
Article:
Akawi, N., McRae, J., Ansari, M. et al. (40 more authors) (2015) Discovery of four recessive developmental disorders using probabilistic genotype and phenotype matching among 4,125 families. Nature Genetics, 47 (11). pp. 1363-1369. ISSN 1061-4036

https://doi.org/10.1038/ng.3410

Published in final edited form as:
Nat Genet. 2015 November ; 47(11): 1363–1369. doi:10.1038/ng.3410.

Discovery of four recessive developmental disorders using probabilistic genotype and phenotype matching among 4,125 families

Nadia Akawi[#1], **Jeremy McRae**[#1], **Morad Ansari**[2], **Meena Balasubramanian**[3], **Moira Blyth**[4], **Angela F. Brady**[5], **Stephen Clayton**[1], **Trevor Cole**[6], **Charu Deshpande**[7], **Tomas W. Fitzgerald**[1], **Nicola Foulds**[8], **Richard Francis**[9], **George Gabriel**[9], **Sebastian S. Gerety**[1], **Judith Goodship**[10], **Emma Hobson**[4], **Wendy D. Jones**[1], **Shelagh Joss**[11], **Daniel King**[1], **Nikolai Klena**[9], **Ajith Kumar**[12], **Melissa Lees**[12], **Chris Lelliott**[1], **Jenny Lord**[1], **Dominic McMullan**[6], **Mary O'Regan**[11], **Deborah Osio**[13], **Virginia Piombo**[1], **Elena Prigmore**[1], **Diana Rajan**[1], **Elisabeth Rosser**[12], **Alejandro Sifrim**[1], **Audrey Smith**[4], **Ganesh J. Swaminathan**[1], **Peter Turnpenny**[13], **James Whitworth**[6], **Caroline F. Wright**[1], **Helen V. Firth**[14], **Jeffrey C. Barrett**[1], **Cecilia W. Lo**[9], **David R. FitzPatrick**[2,$], and **Matthew E. Hurles**[1,$,15] **on behalf of the DDD study**

[1]Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge, CB10 1SA, UK

[2]MRC Human Genetics Unit, MRC IGMM, University of Edinburgh, Western General Hospital, Edinburgh, EH4 2XU, UK

Corresponding Author: Matthew Hurles (meh@sanger.ac.uk), Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge, CB10 1SA, UK, Tel +44 (0)1223 495377; Fax +44 (0)1223 494919.

[$]These authors jointly supervised this work

[15]DDD membership described in Supplementary Note

URLs

Exome Aggregation Consortium: http://exac.broadinstitute.org
PICARD: http://broadinstitute.github.io/picard
Spatial filtering: https://github.com/wtsi-npg/pb_calibration
Variant filtering: https://github.com/jeremymcrae/clinical-filter
Recessive analyses: https://github.com/jeremymcrae/recessiveStats
Phenotype similarity analyses: https://github.com/jeremymcrae/hpo_similarity
Protein Data Bank in Europe: http://pdbe.org
PDBeFold: http://www.ebi.ac.uk/msd-srv/ssm
Annovar: http://www.openbioinformatic.org/annovar/

Accession Codes

Exome sequencing data are accessible via the European Genome-Phenome Archive under accession number EGAS00001000775

Author Contributions

N.A., J.M., S.C., T.W.F., W.D.J., D.K., J.L., Al.S., J.C.B., D.R.F. and M.E.H. developed analytical methods and/or analysed human genotype and phenotype data, Mo.B., A.F.B., Me.B., T.C., C.D., N.F., J.G., E.H., S.J., A.K., M.L., M.O., D.O., E.R., Au.S., P.T. and J.W. phenotyped patients, R.F., G.G., S.S.G., N.K., C.L., V.P. and C.W.L. generated and analysed model organism data, M.A., D.M., E.P. and D.R. performed validation experiments, G.J.S. performed protein structure analysis, C.F.W., H.V.F., J.C.B., D.R.F. and M.E.H. supervised the experimental and analytical work, M.E.H., D.R.F., N.A., J.M. and C.W.L. wrote the manuscript, D.R.F. and M.E.H. jointly supervised the project.

Competing financial interests

MEH is a consultant for and shareholder in Congenica Ltd, which provides genetic diagnostic services

Europe PMC Funders Author Manuscripts

[3]Sheffield Regional Genetics Services, Sheffield Children's NHS Trust, Western Bank, Sheffield, S10 2TH, UK

[4]Yorkshire Regional Genetics Service, Leeds Teaching Hospitals NHS Trust, Department of Clinical Genetics, Chapel Allerton Hospital, Chapeltown Road, Leeds, LS7 4SA, UK

[5]North West Thames Regional Genetics Service, London North West Healthcare NHS Trust, Harrow, HA1 3UJ

[6]West Midlands Regional Genetics Service, Birmingham Women's NHS Foundation Trust, Birmingham Women's Hospital, Edgbaston, Birmingham, B15 2TG, UK

[7]South East Thames Regional Genetics Centre, Guy's and St Thomas' NHS Foundation Trust, Guy's Hospital, Great Maze Pond, London, SE1 9RT, UK

[8]Wessex Clinical Genetics Service, University Hospital Southampton, Princess Anne Hospital, Coxford Road, Southampton, SO16 5YA, UK and Wessex Regional Genetics Laboratory, Salisbury NHS Foundation Trust, Salisbury District Hospital, Odstock Road, Salisbury, Wiltshire, SP2 8BJ, UK and Faculty of Medicine, University of Southampton

[9]Department of Developmental Biology, University of Pittsburgh, Pittsburgh, PA 15201, USA

[10]Northern Genetics Service, Newcastle upon Tyne Hospitals NHS Foundation Trust, Institute of Human Genetics, International Centre for Life, Central Parkway, Newcastle upon Tyne, NE1 3BZ, UK

[11]West of Scotland Regional Genetics Service, NHS Greater Glasgow and Clyde, Institute Of Medical Genetics, Yorkhill Hospital, Glasgow, G3 8SJ, UK

[12]North East Thames Regional Genetics Service, Great Ormond Street Hospital for Children NHS Foundation Trust, Great Ormond Street Hospital, Great Ormond Street, London, WC1N 3JH, UK

[13]Peninsula Clinical Genetics Service, Royal Devon and Exeter NHS Foundation Trust, Clinical Genetics Department, Royal Devon & Exeter Hospital (Heavitree), Gladstone Road, Exeter, EX1 2ED, UK

[14]East Anglian Medical Genetics Service, Box 134, Cambridge University Hospitals NHS Foundation Trust, Cambridge Biomedical Campus, Cambridge, CB2 0QQ, UK

[#] These authors contributed equally to this work.

Abstract

Discovery of most autosomal recessive disease genes has involved analysis of large, often consanguineous, multiplex families or small cohorts of unrelated individuals with a well-defined clinical condition. Discovery of novel dominant causes of rare, genetically heterogenous developmental disorders has been revolutionized by exome analysis of large cohorts of phenotypically diverse parent-offspring trios 1,2. Here we analysed 4,125 families with diverse, rare, genetically heterogeneous developmental disorders and identified four novel autosomal recessive disorders. These four disorders were identified by integrating Mendelian filtering (identifying probands with rare biallelic putatively damaging variants in the same gene) with statistical assessments of (i) the likelihood of sampling the observed genotypes from the general population, and (ii) the phenotypic similarity of patients with the same recessive candidate gene.

This new paradigm promises to catalyse discovery of novel recessive disorders, especially those with less consistent or nonspecific clinical presentations, and those caused predominantly by compound heterozygous genotypes.

We previously noted an exome-wide excess of inherited loss-of-function (LOF) variants in 1,133 individuals with undiagnosed rare developmental disorders (DD)1 but recessive DDs were typically only observed in a single family 3. This N=1 problem for recessive disorders has been well-recognised 4. We hypothesized that by increasing sample size we would observe recurrent instances of both known and novel recessive disorders.

Here we describe an exome analysis of 4,125 mainly simplex families with severe DDs recruited through Clinical Genetics services in the UK and Ireland. Clinical phenotype was systematically recorded using the Human Phenotype Ontology (HPO) 5,6. 1,053 of these families had plausibly causative *de novo* mutations in known disease genes (Methods). To identify candidate recessive genes in the remaining 3,072 families all rare (minor allele frequency, MAF, less than 1%) biallelic (homozygous or compound heterozygous) protein-altering variants were identified in each proband. We observed 74 candidate recessive genes where at least one allele was predicted to be LOF in two or more families.

For four previously weakly implicated (described in one or two families) recessive disorders caused by variants in: *LARP7* 4,7, *LINS* 4,8, *PIGT* 9,10, *COL25A1*11 we identified additional families in whom the phenotypic similarity was sufficient for us to consider that the evidence supporting these new recessive disorders is now compelling (Supplementary Table 1). Furthermore, we identified (*DEPDC5*) or confirmed (*COL9A3* 12) apparently recessive forms of dominant disorders. This phenomenon of rare recessive forms of dominant disorders has been described previously 13,14.

