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

*Nirmal Vadgama[1], Alan Pittman[1], Michael Simpson[2], Niranjanan Nirmalananthan[3], Robin Murray[4], Takeo Yoshikawa[5], Peter De Rijk[6], Elliott Rees[7], George Kirov[7], Deborah Hughes[1], Tomas Fitzgerald[8], Mark Kristiansen[9], Kerra Pearce[9], Eliza Cerveira[10], Qihui Zhu[10], Chengsheng Zhang[10], Charles Lee[10], John Hardy[1], Jamal Nasir[11,12]*

[1]Institute of Neurology, University College London, London WC1N 3BG, UK

[2]Division of Genetics and Molecular Medicine, King's College London, London, UK

[3]St George's University Hospitals NHS Foundation Trust, London, SW17 0QT, UK

[4]Institute of Psychiatry, Psychology, and Neuroscience, King's College, London, UK

[5]RIKEN Brain Science Institute, Wako, Saitama 351-0198, Japan

[6]Applied Molecular Genomics Group, University of Antwerp, Antwerp, Belgium

[7]Centre for Neuropsychiatric Genetics and Genomics, Institute of Psychological Medicine and Clinical Neurosciences, Cardiff University, Cardiff, UK

[8]The European Bioinformatics Institute (EMBL-EBI), Cambridge, UK

[9]UCL Great Ormond Street Institute of Child Health, London WC1N 1EH, UK

[10]Jackson Laboratory for Genomic Medicine, Farmington, CT 06032, USA

[11]Cell Biology and Genetics Research Centre, St. George's University of London.

[12]Present address: Molecular Biosciences Research Group, Faculty of Health & Society, University of Northampton, Northampton, NN1 5PH.

Corresponding author:

Jamal Nasir

Email: jamal.nasir@northampton.ac.uk

**Running Title**

Somatic variation in discordant monozygotic twins

**Conflict of Interest**

The authors declare no conflict of interest.

## Abstract

Recent studies have demonstrated genetic differences between monozygotic (MZ) twins. To test the hypothesis that early post-twinning mutational events associate with phenotypic discordance, we investigated a cohort of thirteen twin pairs (n=26) discordant for various clinical phenotypes using whole-exome sequencing (WES) and screened for copy-number variation (CNV). We identified a *de novo* variant in *PLCB1*, a gene involved in the hydrolysis of lipid phosphorus in milk from dairy cows, associated with lactase non-persistence, and a variant in the mitochondrial complex I gene *MT-ND5* associated with amyotrophic lateral sclerosis (ALS). We also found somatic variants in multiple genes (*TMEM225B, KBTBD3, TUBGCP4, TFIP11*) in another MZ twin pair discordant for ALS. Based on the assumption that discordance between twins could be explained by a common variant with variable penetrance or expressivity, we screened the twin samples for known pathogenic variants that are shared and identified a rare deletion overlapping *ARHGAP11B*, in the twin pair manifesting with either schizotypal personality disorder or schizophrenia. Parent-offspring trio analysis was implemented for two twin pairs to assess potential association of variants of parental origin with susceptibility to disease. We identified a *de novo* variant in *RASD2* shared by 8-year-old male twins with a suspected diagnosis of autism spectrum disorder (ASD) manifesting as different traits. A *de novo* CNV duplication was also identified in these twins overlapping *CD38*, a gene previously implicated in ASD. In twins discordant for Tourette's syndrome, a paternally inherited stop loss variant was detected in *AADAC*, a known candidate gene for the disorder.

**Key Words:** *monozygotic twins; lactose intolerance; ALS; schizophrenia, de novo variants; somatic mosaicism.*

## Introduction

Twin studies have laid the foundation for exploring the genetics of complex traits and common diseases,[1] and heritability estimates for conditions such as schizophrenia (0.79),[2] ALS (0.61),[3] ASD (0.95)[4] and Tourette's syndrome (0.77).[5] The case co-twin design using identical twins discordant for a trait or disorder has thrown light on the determinants of health and disease and causes of individual differences in normal and abnormal behaviour. Recent advancements in DNA sequencing techniques has led to significant methodological improvements, allowing the comparison of twin genomes up to the base pair level, which has challenged the assumption that MZ twins are genetically identical.

The underlying genetic differences between co-twins, including single nucleotide variants (SNVs), indels, gene conversion, CNVs and postzygotic mitotic recombination, can arise during embryonic development. These variations have been proposed as potential genetic mechanisms responsible for discordant MZ twins.[6] Therefore, comparing the genomes of discordant MZ twins signifies a promising opportunity for the search for novel variants implicated in disease, which may ultimately narrow the conceptual gap of missing heritability.

