Mutations in epigenetic regulation genes are a major cause of overgrowth with intellectual disability

Katrina Tatton-Brown[1,2], Chey Loveday[1], Shawn Yost[1], Matthew Clarke[1], Emma Ramsay[1], Anna Zachariou[1], Anna Elliott[1], Harriet Wylie[1], Anna Ardissone[3], Olaf Rittinger[4], Fiona Stewart[5], Karen Temple[6,7], Trevor Cole[8], Childhood Overgrowth Collaboration[1], Shazia Mahamdallie[1], Sheila Seal[1], Elise Ruark[1] and Nazneen Rahman[1,9]*

[1]Division of Genetics and Epidemiology, Institute of Cancer Research, 15 Cotswold Road, London, SM2 5NG, UK

[2]South West Thames Regional Genetics Service, St George's University Hospitals NHS Foundation Trust, London, SW17 0QT, UK

[3]Child Neurology Unit, Foundation IRCCS C Besta Neurological Institute, Milan, Italy

[4]Landeskrankenanstalten Salzburg, Kinderklinik Department of Pediatrics, Klinische Genetik, Salzburg, Austria

[5]Northern Ireland Regional Genetics Service, Belfast City Hospital, Belfast, BT12 6BE, Northern Ireland

[6]Human Development and Health Academic Unit, Faculty of Medicine, University of Southampton, SO17 1BJ and [7]Wessex Clinical Genetics Service, University Hospital Southampton NHS Trust, Southampton, SO16 6YD, UK

[8]West Midlands Regional Genetics Service, Birmingham Women's Hospital NHS Foundation Trust and University of Birmingham, Birmingham Health Partners, B4 6NH, Birmingham

[9]Cancer Genetics Unit, Royal Marsden NHS Foundation Trust, London, SW3 6JJ, UK

*Correspondence: email: rahmanlab@icr.ac.uk twitter: @rahman_nazneen

Abstract

To explore the genetic architecture of human overgrowth syndromes and human growth control we performed experimental and bioinformatic analyses of 710 individuals with overgrowth (height and/or head circumference ≥+2SD) and intellectual disability (OGID). We identified a causal mutation in one of 14 genes in 50% (353/710). This includes *HIST1H1E,* encoding histone H1.4, which has not been associated with a developmental disorder previously. The pathogenic *HIST1H1E* mutations are predicted to result in a product that is less effective in neutralising negatively-charged linker DNA because it has a reduced net charge, and in DNA binding and protein-protein interactions because key residues are truncated. Functional network analyses demonstrated that epigenetic regulation is a prominent biological process dysregulated in individuals with OGID. Mutations in six epigenetic regulation genes, *NSD1*, *EZH2*, *DNMT3A*, *CHD8*, *HIST1H1E* and *EED*, accounted for 44% of individuals (311/710). There was significant overlap between the 14 genes involved in OGID and 611 genes in regions identified in GWAS to be associated with height ($P=6.84x10^{-8}$), suggesting common variation impacting function of genes involved in OGID influences height at a population level. Increased cellular growth is a hallmark of cancer and there was striking overlap between the genes involved in OGID and 260 somatically mutated cancer driver genes ($P=1.75x10^{-14}$). However, the mutation spectra of genes involved in OGID and cancer differ, suggesting complex genotype-phenotype relationships. These data reveal insights into the genetic control of human growth and demonstrate that exome sequencing in OGID has a high diagnostic yield.

Introduction

Human growth control, at the organismal and cellular level, is a complex process essential for health and dysregulated in many developmental disorders and cancers. The mechanistic control of cell size and proliferation has been studied, by diverse approaches, in many different species.[1,2] However, the control of overall size of an organism has been relatively understudied and is still poorly understood. The study of human growth disorders therefore not only improves diagnosis and management of human disease, it also offers an opportunity to enhance knowledge about the fundamental processes governing control of human size.

Human overgrowth syndromes are a nebulous group of conditions defined as having height and/or head circumference ≥2SD above the mean, together with additional phenotypic abnormalities, the most common of which is intellectual disability.[3] Overgrowth syndromes usually occur sporadically within a family and can be caused by several different mechanisms, including gene mutations, imprinting disruption and chromosome dosage abnormalities.[3,4]

Single gene disorders associated with overgrowth and intellectual disability (OGID) are well recognised; Sotos syndrome [MIM:117550] and Weaver syndrome [MIM:277590] are prototypic examples, due to *NSD1* [MIM:606681] and *EZH2* [MIM:601573] mutations respectively.[5,6] OGID syndromes have been increasingly identified over the last decade.[3,4] The advent of next-generation sequencing has been the foremost reason for this progress, and has allowed elucidation of the genetic causes of clinically established syndromes and the delineation of new syndromes.[5-13]

Despite these advances, many individuals with OGID remain without a genetic diagnosis. In addition, the relative contribution of the different genes to OGID is unknown. To better

characterise the genetic landscape of OGID we have here studied 710 affected individuals including 323 parent-proband trios (Table S1).

Methods

Subjects

We recruited participants through the Childhood Overgrowth (COG) Study, which began recruitment in 2005, approved by the London Multicentre Ethics Committee (05/MRE02/17). Informed consent was obtained from all participants and / or parents, as appropriate. Individuals were eligible for this study if they had height and / or head circumference at least two standard deviations above the mean (≥+2SD, UK90 growth data)[14] at some point in childhood, together with intellectual disability. We have termed this condition OGID (overgrowth + intellectual disability). Overgrowth phenotypes that are not associated with intellectual disability, such as Beckwith Wiedemann syndome [MIM 130650] or Marfan syndrome [MIM 154700] were not included. Regional or asymmetric overgrowth phenotypes (e.g. hemihypertrophy) in the absence of increased height or head circumference were not included.

710 individuals with OGID were included. 97% (693) were recruited to the study from Clinical Genetics departments. For 323 individuals, samples from both parents were also available and included. 205 probands had both height and head circumference ≥+2SD, termed 'head+height' in Table S1. 138 had height ≥+2SD with OFC <2SD, termed 'height only' and 109 had OFC ≥+2SD and height <2SD, termed 'head only'. For the remaining 258 individuals, the child was recruited to the study because they had overgrowth, but measurements for both height and head were not provided. The overgrowth category is termed 'unspecified' for these cases in Table S1. Intellectual disability was classified by the referring clinician as severe (77 cases), moderate (228 cases) or mild (229 cases). The referrer did not state the severity of the OGID for 176 individuals (termed 'unspecified' in Table S1).

Control data

We used the Exome Aggregation Consortium (ExAC) data version 3 accessed on 13/11/2015 (excluding the TCGA samples)[15] and the ICR1000 UK exome series[16] as reference data. We generated and analysed the ICR1000 UK exome series data using the same sequencing and analysis pipeline described for the OGID samples.

Targeted gene analyses

We previously reported mutations in *NSD1*, *EZH2*, *DNMT3A* [MIM: 602769] and *PPP2R5D* [MIM: 601646] in 198 cases. The relevant references are in Table S1. Intragenic mutations in these genes were detected with Sanger sequencing. *NSD1* is unusual amongst the 14 OGID genes included in this study in being prone to deletion by a 2Mb 5q35 microdeletion, mediated by flanking low-copy repeats.[17] We used MLPA to identify 5q35 microdeletions encompassing *NSD1.*[18] *NSD1* MLPA is also capable of detecting exon CNVs which account for ~5% of *NSD1* mutations.[18] Microdeletions and exon CNVs in the other genes were not sought, but are unlikely to be a major contributor because the surrounding sequence architecture and/or mechanism of pathogenicity make it much less likely that such events will cause OGID.

Exome sequencing

We performed exome sequencing in all probands in whom no mutation had been identified by targeted gene analyses and in parental samples where available. We performed exome sequencing using the Nextera Rapid Capture Exome Kit (Illumina). We prepared libraries from 50ng genomic DNA using the Nextera DNA Sample Preparation Kit (Illumina). On average 33M reads mapped to the pulldown and 86% of targeted bases had ≥15X coverage. The captured libraries were PCR amplified using the supplied paired-end PCR primers. Exome sequencing in 57 samples was performed before the Nextera Exome Kit was available using the TruSeq Exome Enrichment Kit, which includes the 14 genes involved in OGID. When converting our exome pipeline from TruSeq to Nextera we undertook in-house evaluation and validation to

ensure that the performance was equivalent. Sequencing was performed on an llumina HiSeq 2000 or HiSeq 2500 (high output mode) using v3 chemistry and generating 2 x 101 bp reads.

Variant calling

We used the OpEx v1.0 pipeline to perform variant calling.[19] We converted raw data to FASTQs using CASAVA version 1.8.2 with default settings. The OpEx v1.0 pipeline uses Stampy[20] to map to the human reference genome, Picard to flag duplicates, Platypus[21] to call variants, and CAVA[22] to provide consistent annotation of variants with the HGVS-compliant CSN (Clinical Sequencing Notation) standard v1.0.[22] The transcript information for variant annotation for the 14 relevant genes are given in Table 1.

Variant prioritisation and validation

We excluded variants with MAF>0.5% in either the Exome Aggregation Consortium (ExAC) and / or the ICR1000 UK exome series. For the *de novo* analyses, we identified and validated any high quality (as defined by OpEx[19]) variant in the child, that was not present in either parent. We evaluated and validated all rare variants identified in the 14 genes.

We confirmed all small variants in Table S1 that were called in exomes via Sanger sequencing of M13 tagged PCR products generated from genomic DNA. We performed PCR using the Qiagen Multiplex PCR Kit according to the manufacturer's instructions. We sequenced PCR products using M13 sequencing primers, the BigDye Terminator Cycle Sequencing Kit and an ABI 3730 Genetic Analyser (Applied Biosystems). We analyzed sequences using Mutation Surveyor software v3.20 (SoftGenetics) and verified the outputs by manual inspection by two individuals, independently.

Pathogenic mutation determination

Apart from *HIST1H1E* [MIM:142220] we considered a variant in the other 13 genes pathogenic if it fulfilled one or more of the following: 1) It was a *de novo* mutation in a gene for which such *de novo* mutations were already proven to cause OGID. 2) The inheritance was unknown, because parental samples were unavailable, but it had been previously identified as a pathogenic *de novo* mutation in OGID. 3) It was a protein truncating variant (PTV – frameshifting indels, stop-gain or essential splice-site variants) in an gene in which truncating mutations have been proven to be pathogenic. 4) There was clear evidence from the literature that it was pathogenic. The evidence for *HIST1H1E* mutations being pathogenic is provided in the results.

***HIST1H1E* statistical analyses**

We used the methods described in the DDD study, 2014,[23] to calculate the probability of identifying four *de novo* frameshift mutations in *HIST1H1E* using the gene-specific mutation rates from Samocha *et al.,* 2014.[24] The frameshift mutation rate in *HIST1H1E* (4.18×10^{-7}) was multiplied by twice the number of cases in this study (710) in order to get the expected number of frameshift mutations. We calculated the probability of observing four or more *de novo* frameshift mutations in *HIST1H1E* given the expected number of frameshift mutations using the ppois function in R.