To rigorously assess the evidence supporting each candidate recessive gene, we developed a statistical approach that integrates the probabilities of sampling the observed genotypes and phenotypes by chance (Methods, Figure 1).

We classified protein-altering variants (Methods, Supplementary Figure 1) into two groups of predicted functional consequences: LOF and ‘ functional’ (i.e. protein-altering but not likely LOF, e.g. missense). The probability of drawing N unrelated families with similar biallelic genotypes by chance from the general population, was estimated as the probability of sampling two rare alleles by chance N times from 3,072 random draws using Hardy-Weinberg rules for estimating expected genotype frequencies from allele frequencies (Methods). We estimated this probability separately for two classes of recessive genotype: biallelic LOF and compound heterozygous LOF/functional. First we calculated the cumulative frequency of rare LOF and missense variants using the Exome Aggregation Consortium dataset. The gene-specific cumulative allele frequencies from the ExAC dataset (European, non-Finnish, ancestry) were highly concordant with the frequencies from unaffected parents (of European ancestry) (Supplementary Figure 2), showing that the datasets are comparable. Our method corrects for gene-specific levels of autozygosity (Supplementary Figure 3) and takes account of population structure by matching each proband within the DDD cohort to one of four continental populations within the ExAC

dataset (Methods). We demonstrated that our method is well-calibrated by showing that the number of biallelic rare synonymous genotypes observed in probands closely follows the null distribution (Supplementary Figure 4).

We estimated the probability of sampling N probands from unrelated families with the observed clinical phenotypic similarity using the sum of the maximum Information Content (maxIC) among pairwise comparisons between probands' HPO terms as a summary metric and compared to a null distribution of N probands sampled randomly from among the 4,295 probands studied here (Methods). We demonstrated that this metric is informative by comparing probands sharing protein-altering *de novo* mutations in the same gene to permuted data in which the proband-gene relationship is scrambled (Supplementary Figure 5).

We integrated the p values obtained from the genotypic and phenotypic assessments described above using Fisher's method (Table 1, Supplementary Table 2). In effect, these analyses are testing each and every gene in the genome we can annotate (N=17,370), under two models (biallelic LOF and LOF/functional compound heterozygote), and so we set a conservative threshold for genome-wide significance of 1.44e-6 (0.05/17,370/2). At this threshold, we identified two genes exceeding genome-wide significance, *HACE1* and *KIAA0586*, neither of which had previously been implicated in Mendelian disease in humans.

We identified eight individuals in six families with compound heterozygosity for two apparent LOF variants (N=7) or a LOF and missense variant (N=1) in *KIAA0586* (Figure 2). Five out of the six families had a suspected diagnosis of Joubert syndrome (MIM 213300) with a typical molar tooth sign on brain MR images present in all but one of the affected individuals. Ataxia, hypotonia and Duane anomaly were reported features in several of the affected individuals (Supplementary Table 3).

KIAA0586 encodes TALPID3, which is a centrosomal protein required for ciliogenesis and sonic hedgehog signaling 15. Other genes associated with Joubert syndrome are similarly critical for cilia function 16. Mouse homozygous for a null mutation in the gene encoding Talpid3 lack cilia and are embryonic lethal, dying during organogenesis with randomized left-right patterning; typical characteristics of a ciliopathy 17. All patients were compound heterozygous, sharing the same LOF variant, p.(Arg143Lysfs*4). This variant was observed at a MAF of 0.4% (383/96534) in ExAC, more frequent than the cumulative MAF of other LOF variants in KIAA0586. Therefore we hypothesise that homozygosity for this allele is likely embryonic lethal and that our patients are compound heterozygous for a null allele and a hypomorphic allele.

We identified three individuals from three families with biallelic rare LOF variants in *HACE1*, and an additional family with a homozygous inframe deletion of a single codon in three affected siblings (Figure 3). These variants were associated with intellectual disability and significant abnormality in resting muscle tone; five of the six affected individuals had a combination of hypotonia, dystonia and spasticity. Only one (281381) of the six affected

individuals was ambulant (Supplementary Table 3). Brain MR images show brain atrophy and variable hypoplasia of the corpus callosum (Figure 3)

HACE1 encodes a HECT domain containing E3 ubiquitin ligase, expressed in brain. Notably other brain expressed HECT domain E3 ubiquitin ligases have also been associated with intellectual disability 18. HACE1 regulates Rac1, a small GTPase with diverse roles in signaling. A homozygous knock-out mouse model of *HACE1* has been described 19 which has almost complete pre-weaning lethality, of unknown mechanism.

We also identified four genes (*MMP21*, *PRMT7*, *CSTB* and *COL9A3*) with suggestive initial evidence (p<1e-4) but not meeting genome-wide significance (Table 1), two of which were previously implicated recessive causes of DDs (*CSTB* and *COL9A3*). We further evaluated *MMP21* and *PRMT7* with co-segregation studies, deeper clinical assessment and animal models, and found compelling evidence for recessive causation.

We identified two individuals and one fetus from two families compound heterozygous for different LOF and missense variants in *MMP21,* which encodes a matrix metalloproteinase (Figure 4). Both affected individuals and the affected fetus presented with visceral heterotaxy (situs ambiguous; MIM 306955), including complex heart malformations. The missense mutations lie in close proximity in the conserved zinc-binding site, both predicted to have a major impact on enzymatic activity (Figure 4). Biallelic damaging variants in *MMP21* have been identified in heterotaxy patients in an independent study (Chris Gordon pers comm).

MMP21 has been suggested to function during embryogenesis, and as a matrix metalloproteinase, may modulate cell proliferation and migration through regulating extracellular matrix remodeling 20. Two heterotaxy mutant mouse models, Miri and Koli, were recovered from a phenotype-based mutagenesis screen with pathogenic *Mmp21* missense mutations in the zinc binding domain (Figure 4). Mutants exhibit visceral heterotaxy with heart defects commonly associated with heterotaxy (Figure 4). While heterotaxy can arise from motile cilia defects, videomicroscopy of the embryonic node showed normal cilia motility (Supplementary Movie), suggesting Mmp21 acts downstream of motile cilia.

We identified six affected individuals from three families with compound heterozygous LOF/LOF or LOF/functional variants in *PRMT7.* The associated clinical phenotype is a phenocopy of pseudohypoparathyroidism [PHP; MIM 103580, also known as Albright Hereditary Osteodystrophy, AHO]. Mild intellectual disability with obesity and symmetrical shortening of the digits and posterior metacarpals and metatarsals were observed, similar to the acrodysostosis seen in PHP (Figure 5). Two families shared the same variant at the last base of the first coding exon, which may induce aberrant splicing rather than a missense change (Arg32Thr). *PRMT7* encodes an arginine methyltransferase with several histone substrates 21. Protein modeling indicates that all the observed missense variants are likely to be damaging (Figure 5, Supplementary Figure 6). Shortly after birth (at P10), PRMT7 knock-out mice display significantly reduced body size, reduced weight (-48%), and shortened 5th metatarsal (Supplementary Figure 7). These mice are subviable with only

about 45% the expected number of *Prmt7*$^{tm1a/tm1a}$ pups found at P14. The surviving adult *Prmt7*$^{tm1a/tm1a}$ mice exhibit increased fat mass, reduced length and limb bone anomalies, concordant with the human phenotypes. In addition, reduced bone mineral content and density was observed in *Prmt7*$^{tm1a/tm1a}$ mice, and early onset osteoporosis is observed in AHO. All six affected individuals are female and notably the mouse model exhibits markedly sexually dimorphic phenotypes including bone mineral content, density and 5th metacarpal length changes only significant in females (Figure 5 and Supplementary Table 4). Moreover, a strong female bias has been observed in other AHO-like disorders22,23.

Discovering rare autosomal recessive disorders is challenging, especially for genetically heterogeneous disorders in outbred populations with small families. Only a small fraction of the predicted >1,000 autosomal recessive intellectual disability disorders have yet been discovered 24. New strategies for discovering novel recessive disorders are required. We have proposed the combined use of large-scale ascertainment of small families with diverse clinical presentations, exome sequencing and integrated probabilistic analysis of genotypic enrichment and phenotypic similarity. Integrating genotype and phenotype matching increased power to detect novel recessive disorders while not overly penalizing discovery of new disorders that result in variable or nonspecific clinical presentations (e.g. *HACE1*). Most (10/19) of the families we described here with one of the six recessive disorders shown in Table 1 had a single affected child, nonetheless, these families were enriched for affected siblings compared to the entire cohort (p=6.3e-5, Poisson test), suggesting that including families with affected siblings within our cohort boosts power to detect recessive disorders.

