ONLINE SUPPLEMENT

De novo single nucleotide and copy number variation in discordant monozygotic twins reveals disease-related genes

Supplementary Methods, Supplementary Results, Supplementary Tables, Supplementary Figures and Figure Legends, and Supplementary References.

Supplementary Methods

Description of twin pairs

Coriell Cell Repository DNA was obtained from five pairs of MZ twins discordant for ALS (218 and 318; 421 and 422; 242 and 243), Tourette's syndrome (489 and 490), and Parkinson's disease (PD821 and PD161). DNA was also obtained from the parents of the twins discordant for Tourette's syndrome (487 and 488).

Genomic DNA samples of two twin pairs discordant for schizophrenia were obtained; one pair from King's College London, Institute of Psychiatry (IP16 and IP17), and one pair from the RIKEN Institute, Brain Science Institute (RT1a and RT1b).

A further six discordant MZ twins were recruited, and DNA was extracted from blood and/or saliva samples for molecular/genetic analysis. These twins were discordant for stroke (HG and KG), ALS (LAS and SUS), lactase non-persistence (KEL and KIR), inclusion body myositis (AFF and UNAFF), hereditary spastic paraplegia (VF and LF), and autism spectrum disorder (RP and OH). DNA from the parents of RP and OH was also obtained (DS and DV). Written informed consent was obtained from all participants prior to study entry.

Sample preparation and quantification

DNA extraction from saliva

The Oragene DNA Kit (DNA Genotek Inc., Kanata, Canada) was used to extract total DNA from saliva samples according to the manufacturer's instructions. Briefly, 500µL of the mixed saliva and Oragene solution was placed into a 1.5mL microcentrifuge tube and incubated for 2hrs at 50^{o}C. 20µL of Oragene DNA purifier was added, and the tube was briefly vortexed. The solution was incubated on ice for 10mins, then centrifuged at room temperature for 10mins

at 13,000rpm. The resulting supernatant was transferred into a new 1.5mL microcentrifuge tube, and the remaining pellet was discarded. 95% ethanol was added to the supernatant and the solution was gently mixed by inverting the tube several times; this was left at room temperature for 10mins to allow the DNA to precipitate. The solution was then centrifuged for 2mins at 13,000rpm, and the supernatant was carefully removed and discarded. For further purification, 200µL of 70% ethanol was added and then removed after 1min, taking care not to disturb the pellet. 50µL of Tris-EDTA buffer was added to the tube to dissolve the DNA pellet. The tube was briefly vortexed and incubated for 1hr at 50ºC.

DNA extraction from blood

Blood from subjects was taken in ethylenediaminetetraacetic acid (EDTA) bottles and genomic DNA was extracted from whole blood using a FlexiGene kit (Quiagen) according to manufacturer's instructions. For each sample, 300µL of whole blood was mixed with 750µL of buffer (FG1) to lyse the cells. The sample was centrifuged for 20secs at 13,000rpm, and the supernatant was discarded leaving only the pellet. 150µL of protease-contain buffer (FG2) was added to the tube and then vortexed until the pellet became homogenised. The sample was centrifuged for 10secs and incubated at 65°C for 5secs in a heating block. 150µL of 100% isopropanolol was added to the tube and inverted several times to allow precipitation of the DNA. The sample was centrifuged for 3mins at 13,000rpm and the supernatant was discarded. 150µL of 70% ethanol was added to the sample and vortexed for 5secs before centrifuging for 3mins at 13,000rpm. The supernatant was discarded and the pellet air dried. 200µL of buffer (FG3) was added and the sample and briefly vortexed to dissolve the DNA.

DNA concentration and purity

A NanoDrop ND-1000 Spectrophotometer (NanoDrop Technologies) was used to measure genomic DNA concentration and purity. 1µL of DNase and RNase-free distilled deionised

water (dH2O) was loaded onto the pedestal to obtain a blank reading. 1µL of DNA from all participants was individually loaded to determine the concentration. Absorbance measurements determine molecules absorbing at a specific wavelength. DNA will absorb at 260nm and will contribute to the total absorbance. For quality control, the 260/280nm and 260/230nm absorbance ratios are used to assess the purity of DNA. For DNA, a 260/280nm absorbance ratio of ~1.8 is generally accepted as adequate purity. The 260/230nm absorbance ratio is used as a secondary measure of nucleic acid purity, and a value within the range of 2.0-2.2 is considered acceptable. 1µL of all DNA samples were run on a 1.3% agarose gel to determine the relative concentrations and degree of potential degradation.

Initial screen of known disease-associated variants

Some samples were screened for known pathogenic variants before being put forward for whole-exome sequencing analysis. A pathogenic hexanucleotide repeat expansion within the *C9orf72* gene has been identified as the major cause of ALS. Variation in the hexanucleotide repeat number was first assessed by repeat-primed PCR (rpPCR). A modified Southern blot method was used to confirm the rpPCR detected expansions where sufficient DNA allowed.

Repeat-primed PCR

To provide a qualitative assessment of the presence of an expanded $(GGGGCC)_n$ hexanucleotide repeat in *C9ORF72*, a rpPCR reaction was performed in four twin pairs discordant for ALS. Briefly, 100ng of genomic DNA were used as template in a final volume of 28µL containing 14µL of FastStart PCR Master Mix (Roche Applied Science, Indianapolis, IN, USA), and a final concentration of 0.18mM 7-deaza-dGTP (New England Biolabs Inc., Ipswich, MA, USA), 1x Q-Solution (Qiagen Inc., Valencia, CA, USA), 7% DMSO (Qiagen), 0.9mM $MgCl_2$ (Qiagen), 0.7µM reverse primer consisting of ~four GGGGCC repeats with an

anchor tail, 1.4µM 6FAM-fluorescent labelled forward primer located 280bp telomeric to the repeat sequence, and 1.4µM anchor primer corresponding to the anchor tail of the reverse primer. Primers used for rpPCR are shown in Supplementary Table 1. A touchdown PCR cycling program was used where the annealing temperature was gradually lowered from 70°C to 56°C in 2°C increments with a 3min extension time for each cycle.

The rpPCR is designed so that the reverse primer binds at different points within the repeat expansion to produce multiple amplicons of incrementally larger size. The lower concentration of this primer in the reaction means that it is exhausted during the initial PCR cycles, after which the anchor primer is preferentially used as the reverse primer. Fragment length analysis was performed on an ABI 3730xl genetic analyser (Applied Biosystems Inc., Foster City, CA, USA), and data analysed using GeneScan software (version 4, ABI). Repeat expansions produce a characteristic saw-tooth pattern with a 6-bp periodicity.

Primer name	Primer sequence	Concentration	Modification
ALSFTDf_(6FAM)	6-FAM-AGT CGC TAG AGG CGA AAG C	0.05 µmol	Modified DNA Oligos
ALSFTDr	TAC GCA TCC CAG TTT GAG ACG GGG GCC GGG GCC GGG GCC GGG G	0.05 µmol	Unmodified DNA Oligos
ALSFTDanchor	TAC GCA TCC CAG TTT GAG ACG	0.05 µmol	Unmodified DNA Oligos

Supplementary Table 1. Primers used for rpPCR

Southern blotting

10µg DNA was digested with HindIII and XbaI overnight. DNA fragments were separated on a 0.9% TRIS-Borat-EDTA (TBE) agarose gel, transferred by alkali blotting onto an Amersham Hybond NTM-XL membrane (GE Healthcare, Fisher Scientific, Germany) and hybridised to a

^{32}P-labelled probe overnight. After washing, X-ray films were exposed for 4–6 days at −80°C. BstEII digested lambda DNA was used as a size marker for estimating repeat lengths. Minimal repeat sizes were used for calculation.

Next-generation sequencing panels

Amyotrophic lateral sclerosis

Genomic DNA from one twin pair discordant for ALS (LAS and SUS) was processed on a gene panel, which uses the Illumina TruSeq Custom Amplicon implemented on an Illumina MiSeq platform. This panel utilises PCR amplicon-based target enrichment and screens for variants in 25 ALS disease genes. Sequence analysis involved the full exomes of 10 genes strongly implicated in ALS, and specific genomic areas where disease-causing mutations cluster in 15 other minor or unproven ALS-linked genes. Probes were created using Illumina TruSeq custom amplicon assay DesignStudio v1.6 (http://www.illumina.com/applications/designstudio.ilmn).

Hereditary spastic paraplegia

LF was diagnosed with hereditary spastic paraplegia (HSP) at 34 years presenting initially due to sudden-onset recurrent 20min spasms in her right leg. This was triggered by exercise, driving, caffeine and alcohol. Her identical twin sister (VF) and 16-year-old son may have similar features of spasticity. A diagnosis of complicated HSP superimposed with dystonia was made. The patient did not respond well to L-Dopa (to exclude dopa-responsive dystonia), and there was no diurnal variation in presenting symptoms. Sequencing and analysis of genomic DNA from the twin affected with HSP (LF) was carried out on a multi-gene panel at Sheffield Children's NHS Foundation Trust, using the Agilent SureSelect Neurogenetic panel Version 1 with Illumina MiSeq Analysis pipeline: Version 2. LF was also screened and negative for GLUT1 deficiency syndrome and DYT1 early-onset primary dystonia.

Lactase non-persistence

Additionally, Sanger sequencing was used to detect known SNPs in twins discordant for lactase non-persistence (KIR and KEL) (methods described below).

Next-generation sequencing

DNA library construction of remaining samples

Whole-exome sequencing libraries for the next set of samples (n=16) were prepared with Agilent SureSelect V6 and sequenced on an Illumina HiSeq3000 using a 75-bp paired-end reads protocol.

Whole-exome capture

Alignment of the previous and newly-sequenced exomes to the human reference genome (UCSC hg19), and variant calling and annotation was performed with an in-house pipeline developed by Alan Pittman. Briefly, this involves alignment with NovoAlign, removal of PCR-duplicates with Picard Tools followed by (sample-paired) local realignment around indels and germline variant calling with HaplotypeCaller according to the GATK best practices.

Potentially mosaic variants were identified with GATK MuTect2 (version 2.0) and VarScan2 (version 2.4.3), using each pair as reference to one-another (described below). The raw list of SNVs and indels were then filtered using ANNOVAR. Variants in splicing regions, 5'UTR, 3'UTR and protein-coding regions, such as missense, frameshift, stop-loss and stop gain mutations, were considered. Priority was given to rare variants (<1% in public databases, including 1000 Genomes project, NHLBI Exome Variant Server, Complete Genomics 69, and Exome Aggregation Consortium). Furthermore, we have an in-house set of approximately six

thousand exomes encompassing controls, rare diseases for cross-checking any shortlisted candidate variants, and for sequencing artefact removal.

