\%$ in the combined dataset. These variants were LD-pruned on PLINK v1.9 with parameters –indep-pairwise 50 5 0.2 while excluding extended regions of high LD (chr 6: 25,000,000-35,000,000, and chr 8: 7,000,000-13,000,000) [124]. After filtering, a total of 19,554 high-quality LD-pruned SNPs were available for analysis.

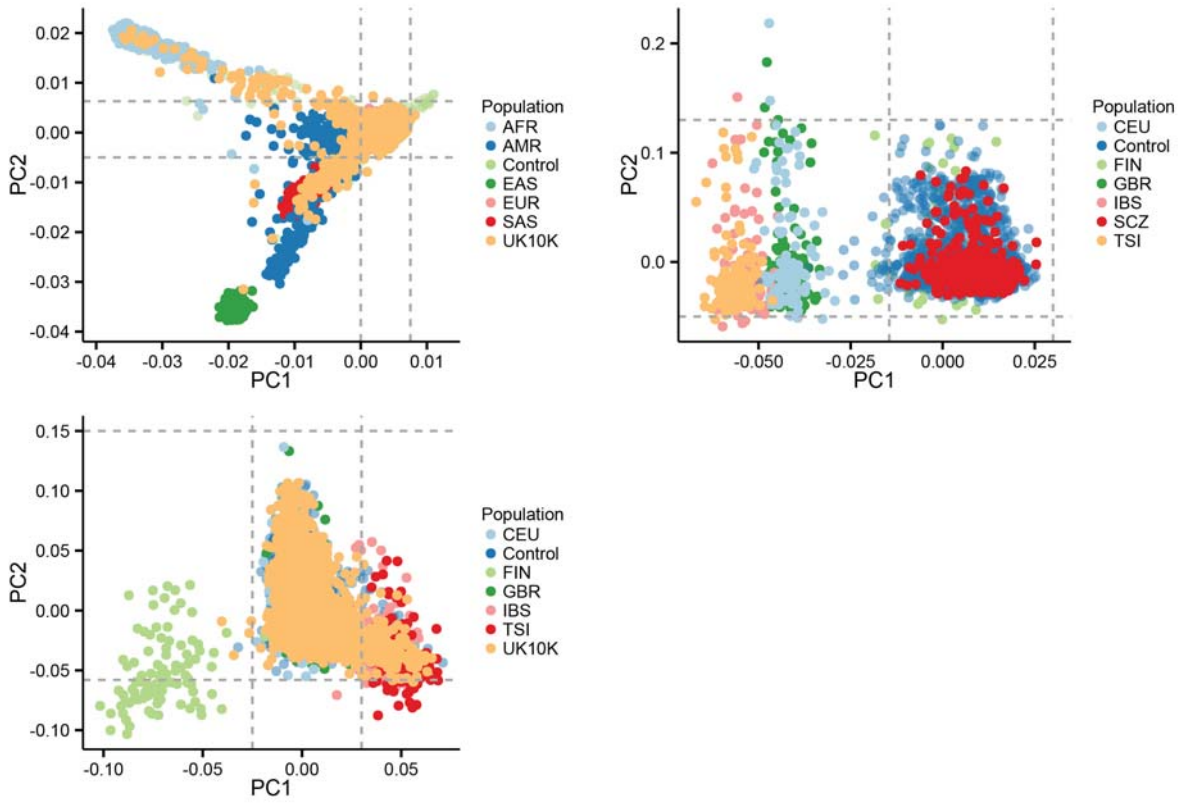

Fig. 2.2 **Principal components analysis of UK and Finnish samples in our UK10K schizophrenia dataset.** Principal components were estimated using 1000 Genomes samples, onto which I projected our cases and controls. I verified if samples had the same population ancestry (UK or Finnish) as reported in 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. **Top left**: Population structure of all UK10K samples, with 1000 Genomes populations used as bases. Samples bracketed by the dotted lines are of European ancestry; **Bottom left**: PCA plot of individuals of non-Finnish European ancestry in the UK10K dataset with 1000 Genomes European populations used as bases. Samples not within the UK cluster (bracketed by the dotted lines) were excluded from analysis; **Top right**: PCA plot of individuals of Finnish ancestry in the UK10K dataset. Samples not in the Finnish cluster (bracketed by the dotted lines) were excluded from analysis. The three-letter symbols describing each population originate from nomenclature in the 1000 Genomes Project. UK10K: samples in our case-control study; SCZ: schizophrenia cases; Control: controls from our study.

Principal components analysis (PCA) was performed using PLINK v1.947 with 1000 Genomes Phase III samples as reference populations. I observed that 407 individuals were of non-European ancestry (Figure 2.2). In a second PCA using only European populations as reference, I observed that our samples were predominantly of UK or North European ancestry, with a small number of cases more related to 1000 Genomes individuals from the Iberian peninsula (Figure 2.2). I retained these individuals, but noted that they may have to be grouped into a separate batch or excluded in later analyses. I estimated kinship coefficients between each sample pair using KING v1.448 [125], and removed 39 duplicate samples and 68 samples with abnormal values likely due to some level of contamination. Individuals in first, second, and third-degree relationships were identified, and 190 samples were selectively removed until the maximum pairwise kinship coefficient within the cohort is 0.09375. In all, 826 samples were removed during QC, resulting in a final cohort of 6,122 UK samples (1,353 cases and 4,769 controls).

2.5.2 Sample-level QC in the Finnish and Swedish case-control data sets

In the UK10K data set, we sequenced the exomes of 399 Finnish individuals with schizophrenia and 2,116 matched controls, and performed variant calling using the GATK pipeline at the Broad Institute (Cambridge, MA). After obtaining unprocessed VCFs containing these samples, I excluded 16 samples with lower-than-expected coverage, and determined that all samples within the Finnish data set were of either non-Finnish European or Finnish ancestry (Figure 2.2). A more detailed projection using 1000 Genomes European individuals revealed that 27 samples were more closely related to non-Finnish Europeans in ancestry, and I excluded these 27 individuals from further analysis. From relatedness analysis, I excluded 67 samples. In all, 103 samples were removed during QC, resulting in a final data set of 2,412 samples (392 cases and 2,020 controls). A similar analysis within the Swedish case-control data set determined that all samples were of non-Finnish European or Finnish ancestry. I excluded 17 samples due to relatedness, resulting in a final data set of 5,073 individuals (2,519 cases and 2,554 controls).

2.6 Variant filtering in case-control data sets

2.6.1 Variant filtering in the UK10K-INTERVAL data set

Standard protocol for variant filtering recommends the use of GATK VQSR for calculating the probability that a variant is real, and selecting a threshold that maintains a desired sensitivity for true variants. The VQSR model trains on the annotation metrics (mapping quality, strand bias, quality by depth) of validated variants from the HapMap project and the 1000 Genomes Project to classify the remaining variants. However, recent studies have suggested that VQSR is less effective in filtering ultra-rare variants, especially those that are seen only once (singletons) or twice (doubletons) in the data set [126]. Notably, VQSR does not filter individual genotypes, which allows low-quality calls to be inaccurately retained if that site on average passes VQSR filtering. The inability to remove these low-quality genotypes within variable sites adds unnecessary noise in downstream analyses. However, recommended thresholds for filtering individual genotypes have not been established.

To complement the GATK filtering step, I empirically derived site and genotype filters by evaluating the sensitivity and specificity of different thresholds using a training set consisting of real rare variants and sequencing artefacts. First, I assumed that rare and singleton ExomeChip genotype calls in 295 UK10K cases (83 in batch 0, 212 in batch 1) represented real variants, and evaluated concordance with corresponding calls in our sequence data to assess sensitivity. Second, I identified inherited variants unique to parent-proband pair (inherited doubletons) and Mendelian inheritance inconsistent variants within DDD parent-proband trios to evaluate SNP and indel filtering thresholds. I computed the percentage of inherited variants retained and putative *de novo* variants removed at various thresholds to evaluate the effectiveness of our variant filtering. Using these data, I explored genotype thresholds across a number of variant and genotype-level metrics, including VQSLOD score, reference allele read depth (DP0), alternate allele read depth (DP1), allelic balance (AB), genotype quality (GQ), and mean genotype quality (GQ_MEAN). Variant thresholds were determined for SNPs and indels separately. In summary, I used rare array-based variants and rare Mendelian inheritance consistent (truth sets) and inconsistent variants from trios (false set) to calibrate variant filtering thresholds.