Our method for identifying novel recessive disorders requires systematic genotype and phenotype data on a known number of families, and cannot be applied when collating only partial genotypic or phenotypic data on selected families. Phenotypic diversity among families is fundamental for estimating the significance of phenotypic similarity for a given candidate gene. The power of this approach is maximal when recording of phenotype terms is complete and consistent. For example, the statistical significance of phenotypic similarity for *PRMT7* (Table 1) is based on the initial HPO terms associated with individual 270360, which did not include annotation of the hand anomalies (Figure 5). Adding the HPO term ‘ short metacarpal’ post-hoc increases the significance of the phenotypic similarity p value considerably from 1.49E-03 to 7.00E-05, and consequently the combined genotypic and phenotypic p value for *PRMT7* becomes genome-wide significant (p=1.97E-07). Unbiased clinical re-evaluation of probands sharing the same candidate gene will likely remain valuable.

Recruiting clinicians also recorded whether affected individuals were reminiscent of a recognized genetic syndrome, and for three of the recessive disorders described above, the same suspected syndrome was annotated recurrently (*KIAA0586* – Joubert; *COL9A3* – Stickler; *PRMT7* – Albrights Hereditary Osteodystropy). The statistical significance (Methods) of shared syndrome annotation (*KIAA0586* – 1e-5 ; *COL9A3* – 4.2e-4; *PRMT7* – 8.2e-4) was greater than for HPO phenotype similarity (Table 1), suggesting that computational phenotype analysis does not yet capture all of the ‘ gestalt’ information utilized by experienced clinicians.

Mouse mutants provided additional support for all four novel recessive disorders described above. In principle, similarity between HPO terms and mouse mutant phenotypes 25 could be incorporated into combinatorial genotypic and phenotypic significance testing 26. However, the current lack of consistent phenotyping data on all mouse mutants and the incomplete and biased gene coverage of mouse mutants could lead to significant biases.

Comprehensive discovery of all autosomal recessive causes of DDs will require much larger datasets. Inevitably this will necessitate international collaboration and harmonizing of phenotypic data. The adoption of standard, interoperable phenotype ontologies, such as HPO, and, crucially, ensuring they are applied consistently, will be essential.

Online Methods

Families

4,295 patients with severe, undiagnosed, developmental disorders and their parents (4,125 families) were recruited and systematically phenotyped at 24 clinical genetics centres within the UK National Health Service and the Republic of Ireland. The study has UK Research Ethics Committee approval (10/H0305/83, granted by the Cambridge South REC, and GEN/284/12 granted by the Republic of Ireland REC). Families gave informed consent for participation and specific additional consent for publication of photographs was sought and given by a subset of families.

Clinical data (growth measurements, family history, developmental milestones, etc.) are collected using a standard restricted-term questionnaire within DECIPHER, and detailed developmental phenotypes for the proband are entered using the Human Phenotype Ontology 5.

Exome sequencing

Genomic DNA (approximately 1 μg) was fragmented to an average size of 150 bp and subjected to DNA library creation using established Illumina paired-end protocols. Adapter-ligated libraries were amplified and indexed via PCR. A portion of each library was used to create an equimolar pool comprising 8 indexed libraries. Each pool was hybridised to SureSelect RNA baits (6541 samples with Agilent Human All Exon V3 Plus with custom ELID # C0338371, and 5903 samples with Agilent Human All Exon V5 Plus) with custom ELID # C0338371) and sequence targets were captured and amplified in accordance with manufacturer’ s recommendations. Enriched libraries were subjected to 75 base paired-end sequencing (Illumina HiSeq) following manufacturer’ s instructions.

Mapping of the short read sequences for each sequencing lanelet was carried out using the BWA (version 0.59) 27 backtrack algorithm with the GRCh37 1000 Genomes phase II reference (also known as hs37d5). PCR and optically duplicated reads were marked using Picard (version 1.98) MarkDuplicates. Lanelets were spatially filtered to account for bubble artifacts and quality controlled (passing thresholds on percentage of reads mapped; percentage of duplicate reads marked; various statistics measuring INDEL distribution against read cycle and an insert size overlap percentage). Lanelets were then merged into BAMs corresponding to sample’ s libraries and duplicates were marked again with Picard

after which they were then merged into BAMs for each sample. Finally sample level bam improvement was carried out using GATK (version 3.1.1) 28 and samtools (version 0.1.19) 29. This consisted of re-alignment of reads around known and discovered INDELs followed by base quality score recalibration (BQSR) both using the GATK, lastly samtools calmd was applied and indexes were created. Known INDELs for realignment were taken from Mills Devine and 1000G Gold set and the 1000G phase low coverage set both part of the Broad' s GATK resource bundle version 2.2. Known variants for BQSR were taken from dbSNP 137 also part of the Broad' s resource bundle. Finally, SNV and INDELs were called using GATK HaplotypeCaller (version 3.2.2), this was run in multisample calling mode using the complete dataset. GATK Variant Quality Score Recalibration (VQSR) was then computed on the whole dataset and applied to the individual sample variant calling format (VCF) files.

Annotation of functional consequence

Minor Allele Frequency—To define the rarity of each SNV and INDEL, the VCF is annotated with minor allele frequency (MAF) data from a variety of different sources. The MAF annotations used include data from 4 different populations of the 1000 Genomes project 30 [AMR, ASN, AFR & EUR], the UK10K cohort, the NHLBI GO Exome Sequencing Project (ESP) and an internal DDD allele frequency generated using unaffected parents. For allele matching of SNVs we use an exact match based on a key generated from four values (chromosome, position, reference allele and alternative allele). For allele matching of INDELS we use a less stringent approach where the key is constructed using a different four values (chromosome, position, slice and direction). This key requires both INDELs to be at the same locus (chromosome and position) while the slice is computed based on the DNA sequence difference between the reference and alternative alleles and direction is either deletion or insertion.

Variant Effect Predictor—To define the functional consequence of each variant (SNVs, INDELs and CNVs) annotations from the Ensembl variant effect predictor (VEP) 31 based on Ensembl gene build 76 are added to the VCF file. VEP produces a number of annotations including, SIFT and Polyphen predictions, ensembl transcripts, HGNC gene names and a prediction of the functional consequence for each variant. The transcript with the most severe consequence is selected and all associated VEP annotations are based on the effect that the variant has on that particular transcript.

We categorised variants into two classes of variation from the VEP consequence predictions:

- **Loss of function (LoF)**: transcript_ablation, splice_donor_variant, splice_acceptor_variant, stop_gained, frameshift_variant
- **Functional**: stop_lost, initiator_codon_variant, transcript_amplification, inframe_insertion, inframe_deletion, missense_variant, coding_sequence_variant

In addition to the standard VEP-predicted consequences, we expanded the definition of splice donor variants to include those at the neighbouring last base of coding exons, if it is guanine in the reference sequence, as these are highly conserved within the position weight matrix of splice donor sites 32 and exhibits a skew towards rare variants – a characteristic signature of functional impact – that in unaffected DDD parents is more akin to canonical

splice donor sites than missense variants (Supplementary Figure 1). We identified bases in the human genome at the 3' end of exons, where the base at that site was a guanine (specific to the strand of the transcript). Exon coordinates were obtained from GENCODE release 19 33, and sequence at these positions were extracted from the hs37d5 reference assembly 30. We only included sites which overlapped synonymous or missense variants to exclude sites outside of the coding sequence.

Mendelian filtering

Of the 4,125 families recruited to the DDD study, we excluded 1,053 families from downstream recessive analyses on the basis of having de novo mutations predicted to alter the coding sequence of genes already robustly linked to dominant or X-linked developmental disorders. For the remaining families, we identified rare (maximum minor allele frequency less than 1%), protein-altering variants in the probands that were consistent with a recessive mode of inheritance. The remaining variants were either compound heterozygous variants or homozygous non-reference variants, averaging 3.2 variants per proband.

The numbers of probands in each of the functional categories (biallelic LoF, or compound heterozygous LoF/functional) were tallied for each gene. For families with affected siblings found to share rare recessive genetic variants within a gene, only the eldest sibling was included in the tally, in order to have counts from independent families.

The variant filtering was conducted in python, this analysis used version 0.2.0.

Statistical genotypic assessment

We tested all genes (coordinates from GENCODE release 19) for enrichment of rare (minor allele frequency < 1%), recessive LoF and functional (see above) genotypes in unrelated families. Because allele frequency can vary by population, and this affects the probability of randomly observing recessive genotypes, we calculated the cumulative frequency of rare alleles in each gene from the ExAC dataset version 0.3 (accessed 2015-02) in four ancestral populations: European (excluding Finns) - NFE; African - AFR; East Asian - EAS and South Asian - SAS. Different alternate alleles at multiallelic sites were examined separately, using the most severe consequence relating to each allele. We classified each undiagnosed DDD proband into one of these four ancestral populations by projecting them onto a principal component analysis of 1000 Genomes populations (using SNPRelate): AFR (N=109), EAS (N=15), NFE (N=2799) and SAS (N=297).

Because rare recessive genotypes are much more likely to be observed in autozygous segments (inheriting DNA from a recent common ancestor from both parents, e.g. in the case of consanguinity), we used bcftools roh (version 1.0) to call autozygous segments across the entire exome, and determined the number of probands autozygous across each gene of interest. We then adjusted the probability of sampling two rare LoF alleles for the number of probands autozygous across each gene.