We modelled the significance of mutation clustering in *HIST1H1E* under a binomial distribution where the probability of observing a mutation in a 12bp region, which comprises 1.8% of the coding sequence, was 0.018.

Protein net charge calculation

We obtained wildtype *HIST1H1E* cDNA (frame 1) sequence from Ensembl (ENST00000304218.5). We generated the HIST1H1E cDNA sequences edited with OGID mutations (frame 2). We used the variant c.430delG to generate the other possible alternative reading frame in *HIST1H1E* (frame 3). We translated the cDNA sequences using the Translate Tool at ExPASy. We calculated the net charge of the carboxy-terminal domain, from p.Lys110 onwards, at neutral pH using the Peptide Property Calculator at the Innovagen website.

Functional network analyses

We performed functional enrichment analysis using g:Profiler (version r1665_e85_eg32).[25] We used the 14 genes in Table 1 as our query set. We looked for enrichment amongst Gene Ontology molecular function terms and KEGG pathway gene sets, requiring the size of the functional category to be between 5 and 500 genes and using the Benjamini-Hochberg false discovery rate as the significance threshold. The FDR q-values presented are the Benjamini-Hochberg critical values.

Phenotypic analyses

We tested for significant difference in the diagnostic yields between different phenotypic groups using the prop.test function in R. We calculated the significance of association between an individual having macrocephaly and their mutation status (either a mutation in a PI3K/AKT pathway gene or a mutation in an epigenetic regulation gene) using a Fisher's exact test, which we implemented with the fisher.test function in R. We calculated the significance of association between an individual having macrocephaly in the absence of increased height and their mutation status, and the significance of association between an individual having increased height in the absence of macrocephaly and their mutation status in the same way. We tested for significant difference in the proportion of individuals with mild intellectual disability for those with

a mutation in a PI3K/AKT pathway OGID gene and those with a mutation in an epigenetic regulation OGID gene using the prop.test function in R.

Height GWAS gene and cancer driver gene comparisons

We obtained the list of 611 genes located in regions associated with human height through GWAS studies from Supplementary Table 1 of Wood *et al.,* 2014.[26] We obtained a list of 260 somatically mutated cancer genes from Supplementary Table 2 of Lawrence *et al.,* 2014[27] and the somatic mutations from the tumor portal website.

We calculated the probability of seeing the observed overlap of the OGID gene set with the GWAS gene set under a hypergeometric probability distribution assuming a total hypothetical size of 20,000 protein-coding genes in the exome using the phyper function in R. We calculated the probability of seeing the observed overlap of OGID gene set with the cancer driver gene set in the same way.

Results

Contribution of gene mutations to OGID

Using exome or targeted gene analyses we identified a pathogenic mutation in one of 14 genes in 357 individuals with OGID, giving a diagnostic yield of 50% (Figure 1). By far the most common cause was a mutation in *NSD1* (240 cases, 34%), followed by *EZH2* (34, 4.8%), *DNMT3A* (18, 2.5%), *PTEN* [MIM:601728] (16, 2.3%), *NFIX* [MIM:164005] (14, 2.0%) *CHD8* [MIM:610528] (12, 1.7%) *BRWD3* [MIM:300553] (7, 1.0%) *HIST1H1E* (5, 0.7%) *PPP2R5D* (3, 0.4%) *EED* [MIM:605984], *GPC3* [MIM:300037]*, MTOR* [MIM:601231] (two cases each) and *AKT3* [MIM:611223], *PIK3CA* [MIM:171834] (one case each)(Table S1). Amongst the 323 parent-proband trios we identified a cause in 191 (59%) of which 179 were *de novo* mutations and 12 were inherited.

Our data allow confirmation that *EED* mutations cause OGID. Two case reports of individuals with a characteristic phenotype that includes overgrowth have been published.[11,28] We here present two additional cases with a *de novo EED* mutation. The individuals have the same facial phenotype to each other and to previously reported cases, with long, narrow palpebral fissures, telecanthus and retrognathia. Notably, EED is a direct binding partner of EZH2[29], which has an established role in causing OGID.[30] Some role in overgrowth was either known, or has been proposed, for the remainder of these, apart from *HIST1H1E*.[7,10,11,13,30-37]

HIST1H1E mutations cause OGID

We present here data showing that certain *HIST1H1E* mutations cause OGID. Through exome sequencing we identified five unrelated probands, COG0405, COG0412, COG0552, COG1739 and COG1832 with heterozygous *HIST1H1E* protein truncating variants (PTVs) (Figure 2, Table 1, Table S1). In four probands the PTV had arisen *de novo.* Parental samples were not

available for the fifth child, but she carried the same mutation as one of the children with a *de novo* mutation. The detection of four *de novo HIST1H1E* mutations in 710 individuals is highly unlikely to have occurred by chance, as determined from gene specific *de novo* mutation rates ($P=5.17x10^{-15}$). None of the mutations are present in the ExAC dataset, nor in 11,677 exomes analysed in-house with similar pipelines. These results strongly support *HIST1H1E* mutations as a cause of OGID.

HIST1H1E encodes histone H1.4. In humans, H1.4 is one of 11 H1 linker histones that mediate the formation of higher order chromatin structures and regulate the accessibility of regulatory proteins, chromatin remodelling factors and histone modifying enzymes to their target sites.[38,39] The five mutations we identified cluster significantly ($P=2.0x10^{-9}$) to a 12bp region in the carboxy-terminal domain (CTD) that is involved in chromatin binding and protein-protein interactions (Figure 2A).[38] PTVs in the intronless histones have been shown to evade nonsense mediated mRNA decay.[40] Thus the OGID causing-mutations are predicted to generate a truncated product.

The CTD of linker histones regulate higher-order chromatin structure through neutralisation of negatively charged linker DNA.[38] The pathogenic *HIST1H1E* mutations all result in the same shift in the reading frame and are predicted to generate similar truncated proteins, with a reduced net charge of 7-9 (compared to 44 for the wildtype protein) (Figure 2A). The mutant protein is thus likely to be less effective in neutralising negatively charged linker DNA. Moreover, the truncation of the c-terminus likely impedes DNA binding and protein-protein interactions. It is also noteworthy that the other possible alteration in reading frame would reduce neither the net charge, nor the length of the protein (Figure 2A). Taken together these data suggest specific *HIST1H1E* mutations, restricted in position and type, cause human overgrowth.

HIST1H1E clinical phenotype

Individuals with *HIST1H1E* mutations had similar facial appearance in childhood with full cheeks, high hairline and telecanthus (Figures 2B, 2C, and 2D). Height, head circumference and degree of intellectual disability were variable, as were the additional clinical features. It is currently unclear whether these additional features are *HIST1H1E* associations or coincidental findings. Individual case descriptions are below.

COG0405, a female individual, was born at term with a weight of 3.58kg (+0.1SD) and a length of 53cm (+1.5SD). She was floppy in the neonatal period. A brain MRI scan at 4 months demonstrated mild ventricular dilatation but no other abnormalities. Her bone age at chronological age of 7 months was advanced at 18-24 months. By 19 months her length was 87cm (+2.0SD) with a weight of 13.4kg (+1.8SD) and she had developed a strabismus. At 13 years of age the individual was reviewed noted to have normal growth with a height of 150.8cm (-0.6SD), a head circumference of 55.8cm (-0.5SD) and a weight of 48.85kg (+0.4SD). She has developed a severe kyphoscoliosis for which she required surgery and has a mild intellectual disability.

COG0412, a male individual, was born at one week post term following an uncomplicated pregnancy and delivery. He weighed 4.75kg (+2.4SD). In the neonatal period he was noted to be floppy; he had poor feeding and undescended testes. At 1.5 years he was very tall at 105cm (+8.3SD) with a weight of 18.8kg (+4.6SD) and a head circumference of 52.5cm (+2.6SD). He was reported to have multiple nevi and redundant skin on the palms of his hands. He had a moderate intellectual disability and no behavioural issues at that time. When he was reviewed at 15.5 years, he was no longer tall with a height of 166.5cm (-0.6SD). His head circumference was 58.7cm (+1.4SD). By this age he had developed an anxiety disorder that was refractory to

medical treatment. He had also developed phobias. In addition, he had major dental problems with crumbling teeth and he had dry, flaky nails.

COG0552, a female individual, was born at term with a weight of 4.79 kg (+2.5SD) and length of 57cm (+3.6SD). She was floppy in the neonatal period with poor feeding. She developed no new medical problems in childhood. At the age of 4.2 years she was reported to be delayed in her development. She had a height of 108cm (+1.2SD); head circumference of 55cm (+3.2SD) and weight of 24kg (+2.7SD).

COG1739, a female individual, was initially thought clinically to have Weaver syndrome. She was born at 37 weeks following an uncomplicated pregnancy and labour with a weight of 3.25kg (+0.8SD), length of 49cm (+0.7SD) and head circumference of 37cm (+3.3SD). She was hypoglycemic and hypertonic in the neonatal period, and was also noted to have camptodactyly. At 1.9 years she was diagnosed with a moderate intellectual disability and had a height of 85cm (mean); head circumference of 51cm (+1.8SD) and weight of 12kg (-0.3SD).

COG1832, a male individual, was born at one week post term weighing 3.74kg (+0.4SD). The pregnancy had been complicated by exposure to chicken pox. At birth, COG1832 was noted to have talipes equinovarus and later in the neonatal period was diagnosed with delayed visual maturation. A brain MRI scan showed a slender corpus callosum and unusual ventricular outline, possibly indicative of a periventricular leukomalacia. At 8.5 years, height was 133.2cm (+0.5SD) with a weight of 33kg (+1.2SD). The head circumference at 6.3 years was 59cm (+3.7SD). He has limited speech but with verbal comprehension markedly ahead of this ability to express himself. He has left amblyopia and astigmatism. His hearing is normal. He suffers from constipation. At times his behaviour is challenging.

Functional network analyses

To investigate the biological processes abrogated by OGID pathogenic mutations we performed functional enrichment analysis using the GO molecular function terms and KEGG pathway gene sets in g:Profiler[25]. The chromatin binding (FDR q-value = $1.58x10^{-6}$) and PI3K/AKT signaling pathway (FDR q-value = $6.80x10^{-5}$) gene sets were significantly enriched.

