Published in final edited form as:

Genet Med. 2018 July ; 20(7): 697–707. doi:10.1038/gim.2017.170.

Identification of Balanced Chromosomal Rearrangements Previously Unknown Among Participants in the 1000 Genomes Project: Implications for Interpretation of Structural Variation in Genomes and the Future of Clinical Cytogenetics

Zirui Dong[1,2,3,#], **Huilin Wang**[1,3,4,#], **Haixiao Chen**[2,5,#], **Hui Jiang**[2,5,#], **Jianying Yuan**[2,5], **Zhenjun Yang**[2,5], **Wen-Jing Wang**[2,5], **Fengping Xu**[2,5,6], **Xiaosen Guo**[2,5], **Ye Cao**[1,3], **Zhenzhen Zhu**[2,5], **Chunyu Geng**[2,5], **Wan Chee Cheung**[1], **Yvonne K Kwok**[1,3], **Huangming Yang**[2,5], **Tak Yeung Leung**[1,3,7], **Cynthia C. Morton**[8,9,10,11,12,*], **Sau Wai Cheung**[7,13,*], and **Kwong Wai Choy**[1,3,8,*]

[1]Department of Obstetrics & Gynaecology, The Chinese University of Hong Kong, Hong Kong, China

[2]BGI-Shenzhen, Shenzhen 518083, China

[3]Shenzhen Research Institute, Department of Obstetrics & Gynaecology, The Chinese University of Hong Kong, Shenzhen, China

[4]Shenzhen Bao'an Maternal and Child Health Hospital, Shenzhen, China

[5]China National Genebank-Shenzhen, BGI-Shenzhen, Shenzhen 518083, China

[6]Department of Biology, University of Copenhagen, Copenhagen, Denmark

[7]The Chinese University of Hong Kong-Baylor College of Medicine Joint Center For Medical Genetics, Hong Kong, China

[8]Department of Obstetrics and Gynecology, Brigham and Women's Hospital, Boston, Massachusetts, USA

[*]Correspondence: richardchoy@cuhk.edu.hk, scheung@bcm.edu and cmorton@partners.org.

[#]These authors contributed equally.

Supplemental Data

Supplemental Data include four tables.

Conflict of interest

The authors declare no conflict of interest.

List of URLs

DGAP: http://dgap.harvard.edu
1000 Genomes Project: http://www.internationalgenome.org/data/
USCS (University of California, Santa Cruz) genome browser: https://genome.ucsc.edu/
ENCODE: https://www.encodeproject.org
DECIPHER (DatabasE of genomiC varIation and Phenotype in Humans using Ensembl Resources): https://decipher.sanger.ac.uk
Integrative Genomics Viewer: http://software.broadinstitute.org/software/igv/
CyDAS (Cytogenetic Data Analysis System): http://www.cydas.org/OnlineAnalysis/
GTEx: http://www.gtexportal.org/home/
Primer3 Web: http://primer3.ut.ee/
NCBI (National Center for Biotechnology Information) Primer-Blast: http://www.ncbi.nlm.nih.gov/tools/primer-blast/

[9]Harvard Medical School, Boston, MA 02115, USA

[10]Program in Medical and Population Genetics, Broad Institute of Harvard and MIT, Cambridge, Massachusetts, USA

[11]Department of Pathology, Brigham and Women's Hospital, Boston, MA 02115, USA

[12]Division of Evolution and Genomic Sciences, School of Biological Sciences, University of Manchester, Manchester Academic Health Science Center, Manchester 03101, UK

[13]Department of Molecular and Human Genetics, Baylor College of Medicine Houston, TX 77030, USA

Abstract

Purpose—Recent studies demonstrate that whole-genome sequencing (WGS) enables detection of cryptic rearrangements in apparently balanced chromosomal rearrangements (also known as balanced chromosomal abnormalities, BCAs) previously identified by conventional cytogenetic methods. We aimed to assess our analytical tool for detecting BCAs in The 1000 Genomes Project without knowing affected bands.

Methods—The 1000 Genomes Project provides an unprecedented integrated map of structural variants in phenotypically normal subjects, but there is no information on potential inclusion of subjects with apparently BCAs akin to those traditionally detected in diagnostic cytogenetics laboratories. We applied our analytical tool to 1,166 genomes from the 1000 Genomes Project with sufficient physical coverage (8.25-fold).

Results—Our approach detected four reciprocal balanced translocations and four inversions ranging in size from 57.9 kb to 13.3 Mb, all of which were confirmed by cytogenetic methods and PCR studies. One of DNAs has a subtle translocation that is not readily identified by chromosome analysis due to similar banding patterns and size of exchanged segments, and another results in disruption of all transcripts of an OMIM gene.

Conclusions—Our study demonstrates the extension of utilizing low-coverage WGS for unbiased detection of BCAs including translocations and inversions previously unknown in the 1000 Genomes Project.

Keywords

balanced chromosomal rearrangement; the 1000 Genomes Project; G-banded chromosome analysis; low-pass whole-genome sequencing

Introduction

A balanced chromosomal rearrangement (or abnormality, BCA) is a type of chromosomal structural variant (SV) involving chromosomal rearrangements (e.g., translocations, inversions and insertions) without cytogenetically apparent gain or loss of chromatin. The incidence of balanced translocations has been estimated to range from 1/500 to 1/625 in the general population[1–3] and the prevalence is well known to be increased in individuals with clinical anomalies[4–7]. Based on the association of increased prevalence with abnormal clinical phenotypes, studies of BCA such as in the Developmental Genome Anatomy Project (DGAP)[5,6,8,9] among others[7,10] have a high yield in identification of genetic disease due to gene disruption or dysregulation.

Current high-resolution methods (*i.e.*, chromosomal microarray analysis and whole exome sequencing) are generally insensitive to BCA that are unaccompanied by sizable genomic imbalances. Thus, detection of BCA relies on conventional cytogenetic methods (*i.e.*, G-banded karyotyping), which are limited to microscopic resolution (~3–10 Mb). More recently, whole-genome sequencing (WGS) using paired-end analysis has enabled molecular delineation of the breakpoints of BCA at base-pair resolution but has been tested and validated only in DNA samples harboring previously recognized BCAs.

By utilizing WGS (mean 7.4-fold base-coverage) and orthogonal techniques (*i.e.*, long-read single-molecule sequencing), the 1000 Genomes Project establishes the most detailed catalogue of human genetic variation, which in turn can be used for association studies relating genetic variation to disease. It provides an unprecedented integrated map of SVs from 2,054 individuals, including copy-number variants, inversions (< 50 kb) and insertions[11,12], and serves as an indispensable reference for geographic and functional studies of human genetic variation. However, no information was available on the frequency of balanced translocations or inversions (> 50 kb in size) in this resource of participants who were healthy at the time of enrollment. Our previous pilot study has shown the feasibility of detecting BCA with low-pass (or low-coverage) paired-end WGS in a blinded fashion[13]. In the present study, we apply our analytical tool to WGS data released by the 1000 Genomes Project[11].