The probability of drawing N unrelated families with recessive genotypes in the same gene by chance was estimated as the binomial probability (including autozygosity) of sampling two rare alleles of a specified functional category N or more times from random draws

matched to our ancestral populations (i.e. 109 AFR, 15 EAS, 2799 NFE, 297 SAS). We determined the entire set of combinations by which N or more families could be distributed across the four ancestral populations, and the probability of observing each combination using population-specific frequencies. We summed these probabilities across all possible combinations to obtain an aggregate probability for sampling N or more families by chance.

An R package was developed to perform these statistical analyses, this analysis used version 0.5.0.

Statistical phenotypic assessment

HPO phenotype similarity—Clinical geneticists referring each proband into the DDD study systematically recorded phenotypes using the Human Phenotype Ontology (HPO) 5. These terms were used to assess the probability that probands sharing the same candidate recessive share more similar clinical phenotypes than expected by chance. Similarity testing used the Human Phenotype Ontology version 2013-11-30.

The similarity of HPO terms between two individuals was estimated as the maximum information content (maxIC) from pairwise comparisons of the HPO terms for the two individuals. For each pairwise comparison of two HPO terms we determined the information content for the most informative common ancestor of the two terms. The information content is calculated as the negative logarithm of the probability of the terms' usage (or any of its descendant terms) within the population of all 4,295 probands in the DDD study.

The summary phenotype similarity score for a set of N probands was estimated as the sum of all the pairwise maxIC scores. The null distribution of this summary phenotype similarity score was simulated by randomly sampling sets of N probands and calculating summary scores as above. The p-value was calculated as the proportion of simulated scores greater than or equal to the observed score.

We demonstrated that this summary phenotype similarity score is both informative and well-calibrated by comparing probands sharing protein-altering de novo mutations in the same gene to permuted data in which the proband-gene relationship is scrambled (Supplementary Figure 5). The permuted data closely follows the expected null distribution.

These analyses were conducted in Python, this analysis used version 0.3.1.

Suspected syndrome similarity—In addition to HPO annotation, recruiting clinicians also annotated whether affected individuals were reminiscent of a recognized phenotypically-defined syndrome. These suspected syndrome annotations were initially standardised to correct, for example, misspellings and abbreviated syndrome names.

The similarity of suspected syndrome terms between two probands was calculated as the negative logarithm of the observed probability of selecting the rarest matched syndrome term between the probands. The observed probability of selecting each term was calculated as the frequency with which the term was used in the probands with at least one suspected syndrome. We compared the similarity of the probands for each candidate recessive gene to

a null distribution constructed by simulating 100000 permutations of randomly sampling probands from the set of probands with at least one suspected syndrome.

Combining genotype-phenotype assessment p values

For each gene, we integrated the p-values obtained from the genotypic and phenotypic assessments described above using Fisher' s method. For the genotypic tests, we had performed tests for two models (biallelic LoF and LoF/functional compound heterozygote). We selected the more significant genotypic test for each gene, since some genes had only biallelic LoF probands, while other genes only had LoF/missense compound heterozygotes. These analyses effectively are testing each and every gene in the genome, under the two models, and so we set a conservative threshold for declaring genome-wide significance of 1.44×10^{-6} (0.05/(17,370*2)).

Code availability

Software for variant filtering analyses available at: https://github.com/jeremymcrae/clinical-filter, this analysis used version 0.2.0.

Software for HPO phenotype similarity analyses available at: https://github.com/jeremymcrae/hpo_similarity, this analysis used version 0.3.1.

Software for recessive genotype enrichment analyses available at: https://github.com/jeremymcrae/recessiveStats, this analysis used version 0.5.0.

Protein modeling

Identification of template protein structures—Protein structures having high sequence homology with the protein to be modelled found by using the sequence similarity search tool at the Protein Data Bank in Europe. Where no hits with a high sequence homology were found, only the relevant domains were used for searching and the sequences of the matches recorded. For modeling PRMT7, a mouse structure with high sequence identity was found for the whole sequence (PDB ID: 4C4A). For MMP21 and HACE1 HECT domain, multiple template sequences were selected for modeling since the sequence identity was borderline for modeling (~30-40%).

Structure Modeling—The Swiss-Model server was used to model the protein structure based on sequence templates chosen in the previous step. The models were downloaded and where multiple template sequence were used, these were superposed using PDBeFold service so as to be on the same coordinate reference for comparison. If the fold of the multiple models were the same, only the model with highest sequence homology was used for further analysis (MMP21: Model built on PDB ID: 1y93; HACE1 HECT Domain: 3tug). Models were built for both the mutated and non-mutated sequence for comparison and validation in the next step.

Mutation Modeling—Mutations were introduced in the modelling programme Chimera 34 and the best rotamer for the new amino acid based on inspection of conformational space around the mutation. The overall structure was optimised using the energy minimize

function in Chimera with 100 cycles of steepest descent and 100 cycles of conjugate gradient energy minimisation. The energy minimised models of the mutated structures were compared with the models built for the mutated sequence in the previous step to validate the modelling strategy.

Model Assessment—Model assessments were done by comparison of the non-mutated and mutated structures in PyMol (The PyMOL Molecular Graphics System, Version 1.7.4 Schrödinger, LLC.). All figures of the structure models were also made using PyMol

Mouse modeling of *Mmp21*

Mouse Mutagenesis Screen and Recovery of *Mmp21* Mutants—*Miri* (B2b873) and *Koli* (b2b2458) mutant lines were recovered from a large-scale recessive mouse *N*-ethyl-*N*-nitrosourea mutagenesis screen conducted in C57BL6/J mice. Fetal echocardiography was used to identify mutants with congenital heart disease (CHD)35,36. Fetuses diagnosed with congenital heart defects were further analyzed by necropsy and imaging using micro computed tomography and/or episcopic confocal microscopy was conducted for detailed histopathological examination of intracardiac anatomy for CHD diagnosis37. This study was approved by the Institutional Animal Care and Use Committee of the University of Pittsburgh and conforms to NIH guidelines.

Whole Exome Sequencing and Mutation Recovery of *Mmp21* Mutations—Genomic DNA from a *Miri* and *Koli* mutant were analyzed by whole mouse exome sequencing analysis using Agilent SureSelect Mouse All Exon Kit V1, followed by Illumina HiSeq 2000 sequencing to achieve a minimum of 50X average target sequence coverage (conducted by BGI Americas). Sequence reads were aligned to the C57BL6 reference genome (mm9) and analysis was carried out using CLCBio Genomic Workbench and GATK software. All variants were annotated with annovar and filtered against dbSNP128 and in-house mouse exome databases with custom scripts. Homozygous coding variants recovered from each mutant line were genotyped across all mutants, with the pathogenic mutation identified as the single mutation consistently homozygous in all mutants. Thus the *Miri* line was shown to harbor a *Mmp21* W177L pathogenic mutation, while *Mmp21* Y325N pathogenic mutation was identified in the *Koli* mutant line. Further breeding of both lines across multiple generations has confirmed the $Mmp21^{W177L}$ and $Mmp21^{Y325N}$ mutations as the pathogenic mutations causing heterotaxy and complex CHD in the *Miri* and *Koli* mutant lines, respectively.

Nodal Cilia Videomicroscopy—The embryonic node from E7.75 embryos was dissected for analysis of cilia motility with 0.35μm microspheres (Polysciences) added to assess flow. Imaging was conducted using a Leica DMIRE2 inverted microscope equipped with a 100x oil objective and differential interference contrast optics. High-speed videos were collected at 200 fps using a Phantom v4.2 camera (Vision Research).

Mouse modeling of *Prmt7*

Mice homozygous for the $Prmt7^{tm1a(EUCOMM)Wtsi}$ allele (hereafter referred to as $Prmt7^{tm1a/tm1a}$) was generated as part of the European Conditional Mouse Mutagenesis

Program and Knockout Mouse Project (EUCOMM/KOMP) projects and Sanger Mouse Genetics Project 38. Mice were generated from embryonic stem cell clone EPD0070_4_F09 and JM8.N4 ES cell line and maintained on a C57BL/6N;C57BL/6-*Tyr*$^{c\text{-}Brd}$. Genotyping was carried out as previously described 39. Animals were housed in specific pathogen-free conditions and placed on a Western high fat diet (Special Diet Services, Witham, UK) from 4 weeks of age with *ad libitum* access to autoclaved, nonacidified water and food and phenotyped according to a previously reported standard pipeline, including Dual-energy X-ray absorptiometry and high-resolution X-ray imaging 40. Bone lengths were derived from the original DICOM files of the left forearm and paw using Sante DICOM Editor 4 program (Santesoft, Athens, Greece). All experiments were performed in accordance with the UK Home Office regulations, UK Animals (Scientific Procedures) Act 1986. Adult phenotyping was performed blind to genotype groups, with a minimum of 7 animals per group.

Prmt7$^{tm1a/tm1a}$ mouse data was analysed using RStudio running R version 3.1.2 and Phenstat package version 2.0.1. This uses a mixed-model framework as described 41 to assess the impact of genotype on phenotype. The analysis was performed by loading mouse body weight into the starting model: Y= Genotype+ Sex+ Genotype*Sex +Weight. Multiple correction testing was performed on the global p-value using the Hochberg correction.