Six genes, *NSD1*, *EZH2*, *DNMT3A*, *EED*, *CHD8* and *HIST1H1E* were in the chromatin binding gene set. All encode proteins involved in epigenetic regulation (Figure 3A). NSD1 is a histone methyltransferase that catalyses methylation of H3K36, and to lesser extent H4K20, and is primarily associated with transcriptional activation.[41] EZH2 and EED are key components of the polycomb repressive complex 2 (PRC2), which catalyses methylation of H3K27, resulting in transcriptional repression of target genes.[29] DNMT3A is a DNA methyltransferase crucial for the establishment of new methylation marks during early embryogenesis and the sex-dependent methylation of imprinted genes.[42,43] *CHD8* encodes an ATP-dependent chromatin remodeler that binds to methylated H3K4, a key histone modification at active promoters.[37] As noted above, H1.4 binds to linker DNA between nucleosomes and has key roles in chromatin compaction and regulation of gene expression.[39] Together, mutations in these six genes accounted for 311 (44%) of our series. Disruption of epigenetic regulation is therefore a prominent molecular mechanism underlying OGID (Figure 1).

Five of the genes, *PTEN*, *AKT3*, *PIK3CA* (which encodes p110α, the catalytic domain of the heterodimeric PI3K lipid kinase), *MTOR* and *PPP2R5D* (which encodes B56δ a regulatory subunit of the heterotrimeric PP2A protein phosphatase) are in the PI3K/AKT pathway, which plays a key role in the regulation of growth (Figure 3B). Activation of the PI3K/AKT pathway results in cellular growth promotion through increased cell metabolism, cell survival, cell turnover and protein synthesis.[44] Together mutations in these genes only made a minor

contribution to our OGID series (23 cases, 3.2%). In part this is because individuals with mutations in these genes are more often diagnosed with other conditions, such as Cowden syndrome [MIM:158350], megalencephaly-capillary malformation syndrome [MIM:602501], or regional overgrowth.[33,36]

The remaining three genes, *NFIX*, *GPC3* and *BRWD3* encode a transcription factor, a proteoglycan and a bromodomain containing protein, respectively[7,32,34] (23 cases, 3.2%). There is currently no clear functional link between these genes and the other genes we report here. However, it is possible *BRWD3* mutations also cause overgrowth through epigenetic regulation dysfunction, as there are data suggesting it is involved in histone H3.3 regulation.[45]

Phenotype analyses

There was enrichment of mutations in individuals with both increased height and head circumference, compared to individuals in whom only one growth parameter was increased, as would be expected. Specifically the diagnostic yield in individuals with both macrocephaly and increased height was 59% (120/205), significantly higher than the diagnostic yields in individuals with only macrocephaly (43%, 47/109, p=0.006) or only increased height (45%, 62/138, p=0.009). There was no significant difference between the diagnostic yields in individuals with only macrocephaly and in those with only increased height (p=0.146). There was also no significant difference between the diagnostic yield in individuals with unspecified growth parameters (50%, 130/258) and any other group.

To further explore the phenotypic spectrum of OGID we compared the growth and intellectual disability severity of the individuals due to mutations in the epigenetic regulation genes and those involved in the PI3K/AKT pathway, using cases for which the relevant phenotypic information was available (217 individuals with complete growth data and 263 individuals with

intellectual disability severity information) (Figure 4). Macrocephaly (i.e. head circumference ≥2SD above the mean) occurred more frequently in individuals with PI3K/AKT pathway gene mutations; all 17 had macrocephaly, compared with 140/200 individuals with OGID due to epigenetic regulation gene mutations ($P=4.1x10^{-3}$; Figure 4A). Furthermore 9/17 of the PIK/AKT pathway cases had macrocephaly without increased height compared with 32/200 of the epigenetic regulation pathway cases ($P=1.0x10^{-3}$; Figure 4A). The remaining 60/200 had increased height without macrocephaly, a combination not present in OGID due to PI3K/AKT pathway gene mutations ($P=4.1x10^{-3}$; Figure 4A). Varying severity of intellectual disability was a feature of both groups, but mild intellectual disability was more common in OGID due to PI3K/AKT pathway gene mutations (14/20) than OGID due to epigenetic regulation gene mutations (101/243; P=0.01) (Figure 4B).

The risk of childhood cancer is one the most controversial areas of OGID management. 8/710 OGID individuals in this study developed cancer in childhood (Table S1). This includes 4/357 with an identified genetic cause, three of whom had an *EZH2* mutation. COG1724 developed neuroblastoma at 46 months, COG0285 developed T-cell non-hodgkins lymphoma at 13 years and COG1521 was diagnosed with both neuroblastoma and acute lymphoblastic leukemia at 13 months. The childhood cancer incidence for *EZH2* mutation carriers in this study was thus 9% (3/34). The remaining child had an *NSD1* microdeletion and T-cell non-hodgkins lymphoma. This information will be useful in family discussions about childhood cancer risk, particularly in relation to surveillance strategies, which are generally of unproven benefit and can be associated with appreciable false positive rates.[46]

Height GWAS loci comparative analyses

We next explored the overlap between the 14 genes and 611 genes implicated through genome-wide association studies (GWAS) to be involved in the control of human height.[26] There

was significant overlap; six genes involved in OGID were also present in height GWAS regions ($P=6.8x10^{-8}$) (Figure S1). The overlap is primarily through the epigenetic regulation genes, all of which, except *EED*, were represented in height GWAS regions. Two separate intronic SNPs in each of *NSD1* and *DNMT3A* were independently associated with height in the GWAS study and there were no other genes within the linkage disequilibrium (LD) blocks of association. This strongly suggests NSD1 and DNMT3A functional impact underlie the height association in these regions (Figure S1). Single SNPs in intron 5 of *CHD8,* intron 9 of *MTOR,* 1kb downstream of *HIST1H1E* and 48kb upstream of *EZH2* were also associated with height.[26] For *HIST1H1E* and *EZH2* there were no other genes in the LD block of association. For *MTOR* the variant associated with a cis-eQTL affecting *MTOR* expression, though the association was better accounted for by an upstream variant (rs2295080) in the *MTOR* promoter region that was in LD with the height SNP (LD r2=0.85).[26] Although the causal SNPs and mechanisms of association are not fully elucidated, these data suggest that common variation in some genes involved in OGID also influence height at a population level.

Cancer somatic driver mutation comparative analyses

Dysregulated cellular growth is a hallmark of cancer, and certain human conditions are associated with both overgrowth and increased cancer risk.[33,47] We therefore next sought to investigate the overlap between the 14 genes and 260 somatically mutated cancer driver genes reported by Lawrence et al.[27] There was significant overlap; 8/14 genes involved in OGID were somatically mutated in a diverse range of cancers (*NSD1*, *EZH2*, *DNMT3A*, *PTEN*, *CHD8*, *HIST1H1E*, *MTOR*, *PIK3CA*; $P=1.7x10^{-14}$). For the PI3K/AKT pathway genes, the mutation spectra are similar in OGID and cancer.[36] By contrast, for the epigenetic regulation genes the mutation spectra in OGID and cancer have substantial, distinctive differences.

Somatic mutations in *HIST1H1E*, *EZH2* and *DNMT3A* occur in haematological malignancies.[27,48-52] *HIST1H1E* and *EZH2* mutations are each present in ~20% of B-cell lymphomas.[50,51] Somatic *HIST1H1E* mutations are nonsynonymous mutations throughout the gene and do not include the clustered PTVs that cause OGID (Figure 5). *EZH2* mutations in B-cell lymphomas are often activating nonsynonymous mutations in the SET domain, the majority of which target a single amino acid p.Tyr646.[50] Nonsynonymous mutations at this residue have not been detected in OGID, and are not present in ExAC, perhaps suggesting germline *EZH2* mutations altering p.Tyr646 are not compatible with life (Figure 5). Inactivating *EZH2* mutations are present in myeloid malignancies and in T-ALL.[48-50] A proportion of these latter mutations overlap with *EZH2* mutations in OGID.

DNMT3A is one of the most frequently mutated genes in AML and mutations also occur less frequently in other haematological malignancies.[27,52] The majority target a single residue, p.Arg882, with the remainder being nonsynonymous variants and PTVs scattered through the gene. Mutations at p.Arg882 have not thus far been reported in OGID (Figure 5). Protein modelling suggests the somatic mutations primarily impact DNA binding, whereas the mutations in OGID are more likely to impact histone binding.[13]

Somatic *NSD1* mutations are seen in ~10% of head and neck squamous cell carcinomas[27,53] and somatic *CHD8* mutations are present in ~3% of glioblastoma multiforme (GBM).[27] For these cancers the mutation pattern is similar to that observed in OGID, with PTVs being the most frequent mutation type (Figure 5).[31] Interestingly, Lawrence et al. found *NSD1* and *CHD8* to each be significant in their pan-cancer analysis, present in 2% of cancers.[27] However, the pan-cancer mutation spectra for each gene was different to that observed in OGID, with most being nonsynonymous mutations scattered throughout the gene (Figure 5).

Discussion

We present here the largest genetic study of overgrowth and intellectual disability performed to date, including 710 affected individuals and 636 parents. We show that OGID is a highly heterogeneous condition, involving at least 14 genes. Perturbation of epigenetic regulation is a prominent mechanism causing OGID and can be caused by mutations in at least six different genes. *NSD1* mutation is by far the most frequent cause of OGID, accounting for 240 (34%) of our series. Notably, *NSD1* is within a 2Mb region flanked by low-copy repeats that mediate a microdeletion, that is one of the commonest causes of Sotos syndrome[17] and was present in 29 individuals. Furthermore exon deletions or duplications (exon CNVs) are reported in ~5% of cases[18] and were present in 9 individuals. We analysed *NSD1* for these types of mutations, using MLPA, as they are not robustly identifiable in our exome data. We did not examine the other genes for microdeletions or exon CNVs. However, they are not known to be a major contributor to pathogenic mutations in the other genes. Even after excluding microdeletions and exon CNVs *NSD1* is still the most common cause of OGID, accounting for 202 (28%) of our series.

The comparative analyses of genes involved in OGID with GWAS height loci and with cancer driver genes highlight intriguing similarities and differences. Our data strongly suggest that common variation impacting epigenetic regulation of gene function influence height at a population level. Further investigation of these GWAS loci would be of considerable interest, particularly in relation to advancing knowledge on how, and why, epigenetic regulation dysfunction impacts human growth.

Several genes involved in OGID are somatically mutated in a diverse range of cancers, but the spectra of mutations, particularly in the epigenetic regulation genes, is different in OGID and cancer. The underlying reasons for these differences will be complex, and may include

embryonic lethality of certain oncogenic mutations when they occur in the germline. Integration of germline and somatic mutational data in future research will be useful, and will likely advance functional and mechanistic understanding of the genes.

One of the most striking results of this study is the high diagnostic yield of genetic testing in OGID; a genetic cause was identified in 50% (357/710) of cases. This is likely to be an underestimate as we have been conservative in attributing pathogenicity to OGID gene variants and additional OGID genes almost certainly exist. Indeed amongst the 132 trios in whom a definitive cause was not found, a *de novo* mutation possibly associated with their phenotype was present in 28; for example two had *de novo* nonsynonymous variants in *XRN1*.

