The American Journal of Human Genetics, Volume *87*

Supplemental Data

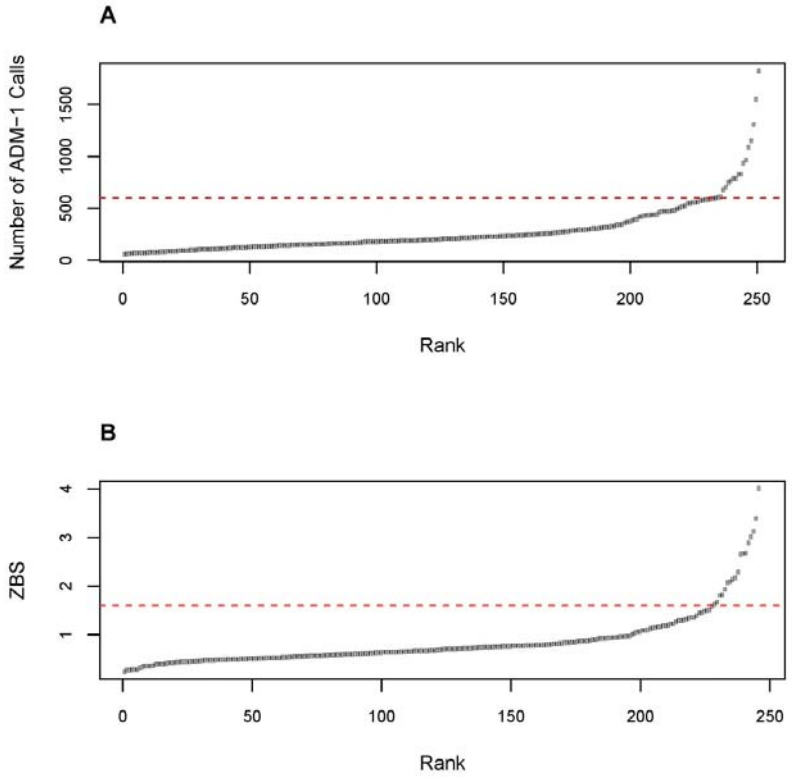
Fine-Scale Survey of X Chromosome

Copy Number Variants and Indels

Underlying Intellectual Disability

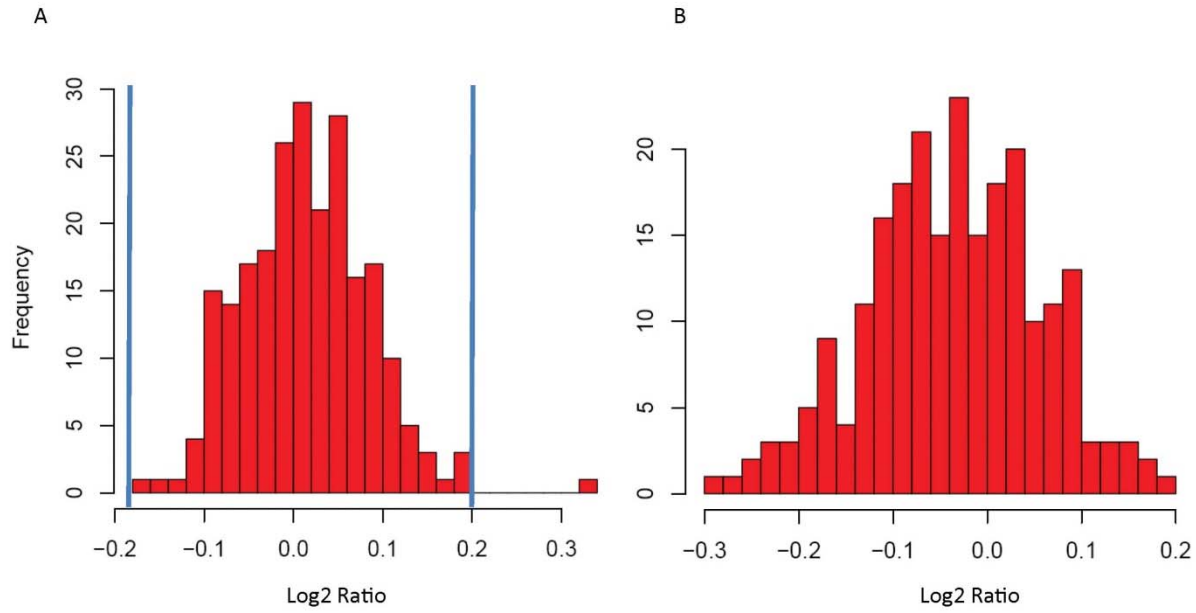
Annabel C. Whibley, Vincent Plagnol, Patrick S. Tarpey, Fatima Abidi, Tod Fullston, Maja K. Choma, Catherine A. Boucher, Lorraine Shepherd, Lionel Willatt, Georgina Parkin, Raffaella Smith, P. Andrew Futreal, Marie Shaw, Jackie Boyle, Andrea Licata, Cindy Skinner, Roger E. Stevenson, Gillian Turner, Michael Field, Anna Hackett, Charles E. Schwartz, Jozef Gecz, Michael R. Stratton, and F. Lucy Raymond

