Nucleotide, Cytogenetic and Expression Impact of the Human Chromosome 8p23.1 Inversion Polymorphism

Nina Bosch[1,2], Marta Morell[1,2], Immaculada Ponsa[1,3,4], Josep Maria Mercader[1,2], Lluís Armengol[1,2,5], Xavier Estivill[1,2,6]*

1 Genetic Causes of Disease Group, Genes and Disease Programme Center for Genomic Regulation (CRG-UPF), Barcelona, Catalonia, Spain, 2 CIBER en Epidemiología y Salud Pública (CIBERESP), Barcelona, Catalonia, Spain, 3 Departament de Biologia Cel·lular, Fisiologia i Immunologia, Facultat de Medicina, Universitat Autònoma de Barcelona, Catalonia, Spain, 4 Institut de Biotecnologia i Biomedicina, Universitat Autònoma de Barcelona, Barcelona, Catalonia, Spain, 5 Quantitative Genomic Medicine Laboratories (qGenomics), Barcelona, Catalonia, Spain, 6 Department of Health and Experimental Life Sciences, Pompeu Fabra University (UPF), Barcelona, Catalonia, Spain

Abstract

Background: The human chromosome 8p23.1 region contains a 3.8–4.5 Mb segment which can be found in different orientations (defined as genomic inversion) among individuals. The identification of single nucleotide polymorphisms (SNPs) tightly linked to the genomic orientation of a given region should be useful to indirectly evaluate the genotypes of large genomic orientations in the individuals.

Results: We have identified 16 SNPs, which are in linkage disequilibrium (LD) with the 8p23.1 inversion as detected by fluorescent in situ hybridization (FISH). The variability of the 8p23.1 orientation in 150 HapMap samples was predicted using this set of SNPs and was verified by FISH in a subset of samples. Four genes (*NEIL2*, *MSRA*, *CTSB* and *BLK*) were found differentially expressed (p<0.0005) according to the orientation of the 8p23.1 region. Finally, we have found variable levels of mosaicism for the orientation of the 8p23.1 as determined by FISH.

Conclusion: By means of dense SNP genotyping of the region, haplotype-based computational analyses and FISH experiments we could infer and verify the orientation status of alleles in the 8p23.1 region by detecting two short haplotype stretches at both ends of the inverted region, which are likely the relic of the chromosome in which the original inversion occurred. Moreover, an impact of 8p23.1 inversion on gene expression levels cannot be ruled out, since four genes from this region have statistically significant different expression levels depending on the inversion status. FISH results in lymphoblastoid cell lines suggest the presence of mosaicism regarding the 8p23.1 inversion.

Introduction

Among the different classes of structural variations, the understanding of the prevalence and spectrum of the inversions in the human genome is still scarce. One of the reasons is that most genome-wide technologies used to discover structural variations are designed to detect gains and losses of genomic material. Only recently, several studies based on fosmid cloning and paired-end sequencing have succeeded in mapping inversion breakpoints in a genome-wide fashion [1–4]. As revealed by these and other studies, which include comparing genomes assembled from different individuals and targeted analyses, the total number of inversion regions in the human genome is now close to 500 (http://projects.tcag.ca/variation/), but this figure will likely increase with deep sequencing data coming from the 1000 genomes project (www.1000genomes.org).

An indirect approach to delineate human inversions, which was first described on *Drosophila* studies [5], is the characterization of

extended blocks of linkage disequilibrium (LD) that are created due to the lack of recombination in heterozygous individuals [6]. Following this criterion, haplotype subgroups can be defined in polymorphic inversions because different alleles are maintained in the different orientations. Such is the case for the polymorphic inversion on chromosome 17q21.31, for which one of the haplotypes associated to one of the conformations has been found to be under positive selection in Europeans [7,8].

Human chromosome 8p23.1 encompasses a 3.8–4.5 Mb polymorphic segment flanked by two large blocks of segmental duplications (SDs). The whole region extends up to 6.5 Mb, which includes the SDs and contains at least 50 genes. This inversion was first described to have a frequency of 26% in the European [9] and 27% in the Japanese [10] populations, assuming that the assembly of the reference sequence corresponds to the non-inverted conformation. However, different studies have found an increased frequency of the inversion, around 60%, in populations of European

ancestry, indicating that the human reference assembly corresponds to the minor allelic orientation of the region [4,11,12]. Moreover, the 8p23.1 region is an intricate DNA segment flanked by two large sets of SDs, named REPP (proximal) and REPD (distal), containing several genes that vary in copy number, such as defensins or *FAM90A* [13–22]. The 8p23.1 region has recently been reported to also contain a high concentration of structural variants in fosmid end-sequencing experiments [4]. Thus, the sizes of the flanking SDs are variable and not fully sequenced (both are defined as gaps in the assembled genome sequences), and it seems plausible that this variability at the SDs could play a key role in the distinct rearrangements affecting the 8p23.1 region.

Although initially considered a neutral polymorphism, the 8p23.1 inversion has been found in mothers of children with an associated phenotype suffering from different rearrangements involving the 8p23.1 region. This phenomenon has also been described in the parents of children carrying other genomic disorders [23] such as Hunter syndrome [24], Williams-Beuren syndrome [25,26], and Prader-Willi or Angelman syndromes [27].

With the aim to identify single nucleotide polymorphisms (SNPs) that could be used as surrogate markers for the 8p23.1 inversion, we performed a genotyping analysis of six individuals that were analyzed by fluorescent in situ hybridization (FISH) and were found to be homozygous for the 8p23.1 inversion (with respect to the reference genome sequence). We have identified two small tracts of SNP in homozygosity, which perfectly correlate with the inversion status of this genomic region. Thus, we predicted the genotype of the inversion for 150 HapMap individuals using the genotype of these SNPs, and we confirmed our predictions by FISH analysis in a subset of these samples. Interestingly, we have detected a variable degree of mosaicism with respect to the inversion within all the samples analyzed by FISH. This phenomenon suggests that the region is mitotically unstable. Finally, we have explored the effect of the inversion rearrangement on 8p23.1 gene expression levels and have found that four genes from this region have statistically significant different expression levels depending on the inversion status.

Methods

Genotype Data and the Delineation of Homozygosity Tracts

A deep SNP genotyping analysis of the 8p23.1 region was performed using the HumanCNV370-Duo chip from Illumina in a sample of six Spanish individuals of the general population that were found to be homozygous for the 8p23.1 inversion FISH, and the HapMap sample NA10861 was used as a control for genotype concordance. Among all markers covered by the array, we focused on 770 SNPs spanning the 8p23.1 region, where two tracts of homozygosity containing 16 SNPs could be defined between 8.5–8.7 Mb and 10.8–11 Mb in the six homozygous inverted individuals. Genotyping was carried out following manufacturer's protocol (Illumina Inc.) at the Barcelona CeGen genotyping center. Illumina's HumanCNV370-Duo chip includes common variation described for the CEU, CHB/JPT, and YRI populations based on HapMap Phase I and II data, and also contains probes that are enriched in CNV and SDs.

