Rare allelic forms of PRDM9 associated with childhood leukemogenesis

Julie Hussin[1,2], Daniel Sinnett[2,3], Ferran Casals[2], Youssef Idaghdour[2], Vanessa Bruat[2], Virginie Saillour[2], Jasmine Healy[2], Jean-Christophe Grenier[2], Thibault de Malliard[2], Stephan Busche[4], Jean-François Spinella[2], Mathieu Larivière[2], Greg Gibson[5], Anna Andersson[6], Linda Holmfeldt[6], Jing Ma[6], Lei Wei[6], Jinghui Zhang[7], Gregor Andelfinger[2,3], James R. Downing[6], Charles G. Mullighan[6], Philip Awadalla[2,3]*

[1]Departement of Biochemistry, Faculty of Medicine, University of Montreal, Canada [2]Ste-Justine Hospital Research Centre, Montreal, Canada [3]Department of Pediatrics, Faculty of Medicine, University of Montreal, Canada [4]Department of Human Genetics, McGill University, Montreal, Canada [5]Center for Integrative Genomics, School of Biology, Georgia Institute of Technology, Atlanta, Georgia, USA [6]Department of Pathology, St. Jude Children's Research Hospital, Memphis, Tennessee, USA [7]Department of Computational Biology and Bioinformatics, St. Jude Children's Research Hospital, Memphis, Tennessee, USA.
*corresponding author : philip.awadalla@umontreal.ca

SUPPLEMENTARY INFORMATION

SUPPLEMENTARY METHODS

SNPs in the ALL Quartet

In order to distinguish germline events from somatic events, each affected child from the ALL quartet was sampled twice, pre- and post-treatment. The pre-treatment samples had less than 20% tumour cells making it difficult to identify tumor-specific mutations in these ALL cases.

Exome SNP Dataset

Exome sequencing was performed as described in Material and Methods. Basic statistics are presented in Table S1A. To have the most accurate set of SNPs possible, we used two approaches that make use of inheritance patterns to call SNPs in the six samples:

(i) Samtools dataset: Mismatches according to the hg18 human reference genome were called using Samtools and were considered as SNPs when the quality score (Phred score) is $\geq$ 20 and when the position is covered in all samples. Extra SNPs with a quality score below 20 in the offspring (pre and post-treatment) were rescued when the position is covered in each sample and the variant allele is confirmed in 30% of the reads in one parent.

(ii) Polymutt dataset: The program Polymutt implements a likelihood-based method for calling SNPs in trio data [1]. Each of the four trios was processed independently and produced a list of genotype calls for SNPs in each individual and a mean quality score for the trio as a whole. We selected the SNPs having a mean quality score $\geq$ 20 and covered in all samples. We removed 1343 SNPs for which the genotype calls were inconsistent for the parents between the different trio analyses.

A total of 30,360 and 47,243 SNPs were inferred by the Samtools and Polymutt approaches, respectively. A total of 28,747 SNPs, called by both methods, were retained for further analysis. Genotypes called by Polymutt at these positions were selected for the six samples, because the genotype mendelian error rate for the Samtools dataset (9115 errors, mendelian error rate of 7.93%) was higher than for the Polymutt dataset (511 errors, mendelian error rate of 0.44%). Finally, we removed 4799 SNPs, homozygous with non-reference alleles in all six samples, and obtained a final exome SNP dataset of 23,437 polymorphic positions.

Genotyping SNP Datasets

The ALL quartet samples were genotyped on two different platforms: the Illumina HumanOmni 2.5-quad BeadChips and the Affymetrix 6.0 SNP arrays.

For the Illumina Omni 2.5 array, 2,383,178 SNPs passed standard quality control filters. The parents were genotyped twice, with a concordance rate of 99,987%. We removed 617 discordant SNPs between duplicated parents, 1339 SNPs with Mendelian errors, 21,532 SNPs with missing data and 1,552,225 SNPs homozygous in all samples. We obtained a final Illumina SNP dataset of 808,082 polymorphic markers.

For the Affymetrix 6.0 array, we obtained 909,515 SNPs after applying standard quality control filters. We subsequently removed 2447 duplicated SNPs, 4403 SNPs with Mendelian errors and 110,423 SNPs with missing data. The final Affymetrix SNP dataset contained a total of 792,242 markers.

Recombination SNP Dataset

To obtain a complete recombination SNP dataset for the ALL quartet, that allows a finer detection of recombination events in coding regions, we merged the exome SNP dataset and the Illumina SNP dataset. Among Illumina Omni2.5 SNPs that passed standard quality control filters, 16,839 SNPs are positioned in Agilent SureSelect targeted regions used in the exome sequencing experiment. Among these, we removed 161 positions detected as polymorphic in the exome sequencing data but homozygous for all samples in the genotyping data. For 3347 polymorphic Illumina SNPs that were not detected in the exome sequencing data, most of which went undetected because the positions were not covered in all samples in the exome data, genotype calls from the Illumina SNP dataset were kept. For the remaining 13,331 coincident positions, the genotype concordance rate was 0.976. We removed 3 SNPs that had different alleles in the two datasets and 1309 SNPs that had the same alleles but at least one sample with discordant genotypes. Combining the concordant SNPs with the remaining SNPs from both datasets, we obtained a final recombination SNP dataset of 816,715 for the ALL quartet (Table S1B).

De novo mutation discovery in the ALL quartet

DND Software [2,3]

To discover *de novo* point mutation in sequencing data, we used a probabilistic approach DND developed by our group [2,3]. The method uses the relatedness between individuals in a pedigree to produce the posterior probability of *de novo* mutation at each genomic site, given the error rate of the sequencing technology, the somatic mutation rate and the population mutation rate for the population from which the samples were drawn.

DND was run on four trios: [M, F, 383R1]; [M, F, 383R2]; [M, F, 610R1]; [M, F, 610R2], where M is the mother, F is the father, R1 denotes pre-treatment samples and R2 denotes post-treatment samples for the two brothers, patients 383 and 610. We used a sequencing error rate of $\varepsilon = 0.005$, a somatic mutation rate of $\mu = 2\times10^{-7}$ and a population mutation rate of $\theta = 0.001$, and examined all sites within targeted regions covered by at least 15 reads in samples from each trios. Simulations revealed that highest confident calls require a read depth of at least 15 in both parents and the non-mutant allele [2].

To infer a germline *de novo* mutation in a child, we considered all positions that had a probability of *de novo* > 0.90 in at least one of the two samples (R1 or R2), with the same alternative allele present in at least one read in the other sample. We found 27 such candidates in 383 and 15 in 610. We removed candidates for which at least two reads in the parents showed the alternative allele: previous validation experiments [3] showed that such *de novo* candidates are very likely to be false positives. Similarly, we removed candidates where at least two reads with the same alternative allele was seen in the sibling. This filter was applied if the children shared at least one parental chromosome at that locus (Figure S10). Although these candidate de novo mutations could reflect mutations that happened in premeiotic divisions in germinal stem cells, it is more likely that the allele was not sampled in the sequencing experiment for that parent or resulted from similar sequencing or mapping errors in both children. Unfortunately, although 7 passed the parental filter, none of the *de novo* candidates passed the sibling filter.

Sample-independant approach

Using the DND software with stringent parameters, we only interrogated regions covered at more than 15X in the parents, which in this case corresponds to ~12Mb of the human exome. To investigate other positions targeted by the sequencing experiments, we used SNPs from the *samtools dataset* called via a sample-independent strategy [3]. We selected positions with at least 8X coverage in the parents (25,286,190 bp) and considered all Mendelian errors. To be called a *de novo* candidate, the variant allele must be seen at least 5 times in one of the child samples, at least twice in the other sample from the same sibling and must not be seen at all in the parental samples or in the other sibling. We identified one such mutation in patient 383 (Figure S1), located on chromosome 5 at position 64 860 544 (hg18). At this locus, the two brothers copied the same paternal chromosome and different maternal chromosomes (Figure S10). If we assumed a binomial distribution with a probability of 0.5 of sequencing the variant allele at a heterozygous position, the probability that this variant was transmitted from the father and that the variant was not sampled in neither the father nor patient 610 (total coverage of 27X) is $p = 7.45\times10^{-9}$. On the other hand, the probability that this variant was transmitted from the mother to 383 (coverage of 10X) is $p = 9.77\times10^{-3}$.

SUPPLEMENTARY RESULTS

Patients' Karyotype

The brothers from the ALL quartet family were both diagnosed, within a 3 years period of time, with B-cell precursor childhood ALL with FAB-L1 morphology at Sainte-Justine Hospital, Montreal, Canada, at the age of 2 for patient 383 and subsequently, at 14 years of age for patient 610. At diagnosis, both siblings showed hyperdiploid leukemia clones. Cytogenetic analyses were performed using standard procedures. G-banded chromosomal analysis for patient 610 revealed clones with the following karyotypic features: 55XY,+X,?del(2q),+5,+8,+10,+14,+17, +18,+21,+21[5]/46XY[21]. By employing the Illumina Omni 2.5 genotyping data, we also detected the following additional chromosomes in patient 610: +4,+?Y. For patient 383, clones were detected with karyotype: 51-53,XY,inv(2)(p11q13),i(17)(q10),+4,+6,+?12,+15,+17,+18, +21[9]/46,XY,inv(2)(p11q13)[23]. The additional chromosomes 17, 18 and 21, shared between siblings' leukemic clones, are frequently gained in ALL hyperdiploidy. Chromosome 2 pericentric inversion in patient 383 is also carried by the mother and occurs at a higher frequency in African Americans compared to individuals of European descent [4]. This aberration is not associated with a specific syndrome and no abnormal phenotype has been described [5]. However, such inversions may lead to recombinants gametes with abnormal karyotes, through crossing over between the normal and inverted homologues.

Fine-scale Dissection of Recombination Events

Recombination analyses were performed separately for the pre-treatment and post-treatment samples, forming two quartets that will be referred to herein as quartet Q1 (pre-treatment) and quartet Q2 (post-treatment). The algorithm used to call recombination events [6] first looks for errors, i.e markers that create double recombinants (Material and Methods). This procedure identified 1214 errors (433 in Q1, 456 in Q2 and 325 in Q1 and Q2). A total of 222 errors came from SNPs from the Illumina SNP dataset and 992 came from the exome SNP dataset. The highest density of sequencing errors detected by the recombination algorithm is located in chromosome 6, between 29 and 34 Mb, in the the complex region of the human leukocyte antigen (HLA) system, very likely caused by mapping errors, given that assembly of this locus using single-end short reads data is difficult.

The algorithm identified a total of 236 switches in the quartet Q1 and 224 in the quartet Q2. Switches not shared between quartets Q1 and Q2 were separated from their closest neighbouring event by at most 5 markers and are very likely to be caused by calling or alignment errors. All shared switches are presented in Figure S10. From this list, switches were called as crossovers only if they were separated from their closest neighbour by more than 2 informative markers. They were divided into two categories: single crossovers, when the nearest crossover is at more than 50Kb and double crossovers, when two crossovers were found within 50Kb of each other (Table S2, Figure 1).

Putative Gene Conversion Events

A region on chromosome 16 was intriguing because both parents showed double crossovers at exactly the same locus (separated by 4 and 8 markers in the father and mother, respectively). Genotyping and sequencing markers both supported the double crossovers. Although these double crossovers could reflect gene conversion events, it is unlikely that such events happened in both parents in the same region. A most likely explanation is that this region is not unique in the genome and that the detected genotypes tag polymorphisms positioned somewhere else in the genome. This is probably the case since these double recombinants occur in gene HYDIN, a gene that has been duplicated very recently, with a nearly identical 360-kb paralogous segment inserted on chromosome 1q21.1 [7]. We therefore removed these double crossovers from our final list of paternal and maternal recombination events.

We identified nine other short regions (<50Kb) flanked by recombination events (Table S2) having no known paralogous segment. All regions comprise genotyping SNPs, they are not resulting from errors in mapping of sequencing reads. These double crossovers could reflect meiotic gene conversions in the same meiosis or recombination events occuring in the same genomic region in the two brothers, since we can not distinguish to which child each event belongs. However, four maternal double crossovers on chromosome 4, 8 and 17 (as well as four additional double crossovers supported by only 2 markers on chromosomes 7, 8, 10 and 18 – see Figure S10) occurred within 1 Mb of another clearly defined recombination event. This would necessarily mean that, in one of the two children, at least two recombination events happened closeby. Under the model of crossover interference, however, the presence of one crossover event in a region reduces significantly the possibility of a second event nearby in the same individual. Therefore, these small double crossovers may reflect unique patterns of recombination or gene conversion, not yet identified in humans.

PRDM9 alleles in the ALL Quartet

Because of the reduced proportion of recombination events overlapping with population hotspots observed in the mother of the ALL quartet, we investigated whether variants in the PRDM9 coding region were identifiable based on the read data. We detected two SNPs in the ZnF array domain of exon 11 of PRDM9, both present in dbSNP v134 (rs74710141 and rs77287813). SNP rs74710141 corresponds to the known C/G substitution in the sixth ZnF repeat, the only difference between PRDM9 allele A and B. It appeared that the hg18 reference allele is the B allele. SNP rs77287813 corresponds to a A/C substitution in the tenth finger, corresponding to the only difference existing between ZnF type *h* and *k*. Due to the structure of the PRDM9 ZnF array, the presence of this SNP is likely to reflect a supplementary repeat instead of a point mutation in the H ZnF repeat.