Comparison of two somatic variant calling methods

VarScan2 and MuTect2 algorithm tools were used to identify de novo variants using the somatic mutation calling method. The union of SNVs called by both variant callers was taken forward for tiering, filtering and manual review. This method was used on discordant MZ twins by treating the affected twin as the 'tumour' sample and the unaffected twin as the 'normal' sample (and viscera to detect somatic mutations present in the unaffected twin but not in the affected twin). MuTect2 uses a Bayesian classifier approach to detect somatic mutations with very low allele fractions, requiring only few supporting reads, followed by carefully tuned filters that ensure high specificity. In this study, MuTect2 was run under the High-Confidence mode with its default parameter settings. Low quality sequenced data was first removed, followed by variant detection in the 'tumour' sample using a Bayesian classifier. A filtering step was then applied, which removes false positive variants caused by sequencing artefacts. Finally, variants are classified as somatic or germline by a second Bayesian classifier.[1]

VarScan2 reads BAM files from 'tumour' and 'normal' samples simultaneously to heuristically call a genotype at positions achieving certain thresholds of coverage and quality. It uses a one-tailed Fisher's exact test to calculate the significance of the difference in allele frequencies of the normal and tumour sample based on the number of reads supporting each allele. We used a cut-off value of Fisher's P-value <0.05. If the resulting p-value meets this significance threshold, variants are classified as somatic (if the tumour call was different from the normal and the normal call was the same as the reference), loss of heterozygosity (if the tumour variant call was not heterozygous but the normal variant call was heterozygous), or unknown (if the

tumour call was different from the normal call and both calls were different from the reference). The variant is classified as germline if the difference does not meet the significance threshold.

In summary, VarScan2 provides sensitive detection of high-quality somatic SNVs, whereas MuTect2 provides sensitive detection of low allelic-fraction. The compatibility of the output VCF files between different methods was examined using Microsoft Excel. Somatic variants which were common to both algorithms were retained for downstream analysis.

Tools	***Version***	***URL***	***Remark***	***Release date***
VarScan	2.4.3	http://dkoboldt.github.io/varscan	Sensitive detection of high-quality somatic SNVs	Dec. 2016
MuTect2	2.0	https://software.broadinstitute.org/gatk/guide/tooldocs/org_broadinstitute_gatk_tools_walkers_cancer_m2_MuTect2.php	Sensitive detection of low allelic-fraction	Nov. 2015

Supplementary Table 2. Details of algorithm tools for somatic SNV detection within NGS data

DNA variant and gene prioritisation

Putative discordant variants called by both MuTect2 and VarScan2 were further filtered according to low variant quality (VQS<90), common variants (MAF>0.01) as reported in public databases (1000g, ExAC, cg69), and variants in regions containing segmental duplications. This stringent filtering criteria provided a manageable list for evaluation of candidate variant sites; thus, variants with genomic locations in exonic, 5'UTR, 3'UTR, splice site and promoter regions were retained.

For the detection of concordant variants, in addition to the filtering criteria above, variants were analysed for their potential deleterious effects using the polymorphism phenotyping v2

(PolyPhen2) and Sorting Tolerant from Intolerant (SIFT) algorithms.[2,3] Within the prioritised variants, those harbouring truncating mutations or mutations predicted to be damaging were considered the most promising candidates. The priority order of variants, from most to least damaging, were as follows: frameshift, nonsense, splice site, missense and non-stop. All missense variants predicted to be benign were removed.

Where parental DNA was available, provisional postzygotic de novo mutations identified in the twins were excluded if they were detected in either parents. To identify germline de novo mutations shared between the twins, parent-offspring trio analysis was performed. Contrary to single sample calling, where samples are analysed individually, joint genotyping was performed on all samples according to GATK best practices. Analysing variants simultaneously across all samples has several advantages, including 1) Being able to better distinguish between homozygous reference sites and sites with missing data; 2) Having higher sensitivity for low-frequency variants. Joint calling enables the 'rescuing' of genotype calls at sites where there's low coverage but other samples within the call set have a confident variant at that location; 3) Being able to more efficiently filter out false positives. Studies have shown that GATK's Variant Quality Score Recalibration (VQSR) provides better calling accuracy than simply using hard filtering.[4] VQSR builds a Gaussian mixture model by looking at the annotation values over a subset of the input call set, then by using machine learning algorithms, determines the annotation profile of good and bad calls, and evaluates all input variants. Joint calling provides a large enough dataset for accurate error modelling and ensures that filtering is applied uniformly across all samples.

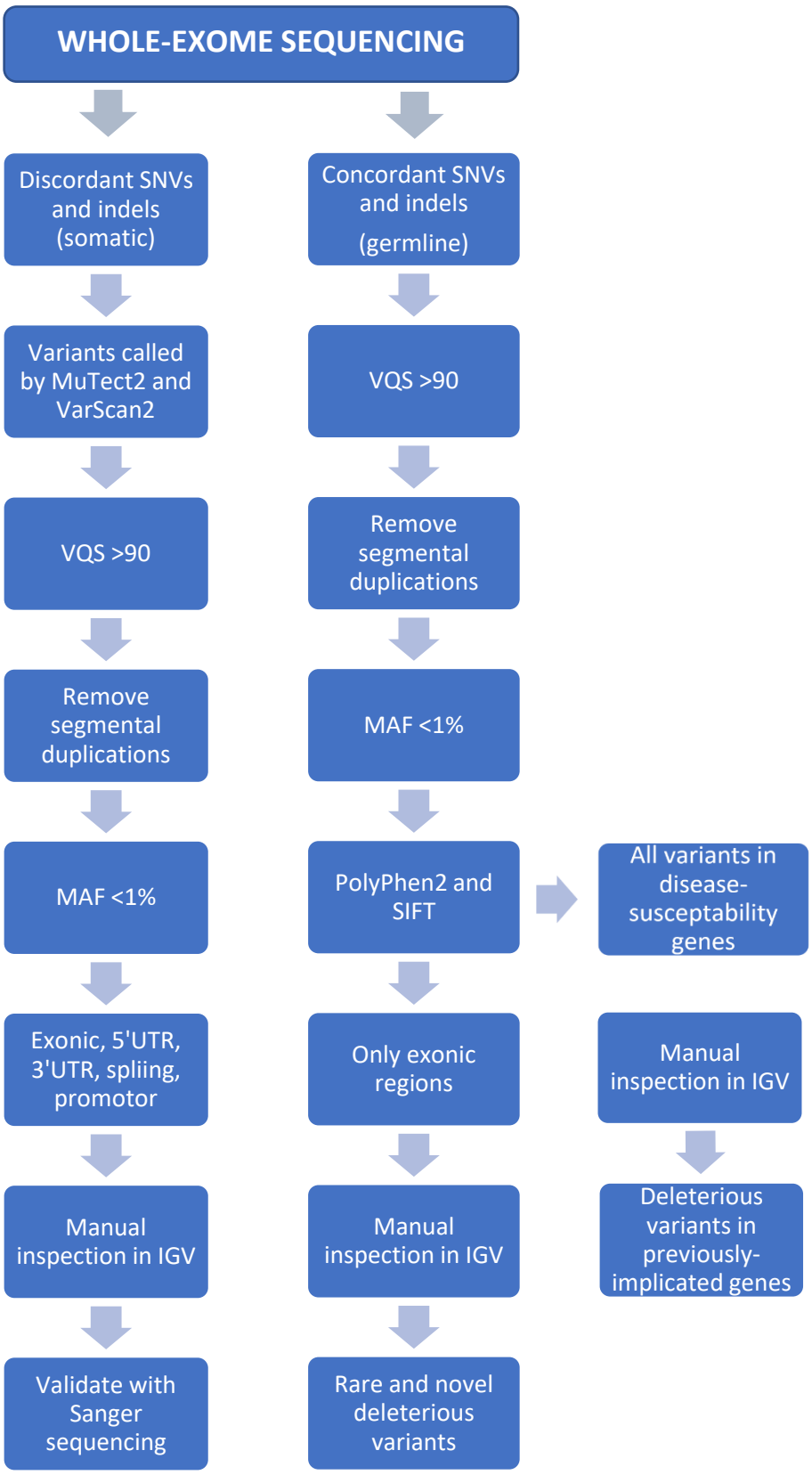
The data were filtered in Excel to identify concordant mutations in co-twins, but absent in the parents and other samples in the dataset. In the twins discordant for Tourette's syndrome (489

and 490), because the father (487) was also affected, inherited pathogenic mutations from the father were also investigated.

PubMed, Online Mendelian Inheritance in Man (OMIM), NIH Genetic Testing Registry (GTR) and DisGeNET were reviewed for previous publications regarding candidate genes. In addition, gene databases for the disorders investigated in this study were also searched, including Amyotrophic Lateral Sclerosis Online Genetics Database (ALSoD), ALSGene, PDGene, Schizophrenia Database (SZDB), Schizophrenia Gene (SZGene). All identified genes for each disorder were pooled together to form a comprehensive list, and was cross-checked with germline variants in each twin pair.

With the possibility of finding potentially-damaging rare variants in novel genes, after prediction of functional effects of the selected variants, all non-exonic variants and variants indicated as synonymous were removed. Genomic evolutionary rate profiling (GERP++) scores, a measure of evolutionary constraint at each base derived by aligning 29 mammalian genomes, were also used to estimate the conservation of each variant. Negative scores indicate a lack of conservation and high positive scores indicate the most conserved nucleotide positions among multiple species.[5]

All concordant and discordant variants between twin siblings were manually reviewed in IGV prior to validation.



Supplementary Figure 1. General overview of the methods used to detect discordant and concordant variants in MZ twins. The MAF filter was not applicable for twins discordant for lactase non-persistence, as common polymorphisms are associated with the condition, and Tourette's syndrome, where modifier variants were more likely to play a role.