Genotype data of 210 unrelated HapMap individuals, including 60 parents of 30 trios from Utah residents with ancestry from northern and western Europe (CEU), 60 parents of 30 trio samples from Yoruba in Ibadan, Nigeria (YRI), 45 unrelated Han Chinese from Beijing, China (CHB), and 45 Japanese from Tokyo, Japanese (JPT) individuals, were also used in the analysis. Due to the significant genetic similarity between Chinese and Japanese groups, we pooled CHB and JPT data and denoted these individuals as Asian (ASN) as in other studies [28].

HapMap phase II data was downloaded from the website (HapMap Data Rel23a/phaseII Mar08; www.hapmap.org). Phased haplotypes were used to predict the genotype of the 150 individuals with respect to the 8p23.1 inversion based on the conserved homozygosity tracts previously described. The 150 individuals analyzed included the 60 parents from CEU samples and the 90 ASN individuals.

FISH Analysis

FISH was performed on metaphase chromosomes prepared from lymphoblastoid cell lines obtained from 24 unrelated Spanish individuals using BAC-derived DNA probes. Two BAC clones that fall within the 8p23.1 inversion and that are free of SDs (RP11-399J23 and RP11-589N15) were used for the experiments. DNA from BAC clones was isolated by standard procedures and labeled with Spectrum green for RP11-399J23, and with Spectrum orange for RP11-589N15. Labeled probes were precipitated and resuspended following standard protocols. Metaphase chromosomes and probes were denatured and hybridized overnight at 37°C. Chromosomes were counterstained with 4′, 6-diamino-2-phenylindole (DAPI). Images were analyzed using the Isis software from Metasystems. At least 20 metaphases with clearly interpretable signals on both chromosomes were counted per individual.

Elongated chromosomes were obtained adding 10 μl of bromodeoxyuridine (10 mg/ml) for three hours incubation time, followed by a 75 minutes incubation time with 10 μl of ethidium bromide (10 mg/ml). This treatment was used for individuals LCL 345, LCL 247, LCL 384, LCL 339, LCL 221, LCL 137, LCL 342, LCL 340, LCL 198, LCL 179, LCL 182, LCL 240 and LCL 195.

HapMap cell lines were obtained from Coriell Repositories (Camdem, NJ). FISH analysis following the same procedure was performed in 9 HapMap samples to confirm the predictions for the 8p23.1 inversion based on the described surrogate SNP markers. Individuals NA11992, NA12057 and NA11839 were used for confirmation of the non-inverted orientation of 8p23.1. NA11993, NA06993 and NA11994 were the heterozygous individuals. Finally, the homozygous status of the inversion was confirmed in individuals NA11831, NA12815 and NA12155.

Association Study of 8p23.1 Inversion and Gene Expression Levels

Normalized gene expression values from CEU and CHB/JPT (ASN) were downloaded from the Sanger Genevar webpage (http://www.sanger.ac.uk/humgen/genevar/). Data retrieved by 31 Illumina™ different probes targeting 26 genes located around the 8p23.1 region (Table S2) were used to explore the correlation between gene expression levels and the three different genotypes of the inversion region (non-inverted homozygous, heterozygous, and homozygous inverted). Individuals included in this analysis were the 150 samples corresponding to the CEU and ASN populations where the status of the inversion could be predicted using surrogate markers. The association between gene expression values and the genotype for the inversion was tested for different genetic models. P-values were derived from likelihood ratio tests, and a significance level of 5% (two sided) was used for the analyses. All these analyses were performed using the SNPassoc R package [29].

Results

Frequency of 8p23.1 Inversion in the Spanish Population

To determine the frequency of 8p23.1 inversion in the Spanish population, we genotyped a total of 24 Spanish control individuals

by FISH. We used BAC clones RP11-399J23 and RP11-589N15, localized at each end of the 3.8 Mb inversion on chromosome 8 and outside the SDs flanking 8p23.1 [10]. Based on the human reference assembly Human NCBI Build 36 we consider the telomere-to-centromere orientation RP11-399J23 (Green) and RP11-589N15 (Red) as the non-inverted conformation. The hybridizations were performed using several slides per individual. For 13 of the 24 individuals we obtained elongated and non-elongated metaphase chromosomes.

From the 24 individuals that were genotyped, 50% (N = 12) were heterozygous for the inversion, 29% (N = 7) homozygous for the inverted conformation, and 21% (N = 5) homozygous for the Build 36 genome assembly conformation (Figure 1). Thus, the 8p23.1 inversion was found in ~80% of control individuals and consequently the reference assembly exemplifies the less frequent orientation of this genomic segment in the Spanish population.