Materials and Methods

WGS data from the 1000 Genomes Project

Alignment files from 2,504 presumably healthy individuals were downloaded from the 1000 Genomes Project. Assessment of data quality and further analysis was processed for each individual independently.

Minimum physical coverage requirement used in this study

As shown in our previous study[13], the minimum read-pair count was used to avoid false negative detection of BCA. This minimum number of read-pairs in a small-insert library (400 to 600 bp) was estimated as 120 million (50 bp), which is equivalent to 4-fold base coverage from whole-genome sequencing. However, read length (35 to 100 bp) and insert size (200 to 600 bp, Figure 1A) were varied among samples from the 1000 Genomes Project[11]. Therefore, physical coverage[14] was used as the required selection criteria instead of the number of read-pairs (Figure 1B).

We defined a chimeric read-pair if two ends aligned to different chromosomes (interchromosomal) or to the same chromosome (intrachromosomal) with an aligned distance larger than 10 kb[13,15]. Physical coverage was estimated by counting the aligned distances from the non-chimeric and uniquely mapped read-pairs[13]. In the present study, the minimum physical coverage of 8.25-fold, estimated based on 90 million read-pairs (data from our previous pilot study)[13], was set to maximize inclusion of 1,166 out of 2,504 samples (Supplementary Table 1). This was based on: (1) only 616 out of 2,504 samples available for this study with 11-fold physical coverage (estimated based on 120 million read-pairs), and (2) an increase in the false negative detection rate in our previous study from 11.1% (1/9) with 90 million read-pairs to 33.3% (3/9) with only 60 million read-pairs[13].

Data quality control and BCA detection

We filtered out low-quality reads (≥ 4% of mismatch rates) and extracted uniquely aligned reads in both ends for further analysis. Detection of chromosomal rearrangement is based on a four-step procedure described in our previous study[13]: briefly, (1) Event clustering: chimeric read-pairs were clustered by sorting the aligned coordinates (GRCh37/hg19) and any two read-pairs were considered to represent two distinct events if they were separated by a distance of >10 kb; (2) Systematic error filtering: Each event was filtered against a control dataset, which was built up by using the events from all the 2,504 samples, and a false positive was filtered out if it was identified in more >5% subjects; (3) Random error filtering: Event was filtered with a cluster property matrix (*i.e.*, supporting read-pair amount and the average number of mismatches) with the reported parameters; and (4) Aligned orientations: each event was filtered based on q/p arm genetic exchange (joining type). As some of the samples were with short read lengths (*i.e.*, 35 bp), we further used Sanger sequencing results to fine map the ligated sequences ate the breakpoints.

Chromosome analysis and FISH validation

Epstein-Barr virus (EBV)-transformed B lymphoblastoid (EBV-B) cell lines were obtained from the Coriell Institute (Camden, NJ) for validation. G-banded chromosome analysis was performed using standard protocols for more than 100 cells in each EBV-B cell line[16]. Fluorescence in situ hybridization (FISH) was performed for NA18612 using standard procedures with BAC clones labeled by nick translation with SpectrumOrange or SpectrumRed, SpectrumGreen dUTP (Abbott Molecular, Des Plaines, IL)[16,17]. BAC clones were selected from the UCSC Genome Browser.

Molecular validation of balanced rearrangements

For samples with translocations and available EBV-B cell lines, genomic DNA was extracted using a commercial DNA extraction kit (Puregene; Qiagen, Hilden, Germany). For samples identified with submicroscopic inversions, DNAs were obtained from the Coriell Institute (Camden). Each DNA was quantified subsequently with the Qubit dsDNA HS Assay Kit (Invitrogen, Life Technologies, Waltham, MA) for DNA quality measurement.

Genomic reference sequences (GRCh37/hg19) at a 1 kb distance from each putative breakpoint region (both upstream and downstream) were used for primer design with Primer3 Web and NCBI Primer-Blast (Supplementary Table 2). PCR amplification was performed simultaneously in cases and control (DNA from YH, a well-characterized normal EBV-B cell line[18]). PCR products were sequenced by Sanger sequencing on an ABI 3730 machine (Applied Biosystems, Thermo Fisher Scientific Inc, Wilmington, DE)[8,13,19] and sequencing results were aligned with BLAT for further confirmation of the balanced rearrangement and for mapping breakpoints at single nucleotide level.

RNA preparation, library construction and sequencing

Total RNA was extracted from each EBV-B cell line with a balanced translocation using TRIzol Reagent (Invitrogen) according to the manufacturer's instructions, and subsequently treated with DNase I (Invitrogen)[20]. For each RNA sample, purity was evaluated with a Nano-Photometer spectrophotometer (Implen, Westlake Village, CA), concentration measured in a Qubit 2.0 Fluorometer (Life Technologies), and RNA integrity verified using an Agilent 2100 BioAnalyzer (Agilent Technologies, Santa Clara, CA).

For library construction, mRNA enrichment was performed with Oligo(dT)25 Dynabeads (Thermo Fisher Scientific) twice and purification was carried out with the Dynabeads® mRNA Purification Kit (Invitrogen, No. 61006). The eluted mRNA was fragmented with Fragmentation Buffer Mix at 94℃ for 10 min. Reverse transcription (RT) was performed with RT Buffer Mix and RT Enzyme Mix followed by double strand cDNA (dscDNA) synthesis with Second Strand Buffer Mix and Second Strand Enzyme Mix. End repair, adaptor (with barcode) ligation and PCR amplification were performed after dscDNA purification. Then, the purified double stranded PCR products were heat denatured to single strand and circularized with Splint Oligo Mix and Ligation Enzyme. The single strand circle DNA (ssCirc DNA) library was rolling circle amplified for constructing the DNA nanoball (DNB), which was substantially loaded into a patterned nanoarray. Paired-end sequencing with 50 bp in each end (PE50) was carried out in a BGISeq-500 platform (BGI, Wuhan, China)[21].

RNA-seq data analysis

Paired-reads that passed standard quality control[13,15] were simultaneously aligned to the human genome (GRCh37/hg19) using HISAT (Hierarchical Indexing for Spliced Alignment of Transcripts)[22], and aligned to human transcriptome (RefSeq) via Bowtie[23]. One base-pair mismatch was set in each alignment. Paired-end aligned reads were used for further analysis. Alignment files were transformed into Pileup files for determination of coverage with Samtools (mpileup). Expression of each gene in each sample was determined based on alignment files from the human transcriptome (RefSeq).

Gene expression of each sample was compared to data reported for 13 EBV-B controls present in the Genotype-Tissue Expression (GTEx) project[24].

Validation of cryptic deletions

Quantitative PCR (qPCR) was performed for validation of the two cryptic deletions. Genomic reference sequences (GRCh37/hg19) of each deleted region were used for primer design with Primer3 Web and NCBI Primer-Blast (Supplementary Table 2). Melting curve analysis was carried out for each pair of primers to ensure specificity of the PCR amplification, and the standard curve method was used to determine PCR efficiency (within a range from 95% to 105%).