Variant filtering thresholds for SNPs

Applying the following filters achieved a reasonable compromise between sensitivity and specificity within our case-control data set (Figure 2.3):

- Exclude variants outside the VQSR tranche with 99.75% sensitivity

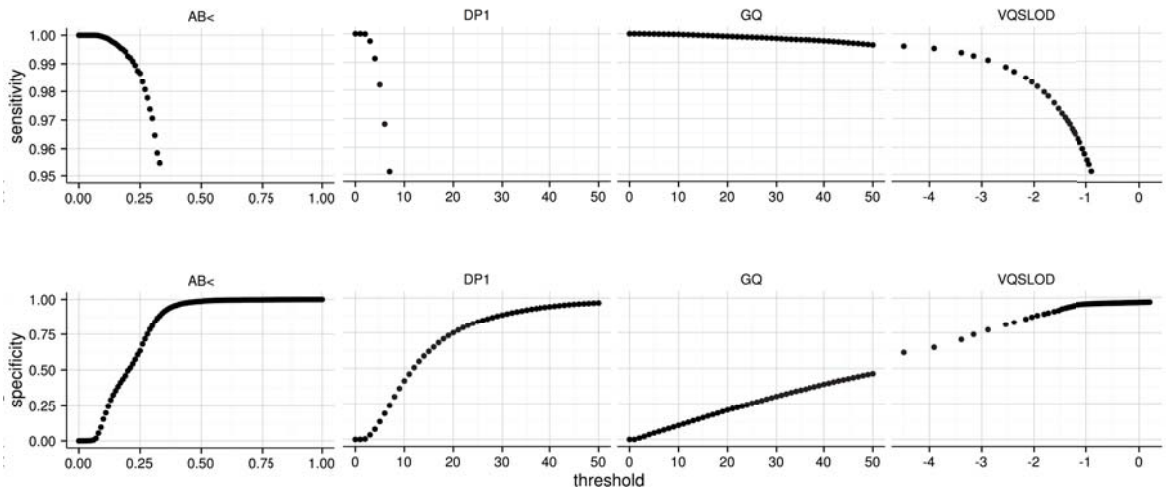

Fig. 2.3 **The evaluation of different variant filtering thresholds using rare DDD inherited variants and Mendelian inconsistent variants as a testing set.** I evaluated the sensitivity and specificity across a range of thresholds for each variant and genotype-metric. AB<: retain variants with allelic balance greater than this threshold; DP1: retain variants with alternate allele read depth greater than this threshold; GQ: retain variants with genotype quality greater than this threshold; VQSR: retain variants with a GATK variant recalibration scores greater than this threshold.

- Exclude variants with mean GQ < 30
- Exclude genotype calls with GQ < 30
- Exclude genotype calls with DP1 < 2
- Exclude genotype calls with AB < 0.2 and AB > 0.8

Using these thresholds, I removed 95.63% of all Mendelian inconsistent genotype calls while retaining 98.38% of all doubleton inherited variants. In the ExomeChip data set, I retained 99.45% of variants seen only once in the UK10K samples, and 99.62% of all heterozygote calls. While GATK recommended a more conservative VQSLOD score threshold (either VQSRTranche99.50 or VQSRTranche99.0), I found that a less stringent VQSR filter combined with genotype-level thresholds retained a larger percentage of rare inherited variants while attaining reasonable specificity. If VQSR were applied without genotype-level filters, only 40.8% of all Mendelian inconsistent genotype calls would be excluded were I to maintain a comparable sensitivity of 98% for doubleton inherited variants. I also removed SNPs with missingness greater than 20%, and tested SNPs for deviation from Hardy-Weinberg equilibrium within each sequencing batch (UK10K batch 0, UK10K batch 1, and INTERVAL) and within the entire data set. The Hardy-Weinberg filter addressed mapping issues that arose from differences in exome baits or decoy sequences used during alignment:

mismapped variants often are seen only as heterozygotes in one batch and homozygotes in another. Any variant that deviated from Hardy-Weinberg equilibrium with χ^2 *P*-values of $< 1 \times 10^{-8}$ in any batch or in the entire data set was excluded. Finally, I excluded variants that resided in low-complexity regions, the 2% of the genome highly enriched for repetitive sequences in which alignment and variant calling is more difficult (see Heng *et al.* [117] for a more precise definition and motivation for its use). At each stage of filtering, I reported the per-sample transition-transversion rate (TiTv), the number of heterozygote calls, the number of non-reference homozygous calls, and the number of variants observed only once within the UK10K-INTERVAL call set (Figure 2.4, 2.5). The variant metrics appeared comparable across the four batches, and the mean sample TiTv was ∼3.26, the expected rate for coding SNPs in European populations.

Variant filtering thresholds for indels

Using the same approach described above for SNPs, I found that the following filters achieved a reasonable compromise between sensitivity and specificity for indel discovery within our case-control data set:

- Exclude variants outside the VQSR tranche with 99.50% sensitivity
- Exclude variants with mean GQ < 90
- Exclude genotype calls with GQ < 90
- Exclude genotype calls with DP1 < 2
- Exclude genotype calls with AB < 0.25 and AB > 0.8

Using these variant and genotype-level thresholds, I removed 92.35% of all unfiltered Mendelian inconsistent indel calls while retaining 93.60% of all doubleton inherited indels. Applying VQSR alone was not sufficient to acquiring a clean indel set: even at a stringent VQSLOD threshold of -0.3151 (VQSRTrancheINDEL0.00to99.00), I only achieved specificity of 40.72% for Mendelian inconsistent indels. I also removed indels with missingness greater than 20%, and those that deviated from Hardy-Weinberg equilibrium with χ^2 *P*-values of $< 1 \times 10^{-8}$. I removed indels that resided in low-complexity, highly repetitive regions (defined in the previous section) that could not be appropriately aligned using short-read technology. Lastly, I excluded all indels that have more than two alternate alleles, or were clustered within 3 bp of another indel. Following these indel filtering steps, the number of indels and frameshift:inframe ratio appeared comparable across all batches (Figure 2.6).

From previous studies of parent-proband trio studies, we expected to find one coding *de novo* mutation per proband [79, 80]. In our DDD trio data set, we observed 92 *de novo* SNVs