To infer which *PRDM9* ZnF allele was carried by the mother, we identified the ZnF types present in the read data for the mother (Material and Methods). Out of 1477 reads that mapped to the PRDM9 ZnF array, 26 (1.76%) aligned specifically to the *k* ZnF type and 8 reads aligned specifically to the *l* ZnF type, which corresponds to 0.54% of the read data mapped in the array, a value below our inference criteria of 1% (Material and Methods). However, the *l* ZnF type, which is one mismatch away from the *e* ZnF type, is two mismatches away from its closest match in the reference: the ZnF type *c* of the *b* allele. This is expected to hamper the mapping of reads sampling the *l* ZnF repeat (for example in cases where there is an error at the end of the read). The two brothers did not copy the same maternal chromosome in this region of chromosome 5. Child 610 had 4417 reads mapping to the PRDM9 ZnF array and very few (a total of 6) aligning to either the *k* or *l* ZnF type. On the other hand, child 383 had 59/2625 reads (2.25%) and 18/2625 reads (0.69%) aligning to the *k* and *l* ZnF type, respectively. The unusual patterns of recombination in the mother and the occurrence of 85 and 26 reads sampling the *k* and *l* ZnF type on one maternal chromosome suggested the presence of the C allele in the mother.

Because SOLiD single-end short reads (50 bp) will not overlap a full ZnF repeat, which is 84 bp long, our data does not allow to determine the order of repeats or insertions in the ZnF array. The A/C genotype in the mother was thus validated by Sanger re-sequencing of the ZnF array (Material and Methods).

Description of *PRDM9* Alleles and Novel ZnF Types

Labeling of the PRDM9 zinc finger (ZnF) alleles and repeat types follows that of Berg and colleagues[8], but to differentiate "allele" from "finger" nomenclature, alleles are in uppercase and fingers are written in lowercase italic characters (for example, allele A has 13 repeats type and is coded: *abcddecfghfij*). The ZnF repeat types *a* to *x* are presented in Figure S6.

In the 22 parental trios of the FCALL cohort, we detected 11 parents for which read data show evidence for the *k* and/or *l* fingers. For 2 additional parents, the *p* and *t* fingers were detected (data not shown), suggesting the presence of ZnF allele L24. We performed Sanger sequencing of the ZnF array for 12 families, which included 9 of the 11 parents with *k, l* or *p* and *t* fingers. Sanger sequencing experiments confirmed these 9 *PRDM9* alternative alleles and revealed the presence of undetected rare alleles in other families: the allele L3 and two alleles not reported in previous studies, L37 and L38 (Table S3). Repeat structure for L37 is *abcddecfghfqj* and for L38, *abcddecughfqj*, with repeat types described in Figure S6. Sanger sequencing of PRDM9 ZnF array was also performed for 76 French-Canadian parents from the FC family cohort and 27 Moroccans (Table S5, Fig S4). We discovered 5 novel alleles in this data denoted L32 to L36. The newly discovered allele L37 was seen three times in Moroccans. The repeat structure for the novel alleles are: L32=*abcvdecfghfij*, L33=*abcdddecfghfij*, L34=*abcddecfghfwj*, L35=*abcddecxghfij* and L36=*abrddecfghfij*, with repeat types described in Figure S6.

In total, we sequenced 258 *PRDM9* ZnF alleles and identified four novel ZnF types: *u*, *v*, *w* and *x* (Figure S6):

- The *u* repeat type is a mutated *f* repeat type, encoding a missense change at a well-conserved position of the repeat sequence: GAG (E) → AAG (K). Only one synonymous mutation was previously observed at this codon (repeat type O, GAG (E) → GAC (E)). This change is not located within the binding unit of the ZnF repeat but it is predicted by SIFT[9] to have a damaging effect on the resulting ZnF array (SIFT score = 0). This repeat type was found in the PRDM9 array from one ALL parent.

- The *v* repeat type is a mutated *d* repeat type, encoding a missense change, TAT (Y) → TTT (F), at a position where only synonymous changes were previously observed (TAT (Y) → TAC (Y) in repeat types *a*, *i*, *j* and *m*). This change is not located within the binding unit of the ZnF repeat and is predicted as tolerated (SIFT score = 0.06). This repeat type was found in the PRDM9 array of two individual of the FC cohort.
- The *w* repeat type is a mutated *i* repeat type, encoding a missense change, CGC (R) → CTC (L), at a position not particularly conserved between repeat types. This change is not near the binding unit of the ZnF repeat and is predicted as tolerated (SIFT score = 0.14). This repeat type was found in the PRDM9 array of one individual of the FC cohort.
- The *x* repeat type is a mutated *f* repeat type, encoding a synonymous change (AAG (K) → AAA (K)). This repeat type was found in the PRDM9 array of one individual in the Moroccan cohort.

Association Testing using Exome Sequencing Control Cohorts

Because we used exome-sequencing data to detect the excess of rare alleles in both the FCALL and the SJALL cohort, we used two sets of ethnically matched exome-sequenced controls to assess the frequencies of PRDM9 alleles, prior to validation with Sanger sequencing.

The FCALL cohort was compared to the FCEXOME cohort, consisting of 68 healthy parents from three disease cohorts (primary immunodeficiencies, schizophrenia and autistic spectrum disorder) recruited at the Sainte-Justine University Hospital (Montreal, Canada)[10]. Exome capture was performed with the *SureSelect Target Enrichment System* from Agilent Technologies optimized for Applied Biosystems' SOLiD sequencing. Exonic sequences were obtained using the SOLiD 3 Plus System and the SOLiD 4 System (Applied Biosystems) technology. Sequence reads were aligned to the NCBI Build 36 reference sequence with BioScope v1.2. PRDM9 alleles were typed from exome sequencing read data as described in Material and Methods. Only 6 individuals showed the presence of *k*-finger alleles. This is significantly lower than in the FCALL cohort, where 12 parents out of 46 showed evidence for *k*-finger alleles in the read data ($p = 0.0182$, two-tailed Fisher's Exact Test).

The SJALL cohort was compared to the exome sequencing data from the CEU population sequenced in the 1000 Genomes Project [11]. Read data aligning to PRDM9 ZnF array was retreived for 99 CEU individuals available on the 1000 Genomes ftp server on February 9th 2012 from the following address: ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/phase1/technical/other_exome_alignment. These alignements were performed at the Broad Institute using the BWA aligner to map the reads to the NCBI Build 36 reference sequence with Illumina data. This data is different from the read data found in the 'exome_alignment' directory, where the bam files were not informative for PRDM9 alleles because of the alignement parameters used (only uniquely mapped reads were kept). Out of the 99 individuals available, 8 individuals showed strong evidence for *k*-finger alleles in the read data : NA11843, NA11894, NA12272, NA12275, NA12287, NA12760, NA12830 and NA12842. This is significantly lower than in the replication SJALL cohort, where 10 out of 50 children showed evidence for *k*-finger alleles in the read data ($p = 0.0353$, one-tailed Fisher's Exact Test).

Association with SNPs Tagging Rare PRDM9 Alleles

Hinch and colleagues [12] identified SNP rs6889665, located within 4 Kb of the *PRDM9* Znf array, for which the ancestral T allele (frequency of 98% in CEU from HapMap) is strongly correlated to *PRDM9* alleles binding to the 13-bp motif (including the common *PRDM9* A and B alleles), whereas the derived C allele at rs6889665 is almost perfectly correlated to alleles predicted to bind the 17-bp motif, including the *PRDM9* C allele. Given the excess of C-like alleles in B-ALL patients, the association between *PRDM9* variants and B-ALL could be detectable using this SNP. Unfortunately, this SNP is not present on Affymetrix arrays, however, its closest SNP on the Affymetrix 6.0 array, rs12153202, is in (imperfect) linkage disequilibrium with rs6889665 in CEU from HapMap3. We thus evaluated the association between B-ALL and rs12153202, using CEU from HapMap3 as controls. Considering all 50 patients, the association is significant overall (p = 0.00265), and for the hypodiploid and infant subgroups (p = 0.01103 and p = 0.00118, respectively). These results are in line with the association results between B-ALL and the *k*-finger *PRDM9* alleles (Table 1), suggesting that SNPs rs12153202 and rs6889665 might be useful tag SNPs to detect the *PRDM9* association in cohorts where genetic data for the *PRDM9* ZnF array is not available, with an increased power if parents (and not only patients) are considered, given the high frequency of C-like alleles in parents of B-ALL patients.

Ancestry Analyses

ALL Quartet

The parents of the ALL quartet are reported to be distant cousins of Moroccan ancestry. To verify their ethnicity and select the appropriate set of controls to use in our analyses, we performed principal components analyses (PCA) of the genetic variation using the smartpca and twstats modules from the Eigensoft package [13]. We first performed a PCA using genotyping data from the parents of the ALL quartet and 163 unrelated individuals from a Moroccan cohort [14]. We selected 27 Moroccan individuals among these with the closest eigenvalues to the 2 first significant PCs (Figure S3A), where DNA were available. We performed a second PCA including 28 unrelated French-Canadians individuals from the FC family cohort (Figure S3B). The results suggest that the parents of the ALL quartet have some Arab ancestry but are genetically closer to the French-Canadians than the Moroccans, although they also are outlier individuals relative to the French-Canadian cluster. Finally, the estimated genome-wide proportion of alleles identical-by-descent between the parents is 0.0267, which means that the parents are likely to be third cousins. This was computed using the --genome option in PLINK [15] to obtain estimates of pairwise IBD sharing.

FCALL cohort and FC family cohort

We performed a PCA of genetic variation to study genetic ancestry of the parents from the FCALL cohort (parents of patients). We include parents from the FC family cohort (controls) and European and African individuals from the HGDP dataset using positions of SNPs in common between exome sequencing variant found in the FCALL parents, genotyped SNPs in HGDP populations and the Affymetrix 6.0 array used to genotype individuals in the FC family cohort. The analysis demonstrate that French-Canadian parents of patients and controls cluster together with the French HGDP individuals, although the FCALL parents do not exactly overlap with the other two groups on the plot, likely because of the small differences in allele frequencies driven by the different technologies used to type genetic variation (exome SNP calling vs genotyping). In any case, the FCALL parents do not show a higher contribution of African genetic background than individuals from the FC family cohort. Therefore, differences in frequencies of PRDM9 ZnF allele C cannot be explained by a higher African ancestry in individuals from the FCALL cohort.

St Jude ALL cohort

The entire B-ALL St Jude cohort includs 61 B-ALL patients for which reported ethnicities were available. We verified these ethnicities by performing a PCA of the patients' genotyping variation genome-wide and removed from subsequent analyses children with an African genetic ancestry, for which we do not have controls. The PCA confirmed the reported ethnicities for most patients, however five individuals reported as "White" are potentially admixed (Figure S8). From the individuals that were reported as "Other", two individuals are likely mixed (black/white), one is likely Asian and one is likely Hispanic or Native American. Fifty patients showed no African component and were included in our analyses, with 39 children clustering with the French-Canadian controls. Therefore, association testing was performed with and without the 11 children with Hispanic, Asian or Native American ancestry.

SUPPLEMENTARY TABLES

Table S1. Coverage and SNPs statistics in the ALL quartet.

R1 and R2 are the two somatic tissues sampled from both brothers. The exome is defined by the regions targeted by *Agilent SureSelect All Exon kit* covering 37,806,033 bp (1,22% of the human genome).

A

	Bioscope assembly		**Exome statistic on coverage**				
Sample	**Total number of mappable reads**	**% of reads aligned**	**% of reads aligned on exome**	**% Exome coverage**	**Mean coverage**	**Average Base Quality**	**Average Mapping Quality**
383 R1	52 802 074	77,49	62,27%	94,63%	**45,96**	28,63	76,81
383 R2	57 835 258	82,11	62,96%	96,46%	**49,93**	29,31	78,61
610 R1	52 729 534	83,07	63,83%	96,99%	**45,89**	29,40	78,82
610 R2	48 659 235	76,98	60,40%	96,84%	**40,14**	28,88	75,05
M	65 453 411	85,72	57,90%	96,73%	**51,81**	28,94	79,36
F	59 254 618	81,74	61,21%	96,02%	**49,95**	28,95	76,46

B

Number of SNPs	**All samples**	**383 R1**		**383 R2**		**610 R1**		**610 R2**		**M**		**F**	
		Hom	**Het**	**Hom**	**Het**	**Hom**	**Het**	**Hom**	**Het**	**Hom**	**Het**	**Hom**	**Het**
Total	816 715	366 234	450 481	366 137	450 578	370 994	445 721	371 094	445 621	363 805	452 910	368 133	448 582
Exome Sequencing Only	9 945	4 038	5 907	3 929	6 016	3 841	6 104	3 949	5 996	3 962	5 983	4 204	5 741
Genotyping Only	794 751	356 905	437 846	356 917	437 834	362 002	432 749	361 994	432 757	354 363	440 388	358 297	436 454
Overlap	12 019	5 291	6 728	5 291	6 728	5 151	6 868	5 151	6 868	5 480	6 539	5 632	6 387

Table S2. Number of maternal and paternal recombination events per chromosome.

Quartet Q1 (parents + children's post-treatment samples) and Q2 (parents + children's pre-treatment samples) were analysed separately. Parameter k is the number of informative markers that separate any two consecutive recombination events. Crossovers shared between quartets are considered to be real recombination events, separated into two categories: single crossovers, if the nearest neighbour is within >50Kb, and double crossovers, if the nearest neighbour is within ≤50Kb (Supplementary Methods). Double crossovers found in chromosome 16 in both parents were likely to be artefacts resulting from the HYDIN duplicated gene and were ignored.