Each reaction was performed in quadruplicate in 10 µl of reaction mixture simultaneously in cases and control (DNA from YH EBV-B cell line[13]) on a StepOnePlus Real-Time PCR System (Applied Biosystems) with SYBR Premix Ex Taq Tli RNaseH Plus (Takara

Biotechnology Co., Ltd., Dalian, Liaoning, China) with the default setting of the reaction condition. The number of copies in each sample was determined by using the ΔΔCt method that compared the Ct (cycle threshold) in case to control[25]. Two independent pairs of primers (Supplementary Table 2) were used in quintuplicate for validation of each deletion.

Accession Number

The accession number for the RNA-seq data reported in this paper is GSE94043 (NCBI Gene Expression Omnibus).

Code availability

All the programs relevant to this pipeline are available at https://sourceforge.net/projects/bca-analysis/files/BCA.tar.gz/download.

Results

We assessed 1,166 samples with at least 8.25-fold physical coverage (Supplementary Table 1) using our reported approach with the same parameters[13]. Four samples (HG02260, HG03729, NA18612 and NA20764) were identified to harbor balanced translocations (Figures 1A), and four samples (NA20759, HG04152, NA18959 and NA21133) were with inversions, the size of which ranged from 57.8 kb to 13.3 Mb (Figures 1A). Among the four cases with balanced translocations, two are female and two are male and they originate from different ethnic populations (Table 1)[11]. For the four cases with inversions, all are males and also originate from different ethnic populations (Table 2).

G-banded karyotypes were observed to be directly consistent with the WGS data for samples HG02260, HG03729, NA20764 and NA20759 (Figure 1B–E), and those of NA18612 were consistent but much less obvious (Figure 2A, described below). Sanger results confirmed each rearrangement in all eight samples with BCAs (Tables 1, 2, next-generation cytogenetic nomenclature[26] shown in Supplemental Table 3). Microhomology sequences were identified in eight of 16 breakpoints suggesting the rearrangements were mediated by microhomology-mediated end joining (MMEJ)[27] (Tables 1, 2). The remaining eight breakpoints represented non-homologous end joining (NHEJ)[28] (Tables 1, 2).

Subtle balanced translocation identified by WGS

Breakpoints of the t(16;17)(q23.1;q24.2) (NA18612, Figure 2B) were located in bands 16q23.1 and 17q24.2, representing translocated segments of 15.0 Mb and 16.2 Mb, respectively. Due to similarity in the G-banding pattern and size of the exchanged segments, chromosome analysis did not readily identify the translocation (Figure 2A). Therefore, metaphase FISH[17] was performed using BAC probes (SpectrumOrange: RP11-7D23 at 16q24.3, SpectrumGreen: RP11-526M7 at 17q25.1 and SpectrumRed: RP11-135N5 at 17p13.3) in more than 100 cells confirming the t(16;17) (Figure 2C).

Gene disruptions by the breakpoints of balanced translocations

Among the four cell lines with balanced translocations, the eight breakpoints disrupted six genes (Table 1, Figures 2D, 2E, 3A, and 3B), four of which resulted in disruption of all transcripts in the derivative chromosomes of the breakpoints. In contrast, none of the breakpoints from the four cases with inversions disrupted any gene (Table 2).

The breakpoint in seq[GRCh37/hg19] 16q23.1(75,336,134_75,336,138) (NA18612, Figure 2D) disrupts the gene encoding craniofacial development protein 1 (*CFDP1*, NM_006324), resulting in aberrant splicing of intron 6 and absence of expression of exon 7. This disruption is supported by observation of RNA-seq reads mapping in the non-exonic region (Figure 2D) and decreased expression of exon 7 (Figure 2E). Although *CFDP1* has been reported to be necessary for cell survival and differentiation during tooth morphogenesis in organ culture[29], it is unlikely to be haploinsufficient [Haploinsufficiency score (HI)=14.9%][30].

In contrast, the seq[GRCh37/hg19] 14q31.1(79,839,173_79,839,174) breakpoint of 46,XX,t(9;14)(q34.2;q31.1) (HG02260) (Figure 3A) disrupts all transcripts of neurexin 3 (*NRXN3*), which is likely to be haploinsufficient (HI=0.3%)[30]. However, expression of *NRXN3* was not detectable among any of the EBV-B cell lines including cases and controls (Figure 3C)[24].

Cryptic deletions

The 3q24 breakpoint of 46,XY,t(3;17)(q24;p13.3) (HG03729) was found to include a 5.2 kb deletion, seq[GRCh37/hg19] 3q24(143,817,430_143,822,651)x1, while the 17p13.3 breakpoint has a 4.4 kb deletion, seq[GRCh37/hg19] 17p13.3(2,910,366_2,914,751)x1 (Table 1, Figure 3D and 3E). Neither deletion was reported previously[11] and both were confirmed by quantitative PCR (Figure 3D and 3E).

Positional effects

Previous studies show that genes in proximity to the breakpoints of a structural variant (*i.e.*, balanced translocation) may be mis-expressed, which is defined as a positional effect[5]. One mechanism for a positional effect is the disruption of topological associated domains (TADs) by the SV's breakpoints[6,9,31]. Here, we used boundaries predicted from the human IMR90 fibroblast cell line (GRCh37/hg19)[31] for our study, as TADs are highly conserved across different cell types and across species[32].

Eight TADs were disrupted by the breakpoints from the four translocations. Thirty-four genes are located in these eight disrupted TADs, and expression was observed in 16 of these genes in normal EBV-B cell lines (Supplemental Table 4). However, mis-expression was not observed in any of these 34 genes from our RNA-seq data (Supplemental Table 4), even though two of these genes are predicted to be likely haploinsufficient (Supplemental Table 4). By using the published ChIP-seq data from Encyclopedia of DNA Elements (ENCODE)[33] in EBV-B cell line GM12878, 22 of the 34 genes have a candidate promoter (indicated by H3K4Me3) near to a potential active regulatory element (indicated by H3K27Ac)[31,33]. In addition, from the accessible chromatin landscape[34], 19 of the 34 genes have highly associated DNA I hypersensitive sites (DHSs) and each of them has at least one DHS located in the same partial TAD as the gene and the predicted promoter.

The breakpoint in seq[GRCh37/hg19] 17q24.2(64,953,078_64,953,079) is likely to be located between the *CACNG4* promoter (indicated by H3K4Me3) and its potential enhancer [indicated by H3K27Ac[31,33] and DHSs[34], Figure 3F] in a human embryonic stem cell line (H1-hESC)[33]. These data suggest that the translocation would likely result in disruption of the interaction between the promoter and enhancer for *CACNG4*[32] in this particular cell line. However, mis-expression of *CACNG4* was not observed in our RNA-seq data from EBV-B cell line (Figure 3G). Both the candidate promoter and enhancer for *CACNG4* were likely located downstream of the breakpoint[33] in EBV-B cell line GM12878 (control EBV-B cell line, Figure 3F).