Studies of *de novo* SNVs in parent-offspring trios have identified a mutation rate estimated to range between (0.82-1.70) x $10^{-8}$ mutations per base per generation.[7–9] Post-twinning mutations result in somatic mosaicism, a phenomenon defined as two or more genetically distinct populations of cells.[10] Francioli et al.[11] compared 350 validated variants in MZ twins and found that ~97% of the variants were germline and ~3% were somatic. In another study, an important fraction of *de novo* mutations presumed to be germline in fact occurred either post-zygotically in the offspring, or were inherited because of low-level mosaicism in one of the parents.[12] The mitotic events during embryogenesis resulting in somatic mosaicism may thus play an important and significant role in human disease and MZ twin discordance. Postzygotic *de novo*

SNVs and CNVs have been found in MZ twins discordant for a range of disorders, including frontometaphyseal dysplasia,[13] Dravet's syndrome,[14] Proteus syndrome,[15] attention problems,[16] and schizophrenia,[17]

We performed WES and CNV analysis of DNA from thirteen twin pairs discordant for a range of complex disorders. For a twin pair discordant for Tourette's syndrome and another pair where both exhibit signs of ASD, we were able to perform parent-offspring trio analysis using DNA obtained from the parents. With the aim to identify potential genetic factors that influence disease manifestation, it was hypothesised that *de novo* genetic mechanisms could increase the risk of disease onset. To investigate this hypothesis, the burden of rare SNVs, indels and CNVs overlapping disease-related genes were analysed.

Considering the estimated somatic mutation rate is low and that these variants can be obfuscated by the relatively high error rate of NGS,[18] a highly sensitive filtering method with high specificity, sequence resolution and coverage should be implemented. We provide evidence for *de novo* SNVs, indels and CNVs in disease-related genes associated with a variety of clinical phenotypes presenting in discordant MZ twins. In addition, the identification of potentially pathogenic variants shared by discordant MZ twins, a possibility which has been overlooked in earlier studies, suggests phenotypic discordance could be explained by a common variant modulated by a secondary genetic or epigenetic mechanism, including variants in regulatory regions or modifier genes.

## Materials and Methods

Sample procurement

Subjects were thirteen twin pairs discordant for a range of clinical phenotypes (see Table 1 and Supplementary Material), including the parents of two twin pairs discordant for ASD and Tourette syndrome. DNA samples of five of the twin pairs were obtained from the Coriell Cell

Repository (https://coriell.org/). Peripheral blood and/or buccal samples were collected from subjects recruited into the study after obtaining written informed consent.

Whole-exome sequencing

WES libraries were prepared with Agilent SureSelect V6 and sequenced on an Illumina HiSeq3000 using a 75-bp paired-end reads protocol (see Supplementary Material).

Sequencing data analysis

We have an in-house set of approximately six thousand exomes (from controls and rare diseases) for cross-checking any shortlisted candidate variants, and sequencing artefacts. Sequence alignment to the human reference genome (UCSC hg19), and variant calling and annotation were performed with our in-house pipeline. 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 Genome Analysis Toolkit (GATK) best practices.[19]

Mosaic variants were identified with GATK MuTect2 (version 2.0) and VarScan2 (version 2.4.3), using each pair as reference to one-another. The raw list of SNVs and indels were then filtered using ANNOVAR. [20] 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 NHLBI Exome Variant Server, Complete Genomics 69 (cg69), Exome Aggregation Consortium (EcAC) and 1000 Genomes). *In silico* prediction of pathogenicity was assessed using SIFT,[21] PolyPhen2,[22] and MutationTaster.[23] Conservation of nucleotides was scored using Genomic Evolutionary Rate Profiling (GERP).[24] For the parent-offspring trio analyses and detection of shared variants, joint genotyping was performed on all samples (n=32). False positive variants were removed based on the Variant Quality Score Recalibration (tranche sensitivity <99.00), including low quality

variants that had low depth of coverage (DP <10) and poor genotype quality (GP <20). Potentially pathogenic variants were submitted to the variant database LOVD (https://databases.lovd.nl/shared) submission IDs 00207897–00207091 and 00210051–00210057.

Variant validation by Sanger sequencing

DNAs were amplified by polymerase chain reaction (PCR), using primers specific to the resulting discordant indels and SNVs (Table 2). Sanger sequencing was performed on an ABI 3730XL Genetic Analyzer (PE Applied Biosystems, Forest City, CA, USA) to validate the variants. Forward and reverse primer sequences for the candidate loci are listed in Supplementary Table 3.

Genome-wide SNP genotyping

Genotyping was performed according to the manufacturer's instructions using the Illumina HumanCore v12 BeadChip (Illumina Inc., San Diego, CA, USA). In the quality control, sample sex and twin zygosity were genotypically confirmed, and no samples were excluded based on a genotyping call rate threshold of >0.997.

Copy number variant detection

Log R Ratios (LRR) and B-allele frequencies (BAF) were generated using Illumina Genome Studio software (v2011.1) and used to call CNVs with PennCNV.[25] CNV calling was performed following the standard protocol and adjusting for GC content. 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. cnvPartition was used as the secondary CNV detection algorithm using the default parameters. All CNV coordinates are according to UCSC build 37/hg19.

Computational validation by ExomeDepth

CNVs called by PennCNV and cnvPartition, were corroborated by computational validation with ExomeDepth. CN calls that were shared by all three calling algorithms (PennCNV, cnvPartition, and ExomeDepth) were considered high-confidence CNVs.