Chr	**Total number of markers**	**Maternal events**							**Paternal events**							
		Informative markers	**Quartet Q1**		**Quartet Q2**		**Shared k>2**		**Informative markers**	**Quartet Q1**		**Quartet Q2**		**Shared k>2**	
			k>1	k>2	k>1	k>2	Single	Double		k>1	k>2	k>1	k>2	Single	Double
1	61451	22517	6	6	8	6	**6**	**0**	20368	4	4	10	6	**4**	**0**
2	66882	23826	11	11	13	11	**11**	**0**	22667	6	4	4	4	**4**	**0**
3	55477	19146	7	7	9	7	**7**	**0**	18703	5	3	3	3	**3**	**0**
4	52750	18497	14	14	14	14	**8**	**2**	18151	3	3	3	3	**3**	**0**
5	50401	17975	5	3	5	3	**3**	**0**	16910	1	1	1	1	**1**	**0**
6	51056	18760	10	6	8	4	**4**	**0**	17441	7	3	3	3	**3**	**0**
7	45728	16101	9	7	9	7	**7**	**0**	15891	5	3	5	3	**1**	**1**
8	43517	14836	9	7	9	7	**5**	**1**	15993	7	5	9	5	**3**	**1**
9	37351	12838	8	6	8	6	**6**	**0**	13168	3	1	1	1	**1**	**0**
10	43005	14528	11	7	11	7	**7**	**0**	15546	1	1	3	1	**1**	**0**
11	40778	13952	3	3	6	3	**3**	**0**	13738	2	2	4	2	**2**	**0**
12	39814	13785	10	10	12	10	**8**	**1**	13674	6	2	4	2	**2**	**0**
13	30103	9605	4	4	4	4	**2**	**1**	11324	3	3	3	3	**3**	**0**
14	28292	10643	3	3	3	3	**1**	**1**	9107	3	1	1	1	**1**	**0**
15	26049	8928	2	2	2	2	**2**	**0**	8984	2	2	2	2	**2**	**0**
16	28506	9827	10	8	8	8	**6**	**(1)**	9602	6	4	8	4	**2**	**(1)**
17	24570	8149	11	7	9	7	**5**	**1**	8804	4	2	10	2	**2**	**0**
18	23846	8080	6	4	6	4	**4**	**0**	8173	2	2	2	2	**2**	**0**
19	19424	6459	1	1	1	1	**1**	**0**	6665	2	2	4	2	**2**	**0**
20	21582	7930	3	3	3	3	**3**	**0**	7241	2	2	2	2	**2**	**0**
21	12137	3917	3	2	2	2	**2**	**0**	4379	2	2	2	2	**2**	**0**
22	13970	5164	1	1	1	1	**1**	**0**	4632	1	1	1	1	**1**	**0**
Total	816689	285463	147	122	151	120	**102**	**7**	281161	77	53	85	55	**47**	**2**

Table S3. *PRDM9* alleles in the ALL quartet and 12 ALL trios based on read data and re-sequencing.

For fathers (F), mothers (M) and patients (N) in each family, the ZnF repeat types from *PRDM9* alleles were first inferred from SOLiD sequencing read data. Repeat types (*a* to *l*) are described in Figure S6. Repeat types with a proportion above 0.01 (highlighted) are inferred to be present in the individuals. Sanger sequencing was subsequently performed in the parents and genotypes with rare alleles are highlighted. All *k* and *l* fingers inferred were validated (alleles C, D, L20). *(next page)*

Table S3 *(continued)*

Family	Individual	Coverage	Proportion of Exome Sequencing Reads Aligning to PRDM9 ZnF Repeats												Validation[a]
			a	*b*	*c*	*d*	*e*	*f*	*g*	*h*	*i*	*j*	*k*	*l*	
Quartet	F	67X	0.0103	0.0304	0.0556	0.0950	0.0274	0.0697	0.0215	0.0438	0.0222	0.0244	0.0014	0.0007	A/A
	M	73X	0.0121	0.0345	0.0697	0.0941	0.0128	0.0724	0.0270	0.0433	0.0291	0.0264	0.0176	0.0054	A/C
	610N	66X	0.0273	0.0287	0.0378	0.0570	0.0178	0.0516	0.0255	0.0278	0.0208	0.0144	0.0011	0.0002	-
	383N	71X	0.0502	0.0499	0.0632	0.096	0.0121	0.0780	0.0300	0.0491	0.0331	0.0236	0.0224	0.0069	-
375	F	11X	0.025	0.0607	0.0142	0.0392	0.0071	0.0821	0.0357	0.025	0.0107	0.0107	0	0	A/A
	M	14X	0.0088	0.0616	0.0352	0.0572	0.0088	0.0616	0.0176	0.0264	0.0220	0.0044	0	0	B/L24
	375N	74X	0.0472	0.0340	0.0546	0.0874	0.0185	0.0941	0.0340	0.0394	0.0374	0.0148	0.0057	0.0010	
380	F	17X	0.0138	0.0635	0.0220	0.0580	0.0165	0.0165	0.0083	0.0110	0.0110	0.0027	0	0.0055	A/A
	M	18X	0.0401	0.0401	0.0321	0.0455	0.0053	0.0642	0.0080	0.0214	0.0160	0.0160	0.0053	0	A/B
	380N	59X	0.0440	0.0343	0.0478	0.0994	0.0292	0.0816	0.0364	0.0575	0.0364	0.0296	0.0017	0.0008	-
390	F	22X	0.0221	0.0287	0.0265	0.0486	0.0110	0.0464	0.0221	0.0221	0.0110	0.0066	0.0022	0	A/A
	M	17X	0.0239	0.0418	0.0269	0.0239	0.0060	0.0537	0.0149	0.0447	0.0269	0.0149	0	0	A/L38
	390N	73X	0.0349	0.0322	0.0728	0.0856	0.0311	0.0802	0.0325	0.0450	0.0349	0.0244	0.0024	0.0003	-
420	F	30X	0.0101	0.0268	0.0385	0.0469	0.0168	0.0519	0.0084	0.0117	0.0101	0.0101	0.0117	0	A/L20
	M	27X	0.0127	0.0362	0.0416	0.0398	0.0036	0.0307	0.0126	0.0325	0.0181	0.0145	0.0036	0.0108	A/C
	420N	11X	0.0963	0.0229	0.0367	0.0505	0.0046	0.0826	0.0229	0.0092	0.0229	0.0092	0.0183	0	-
443	F	32X	0.0296	0.0172	0.0250	0.0499	0.0094	0.0374	0.0109	0.0296	0.0125	0.0047	0	0	A/A
	M	41X	0.0142	0.0303	0.0242	0.0763	0.0085	0.0303	0.0097	0.0097	0.0109	0.0109	0	0.0012	A/L24
	443N	8X	0.0613	0.0123	0.0307	0.0675	0.0184	0.0429	0.0184	0.0308	0.0245	0.0429	0	0	-
579	F	25X	0.0160	0.0321	0.0481	0.0561	0.0200	0.0401	0.0361	0.0341	0.0180	0.0140	0	0	A/A
	M	22X	0.0158	0.0271	0.0249	0.0633	0.0090	0.0520	0.0136	0.0113	0.0158	0.0181	0.0113	0.0023	A/L20
	579N	10X	0.0653	0.0201	0.0402	0.0553	0.0151	0.0503	0.0201	0.0302	0.0050	0.0050	0	0	-
580	F	21X	0.0238	0.0285	0.0404	0.0451	0.0024	0.0618	0.0071	0.0380	0.0166	0.0071	0	0.0071	A/L3
	M	25X	0.0121	0.0382	0.0221	0.0543	0.0101	0.0423	0.0141	0.0201	0.0121	0.0060	0.0020	0	A/L37
	580N	20X	0.0483	0.0266	0.0193	0.0459	0.0097	0.0700	0.0266	0.0242	0.0193	0.0048	0.0024	0	
595	F	37X	0.0134	0.0255	0.0282	0.0523	0.0094	0.0282	0.0121	0.0255	0.0255	0.0188	0.0013	0	A/A
	M	17X	0.0060	0.0150	0.0300	0.0390	0.0030	0.0390	0.0120	0.0270	0.0150	0.0060	0	0	A/A
	595N	26X	0.0594	0.0249	0.0287	0.1054	0.0096	0.0421	0.0115	0.0326	0.0192	0.0230	0	0.0038	-
728	F	18X	0.0088	0.0352	0.0235	0.0323	0.0029	0.0557	0.0176	0.0029	0.0293	0.0088	0.0264	0	A/L20
	M	17X	0.0254	0.0254	0.0226	0.0452	0.0085	0.0565	0.0056	0.0367	0.0226	0.0085	0	0.0028	A/B
	728N	111X	0.0646	0.0704	0.0267	0.0602	0.0120	0.0508	0.0152	0.0218	0.0290	0.0120	0.0201	0.0027	-
752	F	15X	0.0373	0.0271	0.0407	0.0237	0.0102	0.0305	0.0136	0.0237	0.0339	0.0034	0	0	A/B
	M	20X	0.0170	0.0414	0.0365	0.0852	0.0024	0.0341	0.0024	0.0292	0.0073	0.0049	0.0097	0.0146	A/C
	752N	127X	0.1768	0.0613	0.0223	0.0539	0.0055	0.0320	0.0152	0.0133	0.0156	0.0047	0.0141	0.0031	
764	F	13X	0.0197	0.0551	0.0433	0.0354	0.0079	0.0590	0.0079	0.0433	0.0315	0.0157	0	0	A/A
	M	23X	0.0191	0.0404	0.0297	0.0489	0	0.0319	0.0064	0.0063	0.0234	0.0042	0.0021	0.0043	A/A
	764N	92X	0.1136	0.0302	0.0210	0.0668	0.0005	0.0248	0.0081	0.0118	0.0065	0.0172	0.0011	0	-
794	F	15X	0.0130	0.0519	0.0357	0.0519	0.0065	0.0390	0.0130	0.0422	0.0292	0.0195	0.0065	0.0032	A/A
	M	14X	0.0141	0.0141	0.0424	0.0212	0	0.0530	0.0141	0.0247	0.0247	0.0212	0.0212	0	B/D
	794N	40X	0.1038	0.04	0.0375	0.0513	0.0175	0.0425	0.0188	0.0175	0.0075	0.0113	0.0088	0.0013	-

[a] Alleles A, B, C, D, L3, L20 and L24 are described in Berg et al. 2010 [8]. Alleles L37 and L38 are novel alleles, described in Supplementary Results.

Table S4. *PRDM9* alleles in an additional 10 ALL trios with B-ALL children based on read data. ZnF repeat types present in *PRDM9* were inferred based on SOLiD read data, for fathers (F), mothers (M) and patients (N) in each family. Repeat types (*a* to *l*) are described in Figure S6. Repeat types with a proportion above 0.01 (highlighted) are inferred to be present in the individuals.

Family	Individual	Coverage	Proportion of Exome Sequencing Reads Aligning to PRDM9 ZnF Repeats											
			a	*b*	*c*	*d*	*e*	*f*	*g*	*h*	*i*	*j*	*k*	*l*
392	F	30X	0.101	0.053	0.0215	0.0381	0.0033	0.0712	0.0083	0.043	0.0248	0.0182	0	0
	M	27X	0.0731	0.0567	0.0347	0.0475	0.0037	0.0969	0.0183	0.0457	0.0475	0.0201	0	0.0018
	392N	78X	0.0829	0.0383	0.0625	0.074	0.0261	0.0778	0.0274	0.0657	0.0338	0.03	0.0019	0
424	F	33X	0.1161	0.0268	0.0357	0.067	0.0149	0.0818	0.0402	0.0685	0.0327	0.0298	0.006	0
	M	26X	0.0854	0.0645	0.0361	0.0323	0.0114	0.0721	0.0209	0.0361	0.0304	0.019	0.0038	0
	424N	47X	0.1876	0.0486	0.0423	0.0455	0.0095	0.0708	0.0381	0.0476	0.0275	0.019	0.0011	0
614	F	33X	0.1271	0.0651	0.0182	0.0272	0.0076	0.0227	0.0045	0.0605	0.0121	0.0136	0.0182	0.0106
	M	12X	0.1548	0.0502	0.0418	0.0167	0.0084	0.0502	0.0209	0.0209	0.0251	0.0293	0	0
	614N	76X	0.162	0.0567	0.0423	0.0476	0.0078	0.0495	0.015	0.0436	0.0267	0.0137	0.0007	0
617	F	13X	0.1231	0.0423	0.0346	0.0615	0.0346	0.0462	0.0423	0.0385	0.0231	0.0154	0	0
	M	11X	0.129	0.0599	0.0276	0.0276	0	0.0323	0.0184	0.0323	0.0323	0.0046	0.0046	0
	617N	13X	0.1088	0.083	0.0226	0.1132	0.0113	0.0415	0.0226	0.0491	0.0302	0.0113	0	0
657	F	9X	0.1703	0.033	0.022	0.0385	0.0055	0.0604	0.011	0.0604	0.033	0.0165	0.0055	0
	M	12X	0.1331	0.0403	0.0282	0.0524	0.004	0.0605	0.0161	0.0323	0.0403	0.0242	0	0
	657N	5X	0.0612	0.011	0.0659	0.1099	0.0549	0.0659	0	0.0769	0.011	0.022	0.009	0
685	F	24X	0.0763	0.0495	0.0268	0.0392	0.0062	0.0763	0.033	0.033	0.0371	0.0186	0.0144	0.0051
	M	20X	0.2545	0.0375	0.03	0.04	0.0075	0.05	0.025	0.05	0.035	0.02	0.0025	0
	685N	75X	0.1651	0.0617	0.0504	0.0338	0.0179	0.067	0.0239	0.0498	0.0319	0.01	0.002	0.0007
761	F	38X	0.0652	0.0417	0.0248	0.0495	0.0091	0.0717	0.0313	0.0456	0.0183	0.0209	0.0013	0.0013
	M	40X	0.0891	0.054	0.0151	0.0276	0.0025	0.0552	0.0113	0.0452	0.0213	0.0364	0.0025	0
	761N	92X	0.1351	0.0346	0.0362	0.0324	0.0086	0.0789	0.0243	0.0546	0.0286	0.0178	0.0005	0
762	F	25X	0.097	0.0257	0.0436	0.0535	0.0119	0.0614	0.0158	0.0416	0.0238	0.0139	0	0
	M	20X	0.0973	0.0389	0.0414	0.0389	0.0049	0.0681	0.0292	0.0316	0.0292	0.0292	0	0
	762N	81X	0.1193	0.0555	0.0262	0.0366	0.0085	0.0634	0.0128	0.0469	0.0189	0.014	0.0018	0.0006
767	F	21X	0.1253	0.0394	0.0487	0.0557	0.0046	0.0626	0.0116	0.0232	0.0255	0.0093	0.0116	0
	M	32X	0.1144	0.058	0.0329	0.0502	0.0094	0.0533	0.0078	0.0094	0.011	0.0157	0.0204	0.0031
	767N	80X	0.1277	0.0646	0.0292	0.0385	0.0062	0.0752	0.0143	0.0242	0.0311	0.0168	0.0242	0
777	F	18X	0.102	0.0737	0.0283	0.0255	0.0113	0.0567	0.0227	0.0482	0.0198	0.0368	0	0
	M	22X	0.1615	0.0664	0.0221	0.0442	0.0066	0.0708	0.0221	0.0243	0.0243	0.0243	0.0111	0
	777N	59X	0.2017	0.0524	0.0304	0.0355	0.0127	0.0609	0.0144	0.0262	0.0262	0.0144	0.0203	0

Table S5. *PRDM9* alleles in 76 French-Canadian individuals.