Discussion

Balanced chromosomal abnormalities including translocations and inversions, known to cause reproductive problems and/or an abnormal phenotype, currently are mainly detected by G-banded chromosome analysis. However, subtle or cryptic BCAs are not detectable by current methods but may contribute to birth defects in offspring of the carriers due to unbalanced segregations[35]. In the present study, by utilizing existing genomic data from the 1000 Genomes Project, we demonstrate the feasibility of using WGS in the detection of BCAs in samples without prior knowledge of their existence.

In the present study, we set a cutoff of 8.25-fold physical coverage to maximize the inclusion of 1,166 samples out of 2,504 based on the evaluation of the false negative rate in our previous study (11.1% with 90 million read-pairs with insert sizes ranging from 400 to 600 bp)[13]. The exclusion of more than half of all samples (n=1,338) is because of the smaller insert size generated (259.1±93.5 bp, Figure 1A); the number of non-chimeric and uniquely mapped read-pairs was 97.0±40.2 millions, although the base-coverage reached 7.4-fold on average. This indicates that better performance of detecting BCA can be achieved by using larger insert sizes to increase physical coverage, thus, increasing the number of supporting read-pairs for the potential BCAs.

The prevalence of reciprocal balanced translocations in this dataset is one in 291.5 (0.34%, 4/1,166), which is higher than the rate reported estimated by G-banded chromosome analysis[1–3]. This estimate may be biased due to the limited sample size (N=1,166). However, the reported incidences may be underestimated as cryptic or subtle rearrangements, such as that observed for 46,XY,t(16;17)(q23.1;q24.2) (NA18612) may not be readily identified by conventional G-banded chromosome analysis (Figure 3). Another explanation might be that the detection of rearrangements was based on WGS of EBV-B cell line-derived DNAs, which might have EBV-B specific genomic variants owing to the introduction of genomic instability by EBV infection or the conditions of cell culture[36]. However, as the EBV-B specific genomic variants frequently exist as mosaics[36], giving the 100% consistency of more than 100 metaphases in each sample in the present study and the WGS data of these samples used for our analysis were generated from early batches of EBV-Bs in the 1000 Genomes Project, balanced translocations detected probably represent the true events in the subjects' peripheral blood samples. Nonetheless, our approach reports the true events existing in the tested EBV-B cell lines.

In addition to the detection of balanced translocations, the microscopic inversion and three submicroscopic inversions identified (Table 2) were each unique to a single subject among all 1,166 samples analyzed. One explanation for not identifying common or recurrent inversions is that they may be mediated by repetitive elements[28], for which sequencing with small-insert library might not be able to detect[37]. Sequencing with mate-pair library (or large insert library) might be able to overcome such challenge and can also largely reduce the sequencing cost by reducing the read-pair amount required[13,19]. Nonetheless, the identification of both balanced translocations and inversions underscores the importance of using low-pass WGS for nucleotide level precision of chromosomal rearrangements in cytogenetic diagnoses, and brings the future of implementing sequencing a step closer as the first tier test.

Gene disruptions were observed in six out of the eight breakpoints in four cases with balanced translocations, and *NRXN3*, a likely haploinsufficient gene, was disrupted (HG02260). Heterozygous deletion of *NRXN3* is reported in autism spectrum disorder (ASD)[38]. Although this participant in the 1000 Genomes Project is assumed to be healthy at the time of enrollment, a possible explanation for the absence of ASD in this presumably normal individual would be lack of penetrance[38] in the absence of a positive comprehensive medical assessment or some technical failure in the process.

Two cryptic deletions involving both breakpoints were identified in t(3;17)(q24;p13.3) (HG03729) and neither of them were reported previously[11]. Two possible reasons for missed detection in the previous study[11] are: (1) only a limited number of reads mapping in these regions (Figure 3D and 3E) resulting in read-depth differences insufficiently sensitive for identification, and (2) absence of intra-chromosomal aligned read-pairs supporting these two deletions. Genomic imbalance commonly involves the breakpoint of balanced translocations and some of them are known to be pathogenic or likely pathogenic[6], thus, indicating the importance of identification.

No aberration in gene expression resulted from a positional effect, such as disruption of TADs as observed in our EBV-B cell line-derived RNA-seq data from four cases with balanced translocations. One explanation is that expression was only observed in 16 out of the 34 genes in normal EBV-B (Supplemental Table 4), and an effect of dysregulated lower expression cannot be detected for genes without detectable expression in the EBV-B cell lines[24]. In addition, another reason would be the proximate interaction between promoter and enhancer: (1) 22 of the 34 genes have a candidate promoter near to a potential active regulatory element[31,33]; and (2) 19 of the 34 genes have highly associated DHSs[34], and each of them has at least one DHS located in the same partial TAD as the gene and the predicted promoter, indicating some residual interactions remains between promoter and regulatory elements, thus, the disruption of TADs is likely insufficient to alter the gene expression. As data from RNA expression provides evidence for confirming potential effects attributed to a chromosomal rearrangement, it indicates the importance of combining RNA expression analysis with identification of BCAs based on DNA samples for clinical interpretation.

We observed a potential disruption of an interaction between the promoter and an enhancer for the 17q24.2 breakpoint (NA18612) in H1-hESC, which serves as a reference for disease association prediction[31]. However, mis-expression of *CACNG4* was not observed in our RNA-seq data from the EBV-B cell line (Figure 3G). One explanation is that both the candidate promoter and enhancer for *CACNG4* are likely located downstream of the breakpoint[33] in the EBV-B cell lines (GM12878, Figure 3F). As another sample type from this subject is not obtainable for further validation, this argues the common usage of peripheral blood as a valuable sample type for disease studies beyond its simple availability.

Overall, this study is the first reported investigation utilizing low-pass WGS to explore detection of BCAs among samples from the 1000 Genome Project without prior knowledge of a chromosomal abnormality. In addition, disruption of gene, cryptic imbalances and potential disruption of promoter and enhancer interaction were observed in the four cases with balanced translocations, demonstrating the advantage of detecting the breakpoints in BCAs by molecular methods via paired-end sequencing and Sanger sequencing, and has important implications for a new dawn of improved diagnostics in clinical cytogenetics.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

We thank the 1000 Genomes Project for releasing the whole-genome sequencing data for this study, and Drs. Feng Yue and Bo Zhang from Pennsylvania State University for sharing with us the predicted topological association domain from the human IMR90 fibroblast cell line (in GRCh37/hg19 version). This study was approved by the Institutional Review Board from the Chinese University of Hong Kong. This project is supported by the National Natural Science Foundation of China (Project 81370715), the Natural Science Foundation of Guangdong Province (Grant No. 2014A030313795), the Shenzhen Municipal Commission for Development and Reform and Key Laboratory Project in Shenzhen (CXB200903110066A and CXB201108250096A), Health and Medical Research Fund (HMRF, Project 04152666) and the Shenzhen Municipal Government of China (NO.CXZZ20150330171838997). CCM acknowledges the NIH/NIGMS GM061354 for support.