Variant validation by Sanger sequencing

Twin pair	*Gene*	*Forward primer*	*Reverse primer*
LAS and SUS	*DENND5B*	TGAACCTGTCATCTGCCAGA	acagtagttgggagcaaaagt
	LOXHD1	tggggtagccactgtctaac	ggacgcatgtacCTGAGACA
	BTK	tttggtggactctgctacgt	aagggtgctgctttaatgcc
	SLC26A1	TGCGTGAGATGCTGAAGG	GTTGGCGAAGAAGGACGTAT
	MT-ND5	CCACATCATCGAAACCGCAA	CCAGGCGTTTAATGGGGTTT
218 and 318	*FBXO38*	AGTGGTCTTCAGCGTGTAGT	ggccaagtgtgtcagttctg
	RIT2	TTTTGAGACCTCTGCAGCCC	ATTCAGAGAGCGTGAGGAAC
	GRM6	ggggagatggatagagtggg	ATGGGTGAGGTCTTGGCTC
	PPARGC1A	tttgcttttctcctctggcg	CGTAGCTGTCATACctggga
421 and 422	*KIAA1107*	TGCAACTGAAAATGAATGTCGTG	AGATCCAAGACAGTGCTTTCA
	C8orf48	AGAAGACGAGCAGCAGACAT	GGTCATCTGAAAGCCTGGGA
	ATP6V1B2	ccttggtatatctgcgcgtg	ctagggagcaaagaccggag
242 and 243	*TMEM225B*	ctgacctgacccaaacgttg	TGTCCTGGGCTAGATGACTG
	ACTR3C(LRRC61)	AACAAGTTGGGTGGTGGGC	gaagattctggcctctccca
	KBTBD3	GAGGAGACACTTCAGGACTGT	CGAAAGACTAGATCCTGGCAA
	TUBGCP4	tttccttgactagCGCCTGA	tggggccacatagtattaagga
	TFIP11	TTGAGACCAAGGCTGAGGAG	CCTCTTAGGGGCAAGTCTCT
	PHKA2	CTCATCTACGAGGCCAGTGG	ATCCGGAGTCTCAGCATCTC
	GNL3L	ccttttcactgaggcctgac	ggagagaagagggttgccat
KEL and KIR	*PLCB1*	aggaagctggggaaggaaaa	taccactccaagtcctgctc
OH and RP	*RASD2*	ttctctcgctttgttgcagC	CAGGATGTCGAGCTGGTACA
489 and 490	*AADAC*	TGGATTCCATGGAGCATTTTCA	TCTTGCATCGTAAATAGTAGCCA

Supplementary Table 3. Primer pair sequences for variant validation.

Genome-wide SNP genotyping

Wet lab processing

300ng of high-quality genomic DNA from each subject was whole-genome amplified overnight at 37ºC for 20-24hrs in a deep well plate, then fragmented at 37ºC for 1hr15mins in a hybridisation oven, precipitated and resuspended in hybridisation buffer. Samples were

denatured then taken from the plate and loaded onto the chips using a liquid handling robot (Freedom Evo, Tecan Ltd, Switzerland). Hybridisation took place overnight for 16-20hrs at 48°C.

The process of single base extension and staining was carried out by the liquid handling robot. The probes on the chip were extended by a single hapten-labelled dideoxynucleotide (ddNTP) base complementary to the hybridised DNA. ddATP and ddTTP bases were labelled with DNP (2,4- Dinitrophenol), whereas ddCTP and ddGTP were labelled with Biotin. The DNA samples were then stripped off the chip using formamide. The staining procedure involves signal amplification by multi-layer immunohistochemical staining. The haptens were detected simultaneously by Streptavidin and an anti-DNP primary antibody conjugated to green and red fluorophores respectively (STM reagent, Illumina). They were then counterstained with biotinylated anti-steptavidin and a DNP-labelled secondary antibody to the anti-DNP primary antibody (ATM reagent, Illumina) to amplify the fluorescent signals. The last layer of stain was the STM, containing the fluorophores to allow signal detection. Finally, the stained chips were coated in nail varnish to protect the dyes, and scanned using the iScan scanner with autoloader (Illumina Inc, San Diego, USA).

Initial data analysis and quality control

The data were initially analysed using the Illumina Genomestudio software. This generates genotypes, and CN and loss of heterozygosity data (cnvPartition v3.1.6, Illumina). Quality control checks were performed to assess the data quality. Samples were assessed for their call rate, which should be >98% and >99% average across the batch. Specially designed control probes were also checked. Every array contained both sample dependent and sample independent control probes. Sample independent probes assess the quality of the processing,

and sample dependent probes also assess the quality of the DNA. The B-allele frequency (BAF) plots and CN analysis results were checked to identify potentially contaminated samples. The BAF plot would show more than three modes if the sample had been contaminated. A noisy BAF plot may also suggest degradation of the DNA sample. CN data that looks to be duplicated for the whole genome also suggests contamination with at least one other DNA sample.

Copy number variant detection

For all samples, Log R Ratios (LRR) and BAFs were generated using Illumina Genome Studio software (v2011.1) and used to call CNVs with PennCNV.[6] CNV calling was performed following the standard protocol and adjusting for GC content. Samples were excluded if they were found to be an outlier for any one of the following QC metrics: LRR standard deviation, BAF drift, wave factor and total number of CNVs called per person. The LRR represents a measure of magnitude of combined fluorescence-intensity signals, and the BAF denotes the relative ratio of fluorescence signals from one allelic probe compared with another. Duplications can be identified by an increase of LRR and the occurrence of four clusters in BAF. Consequently, a deletion is characterised by a decrease of LRR and lack of heterozygosity (at 0.5) in BAF. CNVs from samples that passed QC were joined together if the distance separating them was <50% of their combined length using an in-house developed open source program (http://x004.psycm.uwcm.ac.uk/~dobril/combine_CNVs/). CNVs were then excluded if they were covered by <3 probes. After CNV merging, the remaining CNVs were visually re-evaluated using the GenomeStudio genotyping module. All CNV coordinates are according to UCSC build 37/hg19.

cnvPartition was used as the secondary CNV detection algorithm using the following default parameters:

Confidence Threshold	|35
Detect extended homozygosity	|True
Exclude intensity only	|False
GC wave adjust	|False
Include sex chromosomes	|True
Minimum homozygous region size	|1000000
Minimum probe count	|3

The CNVs detected by PennCNV and cnvPartition were detected on autosomes only and were based on at least three consecutive probes. Here, these CNVs are referred to as non-mosaic somatic CNVs; that is, acquired somatic CNVs that are present in a sufficiently high proportion of cells to be detected by the applied algorithms.

The evaluation of the non-mosaic structural variants was based on predefined and structured criteria and consisted of the steps shown below.

1. After CNV detection by PennCNV, CNVs of the same type (CN = 0, 1, 2, 3 or 4) were merged if they overlapped with at least 50% of the length of the smaller CNV.

2. Samples from MZ twins, and parents where applicable, were compared. Only CNVs found to be discordant between paired samples, or concordant but overlapping known disease-susceptibility genes, were of main interest.

3. Discordant CNVs were evaluated by visual inspection of LRR and BAF plots in GenomeStudio. If both twins had the same signal intensities, the CNV was classified as concordant. If there was insufficient evidence, the CN call was disregarded.

4. CNVs were independently detected again by cnvPartition in GenomeStudio. CNVs of the same type (deletion or duplication) were merged if they overlapped with at least 50% of the length of the smaller CNV.

5. Only CNVs detected by both algorithms were retained for further analysis. CNVs called by PennCNV were disposed of if the CNV calls made by cnvPartition did not confirm the finding.

6. The LRR and BAF plots of remaining CNVs were visually inspected to select the best candidates for ddPCR validation.

Comparison of CNV/LOH differences between twins.

The output files from cnvPartition and PennCNV were converted into BED format files using a Perl script (see http://genome.ucsc.edu/FAQ/FAQformat.html#format1 for details). The files contain no headers, but all take the same format: 'chr|startPosition (bp)|endPosition(bp)'. BED files are useful as they can be imported as a custom track into the UCSC Genome Browser, allowing one to see all the genes and other features that coincide with the features detailed in the BED file (in this instance, the features in the BED files are the CNV and LOH regions). They can also be viewed and manipulated in Notepad or Excel. The BED files for each pair of twins were compared using BEDTools v2.17, an open source suite of command line operated tools for comparison of BED files (http://bedtools.readthedocs.org/en/latest/content/overview.html). Three comparisons were carried out using the 'intersect' tool:

1. Find features present in twin1 but not in twin2 (filename: {twin1}_not_{twin2}.bed)
2. Find features present in twin2 but not in twin1 (filename: {twin2}_not_{twin1}.bed)
3. Find features that overlap between the two twins (filename: {twin1}_overlap_{twin2}.bed"). This file reports the original feature in twin1 (chr|startPosition(bp)|endPosition(bp)), then the original feature in twin2 (chr|startPosition(bp)|endPosition(bp)), and the final column contains the size of the overlap in bp.

The first two files contain regions of CNV and LOH, which are different between the two twins. The third 'overlap' file was run through a Perl script to remove the overlapping regions

to leave just the regions specific to one twin or the other (filename: {twin1}_{twin2}_unique_regions_from_overlaps.bed). These three files, taken together, contained the CN and LOH differences between each pair of twins. CNVs were then annotated with gene information and compared to previously reported losses, gains, inversions or segmental duplications, and thus categorised as novel or benign polymorphism, using the Database of Genomic Variants (http://dgv.tcag.ca/dgv/app/home). Genes contained within CNVs were then searched on various databases, such as PubMed, OMIM and DisGeNET, to determine pathogenic relevance.

CNV analysis with ExomeDepth

ExomeDepth is an algorithm designed to use read depth data from exome sequencing analysis to call CNVs.[7] It controls technical variability between samples, a feature which ordinarily complicates the analysis and creates spurious CNV calls.

The read count information was extracted from the individual BAM files using the R package Rsamtools. All reads were paired-end. Only reads with a Phred scaled mapping quality ≥20, distance of <1000bp from each other and in the correct orientation, were included. The location was defined by the middle location between the extreme ends of both paired reads. Exons closer than 50bp were merged into a single location owing to the inability to properly separate reads mapping to either of them.

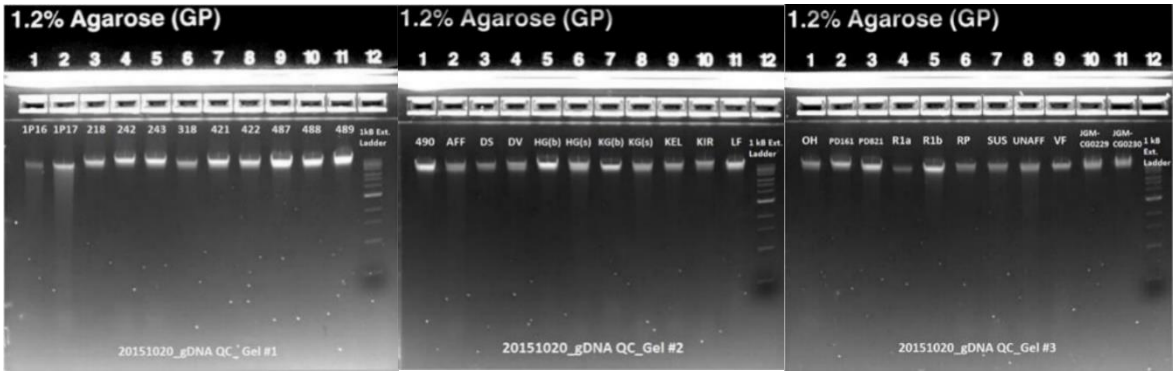
Parameters for ExomeDepth were applied according to the instructions provided by the user guide. To set the threshold for sensitivity and specificity, the correlation between reference and tests count was set to 0.9898. It is advised that this correlation should be >0.97 to avoid a high false positive rate. There is an option, as used in the cancer field, of combining sequence data for healthy and tumour tissue. It is possible to utilise this function by pairing the affected and unaffected twin as ‘tumour’ and ‘normal’, respectively – thus replacing the test sample with

the affected twin sample, and the reference sample with the unaffected twin sample. However, it would be statistically more viable to compare each sample independently against an aggregate reference of all exome samples. This would render shared CN calls between twin pairs, that are not present in other samples, as more likely to be real. This data was used to confirm CNVs detected by the SNP genotyping method. CN calls that were shared by all three calling algorithms (PennCNV, cnvPartition, and ExomeDepth) were considered high confidence CNVs.