Individuals are mothers (M) and fathers (F) of at least 2 children from 49 families **(A)** one parent was sampled per family and with **(B)** both parents sampled per families. *PRDM9* ZnF alleles were assayed by Sanger sequencing. Genotypes with rare alleles are highlighted. Alleles A, B, C, D, L1, L9, L20 and L24 are described in Berg et al. 2010 (8). Alleles L32, L33 and L34 are novel alleles, described in Supplementary Results.

A

Individual ID	Parent	Alleles
8	M	A/A
118	M	A/L24
183	M	A/A
385	M	A/A
210	M	A/A
190	M	A/L9
229	M	A/A
743	M	A/L1
51	M	A/L24
772	M	A/A
608	M	A/B
245	M	A/A
20	F	A/A
257	F	A/L20
270	F	A/A
596	F	A/A
64	F	A/A
713	F	A/A
717	F	A/A
90	F	A/A
944	F	A/A
146	F	A/A

B

Couple ID	Alleles	
	M	F
15_11	A/A	A/B
223_222	A/A	A/A
304_303	A/L32	A/A
348_347	A/A	A/A
38_39	A/A	L20/L24
393_392	A/C	A/A
413_412	A/A	A/A
424_423	A/A	A/A
428_427	A/A	A/A
43_48	A/A	A/A
460_459	A/L9	A/C
584_585	A/A	A/L33
626_625	A/A	A/A
647_651	A/A	A/L34
656_657	A/A	A/B
66_68	A/B	A/L24
692_691	A/L32	A/B
728_311	A/A	A/A
740_739	A/A	A/B
748_749	A/A	A/A
755_756	A/A	A/A
75_74	A/A	A/A
812_815	A/A	A/A
818_817	A/A	A/A
830_823	A/A	A/A
854_853	A/D	A/A
96_95	A/A	A/B

Table S6. B-ALL molecular subtypes for the 24 patients included in this study.

The patients present different subtypes of B-ALL: high hyperdiploid clones (H), clones with translocation (T) and other uncharacterized translocations or genetic defects (O). There is no significant difference between subtypes for the presence of *k*-finger alleles in a family (Freeman-Halton test with 3 categories [16], *p* = 0.268). There is no significant difference between maternal (M) and paternal (F) origin of the *k*-finger alleles (*p* = 0.369, Fisher's exact test).

Child	Sex	Molecular Group	Detected translocations	Leukemic clone ploidy	*k*-finger in Family	Parent Carrier
610	Male	H	None	51-53	Yes	M
383	Male	H	None	55	Yes	M
375	Female	T	t(12;21)	46	No	-
380	Female	O	n/d	46	No	-
390	n/d	H	None	54	No	-
420	n/d	T	t(12;21)	n/d	Yes	M and F
443	Female	T	t(12;21)	46	No	-
579	n/d	O	None	46	Yes	M
580	n/d	T	t(12;21)	n/d	No	-
595	Male	O	None	47	No	-
728	n/d	O	None	n/d	Yes	F
752	Male	T	t(9;12)	45	Yes	M
764	Male	H	None	56	No	-
794	Male	O	None	46	Yes	M
392	Male	T	t(12;21)	46	No	-
424	Male	T	t(1;19)	46	No	-
614	Female	T	t(12;21)	46	Yes	F
617	n/d	O	None	n/d	No	-
657	n/d	T	t(12;21)	n/d	No	-
685	n/d	O	None	n/d	Yes	F
761	Male	H	None	56	No	-
762	n/d	O	None	n/d	No	-
767	Female	O	None	n/d	Yes	M and F
777	Female	H	None	49-54	Yes	M

Table S7. ***PRDM9*** **alleles in 50 children from SJDALL cohort based on read data.**

Patients are separated in 4 B-ALL subtypes: ETV6 translocation (SJETV), hypodiploid (SJHYPO), infant (SJINF) and Philadelphia chromosome-positive (SJPHALL). Reported ethnicities were verified by PCA of genotyped data (Figure S8). Illumina read data from tumor and normal sample sequencing were used. Repeat types found in the read data with a proportion above 0.01 (highlighted) are inferred to be present in the individuals. Repeat types (*a* to *l*) are described in Figure S6.

Sample	**Reported ethnicity (PCA)**	*a*	*b*	*c*	*d*	*e*	*f*	*g*	*h*	*i*	*j*	*k*	*l*
SJETV010	White	0,065	0,052	0,1194	0,1525	0,0437	0,169	0,0686	0,0792	0,0697	0,0922	0	0,0012
SJETV022	White	0,0956	0,0566	0,127	0,1597	0,0314	0,1597	0,0616	0,0541	0,0717	0,073	0,0025	0,0025
SJETV024	White	0,1004	0,0588	0,1136	0,1714	0,0385	0,1572	0,0619	0,0598	0,0619	0,0822	0,003	0
SJETV027	White	0,1155	0,0599	0,1484	0,1284	0,0485	0,1469	0,0728	0,0613	0,0542	0,0942	0,0014	0,0043
SJETV028	White	0,0906	0,0523	0,1115	0,151	0,0441	0,1521	0,0743	0,0476	0,0662	0,0952	0,0023	0,0012
SJETV073	White	0,0871	0,0389	0,1191	0,1294	0,047	0,1569	0,0664	0,0825	0,0573	0,0653	0,0011	0,0034
SJETV085	White	0,0947	0,0579	0,107	0,1719	0,0526	0,1377	0,0588	0,0675	0,0781	0,0895	0,0018	0
SJETV089	White	0,0952	0,0476	0,1125	0,1743	0,0547	0,1358	0,0689	0,0598	0,0912	0,075	0	0,001
SJETV194	White	0,1035	0,0472	0,1068	0,1631	0,0422	0,1548	0,072	0,0522	0,0927	0,0844	0	0,0025
SJHYPO004	White	1,4588	0,0547	0,1258	0,1569	0,0588	0,1422	0,0662	0,0596	0,0743	0,0825	0	0,0016
SJHYPO006	White (adx)	0,0778	0,0488	0,1394	0,1092	0,0174	0,0871	0,0163	0,1254	0,0778	0,1045	0,0476	0,0395
SJHYPO013	Other(Asian)	0,1008	0,0584	0,142	0,1386	0,0309	0,1581	0,0653	0,0561	0,0687	0,0882	0	0,0046
SJHYPO021	White	0,1084	0,0515	0,1041	0,1468	0,0438	0,1391	0,0635	0,0624	0,0756	0,069	0,0011	0,0044
SJHYPO022	White	0,0929	0,0506	0,0988	0,1247	0,0565	0,1412	0,0824	0,0788	0,0624	0,0765	0	0,0024
SJHYPO040	White	0,0978	0,0422	0,0989	0,1578	0,05	0,13	0,0711	0,0222	0,08	0,0844	0,0322	0
SJHYPO042	White	0,0964	0,0321	0,1358	0,1378	0,0394	0,1492	0,0591	0,0808	0,0881	0,0725	0,0259	0,0114
SJHYPO044	White	0,0912	0,0592	0,1208	0,1739	0,0567	0,1406	0,0629	0,0838	0,0826	0,0703	0,0012	0,0012
SJHYPO046	White	0,0922	0,0454	0,121	0,121	0,055	0,1692	0,0646	0,0523	0,0688	0,088	0	0,0028
SJHYPO051	White (adx)	0,1194	0,0525	0,1089	0,1325	0,0617	0,1483	0,0499	0,0604	0,0722	0,0682	0	0,0026
SJHYPO052	White	0,0859	0,0452	0,0914	0,1333	0,0562	0,1773	0,0617	0,0738	0,0804	0,0881	0	0
SJHYPO055	White	0,1009	0,0482	0,1377	0,136	0,0684	0,1675	0,0667	0,0272	0,0614	0,0728	0,0342	0,0018
SJHYPO056	White	0,0883	0,0331	0,1145	0,1283	0,0483	0,2	0,1034	0,0676	0,0717	0,0952	0,0014	0
SJHYPO119	White	0,1138	0,0588	0,1077	0,1457	0,0465	0,12	0,06	0,0575	0,0612	0,0942	0,0037	0,0024
SJHYPO120	Other (Hisp)	0,1115	0,0632	0,0979	0,1586	0,0483	0,1561	0,0595	0,062	0,0595	0,0768	0,0012	0
SJHYPO123	White	0,0903	0,0593	0,1442	0,1226	0,0013	0,1119	0,0472	0,0822	0,0916	0,0836	0,0202	0,0162

Table S7 *(continued)*

Sample	Reported ethnicity (PCA)	*a*	*b*	*c*	*d*	*e*	*f*	*g*	*h*	*i*	*j*	*k*	*l*
SJINF001	White	0,0989	0,053	0,0901	0,1325	0,0336	0,1237	0,0636	0,0601	0,0583	0,0901	0,0018	0
SJINF002	White	0,8073	0,0477	0,1134	0,1314	0,0322	0,1198	0,0619	0,058	0,0464	0,0851	0,0026	0
SJINF003	White	0,1079	0,0539	0,0991	0,1181	0,0466	0,1254	0,0729	0,0466	0,051	0,0802	0,0102	0,0029
SJINF004	White (adx)	0,0971	0,0659	0,0989	0,1099	0,0403	0,1337	0,0659	0,0513	0,0769	0,0842	0	0,0018
SJINF005	White	0,1058	0,0276	0,0982	0,135	0,0537	0,138	0,0613	0,0675	0,0537	0,0813	0,0031	0,0015
SJINF006	White	0,1079	0,0468	0,1079	0,1571	0,0528	0,1439	0,0576	0,0635	0,0743	0,0743	0,0012	0,0012
SJINF007	White	0,081	0,0444	0,1059	0,1752	0,0471	0,1529	0,1007	0,0458	0,0745	0,0706	0,0013	0,0013
SJINF009	White	0,1008	0,0485	0,1263	0,1505	0,0306	0,148	0,0842	0,0268	0,051	0,0663	0,0472	0
SJINF011	White	0,0914	0,054	0,1167	0,1189	0,0474	0,1355	0,0584	0,0727	0,0793	0,0771	0,0011	0,0011
SJINF012	White	0,0885	0,0369	0,1118	0,1339	0,0344	0,1413	0,0725	0,0762	0,0676	0,0909	0	0,0025
SJINF013	White	0,091	0,0556	0,1327	0,0999	0,0164	0,1466	0,0721	0,0708	0,0582	0,0999	0,0025	0
SJINF014	Hispanic	0,098	0,059	0,1359	0,1258	0,0223	0,0991	0,029	0,1102	0,0724	0,0835	0,0379	0,0212
SJINF015	White	0,129	0,0496	0,1042	0,1191	0,0323	0,1377	0,0658	0,0484	0,098	0,0744	0,005	0,0025
SJINF016	Hispanic	0,0943	0,0432	0,1489	0,1	0,017	0,1034	0,0295	0,0693	0,067	0,0886	0,0625	0,0216
SJINF017	Hispanic	0,1032	0,0523	0,0921	0,1004	0,0418	0,1402	0,0635	0,06	0,0676	0,0781	0,0021	0,0014
SJINF019	White	0,118	0,0504	0,0878	0,1468	0,036	0,1583	0,0647	0,0576	0,0619	0,0835	0,0014	0,0014
SJINF020	White	0,1169	0,0438	0,1036	0,1474	0,0372	0,1421	0,073	0,0664	0,0558	0,0704	0,008	0
SJINF022	White	0,9667	0,0516	0,1371	0,117	0,0189	0,1006	0,0239	0,0994	0,0516	0,073	0,0365	0,0352
SJPHALL001	White	0,0713	0,0526	0,132	0,1507	0,0654	0,1507	0,0841	0,0678	0,0759	0,0654	0	0
SJPHALL003	White	0,0906	0,0482	0,1189	0,148	0,0532	0,1654	0,0657	0,0673	0,059	0,0798	0,0017	0,0025
SJPHALL004	Asian	0,0902	0,0486	0,1266	0,1449	0,0466	0,1459	0,0811	0,0669	0,0719	0,0973	0,001	0,0041
SJPHALL005	White (adx)	0,0931	0,0486	0,1275	0,1306	0,0547	0,1528	0,0739	0,0628	0,0648	0,0921	0	0,002
SJPHALL006	White	0,085	0,0385	0,1023	0,1859	0,0332	0,1554	0,0637	0,0531	0,0491	0,085	0,0027	0
SJPHALL007	White (adx)	0,0847	0,0367	0,1195	0,159	0,0555	0,1515	0,0593	0,0724	0,0931	0,0753	0,0028	0
SJPHALL008	White	0,089	0,0503	0,1285	0,1533	0,055	0,154	0,0789	0,0557	0,0875	0,072	0,0008	0,0015

Table S8: Most frequent translocations and fusion genes in ALL.

(A) The most frequent translocations involved in ALL found in the dbCRID and Mitelman database [17,18]. Translocations were selected only if they were reported in more than 10 entries for ALL in these databases. (B) The ALL gene is composed of 38 fusion genes involved in the translocations reported in (A) and found to be implicated in ALL in peer-reviewed publications.