The study also revealed that although a predominant genotype could be established for each studied sample, metaphases
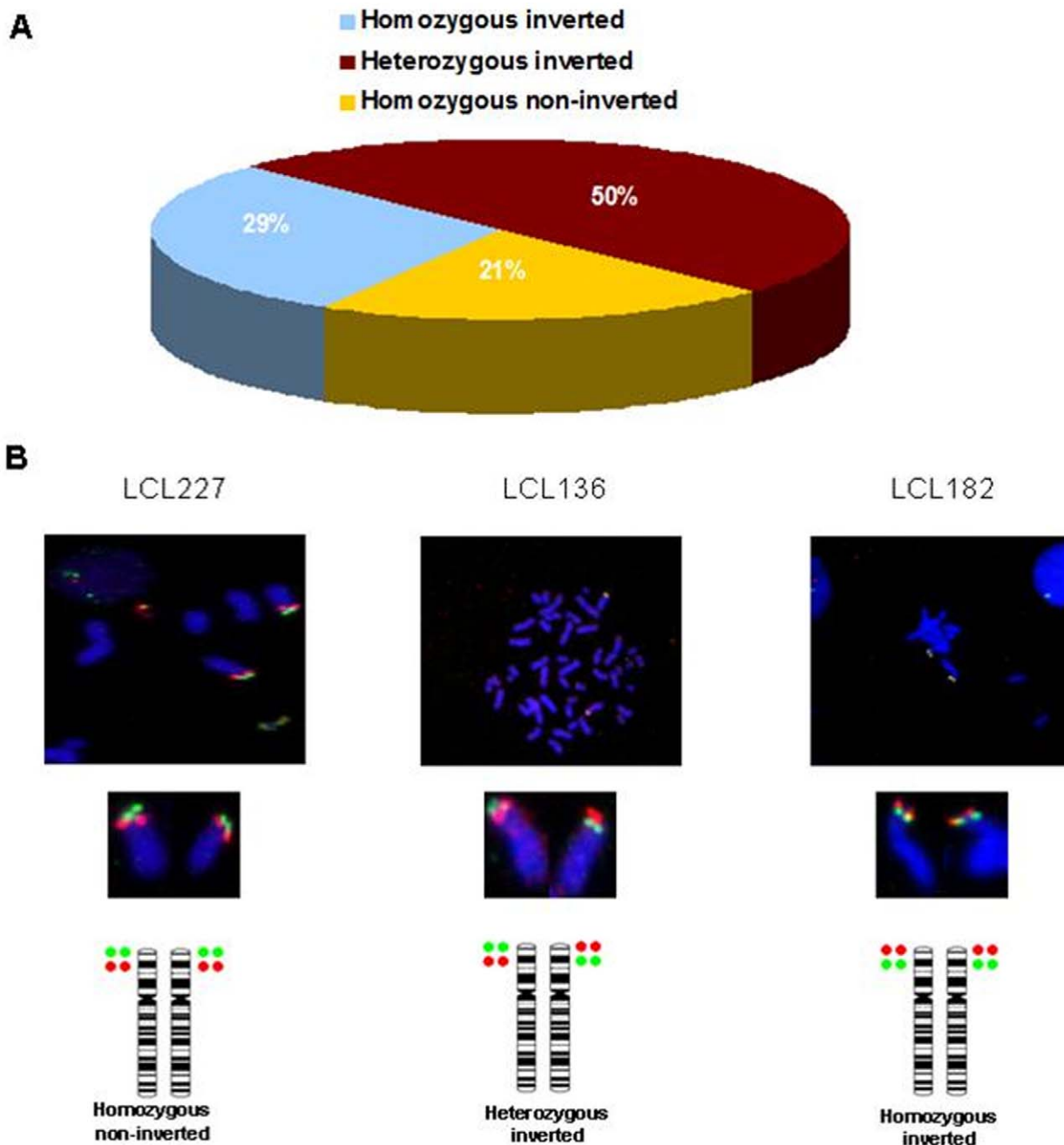
Figure 1. FISH analyses for the human chromosome 8p23.1 inversion in control samples. DNA probes were generated from BAC clones RP11-399J23 (Green) and RP11-589N15 (Red). A Frequencies of each of the three genotypes for the 8p23.1 inversion obtained in 24 Spanish control samples. B Metaphase FISH of three of these samples, LCL227 is "Build-36-non-inverted"; LCL136 is heterozygous for the 8p23.1 inversion, and LCL182 corresponds to homozygous "Build36-inverted".
doi:10.1371/journal.pone.0008269.g001

corresponding to different genotypes were observed in the 24 individuals analyzed (a subset of these samples is shown on Table 1). Therefore, from the 12 samples heterozygous for the inversion, the percentage of metaphases homozygous for either the normal or the inverted conformation ranged from 14% to 38%. Among the seven individuals homozygous for the inversion, 4% to 32% of the metaphases resulted homozygous for the non-inverted conformation or heterozygous for the inversion. Finally, among the 5 individuals that were found to be homozygous for the non-inverted conformation, up to 18% of the metaphases had a heterozygous genotype.

These results suggest instability of this genomic region, probably due to intrachromosomal recombination between the two sets of large SDs at both 8p23.1 extremes of the inversion polymorphism region. The data also highlights the difficulty when interpreting the genotype for the 8p23.1 inversion using FISH. According to our results, a predominant genotype can be assigned to each sample, but somatic mosaicism seems to be common, at least in samples derived from lymphoblastoid cell lines.

Homozygosity Blocks in 8p23.1 Inverted Alleles

If we consider that haplotype subgroups are created by suppression of recombination in the 8p23.1 inverted fragment, we should be able to describe surrogate markers for 8p23.1 inversion. To disclose the surrogate markers, six of the Spanish control individuals, that were found to be homozygous for the 8p23.1 inversion by FISH analysis, were genotyped using the Illumina's HumanCNV370-Duo chip. Homozygous-inverted individuals were chosen as the presence of homozygosity tracts simplifies haplotype estimation and avoids phasing errors. By this procedure we were able to delineate two different homozygosity tracts composed by a total of 16 SNPs within the inverted region, in the proximity of the REPP and REPD duplicons. The first homozygosity tract expands ~172 kb and contains 8 SNPs present in the HumanCNV370-Duo chip and corresponds to the "CGTCGAGG" haplotype in all 6 individuals (Table 2). This genetic "signature" is located at chromosome position 8.5 Mb, close to the REPD distal SDs that flank the 8p23.1 segment. A second block of 8 homozygous SNPs spans ~181 kb, and in this case the conserved haplotype is

"TCACGAGA" (Table 2) and lays at 10.8 Mb, close to REPP, the proximal set of SDs on 8p23.1 (Figure 2A).

Assuming that these 16 SNPs remain as a "signature" which resulted from the recurrent recombination processes that lead to the inversion of the 8p23.1 region, we postulate that these two 8-SNP haplotypes can be used as proxies for the 8p23.1 inversion. We downloaded the phased haplotypes from the 210 HapMap samples (only the parents from the CEU and YRI trios were used) to predict the status of the inversion in these samples. In the population of European ancestry we found that 50% of the individuals are heterozygous for the inversion, 8% are homozygous, and 42% do not have the inverted allele (Figure 2B). Among the population of Asiatic origin the presence of the inversion is extremely high, with 37% of the samples being heterozygous and 48% homozygous for the 8p23.1 inversion (Figure 2). Finally, these two 8-SNP haplotypes were not detected in the YRI samples. This is probably because our predictions are based on Caucasian ancestry individuals, and YRI samples might have different LD patterns and the inversion could not be tagged with these SNPs. Therefore, the profiles segregating with the inversion could not be assessed by our approach, although it has recently been reported that ~60–76% of YRI individuals have the inversion [4,12].

Confirmation of the Predictions for 8p23.1 Inversion by FISH Analysis

In order to confirm the validity of the two 8-SNP haplotypes proxies for the presence of 8p23.1 inversion in Caucasian samples, we have genotyped the inversion by FISH in a subset of HapMap CEU samples. We choose three individuals predicted to be homozygous for the 8p23.1 inversion (NA11831, NA12815 and NA12155), three heterozygous (NA11993, NA06993 and NA1994) and three homozygous for the non-inverted status (NA11992, NA12057 and NA11839) (Figure S1). The predicted genotype was confirmed in the 9 samples, although again some degree of mosaicism was detected but with a predominant genotype (Table 3). These findings are in agreement with the large number of discordant fosmids per individual found in HapMap samples by Kidd et al. (2008), in the 8p23.1 region (Figure S2 and Table S1). Although the discordant fosmid clones could be due to a mapping artifact due to the presence of the SDs

Table 1. FISH analysis of 8p23.1 inversion in Spanish control individuals.

