Published in final edited form as:
Genomics. 2014 April ; 103(4): 276–287. doi:10.1016/j.ygeno.2014.01.001.

One in Four Individuals of African-American Ancestry Harbors a 5.5kb Deletion at chromosome 11q13.1

Kayvan Zainabadi[1,#], **Anuja V. Jain**[1,*], **Frank X. Donovan**[2], **David Elashoff**[3], **Nagesh P. Rao**[4], **Vundavalli V. Murty**[5], **Settara C. Chandrasekharappa**[2], and **Eri S. Srivatsan**[1,±]

[1]Division of General Surgery, Department of Surgery, VAGLAHS West Los Angeles, David Geffen School of Medicine at UCLA, Los Angeles, CA 90073

[2]Genome Technology Branch, National Human Genome Research Institute, National Institutes of Health, Bethesda, MD 20892

[3]Department of Medicine, David Geffen School of Medicine at UCLA, Los Angeles, CA 90025

[4]Department of Pathology and Laboratory Medicine, David Geffen School of Medicine at UCLA, Los Angeles, CA 90025

[5]Department of Pathology and Cell Biology, Columbia University College of Physicians and Surgeons, New York, NY 10032

Abstract

Cloning and sequencing of 5.5kb deletion at chromosome 11q13.1 from the HeLa cells, tumorigenic hybrids and two fibroblast cell lines has revealed homologous recombination between *Alu*Sx and *Alu*Y resulting in the deletion of intervening sequences. Long-range PCR of the 5.5kb sequence in 494 normal lymphocyte samples showed heterozygous deletion in 28.3% of African-American ancestry samples but only in 4.8% of Caucasian samples (p<0.0001). This observation is strengthened by the copy number variation (CNV) data of the HapMap samples which showed that this deletion occurs in 27% of YRI (Yoruba – West African) population but none in non-African populations. The HapMap analysis further identified strong linkage disequilibrium between 5 single nucleotide polymorphisms and the 5.5kb deletion in the people of African

Address for correspondence, Eri S. Srivatsan, Ph.D., Department of Surgery, VAGLAHS/David Geffen School of Medicine at UCLA, Los Angeles, CA 90073, Tel: 310-268-3217, Fax: 310-268-3190, esrivats@ucla.edu.
[#]Current address: The Laboratory of Malaria and Vector Research, National Institute of Allergy and Infectious Diseases, National Institutes of Health, Bethesda, MD 20894
[*]Current address: Department of Pediatrics, Boston Children's Hospital, Boston, MA 02115
[±]Member of Jonsson Comprehensive Cancer Center, UCLA, Los Angeles, CA 90025

Publisher's Disclaimer: This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

Competing Interests

The authors have declared that no competing interests exist

Financial Disclosure

None

Authors' contributions

KZ, SCC, FXD and ESS were responsible for the design and draft of the manuscript. DE performed statistical analysis of the flexibility studies. All authors were involved in the execution of the studies. All authors read and approved the final manuscript.

ancestry. Computational analysis of 175kb sequence surrounding the deletion site revealed enhanced flexibility, low thermodynamic stability, high repetitiveness, and stable stem-loop/hairpin secondary structures that are hallmarks of common fragile sites.

Introduction

We have previously identified a 300kb minimal area of 11q13.1 deletion in cervical tumors, which overlapped with deletions found in other human tumors [1] (Figure 1). Detailed microsatellite analysis of the HeLa cells (a cervical cancer cell line) identified an interstitial deletion of 2.3Mb within the q13.1 locus of one chromosome 11 and a 5.5kb deletion in the second copy of chromosome 11 [1, 2]. Thus, the HeLa cells showed homozygous loss of the 5.5kb sequences. HeLa cell derived tumorigenic hybrid cell lines and a primary cervical tumor also showed homozygous loss of these sequences (represented by the STS marker D11S913) [3]. Surprisingly, we also found heterozygous loss of the same sequences in two fibroblast cell lines (GM00077 and GM00468). However, the molecular nature of the breakpoint sequences leading to these deletions has not yet been determined. As HeLa cells and GM00077 are derived from individuals of African ancestry, it is not known whether the 5.5kb sequence deletion is a cell line phenomenon or is a common event in individuals of this ancestry.

Chromosome 11q13.1 is a region known to contain fragile sites that are prone to breakages. Cytogenetic analysis has revealed two types of fragile sites in the human genome: rare and common. Rare fragile sites are found in less than 5% of cells grown under specific culturing conditions and are molecularly characterized by the expansion of unstable repeats: Folate sensitive rare fragile sites contain CCG repeats, whereas distamycin A and bromodeoxyuridine (BrdU) sensitive sites contain AT-rich mini-satellite sequences [4, 5]. These types of fragile sites only occur in a small percentage of the population, are heritable, and have been associated with diseases such as mental retardation and neurodegenerative diseases. The best examples of rare fragile sites are FRAXA (at Xq27.3), associated with the inactivation of *FMR1*, and FRAXE (at Xq28) associated with the inactivation of *FMR2* [4–6].

In contrast, common fragile sites are found in greater than 5% of treated cells and are considered a normal part of chromosomal structure. Instead of an expansion of triplet repeats, common fragile sites are characterized by repetitive, GC poor and flexible sequences. It has been shown that such sequences facilitate the formation of stable secondary structures during DNA replication, thus causing replication fork stalling and DNA breaks [7–9]. Evidence suggests that a similar mechanism operates at rare fragile sites: the expansion of CCG or AT mini-satellite also facilitates the formation of non-helical secondary structures [7]. Thus the formation of a secondary structure during DNA replication appears to be an important underlying cause of fragility.

We have previously performed sequence analysis and determined that sequences within the cervical cancer deletion locus are GC poor and highly repetitive, similar to those found at other common fragile sites [10]. While such sequences are implicated in genetic instability, evidence for their direct involvement in deletions is not yet known. In this study, we performed molecular and computational analysis to understand the fragile nature of the 11q13.1 locus. We find that 11q13.1 sequences are thermodynamically flexible and unstable, and prone to the formation of complex secondary structures, characteristics that are hallmarks of common fragile sites. We also demonstrate that the HeLa cell 5.5kb deletion is characterized by the presence of two direct *Alu* repeats at the deletion breakpoints, which directly implicates repetitive sequences to the deletion event. We further demonstrate that the heterozygous deletion of the 5.5kb sequence is found in the normal lymphocytes of the general U.S. population and is present in higher frequency in people of African American ancestry.

Material and Methods

Cell Lines and blood samples

Publicly available fibroblast cell lines GM00023, GM00077, GM00468, and GM05399 were obtained from the cell line collection of Coriell Institute for Medical Research (Camden, NJ). While GM00023 was derived from a normal 31-year old Caucasian female, GM05399 was derived from a 1-year old Caucasian male. GM00077 was derived from a 1-year old male of African ancestry affected with the recessive disorder, Tay-Sachs disease. The age and ancestry of the GM00468 cell line derived from a normal male is not known. Cervical cancer cell lines HeLa, C4I, MS751, SiHa, C33A and HT3 were obtained from the ATCC cell line repository (Manassas, VA). HeLa cell derived tumorigenic hybrids were obtained from Dr. Eric J. Stanbridge of UC Irvine, CA. All cell lines were grown in MEM (Minimal Essential Media) medium supplemented with 10% FCS (Fetal Calf Serum). HeLa and C4I contain HPV 18 sequences, and MS751 and SiHa contain HPV 16 sequences. C33A and HT3 cell lines are HPV negative. Blood samples were collected from the blood bank of the VAGLAHS, West Los Angeles using an internal IRB-approved protocol. High molecular weight genomic DNA was isolated using established protocols.

Ethics statement

There was no contact with subjects in the collection of blood samples from the blood bank.

Bacterial artificial chromosomes (BACs)

Bacterial artificial chromosomes (BACs) were isolated from a human genomic BAC library [10, 11]. Sequencing of BACs 41I21 and 152L21 was performed at the University of Oklahoma Genome Center and the sequences of the other BACs were obtained from the NCBI Human Genome database. The Genbank [http://www.ncbi.nlm.nih.gov/]nucleotide accession numbers for the BACs b41I21, b152L21, b755F10, b867G23, and bCTD-3074O7 are AC008102, AC069080, AP000759, AP001107, and AP002748 respectively.

Oligonucleotide primers and polymerase chain reaction (PCR)

The genomic sequence of 700kb of chromosome 11q13.1 was derived from the genome database. Primers were identified from the genomic sequences using the Primer3 program of the MIT Whitehead Institute. Primers identified in our earlier investigations were also used [10]. PCR was performed in 25μl reactions with 100ng of genomic DNA using Amplitaq (Applied Bio-systems, Inc., CA) DNA polymerase. PCR was performed using the established protocol and products were analyzed on 8% polyacrylamide gels, stained with ethidium bromide, and visualized under ultra-violet light [10].

Long-range PCR

For amplification of sequences greater than 1.8kb, a long-range PCR method was employed. The system uses a high fidelity enzyme, Takara Ex Taq DNA Polymerase (PanVera, Inc., WI), longer primer sets (23–26 nucleotides) with higher melting temperatures, and a hot-start PCR method. The forward and reverse primers used for the 7.3kb product are as follows: Forward - 5’ GGA AAG GTC CCA GAA AGA AGA TCA AG 3’ and Reverse - 5’ TGG CCA GTG TGG AGA CCA AGC TGC 3’. Typically, 500ng of genomic DNA or 50ng of BAC DNA were used in a 50μl reaction. The samples were denatured for 3 minutes and a step-down annealing method was used. The cycles consisted of denaturation at 94°C for 45 seconds, annealing at 69°C for 10 cycles, 68°C for 11 cycles, and 66°C for 12 cycles for 30 seconds each, and extension at 72°C for 5 minutes. A final extension was performed at 72°C for 10 minutes. PCR products were analyzed on 1% agarose gels, stained with ethidium bromide, and visualized under ultra-violet light.

Sequencing

PCR products were purified using the QIAquick PCR Purification Kit (Qiagen, CA) and sequenced using the Sanger Dideoxy Sequencing Protocol.

Sequence Analysis

Repeat and GC content of sequences was calculated using the REPEATMASKER program (http://www.repeatmasker.org/) which screens DNA sequences against a library of repetitive elements. The slow/most-sensitive setting was used for all analyses, with large sequences divided into 20kb segments for analysis. DNA helix flexibility and stability was calculated using the FLEXSTAB (http://bioinfo.md.huji.ac.il/marg/Flexstab/) and TWISTFLEX (http://margalit.huji.ac.il/TwistFlex/) programs [7, 12]. In brief, the flexibility parameter that was used measures potential local variation in the DNA structure, expressed as fluctuations in the TWIST angle [13]. The helix stability is based on the sequence-dependent free energy values of the helix-to-coil transition and was expressed in Kcal/mol. Dinucleotide energy values were determined along the window and averaged by the window size. Secondary DNA structure was identified using the MFOLD program (http://mfold.rna.albany.edu/?q=mfold) [13, 14].

