www.oncotarget.com Oncotarget, 2018, Vol. 9, (No. 20), pp: 15275-15291

Research Paper

Analysis of the 9p21.3 sequence associated with coronary artery disease reveals a tendency for duplication in a CAD patient

Natalay Kouprina[1], Mikhail Liskovykh[1,*], Nicholas C.O. Lee[1,*], Vladimir N. Noskov[1], Joshua J. Waterfall[2], Robert L. Walker[2], Paul S. Meltzer[2], Eric J. Topol[3] and Vladimir Larionov[1]

[1]Developmental Therapeutics Branch, National Cancer Institute, Bethesda, MD 20892, USA

[2]Genetics Branch, National Cancer Institute, Bethesda, MD 20892, USA

[3]The Scripps Translational Science Institute, The Scripps Research Institute and Scripps Health, La Jolla, CA 92037, USA

[*]These authors have contributed equally to this work

Correspondence to: *Natalay Kouprina,* ***email:*** *kouprinn@mail.nih.gov*

Keywords: *TAR-cloning; segmental duplication; genome alterations; 9p21; CAD interval*

Received: *November 29, 2017* ***Accepted:*** *February 10, 2018* ***Epub:*** *February 26, 2018* ***Published****: March 16, 2018*

ABSTRACT

Tandem segmental duplications (SDs) greater than 10 kb are widespread in complex genomes. They provide material for gene divergence and evolutionary adaptation, while formation of specific *de novo* SDs is a hallmark of cancer and some human diseases. Most SDs map to distinct genomic regions termed 'duplication blocks'. SDs organization within these blocks is often poorly characterized as they are mosaics of ancestral duplicons juxtaposed with younger duplicons arising from more recent duplication events. Structural and functional analysis of SDs is further hampered as long repetitive DNA structures are underrepresented in existing BAC and YAC libraries. We applied Transformation-Associated Recombination (TAR) cloning, a versatile technique for large DNA manipulation, to selectively isolate the coronary artery disease (CAD) interval sequence within the 9p21.3 chromosome locus from a patient with coronary artery disease and normal individuals. Four tandem head-to-tail duplicons, each ~50 kb long, were recovered in the patient but not in normal individuals. Sequence analysis revealed that the repeats varied by 10-15 SNPs between each other and by 82 SNPs between the human genome sequence (version hg19). SNPs polymorphism within the junctions between repeats allowed two junction types to be distinguished, Type 1 and Type 2, which were found at a 2:1 ratio. The junction sequences contained an Alu element, a sequence previously shown to play a role in duplication. Knowledge of structural variation in the CAD interval from more patients could help link this locus to cardiovascular diseases susceptibility, and maybe relevant to other cases of regional amplification, including cancer.

INTRODUCTION

A growing number of genome-wide associations studies (GWAS) have identified specific regions of the human genome with a strong non-random correlation to complex human traits such as predispositions to diseases [1]. One of such regions has been identified on the INK4b-ARF-INK4a gene cluster located on the human chromosome 9p21.3 (Figure 1A). This region is tightly related with the increase of coronary artery disease (CAD), myocardial infarction [1, 2], ischemic stroke [2, 3] and aortic aneurysm [4]. Recent GWAS has linked single nucleotide polymorphisms (SNPs) at 9p21.3 to CAD and other related similar conditions [2, 3, 5-15]. These associations have been confirmed in multiple independent studies [2, 3, 5-15]. While causal variants

within 9p21.3 have yet to be identified, the risk-associated SNPs cluster together within a ~60 kb region, roughly 100 kb centromeric to the INK4/ARF locus [16]. This 60 kb sequence, named as the CAD interval [17, 18], does not contain protein coding genes.

In addition to CAD, there is a strong correlation of polymorphism at 9p21.3 with predispositions to other diseases, including type II diabetes [19, 20], glioma [21-24], esophageal squamous cell carcinoma (ESCC) [25] and glaucoma [26, 27]. In some cases, such correlation may be linked to mutations in three tumor suppressor genes within the INK4/ARF locus, ARF, INK4b and INK4a. These genes play a central role in cell-cycle arrest, thus affecting key cellular processes such as senescence, apoptosis, and stem cells self-renewal [28, 29]. This locus also contains a fourth gene, MTAP, which has annotated exons overlapping the INK4/ARF locus [30]. MTAP catalyzes the phosphorylation of 5' methyladenosine in the polyamine pathway, and it has also been associated with carcinogenesis [31].