References

1. Mackie Ogilvie C, Scriven PN. Meiotic outcomes in reciprocal translocation carriers ascertained in 3-day human embryos. European journal of human genetics: EJHG. 2002; 10(12):801–806. [PubMed: 12461686]
2. Oliver-Bonet M, Navarro J, Carrera M, Egozcue J, Benet J. Aneuploid and unbalanced sperm in two translocation carriers evaluation of the genetic risk. Mol Hum Reprod. 2002; 8(10):958–963. [PubMed: 12356948]
3. Van Dyke DL, Weiss L, Roberson JR, Babu VR. The frequency and mutation rate of balanced autosomal rearrangements in man estimated from prenatal genetic studies for advanced maternal age. Am J Hum Genet. 1983; 35(2):301–308. [PubMed: 6837576]
4. Warburton D. De novo balanced chromosome rearrangements and extra marker chromosomes identified at prenatal diagnosis: clinical significance and distribution of breakpoints. Am J Hum Genet. 1991; 49(5):995–1013. [PubMed: 1928105]
5. Talkowski ME, Rosenfeld JA, Blumenthal I, et al. Sequencing chromosomal abnormalities reveals neurodevelopmental loci that confer risk across diagnostic boundaries. Cell. 2012; 149(3):525–537. [PubMed: 22521361]
6. Redin C, Brand H, Collins RL, et al. The genomic landscape of balanced cytogenetic abnormalities associated with human congenital anomalies. Nat Genet. 2017; 49(1):36–45. [PubMed: 27841880]
7. De Gregori M, Ciccone R, Magini P, et al. Cryptic deletions are a common finding in “balanced” reciprocal and complex chromosome rearrangements: a study of 59 patients. J Med Genet. 2007; 44(12):750–762. [PubMed: 17766364]

8. Talkowski ME, Ordulu Z, Pillalamarri V, et al. Clinical diagnosis by whole-genome sequencing of a prenatal sample. N Engl J Med. 2012; 367(23):2226–2232. [PubMed: 23215558]

9. Ordulu Z, Kammin T, Brand H, et al. Structural chromosomal rearrangements require nucleotide-level resolution: lessons from next-generation sequencing in prenatal diagnosis. Am J Hum Genet. 2016; 99(5):1015–1033. [PubMed: 27745839]

10. Nilsson D, Pettersson M, Gustavsson P, et al. Whole-genome sequencing of cytogenetically balanced chromosome translocations identifies potentially pathological gene disruptions and highlights the importance of microhomology in the mechanism of formation. Hum Mutat. 2017; 38(2):180–192. [PubMed: 27862604]

11. Sudmant PH, Rausch T, Gardner EJ, et al. An integrated map of structural variation in 2,504 human genomes. Nature. 2015; 526(7571):75–81. [PubMed: 26432246]

12. 1000 Genomes Project Consortium. A global reference for human genetic variation. Nature. 2015; 526(7571):68–74. [PubMed: 26432245]

13. Dong Z, Jiang L, Yang C, et al. A robust approach for blind detection of balanced chromosomal rearrangements with whole-genome low-coverage sequencing. Hum Mutat. 2014; 35(5):625–636. [PubMed: 24610732]

14. Ekblom R, Wolf JB. A field guide to whole-genome sequencing, assembly and annotation. Evol Appl. 2014; 7(9):1026–1042. [PubMed: 25553065]

15. Dong Z, Zhang J, Hu P, et al. Low-pass whole-genome sequencing in clinical cytogenetics: a validated approach. Genet Med. 2016; 18(9):940–948. [PubMed: 26820068]

16. Ou Z, Stankiewicz P, Xia Z, et al. Observation and prediction of recurrent human translocations mediated by NAHR between nonhomologous chromosomes. Genome Res. 2011; 21(1):33–46. [PubMed: 21205869]

17. Gui B, Yao Z, Li Y, et al. Chromosomal analysis of blastocysts from balanced chromosomal rearrangement carriers. Reproduction. 2016; 151(4):455–464. [PubMed: 26825930]

18. Wang J, Wang W, Li R, et al. The diploid genome sequence of an Asian individual. Nature. 2008; 456(7218):60–65. [PubMed: 18987735]

19. Talkowski ME, Ernst C, Heilbut A, et al. Next-generation sequencing strategies enable routine detection of balanced chromosome rearrangements for clinical diagnostics and genetic research. Am J Hum Genet. 2011; 88(4):469–481. [PubMed: 21473983]

20. Zhang B, Zhang W, Nie RE, et al. Comparative transcriptome analysis of chemosensory genes in two sister leaf beetles provides insights into chemosensory speciation. Insect Biochem Mol Biol. 2016; 79:108–118. [PubMed: 27836740]

21. Goodwin S, McPherson JD, McCombie WR. Coming of age: ten years of next-generation sequencing technologies. Nat Rev Genet. 2016; 17(6):333–351. [PubMed: 27184599]

22. Kim D, Langmead B, Salzberg SL. HISAT: a fast spliced aligner with low memory requirements. Nat Methods. 2015; 12(4):357–360. [PubMed: 25751142]

23. Langmead B, Trapnell C, Pop M, Salzberg SL. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biol. 2009; 10(3):R25. [PubMed: 19261174]

24. GTEx Consortium. The Genotype-Tissue Expression (GTEx) project. Nat Genet. 2013; 45(6):580–585. [PubMed: 23715323]

25. Livak KJ, Schmittgen TD. Analysis of relative gene expression data using real-time quantitative PCR and the 2(−Delta Delta C(T)) Method. Methods. 2001; 25(4):402–408. [PubMed: 11846609]

26. Ordulu Z, Wong KE, Currall BB, et al. Describing sequencing results of structural chromosome rearrangements with a suggested next-generation cytogenetic nomenclature. Am J Hum Genet. 2014; 94(5):695–709. [PubMed: 24746958]

27. Sfeir A, Symington LS. Microhomology-mediated end joining: a back-up survival mechanism or dedicated pathway? Trends Biochem Sci. 2015; 40(11):701–714. [PubMed: 26439531]

28. Carvalho CM, Lupski JR. Mechanisms underlying structural variant formation in genomic disorders. Nat Rev Genet. 2016; 17(4):224–238. [PubMed: 26924765]

29. Diekwisch TG, Luan X. CP27 function is necessary for cell survival and differentiation during tooth morphogenesis in organ culture. Gene. 2002; 287(1–2):141–147. [PubMed: 11992732]

30. Huang N, Lee I, Marcotte EM, Hurles ME. Characterising and predicting haploinsufficiency in the human genome. PLoS Genet. 2010; 6(10):e1001154. [PubMed: 20976243]

31. Lupiáñez DG, Kraft K, Heinrich V, et al. Disruptions of topological chromatin domains cause pathogenic rewiring of gene-enhancer interactions. Cell. 2015; 161(5):1012–1025. [PubMed: 25959774]

32. Dixon JR, Selvaraj S, Yue F, et al. Topological domains in mammalian genomes identified by analysis of chromatin interactions. Nature. 2012; 485(7398):376–380. [PubMed: 22495300]