The diagnostic yield in our OGID series is higher than exome sequencing studies in other phenotypes that include intellectual disability, which ranged from 13-35%.[23,54-58] The studies are not directly comparable as most other exome studies included cases in which prior genetic testing was negative. Our study recruitment started prior to the discovery and clinical testing of most of the genes we report here, which allows us to provide a much better estimate of the overall contribution of rare gene mutations to this phenotype.

Given the high success rate, strong consideration should be given to using exome sequencing as a first-line diagnostic test in OGID. Height and head circumference can be easily measured and intellectual disability is readily diagnosable. Therefore, implementation of exome sequencing in OGID should be straightforward. Gene testing would provide important diagnostic and recurrence risk information to many families. Furthermore, it would increase genotype-phenotype data, which are urgently required to improve prognostic information. Of equal importance, exome sequencing in OGID would lead to the identification of new genes and new mutations in known genes. In turn, this will stimulate and facilitate scientific research, enhancing

knowledge of basic biological processes controlling growth, and the diverse pathologies in which human growth control is dysfunctional.

Supplemental Data

Supplemental data include Table S1, Figure S1 and Supplemental Note.

Acknowledgements

We thank the families for their participation and the clinicians that recruited them. The full list of collaborators is in the Supplemental Data. We are grateful to Margaret Warren-Perry for assistance in recruitment. We are grateful to Sandra Hanks, Silvana Powell, Imran Uddin and Ann Strydom for technical and administrative support, and Tara Mills for assistance with the GWAS analyses. We acknowledge support from the NIHR RM/ICR Biomedical Research Centre and Wessex NIHR clinical research network. Katrina Tatton-Brown is supported by funding from the Child Growth Foundation. This work was supported by Wellcome Trust Award 100210/Z/12/Z.

Web Resources

OpEx is available from http://www.well.ox.ac.uk/opex

Picard is available from http://picard.sourceforge.net

Exome Aggregation Consortium (ExAC) is at http://exac.broadinstitute.org

ICR1000 UK exome series is at www.icr.ac.uk/icr1000exomes

ExPASy cDNA Translate Tool is at http://web.expasy.org/translate/

Tumor portal web interface is at http://www.tumorportal.org/

g:profiler web interface is at http://biit.cs.ut.ee/gprofiler/

Protein charge calculator is at http://pepcalc.com/protein-calculator.php

OMIM web interface is at http://www.omim.org/

References

1. Stocker, H., and Hafen, E. (2000). Genetic control of cell size. Curr Opin Genet Dev 10, 529-535.

2. Saucedo, L.J., and Edgar, B.A. (2002). Why size matters: altering cell size. Curr Opin Genet Dev 12, 565-571.

3. Tatton-Brown, K., and Weksberg, R. (2013). Molecular mechanisms of childhood overgrowth. American journal of medical genetics Part C, Seminars in medical genetics 163c, 71-75.

4. Edmondson, A.C., and Kalish, J.M. (2015). Overgrowth Syndromes. Journal of pediatric genetics 4, 136-143.

5. Tatton-Brown, K., Cole, T.R.P., and Rahman, N. (1993). Sotos Syndrome. In GeneReviews(R), R.A. Pagon, M.P. Adam, H.H. Ardinger, S.E. Wallace, A. Amemiya, L.J.H. Bean, T.D. Bird, N. Ledbetter, H.C. Mefford, R.J.H. Smith, et al., eds. (Seattle (WA), University of Washington, Seattle

University of Washington, Seattle. GeneReviews is a registered trademark of the University of Washington, Seattle. All rights reserved.

6. Tatton-Brown, K., Murray, A., Hanks, S., Douglas, J., Armstrong, R., Banka, S., Bird, L.M., Clericuzio, C.L., Cormier-Daire, V., Cushing, T., et al. (2013). Weaver syndrome and EZH2 mutations: Clarifying the clinical phenotype. Am J Med Genet A 161a, 2972-2980.

7. Malan, V., Rajan, D., Thomas, S., Shaw, A.C., Louis Dit Picard, H., Layet, V., Till, M., van Haeringen, A., Mortier, G., Nampoothiri, S., et al. (2010). Distinct effects of allelic NFIX mutations on nonsense-mediated mRNA decay engender either a Sotos-like or a Marshall-Smith syndrome. Am J Hum Genet 87, 189-198.

8. Gibson, W.T., Hood, R.L., Zhan, S.H., Bulman, D.E., Fejes, A.P., Moore, R., Mungall, A.J., Eydoux, P., Babul-Hirji, R., An, J., et al. (2011). Mutations in EZH2 cause Weaver syndrome. Am J Hum Genet 90, 110-118.

9. Cordeddu, V., Redeker, B., Stellacci, E., Jongejan, A., Fragale, A., Bradley, T.E., Anselmi, M., Ciolfi, A., Cecchetti, S., Muto, V., et al. (2014). Mutations in ZBTB20 cause Primrose syndrome. Nature genetics 46, 815-817.

10. Loveday, C., Tatton-Brown, K., Clarke, M., Westwood, I., Renwick, A., Ramsay, E., Nemeth, A., Campbell, J., Joss, S., Gardner, M., et al. (2015). Mutations in the PP2A regulatory subunit B family genes PPP2R5B, PPP2R5C and PPP2R5D cause human overgrowth. Hum Mol Genet 24, 4775-4779.

11. Cohen, A.S., and Gibson, W.T. (2016). EED-associated overgrowth in a second male patient. Journal of human genetics 61, 831-834.

12. Baynam, G., Overkov, A., Davis, M., Mina, K., Schofield, L., Allcock, R., Laing, N., Cook, M., Dawkins, H., and Goldblatt, J. (2015). A germline MTOR mutation in Aboriginal Australian siblings with intellectual disability, dysmorphism, macrocephaly, and small thoraces. Am J Med Genet A 167, 1659-1667.

13. Tatton-Brown, K., Seal, S., Ruark, E., Harmer, J., Ramsay, E., Del Vecchio Duarte, S., Zachariou, A., Hanks, S., O'Brien, E., Aksglaede, L., et al. (2014). Mutations in the DNA methyltransferase gene DNMT3A cause an overgrowth syndrome with intellectual disability. Nature genetics 46, 385-388.

14. Freeman, J.V., Cole, T.J., Chinn, S., Jones, P.R., White, E.M., and Preece, M.A. (1995). Cross sectional stature and weight reference curves for the UK, 1990. Arch Dis Child 73, 17-24.

15. Lek, M., Karczewski, K.J., Minikel, E.V., Samocha, K.E., Banks, E., Fennell, T., O'Donnell-Luria, A.H., Ware, J.S., Hill, A.J., Cummings, B.B., et al. (2016). Analysis of protein-coding genetic variation in 60,706 humans. Nature 536, 285-291.

16. Ruark, E., Munz, M., Renwick, A., Clarke, M., Ramsay, E., Hanks, S., Mahamdallie, S., Elliott, A., Seal, S., Strydom, A., et al. (2015). The ICR1000 UK exome series: a resource of gene variation in an outbred population. F1000Res 4, 883.

17. Kurotaki, N., Stankiewicz, P., Wakui, K., Niikawa, N., and Lupski, J.R. (2005). Sotos syndrome common deletion is mediated by directly oriented subunits within inverted Sos-REP low-copy repeats. Hum Mol Genet 14, 535-542.

18. Douglas, J., Tatton-Brown, K., Coleman, K., Guerrero, S., Berg, J., Cole, T.R., Fitzpatrick, D., Gillerot, Y., Hughes, H.E., Pilz, D., et al. (2005). Partial NSD1 deletions cause 5% of Sotos syndrome and are readily identifiable by multiplex ligation dependent probe amplification. J Med Genet 42, e56.

19. Ruark, E., Munz, M., Clarke, M., Renwick, A., Ramsay, E., Elliott, A., Seal, S., Lunter, G., and Rahman, N. (2016). OpEx - a validated, automated pipeline optimised for clinical exome sequence analysis. Sci Rep 6, 31029.

20. Lunter, G., and Goodson, M. (2011). Stampy: a statistical algorithm for sensitive and fast mapping of Illumina sequence reads. Genome Res 21, 936-939.

21. Rimmer, A., Phan, H., Mathieson, I., Iqbal, Z., Twigg, S.R., Wilkie, A.O., McVean, G., and Lunter, G. (2014). Integrating mapping-, assembly- and haplotype-based approaches for calling variants in clinical sequencing applications. Nature genetics 46, 912-918.

22. Munz, M., Ruark, E., Renwick, A., Ramsay, E., Clarke, M., Mahamdallie, S., Cloke, V., Seal, S., Strydom, A., Lunter, G., et al. (2015). CSN and CAVA: variant annotation tools for rapid, robust next-generation sequencing analysis in the clinical setting. Genome Med 7, 76.

23. Deciphering Developmental Disorders Study. (2015). Large-scale discovery of novel genetic causes of developmental disorders. Nature 519, 223-228.

24. Samocha, K.E., Robinson, E.B., Sanders, S.J., Stevens, C., Sabo, A., McGrath, L.M., Kosmicki, J.A., Rehnstrom, K., Mallick, S., Kirby, A., et al. (2014). A framework for the interpretation of de novo mutation in human disease. Nature genetics 46, 944-950.

25. Reimand, J., Arak, T., Adler, P., Kolberg, L., Reisberg, S., Peterson, H., and Vilo, J. (2016). g:Profiler-a web server for functional interpretation of gene lists (2016 update). 44, W83-89.

26. Wood, A.R., Esko, T., Yang, J., Vedantam, S., Pers, T.H., Gustafsson, S., Chu, A.Y., Estrada, K., Luan, J., Kutalik, Z., et al. (2014). Defining the role of common variation in the genomic and biological architecture of adult human height. Nature genetics 46, 1173-1186.

27. Lawrence, M.S., Stojanov, P., Mermel, C.H., Robinson, J.T., Garraway, L.A., Golub, T.R., Meyerson, M., Gabriel, S.B., Lander, E.S., and Getz, G. (2014). Discovery and saturation analysis of cancer genes across 21 tumour types. Nature 505, 495-501.

28. Cohen, A.S., Tuysuz, B., Shen, Y., Bhalla, S.K., Jones, S.J., and Gibson, W.T. (2015). A novel mutation in EED associated with overgrowth. Journal of human genetics 60, 339-342.

29. Cao, R., Wang, L., Wang, H., Xia, L., Erdjument-Bromage, H., Tempst, P., Jones, R.S., and Zhang, Y. (2002). Role of histone H3 lysine 27 methylation in Polycomb-group silencing. Science 298, 1039-1043.

30. Tatton-Brown, K., Hanks, S., Ruark, E., Zachariou, A., Duarte Sdel, V., Ramsay, E., Snape, K., Murray, A., Perdeaux, E.R., Seal, S., et al. (2011). Germline mutations in the oncogene EZH2 cause Weaver syndrome and increased human height. Oncotarget 2, 1127-1133.