Samples	METAPHASES		
	Homozygous non-inverted	Heterozygous	Homozygous inverted
LCL159	87%	10%	3%
LCL183	82%	18%	-
LCL198	92%	8%	-
LCL146	14%	62%	24%
LCL161	19%	75%	6%
LCL184	13%	84%	3%
LCL241	5%	86%	9%
LCL247	20%	70%	10%
LCL194	4%	8%	88%
LCL339	-	22%	78%

Percentages of each of the three possible genotypes observed within each Spanish individual in a subset of 10 representative samples from the 24 individuals analyzed. The predominant genotype is represented as a grey box.
doi:10.1371/journal.pone.0008269.t001

Table 2. Tracts of homozygosity in 8p23.1 region extracted from the whole genome scan data performed in homozygous inverted individuals.

REPD		REPP	
SNP	Homozygous inverted haplotype	SNP	Homozygous inverted haplotype
rs17627505	C	rs1178061	T
rs10503393	G	rs1178247	C
rs2428	T	rs3885690	A
rs11774860	C	rs2409691	C
rs3827811	G	rs13266785	G
rs17154769	A	rs10282848	A
rs1876836	G	rs10503417	G
rs1039916	G	rs2409719	A

The 8 SNPs that serve as surrogate markers close to the REPD duplicons are shown on the left. The 8 SNPs that serve as surrogate markers close to the REPP duplicons are shown in the right.
doi:10.1371/journal.pone.0008269.t002
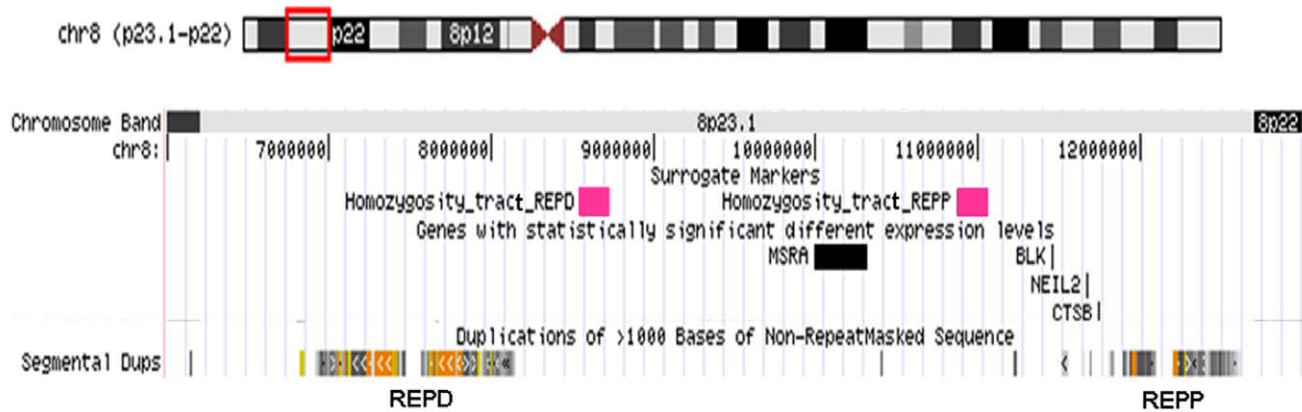
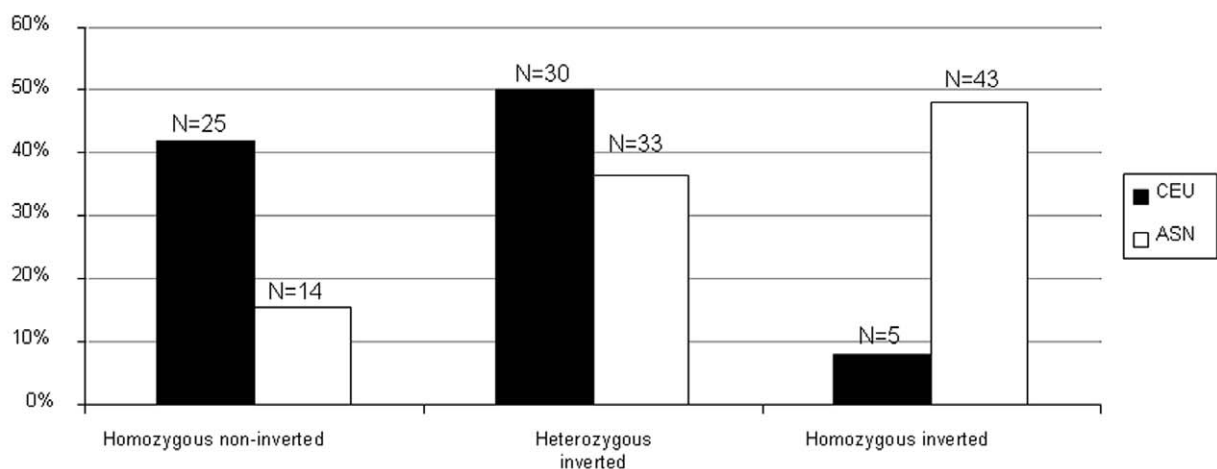Figure 2. Scheme of the localization of the homozygosity tracts used as surrogate markers to predict the status of the human chromosome 8p23.1 inversion in HapMap samples (Homozygsity tract_REPD and Homozygosity tract_REPP). Genes differentially expressed on the association analysis and SDs are also depicted. B Frequencies predicted for the 8p23.1 inversion in CEU (white bars) and ASN (black bars) populations based on haplotypes for two blocks of 8 SNPs inside the region of the polymorphic inversion.
doi:10.1371/journal.pone.0008269.g002

Table 3. FISH analysis in HapMap samples.

HapMap sample	Homozygous non-inverted	Heterozygous	Homozygous inverted
NA11992	95%	5%	-
NA12057	67%	33%	-
NA11839	97%	3%	-
NA11993	-	96%	4%
NA06933	-	91%	9%
NA11994	-	75%	25%
NA11831	-	38%	62%
NA12815	-	35%	65%
NA12155	-	46%	54%

DNA probes were made from BAC clones RP11-399J23 (Green) and RP11-589N15 (Red).
doi:10.1371/journal.pone.0008269.t003

at both ends of the inverted region, several of the fosmid end-sequences are clearly different, unambiguously mapping to the regions reported by Kidd et al. (2008). The results reported here and the fosmid end-sequencing data suggest mitotic rearrangements leading to different population of cells regarding the orientation of the 8p23.1 polymorphic segment.