A

Translocations	Cytogenetic bands		Entries in Databases	
			dbCRID	Mitelman
t(1,11)	1p32/1q23	11q23	0	11
t(1,14)	1p32(p33)	14q11	5	6
	1q21	14q32	8	9
t(1,19)	1q23	19p13.3	7	49
t(4,11)	4q21	11q23.3	20	88
t(5,14)	5q34(q35)	14q11/14q32	3	9
t(6,11)	6q27	11q23	1	10
t(7,9)	7q34/q11	9q34/p13	1	16
t(8,14)	8p24	14q11/q32	3	24
t(9,11)	9p21/q34	11q23.3	5	21
t(9,22)	9q34	22q11.2	19	136
t(10,11)	10p12	11q14/q23	2	13
t(10,14)	10p24.31	14q11.2/q32	8	8
t(11,19)	11q23	19p13.3	5	30
t(17,19)	17q22	19p13	2	11
t(12,21)	12p13.2	21q22.12	29	58

Table S8 (*continued*)

B

Gene	Chr	Start	End	Translocations	Nb of *C* motifs	*C* motifs by Kb
TAL1	1	47454550	47469974	t(1;14)(p32;q11)	5	0,314
EPS15	1	51592522	51757583	t(1;11)(p32;q23)	3	0,018
BCL9	1	145479805	145564639	t(1;14)(q21;q32)	2	0,024
MLLT11	1	149298774	149307597	t(1;11)(q21;q23)	0	-
PBX1	1	162795560	163082934	t(1;19)(q23;p13)	5	0,017
SEPT11	4	78089918	78178792	t(4;11)(q21;q23)	5	0,056
AFF1	4	88075186	88232005	t(4;11)(q21;q23)	9	0,044
RANBP17	5	170221599	170659624	t(5;14)(q34;q11)	16	0,037
TLX3	5	170668892	170671743	t(5;14)(q34;q11/q32)	4	1,403
NKX2-5	5	172591743	172594868	t(5;14)(q34;q32)	0	-
MLLT4	6	167970519	168115552	t(6;11)(q27;q23)	8	0,055
AUTS2	7	68701840	69895821	t(7;9)(q11;p13)	39	0,033
POM121	7	71987871	72059915	t(7;9)(q11;p13)	7	0,097
ELN	7	73080362	73122172	t(7;9)(q11;p13)	12	0,287
TCRB	7	141674678	141987064	t(7;9)(q34;q34)	9	0,046
MYC	8	128816946	128820200	t(8;14)(q24;q11/q32)	0	-
PVT1	8	128875960	129182681	t(8;14)(q24;q11/q32)	16	0,052
MLLT3	9	20334967	20612514	t(9;11)(q21;q23)	9	0,032
PAX5	9	36828530	37024476	t(7;9)(q11;p13)	23	0,117
ABL1	9	132579088	132752883	t(9;22)(q34;q11)	4	0,023
NOTCH1	9	138508716	138560059	t(7;9)(q34;q34)	21	0,409
MLLT10	10	21863107	22072560	t(10;11)(p12;q14/q23)	8	0,038
TLX1	10	102880251	102887526	t(10;14)(p24;q11)	3	0,412
LMO2	11	33836698	33870412	t(7;11)(q34;p13)	2	0,059
PICALM	11	85346132	85457756	t(10;11)(p12;q14)	2	0,018
MLL	11	117812414	117901146	t(1;11)(p32;q23),t(4;11)(q21;q23),t(6;11)(q27;q23), t(9;11)(q21;q23),t(10;11)(p12;q23),t(11;19)(q23;p13)	6	0,066
CCND2	12	4253198	4284777	t(7;12)(q34;p13)	2	0,063
ETV6	12	11694054	11939592	t(12;21)(p13;q22),t(7;12)(q34;p13),t(9;12)(q34;p13)	14	0,052
TRA@ TRD@	14	21432409 21987946	21604421 21995540	t(1;14)(p32;q11), t(5;14)(q34;q11), t(8;14)(q24;q11), t(10;14)(p24;q11), t(11;14)(p13/q23;q11)	17 0	0,030 -
BCL11B	14	98705377	98807575	t(5;14)(q34;q32)	22	0,215
IGH@	14	105124270	105401515	t(1;14)(q21;q32), t(5;14)(q34;q32), t(8;14)(q24;q32)	245	0,884
HLF	17	50697320	50757425	t(17;19)(q22;p13)	0	-
DAZAP1	19	1358583	1386682	t(1;19)(q23;p13)	13	0,463
TCF3	19	1560294	1603328	t(1;19)(q23;p13)	20	0,465
MLLT1	19	6161391	6230959	t(11;19)(q23;p13)	19	0,273
RUNX1	21	35081967	35343511	t(12;21)(p13;q22)	64	0,053
BCR	22	21852551	21990224	t(9;22)(q34;q11)	26	0,189

Table S9. PRDM9 alleles binding motifs in unique and repetitive DNA

Number of motifs *A* and *C,* as presented in Table 2, in coding regions (similar results are obtained for the whole genome) and in the ALL gene list (Table S8). We compared counts using odds ratios (OR), to measure the association between motifs and their occurrence in the ALL gene list. The motif search was performed on the non-degenerate version of the Human Reference Genome (hg18). Repetitive regions were obtained from UCSC tables, with regions found by RepeatMasker [19] and Tandem Repeat Finder [20] programs considered as repetitive DNA. Segmental duplications coordinates were also obtained from UCSC tables.

	Unique DNA		**Repetitive DNA**		**Segmental Duplications**	
	Genes	ALL gene list	Genes	ALL gene list	Genes	ALL gene list
A	3227	13	3726	21	371	3
C	30313	390	64928	458	5489	206
C vs *A* OR [CI]	2.39 [1.50;3.81]**		1.81 [0.69;4.76]		4.78 [1.84;12.46] **	

** significant based on 95% CI (one-tailed $p < 0.025$)

Table S10. Data and Analyses performed in this study.

This study uses genetic information from a total of 639 individuals. The ALL quartet is part of the FCALL cohort but is displayed seperately since many analyses were performed on this family only.

Dataset	Families	Individuals per family	Data	Analyses
ALL quartet	1	2 parents + 2 offsprings	Genotyping on Illumina 2.5M	Ancestry analyses
			Genotyping on Affymetrix 6.0	Recombination Analyses
			Exome Sequencing on SOLiD 4.0	*De novo* Mutation Discovery
			Sanger sequencing of PRDM9 ZnF alleles of parents	PRDM9 ZnF alleles determination
FCALL cohort (Total of 22 parental trios)	22	2 parents + 1 offspring	Exome Sequencing on SOLiD 4.0	Ancestry analyses
	12*	2 parents + 1 offspring	Sanger sequencing of PRDM9 ZnF alleles of parents	PRDM9 ZnF alleles determination
SJDALL cohort (Total of 61 children)	61	1 individual	Genotyping on Affymetrix 6.0	Ancestry analyses
	50*	1 individual	Paired-end WGS on Illumina HiSeq	PRDM9 ZnF alleles determination
FCEXOME cohort (34 pairs of parents)	34	2 parents	Exome Sequencing on SOLiD 4.0	PRDM9 ZnF alleles determination
FC cohort (Total of 69 families)	69	2 parents + 2 offsprings	Genotyping on Affymetrix 6.0	Recombination Analyses
				Ancestry analyses
	27*	2 parents	Sanger sequencing of PRDM9 ZnF alleles of parents	PRDM9 ZnF alleles determination
	22*	1 parent		
1000 Genomes CEU	99 (including pairs of parents and unrelated individuals)		Exome Sequencing on Illumina GA II	PRDM9 ZnF alleles determination
Moroccan cohort (Total of 163 indiv)	163	1 individual	Genotyping on Illumina Human 610-Quad	Ancestry analyses
	27*	1 individual	Sanger sequencing of PRDM9 ZnF alleles	PRDM9 ZnF alleles determination

* A subset of the complete cohort

SUPPLEMENTARY FIGURES

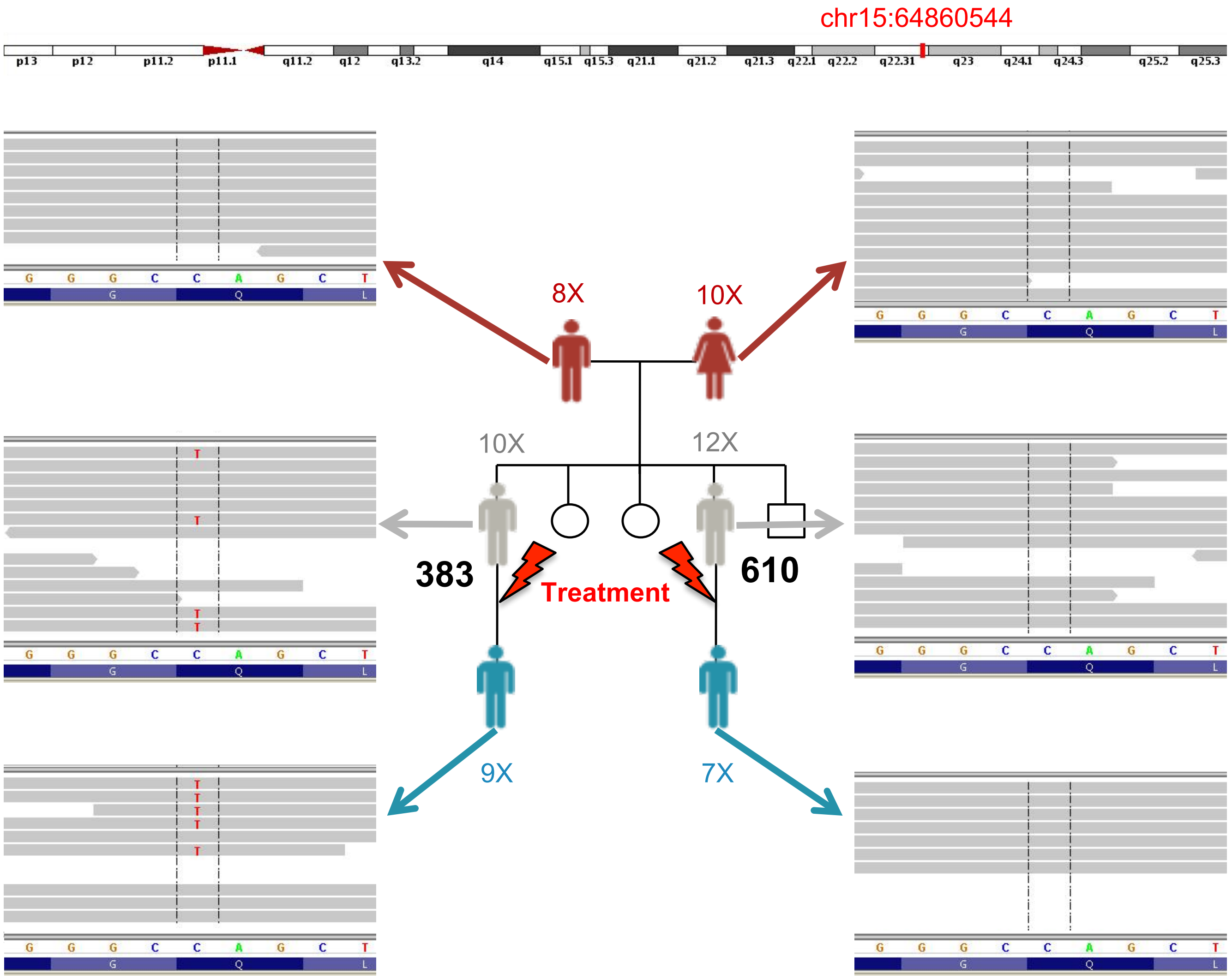


Figure S1. Identification of a *de novo* mutation in the SMAD6 gene on chromosome 15.

The variant allele is evenly sampled in patient 383 samples but is not seen in any of the other samples. Given the data, the probability that this mutation was inherited is $p = 9.77\times10^{-3}$ (Supplementary Methods).

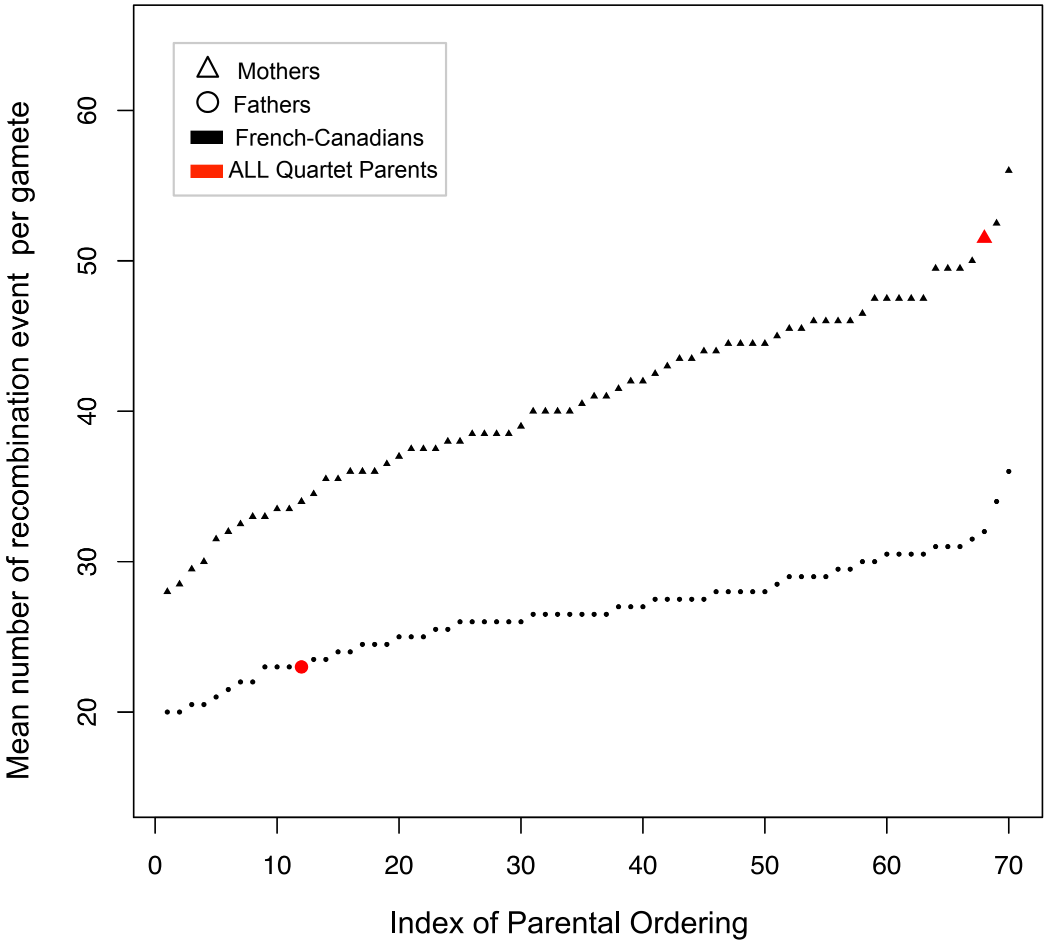


Figure S2. Mean recombination rate in the parents from the ALL quartet and the FC cohort.

