De novo point mutations and copy number variants in discordant monozygotic twins reveals disease-susceptibility genes

Nirmal Vadgama[1]

Alan Pittman[1], *Micheal Simpson*[2], *Niranjanan Nirmalananthan*[3], *Robin Murray, Takeo Yoshikawa, Peter De Rijk, Elliot Rees, George Kirov, Deborah Hughes, Tomas Fitzgerald*[5], *Mark Kristiansen*[6], *Kerra Pearce*[6], *Eliza Cerveira*[4], *Qihui Zhu*[4], *Chengsheng Zhang*[4],

Charles Lee[4], *John Hardy*[1], *Jamal Nasir*[3]

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

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

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

Cell Biology and Genetics Research Centre, St. George's University of London, London, SW17 0RE, UK

Present Address, University of Northampton.

Corresponding author

Jamal Nasir

Email: jnasir@sgul.ac.uk

Conflict of Interest

The authors declare no conflict of interest.

Abstract

Monozygotic (MZ) twins were long thought to be genetically identical, however recent studies have demonstrated genetic differences between them. To test the hypothesis that early post-twinning mutational events leads to phenotypic discordance, we investigated a cohort of eleven twin pairs discordant for a variety of clinical phenotypes and two concordant twin pairs.

Whole-exome sequencing (WES) data was analysed using a union of VarScan2 and MuTect2 variant calling algorithms. Copy-number variation (CNV) analysis from Illumina HumanCore array data was also carried out using PennCNV and cnvPartition, and corroborated by computational validation with ExomeDepth.

Five discordant variants were validated with Sanger sequencing, four of which were in the unaffected twin of an ALS-discordant pair (*TMEM225B, KBTBD3, TUBGCP4, TFIP11*), and one in the affected twin with lactase non-persistence (*PLCB1*). Parent-offspring trio analysis was implemented for two twin pairs to assess potential association of germline de novo mutations with susceptibility to disease. We identified a novel de novo mutation in *RASD2* shared by 8-year-old male twins concordant for a suspected diagnosis of autism spectrum disorder (ASD). *RASD2* is enriched in the striatum and involved in the modulation of dopaminergic transmission. A de novo CNV duplication was also identified in these twins overlapping *CD38*, a gene implicated in social amnesia and ASD. In twins discordant for Tourette's syndrome, an inherited stop loss mutation was detected in *AADAC*, a known candidate gene for the disorder. Moreover, a rare hemizygous deletion in region 15q13.2 was detected in twins with schizophrenia, overlapping *ARHGAP11B*, a human-specific gene involved in basal progenitor amplification and neocortex expansion.

We provide evidence for de novo point mutations and CNVs in disease-related genes associated with a variety of clinical phenotypes present in MZ twins. We also found potentially pathogenic

mutations shared by discordant twins, suggesting additional genetic mechanisms, including mutations in modifier genes and epigenetic changes, might also play a role in the phenotypes.

Introduction

Over the past decade, investigations of de novo mutations have challenged the assumption that MZ twins are genetically identical. It has been shown that the underlying genetic differences between co-twins, SNVs, indels, gene conversion, CNVs and postzygotic mitotic recombination, may arise during embryonic development. These variations have been proposed as potential genetic mechanisms resulting in discordant MZ twins (1).

Post-twinning mutations result in somatic mosaicism, a phenomenon defined as two or more genetically distinct populations of cells in an individual that were developed from a single fertilised egg (2).

Comparing the genomes of discordant MZ twins signifies a promising opportunity for the search of novel candidate variants implicated in disease, which may ultimately narrow the conceptual gap of missing heritability. With the cost of current NGS sequencing methods dramatically lowering, genome-wide comparisons of MZ twins can be made more affordable and readily available. This circumvents the need to focus merely on a set of candidate genes.

Considering the estimated somatic mutation rate is extremely low and that these variants can be obfuscated by the relatively high error rate of NGS (3), a highly sensitive filtering method with high specificity, sequence resolution and coverage should be implemented.

We performed WES and CNV analysis of DNA from eleven twin pairs discordant for a range of complex disorders, including two concordant twin pairs. 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-linked genes were analysed. 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.

Materials and Methods

Subjects

Written informed consent was obtained from all participants prior to study entry. Further details of subjects can be found on Table 1.

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.

Sequencing data analysis

Sequence alignment to the human reference genome (UCSC hg19), and variant calling and annotation was 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.

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. 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.

In silico prediction of pathogenicity was assessed using SIFT (4), PolyPhen2 (5), and MutationTaster (6). Conservation of nucleotides involved by variants was scored using Genomic Evolutionary Rate Profiling (GERP) (7).

Variant validation by Sanger sequencing

The DNA of five twin pairs was amplified by polymerase chain reaction (PCR), using primers specific to the candidate genes that had the resulting discordant 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 samples with genotyping call rates <95% were excluded.