However, GWAS analysis more frequently correlates predisposition to disorders with SNPs mapped to intergenic or non-coding regions rather than to sequences corresponding to annotated genes [32]**.** At present, it is presumed that some diseases risk caused by SNPs at 9p21.3 act through the long non-coding RNA CDKN2B-AS1, commonly referred to as the Antisense Non-coding RNA in the INK4 locus (ANRIL). ANRIL was first identified within a 403 kb germ-line deletion in a family with a history of melanoma and neural system tumors [33]. ANRIL is transcribed as a 3.8-kb-long non-coding RNA from the short arm of human chromosome 9 at the p21.3 locus that overlaps a critical region encompassing three major tumor suppressor loci juxtaposed to the INK4b-ARF-INK4a gene cluster and the MTAP gene [34]. It is transcribed in the direction opposite to the INK4b-ARF-INK4a gene cluster and shares a bidirectional promoter with ARF. The 3' end of ANRIL is terminated at the very end of the CAD interval. Based upon EST assembly, ANRIL has 19 exons with no identified open reading frame. Although cloning a full-length version of the predicted transcript has proven to be difficult, a growing number of alternatively spliced ANRIL transcripts, including circular forms, have recently been reported in the literature [16, 35]. Prior work has shown that long, non-coding RNAs such as Xist, Kcnq1ot1 and HOTAIR can repress genes in cis- or trans- through interaction with Polycomb group (PcG) complexes [36-40]**.** It has also been postulated that ANRIL could play a similar role in PcG-mediated repression of the INK4b-ARF-INK4a gene cluster [41].

In this work, we focused on the analysis of structural variations within the CAD interval sequence using different approaches, including re-isolation of this region by the transformation-associated recombination (TAR) cloning technique [42, 43, 44, 45]. TAR cloning represents a unique tool for selective isolation and manipulation of large DNA molecules. The technique exploits a high level of homologous recombination in the yeast *Sacharomyces cerevisiae*. So far, TAR cloning is the only method available to selectively recover chromosomal segments or genes up to 350 kb in length from simple and complex genomes, including human [42, 43, 44, 45]. In the post-genomic era, TAR cloning has found many applications for functional and structural genomics. For example, it can be used to isolate rearranged chromosomal regions, such as translocations and inversions, from patients and model organisms. TAR cloning allows the assembly and cloning of entire microbe genomes up to several Mb as well as engineering of large metabolic pathways [45].

Here, we used TAR cloning for isolation of the CAD region from a somatic peripheral blood mononuclear cells line derived from a coronary artery disease patient with unique mutations within the CAD interval. As a control, the CAD region was TAR-cloned from blood cells of normal individuals. Our analysis of TAR isolates revealed a previously unknown amplified 50 kb sequence corresponding to the CAD interval in the patient but not in normal individuals. This observation may be important for GWAS studies aiming to link SNPs near the INK4b-ARF-INK4a gene cluster to susceptibility to cardiovascular diseases.

RESULTS

TAR-isolation of the CAD interval sequence from the cell line derived from a patient with the coronary artery disease

Figure 1A illustrates the organization of the INK4b-ARF-INK4a gene cluster at 9p21.3 with the positions of the ANRIL ncRNA and the CAD interval carrying high-risk SNPs. The region chosen for analysis includes the CAD interval with flanking sequences. Figure 1B illustrates a diagram of TAR cloning of the CAD-containing region from the patient derived C087 cell line. The TAR vectors 9p21-1 and 9p21-2 (Supplementary Figure 1) were constructed and used for TAR cloning experiments. The vectors contained a yeast selectable marker HIS3, a yeast centromere from *Saccharomyces cerevisiae* chromosome 6 (CEN6) and two unique targeting sequences (hooks) homologous to the 5' and 3' ends of the targeted genomic region. Vector 9p21-1 was constructed by insertion of two DNA targeting hooks, 171-bp and 209-bp long, corresponding respectively to positions 22,062,540 to 22,062,711 and 22,129,014 to 22,129,313 on human chromosome 9 (human genome sequence; version hg19). The expected size of the captured genomic fragment is 66,763 bp. As for vector 9p21-2, the hooks used were 185-bp and 242-bp long, corresponding to positions 22,061,474 to 22,061,658 and 22,130,089 to 22,130,330 on the chromosome 9 (version hg19) The expected size of the 9p21-2 targeted genomic fragment is 68,856 bp.