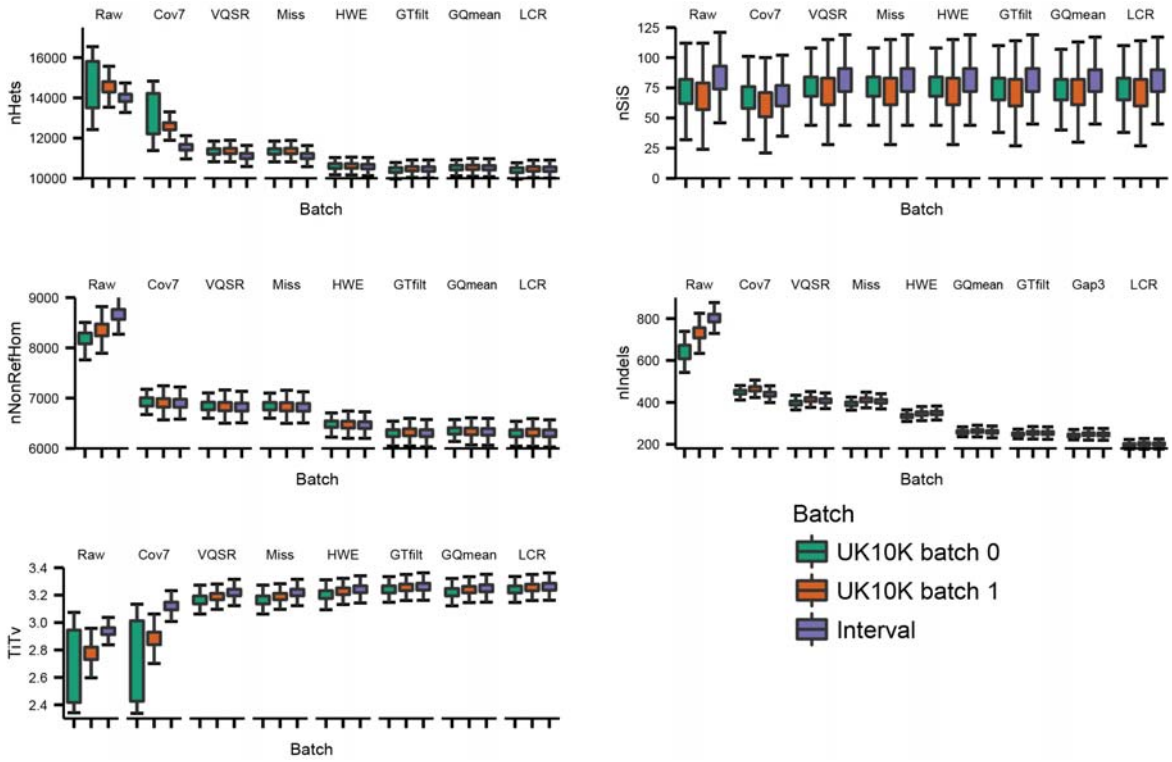

Fig. 2.4 **Variant metrics in the UK10K and INTERVAL datasets after each variant filtering step.** Box plots of per-sample heterozygote count (nHets), non-reference homozygote count (nNonRefHom), TiTv (TiTv), number of singletons (nSiS), and number of indels (nIndels) following each variant QC step. Variant metrics were summarised across all samples in the UK10K and INTERVAL datasets. Raw: no variant QC steps applied; Cov7: restricting to variants with at least 7× 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.

and 12 *de novo* indels per proband prior to variant filtering, and 4 *de novo* SNVs and 0.92 *de novo* indels per proband after variant filtering. The observed *de novo* mutation rate in our data set still exceeded the expected rate of mutation described in previous studies, suggesting that our variant QC was not sufficiently strict to over-filter genuine *de novo* events while vastly reducing the number of false positives.

2.6.2 Variant filtering in the Finnish and Swedish data sets

In the Finnish data set, SNPs and individual genotype calls were excluded according to the following criteria: VQSLOD < -2.6557 (VQSRTrancheSNP99.75), GQ < 30, or GQ_MEAN < 30. Indel sites and genotypes were excluded according to the following criteria: VQSLOD < -0.2731 (VQSRTrancheIndel99.50), GQ < 90, or GQ_MEAN < 30. In addition, I removed variants with missingness greater than 20%, or if they deviated from Hardy-Weinberg equilibrium with χ^2 *P*-values of $< 1 \times 10^{-8}$. All variants within low-complexity regions were excluded. I also removed all indels that have more than two alternate alleles, or were located within 3 base pairs of another indel. After variant and genotype-level QC, the sample TiTv and frameshift:inframe ratio was $\sim$3.29 and $\sim$1.01 respectively, which was comparable across batches of the Finnish data set and with the UK10K-INTERVAL call set (Figure 2.5, 2.6).

I was unable to acquire raw BAMs for the Swedish data set to re-call and perform QC from scratch. However, the Swedish data set as provided already had very stringent filters applied during a previous analysis, and I analysed the dataset with little additional QC. Variant sites and genotypes were filtered out if the Hardy-Weinberg equilibrium χ^2 *P*-values $< 1 \times 10^{-8}$, missingness > 20%, or if they reside within in low-complexity regions. After variant and site QC, the sample TiTv and frameshift:inframe ratio was $\sim$3.28 and $\sim$1.15 respectively, which was comparable across batches and with the UK10K-INTERVAL call set.

2.7 Comparison of population genetics metrics across data sets

Following sample and variant QC, 6,122 UK samples (1,353 cases and 4,769 controls), 2,412 Finnish samples (392 cases and 2,020 controls), and 5,073 Swedish samples (2,519 cases and 2,554 controls) were available for analysis. Variant counts and population genetic metrics between data sets and sequencing batches were harmonized: the sample TiTv (mean $\sim$3.25) and the frameshift:inframe ratio were comparable across all populations and batches

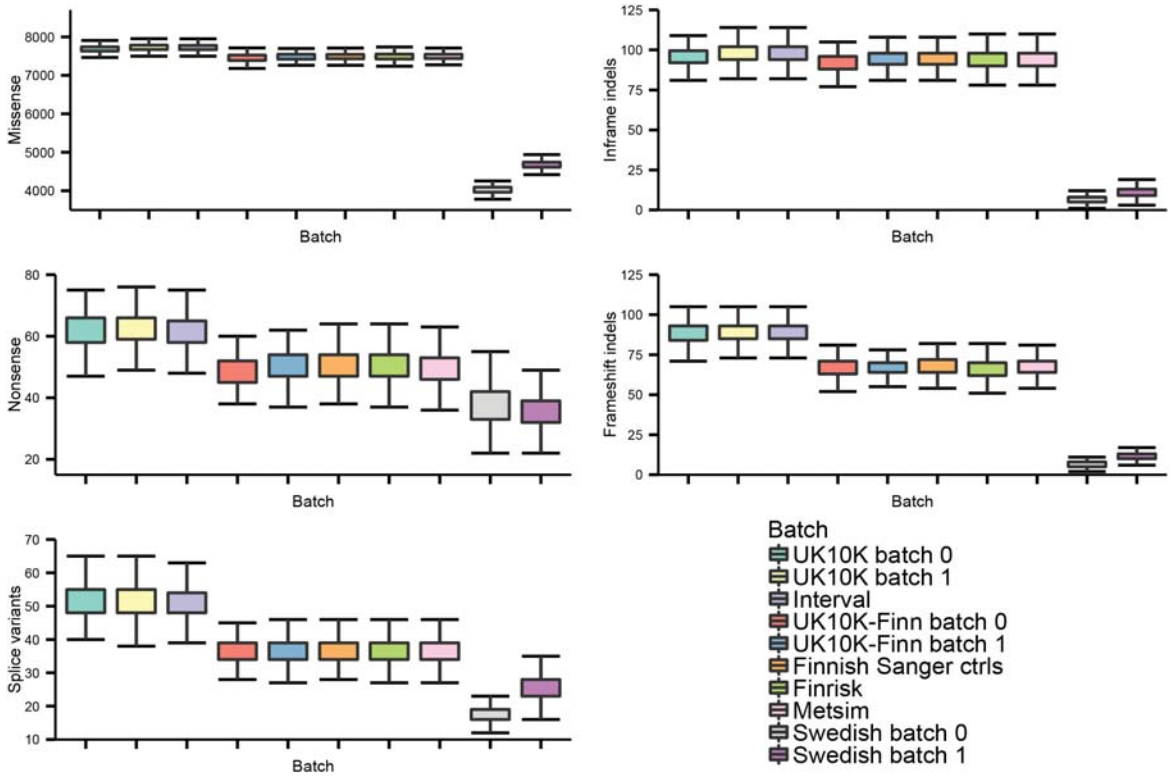

Fig. 2.5 **Variant counts summarised according to variant class and sequencing batch in the UK10K, INTERVAL, Finnish, and Swedish datasets.** Box plots of per-sample variant counts in the UK10K, INTERVAL, Finnish, and Swedish datasets. All samples included in our meta-analysis are represented in the figure. The UK10K datasets was sub-divided according to sequencing batches (batch 0 and batch 1), and sample ancestry (UK and Finnish). The Finnish control datasets was separated by study of origin (Metsim, Finrisk, and Sanger controls). The Swedish case-control dataset 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.

(Figure 2.6). However, I still observed some differences between variant counts between the UK, Finnish, and Swedish data sets (Figure 2.5). The UK, Finnish, and Swedish samples were independently produced and called at different sequencing centres, and the discrepancy in variant counts likely reflected differences in capture, sequencing batch, calling procedure, and quality control. In particular, the Swedish data set we acquired from dbGAP underwent extremely stringent variant filtering, and had per-sample variant counts nearly half of that observed in the UK10K-INTERVAL data set and the 1000 Genomes Phase III data set. These differences would confound rare variant tests and need to be explicitly corrected for. In subsequent analyses, I adjusted for between-population differences by treating them as separate analytical groups. More importantly, cases and controls within each population group appeared to be well-matched, and this was reflected in the null statistics of subsequent variant and gene-based analyses.