CNV 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 (8). 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

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.

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.

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 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 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 can be found in intronic regions, with an MAF greater than 0.01.

After applying our stringent filtering criteria, twenty putative discordant variants were identified, details for which can be found in Supplementary Table 6. However, only five of these variants were validated with Sanger sequencing (Table 2). Four of these discordant loci were in the unaffected twin of an ALS-discordant pair (242 and 243), and a somatic mutation was detected in the twin affected with lactase non-persistence (KEL and KIR) (Figure 1).

Identifying shared pathogenic variants

To test the hypothesis that rare, dominant or recessive variants could contribute to the complex disorders investigated, potentially damaging shared exonic variants were examined. Several variants were identified in known disease-associated loci that could potentially explain disease onset according to a model taking into account the possibility of incomplete penetrance.

After application of the filtering criteria, each co-twin typically had 200 potentially damaging, rare shared variants. These were screened against lists of disease-specific susceptibility genes, which were obtained from various databases, including PubMed, OMIM, NIH GTR, DisGeNET, ALSoD, ALSGene, PDGene, SZDB, and SZGene. This produced a total of 113 variants in the twin cohort; however, by manually reviewing them in IGV we could remove mutations that are obvious artefacts of short-read sequencing and alignment. Mutations that were unambiguous 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. These were therefore not validated with Sanger sequencing.

Parent-offspring trio analysis

A total of 217,290 variants were called in GATK's joint analysis. Variants shared by the MZ twins, but absent in their parents, were considered to be de novo germline mutations. After applying this initial exclusion criteria, a total of 424 and 412 putative de novo SNVs and indels were detected in twin pairs discordant for ASD and Tourette's syndrome, respectively. Variants were further filtered as per similar parameters set for postzygotic de novo mutation detection. Upon manual review in IGV most variants could be excluded on the basis that they were falsely miscalled in one of the parents. However, a nonsynonymous mutation in *RASD2*, a gene encoding for a GTP-binding protein Rhes on chromosome 22 (NM_014310:exon2:c.G170A:p.R57H), was found in the twin pair with ASD (discordant for severity). The 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.

Because the father of the twins discordant for Tourette's syndrome 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.

A 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 shared between the twins. Only one of the variants found to be inherited from the father has previously been implicated in Tourette's syndrome, per our comprehensive list of 138 genes mined from various databases and search of literature. This was a stop loss mutation in *AADAC*, a gene encoding for arylacetamide deacetylase on chromosome 3 (NM_001086:exon5:c.T1198C:p.X400Q). This variant was also found in the shared pathogenic variant analysis. Both variants in the two families were validated with Sanger sequencing (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 twin pairs 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 mutation in *MT-ND5* (c.1260A:p.S420R) was detected in twin SUS, which was not present in the co-twin affected with ALS. 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 called if they are covered by ≥3 probes 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. These putative CNVs were compared 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 shared CNVs, 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 (9). 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 (Table 4).

No CNV differences between co-twins or tissue types within a single individual were found. Four pre-twinning de novo CNVs 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), overlapping >85% proximal of *CD38*. This shared de novo CNV was also called by ExomeDepth, and will be validated with droplet digital PCR as part of a future study.

Discussion

<u>De novo mutation detection in parent-offspring trio analysis</u>

Behavioural abnormalities: ADDitude

A nonsynonymous mutation (c.G170A:p.R57H) within the RASD family member 2 (*RASD2*) was found in both twins with a suspected diagnosis of ASD and behavioural problems (OH and RP), but absent in their parents. *RASD2* belongs to the Ras superfamily of small GTPases and is enriched in the striatum and involved in the modulation of dopaminergic neurotransmission (10). *RASD2* is located on chromosome 22q12.3, a region that harbours numerous susceptibility loci for psychosis (11), and has been suggested to be a vulnerability gene for neuropsychologically defined subgroups of schizophrenic patients (12). Currently, the co-twin has not been officially diagnosed but anecdotally has been showing clinical features of ASD.

Commented [GU1]: Don't understand the subtitle entirely: Nothing about ADD in the paragraph.

In a knockout mice model with targeted deletion of *Rasd2*, Vitucci et al. (10) found that the absence of Rasd2 significantly increases the behavioural sensitivity to motor stimulation with administration of psychotomimetic drugs, such as amphetamine and phencyclidine. Based on these findings, and the postulate that *RASD2* influences prefronto-striatal phenotypes in humans, the authors hypothesise that a genetic mutation resulting in a reduction of this G-protein might play a role in cerebral circuitry dysfunction, resulting in exaggerated psychotomimetic drug responses and the development of specific phenotypes linked to schizophrenia-like symptoms (10).

Tourette's syndrome: What makes one tic?