Figure S1: Sample level QC



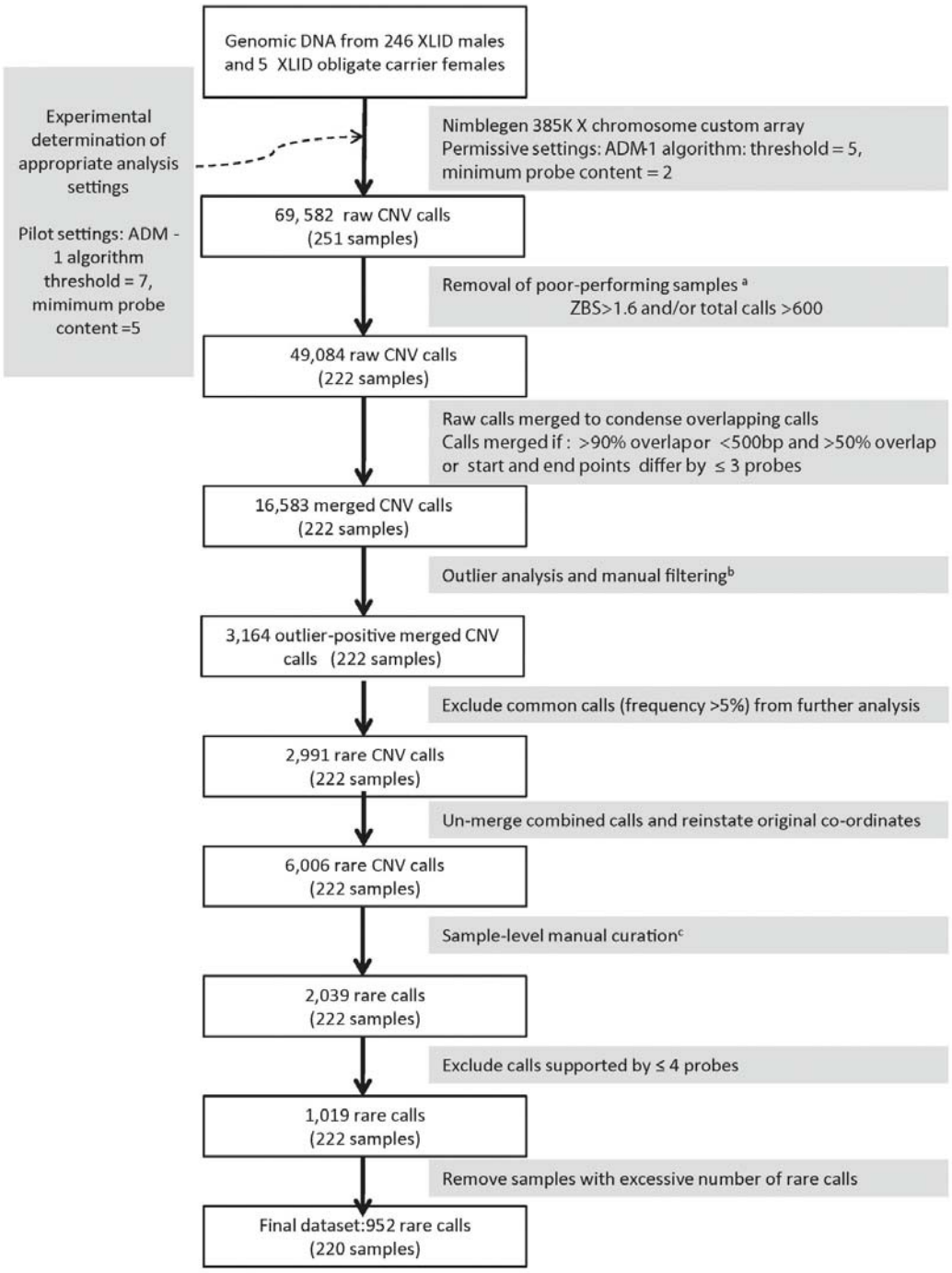
Plots of sample rank against QC indices. Two measures of sample performance were used for QC: (A) Total ADM-1 call number and (B) a z-score based statistic (ZBS). Dashed red lines indicate thresholds above which samples were excluded from high-resolution analysis (total call number >600 and ZBS>1.6). The ZBS was calculated by i) estimating the mean and standard error of the intensity distribution for each probe across the sample population; ii) using these estimates to compute, for each pair of probe and individual, a normalized z-score; iii) to obtain a unique summary statistic per sample we summarized the z-score values across all probes by computing the ratio of the 90% quantile of the squared z-score distribution divided by its theoretical value (90% of a chi-squared distribution on one degree of freedom: 2.7). Female samples were excluded from ZBS analysis.

Figure S2: CNV level QC by outlier analysis



Representative (A) good quality call and (B) rejected call, each supported by 24 probes. CNV-level QC was based on evaluation of the performance of the probes reporting the CNV within the context of the sample population as a whole, after the exclusion of poor-performing samples. For each merged CNV call, histograms of log2 ratio were subjected to automated outlier calling and manual curation, including the review of all large CNVs (>50kb) to assess the probe-level evidence for complex rearrangements. A rare CNV call was considered valid if the average log2 ratio was discrete from the distribution of the main population.

Figure S3: CNV workflow



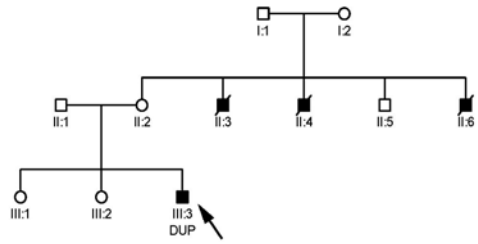
[a] as described in Supplementary Figure 1; [b] as described in Supplementary Figure 2; [c] Overlapping ADM-1 calls within the same sample were condensed. Since outlier analysis was based on mean log2 values but the mean log2 value can be influenced by the contribution of a small number of anomalous probes with extreme log2 values, we ranked the log2 ratio of each probe contained within the CNV and computed the median rank for each sample to identify and exclude low-confidence calls. Only calls supported by an average median rank in the upper or lower 20% of the population were retained.

Figure S4: Pedigrees of XLID families with likely pathogenic copy number variants

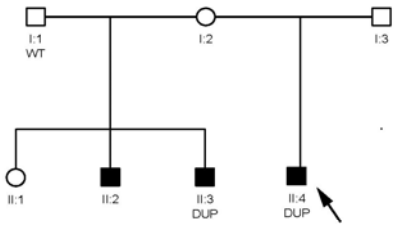
For each family, the individual analyzed by aCGH is indicated by an arrow. (A) Family 57 *MECP2* duplication; (B) Family 340 *MECP2* duplication; (C) Family 344 *MECP2* duplication; (D) Family 389 *MECP2* duplication; (E) Family 495 *MECP2* duplication; (F) Family 509 *MECP2* duplication; (G) Family 538 *HUWE1* duplication; (H) Family 376 Xp22.13-Xp22.11 duplication; (I) Family 505 *ARX* duplication; (J) Family 110 *AFF2* duplication; (K) Family 32 *IL1RAPL1* deletion; (L) Family 398 *SLC16A2* deletion; (M) Family 399 *SLC16A2* deletion; (N) Family 115 *SLC9A6* deletion; (O) Family 147 *MAOA* and *MAOB* deletion.

Figure S4: Continued

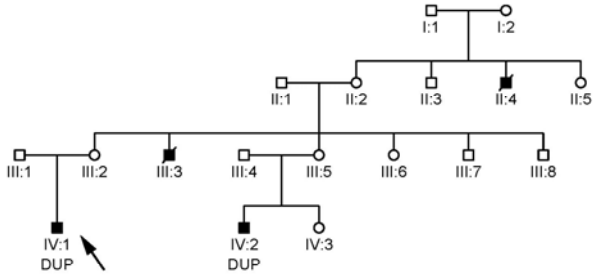
A Family 57



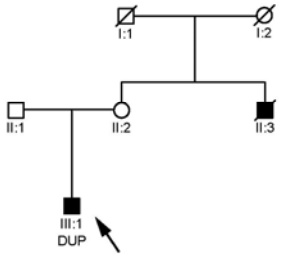
B Family 340



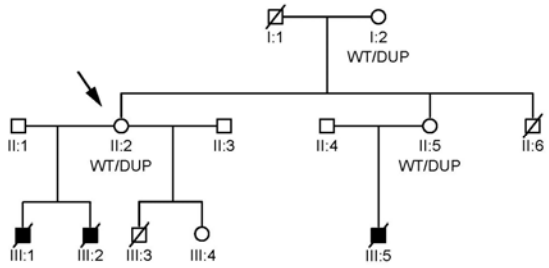
C Family 344



D Family 389



E Family 495



F Family 509

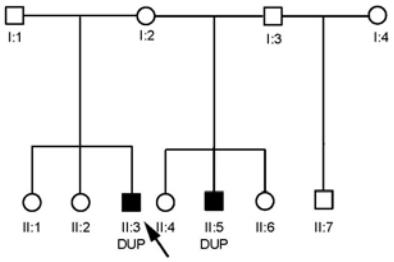
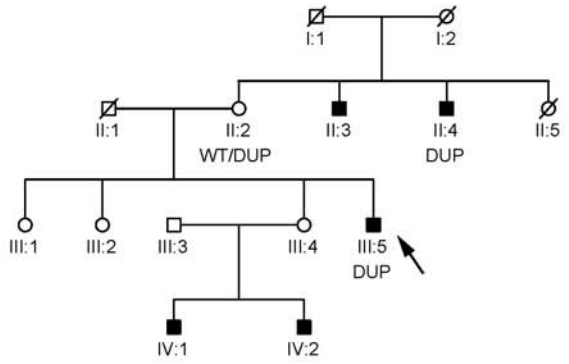
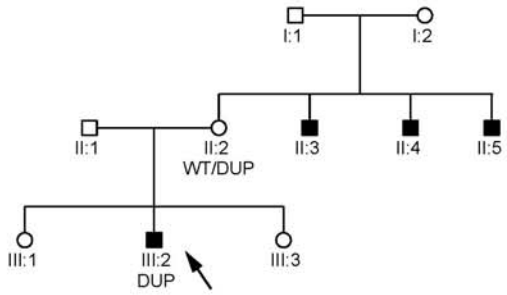