For mouse skeletal staining, mice were killed by lethal IP injection of anesthetic at P10 and confirmation of death was performed by cutting of the femoral artery. Skin and inner organs were removed and full skeletons were first fixed with 100% EtOH for 2 days and then placed in Alcian blue to stain proteoglycans for 24 hours. Then shortly washed with EtOH to remove rest of Alcian blue and cleared in 1%KOH for 24 hours. Skeletons were then placed in Alizarin red staining solution, to detect calcium deposits for at 24 hours and then again in 1%KOH. Skeletons were then embedded in 100% Glycerol, after ascending series of KOH-glycerol mixture (1%KOH:Gly v/v: 80:20, 60:40, 40:60, 20:80). Pictures were acquired with a Leica stereomicroscope and Leica Application System (LAS V4) and bones were measured with Image J. For statistical analysis the Wilcoxon test was performed with R Studio (version 0.98.501)

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgements

We thank the families for their participation and patience. We are grateful to the Exome Aggregation Consortium for making their data available. The DDD study presents independent research commissioned by the Health Innovation Challenge Fund (grant number HICF-1009-003), a parallel funding partnership between the Wellcome Trust and the Department of Health, and the Wellcome Trust Sanger Institute (grant number WT098051). The views expressed in this publication are those of the author(s) and not necessarily those of the Wellcome Trust or the Department of Health. The study has UK Research Ethics Committee approval (10/H0305/83, granted by the Cambridge South REC, and GEN/284/12 granted by the Republic of Ireland REC). The research team acknowledges the support of the National Institute for Health Research, through the Comprehensive Clinical Research Network. The authors wish to thank the Sanger Mouse Genetics Project for generating and providing mouse phenotyping information, Natasha Karp for statistical input on the mouse data and Vagheesh Narasimhan for making the bcftools roh algorithm available. D.R.F. is funded through an MRC Human Genetics Unit program grant to the University of Edinburgh. Work on the Mmp21 mutant mouse models was supported by NIH grant U01-HL098180 to C.W.L. V.P. was funded by a fellowship from the DFG German Research Foundation.

References

1. DDD study. Large-scale discovery of novel genetic causes of developmental disorders. Nature. 2014; 519:223–8. [PubMed: 25533962]
2. De Rubeis S, et al. Synaptic, transcriptional and chromatin genes disrupted in autism. Nature. 2014; 515:209–15. [PubMed: 25363760]
3. Wright CF, et al. Genetic diagnosis of developmental disorders in the DDD study: a scalable analysis of genome-wide research data. Lancet. 2014; 385:1305–14. [PubMed: 25529582]
4. Najmabadi H, et al. Deep sequencing reveals 50 novel genes for recessive cognitive disorders. Nature. 2011; 478:57–63. [PubMed: 21937992]
5. Kohler S, et al. The Human Phenotype Ontology project: linking molecular biology and disease through phenotype data. Nucleic Acids Res. 2014; 42:D966–74. [PubMed: 24217912]
6. Zemojtel T, et al. Effective diagnosis of genetic disease by computational phenotype analysis of the disease-associated genome. Sci Transl Med. 2014; 6:252ra123.
7. Alazami AM, et al. Loss of function mutation in LARP7, chaperone of 7SK ncRNA, causes a syndrome of facial dysmorphism, intellectual disability, and primordial dwarfism. Hum Mutat. 2012; 33:1429–34. [PubMed: 22865833]
8. Akawi NA, Al-Jasmi F, Al-Shamsi AM, Ali BR, Al-Gazali L. LINS, a modulator of the WNT signaling pathway, is involved in human cognition. Orphanet J Rare Dis. 2013; 8:87. [PubMed: 23773660]
9. Kvarnung M, et al. A novel intellectual disability syndrome caused by GPI anchor deficiency due to homozygous mutations in PIGT. J Med Genet. 2013; 50:521–8. [PubMed: 23636107]
10. Nakashima M, et al. Novel compound heterozygous PIGT mutations caused multiple congenital anomalies-hypotonia-seizures syndrome 3. Neurogenetics. 2014; 15:193–200. [PubMed: 24906948]
11. Shinwari JM, et al. Recessive mutations in COL25A1 are a cause of congenital cranial dysinnervation disorder. Am J Hum Genet. 2015; 96:147–52. [PubMed: 25500261]
12. Faletra F, et al. Autosomal recessive Stickler syndrome due to a loss of function mutation in the COL9A3 gene. Am J Med Genet A. 2014; 164A:42–7. [PubMed: 24273071]
13. de Vries BB, Pals G, Odink R, Hamel BC. Homozygosity for a FBN1 missense mutation: clinical and molecular evidence for recessive Marfan syndrome. Eur J Hum Genet. 2007; 15:930–5. [PubMed: 17568394]
14. Van Dijk FS, et al. Compound-heterozygous Marfan syndrome. Eur J Med Genet. 2009; 52:1–5. [PubMed: 19059503]
15. Davey MG, et al. The chicken talpid3 gene encodes a novel protein essential for Hedgehog signaling. Genes Dev. 2006; 20:1365–77. [PubMed: 16702409]
16. Szymanska K, Hartill VL, Johnson CA. Unraveling the genetics of Joubert and Meckel-Gruber syndromes. J Pediatr Genet. 2014; 3:65–78. [PubMed: 25729630]
17. Bangs F, et al. Generation of mice with functional inactivation of talpid3, a gene first identified in chicken. Development. 2011; 138:3261–72. [PubMed: 21750036]
18. Scheffner M, Kumar S. Mammalian HECT ubiquitin-protein ligases: biological and pathophysiological aspects. Biochim Biophys Acta. 2014; 1843:61–74. [PubMed: 23545411]
19. Brown SD, Moore MW. The International Mouse Phenotyping Consortium: past and future perspectives on mouse phenotyping. Mamm Genome. 2012; 23:632–40. [PubMed: 22940749]
20. Ahokas K, et al. Matrix metalloproteinase-21, the human orthologue for XMMP, is expressed during fetal development and in cancer. Gene. 2002; 301:31–41. [PubMed: 12490321]
21. Feng Y, et al. Mammalian protein arginine methyltransferase 7 (PRMT7) specifically targets RXR sites in lysine- and arginine-rich regions. J Biol Chem. 2013; 288:37010–25. [PubMed: 24247247]
22. Leroy C, et al. The 2q37-deletion syndrome: an update of the clinical spectrum including overweight, brachydactyly and behavioural features in 14 new patients. Eur J Hum Genet. 2013; 21:602–12. [PubMed: 23073310]

23. Williams SR, et al. Haploinsufficiency of HDAC4 causes brachydactyly mental retardation syndrome, with brachydactyly type E, developmental delays, and behavioral problems. Am J Hum Genet. 2010; 87:219–28. [PubMed: 20691407]
24. Musante L, Ropers HH. Genetics of recessive cognitive disorders. Trends Genet. 2014; 30:32–9. [PubMed: 24176302]
25. Smedley D, et al. PhenoDigm: analyzing curated annotations to associate animal models with human diseases. Database (Oxford). 2013; 2013 bat025.
26. MacArthur DG, et al. Guidelines for investigating causality of sequence variants in human disease. Nature. 2014; 508:469–76. [PubMed: 24759409]
27. Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics. 2009; 25:1754–60. [PubMed: 19451168]
28. DePristo MA, et al. A framework for variation discovery and genotyping using next-generation DNA sequencing data. Nat Genet. 2011; 43:491–8. [PubMed: 21478889]
29. Li H, et al. The Sequence Alignment/Map format and SAMtools. Bioinformatics. 2009; 25:2078–9. [PubMed: 19505943]
30. Abecasis GR, et al. An integrated map of genetic variation from 1,092 human genomes. Nature. 2012; 491:56–65. [PubMed: 23128226]
31. McLaren W, et al. Deriving the consequences of genomic variants with the Ensembl API and SNP Effect Predictor. Bioinformatics. 2010; 26:2069–70. [PubMed: 20562413]
32. Lim LP, Burge CB. A computational analysis of sequence features involved in recognition of short introns. Proc Natl Acad Sci U S A. 2001; 98:11193–8. [PubMed: 11572975]
33. Harrow J, et al. GENCODE: the reference human genome annotation for The ENCODE Project. Genome Res. 2012; 22:1760–74. [PubMed: 22955987]
34. Pettersen EF, et al. UCSF Chimera--a visualization system for exploratory research and analysis. J Comput Chem. 2004; 25:1605–12. [PubMed: 15264254]
35. Li Y, et al. Global genetic analysis in mice unveils central role for cilia in congenital heart disease. Nature. (In press).
36. Liu X, et al. Interrogating congenital heart defects with noninvasive fetal echocardiography in a mouse forward genetic screen. Circ Cardiovasc Imaging. 2014; 7:31–42. [PubMed: 24319090]
37. Kim AJ, et al. Microcomputed tomography provides high accuracy congenital heart disease diagnosis in neonatal and fetal mice. Circ Cardiovasc Imaging. 2013; 6:551–9. [PubMed: 23759365]
38. Skarnes WC, et al. A conditional knockout resource for the genome-wide study of mouse gene function. Nature. 2011; 474:337–42. [PubMed: 21677750]
39. Ryder E, et al. Molecular characterization of mutant mouse strains generated from the EUCOMM/KOMP-CSD ES cell resource. Mamm Genome. 2013; 24:286–94. [PubMed: 23912999]
40. White JK, et al. Genome-wide generation and systematic phenotyping of knockout mice reveals new roles for many genes. Cell. 2013; 154:452–64. [PubMed: 23870131]
41. Karp NA, Melvin D, Mott RF. Robust and sensitive analysis of mouse knockout phenotypes. PLoS One. 2012; 7:e52410. [PubMed: 23300663]