Our FISH analysis showed that in individuals NA11992, NA12057 and NA11839 the predominant genotype is the non-inverted status. Individuals NA11993, NA06933 and NA11994 have most metaphases heterozygous for the 8p23.1 inversion. In samples NA11831, NA12815 and NA12155 the most observed genotype is the homozygous inverted. We propose that the alternative conformations for the 8p23.1 region arise by intrachromosomal homologous recombination at mitosis and that these two 8-SNP haplotypes are reliable markers to infer the genomic structure of the 8p23.1 region.

Gene Expression Analysis of 8p23.1 Genes in HapMap Populations

Another aim of this study was to investigate if the inverted conformation has an effect on the expression of the genes contained

in the 8p23.1 region. For this purpose we made use of gene expression levels in lymphoblastoid cells of HapMap individuals [30] and we performed an association study with respect to the inversion status. We used gene expression values from 26 genes (Table S2) located around and within the 8p23.1 region and we searched for any association between gene expression levels and the presence of the inversion. These analyses were carried out with the SNPassoc R package on the CEU and ASN populations, where the two 8-SNP haplotypes signature predicts the inversion status.

Interestingly, four genes, *NEIL2* (p = 0.0003), *MSRA* (p = 0.001), *CTSB* (p = 3.46×10^{-5}) and *BLK* (p = 2.08×10^{-5}) exhibited statistically significant expression level differences after Bonferroni correction for multiple testing (p<0.001) (Figure 2A). Box-plots of the different expression levels for the four genes are shown on figure 3. The model of inheritance under which the differential expression was detected was with the additive model, with the exception of *NEIL2*, which follows a dominant inheritance pattern. Under the additive model the effect of the inversion on gene expression is higher in individuals without the inversion than in the heterozygous samples, than in turn show higher levels that the homozigoulsy inverted individuals for *CTSB*, and the contrary is observed for *BLK* and *MSRA* genes. This data suggest that the inverted conformation has some effect on the expression of the genes embedded in the inverted fragment.

Discussion

Although not as abundant as other structural variations, inversions are common genomic variants in the human genome [1,3,4,31]. A recent study has described that there are at least 35 recurrent inversions that are visible by light microscopy, six of which are known to disrupt a gene [32]. An important issue emerging from these studies is the mechanism underlying such common variants. It is well known that the presence of highly identical genomic sequences (low copy repeats or SDs), arranged in opposite orientations, flanking certain regions of the genome, predispose to inversion events by non-allelic homologous recombination (NAHR). It has also been demonstrated that little requirement for long, identical homology blocks between paralogous DNA fragments is needed to produce exchanges by ectopic recombination [33]. What it is not so often taken into consideration is the presence of mitotic recombination leading to mosaicism in regions showing this type of genomic architecture. Mitotic recombination is an important mechanism under study that can complicate the interpretation of the rearrangements mediated by complex SDs, specially when the material used for study are transformed lymphoblastoid cell lines which could exhibit a high mitotic ratio. This may be the reason why several studies on the 8p23.1 region have lead to some discrepancies [4,12]. In this regard, a recent study on mouse stem cells has shown that CNVs involving gains or losses of millions of base pairs occur frequently during mitotic cell culture divisions, supporting the idea that somatic tissues are composed of variants of the mitotic genome [34]. Similarly, a study of several tissues from different individuals has shown that somatic mosaicism is present for a subset of CNVs [35]. Finally, the analysis of concordant and discordant monozygotic twins has revealed mosaicism for CNVs, indicating somatic mosaicism for some CNV regions [36].
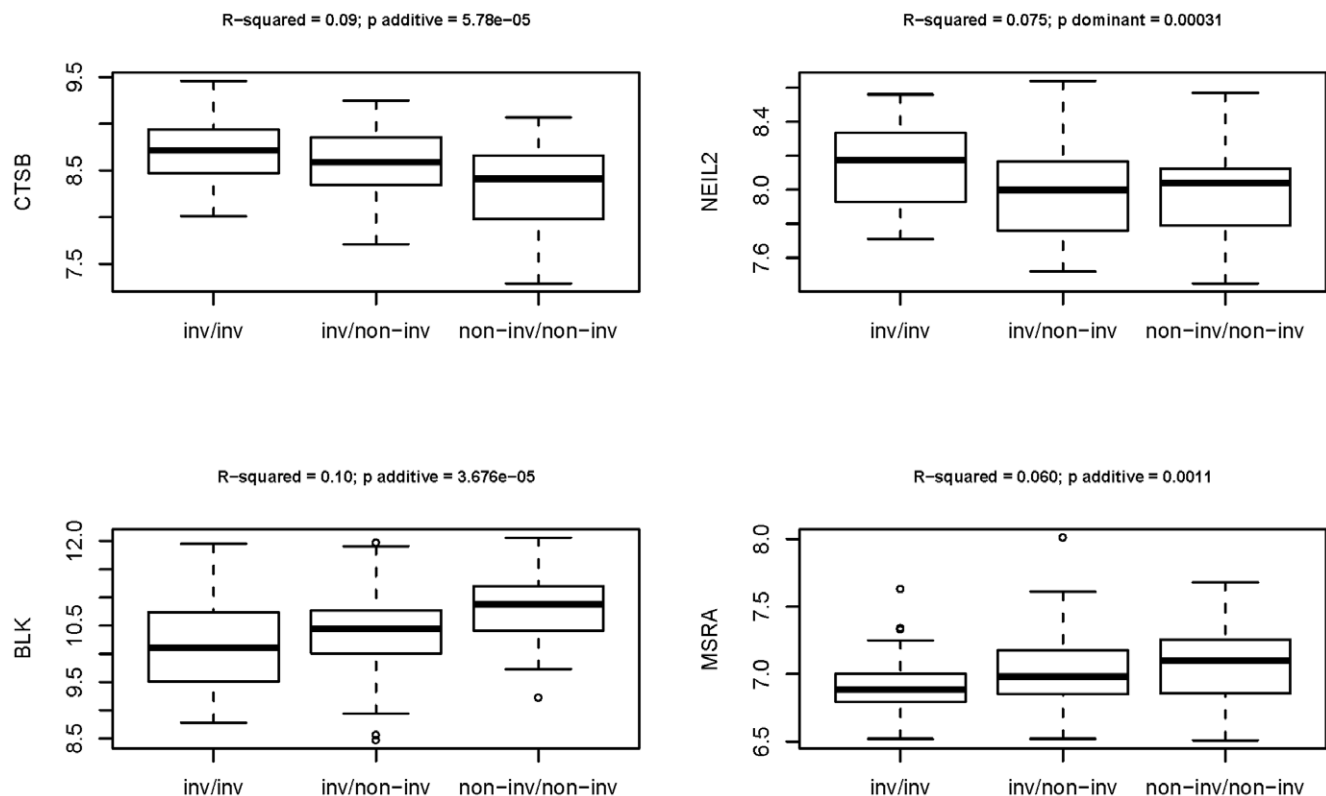
Figure 3. Box-plot representation for the differentially expressed CTSB, NEIL2, BLK and MSRA genes, according to the predicted inversion genotypes. The underlying distributions of expression levels for each of the three possible inversion genotypes: homozygous inverted (inv/inv), heterozygous inverted (inv/non-inv) and homozygous non-inverted (non-inv/non-inv) are shown. On the top of each panel R-squared values derived from general linear model regressions and p-values under the most significant model of inheritance are represented.
doi:10.1371/journal.pone.0008269.g003