33. Rosenbloom KR, Sloan CA, Malladi VS, et al. ENCODE data in the UCSC Genome Browser: year 5 update. Nucleic Acids Res. 2013; 41:D56–63. Database issue. [PubMed: 23193274]

34. Thurman RE, Rynes E, Humbert R, et al. The accessible chromatin landscape of the human genome. Nature. 2012; 489(7414):75–82. [PubMed: 22955617]

35. Ledbetter DH, Martin CL. Cryptic telomere imbalance: a 15-year update. Am J Med Genet C Semin Med Genet. 2007; 145C(4):327–334. [PubMed: 17910073]

36. Shirley MD, Baugher JD, Stevens EL, et al. Chromosomal variation in lymphoblastoid cell lines. Hum Mutat. 2012; 33(7):1075–1086. [PubMed: 22374857]

37. Meyerson M, Gabriel S, Getz G. Advances in understanding cancer genomes through second-generation sequencing. Nat Rev Genet. 2010; 11(10):685–696. [PubMed: 20847746]

38. Vaags AK, Lionel AC, Sato D, et al. Rare deletions at the neurexin 3 locus in autism spectrum disorder. Am J Hum Genet. 2012; 90(1):133–141. [PubMed: 22209245]

39. Shlyueva D, Stampfel G, Stark A. Transcriptional enhancers: from properties to genome-wide predictions. Nat Rev Genet. 2014; 15(4):272–286. [PubMed: 24614317]

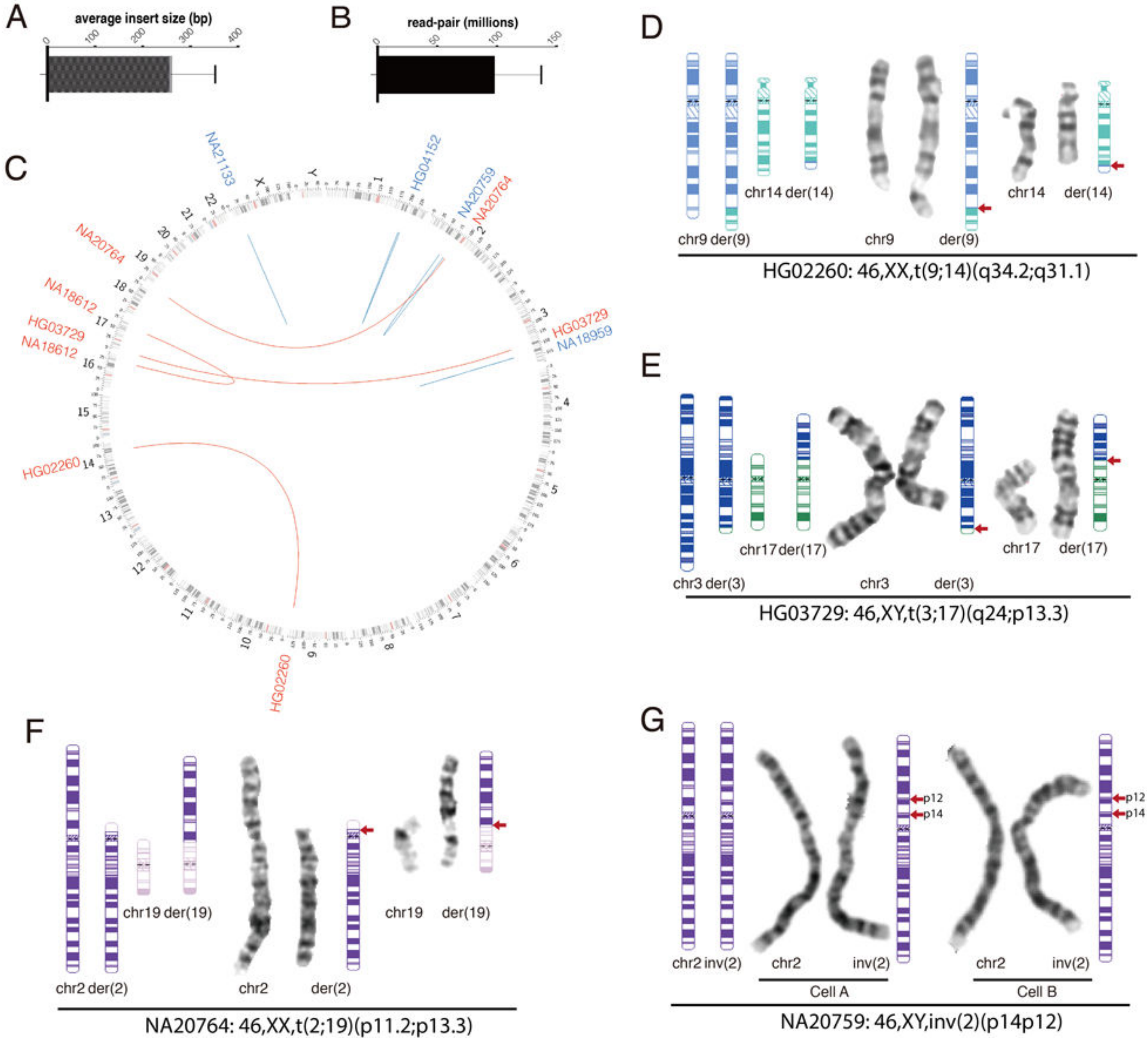

Figure 1. Spectrum of inter- and intra-chromosomal balanced rearrangements and cytogenetic validations

Distributions of the average insert sizes and read-pair amounts of 2,504 samples from the 1000 Genomes Project are shown in **(A)** and **(B)**, respectively. Insert size and read-pair amounts were calculated based on non-chimeric and uniquely mapped read-pairs. **(C)** Spectrum of BCAs. Balanced translocations are indicated with red lines and the corresponding sample IDs are shown in red font in each affected chromosome in the outmost circle. Inversions are indicated in blue lines and sample IDs are shown in blue font. Chromosomal nucleotide positions and bands are shown according to the UCSC Genome Viewer Table Browser. In figures **(D)**, **(E)** and **(F)**, validation of balanced translocations and inversion **(G)** by G-banded chromosome analysis are shown. Ideograms of the balanced rearrangements are shown on the left, while the karyogram images are to the right with the corresponding ideogram of the derivative chromosomes for reference. Breakpoint regions are indicated with red arrows. Sample name and the International System for Human Cytogenomic Nomenclature (ISCN) description are shown below each.