Europe PMC Funders Author Manuscripts

Europe PMC Funders Author Manuscripts

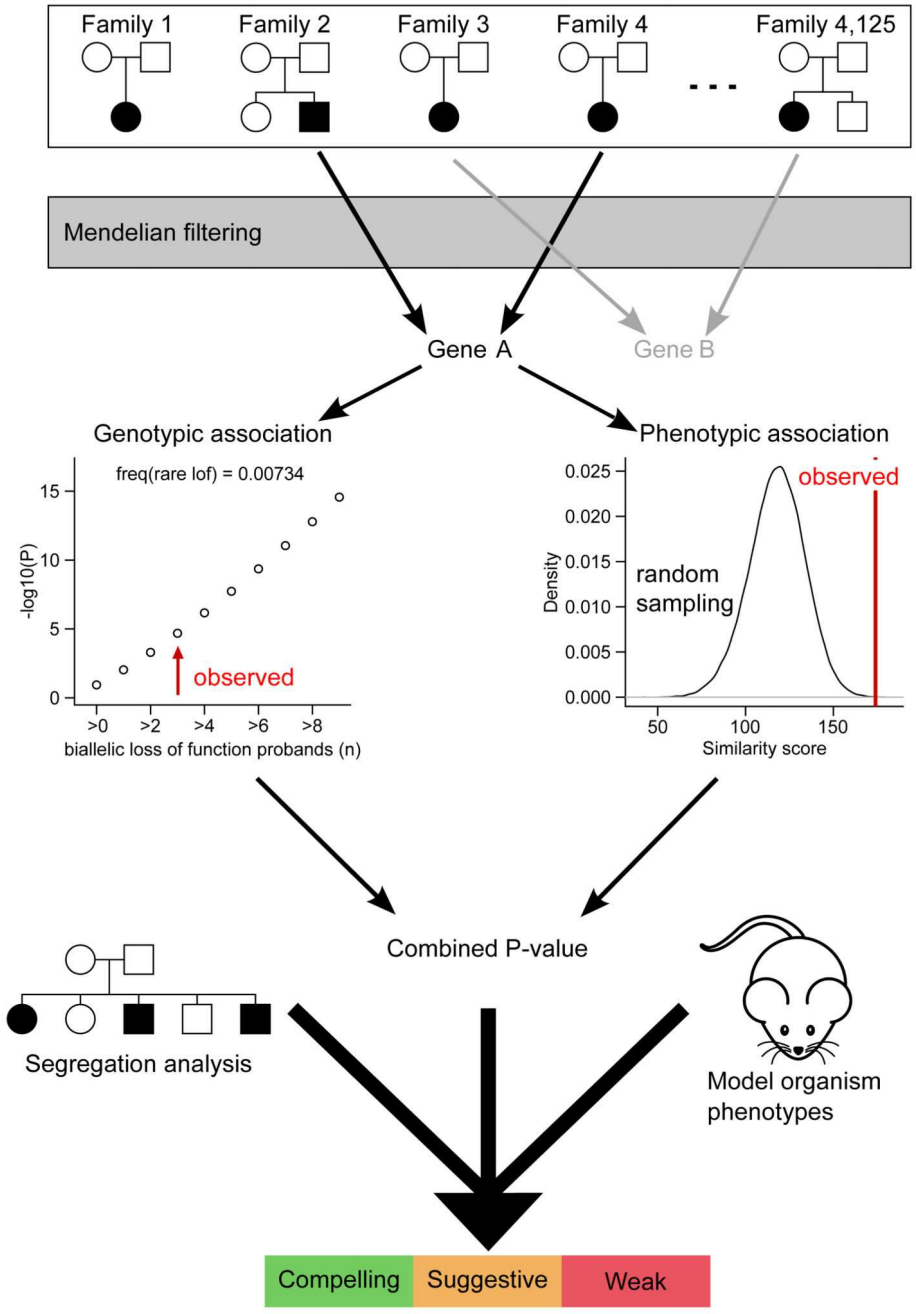

Figure 1. Overview of analytical strategy

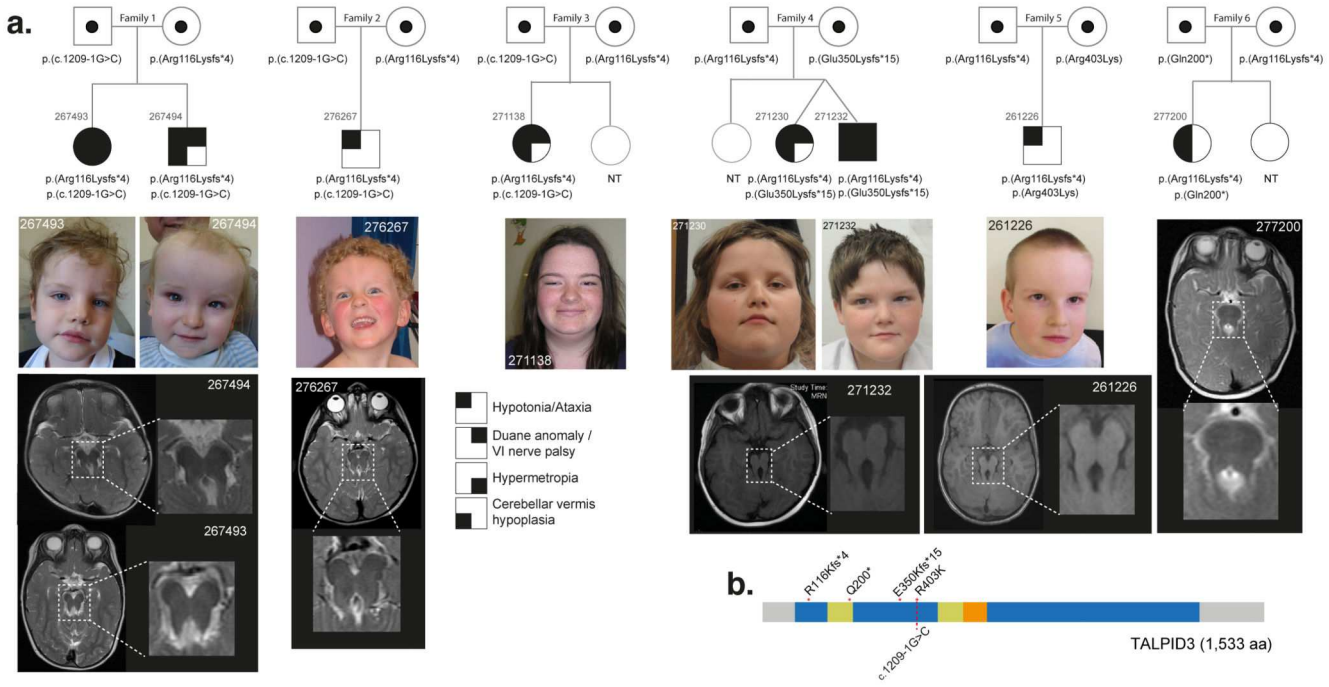

Figure 2. Clinical and neuroradiological features associated with biallelic variants in *KIAA0586*

A. Family structures, genotypes and phenotype key of the six families of the affected individuals with biallelic mutations in *KIAA0586*. The DECIPHER ID is given above and to the left of the pedigree symbol for each affected individual recruited to the DDD study. NT indicates an unaffected individual who has not been tested from the family mutations in *KIAA0586*. Where available an anterio-posterior (AP) facial photographs and a transverse section from the brain MR is given below the family tree. The white dashed box and white lines indicate the expanded region of the same image illustrating the “molar tooth” shape of the brainstem that is considered characteristic of Joubert syndrome in Families 1, 2, 4 and 5. In family 6 the brainstem shape is atypical for Joubert syndrome. The DECIPHER ID is indicated in all images. Informed consent was obtained to publish photographs. **B.** A cartoon depicting the protein domain structure of KIAA0586 (1533 aa) encompassing TALPID3 chain (100-1,343; in blue), which includes a highly conserved Region (467 – 554 aa; in orange) required for centrosomal localization and two coiled-coil (182-232 & 467-501 aa; in yellow) domains. A red asterix (coding region variant) or dashed line (essential splice site) is used to indicate the position of each pathogenic mutations identified in the families above.