Southern blot hybridization analysis

Genomic DNAs were digested with *EcoR1* restriction enzyme, separated on 1.0% agarose gels in Tris-acetate buffer, pH 7.85 and the DNA fragments were transferred onto nylon membranes in 20 X SSC buffer, pH 7.0. Blots were hybridized using 32p-labelled probes in the presence of salmon sperm non specific DNA, washed twice in 2 X SCC, 0.2% SDS buffer, twice in 0.2 X SSC, 0.1% SDS buffer at 65 °C for 30 minutes each time and then exposed to x-ray film for 2 to 7 days [2].

Northern blot hybridization analysis

Pre-made normal tissue and cancer cell line poly (A) RNA blots were purchased from Clontech, Inc. (Palo Alto, CA) and Ambion, Inc. (Austin, TX). Hybridizations were performed with 32p-labelled probes at 42°C and exposed to X-ray films for 2–7 days as described [10].

Fluorescence in situ hybridization

The 5.5kb sequence or the control probes (6.6kb *HRAS* probe localized to the short arm of chromosome 11 (11p15) and a 5kb probe, representing *MLL* gene at 11q23) were labelled and hybridized to metaphase spreads as previously described [2, 15]. Briefly, metaphase spreads were prepared by standard cytogenetic procedures. Probes were labeled by nick translation with digoxigenin- 11-dUTP (Boehringer Mannheim, Indianapolis, IN) following the manufacturer's protocol. Hybridization and washes were performed under identical conditions of stringency. At least 20 metaphase spreads were analyzed.

UCSC Genome browser analysis for the detection of copy number variations (CNVs)

The region of chromosome 11, 65689011–65696252 (hg18), was queried in the UCSC Genome Browser (http://genome.ucsc.edu/). The 'DGV Structural Variation' track, under 'Variation and Repeats', was turned on to view CNVs previously reported to the Database of Genomic Variants. The source (publication) and population distribution of each reported CNV overlapping the queried region was obtained using the associated variation ID.

HapMap Genotype Data and LD Analysis

The HapMap Phase III SNP genotype data was downloaded in PLINK (PED/MAP) format from the International HapMap Project website (http://hapmap.ncbi.nlm.nih.gov/downloads/index.html.en). The genotype data of chromosome 11 (71,098 SNPs) for 166 Yoruba individuals was extracted for analysis.

LD between the deletion and the HapMap Phase III SNPs on chromosome 11 was calculated using PLINK [16] (http://pngu.mgh.harvard.edu/~purcell/plink/). The resulting r^2values for SNPs within 1MB of the deletion were plotted using ggplot2 [17] (http://ggplot2.org/) graphics package for R.

SNP Annotation and Proxy Search (SNAP) [18], from the Broad Institute, was used to evaluate LD within different populations from the HapMap3 (Phase II) data. Six populations were evaluated: three of African ancestry (Yoruba (YRI), Luhya in Webuye, Kenya (LWK), & African ancestry in South-west USA (ASW)) and three of non-African ancestry (Utah residents with Northern and Western European ancestry (CEU), Tuscans in Italy (TSI), & Japanese in Tokyo, Japan/Han Chinese in Beijing, China (JPT/CHB)). The r^2 threshold for LD was set to 0.8 with a distance limit of 500kb from the proxy SNP.

Haploview [19] (http://www.broadinstitute.org/scientific-community/science/programs/medical-and-population-genetics/haploview/haploview) was used to assess LD patterns across a 200kb region of chromosome 11 overlapping the region of interest. Plots were generated for the 6 populations (YRI, LWK, ASW, CEU, TSI, & JPT/CHB) to compare LD patterns between populations of African ancestry and populations of non-African ancestry.

Statistical methods

We evaluated the significance of alterations in the flexibility values observed in the 175kb region, between 225–400kb of the 700kb 11q13.1 genomic region (Figure 1) compared to the region outside this interval using a permutation simulation. The simulation was designed to evaluate the likelihood of identifying a region of this size (175kb) of altered flexibility values assuming we were free to choose the optimal region of difference. In each run of the simulation we first permuted the location labels for each data point. Second, we computed t-statistics comparing sequential regions of size 175kb versus all other regions in the data. For each run this involved examining 30 possible regions of size 175kb starting at each multiple of 25kb. Next, we identified the maximal t-statistic for each simulation run across all 30 regions evaluated. The set of the maximal t-statistics from the set of permutations then formed the permutation distribution for our observed t-statistic. The observed t-statistic for comparing flexibilty values in the 175kb region versus the remaining region had a value 256.6. In 1000 simulations the largest maximal permuted t-statistic was 4.2. This indicates that the p-value of the observed finding is less than 0.001. This permutation simulation controls for the possible bias that could have arisen as the specific region examined was not apriori defined.

Results

Characterization of the HeLa cell 5.5kb deletion breakpoint at chromosome 11q13.1

We have previously localized the HeLa cell homozygous deletion to a ~5.5kb interval of chromosome 11q13.1 [3] (Figure 1). While one of the deletions was 5.5kb, the other deletion was 2.3Mb. To identify the precise deletion breakpoints of the 5.5kb sequence, we designed primers from outside of the deletion (approximately 1kb from each of the retained markers) and attempted amplification of a 7.3kb genomic sequence (Supplemental Figure 1). PCR on the BAC 152L21, which contains the sequences of this area, yielded the expected 7.3kb PCR product (Figure 2A). The 7.3kb product was also observed in four normal lymphocytes, and two different primary fibroblast cell lines, GM00023 and GM05399 (Figures 2A and 2B). Two normal fibroblast cell lines (GM00077 and GM00468) yielded both a 7.3kb product and a 1.8kb product, suggesting heterozygous deletion of the 5.5kb sequence (Figure 2A). However, seven different HeLa cell lines and two HeLa cell derived tumorigenic hybrid cell lines yielded only the smaller 1.8kb product, indicating homozygous deletion of the 5.5kb sequence (Figure 2C). Analysis of five other cervical cancer cell lines showed a normal 7.3kb product (Figure 2D), indicating that 5.5kb deletion is specific to the HeLa cells.

To confirm the PCR results, we performed Southern blot hybridizations. A 2.5kb PCR probe representing the sequences within and outside of the deletion gave rise to an expected 5.5kb *EcoRI* fragment in two normal fibroblast cell lines (GM00023 and GM05339) (Figure 2E). However, only a deletion specific 7kb fragment, indicative of homozygous deletion, was seen in the HeLa cells. The fibroblast cell lines GM00468 and GM00077 yielded the expected 5.5kb and the 7kb fragments, confirming the heterozygous deletion of the 5.5kb sequence in these two cell lines. As a further confirmation, FISH (Fluorescence in situ hybridization) studies were undertaken using the 5.5kb probe. The analysis yielded hybridization of the 5.5kb probe to chromosome 11q13.1 sites in a normal lymphocyte sample (Figure 2F). The fibroblast cell line GM00077 showed 5.5kb probe hybridization to only one chromosome 11 in all the metaphase spreads examined clearly indicating heterozygous deletion of these sequences. While the control probes *HRAS* (11p15) and *MLL* (11q23) hybridized to two chromosome 11s of the HeLa cells, there was no hybridization of the 5.5kb (11q13.1) probe confirming homozygous deletion of these sequences in the HeLa cells.

To characterize the deletion breakpoints, we sequenced the 1.8kb PCR product from the HeLa cells; HeLa cell derived tumorigenic hybrids, and the fibroblast cell lines GM00468 and GM00077. A sequence comparison using the BLAST and Repeat Masker programs revealed a 308bp *Alu* repeat at both the centromeric and the telomeric ends of all the deletions. The two repeats are 85% homologous (with the core *Alu* sequence being 95% identical) and occur in the direct orientation at the break site. Using SNPs from the two *Alu*s we were able to pinpoint the precise sites of breakage to within a few base pairs, indicating that the deletion is indeed 5.5kb (Figure 1). More importantly, sequences at breakpoint sites showed that the deleted samples contained a novel hybrid *Alu* (derived from *Alu*Sx at the centromeric end and *Alu*Y at telomeric end – Figure 1), providing strong evidence for a homologous recombination event giving rise to the observed interstitial deletion. The presence of similar deletion breakpoints in all examined samples indicated that this is either a deletion hotspot or a heritable genetic element.

Identification of the 5.5kb heterozygous deletion in the general population

To determine whether the general population also harbors this deletion, we assayed blood samples obtained from the VA West Los Angeles for the deletion using a long-range PCR protocol. We found that 47 of 494 (9.5%) of the lymphocyte samples showed heterozygous deletion, indicating that nearly 1/10 of the U.S. general population harbors this deletion in one of their chromosome 11s (Table 1, Figure 3A). Semi-quantitative PCR analysis and Southern blotting performed on 100 of these samples confirmed the deletion in these lymphocyte samples (Figures 3B and 3C).

Enrichment of the 5.5kb heterozygous deletion in individuals of African-American ancestry

Since the HeLa cells and GM00077 are both of African ancestry, we hypothesized that this deletion might be found in greater frequency in the African-American (AA) population. We thus analyzed the lymphocyte samples for the relationship to ethnicity. We observed the deletion in 26/92 (28.3%) of AA DNA samples while only in 6/126 (4.7%) of Caucasian samples indicating an enrichment of the deletion in people of AA ancestry ($p<0.0001$) (Table 1). To confirm this observation, we analyzed another subset of 70 lymphocyte samples specifically from people of AA origin by semi-quantitative PCR and Southern blot hybridization. We found heterozygous deletion in 20 (28.6%) of this test case (Figures 3D and 3E), providing further proof that the 5.5kb deletion segregates within this population. These results, therefore, indicate that African-American individuals show a 3-fold enrichment of the deletion in comparison to the general population and a 7-fold enrichment in comparison to the Caucasian population. In the second subset of 70 AA samples we found two of five individuals treated for cancer contained homozygous deletion of the 5.5 kb sequence (Supplemental Table 1). Homozygous deletion was not observed in 15 of the non-tumor samples. Although there is also a higher frequency of heterozygous deletion associated with cardiomyopathy, a large dataset of AA samples is required to confirm and extend an association to a clinical phenotype in the AA population.

Given the high frequency of the deletion in lymphocyte samples, we reasoned that the deletion likely exists as a heritable genetic element (as opposed to a reoccurring *de novo* event). To test this possibility, we analyzed a two-generation family obtained from the Coriell Institute (Camden, NJ) and found the father and children to harbor the heterozygous deletion indicating that the deletion is heritable (Figure 3F). This family was found to be of African ancestry.

Identification of the 5.5kb heterozygous deletion in the YRI population of the HapMap studies

To determine whether the identified deletion interval [chr11:65690475–65695907 (hg18)] has previously been reported in the Database of Genomic Variants (DGV), the region was queried in the UCSC Genome Browser (http://genome.ucsc.edu/) using the 'DGV Structural Variation' track. We noticed that the 5.5kb deletion was seen in previous HapMap investigations [20–25]. The variation IDs were used to gather information about each copy number variation (CNV) from the Database of Genomic Variants (http://projects.tcag.ca/variation/), including the publication source, size, frequency, and population distribution. As reported by Conrad et al [20], from their analysis of 450 HapMap individuals (Figure 4 – variation ID 65975), the 5.5kb deletion was observed in 27% (49 of 180) of the YRI population, in 0% (0/180) of the CEU population, and in 0% (0/90) of the JPT/CHB population (Figure 4). The population distribution of this deletion within the HapMap dataset is consistent with our results that the 5.5kb deletion is prevalent in people of African ancestry.