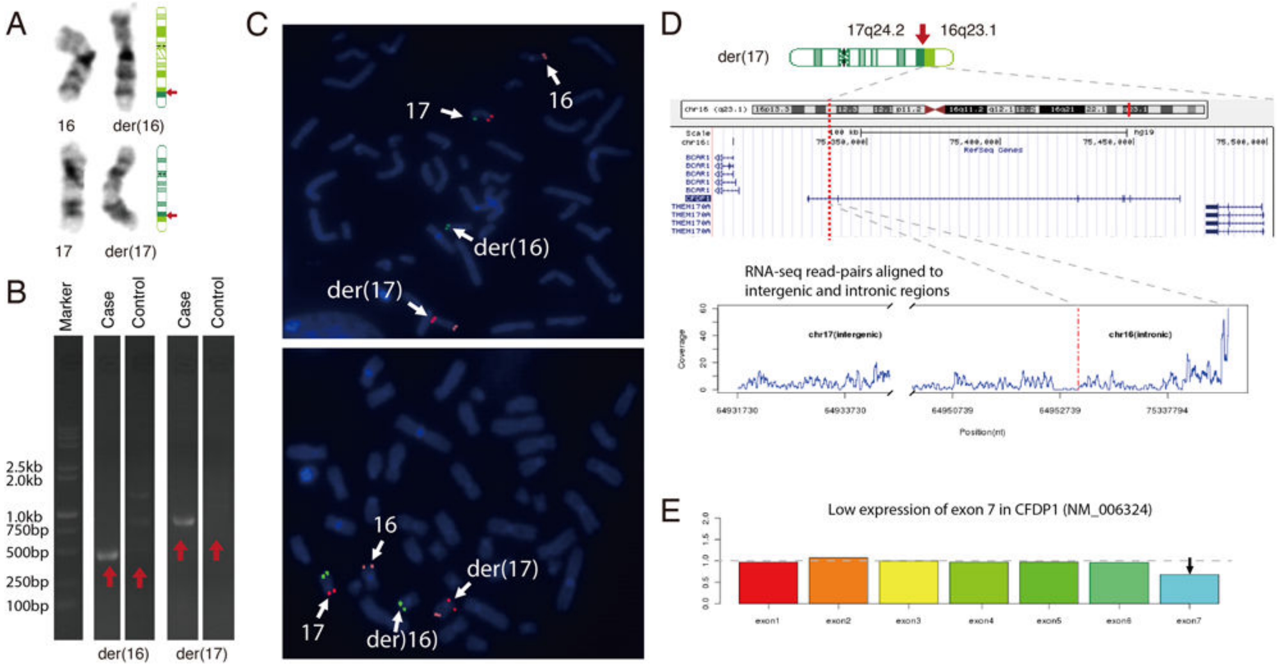

Figure 2. A subtle translocation t(16;17)(q23.1;q24.2) (NA18612) and the aberrant splicing of intron 6 of *CFDP1* (NM_006324)

(A) Validation from chromosome analysis. Ideograms of the derivative balanced translocation chromosomes are shown with the corresponding G-banded chromosome pairs. Breakpoint regions are indicated with red arrows. **(B)** PCR validation of the junction of DNA sequences from the two derivative chromosomes indicated by red arrows while absent in the negative control. **(C)** Validation from FISH. BAC probes are shown with the targeted bands (16q24.3 in SpectrumOrange, 17p13.3 in SpectrumRed and 17q25.1 in SpectrumGreen, respectively). Derivative chromosomes and normal chromosomes are designated with arrows. **(D)** In the der(17), the genomic location of anti-sense gene *CFDP1* (NM_006324) is shown with the breakpoint mapping in intron 6 (red dotted line). RNA-seq read-pairs align to the region (expressed as coverage) proximal to the breakpoint in seq[GRCh37/hg19] 16q23.1(75,336,134_75,336,138). It includes the intergenic region in 17q24.2 and the partial intron 6 of *CFDP1* (NM_006324) in 16q23.1 (two grey dotted lines), indicating the aberrant splicing of intron 6. **(E)** Transcript coverage was plotted with the paired-end aligned reads (RNA-seq). The coverage of each coordinate is divided by the average coverage in this transcript, and subsequently normalized coverage with the average coverage from the other three EBV-B cell lines from the 1000 Genomes Project. Black arrow indicates low expression in exon 7 that is the absence of exon 7 in the disrupted transcript.

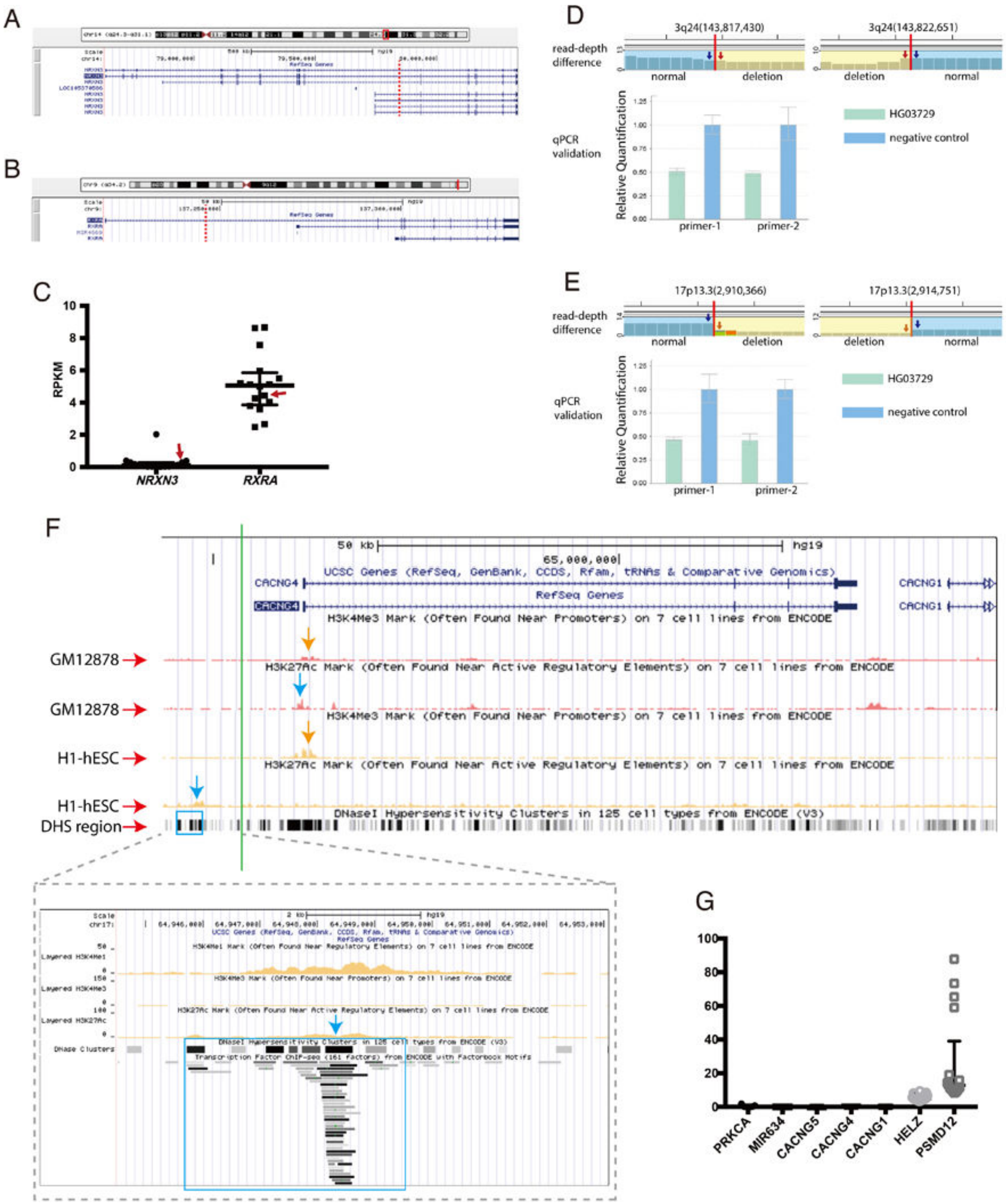

Figure 3. Gene disruption, cryptic deletions and potential disruption of interaction between promoter and enhancer by the breakpoints of balanced translocations