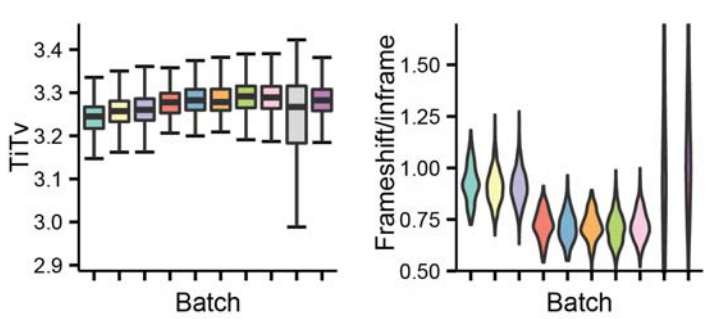

Fig. 2.6 **Distributions of TiTv and frameshift-inframe ratios in the UK10K, INTERVAL, Finnish, and Swedish datasets.** Here, I have a box plot of sample TiTv (left) and violin plot of sample frameshift-to-inframe ratio (right) in the UK10K, INTERVAL, Finnish, and Swedish datasets. All samples included in our meta-analysis are represented in the figure. See 2.5 for the legend, and a description of each batch and sub-study. Following sample and variant QC, the per-sample transition-to-transversion ratio was comparable between all populations (mean $\sim$ 3.25).

2.8 Systematic annotation of coding variants

I used the Ensembl Variant Effect Predictor (VEP) version 75 to annotate coding variants with GENCODE version 19 transcripts as reference [127]. VEP plugins were used to apply *in silico* classifiers to missense variants, such as PolyPhen, SIFT, and CADD [128–130]. For each variant, I assigned a functional consequence on a per-gene basis, aggregating all transcript-level annotations and retaining only the most severe consequence. Coding variants were assigned into the following functional categories:

1. Loss-of-function or disruptive (LoF) variants
    - Frameshift
    - Stop-gained
    - Splice acceptor and donor variants
2. Initiator codon variants
3. Inframe deletion or insertions
4. Missense variants (mis)
5. Synonymous variants

Following other rare variant studies [94, 103, 105], I stratified our analyses into two functional classes: 1. LoF variants, 2. missense or initiator codon variants.

2.9 Evaluating the effectiveness of existing *in silico* predictors of pathogenicity

The use of variant annotation tools to prioritise coding variation has helped increase statistical power for gene discovery [88, 94]. Most variants identified in the coding region reside in the rarer end (MAF $< 0.1\%$) of the allele frequency spectrum (AFS) [78]. If the predicted functional consequence of variants were disregarded, a simple comparison of allele counts between cases and controls would be diluted by large numbers of non-functional variants [88]. Functional annotation tools intend to accurately distinguish the pathogenic, disease-causing variants from neutral polymorphisms, thus enriching our analyses on causal risk variants while decreasing the rate of background noise. However, over-filtering and removing true signals can have a detrimental effect on our power, especially when the allele counts of rare damaging variants are already low due to purifying selection. A delicate balance between specificity and sensitivity in annotating disease-causing variants is required to maximize our power in detecting true associations.

2.9.1 The interpretation of protein-coding consequences

In the simplest case, a variant is annotated as functional based on its effect on a protein product. A true loss-of-function (LoF) variant either drastically reduces levels of the gene product or disrupts a protein's ability to carry out key functions. This can be through truncations, aberrant splicing, shifts in coding sequences, and pre-mature stop codons. A

missense variant causes an amino acid substitution, which can lead to change in protein functionality. Many of these changes are benign and would not be subject to strong selection: missense variants may substitute amino acids without affecting charge and folding or disrupt a domain or peptide that is irrelevant to protein function. On other hand, some missense variants eliminate protein function by disrupt protein folding or modifying the charge of an active site. Thus, even if a variant labelled as missense or loss-of-function by VEP, additional information is needed to properly evaluate its pathogenicity.

2.9.2 A description of existing annotation tools

Because of this, a series of statistical tools have been developed to predict the pathogenicity of missense variants. These missense classifiers primarily differ in the statistical approach applied, the features inputed into the model, and the training and testing data set used for calibration and evaluation (Table 2.3). For instance, PolyPhen2 uses a Bayesian classifier to characterize missense variants based on structural information about the binding site, protein domains, contact with ligands, and subunit interactions [129]. All the calculated features are trained using a Bayes classifier on the HumDiv data set, a curated list of variants causing Mendelian disorders, and the Humvar data set, a more comprehensive list of risk alleles from UniProt [131]. SIFT, another popular tool, models function using a multiple sequence alignment of proteins, and determines which base mutations was most tolerated across similar proteins [128]. A SIFT score of 0.05 indicates the alternate allele was observed in 5% of all alignments and could be considered not as damaging. Other missense classifiers include LRT, MutationTaster, MutationAssessor, FATHMM, Radial SVM, MetaLR, GERP++, PhyloP, Condel2, and SiPhy [132, 133]. Some of these, like GERP++, and PhyloP, classify variants based on the degree of sequence conservation between species, while others, like Radial SVM and Condel2, are ensemble classifiers that integrate results from other tools to annotate variants. CADD differs in its approach completely by simulating its training set and comparing these randomly generated alleles to the set of derived alleles common between the human-chimpanzee ancestral genome [130]. A support vector machine with a large set of features, including SIFT and PolyPhen, was used to model the relative deleteriousness of all possible alleles across the genome. Unlike the other tools, CADD could annotate both coding and non-coding variants.

When evaluating these models, differences in statistical approach, input features, and training and testing data must be carefully considered to prevent issues of circularity and bias (Table 2.3). For example, MutationalTaster incorporates frequency information from 1000 Genomes when determining pathogenicity; a testing set consisting of rare damaging variants as pathogenic and common variants as benign would inflate the classifier's effectiveness.

Ensemble classifiers like CONDEL and Radial SVM incorporated MutationTaster, SIFT, and PolyPhen as features, and indirectly incorporated frequency information. Furthermore, only a few robust variant data sets exist for evaluating the effectiveness of each classifier, and many of these classifiers already use them for training. For instance, PolyPhen, CONDEL, Radial SVM, and MetaLR trained on the same set of curated coding variants provided by Uniprot, while others trained on the Human Gene Mutation Database (HGMD) database. Therefore, a new and wholly independent dataset is best suited for evaluating the performance of these *in silico* classifiers.

2.9.3 Strategy for evaluating variant annotation tools

I evaluated the effectiveness of available annotation tools for LoF and missense variants using a series of novel variant sets previously not used for classifier training. First, I used a clinician-curated set of variants from the DDD study. *De novo* and inherited variants were identified and validated in 1,133 affected probands, and variants disrupting known developmental disorder genes were manually curated to determine if these variants were pathogenic relative to the patient's phenotype. I used all clinically reportable variants as a truth set, and all rejected variants as a false set. For an additional truth set, I accumulated *de novo* mutations from 2,263 trios sequenced as part of the Autism Sequencing Consortium, and 2,500 trios sequenced in the Simon Simplex Collection. I identified all *de novo* missense variants disrupting autism risk genes from Sanders *et al.* [109] as another truth set.

The ExAC database contained coding variants from 60,706 unrelated individuals without severe paediatric diseases joint-called in a single pipeline [112]. It is important to note that this release of ExAC contained a number of individuals with psychiatric phenotypes. I assumed that the fraction of pathogenic variants in this data set was substantially lower than in the DDD and ASD studies, and used missense variants with MAF $< 1\%$ in ExAC as a false set. Finally, I re-annotated a large set of functional, protein-coding variants manually curated by Uniprot for an additional training set. This truth set consisted of variants described by Uniprot as disease-causing and our negative truth set were variants described as general polymorphisms. I applied the following *in silico* classifiers to the missense variants: PolyPhen2, SIFT, LRT, MutationTaster, MutationAssessor, FATHMM, Radial SVM, MetaLR, GERP++, PhyloP, Condel2, CADD, and SiPhy. Using causal variants identified in the DDD and ASD studies, I determined guidelines for prioritising variants in trio and case-control analyses.