To determine whether the 5.5kb deletion in this region was a *de novo* event or part of a heritable event, the pedigree information for the 166 YRI Hapmap samples was obtained from the Hapmap ftp site (ftp://ftp.ncbi.nlm.nih.gov/hapmap/phase_3/relationships_w_pops_051208.txt) and the deletion status for each individual was determined based on the data from Conrad et al [20]. Thirteen of the full trio families showed that this deletion is present in the child (Supplemental Table 2). In all thirteen families, at least one parent carries the deletion as well, providing evidence toward the heritable nature of this deletion event. In one family Y010, the child carries the deletion. However, it is not a full trio because the paternal data is not available. From this information, it can be inferred that the deletion exists predominantly as a stable heritable element supporting our results from the two-generation family.

Linkage Disequilibrium (LD) analysis

Analysis of the HapMap populations has identified the 5.5kb deletion within the 1st intron of the *PACS1* gene, a cytoplasmic trafficking protein [26, 27], on chromosome 11 to be prevalent in the Yoruba African population. To identify any SNPs on chromosome 11 which may be in strong linkage disequilibrium with the 5.5kb deletion, LD analysis was performed with PLINK. The data from Conrad et al [20] provided the copy number of this variation for each individual (Supplemental Table 3). Using this information, knowing that there were not any SNPs residing within the deletion region, each copy number was attributed a genotype for LD analysis: 0 copies = A/A, 1 copy = A/G, 2 copies = G/G. The results included five SNPs on chromosome 11 with r^2> 0.8, and all five SNPs appear to be in perfect LD with the deletion, $r^2 = 1$ (Figure 5A–B). Next, using SNAP, we assessed the LD between these 5 SNPs in other populations. One of the five SNPs (rs11227418) served as a proxy to evaluate the LD in six HapMap (Phase III) populations (three of African ancestry - YRI, LWK, & ASW, and three of non-African ancestry - CEU, TSI, & JPT/CHB). In YRI, LWK, & ASW, all four SNPs were in perfect LD with rs11227418, having r^2 values equal to 1 (Supplemental Figure 2A). However, in the three non-African populations, CEU, TSI, & ASW, there were no data to be plotted because, as it turns out, the five SNPs are monomorphic in these populations and LD cannot be evaluated (Supplemental Figure 2B, Supplemental Table 4).

To get a more in-depth explanation for this, we used Haploview to assess LD in a 200kb region encompassing the 5.5kb deletion, chr11:65,600,000–65,800,000, in all six populations. The haploview plots showed a distinctly different LD pattern in the populations of African ancestry compared to non- African populations (Supplemental Figure 3). There are a greater number of monomorphic SNPs in this region in non-African populations, including the 5 SNPs in strong LD with the 5.5kb deletion. The population specific LD pattern in this region supports the observation of a population specific variation in individuals of African ancestry. The allele frequencies of the five SNPs, obtained from dbSNP (http://ncbi.nlm.nih.gov/projects/SNP/), also convey the population specific variation of these five polymorphisms (Supplemental Table 4). The two West African populations, YRI & LWK, contain all three genotypes, whereas the Africans of Southwest USA population contains two of the genotypes, and the non-African populations are monomorphic.

Evaluation of the specific genotypes in each YRI individual for the five SNPs in strong LD with the deletion showed that a specific haplotype could be observed in individuals carrying the deletion versus individuals that did not carry the deletion (Supplemental Table 3). 122 individuals that did not carry the deletion were homozygous G/G for all five SNPs. All individuals that carried the deletion had a common specific haplotype rs11227418-A, rs495684-T, rs561948-A, rs501690-A, rs501979-A. A haplotype consisting of A/G, T/G, A/G, A/G and A/G for the five SNPs is thus observed in 42 individuals with a heterozygous deletion and a haplotype of A/A, T/T, A/A, A/A and A/A is observed in four individuals (NA18867, NA19146, NA19185, and NA19239) with homozygous deletion of the 5.5kb sequence. The deletion frequencies agreed with the frequencies expected from the Hardy-Weinberg equilibrium calculations. The expected frequencies of heterozygous and homozygous deletions are 42.46 and 3.76 respectively (Supplemental Figure 4). The observed frequencies are 42 and 4 indicating a clear agreement to the expected values.

Identification of fragile-like sequences at the 11q13.1 locus

We have previously reported that the sequences surrounding the homozygous deletion are GC poor and highly repetitive, similar to those found at other common fragile sites (Figure 6A) [10]. Since thermodynamically unstable and flexible sequences are also associated with common fragile sites, we analyzed a 700kb region encompassing 300kb minimal area of primary cervical tumor deletion using the FLEXSTAB and TWISTFLEX computational programs. These programs have been used in previous fragile site studies to determine potential local variations in DNA structure and stability based on sequence dependent dinucleotide twist angles and free energy values [7, 12, 13]. Our analysis shows significant DNA instability and flexibility for a 175kb region surrounding the 5.5kb deletion (near marker D11S913) similar to that observed at other common fragile sites ($p<0.001$) (Figures 6B and 6C). Notably, this 175kb locus overlaps with a 120kb region we have previously shown to contain the lowest GC content and the highest overall repeat content for the entire 700kb analyzed sequence [10].

Since such sequences can result in the formation of secondary structures during DNA replication that are thought to be the molecular basis of fragility, we analyzed the 120kb region for such structures using the MFold program [7, 9]. Our analysis reveals a large number of inverted repeats (including within the 5.5kb deletion locus) that are predicted to form highly stable stem-loop/hair-pin structures (Figure 7A–C). Due to the highly repetitive nature of the locus, secondary structures of unprecedented size and complexity occur, often spanning hundreds or even thousands of base pairs (Figure 7B–D).

Our attempts at the functional analysis of the 5.5kb sequence were unsuccessful due to the instability of this sequence cloned in various vectors. Also, we could not identify coding sequences for a miRNA within this 5.5kb sequence. Thus, we focused our attention on the expression of *PACS1* gene mapped to this locus. We have previously shown that the expected 4.5kb transcript of the *PACS1* gene was present in the HeLa cells and other tumor cell lines [10]. We have also shown the expression of an 8.0kb transcript in the HeLa cells and tumor cell lines indicating the presence of spliced variants of the *PACS1* gene. Due to low level expression of this 8.0kb variant, we could not clone this transcript. A 2.5kb and a 121bp probe mapped 16kb and 1kb proximal to 5.5kb deletion showed the presence of transcripts in normal tissues confirming the possibility that spliced variants of *PACS1* exist in normal tissues (Supplemental Figure 5). Protein analysis has shown that the *PACS1* protein is over expressed in cervical cancer cell lines and primary tumors (data not shown). Thus, the available data suggests that *PACS1* is aberrantly expressed in the HeLa cells and primary cervical tumors.

Discussion

Regions enriched for repetitive elements have previously been shown to be genetically unstable [28–30]. *Alu* and LINE elements have been implicated in genomic rearrangements leading to numerous human diseases [31, 32]. Thermodynamically unstable and flexible

DNA sequences are more prone to adopt single-stranded conformation and form secondary structures such as hairpins or stem-loops [7]. Such structures are stabilized by the presence of inverted repeats and can slow or stall the replication fork during DNA replication. This can produce double stranded breaks and other genetic abnormalities, and is thought to be molecular mechanisms behind common fragile sites [7, 33]. Cells repair such potentially fatal lesions by NHEJ (non-homologous end joining) or homologous recombination [34, 35]. However, direct repeats also provide ready targets for illegitimate homologous recombination, which can lead to deletions and genomic rearrangements.

Thus it is likely that the repeat rich sequences cause genetic instability at this locus in a two-fold manner: 1) Inverted repeats stabilize secondary single-stranded DNA configuration resulting in replication fork stalling and DNA breaks; 2) Direct repeats act as sites of illegitimate homologous recombination in the repair process, resulting in deletions and genomic rearrangements.

It is interesting to note that cervical and nasopharyngeal tumors with 11q13.1 deletions have p53 and/or other repair proteins inactivated, and thus might be prone to such rearrangements. For instance, cells lacking p53 have been shown to have a substantially higher frequency of illegitimate homologous recombination which can cause deletions such as the one in the HeLa cells [36]. Further, loss of the ATR (ATM related) kinase, another important DNA repair protein, appears to be sufficient for induction of common fragile sites under normal conditions [37]. Our preliminary studies indicate that *PACS1* protein might be playing a role in DNA replication and overexpression of this protein could therefore be a cause of tumor associated genomic rearrangements (data not shown).

Here we provide evidence that the direct repeats have recombined in the HeLa cells, resulting in the deletion of the interstitial sequences and giving rise to a hybrid *Alu* containing SNPs from the distal and proximal *Alu* repeats. Moreover, we show that the surrounding sequences are prone to the formation of singe-stranded secondary structures. Additionally, we have found that the 5.5kb deletion also occurs in 9% of the general US population but in 28% of people of African-American ancestry and is heritable. These results are also supported by the observation of this deletion in YRI individuals of the HapMap studies.

Since chromosome 11q13.1 contains fragile sites, an inevitable question arises: do the 120kb sequences represent common or rare (or both) fragile sites? From our data, it appears that these sequences have all of the major hallmarks of a common fragile site: high frequency of repeats, low GC content, flexible sequences, potential secondary structures, and high incidence of genetic alterations in cancer. Moreover, much like the FRA3B and FRA16D common fragile sites, the repeat rich and GC poor nature of the 120 kb sequences are also conserved in the mouse and rat [10, 35,38]. Therefore, our results indicate that the deletion sequences identified here most likely represent a common fragile site. This is further strengthened by the report that the rare fragile site FRA11A is caused by expansion of a CCG repeat in the 5' end of the *C11orf80* gene, located 660kb telomeric from D11S913 [39]. Further, loss of heterozygosity studies has identified a number of tumor suppressor genes at chromosome 11q13.1 [40–43]. We have also mapped cystatin E/M tumor suppressor gene to sequences 200kb proximal to 5.5kb sequence and have observed homozygous deletion of exonic sequences in primary tumors confirming fragility of this region [43]. While the 5.5kb deletion may be heritable in the AA population, sequences surrounding this site may be fragile in the general population