Supplementary Results

Quality control and pre-analysis

Whole gDNA obtained was evaluated for quality and quantity by densitometry analysis using 1.2% agarose gel electrophoresis, as shown in Supplementary Figure 2, and spectrophotometric measurement. Agarose gel electrophoresis showed that the molecular weight extracted was more than 10kb with uniform brightness and all samples had a 260/280 ratio of ~1.8, suggesting that the DNA was integrated, stable and that the extraction was successful.



Supplementary Figure 2. Agarose gel electrophoresis of gDNA. DNA samples had a tight band with minimal smearing, therefore passing quality control.

Hereditary spastic paraplegia

HSP-discordant twins (VF and LF) were recruited into the study through collaboration with Niranjanan Nirmalananthan (St George's University Hospitals NHS Foundation). VF was diagnosed with a complex form of HSP and was therefore assigned to an NGS panel of 41 HSP-linked genes. DNA sequencing was performed on the following genes: *AFG3L2, ALS2, AP5Z1, ATL1, B4GALNT1, BSCL2, C12orf65, CYP27A1, CYP2U1, CYP7B1, DDHD1, DDHD2, FA2H* (excluding exon 1), *FIG4, GBA2, GCH1, HSPD1, KIAA0196, KIF1A, KIFSA, L1CAM, MTPAP, NIPA1, PLP1, PSEN1, REEP1, RTN2, SACS, SIGMAR1, SLC16A2,*

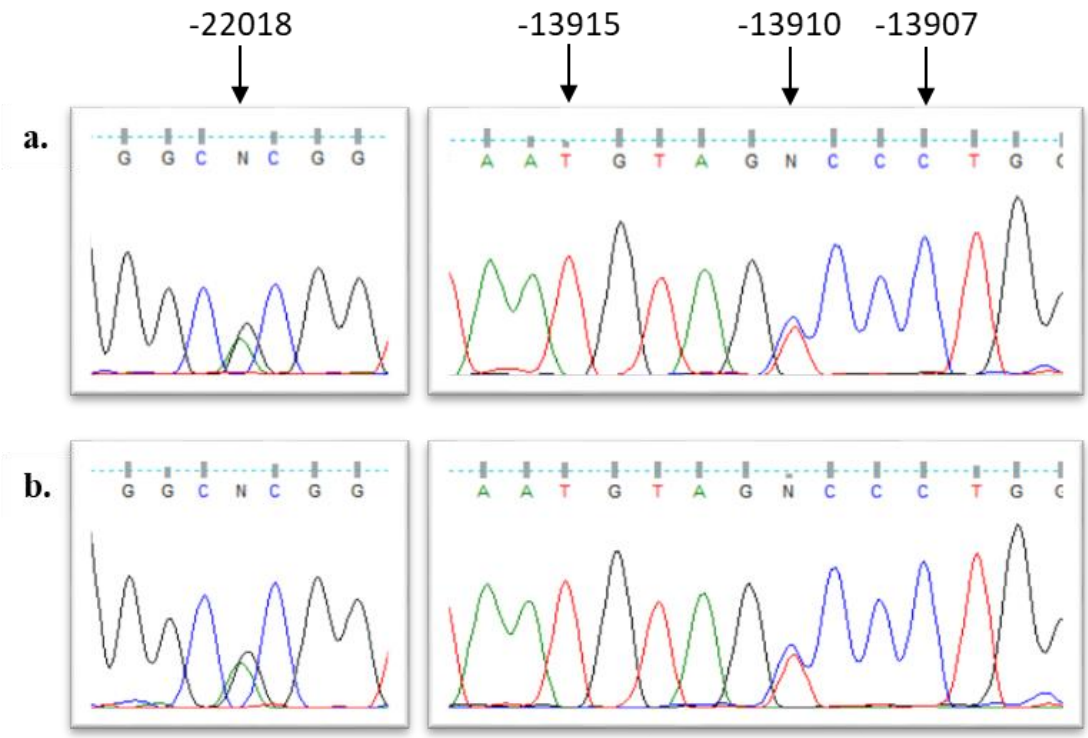
SLC2A1, *SPAST*, *SPG11*, *SPG20*, *SPG21*, *SPG7*, *VAMP1*, *VPS37A*, *WDR45*, *ZFYVE26*, and *ZFYVE27*.

VF was also screened for GLUT1 deficiency syndrome and DYT1 early-onset primary dystonia. However, no clearly pathogenic mutations were detected in VF. This result reduces the likelihood that the symptoms seen in this individual are caused by pathogenic mutations in these genes, suggesting that other rare mutational mechanisms not detectable by this analysis may be present. The difficulty in determining the aetiological basis of complex movement disorders means that a diagnosis of HSP must currently rest on clinical grounds alone.

Lactase persistence

The *LCT* gene is 49.3 kb in length and located on the chromosome 2q21. It contains 17 exons and is translated into a 6 kb transcript (NCBI Reference Sequence NG_008104.1). There are at present five different functional alleles that have been associated with lactase persistence: -14010G>C (rs145946881); -13915T>G (rs41380347); -13910C>T (rs4988235); -13907C>G (rs41525747); and -22018G>A (rs182549). To determine the genotype of these alleles in the twins discordant for lactase persistence (KEL and KIR), approximately 300bp surrounding -13910C>T and ~200bp surrounding -22018G>A was analysed by Sanger sequencing. Both twins had an identical genotype for these SNPs, and no mosaicism could be seen in the sequence traces. The potential lactase non-persistence genotypes that have been reported in people of Northern European ancestry[8,9] were absent in the twins. The SNPs associated with lactase non-persistence in African populations and other genetically diverse groups were also absent.[10–12] These results suggest that both twins are genetically lactase-persistent. Known causes of secondary lactase deficiency could be ruled out based on the clinical history of the twins, such as gastroenteritis, coeliac disease, Crohn's disease, ulcerative colitis, chemotherapy,

or long courses of antibiotics. Thus, other yet unidentified SNPs associated with lactose intolerance could be present in the affected twin.



Supplementary Figure 3. Sanger sequencing in individuals KIR (a) and KEL (b) reveals heterozygosity for variants -13910C>A and -22018G>A, and homozygosity for variants -13907C>G and -13915T>G. Variant -14010G>C was not checked using Sanger sequencing, however it was later confirmed to be homozygous for the G allele with exome sequencing and SNP array data.

Amyotrophic lateral sclerosis

C9orf72 hexanucleotide repeats were screened in four ALS-discordant twin pairs using rpPCR. One MZ twin pair of 58-year-old females of European descent (421 and 422) had abnormally-enlarged (>30) *C9orf72* repeat expansions, although they were reported to be discordant for the disease. A twin pair of 35-year-old males of European descent (242 and 243), had normal and equal numbers of *C9orf72* repeats (n=2). One twin pair (218 and 318) had normal, but different, numbers of *C9orf72* repeats (n=2 and n=8, respectively). Southern blotting was used to confirm and quantify the expansion in subjects 421 and 422. Further, a female twin pair of European descent (LAS and SUS) were screened using an ALS multigene panel; however, no

pathogenic or discordant variants were detected. The affected twin had a 2-year history of bulbar-onset ALS and died at the age of 71. The unaffected, now 75-years-old, has no neuromuscular deficit.

Whole-exome sequencing analysis: VarScan2 and MuTect2

Whole-exome sequencing data were reanalysed by means of VarScan2 and MuTect2 using the annotated variant and genotype attained by the Haplotype Caller-based analysis as reference to explore the possible occurrence of low-frequency variants compatible with a mosaicism state.

As there is a possibility of the unaffected twin having a de novo mutation that is not present in the affected twin, a reverse pairwise analysis was also performed where the affected twin was classified as the 'normal' sample and unaffected twin as the 'tumour' sample (Supplementary Table 5).

The resulting discordant variants were further filtered by excluding those variants that were likely to be non-functional, e.g., synonymous variants and/or variants outside the exonic regions. There are exceptions to this rule, such as for the twin pair discordant for lactase non-persistence (KEL and KIR), where causally-linked variants are likely to be found in intronic regions, with an MAF greater than 0.01.

Consistent with the literature, DNA samples that were LCL-derived yielded a higher-than-average discordant call rate (218 and 318; 421 and 422; 242 and 243). To reduce potential for bias in these samples, only variants found in genes previously implicated in ALS were considered for downstream analysis.

Description	Twin pair	Total germline variants detected	Discordant MuTect2	Discordant VarScan2	Shared discordant variants	VQS >90	Not within segmental duplications	MAF <1%	Exonic and non-synonymous
ALS	LAS	91,434	6,964	5,964	555	63	33	16	4
	SUS	112,182							
ALS	218	97,037	18,302	12,465	10,847	7,857	7,616	651	159
	318	93,543							
ALS	421	85,247	23,796	822	520	257	248	53	11
	422	107,451							
ALS	242	92,502	2,851	2442	346	130	107	32	13
	243	96,855							
Stroke	KG(s)	103,110	858	619	64	4	2	0	0
	HG(s)	98,053							
	KG(b)	105,715	1,302	735	95	8	4	1	0
	HG(b)	105,223							
Lactose intolerance	KEL	94,388	1,573	1,186	116	18	8	2	1
	KIR	108,944							
Inclusion body myositis	AFF	103,017	902	515	58	0	0	0	0
	UNAFF	92,857							
Autism spectrum disorder	RP	100,518	1,895	2,173	115	16	8	0	0
	OH	100,085							
Tourette's syndrome	490	95,279	1,088	1,412	99	7	4	0	0
	489	87,799							
Parkinson's disease	PD821	102,750	1,498	1,148	111	7	7	0	0
	PD161	104,715							
Hereditary spastic paraplegia	LF	101,483	1,537	1,249	99	15	6	2	0
	VF	103,849							
Schizophrenia	RT1b	78,727	1,870	1,490	127	8	6	1	0
	RT1a	104,366							
Schizophrenia	IP16	87,063	1,246	1,243	95	10	8	2	0
	IP17	92,607							

Supplementary Table 4. Filtering protocol for SNVs and indels used to identify differences of functional variants between twin siblings. Discordant variants that were not shared between MuTect2 and VarScan2 were filtered out, and the shared discordant calls were then further filtered according to our exclusion criteria: VQS <90, within segmental duplications (SDs), MAF >1% (as per 1000g, cg69, ExAC), exonic, and non-synonymous.