Our FISH results in lymphoblastoid cells prepared from Spanish control individuals as well as HapMap lymphoblastoid cell lines, points out the presence of a wide range of mosaicism regarding the 8p23.1 inversion that could have arisen during the lympholastoid cell cultures. Although previous data on this subject suggest that the effects of lymphoblastoid cells generation, as well as the effects of passage, on genotype and genetic architecture are minimal [37], it has been reported that somatic events occur during cell culture and that the mitotic events are responsible for these changes [38,39]. It is clear that the locus analyzed here has a very complex genomic architecture due to the complex SDs it contains. Despite somatic rearrangements are a priori expected events due to the extension and high level of identity between the SDs flanking the 8p23.1 inversion polymorphism, little is known about the incidence of these events in the genome. Several studies have suggested that this phenomenon, exemplified in the case of the alpha-globins where somatic deletions encompassing these genes arise by intrachromosomal homologous exchange, is common in blood and sperm [33]. Results from Flores [40] also indicate that some cells within blood samples from normal individuals can undergo genomic rearrangements such as inversions and create genomic structural mosaicism. These findings should expand our current view about the plasticity of the genome and the homogeneity of the genetic material that we characterize in genetic studies at the SNP level.

Another aspect that arises from our study of the 8p23.1 inversion is the high frequency of the inverted allele. The possibility of these results being a misinterpretation of FISH signals due to the scarcely ~3.5 Mb of distance between the two probes it is unlikely, since no differences regarding the levels of mosaicism were observed when the chromosomes were elongated with ethidium bromide. By means of dense SNP genotyping of the region, haplotype-based computational analyses and FISH experiments we could infer and verify the orientation status of alleles in the 8p23.1 region. Chromosomes were initially studied by FISH and surrogate markers for the inversion were identified by analyzing allelic association of 16 SNPs, which delineated two tracts of homozygosity, in a group of individuals with known status for the 8p23.1 inversion. These 16 SNP markers tag the inversion and perfectly correlate in European and Asian ancestry HapMap samples. These 16 markers were initially selected from a sample of Spanish individuals, but they were also useful to predict the inversion in individuals of CEU and ASN origin. Thus, these SNPs can be used in CEU and ASN population to tag the inversion. However, in the YRI population, which exhibits a higher SNP diversity [31], the absence of LD at these 16 surrogate markers does not allow to use them to predict the inversion in these subjects. This argues in favor of several origins for this genomic rearrangement or of an ancient origin of it. The fact that the SNP haplotype "signature" is not maintained in the YRI population is probably due to the overall higher genetic diversity carried by individuals of African descent.

We should also take into consideration that the number of SNPs found in LD with the inversion is subjected to the polymorphisms that share in common the two resources that have been used, the Illumina's HumanCNV370-Duo chip for the Spanish individuals in first instance, and subsequently the genotyping data present in the HapMap database. Thus, the 16 SNP "signature" that we have identified is based on SNPs shared by both resources, meanwhile tracts of homozygosity comprising as much as ~100 SNPs in individuals predicted to be inverted in both populations (CEU and ASN), do overlap with the distal and proximal homozygosity tracts that we describe when we only consider the high coverage data present on HapMap phase III (data not shown). These observations do support the presence of large homozygosity tracts that are in LD with the inverted conformation of 8p23.1 region and the reliability of these 16 surrogate markers.

The analysis performed on HapMap individuals revealed that 37% of ASN population and 50% of the CEU samples were heterozygous for the inverted allele. This frequency is similar to the findings in the Spanish controls genotyped by FISH, where 50% are heterozygous for the inversion, and it is in accordance with previous published results [4,12]. Regarding the homozygous status of the inversion, this is present in 48% of the ASN HapMap individuals, in 8% of the CEU HapMap samples and in 29% of Spanish control samples.

The fact that there is not a long stretch of SNPs along the ~3.8 Mb region that is flanked by the SDs that involve this polymorphic inversion, and only a region of about 170 kb on both sides showing homozygosity for the SNPs, indicates that the inversion has likely been generated many generations ago, leading to many recombination events within the region. It is unknown if these stretches of homozygosity are just relics of the ancestral haplotypes in which the inversion arose or if they might have a role in facilitating its recurrence. The results presented here indicate that the recurrence of the 8p23.1 inversion is much more frequent than previously reported [9], at least in an European population, and that the current human genome reference assembly (Build 36) corresponds to the less common orientation of the 8p23.1 region, as it has been reported for several other structural variants [31].

Although inversions are generally considered as neutral variants regarding phenotype, there are several exceptions where specific genes at the breakpoints are interrupted [41–46]. Moreover, little is known about the effects on the regulation of the transcription of the genes close to the breakpoints where inversions occur. To investigate the possible effects on the genes of 8p23.1 region on the allelic variation of the inverted and non-inverted forms, we performed and association study of gene expression levels and the genotype of the inversion in 150 HapMap individuals. We found four genes (NEIL2, MSRA, CTSB and BLK) that show statistically significant different expression levels (p<0.0005). Two of these genes, NEIL2 and MSRA, are related to repair of oxidative damage [47,48], and CTSB has been suggested to play a role in Alzheimer's disease [49]. To which extent these gene expression differences can influence the function of the respective proteins and if they are directly related to the inversion of the region remains to be proved. We cannot exclude that the degree of mosaicism present in the lymphoblastoid cell lines can modulate the overall tissue-specific expression levels of these genes. In addition we still have no evidences on how this inversion accompanied by any degree of mosaicism can affect the regulation of the transcription of the genes contained in the region. *Another possibility would be that these linkage disequilibrium blocks may contain regulatory variants which are influencing the expression level of these genes.*

In summary, using dense SNP genotyping analysis of the chromosome 8p23.1 region in homozygous inverted samples previously genotyped by FISH, we have been able to describe surrogate markers that tag the 8p23.1 inversion in CEU and ASN populations. Moreover, among the 26 genes we analyzed we have observed gene expression differences in four genes depending on the alternative genomic conformation of the region. We also highlight the presence of mosaicism regarding the inversion in most of the individuals genotyped by FISH. This mosaicism will agree with the different types of discordant fosmid clones with an inverted orientation detected for this region in the analysis performed by Kidd et al. [30]. We postulate that this is a phenomenon that could also involve other regions of the genome

flanked by complex or high level identity SDs and L1 transposons, as previously described [35,50,51]. It is currently unknown if the genomic isoforms and their variable levels of expression have consequences at the phenotypic and functional levels of the individual, deserving future investigation.