31. Tatton-Brown, K., Douglas, J., Coleman, K., Baujat, G., Cole, T.R., Das, S., Horn, D., Hughes, H.E., Temple, I.K., Faravelli, F., et al. (2005). Genotype-phenotype associations in Sotos syndrome: an analysis of 266 individuals with NSD1 aberrations. Am J Hum Genet 77, 193-204.

32. Field, M., Tarpey, P.S., Smith, R., Edkins, S., O'Meara, S., Stevens, C., Tofts, C., Teague, J., Butler, A., Dicks, E., et al. (2007). Mutations in the BRWD3 gene cause X-linked mental retardation associated with macrocephaly. Am J Hum Genet 81, 367-374.

33. Eng, C. (1993). PTEN Hamartoma Tumor Syndrome. In GeneReviews(R), R.A. Pagon, M.P. Adam, H.H. Ardinger, S.E. Wallace, A. Amemiya, L.J.H. Bean, T.D. Bird, C.T. Fong, H.C. Mefford, R.J.H. Smith, et al., eds. (Seattle (WA), University of Washington, Seattle

34. Cottereau, E., Mortemousque, I., Moizard, M.P., Burglen, L., Lacombe, D., Gilbert-Dussardier, B., Sigaudy, S., Boute, O., David, A., Faivre, L., et al. (2013). Phenotypic spectrum of Simpson-Golabi-Behmel syndrome in a series of 42 cases with a mutation in GPC3 and review of the literature. American journal of medical genetics Part C, Seminars in medical genetics 163c, 92-105.

35. Saxena, A., and Sampson, J.R. (2014). Phenotypes associated with inherited and developmental somatic mutations in genes encoding mTOR pathway components. Seminars in cell & developmental biology 36, 140-146.

36. Mirzaa, G., Timms, A.E., Conti, V., Boyle, E.A., Girisha, K.M., Martin, B., Kircher, M., Olds, C., Juusola, J., Collins, S., et al. (2016). PIK3CA-associated developmental disorders exhibit distinct classes of mutations with variable expression and tissue distribution. JCI insight 1.

37. Barnard, R.A., Pomaville, M.B., and O'Roak, B.J. (2015). Mutations and Modeling of the Chromatin Remodeler CHD8 Define an Emerging Autism Etiology. Nucleic Acids Res 9, 477.

38. Harshman, S.W., Young, N.L., Parthun, M.R., and Freitas, M.A. (2013). H1 histones: current perspectives and challenges. Nucleic Acids Res 41, 9593-9609.

39. Kalashnikova, A.A., Rogge, R.A., and Hansen, J.C. (2016). Linker histone H1 and protein-protein interactions. Biochim Biophys Acta 1859, 455-461.

40. Maquat, L.E., and Li, X. (2001). Mammalian heat shock p70 and histone H4 transcripts, which derive from naturally intronless genes, are immune to nonsense-mediated decay. Rna 7, 445-456.

41. Qiao, Q., Li, Y., Chen, Z., Wang, M., Reinberg, D., and Xu, R.M. (2011). The structure of NSD1 reveals an autoregulatory mechanism underlying histone H3K36 methylation. J Biol Chem 286, 8361-8368.

42. Okano, M., Bell, D.W., Haber, D.A., and Li, E. (1999). DNA methyltransferases Dnmt3a and Dnmt3b are essential for de novo methylation and mammalian development. Cell 99, 247-257.

43. Kaneda, M., Okano, M., Hata, K., Sado, T., Tsujimoto, N., Li, E., and Sasaki, H. (2004). Essential role for de novo DNA methyltransferase Dnmt3a in paternal and maternal imprinting. Nature 429, 900-903.

44. Engelman, J.A., Luo, J., and Cantley, L.C. (2006). The evolution of phosphatidylinositol 3-kinases as regulators of growth and metabolism. Nat Rev Genet 7, 606-619.

45. Chen, W.Y., Shih, H.T., Liu, K.Y., Shih, Z.S., Chen, L.K., Tsai, T.H., Chen, M.J., Liu, H., Tan, B.C., Chen, C.Y., et al. (2015). Intellectual disability-associated dBRWD3 regulates gene expression through inhibition of HIRA/YEM-mediated chromatin deposition of histone H3.3. EMBO Rep 16, 528-538.

46. Katanoda, K. (2016). Neuroblastoma Mass Screening--What Can We Learn From It? Journal of epidemiology 26, 163-165.

47. Lapunzina, P. (2005). Risk of tumorigenesis in overgrowth syndromes: a comprehensive review. American journal of medical genetics Part C, Seminars in medical genetics 137c, 53-71.

48. Ernst, T., Chase, A.J., Score, J., Hidalgo-Curtis, C.E., Bryant, C., Jones, A.V., Waghorn, K., Zoi, K., Ross, F.M., Reiter, A., et al. (2010). Inactivating mutations of the histone methyltransferase gene EZH2 in myeloid disorders. Nature genetics 42, 722-726.

49. Ntziachristos, P., Tsirigos, A., Van Vlierberghe, P., Nedjic, J., Trimarchi, T., Flaherty, M.S., Ferres-Marco, D., da Ros, V., Tang, Z., Siegle, J., et al. (2012). Genetic inactivation of the polycomb repressive complex 2 in T cell acute lymphoblastic leukemia. Nature medicine 18, 298-301.

50. Bodor, C., Grossmann, V., Popov, N., Okosun, J., O'Riain, C., Tan, K., Marzec, J., Araf, S., Wang, J., Lee, A.M., et al. (2013). EZH2 mutations are frequent and represent an early event in follicular lymphoma. Blood 122, 3165-3168.

51. Okosun, J., Bodor, C., Wang, J., Araf, S., Yang, C.Y., Pan, C., Boller, S., Cittaro, D., Bozek, M., Iqbal, S., et al. (2014). Integrated genomic analysis identifies recurrent mutations and evolution patterns driving the initiation and progression of follicular lymphoma. Nature genetics 46, 176-181.

52. Yang, L., Rau, R., and Goodell, M.A. (2015). DNMT3A in haematological malignancies. Nature reviews Cancer 15, 152-165.

53. The Cancer Genome Atlas Network. (2015). Comprehensive genomic characterization of head and neck squamous cell carcinomas. Nature 517, 576-582.

54. Rauch, A., Wieczorek, D., Graf, E., Wieland, T., Endele, S., Schwarzmayr, T., Albrecht, B., Bartholdi, D., Beygo, J., Di Donato, N., et al. (2012). Range of genetic mutations associated with severe non-syndromic sporadic intellectual disability: an exome sequencing study. Lancet 380, 1674-1682.

55. de Ligt, J., Willemsen, M.H., van Bon, B.W., Kleefstra, T., Yntema, H.G., Kroes, T., Vulto-van Silfhout, A.T., Koolen, D.A., de Vries, P., Gilissen, C., et al. (2012). Diagnostic exome sequencing in persons with severe intellectual disability. N Engl J Med 367, 1921-1929.

56. Gilissen, C., Hehir-Kwa, J.Y., Thung, D.T., van de Vorst, M., van Bon, B.W., Willemsen, M.H., Kwint, M., Janssen, I.M., Hoischen, A., Schenck, A., et al. (2014). Genome sequencing identifies major causes of severe intellectual disability. Nature 511, 344-347.

57. Vissers, L.E., Gilissen, C., and Veltman, J.A. (2016). Genetic studies in intellectual disability and related disorders. Nat Rev Genet 17, 9-18.

58. Martinez, F., Caro-Llopis, A., Rosello, M., Oltra, S., Mayo, S., Monfort, S., and Orellana, C. (2016). High diagnostic yield of syndromic intellectual disability by targeted next-generation sequencing. J Med Genet.

Figure Titles and Legends

Figure 1. Causal mutation identified in 50% of OGID probands. Proportion of pathogenic mutations identified in 710 individuals with OGID. Epigenetic regulation genes (red), including *NSD1* which is the predominant gene, constitute the major gene set. PI3K/AKT pathway genes (blue) also significantly contribute to OGID.

Figure 2. ***HIST1H1E* mutations cause OGID.** (A) *HIST1H1E* mutations cluster within 12bp region in the carboxy-terminal domain (CTD) and have a similar predicted impact on protein function. The three different frameshift mutations generate the same open reading frame (Frame 2), which is predicted to reduce the length and net charge (at pH7) of the CTD compared to the wildtype (Frame 1). The other possible alternate reading frame (Frame 3) increases the protein length and net charge. CTD, carboxy-terminal domain; NTD, amino-terminal domain. (B, C and D) Facial images of three individuals with *HIST1H1E* mutations showing full cheeks and high hairline.

Figure 3. **Schematic of key biological processes impacted in OGID** . (A) Epigenetic regulation. NSD1, EED and EZH2 directly methylate specific histone tail lysine residues. DNMT3A is a *de novo* DNA methyltransferase, CHD8 is a chromatin remodeling complex protein that binds methylated lysine 4 of histone H3. H1.4 (encoded by *HIST1H1E*) stabilises higher order chromatin structures. All OGID mutations are predicted to lead to reduced function (B) PI3K/AKT pathway. The PI3K/AKT pathway positively regulates growth. AKT3, MTOR and p110α (encoded by *PIK3CA*) are pathway activators. PTEN and B56δ (encoded by *PPP2R5D*) are pathway suppressors. OGID mutations in *AKT3*, *MTOR* and *PIK3CA* are activating, whereas OGID mutations in *PTEN* and *PPP2R5D* are inactivating.