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.

## Results

Identifying discordant variants

WES data were analysed by 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* variant 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 Tables 4 and 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 have been reported in intronic regions, with an MAF greater than 0.01.[26]

After applying our stringent filtering criteria, twenty putative discordant variants were identified across all 13 twin pairs (Supplementary Table 6). However, only five of these variants could be validated with Sanger sequencing (Table 2), including a somatic variant in the *PLCB1* gene in the twin affected with lactase non-persistence (KEL and KIR) and variants in four genes (*TMEM225B, KBTBD3, TUBGCP4, TFIP11*) in the unaffected twin of an ALS-discordant pair (242 and 243 (Figure 1). The success rate of 20% is due in part to the limited sensitivity of Sanger sequencing to detect alleles with a frequency of less than 15%.[27] A comparison of saliva and blood tissues for the twin pair discordant for ischaemic stroke did not reveal any evidence of somatic mosaicism.

Identifying shared function-altering variants

To test the hypothesis that the same rare, dominant or recessive variants could contribute to the phenotype for given twin pairs where no discordant variant was found, all potentially damaging shared exonic variants were examined, including known disease-associated loci that could explain disease onset according to a model taking into account the possibility of incomplete penetrance or variable expressivity.

After application of the filtering criteria (see Materials and Methods), each twin pair shared approximately 200 rare variants that are predicated to affect protein function. These were screened against lists of disease-specific susceptibility genes, which were obtained from various databases, including PubMed, OMIM, NIH GTR, DisGeNET, DISEASES, ALSoD, ALSGene, PDGene, SZDB, and SZGene. This produced a total of 113 variants for the entire cohort. These variants were further reduced by removing those found in multiple other samples, repetitive sequences or systematic mismapping of paralogous sequences. Variants that were unambiguous upon manual inspection in IGV were retained. In total, 23 shared variants were identified in known disease-susceptibility genes. These variants with their functional categories

are shown in Table 3. Considering that the shared variants between co-twins were absent in all other samples, it would be extremely unlikely to obtain false-positives in the same gene location in both twin siblings.

Parent-offspring trio analysis

The availability of parental DNAs for two pairs of twins discordant for ASD and Tourette's syndrome respectively, allowed us to further investigate the origin of the shared variants. A total of 217,290 variants were called in GATK's joint analysis in the entire cohort. Variants shared by the MZ twins, but absent in their parents' blood samples and in the other twins, were considered to be *de novo* germline variants of parental origin. After applying this initial exclusion criteria, a total of 424 and 412 putative *de novo* SNVs and indels were detected in the two twin pairs discordant for ASD and Tourette's syndrome, respectively. These variants were filtered using the same parameters set for postzygotic *de novo* variant detection. Upon manual review in IGV, variants could be further excluded on the basis that they were miscalled in one of the parents.

A nonsynonymous variant in *RASD2*, a gene encoding for a GTP-binding protein Rhes on chromosome 22 (NM_014310.3:c.170G>A:p.(Arg57His)), was found in the twins with ASD discordant for severity. This variant is not reported in the dbSNP, 1000 Genomes, cg69 nor in the in-house database of 6,000 exomes. In the ExAC database containing more than 60,000 human exome data, the variant was found with an allele frequency of 8.13E-06 in the total population (allele count of 1/121112). The variant is also highly conserved across multiple species and predicted to be deleterious in online available bioinformatics tools.[21–23]

No shared *de novo* variant of parental origin was detected in the twins discordant for Tourette's syndrome after application of our stringent filtering criteria described above. Nevertheless, as the father of the twins also had the condition, it is likely that both twins had inherited variants

associated with the disorder. We focused our attention on variants consistent with a dominant mode of inheritance – namely, variants that are homozygous or heterozygous in the affected father, absent in the mother, and heterozygous in the twins.

After filtering for rare or novel variants that were predicted to be damaging by at least one of the pathogenicity prediction tools led to the identification of 41 variants inherited from the father. Only one of these variants have previously been implicated in Tourette's syndrome, from our comprehensive list of 138 genes mined from various databases and search of literature. This stop loss variant in *AADAC*, a gene encoding for arylacetamide deacetylase on chromosome 3 (NM_001086.2:c.1198T>C:p.(*400Glnext*1)), plus the *de novo* variant associated with ASD, were validated with Sanger sequencing in the two families (Figure 2).

Mitochondrial DNA analysis

We next tested the hypothesis that different levels of mtDNA heteroplasmy might account for the phenotypic discordance between the twins. After applying a minimum read count threshold of 10, a total of 399 shared and discordant variants were identified between the twins. These variants included 34 heteroplasmic variants and 365 homoplasmic variants. A total of 36 variants unique to either the affected or unaffected twin were verified using IGV. Among these, 23 were distributed on 12 genes throughout the mitochondrial genome, and 8 were localised at the hypervariable segments HV1 (16024–16383) and HV2 (57–372). Most of the discordant variants came from the twin pair 421 and 422. These samples were excluded from further analysis due to the likelihood of artefacts created from high passage transformed immortalised cell lines. Variants in hypervariable segments were also removed. A novel nonsynonymous variant in *MT-ND5* (m.1260C>A:p.(Ser420Arg)), the complex I mitochondrial gene, was detected in the unaffected twin SUS, but was absent in the co-twin affected with ALS, suggesting it might play a protective role. Although this discordant variant had a high depth of

coverage (with the number of reads ranging from 130 to 220), it could not be excluded or confirmed with Sanger sequencing due to a low allele fraction of 9%.