No pathogenic germline de novo mutations were identified in the twin pair. However, a shared stop loss mutation in *AADAC*, a gene encoding for arylacetamide deacetylase on chromosome 3 (NM_001086:exon5:c.T1198C:p.X400Q), was inherited from the father.

In a meta-analysis of 1181 patients and 118,730 control subjects, Bertelsen et al. (13) 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 (13). CNVs overlapping *AADAC* are the first to be successfully associated with Tourette's syndrome. More recently, Yuan et al. (14) found that variants in *AADAC* may be a candidate factor for Tourette's syndrome development in a Han Chinese cohort.

Remarkably, 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. Considering the above evidence, the stop loss mutation detected in the father and twins warrants functional studies to investigate the role of *AADAC* in the pathogenesis of this disorder.

Discordant single nucleotide and indel variants

WES analysis yielded twenty high-confidence discordant variants within the thirteen twin pairs investigated (Supplementary Table 6). We successfully validated five of these in two twin pairs. The success rate of 20% may be due to the limited sensitivity of Sanger sequencing, and thus the variants that were not confirmed may not all be false positives. Validation with a method capable of detecting low-level mosaicism is warranted, such as ddPCR, which can detect mutations with allele fractions of 0.001% (15).

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 mutations 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 mutations contributed to the phenotypic discordance by having disease-modifying effects in the unaffected twin.

Commented [GU2]: protective?

A somatic frameshift deletion in *PLCB1* was also detected in a buccal sample derived from the affected twin with lactase non-persistence. PLCB1 is expressed predominantly in neurons of the central nervous system and is not abundantly expressed in other tissues or cell types (16). 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 (17). However, the role of this gene in digestive system disorders remains unclear, thus warranting further investigations to verify the significance of this mutation, if any, in lactase non-persistence.

Shared single nucleotide and indel variants

In addition to identifying discordant variants, we sought to examine shared variants with predicted pathogenicity. This included rare homozygous and heterozygous variants, and those in known disease-susceptibility genes. Phenotypic discordance between twin pairs could be explained in several ways. For instance, if the variants exhibited incomplete penetrance, the unaffected twin could develop the disorder later. Of note, we document a shared pathogenic hexanucleotide repeat expansion in ALS-discordant twins 421 and 422 (Supplementary Results). Oligogenic mechanisms may also be a factor.

Shared variants in exonic, splice site, promotor, 5'UTR and 3'UTR regions were cross-compared against a list of disease-linked genes formulated via a comprehensive search of literature and gene databases. The number of concordant variants could be further reduced by removing those found in multiple other samples, repetitive sequences or systematic mismapping of paralogous sequences.

As there will likely be a larger number of shared variants in novel genes (previously not related to the primary diagnosis of the affected twin), the search was restricted to only nonsynonymous variants in exonic regions (data not shown). The concordant variants found in WES should ideally be validated with an independent method. Nevertheless, it would be very unlikely for

both twins to receive false-positive readings at the same gene location. The lists of concordant variants identified within each twin pair will thus be of use to future genetic studies related to the disorders investigated.

Copy number shared variants

A shared de novo CNV duplication was detected in twins with a suspected diagnosis of ASD (discordant for severity), overlapping *CD38*, a gene implicated in ADHD (18), social memory, amnesia and ASD (19). Although our results don't demonstrate 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: Mind the GAP.

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). *ARHGAP* gene products belong to the Rho family of GTP-binding proteins, which are involved in membrane/cytoskeletal reorganisation events. There are approximately 80 distinct RhoGAP domain-containing proteins that are encoded in human DNA.

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 (20). 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 (21). In addition, *ARHGAP5* and *ARHGAP11B* are contained within regions 14q12 and 15q13.2, respectively, which have previously been associated with schizophrenia (22,23).

Of special mention is the gene *ARHGAP11B*, which 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 (24,25). *ARHGAP11B* arose from partial duplication of *ARHGAP11A* in the human lineage, approximately one million years after divergence from chimpanzees, but before divergence from Neanderthals (26). 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 (27). Remarkably, all events were related to the chromosome 15 core duplicon containing *GOLGA*, suggesting that these sequences have a fundamental role in the cycles of chromosomal rearrangement and segmental duplication expansions (24). *GOLGA* sequences might possess favoured sites for microhomology-mediated break-induced replication, mechanisms which may induce segmental duplication formations (27).

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 (26).

The duplicated 8 exons of *ARHGAP11A* is almost identical to the paralogous sequence of *ARHGAP11B*, and thus is not completely queried in high throughput genetic studies. 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 non-human animals – present challenges for its study in disease.

