SUPPLEMENTARY METHODS

Whole exome sequencing in research settings

Individuals 12-19; Individual 31

Of the 10 UK patients reported in this cohort, nine were diagnosed through the Deciphering Developmental Disorders (DDD) research study (http://www.ddduk.org). These nine patients were recruited through regional Clinical Genetics services. The DDD research study is investigating children with severe, undiagnosed developmental delay, and their parents, using a combination of genome-wide assays to detect all major classes of genetic variation in the protein coding portion of the genome. They have recorded clinical information and phenotypes using the Human Phenotype Ontology[1] via a secure web portal within the DECIPHER database.[2]

DNA samples from patients and their parents were analysed by the Wellcome Trust Sanger Institute using high-resolution microarray analysis (array-comparative genomic hybridisation (CGH) and SNP-genotyping) to investigate CNVs in the child, and exome sequencing to investigate SNPs and small insertions/deletions (indels). Putative *de novo* sequence variants were validated using targeted Sanger sequencing. The population prevalence (minor allele frequency) of each variant in nearly 15,000 samples from diverse populations was recorded, and the effect of each genomic variant was predicted using the Ensembl Variant Effect Predictor.[3] Likely diagnostic variants in known developmental disorder genes were fed back to the referring clinical geneticists for validation and discussion with the family via the patient's record in DECIPHER, where they can be viewed in an interactive genome browser. Full genomic datasets were also deposited in the European Genome–Phenome Archive (http://www.ebi.ac.uk/ega).

Individual 29

DNA was extracted from blood samples using established methods and run on a gel to ensure there was no degradation. Concentration and purity of the DNA was quantified and 1ug of genomic DNA was fragmented using sonication, and optimized to give a distribution of 200-500 base pairs. Library preparation was done using Kapa DNA HTP Library Preparation Kit (KAPA Biosystems, 07138008001). Hybridization of the adapter ligated DNA was performed at 47^{o}C, for 64 to 72 hours, to a biotin-labelled probe included in the Nimblegen SeqCap EZ Human Exome Kit (Roche, 06465692001). Libraries were sequenced using the Illumina Hiseq 2500 sequencing system and paired-end 101bp reads were generated for analysis.

Alignment of raw reads was performed using BWA-MEM algorithm onto GRCh37 reference genome. Read groups were added to BAM prior to marking duplicates and sorting using Picard Tool version 1.48. Local realignment was performed around using IndelRealigner module from GATK version 2.7 to reduce the number of mismatching bases across reads. Next, base quality score recalibration (BQSR) module was applied to identify systematic error from sequencing and readjusted for the Phred score. Subsequent variant calling was done using HaplotypeCaller, capable of calling SNPs and indels at the same time using local de-novo assembly of haplotypes in regions that have variability. Join-genotyping was executed by GenotypeGVCFs on the previous VCF output to generate accurate genotype likelihoods by re-genotyping the merged record. Variants were separated to SNPs and indels for variant recalibration using different algorithms. For WES trio analysis, we applied PhaseByTransmission (PBT) to compute the most likely genotype combination given the genotype likelihoods. Finally, the variants were annotated using ANNOVAR version released in March 2015.

From the list of annotation, we retained variants that were rare (allele frequency <1%), coding/protein altering and predicted to be pathogenic by at least one algorithm. The variants were prioritised by further categorising based on inheritance rules such as compound heterozygous, de-novo, homozygous and x-linked recessive variants.

This study was approved by the Singhealth Centralised Institutional Board (CIRB; reference number 2013/798/F).

Computational analysis of facial photographs

We used an ensemble of regression trees to detect a constellation of 68 landmarks on the face.[4] These landmarks were used to create a face mesh using Delaunay triangulation. The averaging algorithm was initialized with a target face mesh, which was created by averaging the facial feature point constellations for 2000 healthy individuals. The face mesh of each patient was aligned to this target mesh with respect to the points across the middle of the face (from forehead to chin). The average face was created by morphing the image of each patient's face onto the average face mesh. To avoid biases towards individuals, for which more images were available, each patient's contribution to the average face mesh was equally weighted. Finally, to avoid variances in illumination between images, which could cause any image in the composite to dominate, we normalized the pixel values within the face to an average value across all faces for each average.

References

1. Robinson PN, Mundlos S. The human phenotype ontology. Clin Genet 2010;**77**(6):525-34 doi: 10.1111/j.1399-0004.2010.01436.x[published Online First: Epub Date]|.
2. Bragin E, Chatzimichali EA, Wright CF, et al. DECIPHER: database for the interpretation of phenotype-linked plausibly pathogenic sequence and copy-number variation. Nucleic Acids Res 2014;**42**(Database issue):D993-D1000 doi: 10.1093/nar/gkt937[published Online First: Epub Date]|.
3. McLaren W, Pritchard B, Rios D, et al. Deriving the consequences of genomic variants with the Ensembl API and SNP Effect Predictor. Bioinformatics 2010;**26**(16):2069-70 doi: 10.1093/bioinformatics/btq330[published Online First: Epub Date]|.
4. Kazemi V, Sullivan J. One Millisecond Face Alignment with an Ensemble of Regression Trees. Proceedings of the 2014 IEEE Conference on Computer Vision and Pattern Recognition: IEEE Computer Society, 2014:1867-74.