Supporting Information

Figure S1 FISH analysis of the human chromosome 8p23.1 inversion in HapMap samples. Metaphase FISH of three HapMap individuals, NA12057 as an example of non-inverted individual; NA06993 is heterozygous for the 8p23.1 inversion and NA12815 corresponds to a homozygous inverted individual.

Found at: doi:10.1371/journal.pone.0008269.s001 (1.20 MB TIF)

Figure S2 Scheme of the different sizes of 8p23.1 inversion within nine human cell lines as a result of fosmid-end cloning and sequencing. For each cell line fosmid library (ABC7 to G248) several end-sequenced fosmid clones were discordant for the mapping of the end sequences, showing an inversion with respect to the reference genome, and also showing different mapping positions. The abundance of each fosmid clone and the approximate sizes of the rearrangements are shown in Supplementary Table 1 (data extracted from Kidd et al., 2008). The filed orange-yellow boxes correspond to the segmental duplications (SD)

that flank and are within the inverted polymorphic region. Nucleotide positions are in megabases (Mb).

Found at: doi:10.1371/journal.pone.0008269.s002 (0.62 MB TIF)

Table S1 Sizes of large inversions detected in human cell lines in fosmid-end cloning and sequencing

Found at: doi:10.1371/journal.pone.0008269.s003 (0.03 MB XLS)

Table S2 8p23.1 Gene expression levels. Genes analyzed in the association study between gene expression levels and the genotype for the 8p23.1 inversion, and their corresponding probes (http://www.sanger.ac.uk/humgen/genevar/).

Found at: doi:10.1371/journal.pone.0008269.s004 (0.03 MB DOC)

Author Contributions

Conceived and designed the experiments: NB XE. Performed the experiments: NB MM IP JMM. Analyzed the data: NB IP JMM. Contributed reagents/materials/analysis tools: NB LA. Wrote the paper: NB.

References

1. Tuzun E, Sharp AJ, Bailey JA, Kaul R, Morrison VA, et al. (2005) Fine-scale structural variation of the human genome. Nat Genet.
2. Eichler EE, Nickerson DA, Altshuler D, Bowcock AM, Brooks LD, et al. (2007) Completing the map of human genetic variation. Nature 447: 161–5.
3. Korbel JO, Urban AE, Grubert F, Du J, Royce TE, et al. (2007) Systematic prediction and validation of breakpoints associated with copy-number variants in the human genome. Proc Natl Acad Sci U S A 104: 10110–5.
4. Antonacci F, Kidd JM, Marques-Bonet T, Ventura M, Siswara P, et al. (2009) Characterization of six human disease-associated inversion polymorphisms. Hum Mol Genet.
5. Navarro A, Barbadilla A, Ruiz A (2000) Effect of inversion polymorphism on the neutral nucleotide variability of linked chromosomal regions in Drosophila. Genetics 155: 685–98.
6. Pritchard JK, Przeworski M (2001) Linkage disequilibrium in humans: models and data. Am J Hum Genet 69: 1–14.
7. Stefansson H, Helgason A, Thorleifsson G, Steinthorsdottir V, Masson G, et al. (2005) A common inversion under selection in Europeans. Nat Genet 37: 129–37.
8. Zody MC, Jiang Z, Fung HC, Antonacci F, Hillier LW, et al. (2008) Evolutionary toggling of the MAPT 17q21.31 inversion region. Nat Genet 40: 1076–83.
9. Giglio S, Broman KW, Matsumoto N, Calvari V, Gimelli G, et al. (2001) Olfactory receptor-gene clusters, genomic-inversion polymorphisms, and common chromosome rearrangements. Am J Hum Genet 68: 874–83.
10. Sugawara H, Harada N, Ida T, Ishida T, Ledbetter DH, et al. (2003) Complex low-copy repeats associated with a common polymorphic inversion at human chromosome 8p23. Genomics 82: 238–44.
11. Chen GK, Slaten E, Ophoff RA, Lange K (2006) Accommodating chromosome inversions in linkage analysis. Am J Hum Genet 79: 238–51.
12. Deng L, Zhang Y, Kang J, Liu T, Zhao H, et al. (2008) An unusual haplotype structure on human chromosome 8p23 derived from the inversion polymorphism. Hum Mutat 29: 1209–16.
13. Hollox EJ, Huffmeier U, Zeeuwen PL, Palla R, Lascorz J, et al. (2008) Psoriasis is associated with increased beta-defensin genomic copy number. Nat Genet 40: 23–5.
14. Bosch N, Escaramis G, Mercader JM, Armengol L, Estivill X (2008) Analysis of the multi-copy gene family FAM90A as a copy number variant in different ethnic backgrounds. Gene 420: 113–7.
15. Ballana E, Gonzalez JR, Bosch N, Estivill X (2007) Inter-population variability of DEFA3 gene absence: correlation with haplotype structure and population variability. BMC Genomics 8: 14.
16. Fellermann K, Stange DE, Schaeffeler E, Schmalzl H, Wehkamp J, et al. (2006) A chromosome 8 gene-cluster polymorphism with low human beta-defensin 2 gene copy number predisposes to Crohn disease of the colon. Am J Hum Genet 79: 439–48.
17. Barber JC, Maloney V, Hollox EJ, Stuke-Sontheimer A, du Bois G, et al. (2005) Duplications and copy number variants of 8p23.1 are cytogenetically indistinguishable but distinct at the molecular level. Eur J Hum Genet.