Recombination events were called using genotyping data from the Affymetrix 6.0 array. The FC cohort is composed of 69 French-Canadian quartets. For the ALL quartet, only children's post-treatment samples were used to infer recombination. Mothers (triangles) and fathers (circles) are ordered according to their recombination rate. Genetic markers from Affymetrix 6.0 platform were used for all individuals, including the ALL quartet parents.

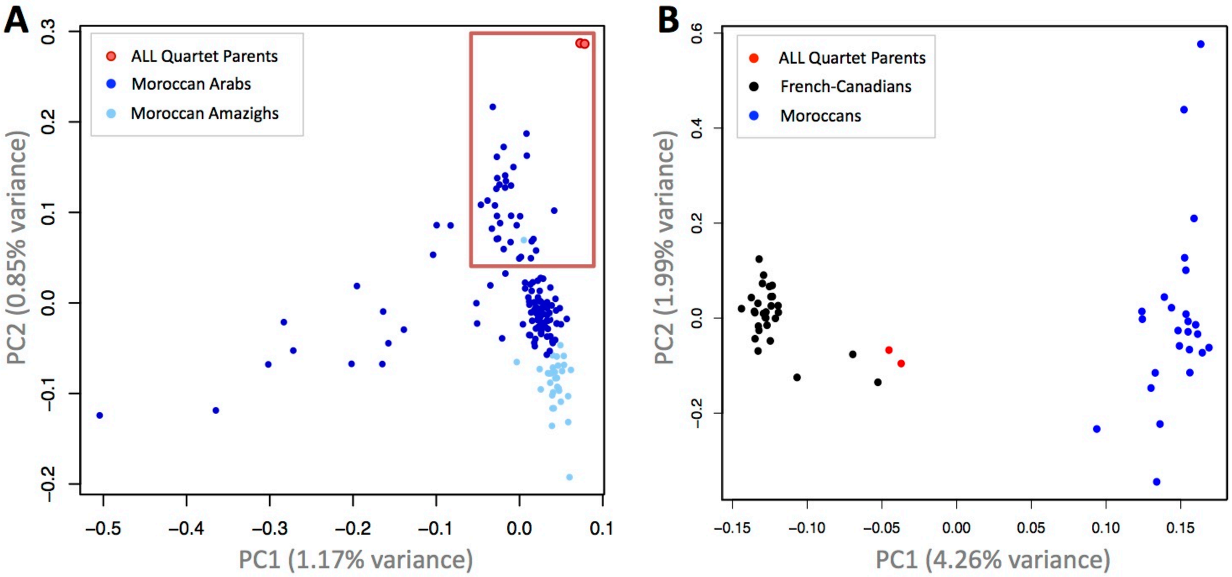


Figure S3. Genetic Ancestry of the ALL Quartet Parents.

We preformed a Principal Component Analysis of genetic variation using 61,454 SNPs from (A) the ALL quartet parents and 163 unrelated Moroccans; (B) the ALL quartet parents, the 27 unrelated Moroccans, chosen to be the closest to the ALL quartet parents (falling in the red rectangle in panel A), and 28 unrelated French-Canadians (Supplementary Results). The ALL quartet parents are closer to Moroccans of Arab ancestry then to Moroccan Berbers (Amazighs).

Individual ID	Gender	Alleles
A55	Female	A/A
A51	Female	A/A
A41	Female	A/L35
A109	Female	A/A
B04	Female	L36/L14
B08	Female	A/L20
I06	Female	A/A
I54	Female	A/A
I59	Female	A/A
I71	Female	A/A
I62	Female	A/A
I66	Female	B/L20
A60	Female	A/E
I09	Female	A/A
I28	Female	A/A
A68	Female	A/A
A65	Female	A/L37
I52	Female	A/A
A46	Male	A/A
A32	Male	A/A
A145	Male	A/A
A147	Male	A/A
A150	Male	A/A
A117	Male	L37/L20
A138	Male	A/L37
I19	Male	A/L37
A112	Male	A/A

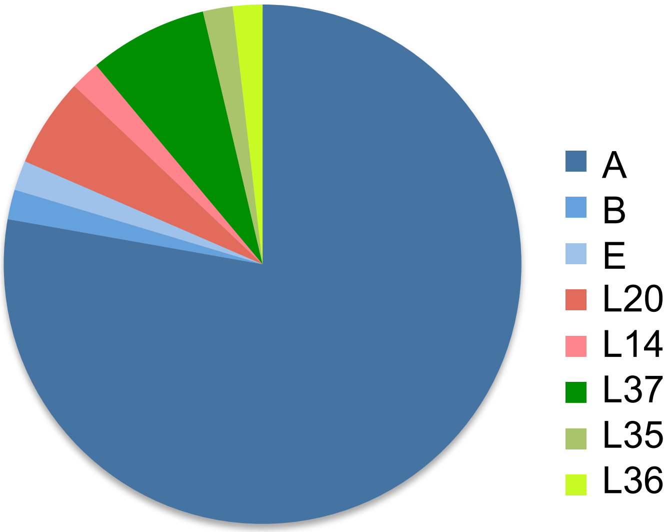


Figure S4. PRDM9 ZnF alleles in 27 unrelated Moroccan individuals.

Individuals were selected from the Moroccan cohort [14] based on ancestry, chosen to be the closest to the ALL quartet parents (Figure S3). PRDM9 ZnF alleles were assayed by Sanger sequencing. Alleles A, B, E, L14 and L20 are described previously [8]. Alleles L35, L36 and L37 are novel alleles, described in Supplementary Results.

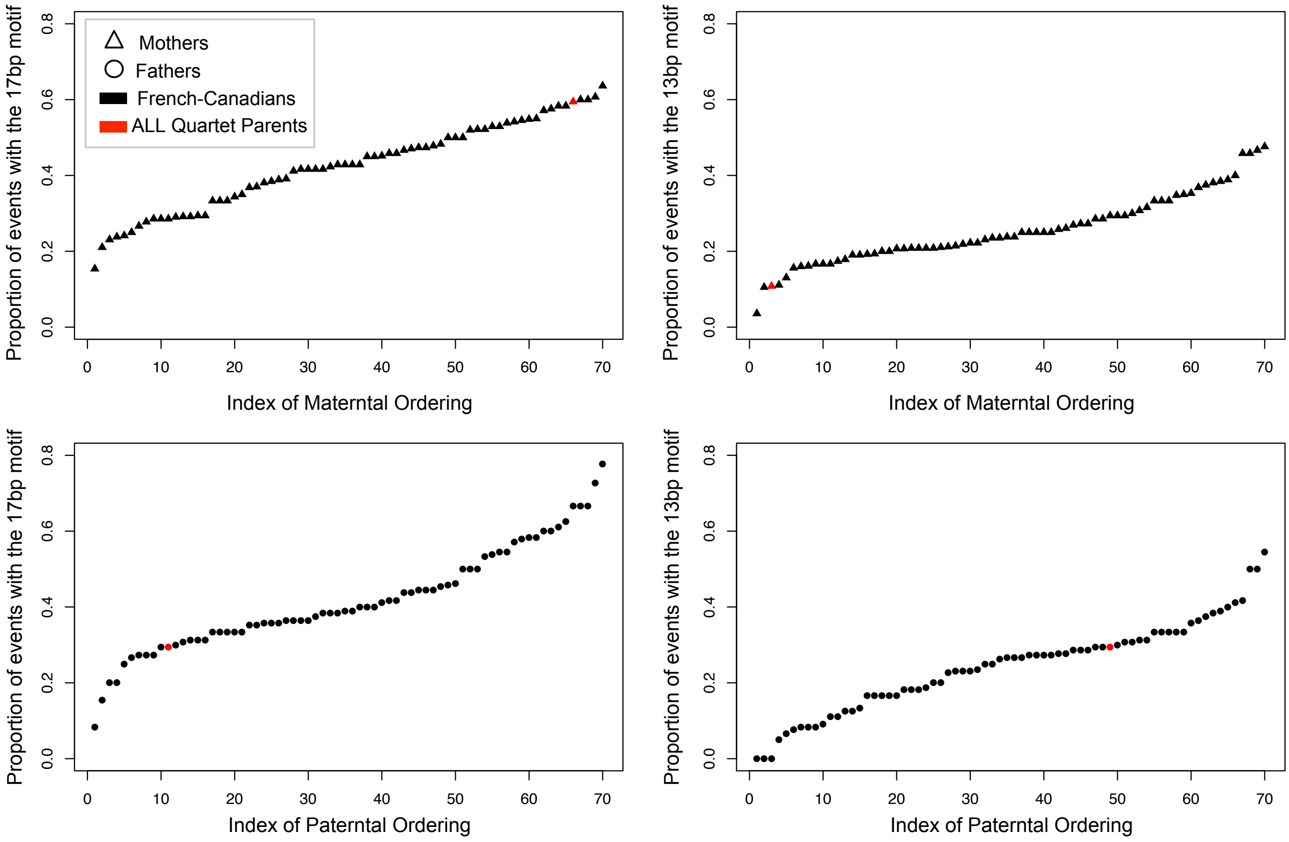


Figure S5. Proportion of recombination events called near *PRDM9* binding motifs.

We computed the proportion of recombination events overlapping the 17-bp and 13-bp predicted to be recognized by the C and A alleles of *PRDM9*, respectively, for parents of the ALL quartet, compared to parents from 69 French-Canadian quartets. The 13-bp motif, CCNCCNTNNCCNC is enriched in linkage disequilibrium-based hotspots inferred from HapMap data, whereas a 17-bp motif, CCNCNNTNNNCNNNNCC, is associated with African-enriched hotspots. Recombination events were called using genetic markers from Affymetrix 6.0 platform for all individuals, including the ALL quartet parents for which only children's post-treatment samples were used to call recombination events. Mothers (triangles) and fathers (circles) are ordered according to the proportion of motifs found near their recombination events.