Description	Twin pair	Total germline variants detected	Discordant MuTect2	Discordant VarScan2	Shared discordant variants	VQS >90	Not within segmental duplications	MAF <1%	Exonic and non-synonymous
ALS	LAS	91,434	2,158	3,600	525	13	10	10	4
	SUS	112,182							
ALS	218	97,037	3,781	3,862	566	243	220	20	4
	318	93,543							
ALS	421	85,247	15,600	-	-	-	-	-	-
	422	107,451							
ALS	242	92,502	14,328	9,976	8,072	5,528	5,343	376	84
	243	96,855							
Stroke	KG(s)	103,110	1,430	914	100	13	8	1	0
	HG(s)	98,053							
	KG(b)	105,715	1,189	676	70	6	5	2	1
	HG(b)	105,223							
Lactose intolerance	KEL	94,388	745	558	85	7	4	4	1
	KIR	108,944							
Inclusion body myositis	AFF	103,017	1,688	1,300	92	11	7	3	0
	UNAFF	92,857							
Autism spectrum disorder	RP	100,518	1,933	2,175	109	13	7	0	0
	OH	100,085							
Tourette's syndrome	490	95,279	2,081	1,941	149	20	10	1	0
	489	87,799							
Parkinson's disease	PD821	102,750	1,276	1,082	106	8	7	2	1
	PD161	104,715							
Hereditary spastic paraplegia	LF	101,483	1,389	1,104	102	12	8	0	0
	VF	103,849							
Schizophrenia	RT1b	78,727	1,354	1,145	115	6	4	1	1
	RT1a	104,366							
Schizophrenia	IP16	87,063	860	912	20	0	0	0	0
	IP17	92,607							

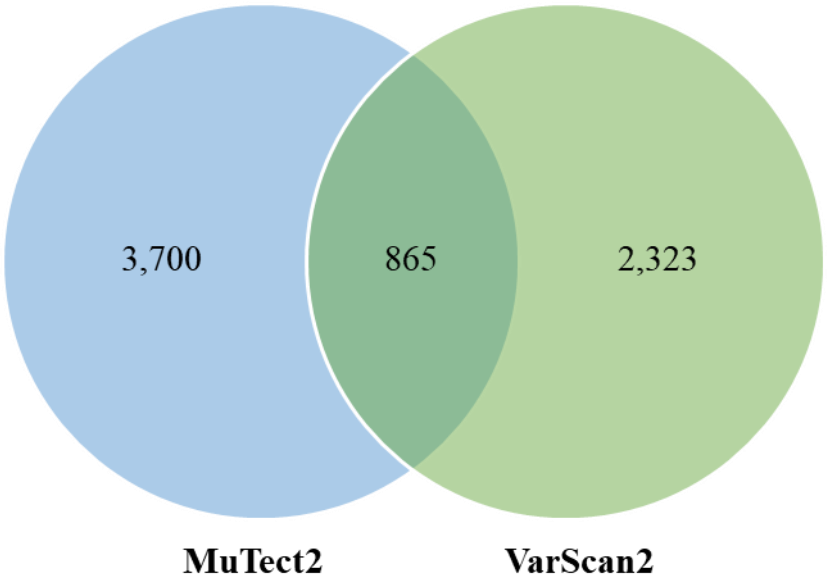
Supplementary Table 5. Somatic variants specific to the affected twin were assumed to be of biological significance, but to determine the total amount of somatic mutations, variants specific to the unaffected twin were also checked. Discordant variants that were not shared between MuTect2 and VarScan2 were filtered out, and the shared discordant calls were then further filtered according to our exclusion criteria: VQS <90, within segmental duplications (SDs), MAF >1% (as per 1000g, cg69, ExAC), exonic, and non-synonymous. The VarScan2 analysis of twins 421 and 422 failed due to technical issues that are currently being resolved. This data is currently omitted.

The bam files for each twin were loaded into IGV and the short read genomic alignments for the potential discordant variants were manually inspected to remove additional artefacts that bypassed prior filtering parameters. These included variants located at the start and end position of reads, base quality scores that were less than 20 on average, homopolymer runs, and variants seen in genomic neighbourhoods with multiple nearby rare variants (suggestive of alignment artefacts caused by nearby indels). These variants were further inspected for intrinsic genome characteristics, such as segmental duplication, micro-satellites, and simple tandem repeats using online genome browsers Ensembl and UCSC.

Twin pair	*Gene*	*Region*	*Type*	*Chr*	*Position*	*Genotype*		*Depth of coverage (proband)*		*Variant frequency (proband)*	*Depth of coverage (MZ twin)*		*Variant frequency (MZ twin)*
						Reference	**Variant**	**Reference**	**Variant**		**Reference**	**Variant**	
LAS and SUS	*DENND5B*	Exon	Stopgain	12	31632586	C	A	48	4	8%	12	0	0%
	LOXHD1	Exon	Nonsynonymous	18	44157785	G	A	150	9	6%	40	0	0%
	BTK	5'UTR	Substitution	X	100645610	C	A	58	0	0%	14	3	18%
	SLC26A1	Exon	Nonsynonymous	4	985253	C	A	70	0	0%	22	3	12%
	MT-ND5	Exon	Nonsynonymous	MT	13596	C	A	227	0	0%	136	10	9%
218 and 318	*FBXO38*	Exon	Nonsynonymous	5	147805180	G	A	11	12	52%	12	0	0%
	RIT2	Exon	Nonsynonymous	18	40323567	C	T	15	11	42%	8	0	0%
	GRM6	Exon	Nonsynonymous	5	178417745	C	T	11	10	48%	14	0	0%
	PPARGC1A	Exon	Nonsynonymous	4	23814713	T	C	69	0	0%	18	5	22%
421 and 422	*KIAA1107*	Exon	Nonframeshift deletion	1	92647643	CTT	-	16	0	0%	26	4	14%
	C8orf48	Exon	Frameshift deletion	8	13425060	AG	-	13	0	0%	22	9	29%
	ATP6V1B2	Exon	Nonframeshift deletion	8	20054932	GATGCGGGG	-	12	0	0%	16	5	24%
242 and 243	*GS1-259H13.2*	Exon	Frameshift deletion	7	99208104	G	-	11	0	0%	15	9	38%

	ACTR3C	5'UTR	Substitution	7	150020654	G	A	10	0	0%	10	8	44%
	KBTBD3	Exon	Nonsynonymous	11	105924526	A	G	20	0	0%	45	7	14%
	TUBGCP4	Exon	Nonsynonymous	15	43678059	C	T	20	0	0%	43	15	26%
	TFIP11	Exon	Frameshift deletion	22	26888040	A	-	70	0	0%	83	48	36%
	PHKA2	Exon	Nonsynonymous	X	18924724	C	T	14	0	0%	28	3	10%
	GNL3L	Exon	Nonsynonymous	X	54581036	A	G	11	0	0%	20	15	42%
KEL and KIR	*PLCB1*	Exon	Frameshift deletion	20	8637865	A	-	37	12	26%	60	0	0%

Supplementary Table 6. Details of candidate discordant variants from WES data after cumulative application of the filters and manual reviewing using IGV. Coordinates refer to human genome build UCSC hg19/GRCh37.



Supplementary Figure 4. Venn diagram illustrating the overlap between MuTect2 and VarScan2. The two somatic mutation callers were used to detect differences between co-twins; only those that were shared were considered for downstream analysis. The figure shows the average number variants called in the total 28 pairwise comparisons.

DNA from the entire twin cohort (n=26) and two sets of parents (n=4) recruited in this study were processed on the Illumina HumanCore BeadChip 12v1. All processing was carried out in accordance with the Infinium HD Ultra Assay protocol (Rev B, 2010, Illumina Inc, San Diego, USA).

CNVs affecting disease-susceptibility genes

After CNV merging, a total of four discordant *de novo* CNV duplications were found in three subjects – namely, chr1:231711489-231803604 in UNAFF; chr4:15776181-15840839 and chr4:139790626-139915883 in RP; and chr8:105981846-106108593 in 490. However manual inspection revealed that these variants are shared with each of the corresponding co-twins. In other words, the apparent discordant CNVs were undercalled (false negative) in the twin that originally didn't have the CNV called by pennCNV. CNVs were called if they are covered by ≥10 probes, but in some instances, they spanned <10 probes so were filtered out of the data.

Next, we lowered the probe threshold to ≥3 to detect smaller CNVs that would potentially be filtered out of the data. As this is expected to result in a higher frequency of false positive calls, CNVs were also called using cnvPartition, and CN segments were only included in further analysis if the CN calls agreed between both algorithms. The results obtained from SNP array analysis are summarised in Supplementary Table 7

These putative CNVs were compared against exome sequencing CNV calls. ExomeDepth yielded approximately 130 CNVs per sample, of which about 90% calls per sample are known in the population. This estimate was determined by comparing the probe locations matching with published CNVs.[13,14] Most of the samples had around a 1:1 deletions/duplications ratio.

We focused on subsets of genes that are associated with known phenotypes in disease databases such as OMIM and DisGeNET, or genes that are intolerant to LoF mutations based on the Residual Variation Intolerance Score (RVIS) or the probability of being loss-of-function intolerant (pLI) score[15] (Table 4). An RVIS <0.0 means that a given gene has less common functional variation than expected, and is referred to as 'intolerant'; whereas an RVIS >0.0 indicates that a gene has more common functional variation. Genes with high pLI scores (pLI ≥ 0.9) are extremely LoF intolerant, whereas genes with low pLI scores (pLI ≤ 0.1) are LoF tolerant.