Figure S4: Continued

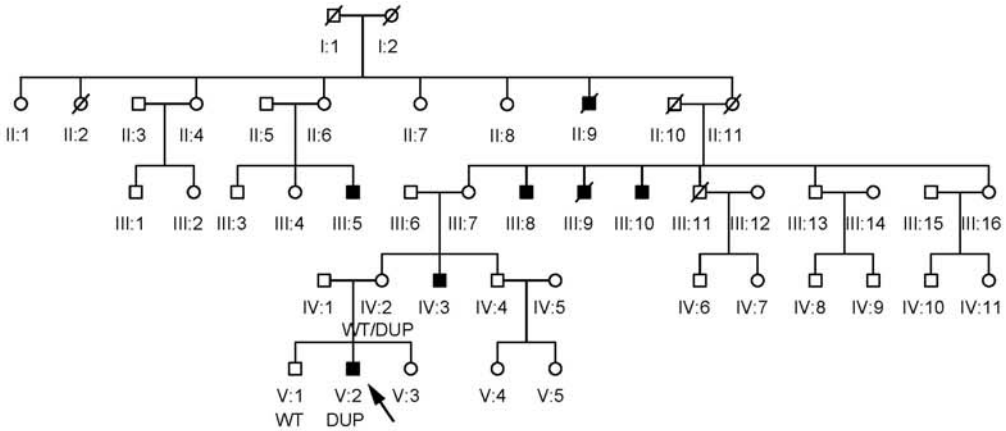
G Family 538



H Family 376



I Family 505



J Family 110

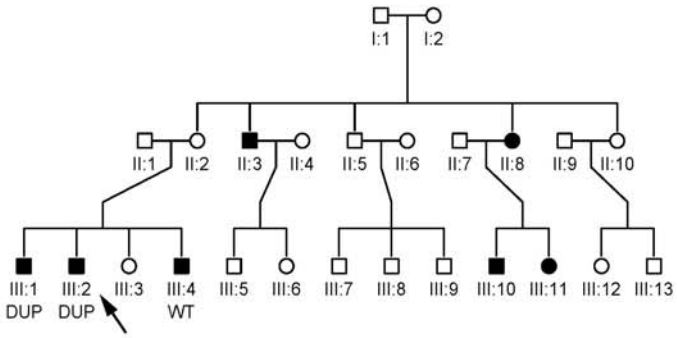


Figure S4: Continued

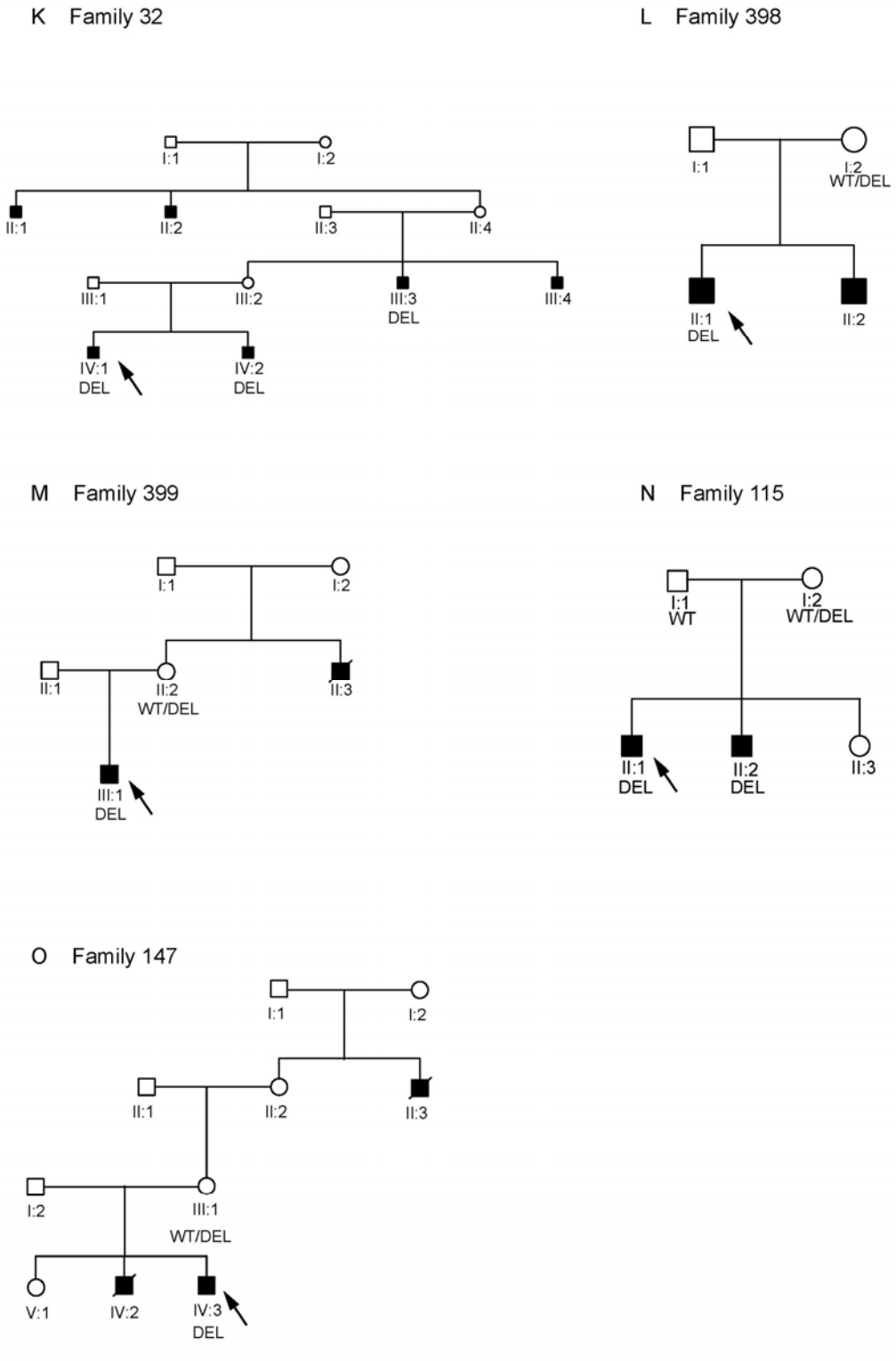
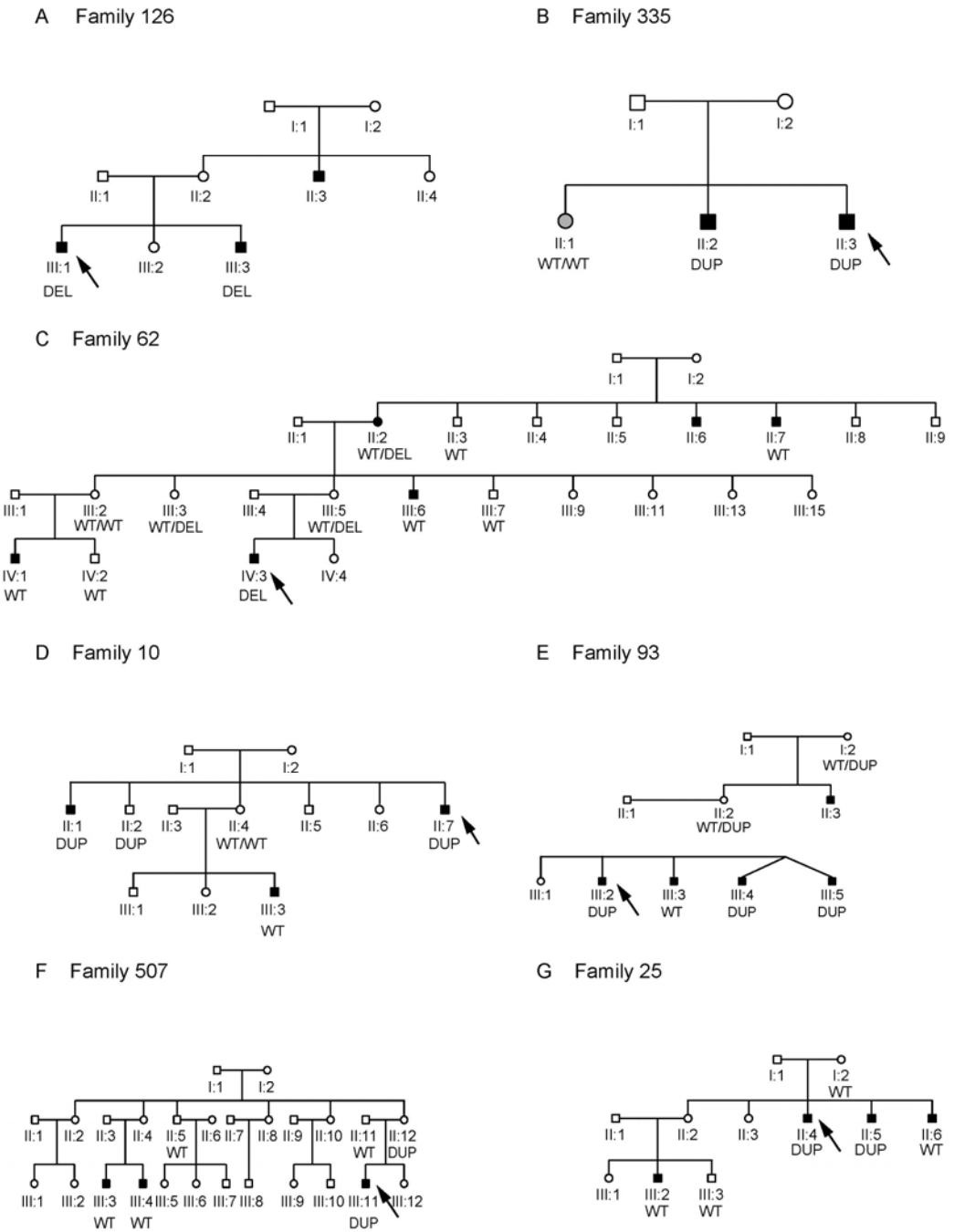
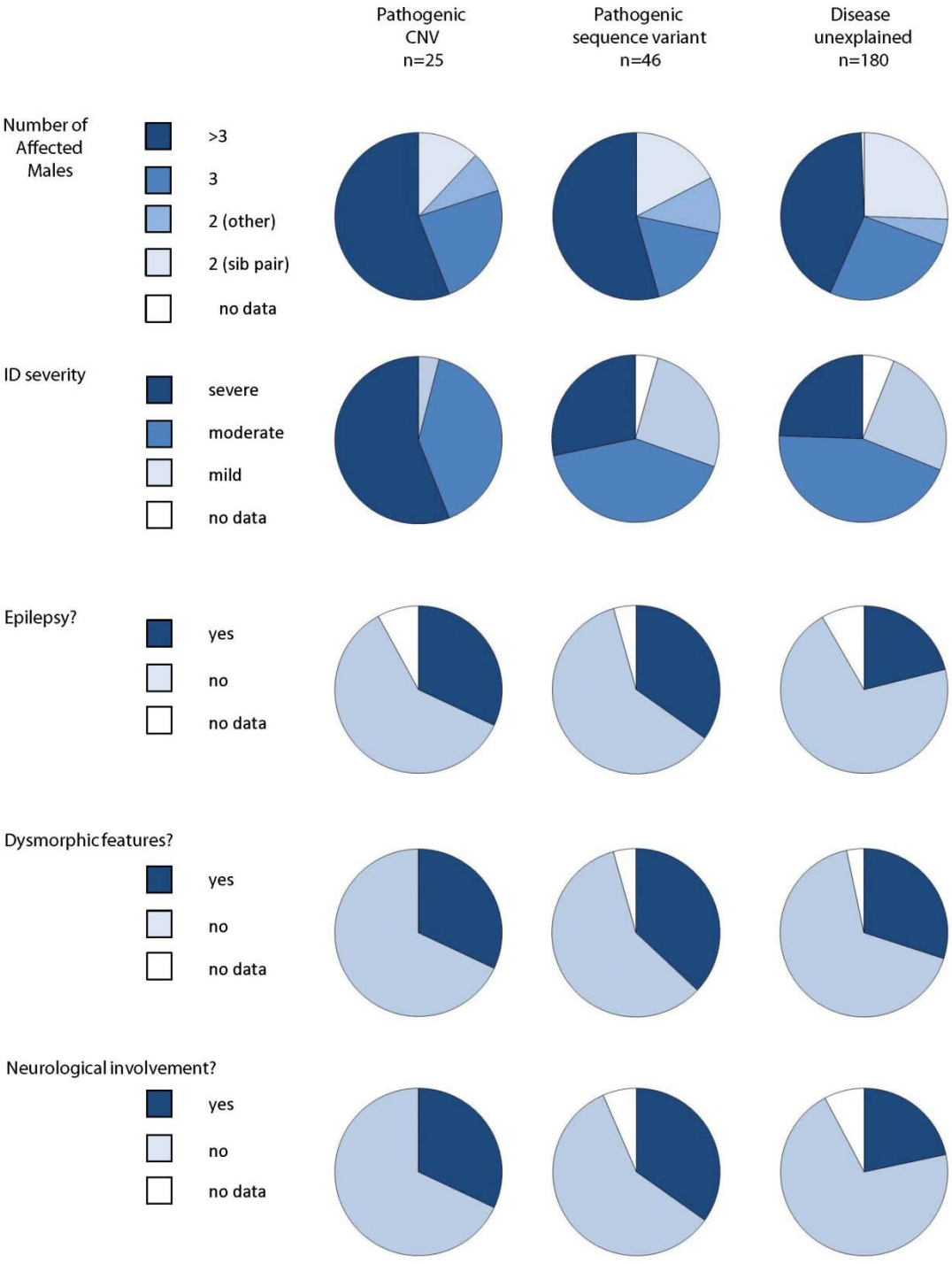


Figure S5: Pedigrees and segregation data for families with CNVs of unknown significance or without disease association



For each family, the individual analyzed by aCGH is indicated by an arrow. (A) Family 126 *NLGN4X* non-coding deletion; (B) Family 335 PCDH11X non-coding deletion. Grey shading indicates mildly affected female; (C) Family 62 *AWAT1* deletion; (D) Family 10 *IDS* duplication; (E) Family 93 both duplication calls (containing *XIAP* and *STAG2*) co-segregate; (F) Family 507 Xp22.33 duplication; (G) Family 25 Xp22.33 duplication.