Copy number variant analysis

CNVs were initially called in PennCNV if they are covered by ≥3 probes in order to detect small 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. Putative CNVs were further confirmed against WES CNV calls using ExomeDepth. CN calls that were shared by all three calling algorithms were considered for downstream analysis (Table 4).

For the CNVs shared between co-twins, 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 variants based on the Residual Variation Intolerance Score (RVIS) or the probability of being loss-of-function intolerant (pLI).[28] 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 than expected. Genes with high pLI scores (pLI ≥ 0.9) are extremely LoF intolerant, whereas genes with low pLI scores (pLI ≤ 0.1) are LoF tolerant (Table 4).

No CNV differences between co-twins or tissue types were found. However, four CNVs of parental germline origin were identified by the CN calling algorithms used but were not experimentally validated. Three were found to not overlap any genes or regulatory regions, and one was a CNV duplication found in the twins exhibiting signs of ASD (OH and RP) but

differing in severity and overlapping >85% proximal of *CD38*. This shared *de novo* CNV was also called by ExomeDepth.

The 138kb deletion in 15q13.2 spanning *ARHGAP11B* in the schizophrenia-discordant twins RT1a/RT1b was of particular interest as gene ontology terms include cerebral cortex development. CNV deletions containing *ARHGAP11B* have previously been associated with schizophrenia[29] in numerous studies. The twins are discordant as the co-twin of the proband was diagnosed with schizotypal personality disorder.

## Discussion

We report the successful detection of genomic differences between phenotypically discordant MZ twins. Numerous studies have failed to find somatic variants using NGS technology between discordant twins, with variants often masked by a substantial number of false positives. Such studies include twins discordant for multiple sclerosis,[30] Crohn's disease,[31] congenital hypothyroidism,[32] and ALS.[33] By taking a union of MuTect2 and VarScan2 to identify somatic variants, we offer a proof of concept for assessing the genetic aetiology of complex traits in discordant twins. A comparison of variant detection tools showed that MuTect2 identifies more low coverage somatic variants and has excellent capability in both control of false calls and discovery of potential true positives. VarScan2 was as efficient in low-frequency variants detection but exhibited an advantage in discovering somatic SNVs with relatively high frequencies, which makes it a beneficial supplement of MuTect2.[34]

Our data suggest new variants are relatively common and involve a variety of pre and postzygotic mechanisms affecting nuclear and mitochondrial genes, including epistasis. For some of the twin pairs, no discordant variants were identified, and the shared variants present in disease-susceptibility genes do not readily explain the twins' discordant phenotypes. We postulate discordance could be explained by shared pathogenetic variants resulting in

incomplete penetrance or variable expressivity followed by secondary genetic or epigenetic changes affecting gene regulation. Other potential contributory factors might involve tissue-specific somatic mutations.

Discordant single nucleotide and indel variants

A somatic frameshift deletion in *PLCB1* was detected in a buccal sample, containing epithelial cells of common developmental origin to the cells lining the gut, derived from the affected twin with lactase non-persistence. PLCB1 is highly expressed in the cardia and colon, and in dairy cows this protein has been shown to hydrolyse most of the lipid phosphorus in the low- and high-density lipoprotein fractions of milk.[35] However, the role of this gene in digestive system disorders remains unclear, but warrants further investigations to verify the significance of this variant, if any, in lactase non-persistence.

In the ALS-discordant pair, two discordant nonsynonymous variants were identified in *KBTBD3* and *TUBGCP4*, and two frameshift deletions in *TMEM225B* and *TFIP11* (Table 2). Although these variants were predicted to disrupt protein function, this cannot be reconciled with the fact that they were detected in the unaffected twin. Nevertheless, it is possible that these somatic variants contributed to the phenotypic discordance by having protective effects in the unaffected twin. It is important to note that the DNA samples used for this pair are LCL-derived, and *de novo* mutations are known to be caused by the cell line transformation and culturing. However, several studies have suggested low-passage LCLs to be an appropriate representation of the donor's genome.[36] Nevertheless, we acknowledge that independent validation on DNA from uncultured sources is ideal.

Shared single nucleotide and indel variants

We examined shared variants with predicted pathogenicity. This included rare homozygous and heterozygous variants, and those in known disease-susceptibility genes.