Name	Method	Features	Training set	Classifies
CADD	Support vector machine.	Conservation metrics, functional genomic data, transcript information, and protein-level scores. 63 annotations in total.	Derived alleles common between the human-chimpanzee ancestral genome and simulated alleles.	All variants (SNPs and indels of all classes)
CONDEL2	Weighted average of other classifier scores.	Other variant classifiers, including PolyPhen2, Mutation Assessor, and SIFT.	HumVar and HumDiv variant databases.	Missense variants
FATHMM	Hidden Markov Model.	Alignment of homologous sequences across species, along with an overlaying of protein domain information.	HGMD and UniProt databases.	Missense variants
GERP++ RS	Maximum likelihood estimation and dynamic programming to define constrained elements.	Multiple alignment to study conservation across species. A comparative genomics approach.	Genomes of humans and other species. Primarily mammalian.	Missense variants
LR	Logistic regression.	11 other classifiers, including PolyPhen, SIFT, MutationAssessor, FATHMM.	UniProt database.	Missense variants
LRT	Likelihood ratio test.	Multiple alignment of human genes with other species to identify conserved regions.	Genomes of 32 vertebrate species.	Missense variants
MutationAssessor	Computing a relative conservation score, and calibrating on the training data.	Multiple alignment of gene and protein families within and between species.	UniProt database.	Missense variants
PhyloP	Unsupervised phylogenetic methods to estimate probabilities.	Sequence alignment across species to identify departures from neutral rate of substitution.	Genome-wide alignment of36 species.	Missense variants
PolyPhen2 HDIV	Naive Bayes clasifer.	Eight sequence-based and three structure-based predictive features from multiple sequence alignment.	HumDiv database.	Missense variants
Polyphen2 HVAR	Naive Bayes clasifer.	Eight sequence-based and three structure-based predictive features from multiple sequence alignment.	HumVar database.	Missense variants
RadialSVM score	Support vector machine.	11 other classifiers, including PolyPhen, SIFT, MutationAssessor, FATHMM.	UniProt database.	Missense variants
SIFT	Scores computed from sequencing alignment conservative. A scaled probability calculated for each position.	Sequence alignment across species to identify evolutionary conservation of amino acids.	Genomes of humans and other species.	Missense variants
SiPhy	Hidden Markov Model.	Inferring site-specific substitution biases directly from sequence alignments. Conservation-based method.	Genomes of humans and other species.	Missense variants
VEST3	Random forest.	86 features, including amino acid properties and conservation scores.	HGMD database and common variants in the NHLBI exome sequencing project.	Missense variants

Table 2.3 Description and summary of statistical tools developed to predict the pathogenicity of coding variants. The statistical method, features, and training set of each missense classifier were described. More information on these tools could be found in the annotation database dbNSFP [132, 133].

2.9.4 Preparation of annotation files

I used the Annovar tool and the dbNSFP v2.7 database [132, 133] to annotate all missense variants with the following classifiers: PolyPhen2, SIFT, LRT, MutationAssessor, FATHMM, Radial SVM, MetaLR, GERP++, PhyloP, CADD, and SiPhy. I used VEP to annotate variants with CADD and Condel2 scores. Condel2 scores were separately downloaded from FannsDB and parsed to be compatible with VEP.

2.9.5 Classifiers display variable performance depending on test data

First, I tested the effectiveness of the 14 classifiers in identifying pathogenic and benign variants in the UniProt data set. I found that ensemble classifiers, such as LR score, Radial SVM, VEST3, and CONDEL, had the greatest area under the curve (AUC), and reached a sensitivity and specificity of just under 90% (Figure 2.7). These classifiers used PolyPhen, SIFT and conservation scores as features to train more flexible statistical methods like the random forest and support vector machine. This was followed by CADD and PolyPhen that reached a sensitivity and specificity of just under 80%. SIFT and annotation methods based on conservation did not perform as well as the other classifiers.

I next evaluated the classifiers using pathogenic *de novo* mutations from the DDD and ASD studies as positive testing data. I first used UniProt benign polymorphisms as negative testing data. As seen in Figure 2.8, I found that the missense classifiers performed substantially worse when classifying *de novo* mutations when compared to UniProt pathogenic variants. None of the classifiers had a discrimination threshold that simultaneously achieved a sensitivity and specificity of greater than 82%. Ensemble classifiers like LR score and Radial SVM still outperformed the remaining classifiers. Along with CADD, these more flexible methods outperformed PolyPhen, SIFT, and other conservation-based annotations. Finally, I used ExAC missense variants with MAF $< 1\%$ as an alternate negative testing data set, while still using diagnostic *de novo* mutations as the positive testing set. Again, the ensemble classifiers massively outperformed the remaining annotation tools, with LR score and Radial SVM leading with the highest AUC (Figure 2.9).

I attempted to identify optimal discrimination thresholds for each missense classifier using Youden's J-statistic. Surprisingly, the optimal discrimination threshold for each annotation tool was highly sensitive to the testing data set used. For LR score, the optimal threshold for the Uniprot testing data set was 0.28, and resulted in a sensitivity and specificity of 0.92 and 0.87 respectively. However, this same threshold resulted in a sensitivity and specificity of 0.68 and 0.98 when classifying *de novo* mutations and ExAC common variants. The optimal threshold for this data set was instead 0.037, which yielded a sensitivity and specificity of

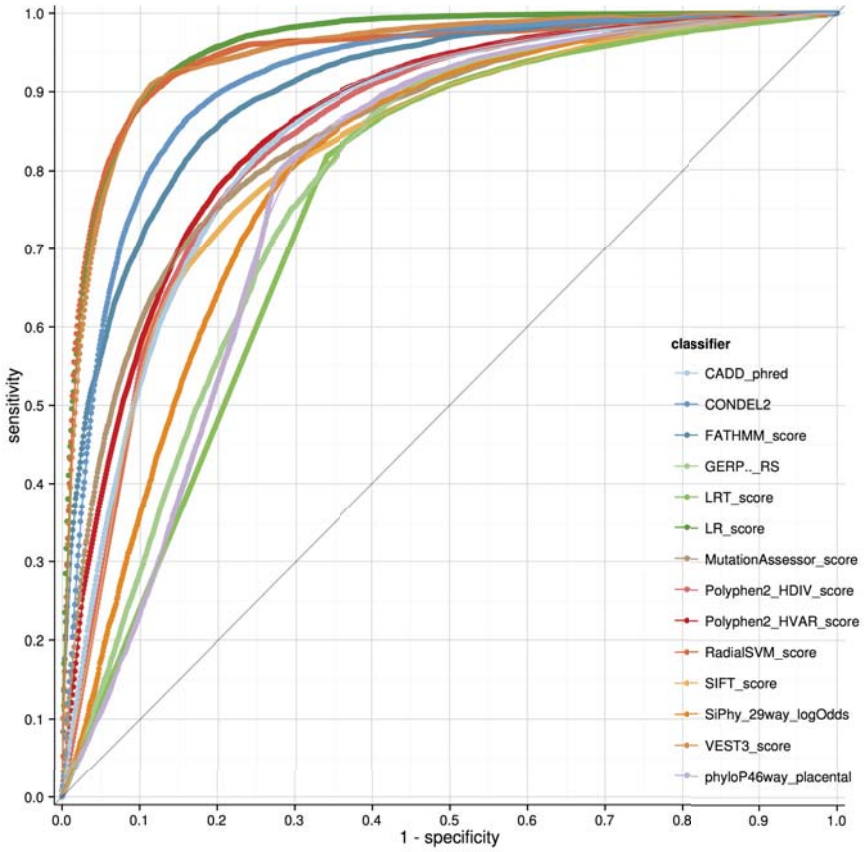

Fig. 2.7 **ROC curve evaluating the performance of missense classifiers on UniProt pathogenic-benign variants.** UniProt pathogenic variants were used as the positive testing set, while UniProt polymorphic (benign) variants were used as the negative testing set. The sensitivity and 1− specificity was plotted at various threshold settings for each classifier.