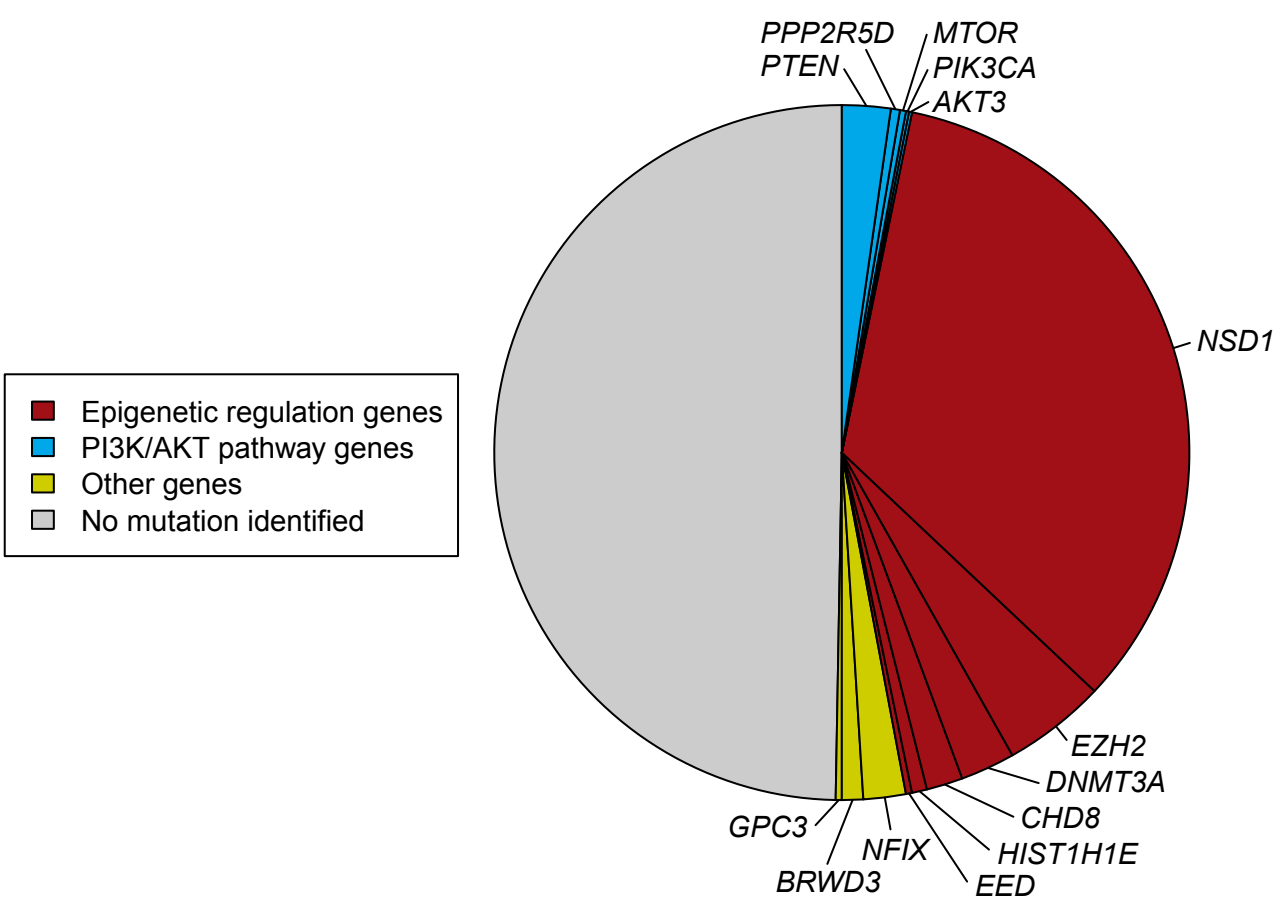
Figure 4. **Phenotypic differences between OGID due to mutations in epigenetic regulation genes compared to PI3K/AKT pathway genes.** Comparison of the distribution of (A) overgrowth categories and (B) degree of intellectual disability in cases with epigenetic regulation gene mutations (red) compared with PI3K/AKT pathway gene mutations (blue).

Figure 5. **Mutations in epigenetic regulation genes in OGID and cancers**. Protein schematics showing the position of mutations in *HIST1H1E, EZH2, DNMT3A, NSD1* and *CHD8* in OGID (below the gene) and specific cancers (above the gene). The somatic cancer driver mutations are from Lawrence et al.[27] AML, acute myeloid leukemia; CLL, chronic lymphocytic leukemia; GBM, glioblastoma multiforme; HNSC, head and neck squamous cell carcinoma; OGID, overgrowth-intellectual disability.

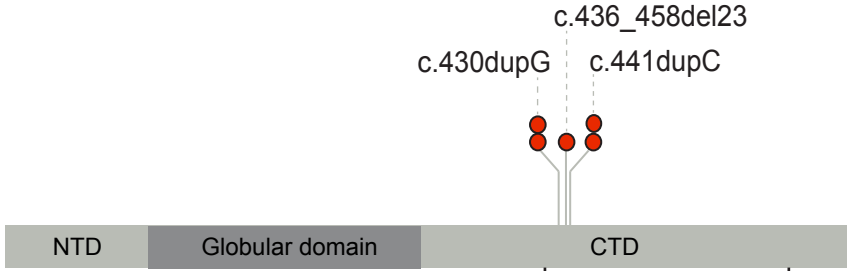
Tables

Table 1. Gene and transcript information for 14 genes involved in OGID

GENE	MIM Number	HGNC ID	Ensembl Transcript	RefSeq Transcript
AKT3	611223	HGNC:393	ENST00000366539	NM_005465
BRWD3	300553	HGNC:17342	ENST00000373275	NM_153252
CHD8	610528	HGNC:20153	ENST00000399982	NM_020920
DNMT3A	602769	HGNC:2978	ENST00000264709	NM_022552
EED	605984	HGNC:3188	ENST00000263360	NM_003797
EZH2	601573	HGNC:3527	ENST00000320356	NM_004456
GPC3	300037	HGNC:4451	ENST00000370818	NM_004484
HIST1H1E	142220	HGNC:4718	ENST00000304218	NM_005321
MTOR	601231	HGNC:3942	ENST00000361445	NM_004958
NFIX	164005	HGNC:7788	ENST00000360105	NM_002501
NSD1	606681	HGNC:14234	ENST00000439151	NM_022455
PIK3CA	171834	HGNC:8975	ENST00000263967	NM_006218
PPP2R5D	601646	HGNC:9312	ENST00000485511	NM_006245
PTEN	601728	HGNC:9588	ENST00000371953	NM_000314


Epigenetic regulation genes
PI3K/AKT pathway genes
Other genes
No mutation identified
PPP2R5D
PTEN
MTOR
PIK3CA
AKT3
NSD1
EZH2
DNMT3A
CHD8
HIST1H1E
EED
NFIX
BRWD3
GPC3

A



Frame	Mutation	Sequence	Length	Charge
1	Wildtype	...KKATGAATPKKSAKKTPKKAKKPAAAAGAKKAKSPKKAKAAKPKKAPKSPAKAKAVKPK...	219	44
2	c.430dupG	...KKATGGGHPQEERQEDPKEGEEAGCSCWSQKSEKPEKGESSQAKKGAQEPSEGQSS	194	7
2	c.441dupC	...KKATGAATPQEERQEDPKEGEEAGCSCWSQKSEKPEKGESSQAKKGAQEPSEGQSS	194	7
2	c.436_458del23	...KKATGAA--------DPKEGEEAGCSCWSQKSEKPEKGESSQAKKGAQEPSEGQSS	186	9
3		...KKATGRPPPRRAPRRPQRRRRSRLQLLEPKKRKARKRRKQPSQKRRPRAQRRPKQLNPR...	227	48