Disease status	ID	Chr	Start position	End position	CN 1(0or1) 2(3or4)	Size (bp)	No of probes	Genes
ALS	LAS	2	203553836	203665782	2	111946	8	*FAM117B, ICA1L*
		12	31266287	31406907	2	140620	19	*OVOS2*
	SUS	2	203553836	203665782	2	111946	8	*FAM117B, ICA1L*
		12	31296219	31406907	2	110688	15	*OVOS2*
ALS	218	3	160154258	160156933	2	2675	8	*TRIM59*
		21	14669931	14844368	1	174437	9	*FGF7P2, C21orf110*
	318	3	160154258	160156933	2	2675	8	*TRIM59*
		21	14669931	14844368	1	174437	9	*FGF7P2, C21orf110*
ALS	421	-	-	-	-	-	-	-
	422	-	-	-	-	-	-	-
ALS	242	4	116880079	117095295	1	215216	14	-
		4	131965131	132177937	2	212806	17	-
		22	33875161	33928049	2	52888	7	*LARGE1*
	243	4	116880079	117095295	1	215216	14	-
		4	131965131	132218117	2	252986	19	-
		22	33875161	33928049	2	52888	7	*LARGE1*
Stroke	KG(s)	4	111079337	111173378	2	94041	12	*ELOVL6*
		12	106077858	106177127	2	99269	18	-
		20	23677829	23725019	2	47190	8	
	HG(s)	4	111079337	111173378	2	94041	12	*ELOVL6*
		12	106077858	106154088	2	76230	16	
		20	23671740	23725019	2	53279	9	
	KG(b)	4	111079337	111173378	2	94041	12	*ELOVL6*
		12	106077858	106177127	2	99269	12	-
		20	23671740	23725019	2	53279	9	
	HG(b)	4	111079337	111173378	2	94041	12	*ELOVL6*
		12	106077858	106177127	2	99269	10	-
		20	23677829	23725019	2	47190	8	
Lactose Intolerance	KEL	1	49580287	49672039	1	91752	8	*AGBL4*
		19	6904195	7103542	2	199347	40	*ZNF557, MBD3L2, MBD3L5, ADGRE4P, FLJ25758, MBD3L4, MBD3L3, ADGRE1*
	KIR	1	49580287	49672039	1	91752	8	*AGBL4*
		19	6904195	7103542	2	199347	40	*ZNF557, MBD3L2, MBD3L5, ADGRE4P,*

								FLJ25758, MBD3L4, MBD3L3, ADGRE1
Inclusion body myositis	AFF	1	231730121	231767890	2	37769	7	*DISC1*
		7	69966192	70022011	2	55819	8	*AUTS2*
		7	70137165	70295629	2	158464	19	*AUTS2*
		19	54731679	54740705	2	9026	5	*LILRB3*
	UNAFF	1	231711489	231803604	2	92115	11	*DISC1*
		7	69966192	70022011	2	55819	8	*AUTS2*
		7	70014509	70295629	2	281120	31	*AUTS2*
		19	54731679	54740705	2	9026	5	*LILRB3*
Autism spectrum disorder	DS	4	161953729	162002921	2	49192	6	-
	DV	-	-	-	-	-	-	-
	RP	4	139790626	139915883	2	125257	16	-
		4	15776181	15840839	2	64658	10	*CD38*
		4	161953729	162002921	2	49192	6	-
		13	64378676	64390470	2	11794	3	-
	OH	4	139790626	139915883	2	125257	9	-
		4	15776181	15840839	2	64658	10	*CD38*
		4	161953729	162002921	2	49192	6	-
		13	64346536	64390470	2	43934	5	-
Tourette's Syndrome	487	2	90108545	90109261	1	716	5	-
	488	2	90108545	90109261	1	716	5	-
		3	173259356	173289281	2	29925	4	*NLGN1*
		8	105981846	106021780	2	39934	4	
		8	106095659	106115255	2	19596	6	
		10	45218841	45359483	2	140642	14	
		12	8035139	8101326	2	66187	9	*SLC2A3, NANOGP1, NECAP1*
		22	22313954	22560977	2	247023	37	*IGLV4-69, IGLV4-60, IGLV8-61, IGLV6-57, TOP3B*
	490	2	89987044	90109261	1	122217	8	*IGKV2D-29, IGKV2D-28, IGKV2D-26, IGKV3D-20*
		3	173259356	173289281	2	29925	4	*NLGN1*
		5	113429984	113435957	1	5973	4	-
		8	105981846	106108593	2	126747	14	-
		10	45218841	45359483	2	140642	14	-
		22	22313954	22560977	2	247023	33	*IGLV4-69, IGLV4-60, IGLV8-61, IGLV6-57, TOP3B*
	489	2	89987044	90109261	1	122217	8	*IGKV2D-29, IGKV2D-28, IGKV2D-26, IGKV3D-20*
		3	173259356	173289281	2	29925	4	*NLGN1*

		5	113429984	113435957	1	5973	4	-
		8	105981846	106052343	2	70497	9	-
		10	45218841	45359483	2	140642	14	-
		22	22313954	22550078	2	236124	30	*IGLV4-69, IGLV4-60, IGLV8-61, IGLV6-57, TOP3B*
Parkinson's disease	PD821	3	173259356	173289281	2	29925	4	*NLGN1*
		17	34458934	34461869	2	2935	4	-
	PD161	3	173259356	173289281	2	29925	4	*NLGN1*
		17	34450463	34461869	2	11406	5	-
HSP	VF	1	248789519	248813267	1	23748	9	*OR2T11, OR2T35, OR2T27*
		2	40447587	40509154	1	61567	14	*SLC8A1*
		22	25669569	25875573	2	206004	7	*LRP5L*
	LF	1	248789519	248813267	1	23748	9	*OR2T11, OR2T35, OR2T27*
		2	40447587	40509154	1	61567	14	*SLC8A1*
		22	25669569	25905668	2	236099	10	*LRP5L*
Schizophrenia	RT1b	4	161861640	161924832	2	63192	6	-
		13	23548470	23586366	2	37896	18	-
		15	30950529	31088443	1	137914	9	*ARHGAP11B*
		15	32513176	32514341	2	1165	4	-
		22	49565404	49570587	2	5183	6	-
	RT1a	4	161861640	161924832	2	63192	6	-
		13	23548470	23586366	2	37896	18	-
		15	30950529	31088443	1	137914	9	*ARHGAP11B*
		15	32513176	32514341	2	1165	4	-
		22	49566426	49570587	2	4161	5	-
Schizophrenia	IP16	14	32204263	32563640	2	359377	40	*NUBPL, ARHGAP5*
		17	34450463	34461869	2	11406	5	-
		19	54731679	54844626	2	112947	17	*LILRA6, LILRB5, LILRB2, LILRA3, LILRA5*
	IP17	14	32164373	32563640	2	399267	41	*NUBPL, ARHGAP5*
		17	34450463	34461869	2	11406	5	-
		19	54749011	54844626	2	95615	12	*LILRA6, LILRB5, LILRB2, LILRA3, LILRA5*

Supplementary Table 7. CNVs identified by PennCNV and cnvPartition where merged and manually screened in GenomeStudio. False negative CNV calls are included in the list. CN 1 = deletion, CN 2 = duplication. Putative de novo CNVs are highlighted in yellow. Coordinates refer to human genome build UCSC hg19/GRCh37.

Inclusion body myositis (AFF and UNAFF)

AFF presented with a 6-year history of progressive asymmetrical (left worse than right) leg and then arm weakness, with particular weakness in the quadriceps and the finger flexors. Shortly before presentation he had developed dysphagia. Muscle biopsy confirmed the clinical diagnosis of inclusion body myositis. His identical twin brother UNAFF had no clinical symptoms or signs of a neuromuscular disorder on assessment. Analysis of these twins yielded four CNVs. This included a 92115bp CNV duplication containing *DISC1* (chr1:231711489-231803604), located in the region 1q42.2; two CNV duplications overlapping *AUTS2,* which is likely a larger CNV falsely called as two (chr7:69966192-70295629); and finally, a CNV duplication (chr19:54731679-54740705) in the region 19q13.42 overlapping *LILRB3*. An important paralog of this gene is *LILRB1*, which has been associated with idiopathic inflammatory myopathies.[16] Although *DISC1* and *AUTS2* are known to be strongly implicated in schizophrenia[17] and autism[18] respectively, neither twin presented with psychiatric symptoms to our knowledge. CNVs overlapping *AUTS2* and *LILRB3* were validated with ExomeDepth.

Autism and behavioural abnormalities (RP and OH)

Three putative de novo CNVs were detected in both twins but were absent in the parents (Supplementary Table 7). The CNV duplication on chromosome 4p15.32 covers >85% proximal of *CD38,* a gene encoding for a multifunctional ectonucleotidase involved in signal transduction, cell adhesion and calcium signalling. Interestingly, *CD38* has previously been implicated in ADHD,[19] including social memory, amnesia and autism spectrum disorder.[20] This CNV was confirmed by ExomeDepth, but has not been experimentally validated.

Tourette's syndrome (489 and 490)

For an individual to be diagnosed with Tourette's syndrome, they must display multiple motor tics (e.g. blinking or shrugging the shoulders) and at least one vocal tic (e.g. humming, throat clearing, or shouting out a phrase), both of which must be ongoing for at least one year. Tics are abrupt, rapid, non-rhythmic, persistent, stereotyped motor movements or vocalisations. Individuals who have either motor or verbal tics (but not both) for more than a year are given diagnoses of chronic motor tics or chronic verbal tics, respectively.

The forty-four-year-old father (487) of American Indian descent was given a diagnosis of Tourette's syndrome at seven years of age, but experienced motor and vocal tics from the age of six. This follows the typical development of Tourette's syndrome, which manifests in early childhood with symptoms peaking before puberty. Tourette's syndrome can often co-occur with other neuropsychiatric disorders, such as attention deficit hyperactivity disorder (ADHD) and obsessive-compulsive disorder (OCD), and it is often these co-occurring conditions that bring affected individuals to seek medical attention. In line with this, the father had a secondary lifetime clinical diagnosis of OCD (onset nine years) and ADHD (onset twelve years). Although there is no prior family history of Tourette's syndrome, OCD or ADHD, his mother had a diagnosis of anxiety disorder.

The affected twin (490) similarly experienced tics from the age of seven, and was given a diagnosis of Tourette disorder at eight. No other conditions (such as ADHD or OCD) were evident in the affected twin by fifteen years of age, when DNA samples were collected from the family. The unaffected twin (489), moreover, remained asymptomatic for all conditions. The twins have a fourteen-year-old brother who is also asymptomatic

(DNA samples for this individual was not obtained). The forty-four-year-old mother (488) of Caucasian descent has no primary lifetime clinical diagnosis or family history of related disorders. The racial category in the medical notes classify the twins as 'Caucasian', however it is evident that twins have mixed heritage parents.

The twins inherited four CNVs from the mother. An apparent CNV deletion (CN=1) in the mother and father (chr2:90108545-90109261) was also present in the twins, but was expanded by 121,501bp. This resulted in a loss of immunoglobulin kappa variable genes in the twins. CNVs have been found to undergo modification in size when transmitted from parent to offspring (South et al., 2008). However, this CNV should be at best regarded as tentative, as it is in close proximity to the centromere. Various CNV analysis protocols recommend removing called CNVs in HLA regions, and genomic regions that are near centromeres and telomeres. Conversely, enrichments of germline CNVs near assembly gaps and in regions of low-mappability have shown to be reliable (Monlong et al., 2015), as they can be the result of reduced selection pressure, rather than faults of the detection tools.