18. Aldred PM, Hollox EJ, Armour JA (2005) Copy number polymorphism and expression level variation of the human alpha-defensin genes DEFA1 and DEFA3. Hum Mol Genet.
19. Taudien S, Galgoczy P, Huse K, Reichwald K, Schilhabel M, et al. (2004) Polymorphic segmental duplications at 8p23.1 challenge the determination of individual defensin gene repertoires and the assembly of a contiguous human reference sequence. BMC Genomics 5: 92.
20. Linzmeier RM, Ganz T (2005) Human defensin gene copy number polymorphisms: Comprehensive analysis of independent variation in alpha- and beta-defensin regions at 8p22-p23. Genomics.
21. Abu Bakar S, Hollox EJ, Armour JA (2009) Allelic recombination between distinct genomic locations generates copy number diversity in human beta-defensins. Proc Natl Acad Sci U S A 106: 853–8.
22. Alkan C, Kidd JM, Marques-Bonet T, Aksay G, Antonacci F, et al. (2009) Personalized copy number and segmental duplication maps using next-generation sequencing. Nat Genet 41: 1061–1067.
23. Lupski JR (1998) Genomic disorders: structural features of the genome can lead to DNA rearrangements and human disease traits. Trends Genet 14: 417–22.
24. Bondeson ML, Dahl N, Malmgren H, Kleijer WJ, Tonnesen T, et al. (1995) Inversion of the IDS gene resulting from recombination with IDS-related sequences is a common cause of the Hunter syndrome. Hum Mol Genet 4: 615–21.
25. Bayes M, Magano LF, Rivera N, Flores R, Perez Jurado LA (2003) Mutational mechanisms of Williams-Beuren syndrome deletions. Am J Hum Genet 73: 131–51.
26. Osborne LR, Li M, Pober B, Chitayat D, Bodurtha J, et al. (2001) A 1.5 million-base pair inversion polymorphism in families with Williams-Beuren syndrome. Nat Genet 29: 321–5.
27. Gimelli G, Pujana MA, Patricelli MG, Russo S, Giardino D, et al. (2003) Genomic inversions of human chromosome 15q11-q13 in mothers of Angelman syndrome patients with class II (BP2/3) deletions. Hum Mol Genet 12: 849–58.
28. Voight BF, Kudaravalli S, Wen X, Pritchard JK (2006) A map of recent positive selection in the human genome. PLoS Biol 4: e72.
29. Gonzalez JR, Armengol L, Sole X, Guino E, Mercader JM, et al. (2007) SNPassoc: an R package to perform whole genome association studies. Bioinformatics 23: 644–5.
30. Stranger BE, Forrest MS, Dunning M, Ingle CE, Beazley C, et al. (2007) Relative impact of nucleotide and copy number variation on gene expression phenotypes. Science 315: 848–53.
31. Kidd JM, Cheng Z, Graves T, Fulton B, Wilson RK, Eichler EE (2008) Haplotype sorting using human fosmid clone end-sequence pairs. Genome Res 18: 2016–23.
32. Thomas NS, Bryant V, Maloney V, Cockwell AE, Jacobs PA (2008) Investigation of the origins of human autosomal inversions. Hum Genet 123: 607–16.
33. Lam KW, Jeffreys AJ (2006) Processes of copy-number change in human DNA: the dynamics of v-globin gene deletion. Proc Natl Acad Sci U S A 103: 8921–7.

34. Liang Q, Conte N, Skarnes WC, Bradley A (2008) Extensive genomic copy number variation in embryonic stem cells. Proc Natl Acad Sci U S A 105: 17453–6.

35. Piotrowski A, Bruder CE, Andersson R, de Stahl TD, Menzel U, et al. (2008) Somatic mosaicism for copy number variation in differentiated human tissues. Hum Mutat 29: 1118–1124.

36. Bruder CE, Piotrowski A, Gijsbers AA, Andersson R, Erickson S, et al. (2008) Phenotypically concordant and discordant monozygotic twins display different DNA copy-number-variation profiles. Am J Hum Genet 82: 763–771.

37. Simon-Sanchez J, Scholz S, Fung HC, Matarin M, Hernandez D, et al. (2007) Genome-wide SNP assay reveals structural genomic variation, extended homozygosity and cell-line induced alterations in normal individuals. Hum Mol Genet 16: 1–14.

38. Banchs I, Bosch A, Guimera J, Lazaro C, Puig A, et L., (1994) New alleles at microsatellite loci in CEPH families mainly arise from somatic mutations in the lymphoblastoid cell lines. Hum Mutat 3: 365–372. Erratum in: Hum Mutat 1994 4: 232.

39. Royle NJ, Armour JA, Crosier M, Jeffreys AJ (1993) Abnormal segregation of alleles in CEPH pedigree DNAs arising from allele loss in lymphoblastoid DNA. Genomics 15: 119–22.

40. Flores M, Morales L, Gonzaga-Jauregui C, Dominguez-Vidana R, Zepeda C, et al. (2007) Recurrent DNA inversion rearrangements in the human genome. Proc Natl Acad Sci U S A 104: 6099–106.

41. Tadin-Strapps M, Warburton D, Baumeister FA, Fischer SG, Yonan J, et al. (2004) Cloning of the breakpoints of a de novo inversion of chromosome 8, inv (8)(p11.2q23.1) in a patient with Ambras syndrome. Cytogenet Genome Res 107: 68–76.

42. Sood R, Bader PI, Speer MC, Edwards YH, Eddings EM, et al. (2004) Cloning and characterization of an inversion breakpoint at 6q23.3 suggests a role for Map7 in sacral dysgenesis. Cytogenet Genome Res 106: 61–7.

43. Beiraghi S, Zhou M, Talmadge CB, Went-Sumegi N, Davis JR, et al. (2003) Identification and characterization of a novel gene disrupted by a pericentric inversion inv(4)(p13.1q21.1) in a family with cleft lip. Gene 309: 11–21.

44. Saito-Ohara F, Fukuda Y, Ito M, Agarwala KL, Hayashi M, et al. (2002) The Xq22 inversion breakpoint interrupted a novel Ras-like GTPase gene in a patient with Duchenne muscular dystrophy and profound mental retardation. Am J Hum Genet 71: 637–45.

45. Iida A, Emi M, Matsuoka R, Hiratsuka E, Okui K, et al. (2000) Identification of a gene disrupted by inv(11)(q13.5;q25) in a patient with left-right axis malformation. Hum Genet 106: 277–87.

46. Lakich D, Kazazian HH, Jr., Antonarakis SE, Gitschier J (1993) Inversions disrupting the factor VIII gene are a common cause of severe haemophilia A. Nat Genet 5: 236–41.

47. Bandaru V, Sunkara S, Wallace SS, Bond JP (2002) A novel human DNA glycosylase that removes oxidative DNA damage and is homologous to Escherichia coli endonuclease VIII. DNA Repair (Amst) 1: 517–29.

48. Moskovitz J, Jenkins NA, Gilbert DJ, Copeland NG, Jursky F, et al. (1996) Chromosomal localization of the mammalian peptide-methionine sulfoxide reductase gene and its differential expression in various tissues. Proc Natl Acad Sci U S A 93: 3205–8.

49. Esch FS, Keim PS, Beattie EC, Blacher RW, Culwell AR, et al. (1990) Cleavage of amyloid beta peptide during constitutive processing of its precursor. Science 248: 1122–4.

50. Coufal NG, Garcia-Perez JL, Peng GE, Yeo GW, Mu Y, et al. (2009) L1 retrotransposition in human neural progenitor cells. Nature 460: 1127–1131.

51. Kano H, Godoy I, Courtney C, Vetter MR, Gerton GL, Ostertag EM, Kazazian HH, Jr.: L1 retrotransposition occurs mainly in embryogenesis and creates somatic mosaicism. Genes Dev 2009, 23(11): 1303–1312.