*De novo* variant detection in parent-offspring trio analysis

*Autism spectrum disorder*

A nonsynonymous variant (NM_014310.3:c.170G>A:p.(Arg57His)) of parental germline origin within the RASD family member 2 (*RASD2*) gene was found in both twins with a suspected diagnosis of ASD and behavioural problems (OH and RP) but showing different degrees of severity. *RASD2* belongs to the Ras superfamily of small GTPases and is enriched in the striatum and involved in the modulation of dopaminergic neurotransmission.[37] *RASD2* is located on chromosome 22q12.3, a region that harbours numerous susceptibility loci for psychosis,[38] and has been suggested to be a vulnerability gene for neuropsychologically defined subgroups of schizophrenic patients.[39] Currently, the co-twin has not been diagnosed but anecdotally has been showing clinical features of ASD.

*Tourette's syndrome*

No function-altering germline *de novo* variants were identified in the twin pair. However, a shared novel stop loss variant in *AADAC*, a gene encoding for arylacetamide deacetylase on chromosome 3 (NM_001086.2:c.1198T>C:p.(*400Glnext*1)), was inherited from the father. In a meta-analysis of 1181 patients and 118,730 control subjects, Bertelsen et al.[40] determined a significant association between *AADAC* and Tourette's syndrome. Further, functional studies demonstrated that *AADAC* is expressed in several brain regions previously implicated in the pathophysiology of Tourette's syndrome, including the Purkinje cell layer of the human cerebellum.[40] CNVs overlapping *AADAC* are the first to be successfully associated with Tourette's syndrome. More recently, Yuan et al.[41] found that variants in *AADAC* may be a candidate factor for Tourette's syndrome development in a Han Chinese cohort.

Transcriptome profiling data from The BrainSpan Atlas of the Developing Human Brain (http://www.brainspan.org) illustrates that *AADAC* expression peaks in the striatum between

birth and adolescence. This is consistent with the typical clinical time course of tic onset, and indeed the age of onset of Tourette's syndrome in the father and the affected twin investigated herein. However, other mechanisms such as epigenetics must be considered to explain the asymptomatic twin. Considering the above evidence, the stop loss variant detected in the father and twins warrants functional studies to investigate the role of *AADAC* in the pathogenesis of this disorder.

Copy number shared variants

A shared *de novo* CNV duplication of parental germline origin was detected in twins with a suspected diagnosis of ASD (discordant for severity), overlapping *CD38*, a gene implicated in ADHD,[42] social memory, amnesia and ASD.[43] Although our results don't prove CNV contribution to phenotypic MZ discordance, the pre-twinning structural events detected in this twin cohort could represent a susceptible genetic background (Table 4).

*Schizophrenia*

A deletion in *ARHGAP11B* (CN = 1) and a duplication in *ARHGAP5* (CN = 3) were identified in schizophrenia-discordant twin pairs RT1a/RT1b and IP16/IP17, respectively (Supplementary Figures 7 and 8). Although RT1b was diagnosed with schizophrenia, his co-twin (RT1a) had schizotypal personality features.

The *ARHGAP5* gene product (a GTPase-activating protein for Rho family members) is linked to Ras, and thus to EGF receptor-mediated proliferation, migration and differentiation of forebrain progenitors.[44] Therefore, an *ARHGAP5* duplication in an MZ twin pair discordant for schizophrenia might point to an aetiological basis, because schizophrenia has been linked to altered prenatal neurogenesis of cortical neurons.[45] In addition, *ARHGAP5* and *ARHGAP11B* are contained within regions 14q12 and 15q13.2, respectively, which have previously been associated with schizophrenia.[46,47]

*ARHGAP11B*, resides on chromosome 15q13.2, one of the most complex and unstable loci in the human genome. Several neurodevelopmental disorders have been linked to structural variants in this and nearby regions.[48] *ARHGAP11B* arose from partial duplication of *ARHGAP11A* in the human lineage, approximately one million years after divergence from chimpanzees, but before divergence from Neanderthals.[49] This led to the formation of large and complex human-specific segmental duplications, mediating recurrent rearrangements contributing to 15q13.3 microdeletion syndrome associated with intellectual disability, epilepsy and schizophrenia.[50] *ARHGAP11B* is, to date, the only human-specific gene shown to promote basal progenitor generation and proliferation, including cortical plate augmentation and gyrification induction, and has been proposed to play an important role in the evolutionary expansion of the human neocortex.[49]

The duplicated 8 exons of *ARHGAP11A* are almost identical to the paralogous sequence of *ARHGAP11B*, and thus not completely queried in high throughput genetic studies at the same locus and may have been mixed complex rearrangements. Indeed, variations in this region have flown below the radar of available genome-wide technologies, which likely has downplayed its hypothesised associations with neurodevelopmental disorders. Because of the genomic complexity of the region, the extent of human structural diversity and breakpoints of most rearrangement events are poorly understood at the molecular genetic level. Moreover, the wide expression of *ARHGAP11B*, its multiple functions and modes of regulation – not to mention its absence in other species – present challenges for its study in disease.