The multiple CNVs in the twins include maternally-inherited duplications of *NLGN1* and *TOP3B*. A CNV duplication on chromosome 22q11.22 (22313954-22550078) overlapped >90% of *TOP3B*, a gene that has been implicated in neurodevelopmental disorders.[21] Further, a CNV duplication on chromosome 3q26.31 contained *NLGN1*, a gene involved in forming excitatory synapses and maintaining synaptic plasticity.[22]

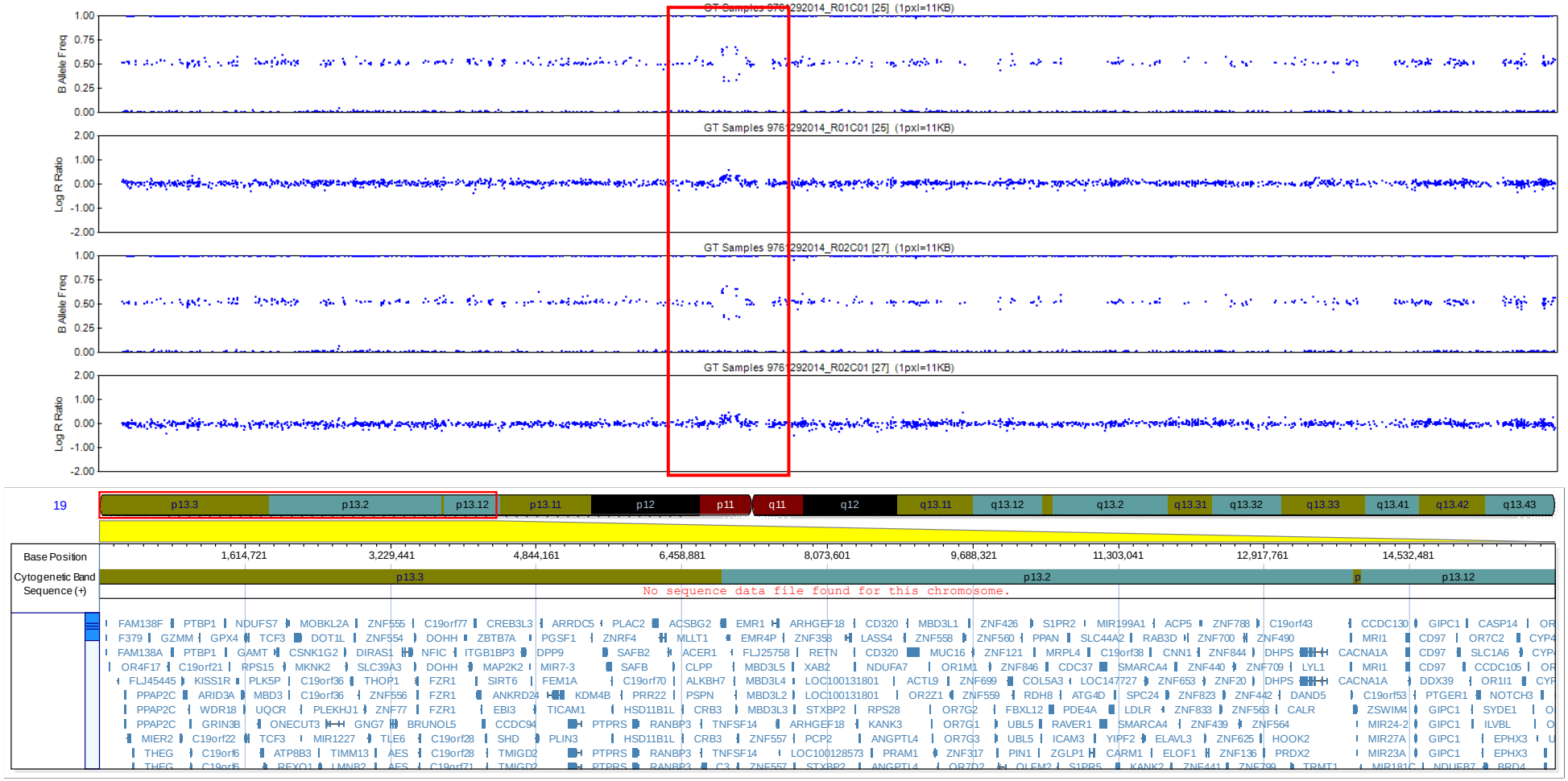
Moreover, both twins have a putative de novo CNV deletion in a non-genic region (chromosome 5q22.3; location 113429984-113435957). These CNVs were not computationally or experimentally validated.

Ischaemic stroke (KG and HG)

DNA from this twin pair was extracted from blood ($HG_{(b)}$ and $KG_{(b)}$) and saliva ($HG_{(s)}$ and $KG_{(s)}$). A CNV duplication on chromosome 4q25 was found in both twins, containing *ELOVL6*. No intra-tissue CNV differences were detected in the SNP array analysis and ExomeDepth analysis confirmed this CNV. *ELOVL6* plays an important role in fat metabolism and insulin sensitivity,[23] and has been associated with heart failure, obesity, atherosclerosis, psoriasis, and atopic dermatitis.[24] These twins have a family history of hypertension, depression and psoriasis, and where diagnosed with atopic dermatitis (HG) and seborrheic dermatitis capitis (KG).

Lactose intolerance (KIR and KEL)

A large CNV duplication spanning 40 probes was detected on chromosome 19p13.2 (6904195-7103542) in both twins. Interestingly, the CNV spans genes that are related to gastrointestinal disorders, including *MBD3L2* (stomach neoplasms), *ADGRE4P* (colorectal cancer metastatic), and *ADGRE1* (liver cirrhosis). Lactose intolerance has shared genes with colorectal cancer (*LCT, CASR, GLB1*), and liver cirrhosis (*CASR and GLB1*).[25] This CNV was validated with ExomeDepth analysis.



Supplementary Figure 5. Results of chromosomal microarray analysis in KEL and KIR. A duplication (CN = 3) is depicted by the BAF plot splitting into two new populations of data points representing the allelic ratios 1:2 and 2:1 (genotypes ABB and AAB). The red rectangle contains the identical 199 kb duplicated genomic region on chromosome 19p13.2.

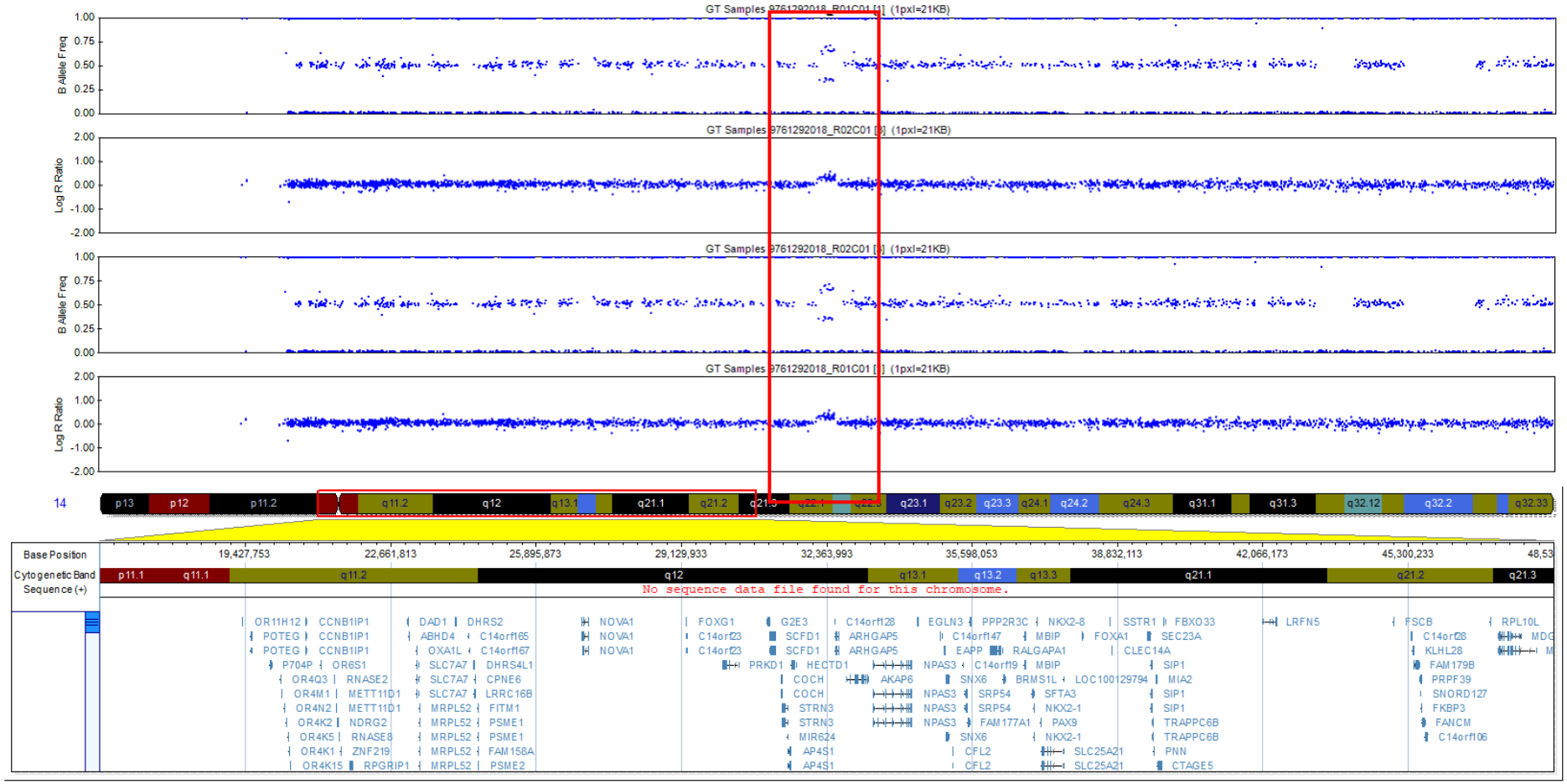
Hereditary spastic paraplegia (VF and LF)

VF was diagnosed with predominant HSP superimposed with dystonia. Her identical twin sister (LF) and sixteen-year-old son later presented with similar, but milder, symptoms. Further clinical details of VF can be found above. A shared hemizygous deletion (CN = 1) of ~60 kb on chromosome 2p22.1 was found in both twins, containing *SLC8A1* (Supplementary Figure 6). Unfortunately, DNA of the proband's son, who was also affected, was not available for segregation analysis.