Figure S6: Cohort comparisons



Pie charts showing pedigree features and clinical characteristics of IGOLD cohort families subdivided according to whether disease can be attributed to a pathogenic CNV, sequence variant or remains unexplained.

Figure S7: Sequence alignments of deletion and duplication breakpoint junctions

For each sequenced breakpoint, junction fragment sequence (JF) is aligned to proximal (A) and distal (B) reference sequence matches, shown in blue and orange. Short stretches of microhomology at the breakpoint, which could be derived from either proximal or distal sequence, are in bold underlined black type. Inserted sequence found at the breakpoint in family 206 is indicated in green and likely results from a serial replication slippage event.

A) FAMILY 110 (DUPLICATION)

```
110_A   GTCACTACAGAGGCTATGTAAATACTCCAAGTTCCAAAGACAAAGGGATTTCCAGTACTTCCTCAAGGCCATCTCTCATA

JF      GTCACTACAGAGGCTATGTAAATACTCCAAGTTCCAAAGACTAATTATGTGCTGATTGTTGAATGGTGAACTTCTTGAGG

110_B    GGAAGGTAGAGTGCTTACTCTGGCGACCTGGACTGGACTACTAATTATGTGCTGATTGTTGAATGGTGAACTTCTTGAGG
```

B) FAMILY 121 (DELETION)

```
121_A   TAGTAGAGGACTTTTTTCTTTCTTAAAGCAAATTATATGCCTTCAGCAGTACTTATTTTTACAAAATGCAGAAGACTGTA

JF      TAGTAGAGGACTTTTTTCTTTCTTAAAGCAAATTATATGCCAGATCTCACATGAACTGAGAGTGAGGACTCATTCATTAT

121_B   GAGGGAGAAAGAGTGGGGAGGTGCCACAGTTTTAAACAACCAGATCTCACATGAACTGAGAGTGAGGACTCATTCATTAT
```

C) FAMILY 147 (DELETION)

```
147_A   AGCCTCCCGAGTAGCTGGGACTACAGGTGCCCACCACCATGCCCAGCTAATTTTTTGTATTTTTAGTAGAGATGGGGTTT

JF      AGCCTCCCGAGTAGCTGGGACTACAGGTGCCCACCACCATAGCAGCATGATTTATAATCCTTTGGGTATACACCCAGTAA

147_B   AACAGTGCTGCAATAACCATACGTGTGCATGTGTCTTTATAGCAGCATGATTTATAATCCTTTGGGTATACACCCAGTAA
```

D) FAMILY 206 (DELETION)

```
206_A   AGGGACCAGCCAGTCGCATCAACAGCAGCCCCACTGTGGTCAGACTCAAGGAGGGCCTGGCTGGAAGCAGGATGATGGTC

JF      AGGGACCAGCCAGTCGCATCAACAGCAGCCCCACTGTGGTCTCAAACCTCCCCAAGCCATGCTAGGACCTCACACCTCGG
                                                 ↑
                                          CTCAAACCTCCC

206_B   CAAAACAGCCAGGTGACCTCCCATCTAGGGATGGGGCCCCCTCAAACCTCCCCAAGCCATGCTAGGACCTCACACCTCGG
```

E) FAMILY 317 (DUPLICATION)

317_A CAATGTGAATGGCATCATCTAATCAGCTGAAGGCCTAGATAGAATCAAAAGATAAAGGAAAAAGCAAACTCTCTCTCTCT

JF CAATGTGAATGGCATCATCTAATCAGCTGAAGGCCTAG**ATA**CTGGAAACAGCCTTATGCAATGGCAGGGTTAACTTTGTG

317_B CTGCTGGGTTCTGTTTGGGTCTCTTCTCCTTGTGCTGCATACTGGAAACAGCCTTATGCAATGGCAGGGTTAACTTTGTG

F) FAMILY 398 (DELETION)

398_A GATTACAGGCGTGAGCCACCACACCCGGCAACTCCATTTCTTTTCTGTAGGATTTATTGTAAACATTGATTCACGCATTA

JF GATTACAGGCGTGAGCCACCACACCCGGCAACTCC**ATTTCTTTT**GGGAAGGGGAAGCCATCCCAGCCTTTACCACCCAAT

398_B AAGGTACAGGCAGCAGTTGCACATAGGGTTGGAAAATTTCTTTTGGGAAGGGGAAGCCATCCCAGCCTTTACCACCCAAT

G) FAMILY 399 (DELETION)

399_A AAAGTTTCTGGAGAAACTTTAGAGAAATTTTACTGGTCTTTCAGATTGAAATTTTGGGGAAGCATTGTTTTGGTTTTCTG

JF AAAGTTTCTGGAGAAACTTTAGAGAAATTTTACTGGTC**TT**GGAGATTTCTCAAAGAATTTAAAACAGAACTACCGTTTGA

399_B TGGGATTACAGGCGTGTGCCACTGCACCCAGTCTCAGTTTGGAGATTTCTCAAAGAATTTAAAACAGAACTACCGTTTGA

H) FAMILY 463 (DUPLICATION)

463_A ATAACCAAAATCAGCAAGGGAAAGGGAAAAGAAAAGAAGAACCTGCTAAATGAACATGCCATGTATCAGAACAAACAGAA

JF ATAACCAAAATCAGCAAGGGAAAGGGAAAAGAAAAGAA**GAA**AATTACAAGAAAAATAAAACGCTACAGCCGAATAAAAGA

463_B GGTATGTCAACATTAGGTAAAGAGGACTTCATCGCAACGAAAATTACAAGAAAAATAAAACGCTACAGCCGAATAAAAGA

I) FAMILY 505 (DUPLICATION)

505_A **GGGTTAGATAGCGGGTTATAACGGATATTATTGCGATCTTTGTGCCTTTCTGCCTCCCTTGGTTGCCGGCTGCCGGCTCC**

JF **GGGTTAGATAGCGGGTTATAACGGATATTATTGCGATCT**T**GTATCCCAGAACTTAAAAAAGAAAACAAAGAAACCCTAAT**

505_B ACATGTGTACCTATGTAACAAACCTGCACGTTCTGCACATGTATCCCAGAACTTAAAAAAGAAAACAAAGAAACCCTAAT

J) FAMILY 506 (DELETION)

506_A AAGGGGGGGGCTACACCGGGGGAATGGGAGGGTTTGGGAAGGCGGATAGGCTGACACCAGGAGTGAGCAGAACGAGGGGG

JF AAGGGGGGGGCTACACCGGGGGAATGGGAGGGTTTGGG**AA**TTTTGTGTTTTTAGTAGAGACGGGGTTTCTCCATGTTGGC

506_B GTAGCTGGGATTACAGGCATGCAGCACCACGCCTGGCTAATTTTGTGTTTTTAGTAGAGACGGGGTTTCTCCATGTTGGC

K) FAMILY 540 (DELETION)