One unanswered question still remains: exactly how far does the fragility at 11q13.1 extend? Experimental data suggest that it is not limited to the 120kb interval examined. For instance, the 300kb minimal area of cervical cancer deletion extends an additional 175kb centromeric to this region. Moreover, viral integrations, which represent possible locations of common fragile sites, have also been reported for different regions of 11q13.1. Notably, the number one hot spot of integration for HIV and ERV9 (endogenous retro virus 9) has been mapped to a region approximately 1.5Mb centromeric of D11S913 [44, 45]. Moreover, the Hepatitis B Virus (HBV) has also been shown to repeatedly integrate at a site approximately 1.1Mb telomeric of D11S913 [46]. Since viral integrations are hallmarks of common fragile sites, these results suggest that the common fragile site FRA11H at 11q13.1 (http://www.ncbi.nlm.nih.gov/projects/mapview/maps.cgi?TAXID=9606&CHR=11&MAPS=ugHs%2Cgenes%2Cgenec-r&QSTR=2446%5Bgene_id%5D&QUERY=uid%28-2146581571%29&BEG=11q13.1&END=11q13.1&oview=default) could lie in a 2.6Mb distance between the two viral integration sites. Sequence analysis of this larger interval indeed reveals the presence of two additional flexible/unstable regions (75kb and 100kb in length) that are also highly repetitive and GC poor (data not shown). These sites occur 650kb and 1Mb telomeric of D11S913 and show remarkable similarities to the unstable 120kb sequences reported in this investigation. Additionally, sequences near the viral integration sites also appear to be particularly flexible: a 50kb region near the HBV integration site and a 100kb region near the HIV/ERV9 integration site each contain five peaks of significant flexibility. It is likely that these interrupted local hotspots of flexible/unstable and repetitive sequences in the 2.6Mb regions contribute to the overall fragility of the 11q13.1 locus.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

We are deeply indebted to Neta Ben-Porat from Dr. Batsheva Kerem's group at The Hebrew University for her generous technical assistance with the FLEXSTAB/TWISTFLEX analysis, and to Dr. Michael Zuker of Washington University for his assistance with the MFold program. We would also like to thank Dr. Marilene Wang for the proof reading of the manuscript. FXD and SCC gratefully acknowledge the Intramural Program of the National Human Genome Research Institute, Bethesda, MD. This work was supported by: UCLA Student Research Program; American Association for Cancer Research (AACR); Frederick R. Waingrow; and Howard Hughes research fellowships to KZ and by funds from VAGLAHS, West Los Angeles Surgical Education Research Center; and Merit grant from the Veterans Administration, Washington, DC to ESS.

References

1. Srivatsan ES, Chakrabarti R, Zainabadi K, Pack SD, Benyamini P, Mendonca MS, Yang PK, Kang K, Motamedi D, Sawicki MP, Zhuang Z, Jesudasan RA, Bengtsson U, Sun C, Roe BA, Stanbridge EJ, Wilczynski SP, Redpath JL. Localization of deletion to a 300 Kb interval of chromosome 11q13 in cervical cancer. Oncogene. 2002; 21:5631–5642. [PubMed: 12165862]

2. Srivatsan ES, Bengtsson U, Manickam P, Benyamini P, Chandrasekharappa SC, Sun C, Stanbridge EJ, Redpath JL. Interstitial deletion of 11q13 sequences in HeLa cells. Genes, chromosomes & cancer. 2000; 29:157–165. [PubMed: 10959095]
3. Mendonca MS, Farrington DL, Mayhugh BM, Qin Y, Temples T, Comerford K, Chakrabarti R, Zainabadi K, Redpath JL, Stanbridge EJ, Srivatsan ES. Homozygous deletions within the 11q13 cervical cancer tumor-suppressor locus in radiation-induced, neoplastically transformed human hybrid cells. Genes, chromosomes & cancer. 2004; 39:277–287. [PubMed: 14978789]
4. Sutherland GR. Rare fragile sites. Cytogenetic and genome research. 2003; 100:77–84. [PubMed: 14526166]
5. Durkin SG, Glover TW. Chromosome fragile sites. Annual review of genetics. 2007; 41:169–192.
6. Lukusa T, Fryns JP. Human chromosome fragility. Biochimica et biophysica acta. 2008; 1779:3–16. [PubMed: 18078840]
7. Zlotorynski E, Rahat A, Skaug J, Ben-Porat N, Ozeri E, Hershberg R, Levi A, Scherer SW, Margalit H, Kerem B. Molecular basis for expression of common and rare fragile sites. Molecular and cellular biology. 2003; 23:7143–7151. [PubMed: 14517285]
8. Schwartz M, Zlotorynski E, Kerem B. The molecular basis of common and rare fragile sites. Cancer letters. 2006; 232:13–26. [PubMed: 16236432]
9. Burrow AA, Marullo A, Holder LR, Wang YH. Secondary structure formation and DNA instability at fragile site FRA16B. Nucleic acids research. 2010; 38:2865–2877. [PubMed: 20071743]
10. Zainabadi K, Benyamini P, Chakrabarti R, Veena MS, Chandrasekharappa SC, Gatti RA, Srivatsan ES. A 700-kb physical and transcription map of the cervical cancer tumor suppressor gene locus on chromosome 11q13. Genomics. 2005; 85:704–714. [PubMed: 15885497]
11. Kim UJ, Birren BW, Slepak T, Mancino V, Boysen C, Kang HL, Simon MI, Shizuya H. Construction and characterization of a human bacterial artificial chromosome library. Genomics. 1996; 34:213–218. [PubMed: 8661051]
12. Mishmar D, Rahat A, Scherer SW, Nyakatura G, Hinzmann B, Kohwi Y, Mandel-Gutfroind Y, Lee JR, Drescher B, Sas DE, Margalit H, Platzer M, Weiss A, Tsui LC, Rosenthal A, Kerem B. Molecular characterization of a common fragile site (FRA7H) on human chromosome 7 by the cloning of a simian virus 40 integration site. Proceedings of the National Academy of Sciences of the United States of America. 1998; 95:8141–8146. [PubMed: 9653154]
13. Sarai A, Mazur J, Nussinov R, Jernigan RL. Sequence dependence of DNA conformational flexibility. Biochemistry. 1989; 28:7842–7849. [PubMed: 2611216]
14. Zuker M. Mfold web server for nucleic acid folding and hybridization prediction. Nucleic acids research. 2003; 31:3406–3415. [PubMed: 12824337]
15. Winokur ST, Bengtsson U, Vargas JC, Wasmuth JJ, Altherr MR, Weiffenbach B, Jacobsen SJ. The evolutionary distribution and structural organization of the homeobox-containing repeat D4Z4 indicates a functional role for the ancestral copy in the FSHD region. Human molecular genetics. 1996; 5:1567–1575. [PubMed: 8894690]
16. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MA, Bender D, Maller J, Sklar P, de Bakker PI, Daly MJ, Sham PC. PLINK: a tool set for whole-genome association and population-based linkage analyses. American journal of human genetics. 2007; 81:559–575. [PubMed: 17701901]
17. Wickham, H. ggplot2; Elegant Graphics for Data Analysis. New York: Springer; 2009.
18. Johnson AD, Handsaker RE, Pulit SL, Nizzari MM, O'Donnell CJ, de Bakker PI. SNAP: a web-based tool for identification and annotation of proxy SNPs using HapMap. Bioinformatics. 2008; 24:2938–2939. [PubMed: 18974171]
19. Barrett JC. Haploview: Visualization and analysis of SNP genotype data. Cold Spring Harbor protocols. 2009; 2009 pdb ip71.
20. Conrad DF, Pinto D, Redon R, Feuk L, Gokcumen O, Zhang Y, Aerts J, Andrews TD, Barnes C, Campbell P, Fitzgerald T, Hu M, Ihm CH, Kristiansson K, Macarthur DG, Macdonald JR, Onyiah I, Pang AW, Robson S, Stirrups K, Valsesia A, Walter K, Wei J, C. Wellcome Trust Case Control. Tyler-Smith C, Carter NP, Lee C, Scherer SW, Hurles ME. Origins and functional impact of copy number variation in the human genome. Nature. 2010; 464:704–712. [PubMed: 19812545]

21. Kidd JM, Cooper GM, Donahue WF, Hayden HS, Sampas N, Graves T, Hansen N, Teague B, Alkan C, Antonacci F, Haugen E, Zerr T, Yamada NA, Tsang P, Newman TL, Tuzun E, Cheng Z, Ebling HM, Tusneem N, David R, Gillett W, Phelps KA, Weaver M, Saranga D, Brand A, Tao W, Gustafson E, McKernan K, Chen L, Malig M, Smith JD, Korn JM, McCarroll SA, Altshuler DA, Peiffer DA, Dorschner M, Stamatoyannopoulos J, Schwartz D, Nickerson DA, Mullikin JC, Wilson RK, Bruhn L, Olson MV, Kaul R, Smith DR, Eichler EE. Mapping and sequencing of structural variation from eight human genomes. Nature. 2008; 453:56–64. [PubMed: 18451855]

22. Shaikh TH, Gai X, Perin JC, Glessner JT, Xie H, Murphy K, O'Hara R, Casalunovo T, Conlin LK, D'Arcy M, Frackelton EC, Geiger EA, Haldeman-Englert C, Imielinski M, Kim CE, Medne L, Annaiah K, Bradfield JP, Dabaghyan E, Eckert A, Onyiah CC, Ostapenko S, Otieno FG, Santa E, Shaner JL, Skraban R, Smith RM, Elia J, Goldmuntz E, Spinner NB, Zackai EH, Chiavacci RM, Grundmeier R, Rappaport EF, Grant SF, White PS, Hakonarson H. High-resolution mapping and analysis of copy number variations in the human genome: a data resource for clinical and research applications. Genome research. 2009; 19:1682–1690. [PubMed: 19592680]

23. Wang K, Li M, Hadley D, Liu R, Glessner J, Grant SF, Hakonarson H, Bucan M. PennCNV: an integrated hidden Markov model designed for high-resolution copy number variation detection in whole-genome SNP genotyping data. Genome research. 2007; 17:1665–1674. [PubMed: 17921354]

24. Cooper GM, Zerr T, Kidd JM, Eichler EE, Nickerson DA. Systematic assessment of copy number variant detection via genome-wide SNP genotyping. Nature genetics. 2008; 40:1199–1203. [PubMed: 18776910]

25. Matsuzaki H, Wang PH, Hu J, Rava R, Fu GK. High resolution discovery and confirmation of copy number variants in 90 Yoruba Nigerians. Genome biology. 2009; 10:R125. [PubMed: 19900272]

26. Scott GK, Gu F, Crump CM, Thomas L, Wan L, Xiang Y, Thomas G. The phosphorylation state of an autoregulatory domain controls PACS-1-directed protein traffic. The EMBO journal. 2003; 22:6234–6244. [PubMed: 14633983]

27. Youker RT, Shinde U, Day R, Thomas G. At the crossroads of homoeostasis and disease: roles of the PACS proteins in membrane traffic and apoptosis. The Biochemical journal. 2009; 421:1–15. [PubMed: 19505291]

28. Smith DI, Huang H, Wang L. Common fragile sites and cancer (review). International journal of oncology. 1998; 12:187–196. [PubMed: 9454904]

29. Ried K, Finnis M, Hobson L, Mangelsdorf M, Dayan S, Nancarrow JK, Woollatt E, Kremmidiotis G, Gardner A, Venter D, Baker E, Richards RI. Common chromosomal fragile site FRA16D sequence: identification of the FOR gene spanning FRA16D and homozygous deletions and translocation breakpoints in cancer cells. Human molecular genetics. 2000; 9:1651–1663. [PubMed: 10861292]

30. Calabretta B, Robberson DL, Barrera-Saldana HA, Lambrou TP, Saunders GF. Genome instability in a region of human DNA enriched in Alu repeat sequences. Nature. 1982; 296:219–225. [PubMed: 6278320]

31. Kazazian HH Jr, Goodier JL. LINE drive. retrotransposition and genome instability. Cell. 2002; 110:277–280. [PubMed: 12176313]

32. Kolomietz E, Meyn MS, Pandita A, Squire JA. The role of Alu repeat clusters as mediators of recurrent chromosomal aberrations in tumors. Genes, chromosomes & cancer. 2002; 35:97–112. [PubMed: 12203773]

33. Cimprich KA. Fragile sites: breaking up over a slowdown. Current biology : CB. 2003; 13:R231–R233. [PubMed: 12646149]

34. Michel B, Flores MJ, Viguera E, Grompone G, Seigneur M, Bidnenko V. Rescue of arrested replication forks by homologous recombination. Proceedings of the National Academy of Sciences of the United States of America. 2001; 98:8181–8188. [PubMed: 11459951]

35. Krummel KA, Denison SR, Calhoun E, Phillips LA, Smith DI. The common fragile site FRA16D and its associated gene WWOX are highly conserved in the mouse at Fra8E1. Genes, chromosomes & cancer. 2002; 34:154–167. [PubMed: 11979549]

36. Gebow D, Miselis N, Liber HL. Homologous and nonhomologous recombination resulting in deletion: effects of p53 status, microhomology, and repetitive DNA length and orientation. Molecular Cell Biology. 2000; 20:4028–4035.