0.96 and 0.94 respectively. Unfortunately, this pattern was also observed for Radial SVM and VEST3. While the ensemble classifiers appear to outperform the other classifiers, identifying the discrimination thresholds at which this generally occurs is not at all straightforward. While it is difficult to explain this variability in optimal thresholds, these ensemble classifiers directly or indirectly incorporate allele frequency as a feature in their models, and this may lead to biases in evaluation depending on the proportion of common and rare variants in the testing data sets.

Ultimately, I selected CADD > 15 to classify missense variants as damaging in our case-control analysis. While CADD had an AUC lower than LR pred and Radial SVM, it had optimal discrimination thresholds that were highly comparable across the different testing data sets. The sensitivity and specificity at these optimal thresholds did not vary significantly between different testing data sets. For *de novo* — ExAC common variant data set, the optimal threshold was 14.1, the sensitivity was 0.84, and specificity was 0.86; for the *de novo* — Uniprot benign data set, the optimal threshold was 16.3, with a sensitivity of 0.76, and specificity of 0.79; for the Uniprot pathogenic-benign data set, the threshold was 15.4, with a sensitivity of 0.82 and specificity of 0.75. CADD performed robustly across each of our testing data sets, and its performance was superior to both PolyPhen, SIFT, and the other conservation scores. Finally, its continuous score extended to synonymous, splice, LoF, intronic, and intergenic variants, which may be useful for analyses that extended beyond missense variants.

2.9.6 A comparison of annotation approach with other whole-exome sequencing studies

While the annotation approach described here does not differ drastically with approaches used by other whole-exome sequencing studies, it does differ in some notable aspects, which I discuss here. Nearly all studies grouped functional coding variants into two categories for analysis: loss-of-function variants (defined as nonsense, essential splice, and frameshift variants), and nonsynonymous variants (defined as missense and inframe indels) [98, 103, 105, 118, 94, 112, 134, 135]. Variants were annotated based on the most severe consequence on any transcript. Where studies generally differed was in the tool used to annotate variants, the transcript reference database, and the *in silico* classifiers used to prioritise pathogenic missense variants. For instance, Purcell *et al.* and Fromer *et al.* used PLINK/SEQ to annotate variants according to the RefSeq transcript definitions; Do *et al.* and De Rubeis *et al.* used SnpEff and also according to RefSeq transcripts; the DDD and ExAC studies used VEP according to Ensembl GENCODE transcript definitions; Genovese *et al.* used SnpEff

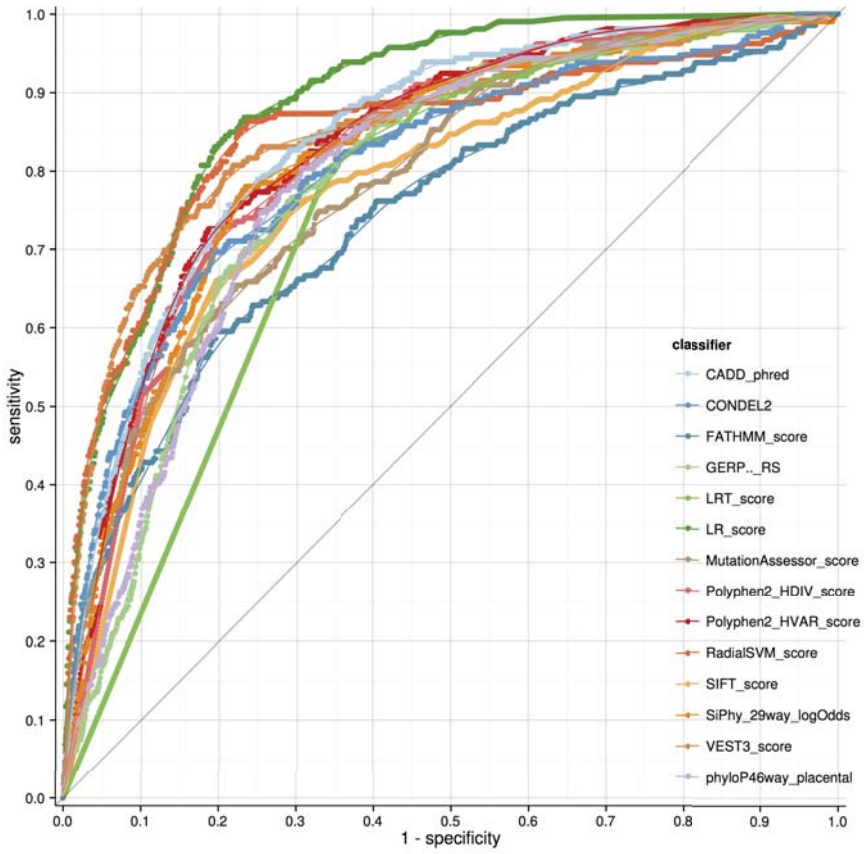

Fig. 2.8 **ROC curve evaluating the performance of missense classifiers on pathogenic *de novo* mutations and benign variants from UniProt.** Pathogenic *de novo mutations* from the DDD and autism studies were used as the positive testing set, while UniProt polymorphic (benign) variants were used as the negative testing set. The sensitivity and 1 - specificity was plotted at various threshold settings for each classifier.

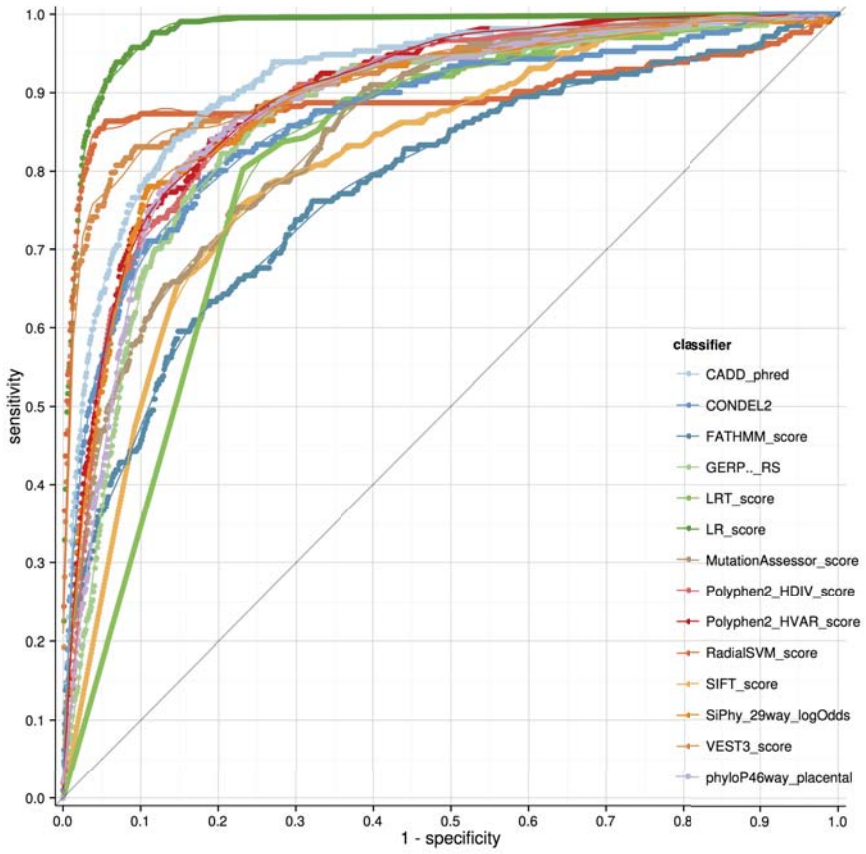

Fig. 2.9 **ROC curve evaluating the performance of missense classifiers on pathogenic *de novo* mutations and ExAC missense variants with MAF** $> 1\%$**.** Pathogenic *de novo mutations* from the DDD and autism studies were used as the positive testing set, while ExAC missense variants with MAF $> 1\%$ variants were used as the negative testing set. The sensitivity and 1 - specificity was plotted at various threshold settings for each classifier.

according to GENCODE transcripts, and Fuchsberger *et al.* used a combination of annotation tools that included SnpEff and ANNOVAR according to GENCODE transcripts. In many of these studies, little justification was provided for the choice of annotation tool and transcript database. Notably, when 80 million variants were annotated using multiple approaches, only a 83% agreement was observed in the annotation for exonic variants when using the RefSeq or Ensembl GENCODE transcript sets as references [136]. In the end, I decided to annotate variants using VEP with GENCODE transcripts as reference because a number of data sets and resources used in our analyses, such as ExAC database, GTeX database, and the DDD study, followed this approach. In addition, as seen in the following section, the GENCODE transcript reference contained a more complete set of coding genes, which permitted the analysis of an additional 1,067 protein-coding genes. However, as discussed in [136], variant annotation remained an unsolved problem, and no single annotation software or transcript set was identified as directly superior to the others.