Several RhoGAPs have been linked to schizophrenia. For example, a study reported an association between variation in *ARHGAP32*, which encodes a neuron-associated GTPase-activating protein, and schizophrenia and schizotypal personality traits.[51] *ARHGAP33* regulates synapse development and autistic-like behaviour.[52] A missense polymorphism in *ARHGAP3* has been associated with schizophrenia in men.[53] Further, in a genome-wide association study

from the Han Chinese population, Wong et al.[54] identified a schizophrenia susceptibility locus on Xq28, which harbours the gene *ARHGAP4*. This study shows segmental duplications play an important role in normal variation as well as in genomic disease defining hotspots of rearrangement that are susceptible to variation among the normal population. Considering the above findings, we propose that both *ARHGAP5* and *ARHGAP11B* are potentially associated with neuropsychiatric disorders, and this preliminary study provides the necessary baseline to begin future studies on disease populations.

## Conclusion

We have successfully identified pre and postzygotic variants in twin pairs discordant for complex traits. Parent-offspring trio analysis revealed a novel candidate gene for ASD and a novel variant in a gene implicated in Tourette's syndrome.

This study also sheds new light on the genetics of complex disorders including lactase non-persistence, ALS and schizophrenia. We show *de novo* variants are relatively common, could involve multiple genes and invoke multiple mechanisms affecting both nuclear and mitochondrial genes. We also show that discordant MZ twins may share common underlying variants and postulate that additional genetic events, such as epigenetic changes, might lead to phenotypic discordance. Thus, previous studies, mostly based on the premise of a *de novo* mutation in the affected twin, may have missed the opportunity to detect shared variants originating from the parents. Of note, we document a shared pathogenic hexanucleotide repeat expansion in the ALS-discordant twin pair 421 and 422 (see Supplementary Material), which have been previously reported.[55] However, we haven't been able to follow up these and other twins to determine whether there may be delayed onset in the 'unaffected' co-twin. In light of this, it would be worthwhile to re-evaluate earlier twin studies. In addition, functional

validation of the variants reported herein is warranted, a lack of which is acknowledged as a limitation.

## Acknowledgements

We are grateful to the study participants and The Leverhulme Trade Charities Trust for a bursary to NV and The Dowager Countess Eleanor Peel Trust for generously supporting the work (JN).

## References

1 van Dongen J, Slagboom PE, Draisma HHM, Martin NG, Boomsma DI. The continuing value of twin studies in the omics era. *Nat Rev Genet* 2012; **13**: 640–653.

2 Hilker R, Helenius D, Fagerlund B, Skytthe A, Christensen K, Werge TM *et al.* Heritability of Schizophrenia and Schizophrenia Spectrum Based on the Nationwide Danish Twin Register. *Biol Psychiatry* 2018; **83**: 492–498.

3 Al-Chalabi A, Fang F, Hanby MF, Leigh PN, Shaw CE, Ye W *et al.* An estimate of amyotrophic lateral sclerosis heritability using twin data. *J Neurol Neurosurg Psychiatry* 2010; **81**: 1324–6.

4 Colvert E, Tick B, McEwen F, Stewart C, Curran SR, Woodhouse E *et al.* Heritability of autism spectrum disorder in a UK population-based twin sample. *JAMA Psychiatry* 2015; **72**: 415–423.

5 Mataix-Cols D, Isomura K, Pérez-Vigil A, Chang Z, Rück C, Larsson KJ *et al.* Familial Risks of Tourette Syndrome and Chronic Tic Disorders. *JAMA Psychiatry* 2015; **72**: 787.

6 Ketelaar ME, Hofstra EMW, Hayden MR. What monozygotic twins discordant for phenotype illustrate about mechanisms influencing genetic forms of neurodegeneration. *Clin Genet* 2012; **81**: 325–33.

7 Dal GM, Ergüner B, Sağıroğlu MS, Yüksel B, Onat OE, Alkan C *et al.* Early postzygotic mutations contribute to de novo variation in a healthy monozygotic twin pair. *J Med Genet* 2014; **51**: 455–9.

8 Kong A, Frigge ML, Masson G, Besenbacher S, Sulem P, Magnusson G *et al.* Rate of de novo mutations and the importance of father's age to disease risk. *Nature* 2012; **488**: 471–475.

9 Campbell CD, Chong JX, Malig M, Ko A, Dumont BL, Han L *et al.* Estimating the human mutation rate using autozygosity in a founder population. *Nat Genet* 2012; **44**: 1277–1281.

10 Freed D, Stevens EL, Pevsner J. Somatic mosaicism in the human genome. *Genes (Basel)* 2014; **5**: 1064–94.

11 Francioli LC, Polak PP, Koren A, Menelaou A, Chun S, Renkens I *et al.* Genome-wide patterns and properties of de novo mutations in humans. *Nat Genet* 2015; **47**: 822–826.

12 Acuna-Hidalgo R, Bo T, Kwint MP, van de Vorst M, Pinelli M, Veltman JA *et al.* Post-zygotic Point Mutations Are an Underrecognized Source of De Novo Genomic Variation. *Am J Hum Genet* 2015; **97**: 67–74.