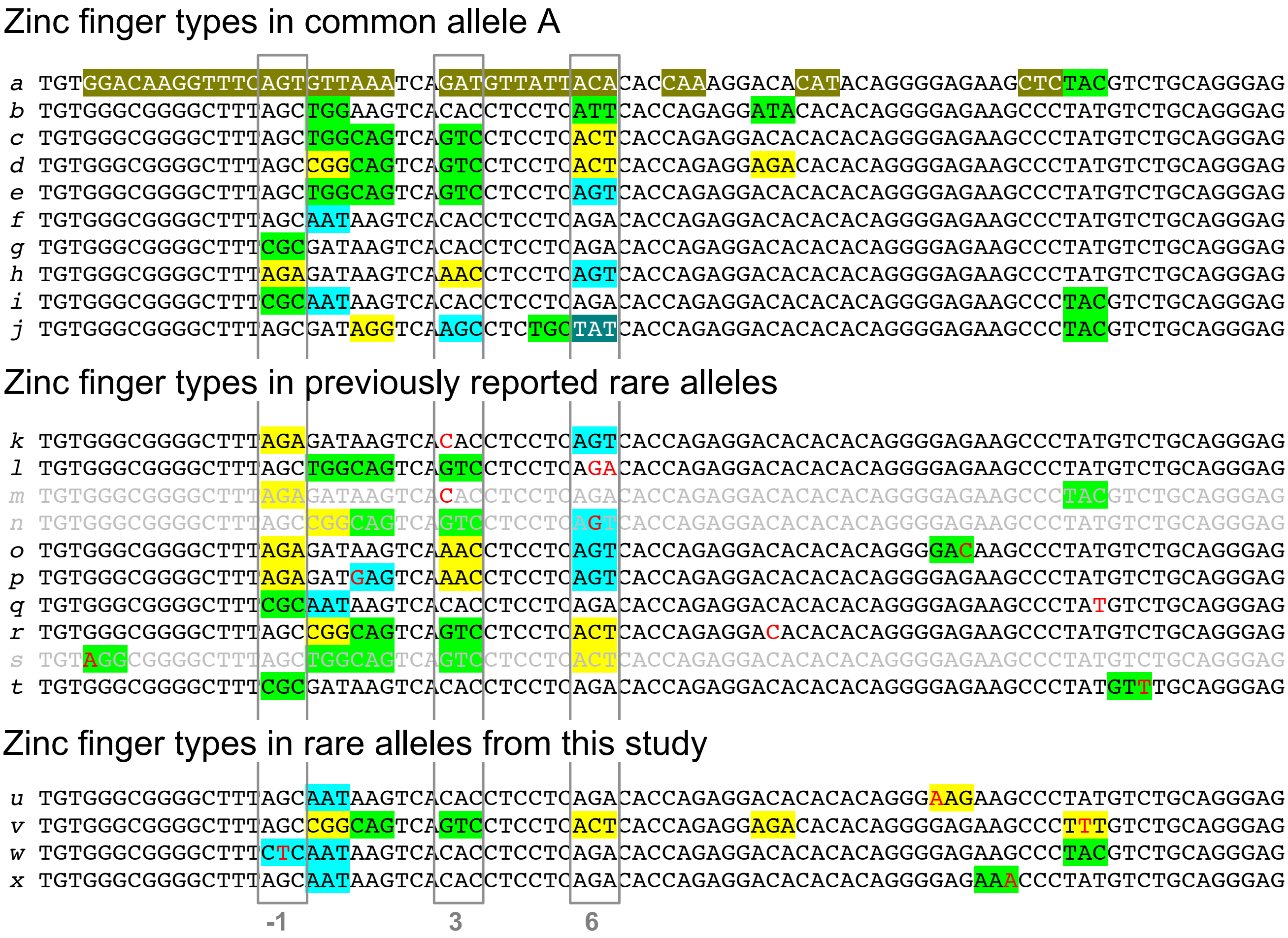

Zinc finger types in common allele A
a TGTGGACAAGGTTTCAGTGTTAAATCAGATGTTATTACACACCAAAGGACACATACAGGGGAGAAGCTCTACGTCTGCAGGGAG
b TGTGGGCGGGGCTTTAGCTGGAAGTCACACCTCCTCATTCACCAGAGGATACACACAGGGGAGAAGCCCTATGTCTGCAGGGAG
c TGTGGGCGGGGCTTTAGCTGGCAGTCAGTCCTCCTCACTCACCAGAGGACACACACAGGGGAGAAGCCCTATGTCTGCAGGGAG
d TGTGGGCGGGGCTTTAGCCGGCAGTCAGTCCTCCTCACTCACCAGAGGAGACACACAGGGGAGAAGCCCTATGTCTGCAGGGAG
e TGTGGGCGGGGCTTTAGCTGGCAGTCAGTCCTCCTCAGTCACCAGAGGACACACACAGGGGAGAAGCCCTATGTCTGCAGGGAG
f TGTGGGCGGGGCTTTAGCAATAAGTCACACCTCCTCAGACACCAGAGGACACACACAGGGGAGAAGCCCTATGTCTGCAGGGAG
g TGTGGGCGGGGCTTTCGCGATAAGTCACACCTCCTCAGACACCAGAGGACACACACAGGGGAGAAGCCCTATGTCTGCAGGGAG
h TGTGGGCGGGGCTTTAGAGATAAGTCAAACCTCCTCAGTCACCAGAGGACACACACAGGGGAGAAGCCCTATGTCTGCAGGGAG
i TGTGGGCGGGGCTTTCGCAATAAGTCACACCTCCTCAGACACCAGAGGACACACACAGGGGAGAAGCCCTACGTCTGCAGGGAG
j TGTGGGCGGGGCTTTAGCGATAGGTCAAGCCTCTGCTATCACCAGAGGACACACACAGGGGAGAAGCCCTACGTCTGCAGGGAG
Zinc finger types in previously reported rare alleles
k TGTGGGCGGGGCTTTAGAGATAAGTCACACCTCCTCAGTCACCAGAGGACACACACAGGGGAGAAGCCCTATGTCTGCAGGGAG
l TGTGGGCGGGGCTTTAGCTGGCAGTCAGTCCTCCTCAGACACCAGAGGACACACACAGGGGAGAAGCCCTATGTCTGCAGGGAG
m TGTGGGCGGGGCTTTAGAGATAAGTCACACCTCCTCAGACACCAGAGGACACACACAGGGGAGAAGCCCTACGTCTGCAGGGAG
n TGTGGGCGGGGCTTTAGCCGGCAGTCAGTCCTCCTCAGTCACCAGAGGACACACACAGGGGAGAAGCCCTATGTCTGCAGGGAG
o TGTGGGCGGGGCTTTAGAGATAAGTCAAACCTCCTCAGTCACCAGAGGACACACACAGGGGACAAGCCCTATGTCTGCAGGGAG
p TGTGGGCGGGGCTTTAGAGATGAGTCAAACCTCCTCAGTCACCAGAGGACACACACAGGGGAGAAGCCCTATGTCTGCAGGGAG
q TGTGGGCGGGGCTTTCGCAATAAGTCACACCTCCTCAGACACCAGAGGACACACACAGGGGAGAAGCCCTATGTCTGCAGGGAG
r TGTGGGCGGGGCTTTAGCCGGCAGTCAGTCCTCCTCACTCACCAGAGGACACACACAGGGGAGAAGCCCTATGTCTGCAGGGAG
s TGTAGGCGGGGCTTTAGCTGGCAGTCAGTCCTCCTCACTCACCAGAGGACACACACAGGGGAGAAGCCCTATGTCTGCAGGGAG
t TGTGGGCGGGGCTTTCGCGATAAGTCACACCTCCTCAGACACCAGAGGACACACACAGGGGAGAAGCCCTATGTTTGCAGGGAG
Zinc finger types in rare alleles from this study
u TGTGGGCGGGGCTTTAGCAATAAGTCACACCTCCTCAGACACCAGAGGACACACACAGGGAAGAAGCCCTATGTCTGCAGGGAG
v TGTGGGCGGGGCTTTAGCCGGCAGTCAGTCCTCCTCACTCACCAGAGGAGACACACAGGGGAGAAGCCCTTTGTCTGCAGGGAG
w TGTGGGCGGGGCTTTCTCAATAAGTCACACCTCCTCAGACACCAGAGGACACACACAGGGGAGAAGCCCTACGTCTGCAGGGAG
x TGTGGGCGGGGCTTTAGCAATAAGTCACACCTCCTCAGACACCAGAGGACACACACAGGGGAGAAACCCTATGTCTGCAGGGAG
-1
3
6

Figure S6. ZnF repeat types of PRDM9 alleles.

Modified from Berg et al. 2010 [8]. Variable codons are colored. Red bases in ZnF repeat types from K to X are differences with respect to the closest repeat type present in the hg18 reference genome (note that the hg18 reference genome does not encode the E finger and therefore, the L finger is two differences away from the C repeat type). Repeat types M, N and S [8,21] were not found in this study. The 3 codons coding for the binding unit of each repeat (positions -1, 3 and 6 of the ZnF alpha helices) are indicated.

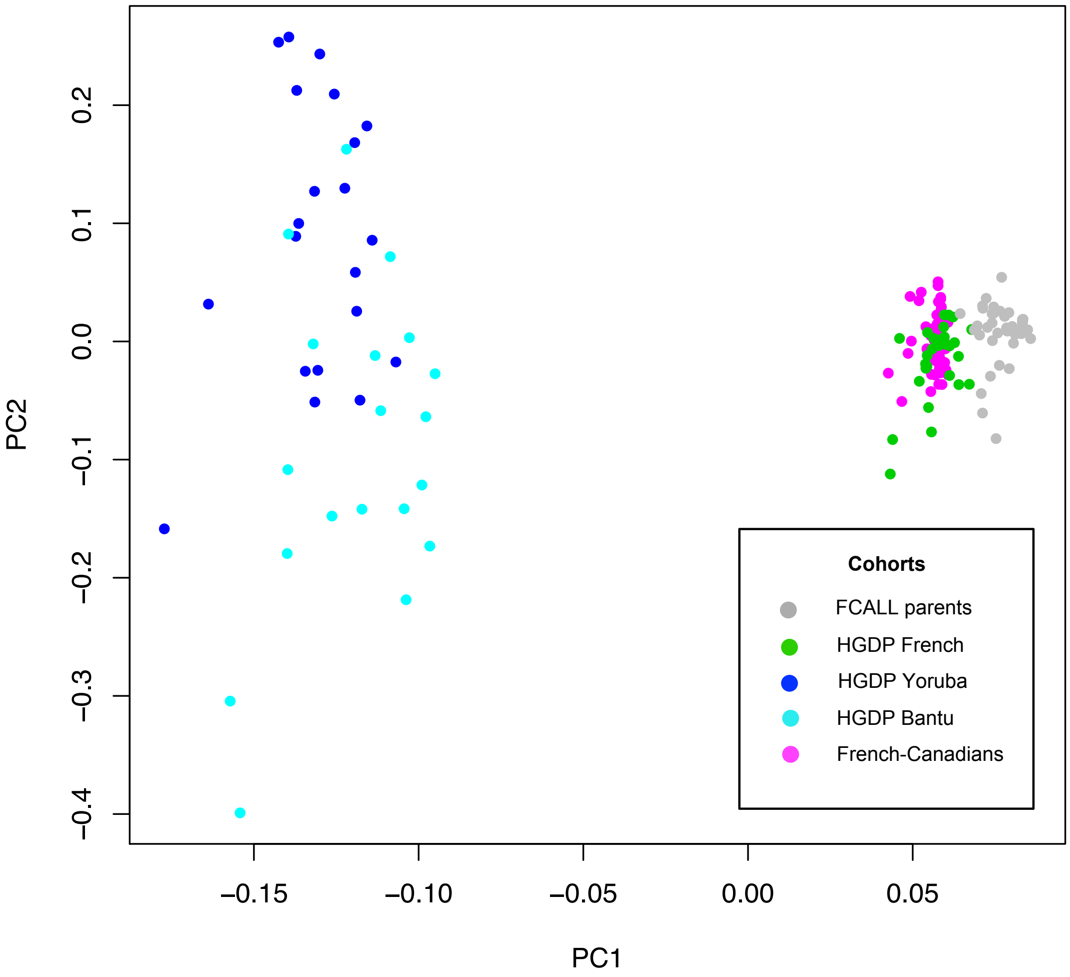


Figure S7. Genetic ancestry of parents from the FCALL cohort.

We performed a Principal Component Analysis of genetic variation using 358 SNPs in common between exome sequencing SNPs from the FCALL parents and genotyped SNPs from HGDP populations and the FC family cohort.

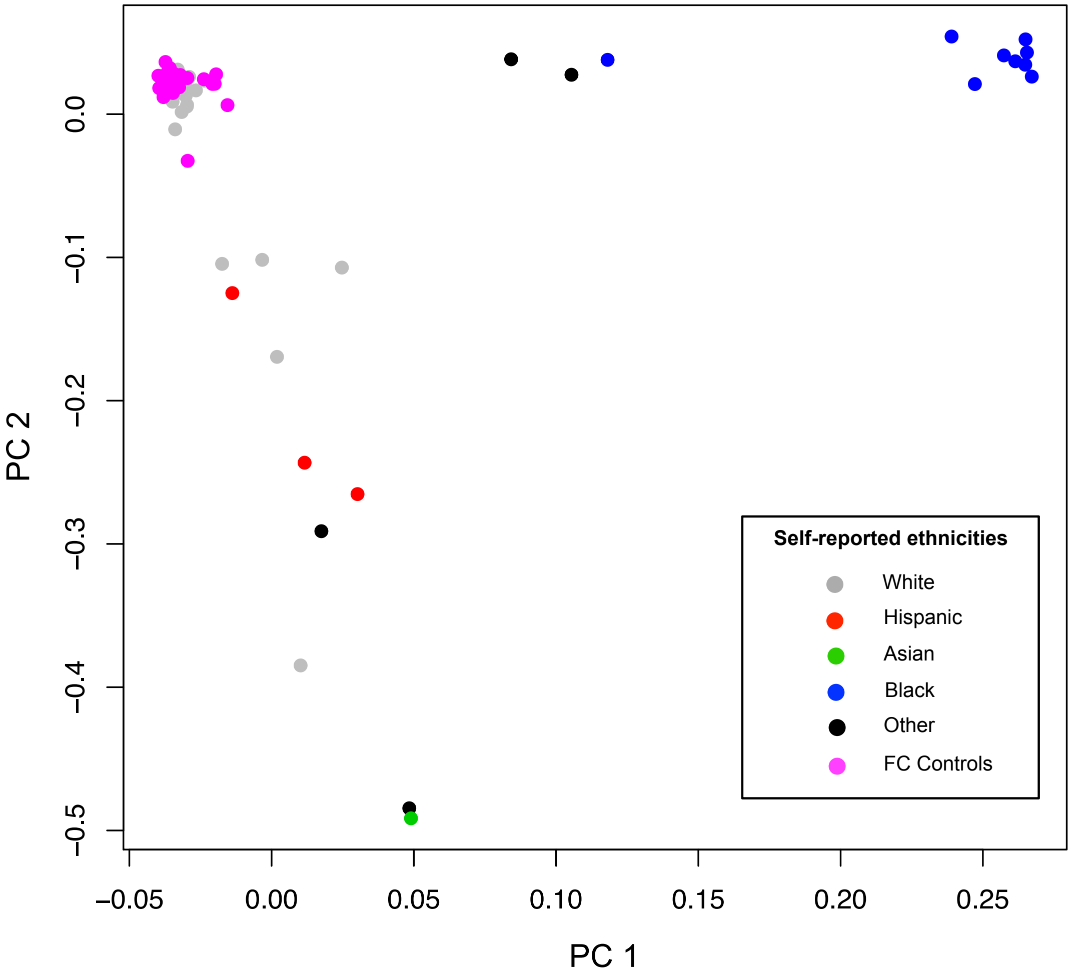
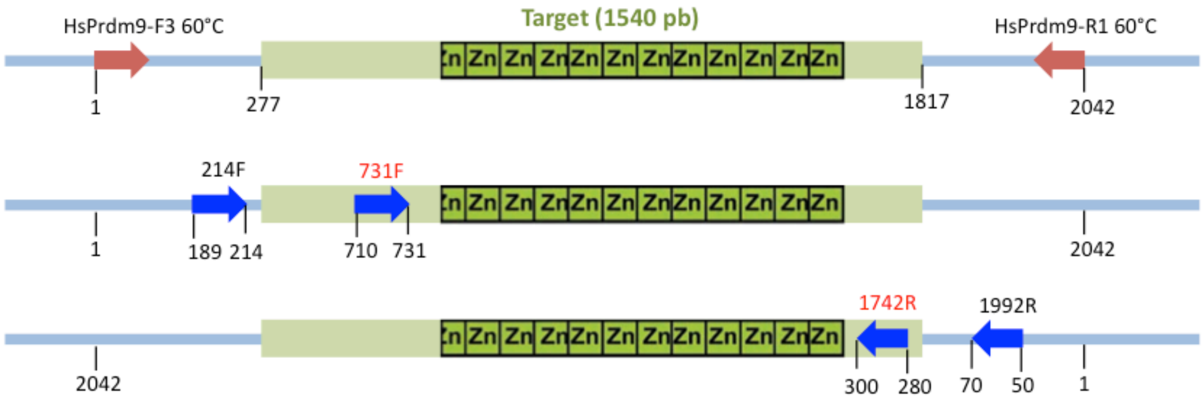


Figure S8. Genetic ancestry of SJDALL patients.