Commented [GU3]: would rather use "in other species" than "in non-human animals"

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 (28). *ARHGAP33* regulates synapse development and autistic-like behaviour (29). A missense polymorphism in *ARHGAP3* has been associated with schizophrenia in men (30). Further, in a genome-wide association study from the Han Chinese population, Wong et al. (31) identified a schizophrenia susceptibility locus on Xq28, which harbours the gene *ARHGAP4*.

This study shows that 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

This study supports the polygenic nature of the complex disorders investigated and the threshold model for their manifestation. The discordant MZ twin strategy employed in this study to identify candidate genes involved in the pathogenesis of complex traits is reasonable and practical. Notwithstanding, the actual search for discordant variants is not an easy task. For some of the twin pairs, no discordant variants were identified, and the shared variants present in disease-susceptibility genes do not explain the twins' discordant phenotypes. This suggests that genetic alterations in these twins might lie outside the accessible exonic regions.

Future research directions include systematic whole-genome, methylome, transcriptome and proteome analyses on discordant MZ twins, to identify novel disease-causing candidate genes and elucidate the role of differential gene expression in complex disorders.

Acknowledgements

We are grateful to the study participants and The Leverhulme Trade Charities Trust for a bursary to NV.

References

1. Ketelaar ME, Hofstra EMW, Hayden MR. What monozygotic twins discordant for phenotype illustrate about mechanisms influencing genetic forms of neurodegeneration. Clin Genet. 2012 Apr;81(4):325–33.
2. Freed D, Stevens EL, Pevsner J. Somatic mosaicism in the human genome. Genes (Basel). 2014 Dec 11;5(4):1064–94.
3. Kuhlenbäumer G, Hullmann J, Appenzeller S. Novel genomic techniques open new avenues in the analysis of monogenic disorders. Hum Mutat. 2011 Feb;32(2):144–51.
4. Kumar P, Henikoff S, Ng PC. Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. Nat Protoc. 2009 Jul 25;4(7):1073–81.
5. 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 Apr 1;7(4):248–9.
6. Schwarz JM, Rödelsperger C, Schuelke M, Seelow D. MutationTaster evaluates disease-causing potential of sequence alterations. Nat Methods. 2010 Aug 1;7(8):575–6.
7. 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++. Wasserman WW, editor. PLoS Comput Biol. 2010 Dec 2;6(12):e1001025.
8. 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 Nov 1;17(11):1665–74.

9. Petrovski S, Wang Q, Heinzen EL, Allen AS, Goldstein DB. Genic Intolerance to Functional Variation and the Interpretation of Personal Genomes. Williams SM, editor. PLoS Genet. 2013 Aug 22;9(8):e1003709.

10. 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 Feb;41(3):916–27.

11. 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 Apr;160(4):680–6.

12. 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 Nov 1;64(9):789–96.

13. 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 Mar 1;79(5):383–91.

14. 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 Feb 14;666:24–7.

15. Hindson BJ, Ness KD, Masquelier DA, Belgrader P, Heredia NJ, Makarewicz AJ, et al. High-Throughput Droplet Digital PCR System for Absolute Quantitation of DNA Copy Number. Anal Chem. 2011 Nov 15;83(22):8604–10.

16. Lu G, Chang JT, Liu Z, Chen Y, Li M, Zhu J-J. Phospholipase C Beta 1: a Candidate Signature Gene for Proneural Subtype High-Grade Glioma. Mol Neurobiol. 2016;53(9):6511–25.

17. 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 Jul 31;9(07):1104–12.

18. Ebstein RP, Monakhov M, Lai PS, Chew SH. CD38 Gene Expression and Human Personality Traits: Inverse Association with Novelty Seeking. Messenger. 2014 Jun 1;3(1):72–7.

19. 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 Nov;61(6):828–38.

20. 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 Dec 19;97(26):14686–91.

21. 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 Mar;50(3):169–77.

22. 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 Aug 3;14(8):804–19.

23. 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 May 31;6(5):e824–e824.

24. 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 Dec 19;46(12):1293–302.

25. El-Hattab AW, Smolarek TA, Walker ME, Schorry EK, Immken LL, Patel G, et al. Redefined genomic architecture in 15q24 directed by patient deletion/duplication breakpoint mapping. Hum Genet. 2009 Oct;126(4):589–602.

26. 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 Mar 27;347(6229):1465–70.

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

28. 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. Hashimoto K, editor. PLoS One. 2012 Apr 18;7(4):e35696.

29. 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 Sep 14;20(9):1120–31.

30. 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 Mar 1;73(2–3):383–5.

31. 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

Jul;40(4):777–86.

Titles and Legends to Figures

Figure 1. IGV screenshots and electropherograms confirming somatic mutations 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 mutation. **a**. A germline de novo mutation in *RASD2* in twins with behavioural issues and suspected autism (OH and RP), which is absent in their parents (DS and DV). **b.** A stop loss mutation detected in *AADAC* in twins discordant for Tourette's syndrome, inherited from the affected father.