B COG0405

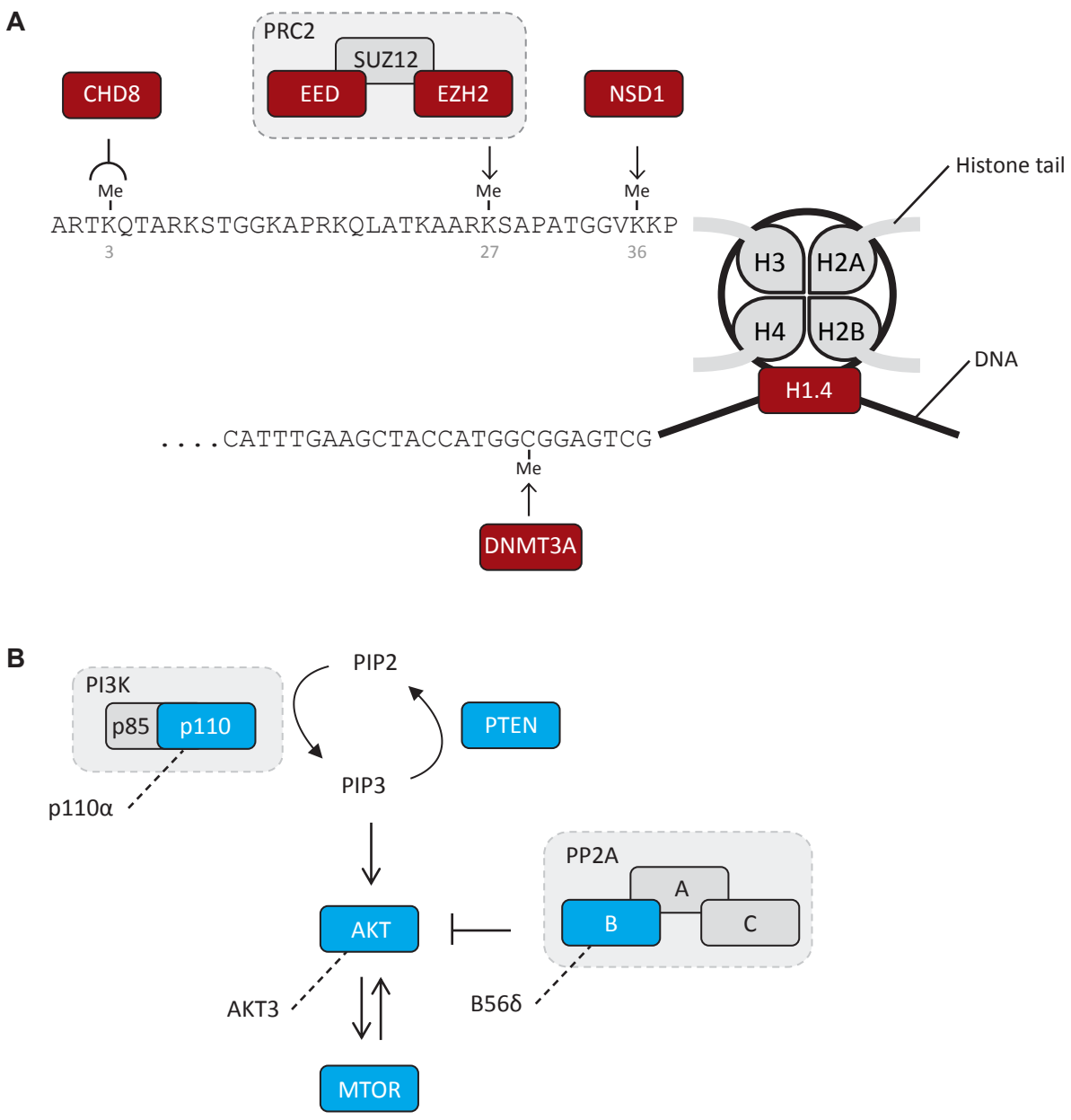
1.5 years 13 years

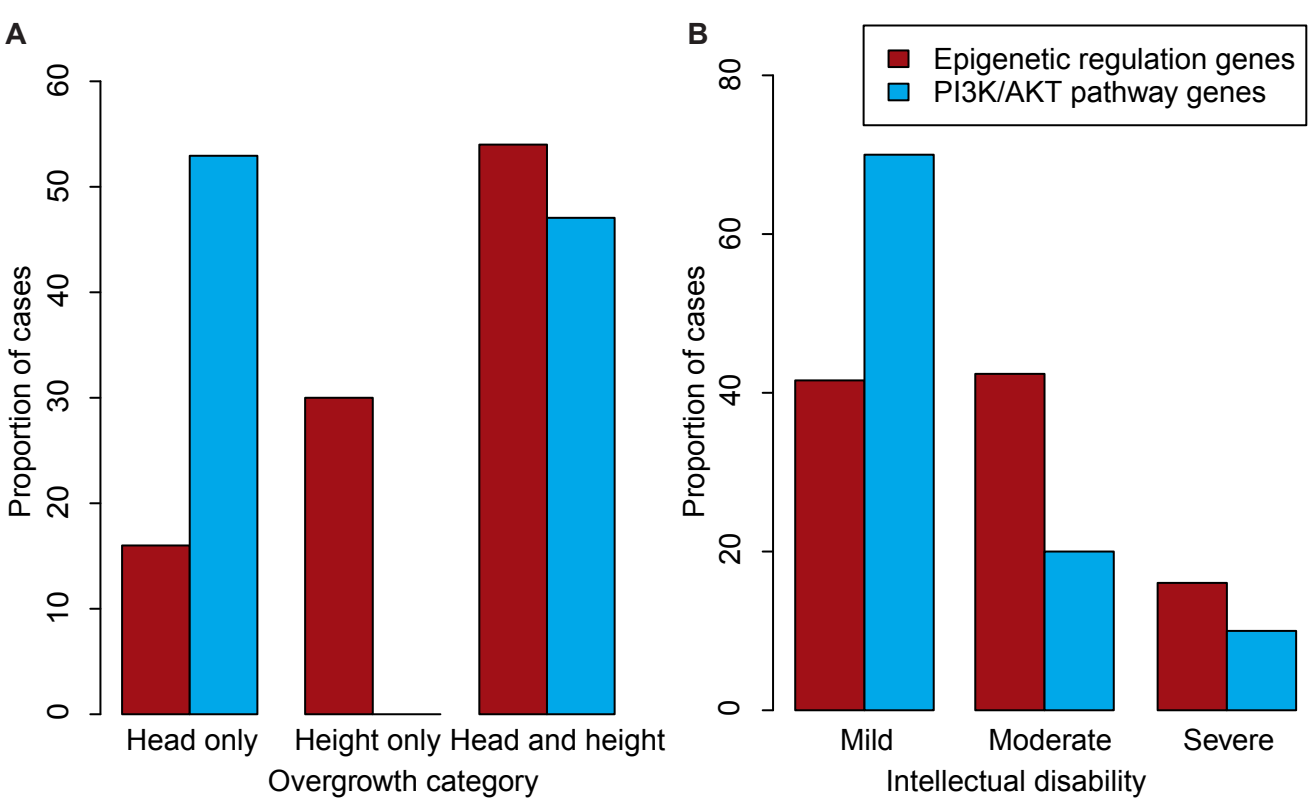
C COG0412

1.5 years 15.5 years

D COG1832

4 years


A
PRC2
SUZ12
EED
EZH2
CHD8
NSD1
Me
Me
Me
ARTKQTARKSTGGKAPRKQLATKAARKSAPATGGVKKP
3
27
36
Histone tail
H3
H2A
H4
H2B
DNA
H1.4
....CATTTGAAGCTACCATGGCGGAGTCG
Me
DNMT3A
B
PI3K
p85
p110
p110α
PIP2
PIP3
PTEN
AKT
AKT3
MTOR
PP2A
A
B
C
B56δ


A
Proportion of cases
0
10
20
30
40
50
60
Head only
Height only
Head and height
Overgrowth category
B
Epigenetic regulation genes
PI3K/AKT pathway genes
Proportion of cases
0
20
40
60
80
Mild
Moderate
Severe
Intellectual disability

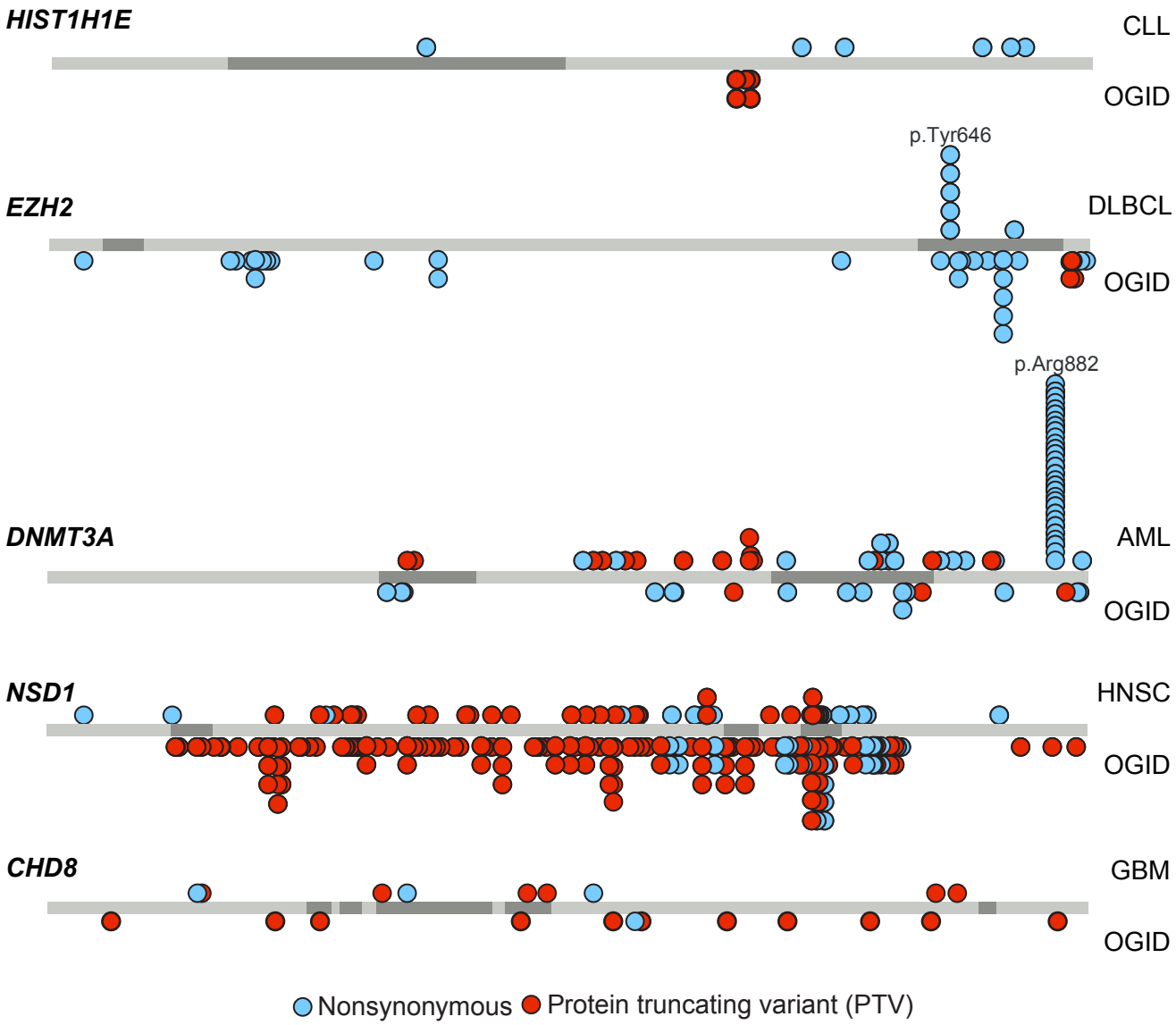
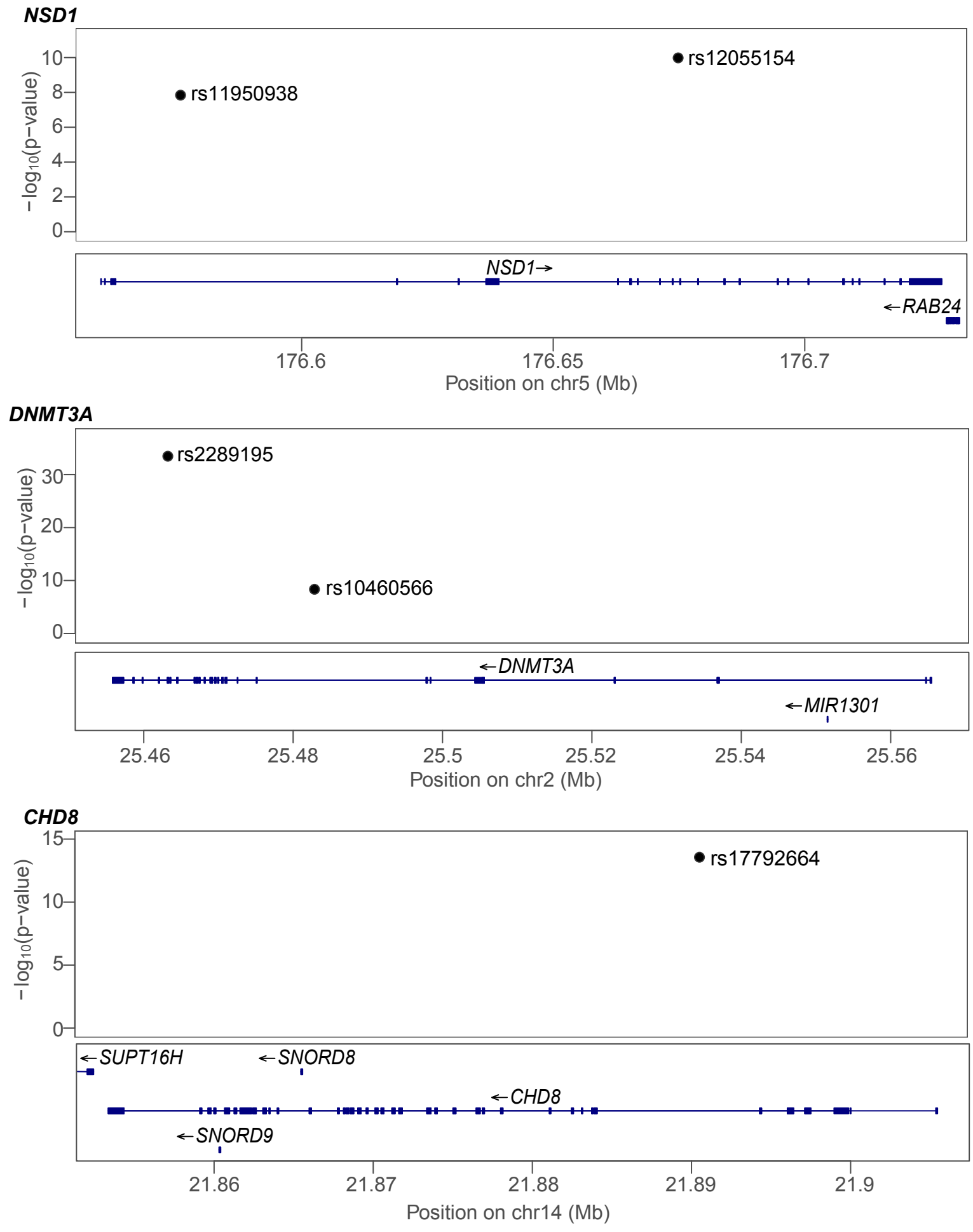
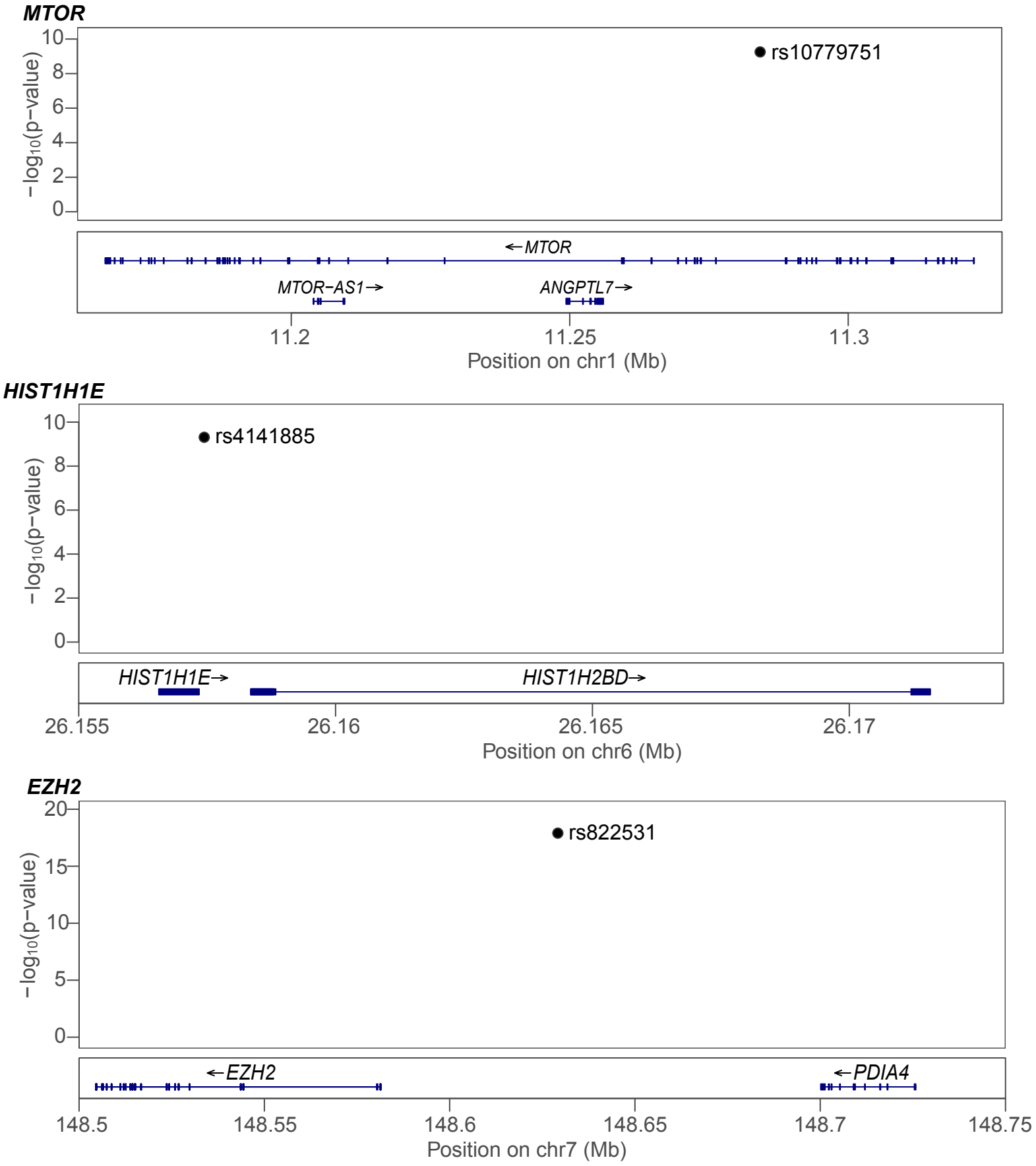