Europe PMC Funders Author Manuscripts

Europe PMC Funders Author Manuscripts

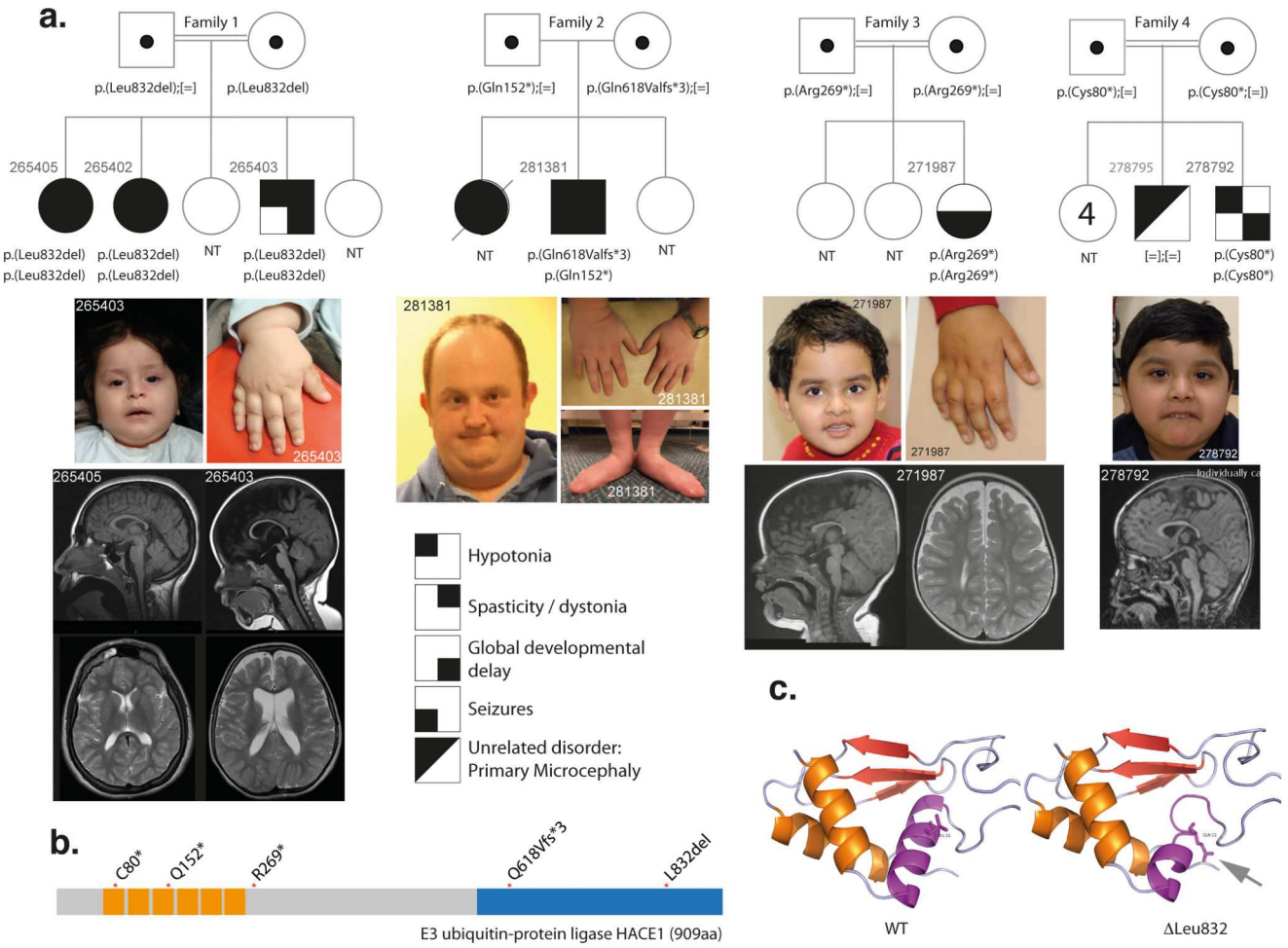

Figure 3. Clinical and neuroradiological features associated with biallelic variants in *HACE1*

A. Family structures and genotypes of four families. The DECIPHER ID is given for each recruited affected individual. NT indicates an individual who has not been tested from the family mutations in *HACE1*. Where available an anterio-posterior (AP) facial photographs and a transverse section from the brain MR is given below the family tree. The elder sibling in Family 4 is also recruited to DDD but on recruitment was considered to have a different disorder from his brother, primary microcephaly. Where available photographs of the face, hands and feet are presented. Saggital and transverse sections from the brain MR is given below the family tree. The saggital images show hypoplasia of the corpus callosum and the transverse images show an apparently reduced brain volume due to paucity of white matter. Informed consent was obtained to publish photographs.

B. Protein domain structure of HACE1 (909 aa) encompassing 6 Ankyrin repeats (in orange) and 1 HECT (E6AP-type E3 ubiquitin-protein ligase) domain (574 – 909 aa; in blue). A red asterix indicates the position of each pathogenic mutation.

C. Prediction of mutation consequence on tertiary protein structure based on PDB IDs: 4bbn and 3tug. The HECT domain is highly conserved and all models share the same fold. The deletion of Leu 832 has significant effect on the fold of the protein. This mutation disrupts

the helix by perturbing the hydrophobic core of the domain. This suggests that the mutated protein (ΔLeu832) is unlikely to have the same fold.

Europe PMC Funders Author Manuscripts

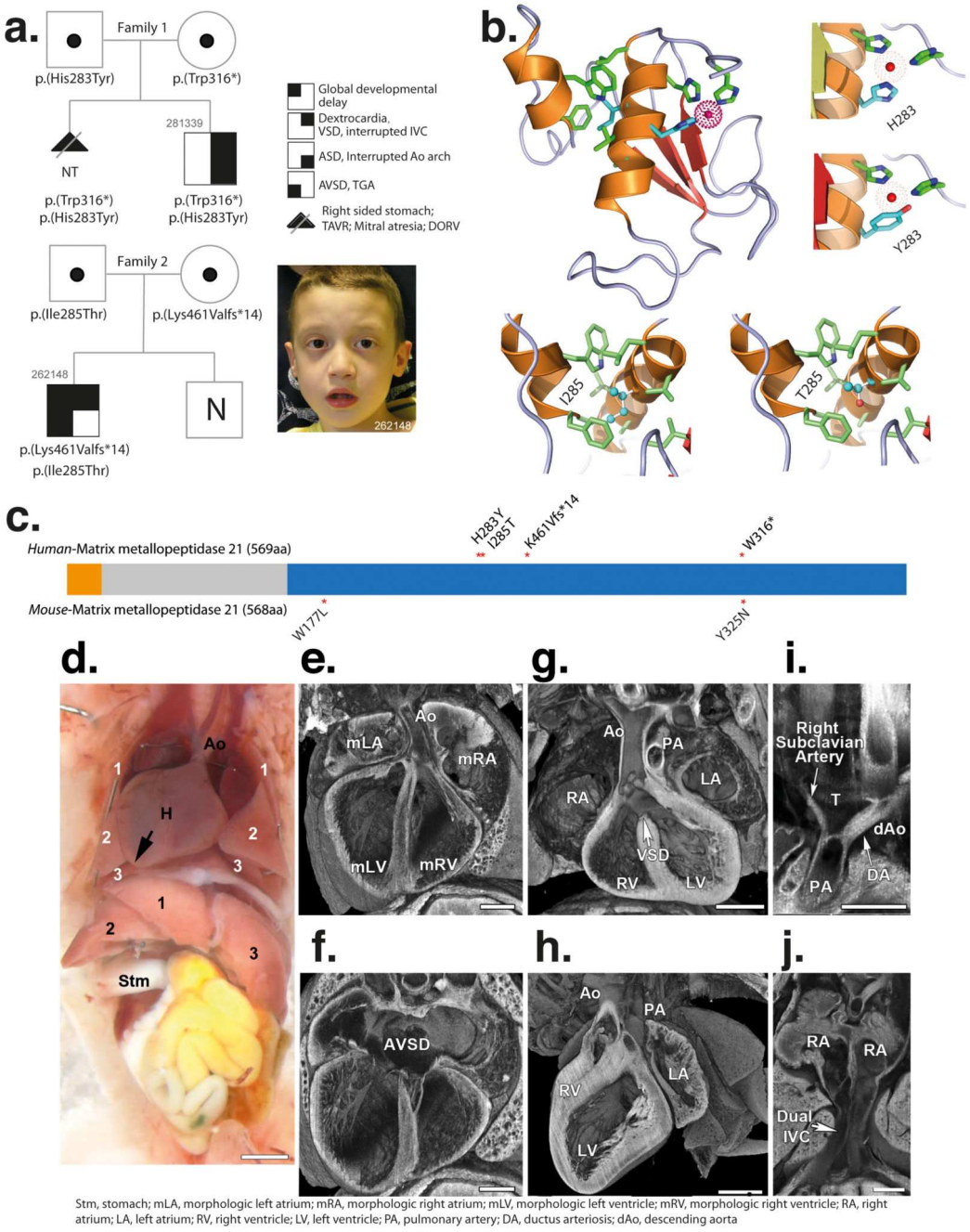

Figure 4. Features associated with biallelic variants in *MMP21/Mmp21* in humans and mice

A. Pedigrees and genotypes of two families, with DECIPHER IDs. Photograph showing mild craniofacial dysmorphisms including hypoplasia of the malar region and supraorbital ridges and prominent lips. Informed consent was obtained to publish the photograph.