37. Casper AM, Nghiem P, Arlt MF, Glover TW. ATR regulates fragile site stability. Cell. 2002; 111:779–789. [PubMed: 12526805]

38. Matsuyama A, Shiraishi T, Trapasso F, Kuroki T, Alder H, Mori M, Huebner K, Croce CM. Fragile site orthologs FHIT/FRA3B and Fhit/Fra14A2: evolutionarily conserved but highly recombinogenic. Proceedings of the National Academy of Sciences of the United States of America. 2003; 100:14988–14993. [PubMed: 14630947]

39. Debacker K, Winnepenninckx B, Longman C, Colgan J, Tolmie J, Murray R, van Luijk R, Scheers S, Fitzpatrick D, Kooy F. The molecular basis of the folate-sensitive fragile site FRA11A at 11q13. Cytogenetic and Genome research. 2007; 119:9–14. [PubMed: 18160775]

40. Tanaka C, Yoshimoto K, Yamada S, Nishioka H, Ii S, Moritani M, Yamaoka T, Itakura M. Absence of germ-line mutations of the multiple endocrine neoplasia type 1 (MEN1) gene in familial pituitary adenoma in contrast to MEN1 in Japanese. Journal of Clinical Endocrinology & Metabolism. 1998; 83:960–965. [PubMed: 9506756]

41. Cheng Y, Chakrabarti R, Garcia-Barcelo M, Ha TJ, Srivatsan ES, Stanbridge EJ, Lung ML. Mapping of nasopharyngeal carcinoma tumor suppressive activity to a 1.8 megabase region of chromosome 11q13. Genes Chromosomes & Cancer. 2002; 34:97–103. [PubMed: 11921287]

42. Yi Y, Nowak NJ, Pacchia AL, Morrison C. Chromosome 11 genomic changes in parathyroid adenoma and hyperplasia: array CGH, FISH, and tissue microarrays. Genes Chromosomes & Cancer. 2008; 47:639–648. [PubMed: 18398822]

43. Veena MS, Lee G, Keppler D, Mendonca MS, Redpath JL, Stanbridge EJ, Wilczynski SP, Srivatsan ES. Inactivation of cystatin M tumor suppressor gene in cervical cancer. Genes Chromosomes & Cancer. 2008; 47:740–754. [PubMed: 18506750]

44. Schroder AR, Shinn P, Chen H, Berry C, Ecker JR, Bushman F. HIV-1 integration in the human genome favors active genes and local hotspots. Cell. 2002; 110:521–529. [PubMed: 12202041]

45. Svensson AC, Raudsepp T, Larsson C, Di Cristofano A, Chowdhary B, La Mantia G, Rask L, Andersson G. Chromosomal distribution, localization and expression of the human endogenous retrovirus ERV9. Cytogenetics and cell genetics. 2001; 92:89–96. [PubMed: 11306803]

46. Shamay M, Agami R, Shaul Y. HBV integrants of hepatocellular carcinoma cell lines contain an active enhancer. Oncogene. 2001; 20:6811–6819. [PubMed: 11687960]

Research highlights

We have demonstrated heterozygous deletion of 5.5kb sequence on chromosome 11q13 in a specific population, here in African American (AA) population.

This polymorphism is a founder effect from the Yoruban population, an ethnic group of southwestern Nigeria.

We have also shown single nucleotide polymorphisms (SNPs) closely linked to the heterozygous deletion.

The deletion sequences or the SNPs could be used for association studies in diseases prevalent in the AA population.

We also demonstrate that the region surrounding the heterozygous deletion sequence is fragile, amenable to breakage from carcinogens.

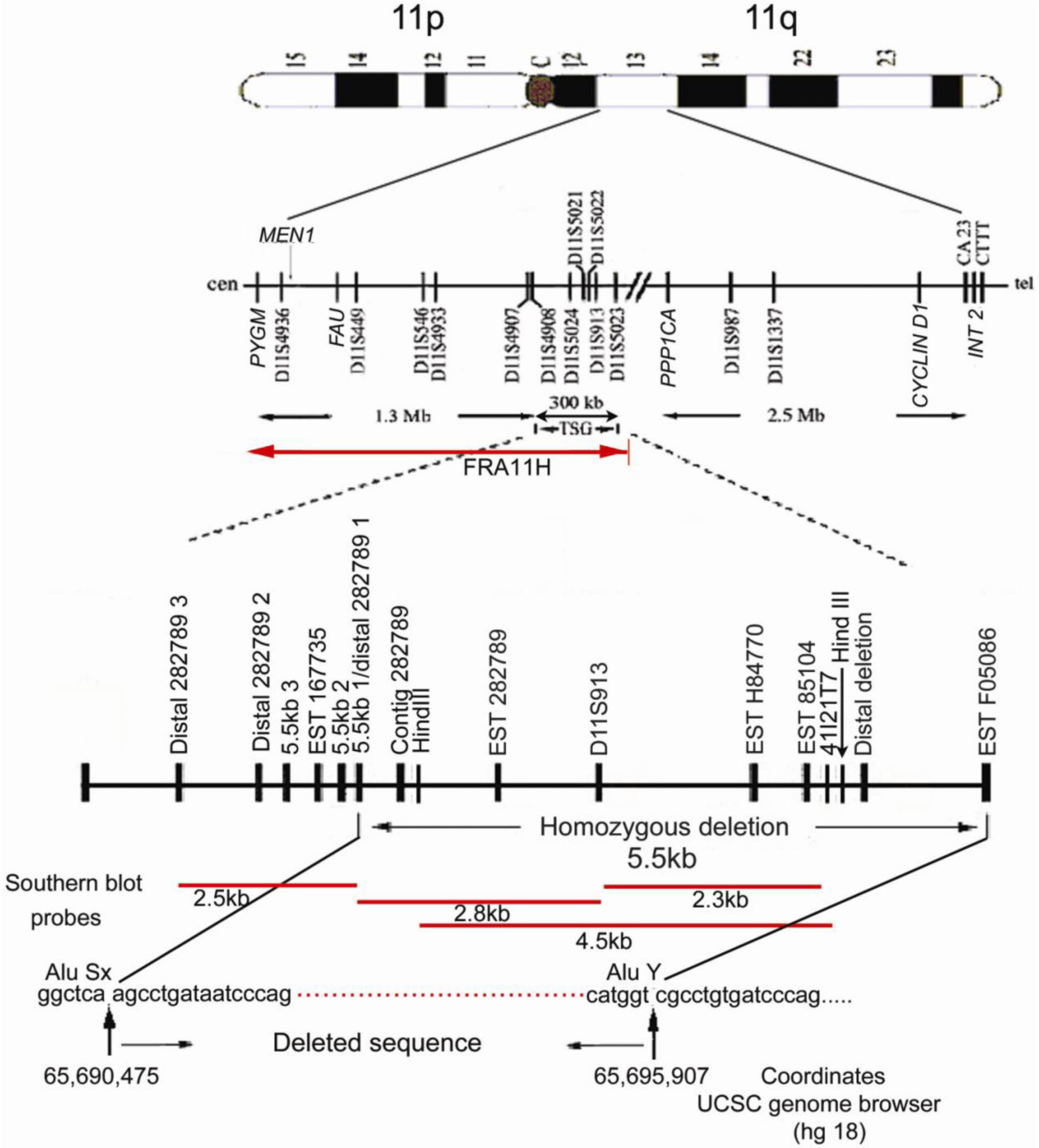


Figure 1.

HeLa cell deletion breakpoints at proximal chromosome 11q13.1. We have previously mapped one of the HeLa cell allelic deletions to a 2.3Mb and a second allelic deletion between markers 5.5kb1 and EST F05086, an interval of 5.5kb. These two deletions overlap yielding a homozygous deletion of 5.5kb sequence in the HeLa cells. The *Alu*Sx and *Alu*Y sequences at the 5.5kb breakpoints are represented in bold letters. Note the homologous repeat sequence at the centromeric and telomeric ends of the deletion. The vertical upward arrows represent the breakpoint sites. The red dots represent the intervening sequences. The coordinates of deletion in the UCSC browser (hg18) are also shown. Known ESTs (expressed sequence tags) are also indicated. Probes used for Southern blot hybridizations are shown in red horizontal lines. FRA11H is mapped between proximal to *MEN1* and 100kb distal to *PACS1*, a distance of 2.6Mb. Map is not drawn to scale.