13 Robertson SP, Thompson S, Morgan T, Holder-Espinasse M, Martinot-Duquenoy V, Wilkie AOM *et al.* Postzygotic mutation and germline mosaicism in the otopalatodigital syndrome spectrum disorders. *Eur J Hum Genet* 2006; **14**: 549–554.

14 Vadlamudi L, Dibbens LM, Lawrence KM, Iona X, McMahon JM, Murrell W *et al.* Timing of De Novo Mutagenesis — A Twin Study of Sodium-Channel Mutations. *N Engl J Med* 2010; **363**: 1335–1340.

15 Lindhurst MJ, Sapp JC, Teer JK, Johnston JJ, Finn EM, Peters K *et al.* A Mosaic Activating Mutation in AKT1 Associated with the Proteus Syndrome. *N Engl J Med* 2011; **365**: 611–619.

16 Ehli EA, Abdellaoui A, Hu Y, Hottenga JJ, Kattenberg M, van Beijsterveldt T *et al.* De novo and inherited CNVs in MZ twin pairs selected for discordance and concordance on Attention Problems. *Eur J Hum Genet* 2012; **20**: 1037–1043.

17 Tang J, Fan Y, Li H, Xiang Q, Zhang D-F, Li Z *et al.* Whole-genome sequencing of monozygotic twins discordant for schizophrenia indicates multiple genetic risk factors for schizophrenia. *J Genet Genomics* 2017; **44**: 295–306.

18 Kuhlenbäumer G, Hullmann J, Appenzeller S. Novel genomic techniques open new avenues in the analysis of monogenic disorders. *Hum Mutat* 2011; **32**: 144–151.

19 McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A *et al.* The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res* 2010; **20**: 1297–1303.

20 Wang K, Li M, Hakonarson H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res* 2010; **38**: e164–e164.

21 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.

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

23 Schwarz JM, Rödelsperger C, Schuelke M, Seelow D. MutationTaster evaluates disease-causing potential of sequence alterations. *Nat Methods* 2010; **7**: 575–576.

24 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.

25 Wang K, Li M, Hadley D, Liu R, Glessner J, Grant SFA *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.

26 Labrie V, Buske OJ, Oh E, Jeremian R, Ptak C, Gasiūnas G *et al.* Lactase nonpersistence is directed by DNA-variation-dependent epigenetic aging. *Nat Struct Mol Biol* 2016; **23**: 566–73.

27 Beicht S, Strobl-Wildemann G, Rath S, Wachter O, Alberer M, Kaminsky E *et al.* Next generation sequencing as a useful tool in the diagnostics of mosaicism in Alport syndrome. *Gene* 2013; **526**: 474–477.

28 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.

29 Levinson DF, Duan J, Oh S, Wang K, Sanders AR, Shi J *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.

30 Baranzini SE, Mudge J, van Velkinburgh JC, Khankhanian P, Khrebtukova I, Miller NA *et al.* Genome, epigenome and RNA sequences of monozygotic twins discordant for multiple sclerosis. *Nature* 2010; **464**: 1351–1356.

31 Petersen B-S, Spehlmann ME, Raedler A, Stade B, Thomsen I, Rabionet R *et al.* Whole genome and exome sequencing of monozygotic twins discordant for Crohn's disease. *BMC Genomics* 2014; **15**: 564.

32 Magne F, Serpa R, Van Vliet G, Samuels ME, Deladoëy J. Somatic mutations are not observed by exome sequencing of lymphocyte DNA from monozygotic twins discordant for congenital hypothyroidism due to thyroid dysgenesis. *Horm Res Paediatr* 2015; **83**: 79–85.

33 Meltz Steinberg K, Nicholas TJ, Koboldt DC, Yu B, Mardis E, Pamphlett R. Whole genome analyses reveal no pathogenetic single nucleotide or structural differences

between monozygotic twins discordant for amyotrophic lateral sclerosis. *Amyotroph Lateral Scler Front Degener* 2015; **16**: 385–392.

34 Cai L, Yuan W, Zhang Z, He L, Chou K-C. In-depth comparison of somatic point mutation callers based on different tumor next-generation sequencing depth data. *Sci Rep* 2016; **6**: 36540.

35 Cecchinato A, Chessa S, Ribeca C, Cipolat-Gotet C, Bobbo T, Casellas J *et al.* Genetic variation and effects of candidate-gene polymorphisms on coagulation properties, curd firmness modeling and acidity in milk from Brown Swiss cows. *animal* 2015; **9**: 1104–1112.

36 Nickles D, Madireddy L, Yang S, Khankhanian P, Lincoln S, Hauser SL *et al.* In depth comparison of an individual's DNA and its lymphoblastoid cell line using whole genome sequencing. *BMC Genomics* 2012; **13**: 477.

37 Vitucci D, Di Giorgio A, Napolitano F, Pelosi B, Blasi G, Errico F *et al.* Rasd2 Modulates Prefronto-Striatal Phenotypes in Humans and 'Schizophrenia-Like Behaviors' in Mice. *Neuropsychopharmacology* 2016; **41**: 916–927.