HIST1H1E
CLL
OGID
p.Tyr646
EZH2
DLBCL
OGID
p.Arg882
DNMT3A
AML
OGID
NSD1
HNSC
OGID
CHD8
GBM
OGID
Nonsynonymous
Protein truncating variant (PTV)

Figure S1. Height GWAS regional association plots for OGID genes





Regional association plots showing variants with significant associations in the height GWAS study by Wood et al.[27] that are in the vicinity of OGID genes.

Supplemental Note. The Childhood Overgrowth Collaboration

The following individuals coordinated recruitment and collection of the families and samples.

M-C. Addor, M. Akgul, L. Aksglaede, M. Ahmed, D. Amor, K. Anderson, R. Anderson, S. Andries, H. Archer, R. Armstrong, P. Ashton-Prolla, M. Bahceci, M. Balasubramanian, D. Baralle, D. Barge, A. Barnicoat, M. Barrow, J. Barwell, G. Baujat, G. Baynam, P. Beales, K. Becker, E. Beckh-Arnold, A. Ben-Yehuda, J. Berg, B. Bernhard, S. Bhal, M. Bhat, J. Birch, L. Bird, M. Bitner-Glindzicz, E. Blair, J. Bliek, M. Blyth, A. Bottani, M. Bouma, M. Boxill, F. L. Bradley, A. Brady, Breatnach, G. Brice, B. Buehler, A. Burke, J. Burn, J. Campbell, N. Canham, B. Castle, K. Chandler, R. Chandrasena, E. Chang, C. Christenden, C. Chu, D. Cilliers, A. Clarke, J. Clayton-Smith, C. Clericuzio, V. Clowes, T. Cole, A. Colley, A. Collins, F. Connell, J. Cook, I. Cordeiro, E. Crocker, Y. Crow, V. Culic, T. Cushing, T. Dabir, A. Dalton, S. Danda, R. Davidson, S. Davies, R. Day, D. Dearnaley, M-A. Delrue, M. De Roy, V. de Soberanis, M. de Ville, N. Dennis, C. Deshpande, B. Desouza, L. Devlin, A A. Dieckmann, -M. Differ, R. Dinwiddie, A. Dixit, A. Dobbie, J. Dominguez, A. Donaldson, D. Donnai, D. Donnelly, H. Dorkins, M. Doz, J. Dupont, D. Eastwood, M. Edwards, I. Ellis, F. Elmslie, L. Escobar, R. Evans, F. Faravelli, C. Fauth, H. Firth, R. Fisher, T. Fiskerstrand, D. Fitzpatrick, A. Flanagan, F. Flinter, P. Foley, A. Foster, N. Foulds, W. Foulkes, J. Franklin, A. Fryer, H. Fryssira, A. Gallagher, S. Garcia, C. Gardiner, M. Gardner, C. Garrett, B. Gener, M. Gerrard, R. Gibbons, Y. Gillerot, H. Goel, D. Goudie, K. Gowrishankar, C. Graham, A. Green, N. Gregersen, J. Hale, M. Hamilton, J. Harper, R. Harrison, V. Harrison, A. Henderson, P. Henman, R. Hennekam, E. Hobson, S. Hodgson, M. Holder, S. Holder, T. Homfray, D. Horovitz, H. Hughes, Z. Huma, M. Hunter, J. Hurst, W-L. Hwu, A. Irvine, M. Irving, L. Izatt, M-L. Jacquemont, S. Jagadeesh, L. Jenkins, U. Jensen, C. Jessen, D. Johnson, J. Johnson, E. Jones, L. Jones, A. Jorgensen, D. Josifova, S. Joss, Dr. Kanabar, P. Kannu, K. Keppler-Noreuil, B. Kerr, H. Kingston, J. Kingston, U. Kini, E. Kinning, A. Krause, V. Krishnamurthy, A. Kumar, D. Kumar, A. Medeira, V. Meiner, C. Mercer, K. Milstein, Y. Miyoshi, E. Moran, K. Lachlan, W. Lam, P. Lapunzina, M. Lees, N. Leonard, G. Levitt, I. Lewis, J. Liebelt, A. Livesey, C. Longman, T. Lopponen, Dr Lozano, A. Lucassen, P. Lunt, S-A Lynch, S. Lyonnet, J. MacDonnell, A. Magee, E. Maher, S. Maitz, A. Male, S. Mansour, C. Marcelis, E. McCann, V. McConnell, T. McDevitt, M. McEntagart, J. McGaughran, G. McGillivray, R. McGowan, S. McKee, C. McKeown, C. Meany, S. Mehta, K. Metcalfe, Z. Miedzybrodzka, S. Mohammed, G. Monaghan, T. Montgomery, A. Morgan, B. Morland, P. Morrison , J. Morton, R. Mudgal, A. Munaza, V. Murday, S. Nampoothiri, K. Nathanson, K. Neas, A. Nemeth, G. Neri, R. Newbury-Ecob, C. Nur Semerci, C. Ockeloen, C. Oley, C. Owen, K. Ozono, Panarello, S-M. Park, M. Parker, C. Patel, M. Patton, S. Payne, M. Pearson, J. Piard, D. Pilz, M. Pinkney, B. Plecko, M. Pocha, G. Poke, R. Posmyk, C. Pottinger, K. Prescott, S. Price, K. Pritchard-Jones, A. Proctor, V. Puthi, O. Quarrell, A. Raas-Rothchild, E. Rahikkala, W. Raith, J. Rankin, L. Raymond, G. Rea, L. Read, W. Reardon, E. Reid, H.Rees, N. Revencu, O. Rittinger, M. Robards, A. Roposch, E. Rosser, D. Rourke, D. Ruddy, A. Saggar, N. Saleh, V. Saletti, J. Sampson, R. Sandford, H. Santos, A. Sarkar, R. Scott, I. Scurr, C. Searle, A. Selicorni, R. Semple, S. Sharif, A. Shaw, C. Shaw-Smith, D. Shears, J. Shelagh, N. Shur, L. Side, M. Simon, F. Skovby, G. Smith, S. Smithson, M. Splitt, M. Stevens, A. Stewart, F. Stewart, H. Stewart, K. Stopps, C. Stumpel, K. Stuurman, D. Subramanian, M. Suri, A. Swain, E. Sweeney, K. Szakszon, Y. Sznajer, G. Tanateles, A. Taylor, C. Taylor, M. Teixeira, I.K. Temple, E. Thomas, E. Thompson, F. Thonney, M. Tischowitz, J. Tolmie, S. Tomkins, S. Turkmen, A. Turner, P. Turnpenny, M. Van-Haelst, L. Van Maldergem, P. Vasudevan, I. Veenstra-Knol, C. Verellen, I.C. Verma, J. Vigneron, E. Wakeling, L. Wainwright L. Walker, D. Weaver, P. Wheeler, K. White, S. White, M. Whiteford, D. Williams, L. Wilson, R. Winter, G. Woods, M. Wright, N. Yachelevich, A. Yeung, A. Zankl