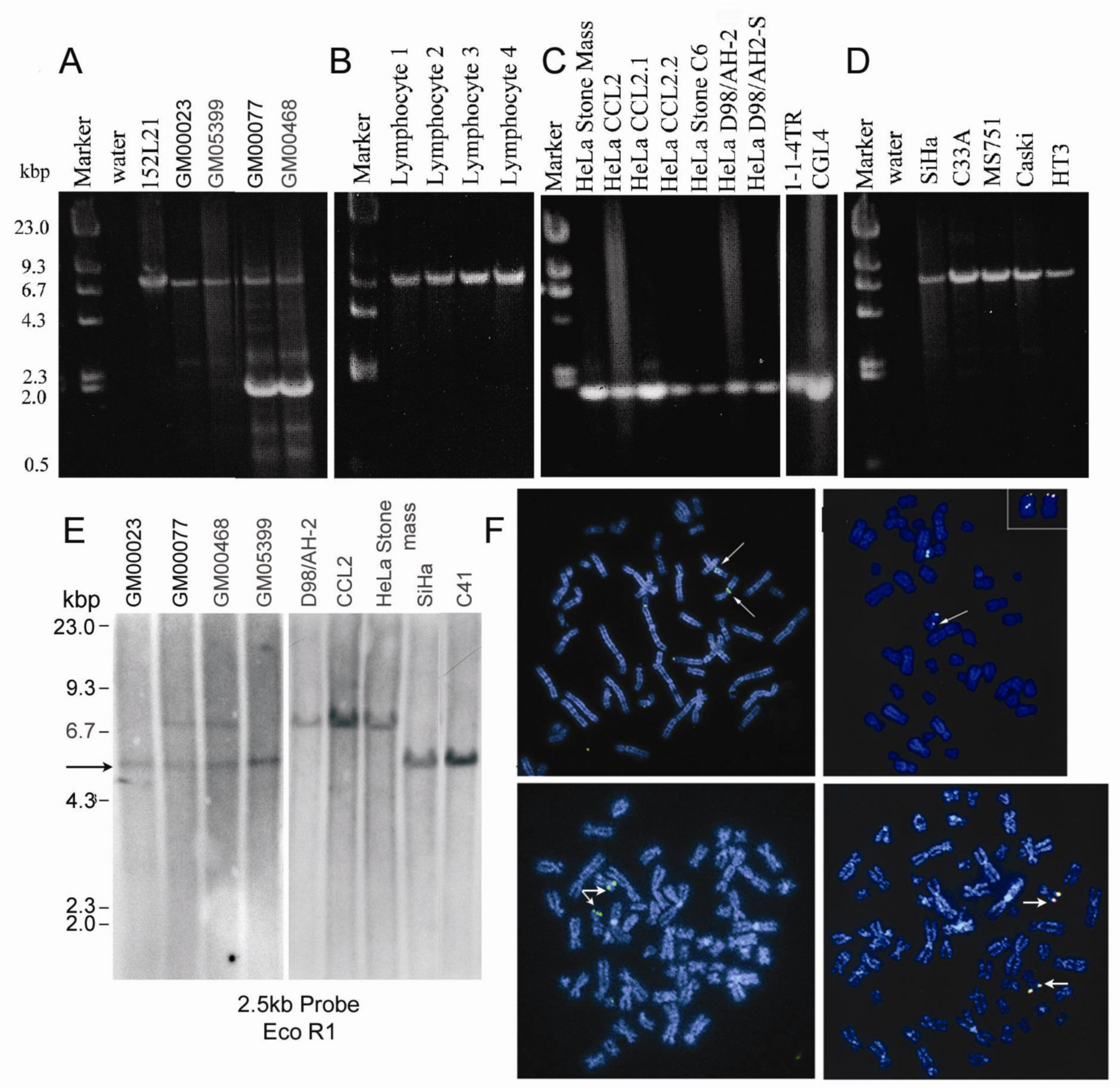


Figure 2.

Long-range PCR confirming 5.5kb deletion in the fibroblast and HeLa cell lines. PCR performed with primers localized outside of the 5.5kb deletion shows A) the presence of the expected 7.3kb product in the BAC 152L21 and in normal fibroblast cell lines GM00023 and GM05399. Two other fibroblast cell lines GM00077 and GM00468 contain a 1.8kb product in addition to the 7.3kb product. The high intensity of the 1.8kb product in GM00077 and GM00468 represents a preferential amplification of the shorter PCR product. The results therefore indicate heterozygous deletion of 5.5kb sequences from these two cell lines. B) PCR product of 7.3kb is also seen in four normal lymphocytes, C) Seven different isolates of HeLa cells and tumorigenic hybrids 1-1-4TR and CGL4 contain only the 1.8kb product confirming homozygous loss of the 5.5kb sequences in these cell lines, D) Five other cervical cancer cell lines show the presence of 7.3kb product indicating absence of deletion in these cells. E) Southern blot hybridization to a 2.5kb probe representing sequences outside and inside of deletion shows the presence of the expected 5.5kb fragment in the two normal fibroblast (GM00023, GM05399) and in two cervical cancer (SiHa, C41) cell lines (indicated by the arrow). The HeLa cell DNAs hybridize to a rearranged 7kb fragment representing the deletion. Two other fibroblast cell lines (GM00077, GM00468) show hybridization to both the expected 5.5kb fragment and the aberrant 7kb fragment. These two fibroblast cell lines therefore show heterozygous deletion for the sequences that are homozygously deleted in HeLa cells. F) FISH analysis using the 5.5kb probe sequence shows hybridization to the 11q13.1 locus in a normal lymphocyte as indicated by the arrows (top left panel). Fibroblast cell line GM00077 contains two copies of the control probe *HRAS*, (11p15) and only one copy of the 5.5kb probe (top right panel). The inset again shows the hybridization of a single copy of the 11q13.1 probe. The results therefore suggest heterozygous deletion of the 5.5kb sequences in this cell line. The HeLa cells (bottom panels). show hybridization to the control 11p15- *HRAS* and 11q23-*MLL* probes (bold arrows), left and right panels respectively, but no hybridization to the 5.5kb probe, indicating homozygous deletion of the 5.5kb probe sequences (bottom panels).

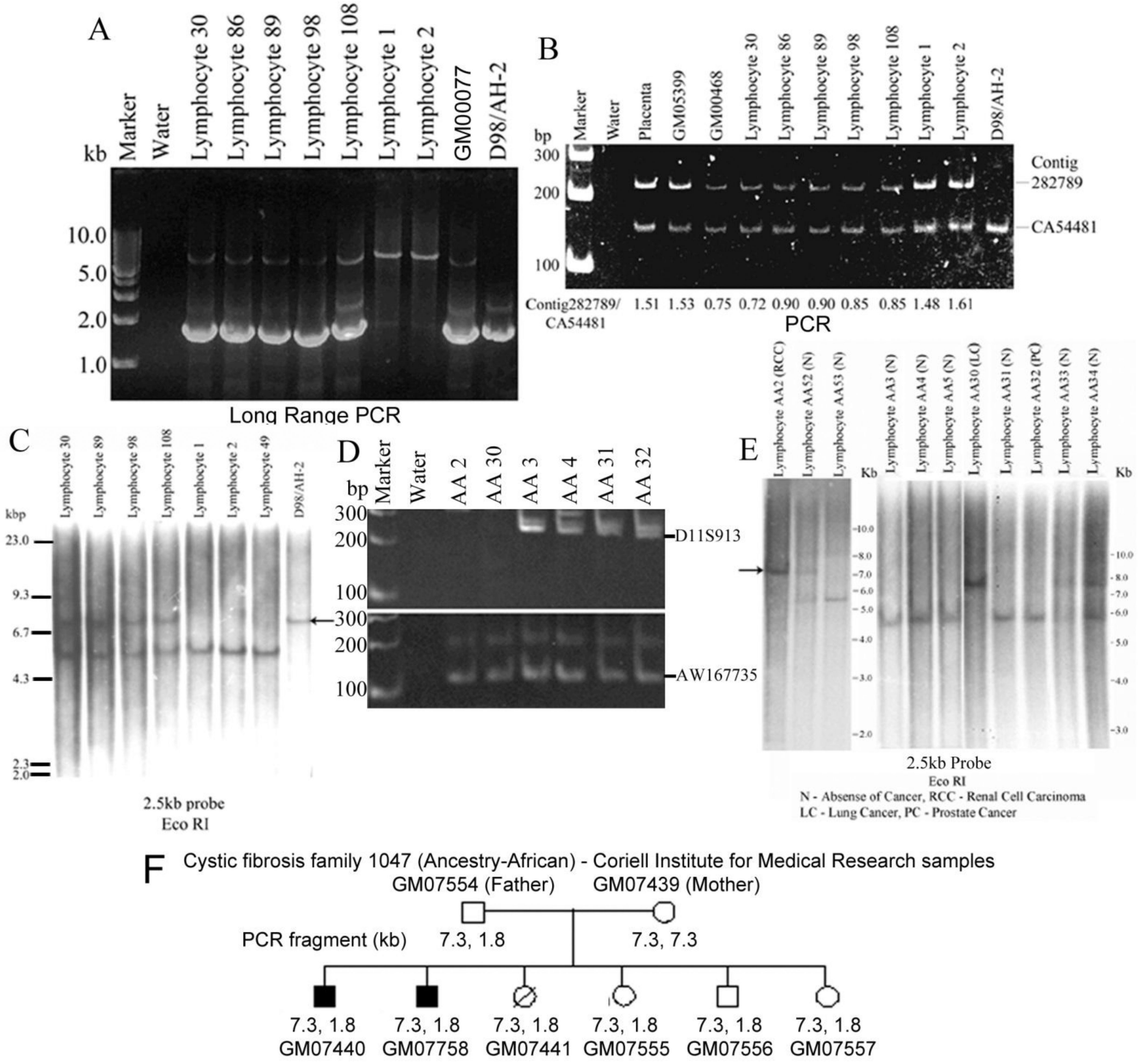
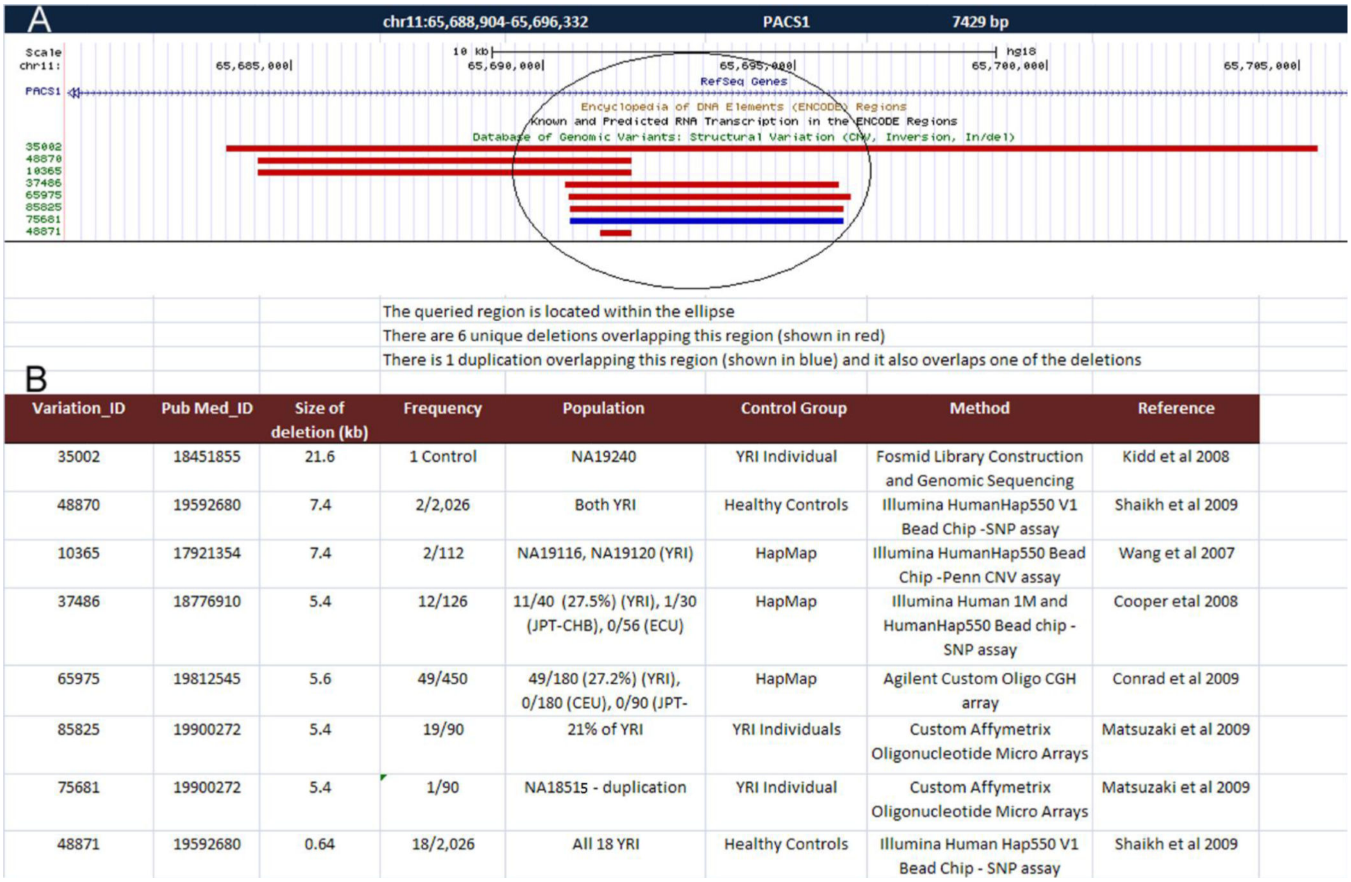