Whole-exome sequencing studies also differed in the tools used for classifying missense variants as pathogenic and benign. Fromer *et al.*, Do *et al.* and De Rubeis *et al.* used PolyPhen-2, while Purcell *et al.*, Fuchsberger *et al.*, and Genovese *et al.* used an ensemble approach in which missense variants classified as damaging by multiple tools were defined as pathogenic. In the previous sections, I demonstrated that a number of the classifiers, including CADD, outperformed PolyPhen-2 and SIFT. On the other hand, the ensemble approach incorporated a number of tools that did not perform well in our evaluation (such as LRT), or was not very robust and had very different optimal discrimination thresholds depending on the testing set used . Furthermore, in Table 2.3, I described complicated interdependencies between the different annotation tools, in which the same data sets were used for training and evaluation, and some tools even incorporated SIFT and PolyPhen as features during training. Thus, I decided to use CADD to annotate missense variants in our analysis, which achieved reasonable sensitivities and specificities while robust to the choice of the testing data set. I did not apply LOFTEE, as no other case-control or trio study performed additional filtering on loss-of-function variants. However, this remains an unsolved problem, and no single approach could be suggested as directly superior to the others.

2.10 A meta-analysis of published schizophrenia parent-proband trio studies

Recent studies have leveraged whole-exome sequencing to identify *de novo* mutations in parent-proband trios. These mutations are very rare germline events that arose in a single

generation, and their unlikely occurrence in individual genes have been used to implicate risk genes for severe Mendelian disorders. In these disorders, gene discovery did not require a well-calibrated statistical model: for instance, five of the six probands sequenced with Wiedemann-Steiner syndrome had LoF mutations in *KMT2A* [137], while nine of the ten probands sequenced with Kabuki syndrome had *de novo* truncating events in *KMT2D* [82]. However, for more complex and heterogeneous disorders, the burden of *de novo* mutations was likely spread over many genes. Early sequencing studies of hundreds of schizophrenia and autism probands successfully demonstrated that a genome-wide excess of *de novo* mutations existed in cases compared to controls [80, 96], but were underpowered to identify individual genes.

Because recent studies have suggested that case-control and *de novo* data appeared to implicate an overlapping set of genes [105], I aggregated validated de novo mutations identified in schizophrenia trios from seven published studies for analysis with our case-control cohort [98, 99, 95, 97, 100–102]. I ensured that all *de novo* mutations included our analysis had been validated with Sanger sequencing, and that each parent-proband trio was included only once in our analysis (Table 2.4). For example, the Xu *et al.* 2011 and 2012 studies and the Takata et al. 2014 study analysed trios from the same underlying cohort. After excluding sample duplicates, I identified 118 LoF and 662 missense *de novo* mutations in 1,077 schizophrenia probands for subsequent analysis.

2.11 Gene-specific mutation rates based on GENCODE transcripts

To implicate individual genes using *de novo* mutations, a robust method of evaluating the excess of *de novo* events is needed. One approach to evaluating the excess of *de novo* mutations is to first estimate the expected per-generation rate of new mutations in gene g (μ_g). Given this gene-specific rate, the probability of observing X new mutations in gene g as observed in N trios can be modelled using the following Poisson distribution:

$$X \sim Pois(2N\mu_g)$$
$$P(X \geq x) = 1 - \sum_{i=0}^{x-1} P(X = i)$$

where X is number of *de novo* mutations in gene g, μ_g is the gene-specific mutation rate, and N is the number of trios in our study. However, establishing robust gene-specific mutation

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 2.4 Published studies identifying *de novo* mutations in schizophrenia parent-proband trios using whole-exome sequencing. The Xu *et al.* and Takata *et al.* studies analysed the trios from the same underlying cohort. After excluding sample duplicates, 1,077 schizophrenia trios were available for analysis.

rates is challenging: genes differ significantly in both total coding length and local sequence context, resulting in substantial differences in their mutability.

A recent study generated gene-specific mutation rates by considering the tri-nucleotide context of each base change, and integrating these locally adjusted rates across an entire gene [138]. The probabilities of each of the 192 possible mutational changes were described as constant values in a mutation rate table. To calculate a gene-specific mutation rate for different types of mutations (LoF, missense, synonymous), the authors determined all possible mutational changes in the gene that would introduce a change of that particular class, and added the tri-nucleotide probabilities of all of these theoretical events. As a robustness check, the study showed that the correlation between the number of rare synonymous variants in each gene and the probability of a synonymous mutation as defined by the mutational rate model was 0.94.

Because of the reliability of this model as demonstrated in its use in previous studies of autism and developmental disorders [105, 118], I chose to incorporate it in our analysis of schizophrenia trios, with a few minor adjustments. First, the gene-specific mutation rates in Samocha *et al.* were calculated based on canonical transcripts as defined by an older version of NCBI RefSeq database (pre-2014), which described fewer protein-coding genes and transcripts per gene than the GENCODE database [123]. Second, the missense mutation rates did not incorporate *in silico* annotations to prioritise more damaging events, and

restricting our analysis to only CADD $\geq$ 15 missense variants further reduces the mutational target of each gene and improves power [130]. To address these limitations, I identified the canonical GENCODE v.19 coding transcript of each gene as defined by the APPRIS annotation pipeline. APPRIS incorporated information from protein structure, functional information, and evolutionary evidence to identify one transcript per gene as the principal functional isoform. In the case of multiple principal transcripts, I conservatively selected the longest APPRIS principal transcript. Gene-specific mutation rates for LoF, missense, and synonymous variants for each GENCODE transcript were computed using the tri-nucleotide mutation rates and method previously described in Samocha *et al*, by adding the probabilities of all theoretical mutational events. I then annotated all possible missense mutations with CADD scores, and calculated a gene-specific mutation rate for missense variants with CADD PHRED score $\geq$ 15.

For genes that existed in both transcript references (RefSeq and GENCODE), our mutation rates based on GENCODE transcripts correlated well with those described in Samocha *et al.*, with a correlation coefficient of 0.97 and 0.98 for missense and LoF mutations respectively (Figure 2.10). Notably, our gene mutation rates were on average greater than the published rates since I conservatively selected for longer transcripts when multiple principal isoforms are available. By using GENCODE over RefSeq, I generated rates for an additional 1,067 protein-coding genes, enabling statistical tests on a more comprehensive set of genes. I also found that only 44% of all possible missense variants had CADD $\geq$ 15, resulting in a substantial reduction in the mutational target for most genes in the genome (Figure 2.11). Interestingly, there was substantial variability in the fraction of CADD damaging sites in different genes: I found that missense damaging sites were nearly completely absent in around $\sim$1,500 genes, while in other genes, more than 75% of all missense sites can be prioritised as damaging. This variability appears to be a property of gene function, since olfactory receptors as a class appear to have the lowest proportion of missense damaging sites. As later shown in Section 4.3.2, these classifier-adjusted rates increased our power to distinguish patterns in *de novo* burden across neurodevelopmental and psychiatric disorders.