We preformed a Principal Component Analysis of genetic variation using 201 474 SNPs from 61 St Jude patients and 76 French-Canadians controls. The first Principal Component (PC1) separates individuals of European descent from individuals from African descent: the 11 individuals showing African ancestry (PC1> 0.05) were removed from analyses.



Primer name	Sequence
HsPrdm9-F3	TGTAAGGAATGACACTGCCCTGA
HsPrdm9-R1	ATGTCCCCCGAACACTTACAGAA
214F	TGATTGTTTCTTCATTTGATCTTCA
731F	TGGAGAGTGTGGACAAGGTTT
1742R	AGCAGAGGCTTGACCTATCG
1992R	GTCATGAAAGTGGCGGATTT

Figure S9. PCR primers used for amplifying and sequencing PRDM9 ZnF alleles.

The ZnF array in exon 11 of PRDM9 was amplified as described in Baudat et al. 2010 [21]. Sanger sequencing was performed with primers 214F, 731F, 1742R and 1992R.

Figure S10. Chromosomal crossover breakpoints and shared haplotypes in the ALL quartet.

Graphical view of all paternal (blue) and maternal (red) crossover breakpoints that occurred in affected children inferred based on exome and genotyping SNPs (Supplementary Methods). When the lines are a part, the two brothers copied different parental chromosomes and when they are close together, they copied the same parental chromosome and share the same haplotype. White spaces represent regions were no informative markers were available. Small dots between parental tracks represent single markers that caused double crossover events and are likely to be SNP calling errors[6]. For small double crossovers, occurring within ≤50Kb and resulting from more than one marker, we indicated the number of informative markers separating them. Small double recombinants are likely to be false positive or reflect gene conversion events (see Supplementary Methods for details on checks performed in order to control for false positive breakpoints). The y-axis shows the position on the chromosome in tens of Mb. (*next page*)

Figure S10 (*continued*)

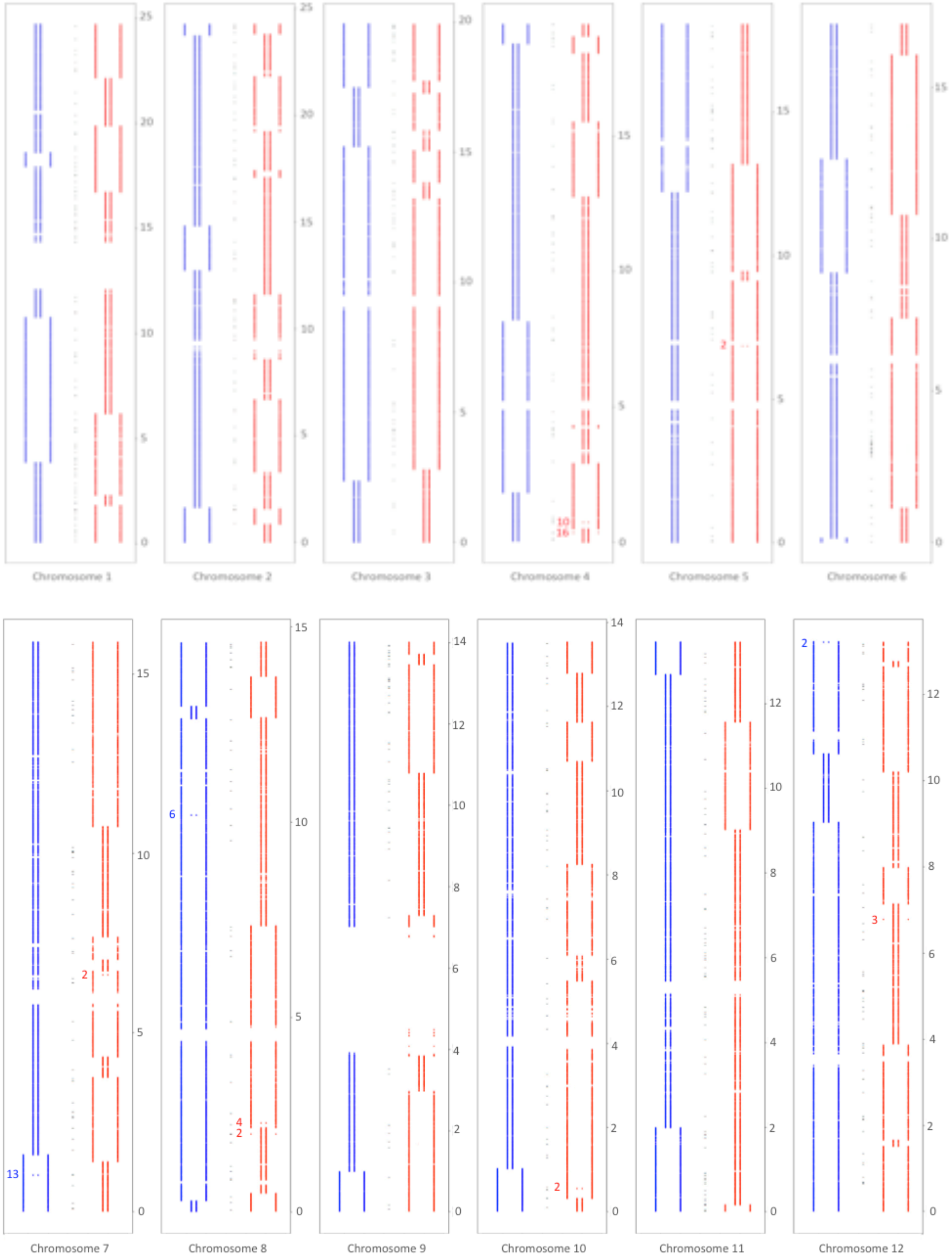
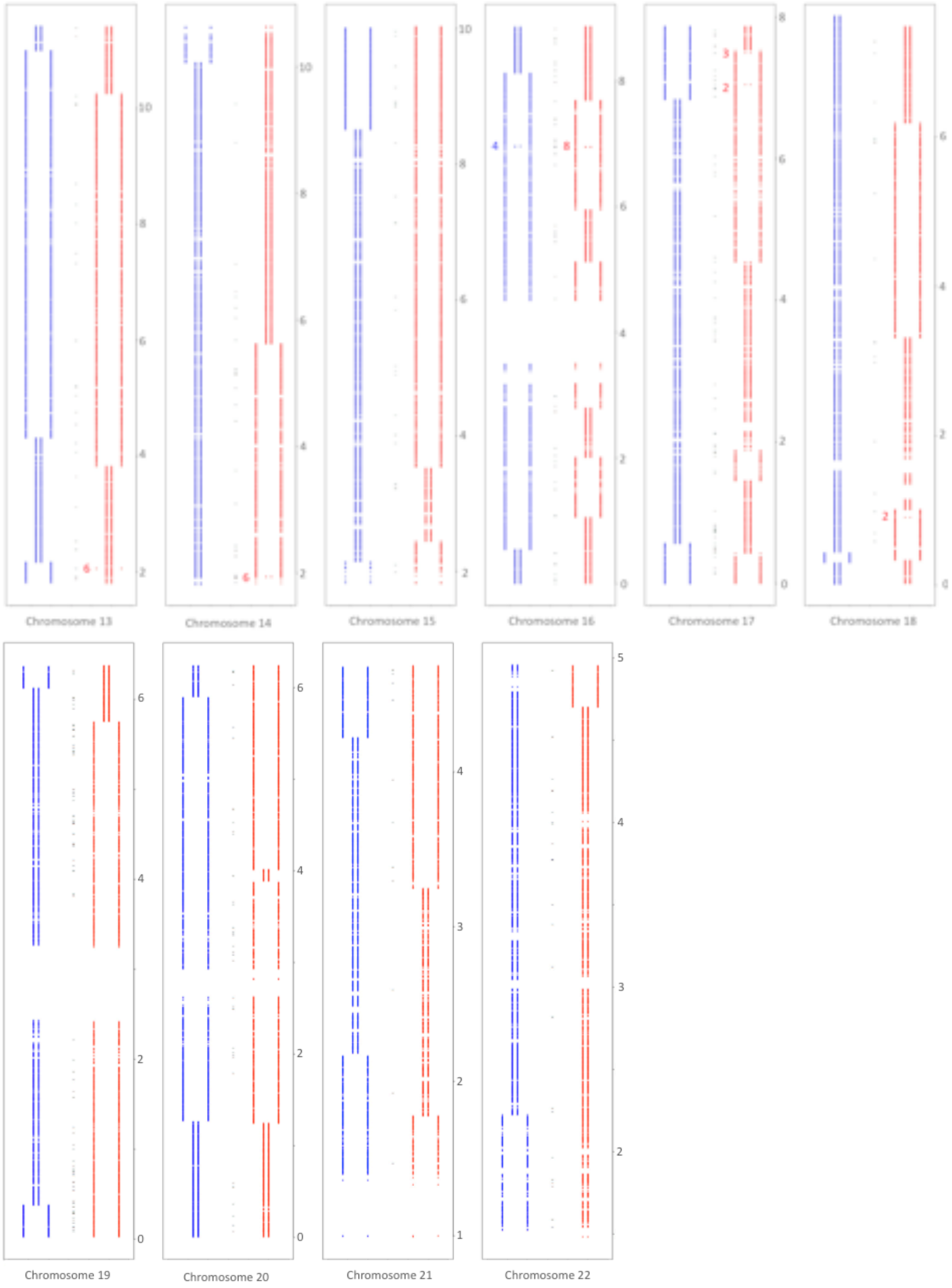


Figure S10 (*continued*)



Chromosome 13 Chromosome 14 Chromosome 15 Chromosome 16 Chromosome 17 Chromosome 18

Chromosome 19 Chromosome 20 Chromosome 21 Chromosome 22

REFERENCES

1. Li B, Abecasis G (2011) Polymutt: a tool for calling polymorphism and de novo mutations. http://genomesphumichedu/wiki/Polymutt.
2. Cartwright RA, Hussin J, Keebler JEM, Stone EA, Awadalla P (2011) A family-based probabilistic method for capturing de novo mutations from high- throughput short-read sequencing data. Statistical Applications in Genetics and Molecular Biology (in press).
3. Conrad DF, Keebler JE, DePristo MA, Lindsay SJ, Zhang Y, et al. (2011) Variation in genome-wide mutation rates within and between human families. Nat Genet 43: 712-714.
4. Phillips RB (1978) Pericentric inversions inv(2)(p11q13) and inv(2)(p13q11) in 2 unrelated families. J Med Genet 15: 388-390.
5. Srebniak M, Wawrzkiewicz A, Wiczkowski A, Kazmierczak W, Olejek A (2004) Subfertile couple with inv(2),inv(9) and 16qh+. J Appl Genet 45: 477-479.
6. Hussin J, Roy-Gagnon MH, Gendron R, Andelfinger G, Awadalla P (2011) Age-dependent recombination rates in human pedigrees. PLoS Genet 7: e1002251.
7. Doggett NA, Xie G, Meincke LJ, Sutherland RD, Mundt MO, et al. (2006) A 360-kb interchromosomal duplication of the human HYDIN locus. Genomics 88: 762-771.
8. Berg IL, Neumann R, Lam KW, Sarbajna S, Odenthal-Hesse L, et al. (2010) PRDM9 variation strongly influences recombination hot-spot activity and meiotic instability in humans. Nat Genet 42: 859-863.
9. Ng PC, Henikoff S (2003) SIFT: Predicting amino acid changes that affect protein function. Nucleic Acids Res 31: 3812-3814.
10. Casals F, Hodgkinson A, Idaghdour Y, Hussin J, Bruat V, et al. (In review) Whole-exome sequencing in 109 individuals reveals an excess of rare and functional variants in the French-Canadian population.
11. (2010) A map of human genome variation from population-scale sequencing. Nature 467: 1061-1073.
12. Hinch AG, Tandon A, Patterson N, Song Y, Rohland N, et al. (2011) The landscape of recombination in African Americans. Nature 476: 170-175.
13. Price AL, Patterson NJ, Plenge RM, Weinblatt ME, Shadick NA, et al. (2006) Principal components analysis corrects for stratification in genome-wide association studies. Nat Genet 38: 904-909.
14. Idaghdour Y, Czika W, Shianna KV, Lee SH, Visscher PM, et al. (2010) Geographical genomics of human leukocyte gene expression variation in southern Morocco. Nat Genet 42: 62-67.
15. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MA, et al. (2007) PLINK: a tool set for whole-genome association and population-based linkage analyses. Am J Hum Genet 81: 559-575.
16. Freeman GH, Halton JH (1951) Note on an exact treatment of contingency, goodness of fit and other problems of significance. Biometrika 38: 141-149.
17. Kong F, Zhu J, Wu J, Peng J, Wang Y, et al. (2011) dbCRID: a database of chromosomal rearrangements in human diseases. Nucleic Acids Res 39: D895-900.
18. Mitelman F, Johansson B, Mertens F (2011) Mitelman Database of Chromosome Aberrations and Gene Fusions in Cancer. http://cgapncinihgov/Chromosomes/Mitelman.
19. Smit A, Hubley, R & Green, P. (1996-2012) RepeatMasker Open-3.0 http://www.repeatmasker.org.
20. Benson G (1999) Tandem repeats finder: a program to analyze DNA sequences. Nucleic Acids Res 27: 573-580.

21. Baudat F, Buard J, Grey C, Fledel-Alon A, Ober C, et al. (2010) PRDM9 is a major determinant of meiotic recombination hotspots in humans and mice. Science 327: 836-840.