Supplementary Figure 6. Results of chromosomal microarray analysis in VF and LF. A hemizygous deletion (CN = 1) is depicted as a loss of heterozygotes in the BAF plot and loss of signal intensity in the LRR plot. The red rectangle contains the identical 62 kb deleted genomic region on chromosome 2p22.1.

Schizophrenia (IP16 and IP17)

The largest CNV identified by SNP array analysis was a 399 kb duplication on chromosome 14q12 (32164373-32563640) overlapping genes *NUBPL* and *ARHGAP5*. This was confirmed with exome sequencing CNV calling using ExomeDepth. According to the Database of ClinGen Dosage Sensitivity Map, the haploinsufficiency score of 8.15% (high rank = 0-10%) and pLI score of 0.99 suggests that duplication of *ARHGAP5* may be pathogenic. *ARHGAP5* also appears to be intolerant of variation with a genic intolerance score of -1.31, while being among the top 8% of most intolerant genes.



Supplementary Figure 7. Results of chromosomal microarray analysis in IP16 and IP17. A duplication (CN = 3) is depicted by the BAF plot splitting into two new populations of data points representing the allelic ratios 1:2 and 2:1 (genotypes ABB and AAB). The red rectangle contains the identical 359 kb duplicated genomic region on chromosome 14q12.

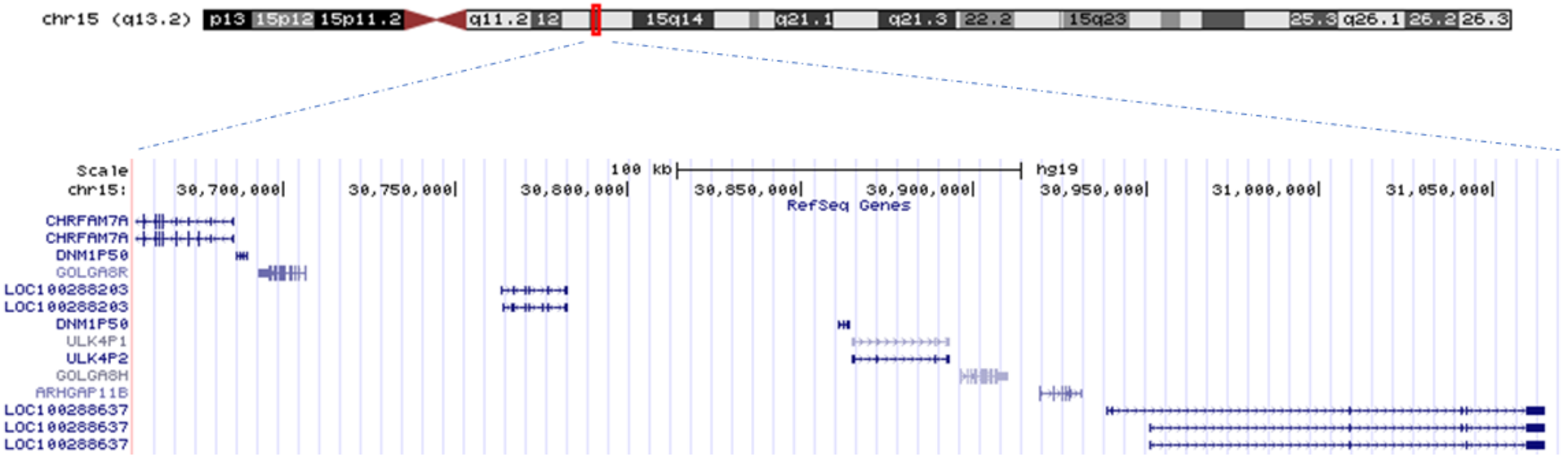
Schizophrenia (RT1a and RT1b)

Both SNP array and exome sequencing CNV calling revealed a 138 kb hemizygous deletion in 15q13.2 (chr15:30950529-31088443) covering *ARHGAP11B* in both twins. Due to lack of probe coverage in this region, it was difficult to ascertain an accurate size of the deletion. The CNV could potentially have extended to left-flanking neighbouring genes, including *CHRFAM7A*.

ARHGAP11B was of particular interest as gene ontology terms include cerebral cortex development. It promotes development and evolutionary expansion of the brain neocortex and it is able to promote amplification of basal progenitors in the subventricular zone, producing more neurons during foetal corticogenesis. CNV deletions containing *ARHGAP11B* has previously been associated with schizophrenia.[26]

The segment overlaps with CNVs recorded in the Database of Genomic Variants (http://dgv.tcag.ca/dgv/app/home). There are various syndromes, including schizophrenia, associated with this gene in the Database of Chromosomal Imbalance and Phenotype in Humans Using Ensemble Resources (DECIPHER), but the region deleted in this case is unique in that it encompasses the genes that were not associated with schizophrenia on their own. ClinVar has 93 records that mention *ARHGAP11B*; 38 of which are deletions, and of these 36 are considered pathogenic. No variants are confined to only *ARHGAP11B*.

Supplementary Figure 8. Results of chromosomal microarray analysis in RT1a and RT1b. A hemizygous deletion (CN = 1) is depicted as a loss of heterozygotes in the BAF plot and loss of signal intensity in the LRR plot. The red rectangle contains the identical 138 kb deleted genomic region on chromosome 15q13.2. The full extent of the deletion cannot be determined due to the absent probes in the left flanking region.



Supplementary Figure 9. Visual representation of protein-coding RefSeq genes that could potentially be included in the CN deletion detected by the SNP array. Screen capture of the deleted region in 15q13.2 from UCSC Genome Browser GRCh37/hg19.

Supplementary References

1 Cibulskis K, Lawrence MS, Carter SL *et al.* Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples. *Nat Biotechnol* 2013; **31**: 213–219.

2 Kumar P, Henikoff S, Ng PC. Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. *Nat Protoc* 2009; **4**: 1073–1081.

3 Adzhubei IA, Schmidt S, Peshkin L *et al.* A method and server for predicting damaging missense mutations. *Nat Methods* 2010; **7**: 248–249.

4 Pirooznia M, Kramer M, Parla J *et al.* Validation and assessment of variant calling pipelines for next-generation sequencing. *Hum Genomics* 2014; **8**: 14.

5 Davydov E V., Goode DL, Sirota M, Cooper GM, Sidow A, Batzoglou S. Identifying a High Fraction of the Human Genome to be under Selective Constraint Using GERP++. *PLoS Comput Biol* 2010; **6**: e1001025.

6 Wang K, Li M, Hadley D *et al.* PennCNV: An integrated hidden Markov model designed for high-resolution copy number variation detection in whole-genome SNP genotyping data. *Genome Res* 2007; **17**: 1665–1674.

7 Plagnol V, Curtis J, Epstein M *et al.* A robust model for read count data in exome sequencing experiments and implications for copy number variant calling. *Bioinformatics* 2012; **28**: 2747–2754.

8 Tishkoff SA, Reed FA, Ranciaro A *et al.* Convergent adaptation of human lactase persistence in Africa and Europe. *Nat Genet* 2007; **39**: 31–40.

9 Ingram CJE, Mulcare CA, Itan Y, Thomas MG, Swallow DM. Lactose digestion and the evolutionary genetics of lactase persistence. *Hum Genet* 2009; **124**: 579–591.

10 Friedrich DC, Santos SEB, Ribeiro-dos-Santos ÂKC, Hutz MH. Several Different Lactase Persistence Associated Alleles and High Diversity of the Lactase Gene in the Admixed Brazilian Population. *PLoS One* 2012; **7**: e46520.

11 Ranciaro A, Campbell MC, Hirbo JB *et al.* Genetic Origins of Lactase Persistence and the Spread of Pastoralism in Africa. *Am J Hum Genet* 2014; **94**: 496–510.

12 Jones BL, Raga TO, Liebert A *et al.* Diversity of Lactase Persistence Alleles in Ethiopia: Signature of a Soft Selective Sweep. *Am J Hum Genet* 2013; **93**: 538–544.

13 Conrad DF, Pinto D, Redon R *et al.* Origins and functional impact of copy number variation in the human genome. *Nature* 2010; **464**: 704–712.

14 Durbin RM, Altshuler DL, Durbin RM *et al.* A map of human genome variation from population-scale sequencing. *Nature* 2010; **467**: 1061–1073.

15 Petrovski S, Wang Q, Heinzen EL, Allen AS, Goldstein DB. Genic Intolerance to Functional Variation and the Interpretation of Personal Genomes. *PLoS Genet* 2013; **9**: e1003709.

16 Schleinitz N, Cognet C, Guia S *et al.* Expression of the CD85j (leukocyte Ig-like receptor 1, Ig-like transcript 2) receptor for class I major histocompatibility complex molecules in idiopathic inflammatory myopathies. *Arthritis Rheum* 2008; **58**: 3216–3223.

17 Ayalew M, Le-Niculescu H, Levey DF *et al.* Convergent functional genomics of schizophrenia: from comprehensive understanding to genetic risk prediction. *Mol Psychiatry* 2012; **17**: 887–905.

18 Oksenberg N, Ahituv N. The role of AUTS2 in neurodevelopment and human evolution. *Trends Genet* 2013; **29**: 600–608.

19 Ebstein RP, Monakhov M, Lai PS, Chew SH. CD38 Gene Expression and Human Personality Traits: Inverse Association with Novelty Seeking. *Messenger* 2014; **3**: 72–77.

20 Higashida H, Yokoyama S, Huang J-J *et al.* Social memory, amnesia, and autism: Brain oxytocin secretion is regulated by NAD+ metabolites and single nucleotide polymorphisms of CD38. *Neurochem Int* 2012; **61**: 828–838.

21 Stoll G, Pietiläinen OPH, Linder B *et al.* Deletion of TOP3β, a component of FMRP-containing mRNPs, contributes to neurodevelopmental disorders. *Nat Neurosci* 2013; **16**: 1228–1237.

22 Hoy JL, Haeger PA, Constable JRL *et al.* Neuroligin1 Drives Synaptic and Behavioral Maturation through Intracellular Interactions. *J Neurosci* 2013; **33**: 9364–9384.

23 Matsuzaka T, Shimano H. Elovl6: a new player in fatty acid metabolism and insulin sensitivity. *J Mol Med* 2009; **87**: 379–384.

24 Uchida Y. The role of fatty acid elongation in epidermal structure and function. *Dermatoendocrinol* 2011; **3**: 65–9.

25 Andrzej P, Piotr M, Borun P *et al.* Influence of lactose intolerance on colorectal cancer incidence in the Polish population. *Hered Cancer Clin Pract* 2015; **13**: A7.

26 Levinson DF, Duan J, Oh S *et al.* Copy Number Variants in Schizophrenia: Confirmation of Five Previous Findings and New Evidence for 3q29 Microdeletions and VIPR2 Duplications. *Am J Psychiatry* 2011; **168**: 302–316.