B. Modeling of the effect of variants on the tertiary structure of MMP21, based on PDB IDs 1Y93 and 1FBL. Wildtype and mutated residues shown in cyan. The His283Tyr mutation severely reduces Zn2+ binding. The Ile285Thr mutation causes loss of hydrophobicity and weaker interactions with neighbouring non-polar amino acids (green) and should cause

movements of surrounding residues potentially leading to conformational shifts that affect Zn2+ binding.

C. Protein structure (569aa) encompassing signal peptide (1-24 aa; in orange), propeptide (25- 144 aa; in grey) and matrix metalloproteinase-21 chain (145 – 569 aa; in blue) that includes 4 Hemopexin repeats, with the positions of the human mutations shown above, and positions of ENU-induced mutations in mice shown below.

D. Necropsy picture of a *Miri* mutant with dextrocardia with anterior positioning of the aorta (Ao), indicating transposition of the great arteries (TGA), right lung isomerism, inverted liver lobation (1-3), and dextrogastria (Stm). Confocal episcopic microscopy showed: *Miri* mutant exhibiting dextrocardia with TGA (**E**) and atrioventricular septal defect (AVSD) (**F**), *Koli* mutant with double outlet right ventricle (**G**), *Miri* mutant with dextrocardia with TGA and hypoplastic right ventricle (**H**), mutant with anomalous right subclavian artery from pulmonary trunk (**I**), and mutant with duplicated inferior vena cava (IVC) draining into bilaterally symmetric right atria, indicating right atrial isomerism (**J**).

Europe PMC Funders Author Manuscripts

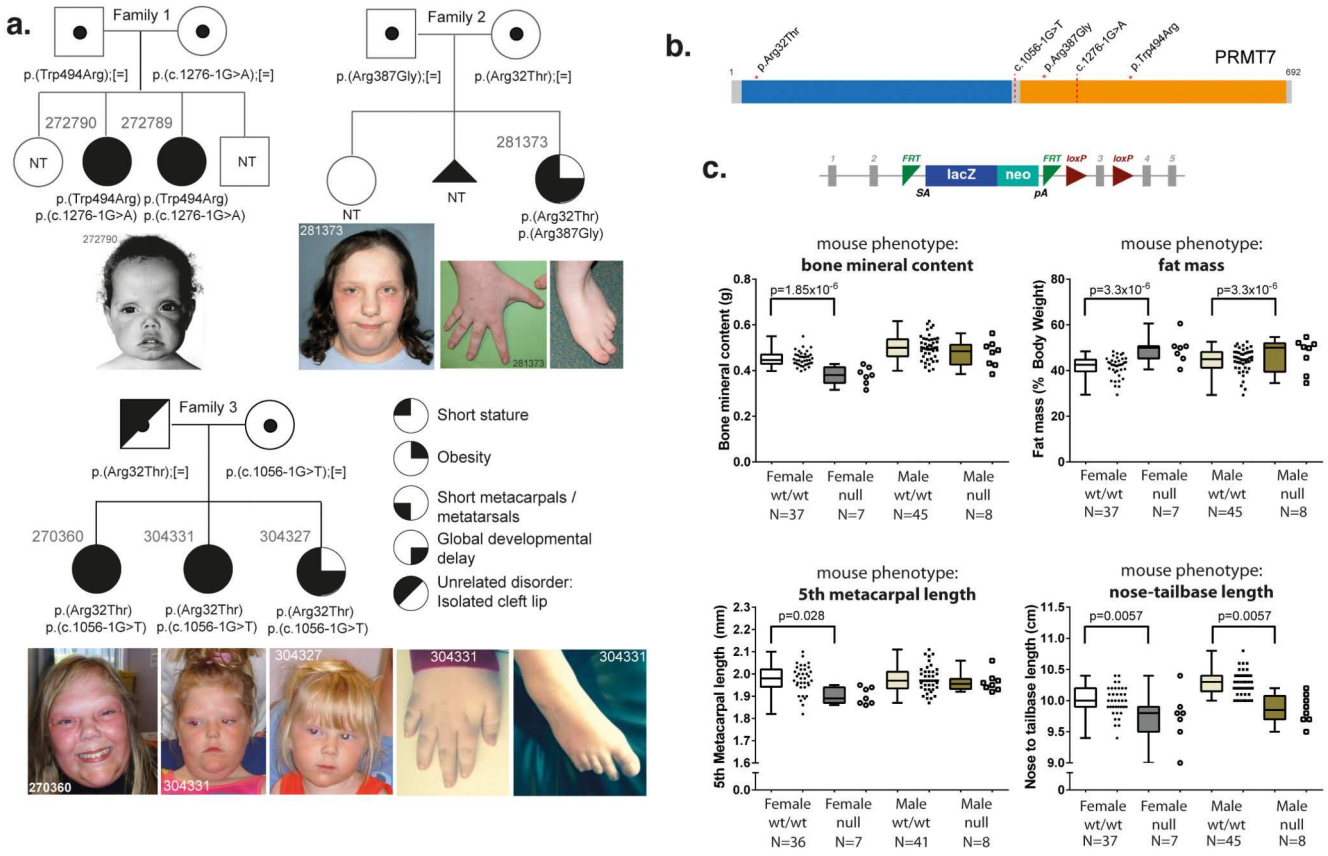

Figure 5. Features associated with biallelic variants in *PRMT7/Prmt7* in humans and mice

A. Pedigrees and genotypes of three families. The DECIPHER ID is given for each recruited affected individual. The father in Family 3 had a cleft lip which is thought to be coincidental. AP facial photographs of the affected individuals are given below the pedigrees. In Family 2 and 3 pictures of the foot show shortened posterior metatarsals. In Family 3 a photo of the hand of an affected individual shows brachydactyly with short metacarpals. Informed consent was obtained to publish photographs.

B. Protein domain structure of PRMT7 (692 aa) encompassing two active S-adenosylmethionine-dependent methyltransferases (SAM or AdoMet-MTase) PRMT-type domains (14 – 345 aa; in blue & 358 – 684 aa; in orange). Red asterixes indicate the positions of pathogenic mutations.

C. Data derived from mice homozygous for a targeted inactivation of *Prmt7* (null; *Prmt7*$^{tm1a/tm1a}$). DEXA scanning at 14 weeks of age indicates: female null mice have a reduced bone mineral content (top left graph), both male and female null mice have elevated fat mass (as a percentage of total body mass; top right graph) and reduced body length as determined by distance from nose to the base of the tail (bottom right graph). X-Ray images from 14 week old mice show that the length of the 5th metacarpal was reduced in female null mice only. Box-and-whiskers plots show min-mean-max values. P-values presented are either global adjusted p-values for genotype, or (for sexual dimorphism) the p-value for the interaction between sex and genotype.

Europe PMC Funders Author Manuscripts

Europe PMC Funders Author Manuscripts

Table 1

Integrated genotype-enrichment and phenotype-matching: genes with combined p-value less than 1e-4

Gene	# LOF/ LOF[a]	#LOF/ Func[b]	P(geno)[c]	P(pheno)[d]	Combined[e]	Evidence	Additional evidence
KIAA0586	6	1	1.42E-06	5.60E-04	1.75E-08	Genome-wide significant	Co-segregation in two affected sibs Mouse and chick mutants with ciliary phenotypes
HACE1	3	0	2.39E-08	1.11E-01	5.50E-08	Genome-wide significant	Co-segregation in one affected sib Three affected sibs with homozygous inframe deletion Mouse mutant with early lethality phenotype
PRMT7	1	2	1.45E-04	1.49E-03	3.53E-06	Suggestive	Co-segregation in three affected sibs Concordant mouse mutant with AHO-like phenotype
CSTB	2	0	3.16E-05	1.76E-02	8.56E-06	Suggestive	Previously implicated gene
COL9A3	2	0	3.89E-05	4.08E-02	2.28E-05	Suggestive	Previously implicated gene
MMP21	0	2	2.05E-03	1.03E-03	2.97E-05	Suggestive	Co-segregation with affected sib Two mouse mutants with heterotaxy

[a] Number of families with biallelic LOF variants

[b] Number of families compound heterozygous for LOF and functional variants

[c] Binomial p value of sampling N families, minimum of LOF/LOF and LOF/functional tests

[d] Phenotype matching p value assessed by permutation

[e] Genotype and Phenotype p values combined using Fisher' s methods