540_A AATTTCTCTACTGGTTTGTACTTGTTGAAGATAAGCATGTGTCCTGAACATTTATTATATGTTATACGTCCTGTAGATCT

JF AATTTCTCTACTGGTTTGTACTTGTTGAAGATAAGCATGTGCCCGGCAAGGGCTTTGTTCATTAGGATCAACAAGGTGCT

540_B GGCCTCCCAAAGTGCTGGGATTACAGGCGTGAGCCACCGTGCCCGGCAAGGGCTTTGTTCATTAGGATCAACAAAGTGCT

L) FAMILY 126 (DELETION)

126_A AGGAATTAAAGACACACAGAAATATAGAGGTGTGGAGTGGGAAATCAGGGGTCTCACAGCCTTCAGAGGTGAGAGCCTCA

JF AGGAATTAAAGACACACAGAAATATAGAGGTGTGGAGTGGAAAACAACAAAATCTCTAATACCCTATGTACTGCATTTAC

126_B AGCCTGGGTTGATAGAGCAAGACCCAGTCTCTAAAAAAATAAAACAACAAAATCTCTAATACCCTATGTACTGCATTTAC

Table S1: Pathogenic CNVs identified in parallel analyses of the IGOLD families

Family	Mutation	Mode of Identification	Reference
77	*HSD17B10* and *HUWE1* duplication	BAC Array	[1]
121	*IL1RAPL1* intragenic deletion	BAC Array	G.Froyen, unpublished data
164	*MECP2* duplication	MLPA	[2; 3]
185	*MECP2* duplication	MLPA	[2; 3]
241	*MECP2* duplication	MLPA	unpublished data
304	*HSD17B10 and HUWE1* duplication	BAC Array	[1]
317	Xq24-Xq25 duplication	BAC Array	G.Froyen, unpublished data
340	*MECP2* duplication	MLPA	[4]
359	*HSD17B10* and *HUWE1* duplication	BAC Array	[1]
422	Xq13.1-q21.1 duplication	BAC Array	[5], G.Froyen, unpublished data
495	*MECP2* duplication	MLPA	[4]
509	*MECP2* duplication	MLPA	unpublished data

MLPA: multiplex ligation-dependent probe amplification; BAC: Bacterial artificial chromosome.

Table S2: Clinical summary of IGOLD cohort

	No. families (n=251)	%
No. affected males		
2 (sib pair)	57	22.7%
2 (other)	17	6.8%
3	61	24.3%
>3	116	46.2%
Ancestry		
European	228	90.8%
Black African	4	1.6%
Asian	5	2.0%
European/Aboriginal	1	0.4%
European/Asian	1	0.4%
Hispanic	1	0.4%
No data	11	4.4%
Severity of intellectual disability		
Severe (IQ 20-34)	71	28.3%
Moderate (IQ 35-49)	109	43.4%
Mild (IQ 50-69)	58	23.1%
No data	13	5.2%
Head circumference		
Macrocephaly	34	13.5%
Microcephaly	31	12.4%
Normal	162	64.5%
No data	24	9.6%
Epilepsy		
Yes	62	24.7%
No	170	67.7%
No data	19	7.6%
Speech and language		
Absent	32	12.7%
Delayed	206	82.1%
Normal	2	0.8%
No data	11	4.4%
Dysmorphic features		
Yes	79	31.5%
No	164	65.3%
No data	8	3.2%
Neurological features		
Yes	63	25.1%
No	171	68.1%
No data	17	6.8%

Table S3: Array design and probe allocation

Design	Feature	Number of Probes	Mean Probe Spacing (bp)	Median Probe spacing (bp)
Targeted Design	Coding Sequences (742 genes)	103 481	59	36
	Ultra Conserved Elements (n=27)			
	Highly Conserved Elements (n=181)			
	Autosomal Sub-Telomeres	8 074	Nd	nd
Backbone Design	X chromosome	271 126	506	463
	Random X chromosome	2 119	756	476

The X chromosome design comprised two parts: targeted high-density coverage of regions of interest and a genomic backbone covering all 155Mb of the X chromosome. In addition to the coding sequences of Vega-annotated genes, we targeted the X chromosome members of the ultra conserved elements identified by Bejerano et al [6] and 181 X chromosome members of the top 5000 highly conserved elements defined by Siepel et al [7]. Probes for autosomal sub-telomeres were also included, with no imbalances detected. The probe spacing of the targeted design was calculated by dividing the target size by the number of probes contained within it. The probe spacing of the backbone design was calculated from the lengths of uninterrogated sequences between adjacent backbone probes, with the centromere excluded. A maximum of two unique matches per oligonucleotide was permitted in order to effectively cover pseudoautosomal and other X-Y homologous regions. nd= not determined.

Table S4: Experimental validation of analysis settings

Deletion	Call extent	Number of Probes	Gene content	Pilot Analysis Settings			Permissive Analysis Settings		
				Number of calls	FN	FP	Number of calls	FN	FP
Vdel_1	19376389-19378670	2-8	*MAP3K15* (intronic)	5	10	0	14	1	0
Vdel_2	33952795-33980684	53-57	Intergenic	3	2	0	5	0	0
Vdel_3	69648049-69650096	3-7	Intergenic	3	9	0	12	0	0
Vdel_4	131766775-131769279	4-8	*HS6ST2* (intronic)	7	12	2	14	5	2
Vdel_5	134634881-134636316	12-18	Intergenic	4	2	0	5	1	0
Vdel_6	143435664-143445262	5-30	Intergenic	15	6	0	20	1	0
Vdel_7	153283848-153284100	3-10	*DNS1L1* (3'UTR)	6	2	0	9	0	1
Vdel_8	154044709-154057314	23	Intergenic	4	1	0	4	1	0
TOTAL				47	44	2	83	9	3

Analysis thresholds were evaluated by PCR validation of a sample of 8 small polymorphic deletions which had been identified in the analysis. FN = number of false negative calls, FP= number of false positive calls. The pilot settings were ADM-1 threshold = 7, minimum probe content = 5. By reducing the analysis stringency to the permissive settings (ADM-1 threshold = 5, minimum probe content = 2), the calling of the polymorphic deletions was improved: the false negative rate was reduced to 10% (9/89 calls) at permissive settings, compared to 49% (44/89 calls) with pilot settings. There was minimal impact on false positive calls: using pilot settings, false positive rate was 4.2% (2/47 calls) and 3.6% (3/83 calls) using permissive settings. The detection efficiency varied considerably between loci and fell close to the detection limit of the array in terms of probe content and/or size.

Table S5: Experimentally verified CNVs in QC fail samples

Family	**Type**	**Genes**	**Genomic Co-ordinates**	**Extent (kb)**
495 (female sample)	Dup	Several, including *MECP2*	152,463,832-154,426,868 (M)	1,963
463	Dup	*GSPT2*	51,469,871-51,509,041 (S)	39
717	Dup	Several, including *VCX*, *VCX2*, *VCX3A* and *STS*	6,458,251-8,098,324 (E)	1,640
422	Dup	Several, including *MED12*, *NLGN3*, *SLC16A2*, K*IAA2022*, *ATRX* and *BRWD3*	70,134,868-81,653,582 (E)	11,519
350 (female sample)	Dup	Several, including *ZNF630*	47,747,299-47,887,027 (E)	140
763	Del	Several, including *ZNF630*	47, 882,521-47,748,841 (E)	134

Letters in parentheses after genomic co-ordinates specify whether CNV bounds are ADM-1 estimates (E) or have been adjusted after breakpoint sequencing (S) or qPCR and manual inspection of probe log2 ratios (M).