2.12 Discussion

Using whole-exome sequence data from the UK10K study, INTERVAL study, Swedish Schizophrenia project, and the SiSU project, I generated a discovery data set of 4,264 schizophrenia cases and 9,343 controls. Despite following standard protocol for alignment and joint calling all samples at the same time, I still observed substantial batch effects from different exome captures used at different time points of the experiment. To address this, I

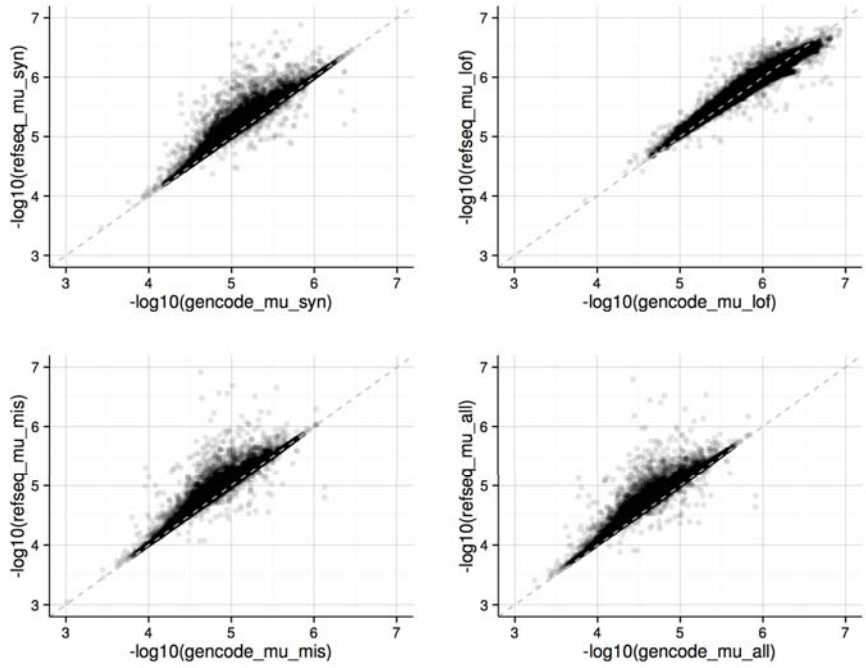

Fig. 2.10 **Correlation between mutation rates generated using GENCODE and RefSeq transcript databases** I compared the LoF, missense, synonymous, and total mutation rates generated using the two different transcript references. Each dot represented a different gene, and mutation rate μ calculated from RefSeq was plotted along the Y-axis, while the rate from GENCODE was plotted along the X-axis.

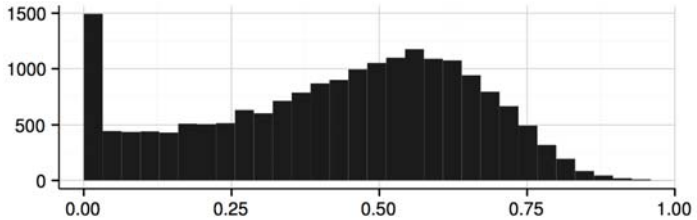

Fig. 2.11 **The ratio of the damaging missense mutation rate to the missense mutation rate of each GENCODE coding gene.** The ratio between missense rate using only CADD damaging sites to the rate from all missense sites was displayed using a histogram. The mean of the bi-modal distribution was 0.44.

restricted our analysis to regions with reasonable coverage in all samples (7× or greater in 80% of each sequencing batch), and then identified appropriate variant- and genotype-level filters using rare inherited and Mendelian inconsistent calls from the DDD study. I found that well-calibrated threshold filters on variant- and genotype-level quality metrics (GQ, DP, and AB) complemented well with a supervised method like GATK VQSR to produce reasonable sensitivity and specificity for rare variant calls. A small number of common coding SNPs was sufficient for sample-level QC aimed at reducing potential biases from ancestry, relatedness, and contamination. Following sample and variant QC, I observed no genome-wide inflation in rare variant tests in subsequent analyses (Section 3.3.1, 3.3.6).

To increase power of collapsing tests of missense variants, I tested the effectiveness of a number of available *in silico* classifiers on a set of *de novo* mutations from the DDD study that were reported back to patients and their families as clinically significant. Ensemble classifiers (LR pred and Radial SVM) performed well when compared to commonly used tools like SIFT and PolyPhen, but a fixed discrimination threshold could not be reliably determined. As a second best option, I decided to annotate missense variants with a CADD score ≥ 15 as damaging, excluding up to 80% of all benign polymorphisms while retaining up to 80% of all diagnostic missense variants. I restricted our subsequent analyses to damaging missense and LoF variants. Lastly, I extended the tri-nucleotide *de novo* model to all canonical GENCODE transcripts, and generated mutation rates for damaging missense variants in addition to all other functional classes. Taken together, the steps highlighted in this Chapter lay the framework for analyses of rare variant data that should also be applicable in future exome sequencing studies.

2.13 Consortia

I would like to acknowledge the following consortia for providing data for the analyses described in this thesis.

2.13.1 UK10K consortium

Richard Anney, Mohammad Ayub, Anthony Bailey, Gillian Baird, Jeff Barrett, Douglas Blackwood, Patrick Bolton, Gerome Breen, David Collier, Paul Cormican, Nick Craddock, Lucy Crooks, Sarah Curran, Petr Danecek, Richard Durbin, Louise Gallagher, Jonathan Green, Hugh Gurling, Richard Holt, Chris Joyce, Ann LeCouteur, Irene Lee, Jouko Lönnqvist, Shane McCarthy, Peter McGuffin, Andrew McIntosh, Andrew McQuillin, Alison Merikangas, Anthony Monaco, Dawn Muddyman, Michael O'Donovan, Michael Owen, Aarno Palotie,

Jeremy Parr, Tiina Paunio, Olli Pietilainen, Karola Rehnström, Tarjinder Singh, David Skuse, Jim Stalker, David St. Clair, Jaana Suvisaari, Hywel Williams

2.13.2 DDD Study

Nadia Akawi, Saeed Al-Turki, Kirsty Ambridge, Jeffrey Barrett, Daniel Barrett, Tanya Bayzetinova, Nigel Carter, Stephen Clayton, Eve Coomber, Helen Firth, Tomas Fitzgerald, David FitzPatrick, Sebastian Gerety, Susan Gribble, Matthew Hurles, Philip Jones, Wendy Jones, Daniel King, Netravathi Krishnappa, Laura Mason, Jeremy McRae, Parker Michael, Anna Middleton, Ray Miller, Katherine Morley, Vijaya Parthiban, Elena Prigmore, Diana Rajan, Alejandro Sifrim, Tarjinder Singh, Adrian Tivery, Margriet van Kogelenberg, Caroline Wright

2.13.3 Swedish Schizophrenia Study

Sarah Bergen, Kimberly Chambert, Menachem Fromer, Christina M. Hultman, Anna K. Kähler, Steve McCarroll, Jennifer L. Moran, Shaun Purcell, Stephan Ripke, Douglas Ruderfer, Edward Scolnick, Pamela Sklar, Patrick F. Sullivan

2.13.4 INTERVAL study

Participants in the INTERVAL randomised controlled trial were recruited with the active collaboration of NHS Blood and Transplant England (www.nhsbt.nhs.uk), which has supported field work and other elements of the trial. DNA extraction and genotyping was funded by the National Institute of Health Research (NIHR), the NIHR BioResource (http://bioresource.nihr.ac.uk/) and the NIHR Cambridge Biomedical Research Centre (www.cambridge-brc.org.uk). The academic coordinating centre for INTERVAL was supported by core funding from: NIHR Blood and Transplant Research Unit in Donor Health and Genomics, UK Medical Research Council (G0800270), British Heart Foundation (SP/09/002), and NIHR Research Cambridge Biomedical Research Centre.

A complete list of the investigators and contributors to the INTERVAL trial is provided in reference [139], and http://www.intervalstudy.org.uk/about-the-study/whos-involved/interval-contributors/.

2.13.5 Sequencing Initiative Suomi project

The Sequencing Initiative Suomi (SISu) project is an international collaboration between research groups aiming to build tools for genomic medicine. These groups are generating

whole genome and whole exome sequence data from Finnish samples and provide data resources for the research community. Key groups of the project are from Universities of Eastern Finland, Oulu and Helsinki and The Institute for Health and Welfare, Finland, Lund University, The Wellcome Trust Sanger Institute, University of Oxford, The Broad Institute of Harvard and MIT, University of Michigan, Washington University in St. Louis, and University of California, Los Angeles (UCLA). The project is coordinated in the Institute for Molecular Medicine Finland at the University of Helsinki.