Figures **(A)**, **(B)** and **(C)** *NRXN3* disruption in 46,XX,t(9;14)(q34.2;q31.1) (HG02260). **(A)** and **(B)** Genomic locations of *NRXN3* and *RXRA* are shown with breakpoints indicated by red dotted lines. **(C)***NRXN3* and *RXRA* expression for the four cases from the 1000 Genomes Project and for 13 reported EBV-B normal control cell lines (the GTEx project). Gene expression for *NRXN3* and *RXRA* in HG02260 are indicated with red arrows. Figures **(D)** and **(E)** cryptic deletions involved at the breakpoints in translocation 46,XY,t(3;17) (q24;p13.3) (HG03729). Two cryptic deletions of seq[GRCh37/hg19] 3q24(143,817,430_143,822,651)×1 and seq[GRCh37/hg19] 17p13.3(2,910,366_2,914,751) ×1 were detected by read-depth difference algorithm and were further confirmed by quantitative PCR. The deleted regions are shown in a yellow background with a red arrow while the normal copy-ratio (diploid) is shown in a blue background with a blue arrow. Two independent pairs of primers (Supplementary Table 2) were used to perform qPCR in quintuplicate for validation of each deletion. The bars in cyan show the relative quantification of HG03729, while the bars in blue indicate the negative control. Figures **(F)** and **(G)** potential disruption of interaction between promoter and enhancer from rearrangement in 46,XY,t(16;17)(q23.1;q24.2) (NA18612) in H1-hESC. **(F)** Genes and the ChIP-seq data from the ENCODE Project are shown in terms of the genomic location. Each cell line with the ChIP-seq data (*i.e.*, H3K4Me3 and H3K27Ac)[33] is labeled with a red arrow. Breakpoint in seq[GRCh37/hg19] 17q24.2(64,953,078_64,953,079) is shown by a green vertical line, while the candidate promoters and enhancers are indicated with orange and blue arrows, respectively. The region of potential enhancer in H1-hESC is highlighted in DNase I Hypersensitivity Clusters[34] in a blue rectangle (DHS region). The figure below is zoomed in on the potential enhancer region in H1-hESC. Enrichment of H3K4Me1 and absence of H3K4Me3 support a potential active enhancer in this region[33,39], while enrichment of DNA-binding sequence motifs also indicates the candidate region of the interaction for regulatory elements[33]. **(G)** Gene expression level (Read Per Kilobase Million) of the four cases and 13 EBV-B normal control samples (GTEx project)[24].

Table 1

Balanced translocations detected in the 1000 Genomes Project

Sample ID	Karyotype	Continental group[#]	Population[*]	der[$]	Breakpoint A		Breakpoint B		Micro-homology	chr	Deletion involving breakpoints			Gene disrupted	Genes in TAD[^]
					chr	Coordinate	chr	Coordinate			Size (bp)	Start	End		
HG02260	46,XX, t(9;14)(q34.2;q31.1)	AMR	PEL	9	9	137230907	14	79839174	–	9	14	137230908	137230922	*RXRA*	*WDR5,RNU6ATAC*
				14	14	79839173	9	137230923	–	14	0	NA	NA	*NRXN3*	–
HG03729	46,XY, t(3;17)(q24;p13.3)	SAN	ITU	3	3	143817430	17	2910366	–	3	5,219	143817431	143822650	–	*SLC9A9,C3orf58*
				17	3	143822651	17	2914751	TT	17	4,383	2910367	2914750	*RAP1GAP2*	*MIR1253*
NA18612	46,XY, t(16;17)(q23.1;q24.2)	ASN	CHB	16	16	75336134	17	64953079	C	16	3	75336135	75336137	*CFDP1*	*CTRB2,CTRB1,BCAR1,TMEM170A,CHST6,CHST5,TMEM231,GABARAPL2,ADAT1*
				17	17	64953078	16	75336138	A	17	0	NA	NA	–	*MIR634,CACNG5,CACNG4,CACNG1,HELZ*
NA20764	46,XX, t(2;19)(p11.2;p13.3)	EUR	TSI	2	2	86491099	19	424310	–	2	1	86491100	86491100	*REEP1*	*KDM3A,CHMP3,RNF103-CHMP3,RNF103,RMND5A,CD8A,CD8B,ANAPC1P1,MRPL35*
				19	19	424308	2	86491101	TG	19	1	424309	424309	*SHC2*	*PPAP2C,MIER2,THEG,C2CD4C, ODF3L2,MADCAM1*

[#] Continental group: AMR, SAN, ASN and EUR refer to American, East Asian, South Asian and European, respectively

[*] Population: PEL, ITU, CHB and TSI refer to Peruvian in Lima (Peru), Indian Telugu in the UK, Han Chinese in Beijing (China) and Toscani in Italy, respectively.

[$] der: derivative chromosome

[^] TAD: topological associated domain; domain information from the human IMR90 fibroblast cell line (hg19)

Table 2

Inversions detected in the 1000 Genomes Project

Sample ID	Karyotype	Continental group[#]	Population[*]	Coordinate		Microhomology	Deletion (bp)		Gene disruption	
				Break A[c]	Break B[c]		Break A[c]	Break B[c]	Break A[c]	Break B[c]
NA20759	46,XY,inv(2) (p14p12)	EUR	TSI	67974196	81338340	G	1	0	–	–
				67974194	81338341	–				
HG04152	46,XY,inv(1) (q32.1q32.1)	SAN	BEB	199531693	203428169	–	65	0	–	–
				199531758	203428141	–				
NA18959	46,XY,inv(3) (q26.1q26.1)	ASN	JPT	166243083	166459774	AT	4	0	–	–
				166243087	166459775	G				
NA21133	46,XY,inv(X) (p22.2p22.2)	SAN	GIH	15629309	15687230	T	0	0	–	–
				15629310	15687231	–				

[#] Continental group: EUR, SAN and ASN refer to European, South Asian and East Asian, respectively

[*] Population: TSI, BEB, JPT and GIH refer to Toscani in Italy, Bengali in Bangladesh, Japanese in Tokyo (Japan) and Gujarati Indian in Houston (TX, USA) respectively.

[c] Breakpoint = breakpoint