Table S6: Rare CNVs and indels identified in the IGOLD cohort using high resolution analysis

Abbreviations

N probes: Number of probes reporting CNV call. Overlaps extracted using the Tables function in the UCSC genome browser, except exon overlap which utilised Ensembl API. LCR overlap: CNV region coincides, at least partially, with Washington SegDup listing. DGV overlap: CNV region coincides, at least partially, with DGV variant listing. SNP overlap: CNV region contains dbSNP listing(s). Experimental Validation: Y indicates deletion or duplication confirmed by further analysis. SNP: No deletion or duplication but single nucleotide difference between test and reference samples detected. For Validation Method; qPCR: quantitative real time PCR, QMPSF: quantitative multiplex PCR of short fragments, PCR: standard PCR, Multiplex PCR: standard PCR incorporating an additional primer pair to control for amplification failure, Junction fragment analysis: PCR fragments spanning the rearrangement breakpoint were amplified by PCR and sequenced, FISH: fluorescent in situ hybridization, RT-PCR: reverse transcriptase PCR and sequence analysis of gene transcripts in RNA extracted from patient LCL.

See separate Excel file available online.

Table S7: Summary of CNV and in dels experimental validations

CNV Size	Deletions				Duplications			
	Number of samples	**Number of loci**	**Confirmed deletion**	**Underlying Sequence Variant**	**Number of samples**	**Number of loci**	**Confirmed duplication**	**Underlying sequence variant**
<1kb	15	9	2/9	5/9	2	2	0/2	**1/2**[a]
1-10kb	20	12	11/12	1/12	8	2	2/2	**0/2**
>10kb	3	3	3/3	0/3	26	24	24/24	**0/24**

CNVs were subdivided into 3 size categories and the contribution of genuine copy number changes and underlying sequence variants were assessed for deletions and duplications. [a] The unconfirmed duplication in this category may be due to altered hybridization caused by an adjacent and confirmed deletion. This table does not include data from samples which failed analysis quality control, such as those listed in Supplementary Table 5.

Table S8: Relationship between common SNP genotype and probe log2 ratio

dbSNP ID	Sequence variant	Number of individuals		Mean log2 ratio		Mean log2 difference	T	Bonferroni adjusted p-value
		Reference allele	Other allele	Reference allele	Other allele			
rs3829990	PIR_c.681G>A p.Q227Q	129	14	-0.116	-0.031	-0.085	-3.57	0.027 *
rs2071308	MAGEB3_c.320G>A p.R107H	130	35	0.03	-0.041	0.071	2.76	0.117
rs4898	TIMP1_c.372T>C p.F124F	117	41	-0.048	-0.056	0.008	0.40	1
rs2073162	TNMD_c.306G>A p.V102V	123	50	-0.02	-0.09	0.07	3.44	0.015*
rs3813933	LONRF3_c.42T>C p.A14A	70	6	-0.014	-0.053	0.039	0.99	1
rs5956583	BIRC4_c.1268A>C p.Q423P	114	45	0.034	0.039	-0.005	-0.28	1
rs4830219	IGSF1_c.2556T>C p.Y852Y	123	47	0.107	0.303	-0.196	-5.93	$1.27e^{-06}$**
rs1129093	ZNF75_c.1434G>A p.T478T	104	69	-0.056	0.041	-0.097	-5.69	1.086^{-06}**
rs5930931	GPR112_c.1103C>Ap.P368H	65	33	-0.013	0.155	-0.168	-5.89	$2.102e^{-06}$**
rs1329546	GPR112_c.7941C>Ap.T2647T	95	66	0.069	0.341	-0.272	-10.73	$<3.3e^{-15}$**
rs5930942	GPR112_c.9117G>Ap.T3039T	75	97	0.023	0.094	-0.071	-3.59	0.075*
rs1190736	GPR101_c.370G>T p.V124L	69	78	-0.093	0.025	-0.118	-5.39	$2.80e^{-07}$**
rs764631	FMR1NB_c.425C>Tp.A142V	156	65	-0.01	-0.052	0.042	2.05	0.626
rs4833	BGN_c.141G>Ap.S47S	86	68	-0.077	-0.011	-0.066	-2.42	0.251
rs2269415	ATP2B3_c.2592G>Cp.V864V	86	87	-0.21	-0.132	-0.078	-2.76	0.097

X chromosome coding SNPS with a minor allele frequency >0.35 were identified using HapMart. Analysis was restricted to SNPs targeted by 5 or more probes and where we had previously obtained genotype information by exon re-sequencing. The mean log2 ratios of the major and minor allele groups were compared using a two-tailed t-test. Bonferroni adjustment was applied to correct for multiple testing. ** significant at $p<0.01$ levels; * significant at $p<0.05$

Table S9: CNV variants identified in the IGOLD cohort that have been reported previously to be predisposing alleles for ID.

Gene(s)	ChrX location (Mb)	Genomic Extent	Type	Number of occurrences	Present in DGV?	Present in Decipher?	Frequency in controls	Published reports
TSPAN7	38.37-38.52	145kb	Dup	2; including family 359 (*HUWE1* duplication)	Single report [8]	One report (male)	nd	[9] ID, but found in conjunction with other pathogenic variants and/or not segregating fully with disease. [10] ASD, but no impact on expression in [11] [12] ASD cases but suggest neutral polymorphism [13] SZ with normal IQ
HDH1A, STS, VCX, VCX2, VCX3B	6.46-8.10	1.6Mb	Dup	1	Yes, although unique duplication boundaries	Reported in males and females	0/130	[14] suggest causative in male with severe MR (maternally inherited)
ASMT	1.68-1.71	23kb	Dup	7 (plus 2 larger duplications)	Yes	No	nd	[12; 15]ID: suggest association, but acknowledge PAR regions are poorly represented on array platforms
ZNF630, SSX6, SPACA5	47.75-47.88	138kb	Del	1	Yes	No	2/152	ID no significant association
			Dup	3	Yes	No	5/152	

Table S10: Characteristics of sequenced breakpoints

Family	Gene(s) within CNV	CNV Type	START[a]	END[a]	ADM-1 Estimated Start	ADM-1 Estimated End	Breakpoint microhomology	Repetitive Element overlap[b]	
								Breakpoint A	Breakpoint B
121	*Il1RAPL1*	Del	28,922,932	29,253,959	28,923,176	29,253,472	CC	-	***MSTB* (LTR)**
147	*MAOA* and *MAOB*	Del	43,426,228	43,666,586	43,426,417	43,665,638	AT	***AluY* (SINE)**	***L1P2* (LINE)**
398	*SLC16A2*	Del	73,666,964	73,669,304	73,666,845	73,668,835	ATTTCTTTT	*AluSg5* (SINE)	-
399	*SLC16A2*	Del	73,552,449	73,567,609	73,553,102	73,567,239	TT (+AGATT)	***Charlie18a* (DNA)**	*AluSz*(SINE) and ***L1P4* (LINE)**
506	*CUL4B*	Del	119,578,701	119,584,448	119,579,350	119,584,201	AA	-	***AluSp*(SINE)**
540	*PTCHD1*	Del	23,239,008	23,329,210	23,238,828	23,327,945	GTG	-	***AluYc*(SINE)**
126	*NLGN4X*	Del	6,027,992	6,037,317	6,028,255	6,036,586	None	***MER9a3* (LTR)**	***AluJr4* (SINE)**
206	*WDR13*	Del	48,345,024	48,348,048	48,345,615	48,348,023	C[c]	-	-
110	*AFF2*	Dup	147,547,319	147,757,141	147,548,998	147,757,224	AC	-	-
317	Several	Dup	119,698,636	125,699,533	119,698,440	125,699,839	ATA	***MLT2B3*(LTR)**	***L1MEc* (LINE)**
505	*POLA1* and *ARX*	Dup	24,902,835	24,943,900	24,902,734	24,940,400	T	-	***L1MD* (LINE)**
463	*GSPT2*	Dup	51,469,871[d]	51,509,041	51,470,000	51,509,654	GAA	(TG)n simple repeat	***L1MD* (LINE)**
494	*FAAH2*	Del	not determined		57,487,858	57,493,349			
115	*SLC9A6*	Del	not determined		134,934,236	134,943,268			
32	*Il1RAPL1*	Del	not determined		28,939,863	29,497,216			
376	Several	Dup	not determined		18,985,933	22,751,175			