Figure 3.

Heterozygous deletion of 5.5kb sequences in constitutional lymphocyte samples. A) Long range PCR carried out on lymphocyte DNAs shows the presence of the expected 7.3kb product in samples 1 and 2. Samples 30, 86, 89, 98 and 108 show the presence of 7.3kb and 1.8kb products indicating the loss of 5.5kb sequence from one of the copies of chromosome 11. Here, the fibroblast cell line GM00077 containing the heterozygous deletion (see Figure 2 above) and HeLa (D98/AH-2) cells containing the 5.5kb homozygous deletion (presence of only the 1.8kb PCR product) are used as controls. The intensity of the 1.8kb product is higher due to preferential amplification of the smaller product. B) A Semi-quantitative PCR shows the intensity ratio of contig 282789/CA54481 in deleted samples to be half of that in non-deleted samples confirming heterozygous deletion. C) Southern blot hybridization shows the expected 7kb and 5.5kb fragments in four of the samples containing heterozygous deletion. Two of the lymphocyte samples without deletion contain the expected 5.5kb fragment. HeLa cells show the presence of the 7kb fragment (indicated by the arrow) confirming homozygous deletion of the 5.5kb sequences. D) PCR analysis on the test case African-American (AA) samples shows homozygous deletion of the 5.5kb sequence in two of the AA samples, AA2 and AA30. E) Southern blot hybridization confirms the presence of homozygous deletion (presence of only the 7kb fragment –indicated by the arrow) in these two samples. Three other samples (AA33, AA34, AA52) contain heterozygous deletion. F) Long range PCR shows heritability of the 5.5kb deletion (presence of 1.8kb fragment) in a 2-generation African ancestral cystic fibrosis family. The proband (crossed circle) and the two affected brothers (dark squares) are also shown.


A
chr11:65,688,904-65,696,332
PACS1
7429 bp
Scale
chr11:
10 kb
hg18
65,685,000
65,690,000
65,695,000
65,700,000
65,705,000
RefSeq Genes
PACS1
Encyclopedia of DNA Elements (ENCODE) Regions
Known and Predicted RNA Transcription in the ENCODE Regions
Database of Genomic Variants: Structural Variation (CNV, Inversion, In/del)
35002
48870
10365
37486
65975
85825
75681
48871

The queried region is located within the ellipse

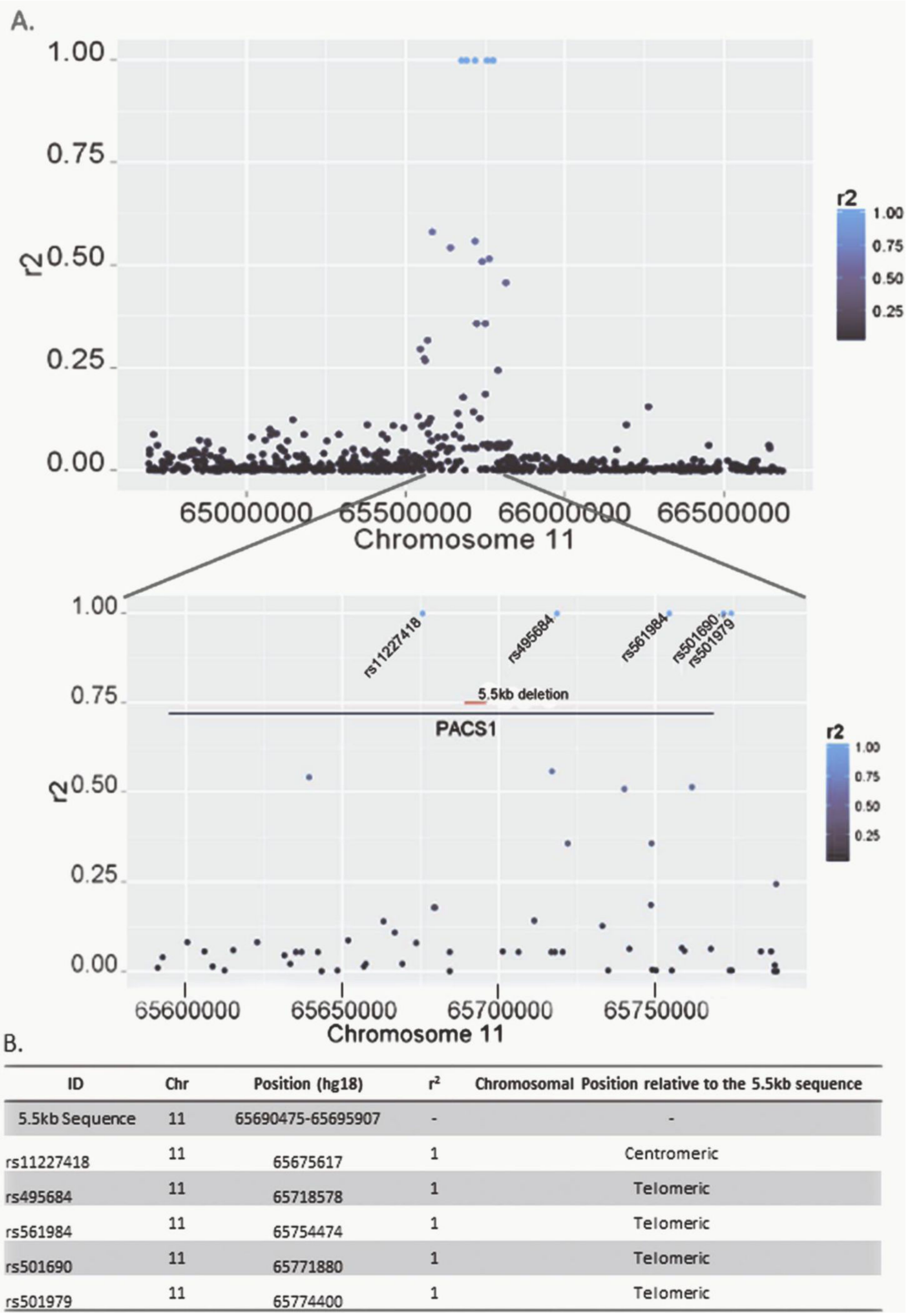
There are 6 unique deletions overlapping this region (shown in red)

There is 1 duplication overlapping this region (shown in blue) and it also overlaps one of the deletions

Variation_ID	Pub Med_ID	Size of deletion (kb)	Frequency	Population	Control Group	Method	Reference
35002	18451855	21.6	1 Control	NA19240	YRI Individual	Fosmid Library Construction and Genomic Sequencing	Kidd et al 2008
48870	19592680	7.4	2/2,026	Both YRI	Healthy Controls	Illumina HumanHap550 V1 Bead Chip -SNP assay	Shaikh et al 2009
10365	17921354	7.4	2/112	NA19116, NA19120 (YRI)	HapMap	Illumina HumanHap550 Bead Chip -Penn CNV assay	Wang et al 2007
37486	18776910	5.4	12/126	11/40 (27.5%) (YRI), 1/30 (JPT-CHB), 0/56 (ECU)	HapMap	Illumina Human 1M and HumanHap550 Bead chip - SNP assay	Cooper etal 2008
65975	19812545	5.6	49/450	49/180 (27.2%) (YRI), 0/180 (CEU), 0/90 (JPT-	HapMap	Agilent Custom Oligo CGH array	Conrad et al 2009
85825	19900272	5.4	19/90	21% of YRI	YRI Individuals	Custom Affymetrix Oligonucleotide Micro Arrays	Matsuzaki et al 2009
75681	19900272	5.4	1/90	NA18515 - duplication	YRI Individual	Custom Affymetrix Oligonucleotide Micro Arrays	Matsuzaki et al 2009
48871	19592680	0.64	18/2,026	All 18 YRI	Healthy Controls	Illumina Human Hap550 V1 Bead Chip - SNP assay	Shaikh et al 2009

Figure 4.

Reported copy number variations overlapping the 5.5kb region are found to exist primarily in the Yoruba (YRI) population. A) The UCSC Genome Browser (hg 18) showing eight copy number variations from the Database of Genomic Variants that overlap with the queried region (deletions = red, duplications = blue). Each CNV has a unique identifier (in green). B) Information pertaining to each variation ID shows that the copy number variations in this region are found primarily in the Yoruba population.



ID	Chr	Position (hg18)	r^2	Chromosomal Position relative to the 5.5kb sequence
5.5kb Sequence	11	65690475-65695907	-	-
rs11227418	11	65675617	1	Centromeric
rs495684	11	65718578	1	Telomeric
rs561984	11	65754474	1	Telomeric
rs501690	11	65771880	1	Telomeric
rs501979	11	65774400	1	Telomeric

Figure 5.

A.) Plot of r^2 values for SNPs on Chromosome 11 in the YRI HapMap3 population. Five SNPs have r^2 values = 1 conveying strong LD with the 5.5kb deletion. A zoomed in view of the plot is also present to show the positions of the SNPs relative to the deletion and *PACS1*. B) A table listing the five SNPs with strong LD and their positions on chromosome 11.