38 Potash JB, Zandi PP, Willour VL, Lan T-H, Huo Y, Avramopoulos D *et al.* Suggestive Linkage to Chromosomal Regions 13q31 and 22q12 in Families With Psychotic Bipolar Disorder. *Am J Psychiatry* 2003; **160**: 680–686.

39 Liu Y-L, Fann CS-J, Liu C-M, Chen WJ, Wu J-Y, Hung S-I *et al.* RASD2, MYH9, and CACNG2 Genes at Chromosome 22q12 Associated with the Subgroup of Schizophrenia with Non-Deficit in Sustained Attention and Executive Function. *Biol Psychiatry* 2008; **64**: 789–796.

40 Bertelsen B, Stefánsson H, Riff Jensen L, Melchior L, Mol Debes N, Groth C *et al.* Association of AADAC Deletion and Gilles de la Tourette Syndrome in a Large European Cohort. *Biol Psychiatry* 2016; **79**: 383–391.

41 Yuan L, Zheng W, Yang Z, Deng X, Song Z, Deng H. Association of the AADAC gene and Tourette syndrome in a Han Chinese cohort. *Neurosci Lett* 2018; **666**: 24–27.

42 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.

43 Higashida H, Yokoyama S, Huang J-J, Liu L, Ma W-J, Akther S *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.

44 Fallon J, Reid S, Kinyamu R, Opole I, Opole R, Baratta J *et al.* In vivo induction of massive proliferation, directed migration, and differentiation of neural cells in the adult mammalian brain. *Proc Natl Acad Sci* 2000; **97**: 14686–14691.

45 Akbarian S, Bunney WE, Potkin SG, Wigal SB, Hagman JO, Sandman CA *et al.* Altered distribution of nicotinamide-adenine dinucleotide phosphate-diaphorase cells in frontal lobe of schizophrenics implies disturbances of cortical development. *Arch Gen Psychiatry* 1993; **50**: 169–77.

46 Lavedan C, Licamele L, Volpi S, Hamilton J, Heaton C, Mack K *et al.* Association of the NPAS3 gene and five other loci with response to the antipsychotic iloperidone identified in a whole genome association study. *Mol Psychiatry* 2009; **14**: 804–819.

47 Chen J, Calhoun VD, Perrone-Bizzozero NI, Pearlson GD, Sui J, Du Y *et al.* A pilot study on commonality and specificity of copy number variants in schizophrenia and bipolar disorder. *Transl Psychiatry* 2016; **6**: e824–e824.

48 Antonacci F, Dennis MY, Huddleston J, Sudmant PH, Steinberg KM, Rosenfeld JA *et al.* Palindromic GOLGA8 core duplicons promote chromosome 15q13.3 microdeletion and evolutionary instability. *Nat Genet* 2014; **46**: 1293–1302.

49 Florio M, Albert M, Taverna E, Namba T, Brandl H, Lewitus E *et al.* Human-specific

gene ARHGAP11B promotes basal progenitor amplification and neocortex expansion. *Science (80- )* 2015; **347**: 1465–1470.

50 Dennis MY, Eichler EE. Human adaptation and evolution by segmental duplication. *Curr Opin Genet Dev* 2016; **41**: 44–52.

51 Ohi K, Hashimoto R, Nakazawa T, Okada T, Yasuda Y, Yamamori H *et al.* The p250GAP Gene Is Associated with Risk for Schizophrenia and Schizotypal Personality Traits. *PLoS One* 2012; **7**: e35696.

52 Schuster S, Rivalan M, Strauss U, Stoenica L, Trimbuch T, Rademacher N *et al.* NOMA-GAP/ARHGAP33 regulates synapse development and autistic-like behavior in the mouse. *Mol Psychiatry* 2015; **20**: 1120–1131.

53 Hashimoto R, Yoshida M, Ozaki N, Yamanouchi Y, Iwata N, Suzuki T *et al.* A missense polymorphism (H204R) of a Rho GTPase-activating protein, the chimerin 2 gene, is associated with schizophrenia in men. *Schizophr Res* 2005; **73**: 383–385.

54 Wong EHM, So H-C, Li M, Wang Q, Butler AW, Paul B *et al.* Common Variants on Xq28 Conferring Risk of Schizophrenia in Han Chinese. *Schizophr Bull* 2014; **40**: 777–786.

55 Pamphlett R, Cheong PL, Trent RJ, Yu B. Can ALS-Associated C9orf72 Repeat Expansions Be Diagnosed on a Blood DNA Test Alone? *PLoS One* 2013; **8**: e70007.

## Titles and Legends to Figures

**Figure 1.** IGV screenshots and electropherograms confirming somatic variation in twins discordant for ALS (242 and 243) and lactase non-persistence (KEL and KIR).

**Figure 2**. IGV screenshots and electropherograms confirming a germline *de novo* and inherited variants. **a**. A parental germline *de novo* variant in *RASD2* in twins with behavioural issues and

ASD (OH and RP), which is absent in their parents (DS and DV). **b.** A stop loss variant detected in *AADAC* in twins discordant for Tourette's syndrome, inherited from the affected father.