For each breakpoint, two breakpoint regions (A and B) were defined as 150bp sequences centred at the breakpoint from proximal and distal reference sequence matches (identified by BLAT searching). [a] Co-ordinates exclude regions of microhomology: start corresponds to the last reference base before any microhomology and end to the first base after any microhomology; [b] Repetitive elements in bold type overlie the rearrangement breakpoint whereas those in regular type are contained within the 150bp breakpoint region but do not extend to the breakpoint itself; [c] The breakpoint in family 206 also contains a 12bp insertion presumed to result from serial replication slippage. [d] The telomeric boundary falls within a region of segmental duplication, and sequence analysis cannot distinguish the two possible reference matches. The breakpoint position reported here corresponds to the duplicon position closest to the array estimate of duplication extent

Table S11: Comparison of genomic features in breakpoint regions and a simulated control dataset

	Breakpoint Regions (n=24)	Simulated Dataset (n=500)	p-value
LCR overlap	1	30	1
Repetitive element overlap			
none	9	184	1
LTR	3	64	1
LINE	4	160	0.174
SINE	5	48	0.204
DNA element	1	10	0.406
Other	1	15	0.533
Multiple elements within 150bp window	1	19	0.615
Non-B DNA structures			
G-quartet	12	130	0.017
Z-DNA	2	3	0.019

Features of the proximal and distal 150bp breakpoint reference sequences for each of the 12 sequenced rearrangements were compared to a control set of 500 sequences, each 150bp in length, generated by random sampling of the X chromosome, following exclusion of sequence gaps and centromeric regions. LCRs were identified using the UCSC genome browser (for LCRs: genomicSuperDups track). Repetitive Elements were identified using RepeatMasker. G-quartet and Z-DNA were detected using QGRS mapper and Z-hunt online respectively. Statistical significance assessed using chi-square contingency test.

References in Supplementary Materials

1. Froyen G, Corbett M, Vandewalle J, Jarvela I, Lawrence O, Meldrum C, Bauters M, Govaerts K, Vandeleur L, Van Esch H, et al. (2008) Submicroscopic duplications of the hydroxysteroid dehydrogenase HSD17B10 and the E3 ubiquitin ligase HUWE1 are associated with mental retardation. Am J Hum Genet 82:432-443
2. Friez MJ, Jones JR, Clarkson K, Lubs H, Abuelo D, Bier JA, Pai S, Simensen R, Williams C, Giampietro PF, et al. (2006) Recurrent infections, hypotonia, and mental retardation caused by duplication of MECP2 and adjacent region in Xq28. Pediatrics 118:e1687-1695
3. Bauters M, Van Esch H, Friez MJ, Boespflug-Tanguy O, Zenker M, Vianna-Morgante AM, Rosenberg C, Ignatius J, Raynaud M, Hollanders K, et al. (2008) Nonrecurrent MECP2 duplications mediated by genomic architecture-driven DNA breaks and break-induced replication repair. Genome Res 18:847-858
4. Clayton-Smith J, Walters S, Hobson E, Burkitt-Wright E, Smith R, Toutain A, Amiel J, Lyonnet S, Mansour S, Fitzpatrick D, et al. (2009) Xq28 duplication presenting with intestinal and bladder dysfunction and a distinctive facial appearance. Eur J Hum Genet 17:434-443
5. Thode A, Partington MW, Yip MY, Chapman C, Richardson VF, Turner G (1988) A new syndrome with mental retardation, short stature and an Xq duplication. Am J Med Genet 30:239-250
6. Bejerano G, Pheasant M, Makunin I, Stephen S, Kent WJ, Mattick JS, Haussler D (2004) Ultraconserved elements in the human genome. Science 304:1321-1325
7. Siepel A, Bejerano G, Pedersen JS, Hinrichs AS, Hou M, Rosenbloom K, Clawson H, Spieth J, Hillier LW, Richards S, et al. (2005) Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. Genome Res 15:1034-1050
8. Zogopoulos G, Ha KC, Naqib F, Moore S, Kim H, Montpetit A, Robidoux F, Laflamme P, Cotterchio M, Greenwood C, et al. (2007) Germ-line DNA copy number variation frequencies in a large North American population. Hum Genet 122:345-353
9. Froyen G, Van Esch H, Bauters M, Hollanders K, Frints SG, Vermeesch JR, Devriendt K, Fryns JP, Marynen P (2007) Detection of genomic copy number changes in patients with idiopathic mental retardation by high-resolution X-array-CGH: important role for increased gene dosage of XLMR genes. Hum Mutat 28:1034-1042
10. Marshall CR, Noor A, Vincent JB, Lionel AC, Feuk L, Skaug J, Shago M, Moessner R, Pinto D, Ren Y, et al. (2008) Structural variation of chromosomes in autism spectrum disorder. Am J Hum Genet 82:477-488
11. Noor A, Gianakopoulos PJ, Fernandez B, Marshall CR, Szatmari P, Roberts W, Scherer SW, Vincent JB (2009) Copy number variation analysis and sequencing of the X-linked mental retardation gene TSPAN7/TM4SF2 in patients with autism spectrum disorder. Psychiatr Genet 19:154-155
12. Cai G, Edelmann L, Goldsmith JE, Cohen N, Nakamine A, Reichert JG, Hoffman EJ, Zurawiecki DM, Silverman JM, Hollander E, et al. (2008) Multiplex ligation-dependent probe amplification for genetic screening in autism spectrum disorders: efficient identification of known microduplications and identification of a novel microduplication in ASMT. BMC Med Genomics 1:50
13. Guilmatre A, Dubourg C, Mosca AL, Legallic S, Goldenberg A, Drouin-Garraud V, Layet V, Rosier A, Briault S, Bonnet-Brilhault F, et al. (2009) Recurrent rearrangements in synaptic and neurodevelopmental genes and shared biologic pathways in schizophrenia, autism, and mental retardation. Arch Gen Psychiatry 66:947-956
14. Wagenstaller J, Spranger S, Lorenz-Depiereux B, Kazmierczak B, Nathrath M, Wahl D, Heye B, Glaser D, Liebscher V, Meitinger T, et al. (2007) Copy-number variations measured by single-nucleotide-polymorphism oligonucleotide arrays in patients with mental retardation. Am J Hum Genet 81:768-779

15. Lugtenberg D, Zangrande-Vieira L, Kirchhoff M, Whibley AC, Oudakker AR, Kjaergaard S, Vianna-Morgante AM, Ruiter TKM, Jehee FS, Ullmann R, et al. (2010) Recurrent deletion of ZNF630 at Xp11.23 is not associated with mental retardation. Am J Med Genet A In Press