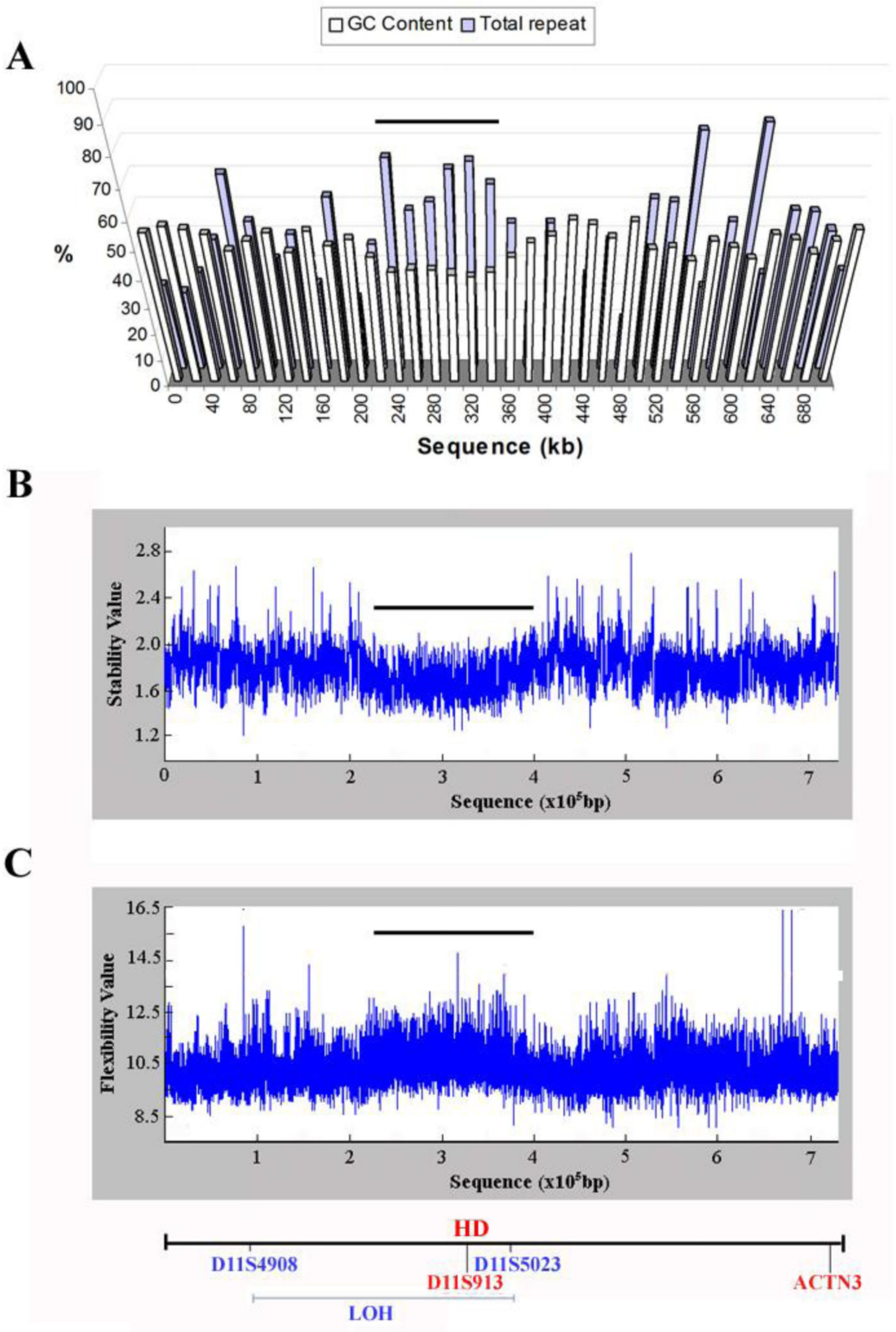


Figure 6.

Repeat content, thermodynamic stability and flexibility of the tumor suppressor locus. A) A 700kb region surrounding the HeLa cell homozygous deletion was divided into 20kb segments and analyzed for repeat and GC content by REPEATMASKER. A rise in overall repeat content, especially in LINE1 elements, and a drop in GC content are observed for a 120kb region surrounding D11S913 (indicated by black horizontal bar). B) An overlapping 175kb region shows absence of stability peaks, indicating that these sequences are thermodynamically unstable than the surrounding sequences (p<0.001; t-test). C)

TWISTFLEX analysis shows that the 175kb sequence contains a significant increase in overall DNA flexibility as compared to the surrounding sequences (p<0.001; t-test). These characteristics are seen in other common fragile sites of the human genome.

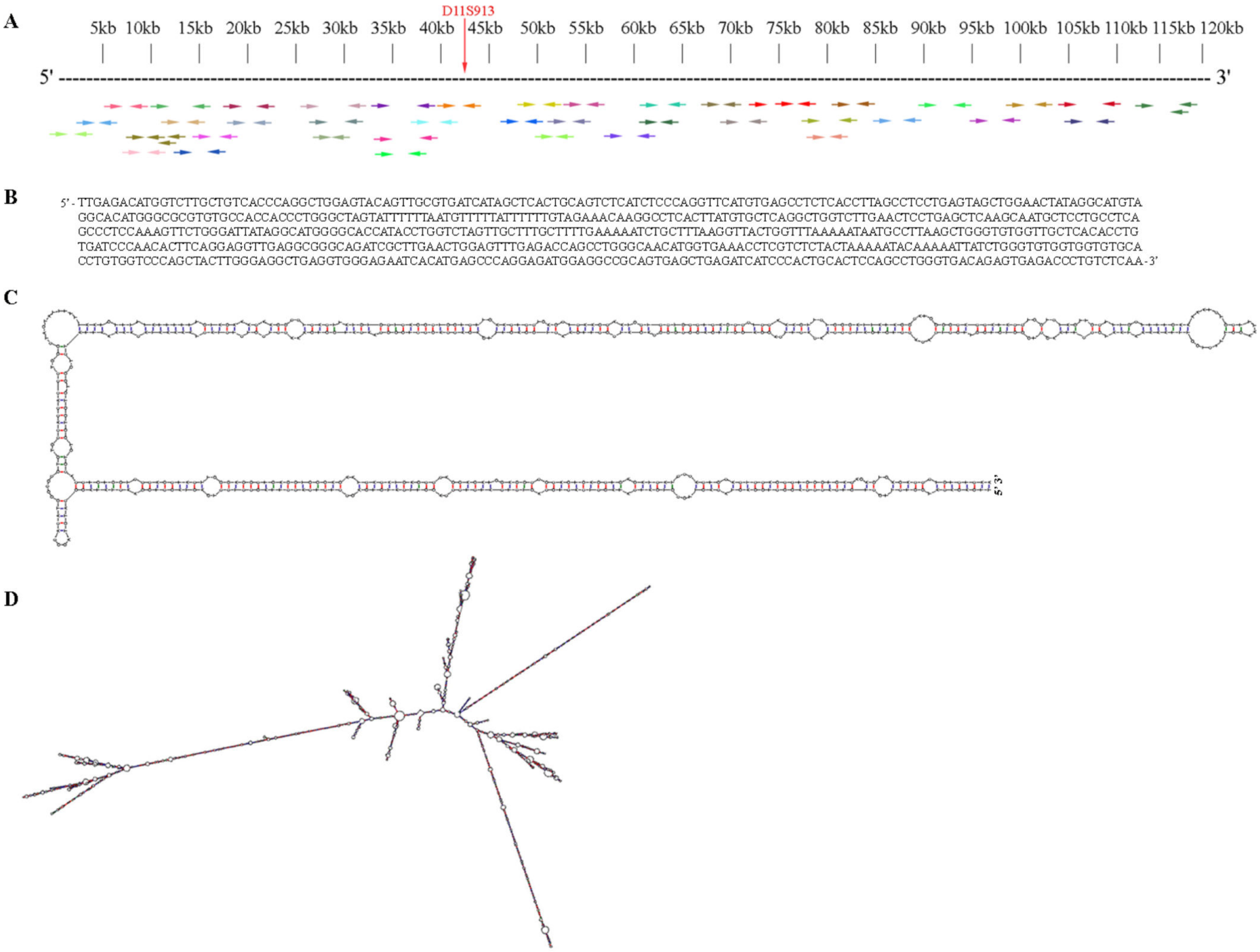

A
D11S913
5kb
10kb
15kb
20kb
25kb
30kb
35kb
40kb
45kb
50kb
55kb
60kb
65kb
70kb
75kb
80kb
85kb
90kb
95kb
100kb
105kb
110kb
115kb
120kb
5'
3'
B
5' - TTGAGACATGGTCTTGCTGTCACCCAGGCTGGAGTACAGTTGCGTGATCATAGCTCACTGCAGTCTCATCTCCCAGGTTCATGTGAGCCTCTCACCTTAGCCTCCTGAGTAGCTGGAACTATAGGCATGTA
GGCACATGGGCGCGTGTGCCACCACCCTGGGCTAGTATTTTTTAATGTTTTTATTTTTTGTAGAAACAAGGCCTCACTTATGTGCTCAGGCTGGTCTTGAACTCCTGAGCTCAAGCAATGCTCCTGCCTCA
GCCCTCCAAAGTTCTGGGATTATAGGCATGGGGCACCATACCTGGTCTAGTTGCTTTGCTTTTGAAAAATCTGCTTTAAGGTTACTGGTTTAAAAATAATGCCTTAAGCTGGGTGTGGTTGCTCACACCTG
TGATCCCAACACTTCAGGAGGTTGAGGCGGGCAGATCGCTTGAACTGGAGTTTGAGACCAGCCTGGGCAACATGGTGAAACCTCGTCTCTACTAAAAATACAAAAATTATCTGGGTGTGGTGGTGTGCA
CCTGTGGTCCCAGCTACTTGGGAGGCTGAGGTGGGAGAATCACATGAGCCCAGGAGATGGAGGCCGCAGTGAGCTGAGATCATCCCACTGCACTCCAGCCTGGGTGACAGAGTGAGACCCTGTCTCAA - 3'
C
5' 3'
D

Figure 7.

Secondary structures for the 120kb unstable, flexible, repeat rich sequences as predicted by MFold. A) Map of the 120kb interval detailing the presence of each of the 44 secondary structures. The colored arrows represent the location of the inverted repeats contributing to secondary structures. One inverted repeat can interact with multiple other repeats to form several possible secondary structures. B) A representative sequence of one of the secondary structures contains two inverted *Alu* repeats located at the 5' and 3' end of the sequence. C) The predicted secondary structure for the above sequence. Alu repeats provide the inverted sequences that help stabilize this stem-loop structure. Such structures are shown to be responsible for fragile site induction (7). D) A large secondary structure for the region 63kb – 69kb which contains numerous over-lapping and nested inverted repeats. The sequences of this structure contain a series of SINE and LINE inverted repeats, demonstrating how multiple inverted repeats can stabilize structures that span several kilobases.

Table 1

Deletion frequency of 5.5kb sequence in the general population

Ancestry analyzed	Samples	Deletions	Deletion frequency (%)
Total	494	47	9.50
Caucasian	126	6	4.76
African American	92	26	28.30
JPT/CHB	17	0	0
Mixed/not-provided	259	15	5.79

p-value <0.0001, calculated for AA population with respect to Caucasian population. The data above indicates that 1) heterozygous deletion is a common event in the general population and 2) higher frequency of deletion is observed in people of African American ancestry compared to those of Caucasian ancestry (28.3% vs 4.76%).