After transformation of a TAR vector and genomic DNA into yeast *S. cerevisiae* cells, recombination between the hooks in the vector and the targeted genomic sequences leads to the rescue of the target region as

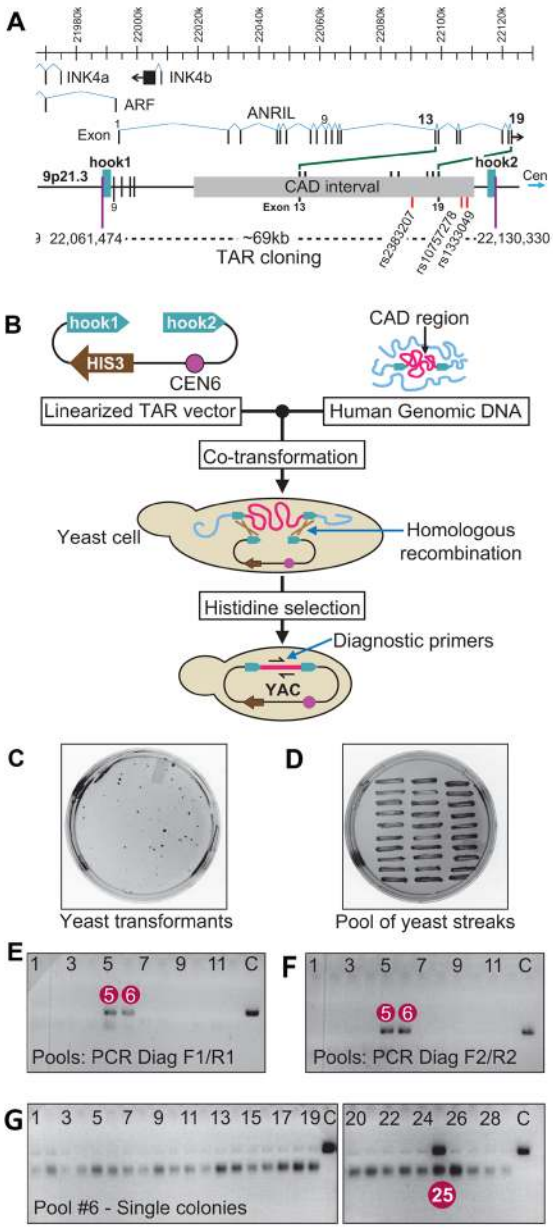

Figure 1: TAR cloning of the CAD interval at 9p21.3 from a patient with the coronary artery disease. (A) A scheme of organization of the 9p21.3 region. Positions of three known genes, ARF, INK4a and INK4b, as well as a long noncoding RNA, ANRIL, are shown. ANRIL transcript covering 126 kb includes the CAD interval. ANRIL consists of 19 exons. Exons 13-19 are within the CAD interval. Positions of three SNPs, rs10757278, rs1333049 and rs2383207, specific to the mutated allele of the CO87 patient are shown. Positions of the targeted sequences (hooks) chosen for TAR cloning are indicated (human genome sequence; version hg19). **(B)** The diagram showing a general scheme of TAR cloning of a region of interest (in red) from total genomic DNA (in blue) with a linearized TAR vector containing a yeast selectable marker HIS3, centromeric sequence from chromosome 6 (CEN6) and two unique targeting sequences (hook1 and hook2) homologous to 5' and 3' ends of the targeted region. After co-transformation into yeast *Saccharomyces cerevisiae*, recombination between targeting sequences in the vector and the targeted sequences of the genomic DNA fragment leads to the rescue of the fragment as a circular YAC (yeast artificial chromosome) molecule. For TAR cloning experiments, the vector DNA is linearized by a unique endonuclease located between the hooks to expose targeting sequences. **(C)** A representative His^- plate with yeast His^+ transformants and **(D)** the plate with 30 streaked randomly chosen His^+ transformants. **(E)**, **(F)** Analysis of 11 pools of yeast transformants by PCR for the presence of the CAD region using two pairs of diagnostic primers, F1/R1 and F2/R2). Pools #5 and #6 are positive for diagnostic PCR. **(G)** Analysis of 30 individual His^+ transformants from pool # 6. Clone #25 is positive for the CAD interval. C-control PCR with human genomic DNA.

a circular TAR/YAC (Yeast Artificial Chromosome) molecule (Supplementary Figure 2A) [42, 43]. To identify the desired region-containing clones, yeast transformants (Figure 1C) were combined into pools (Figure 1D) and then examined with PCR reactions using diagnostic primers, F1/R1 and F2/R2 (Supplementary Table 1). These primers are not present in the TAR vector but are specific to the targeted genomic region. As an example, Figure 1E and 1F illustrates the PCR screening of 11 pools of transformants obtained by the 9p21-2 vector. Pools #5 and #6 are positive for diagnostic primers. Figure 1G illustrates the PCR analysis of 29 transformants of pool #6. Individual transformant #25 is positive for diagnostic primers. Using both vectors, eight CAD-region-positive TAR/YAC clones were identified and used for further analysis.

Physical analysis of TAR clones isolated from C087 cells suggests the presence of segmental duplication

TAR isolates containing the CAD region were characterized in a Yeast Artificial Chromosome (YAC) and a Bacteria Artificial Chromosome (BAC) form. The retrofitting vector BRV1 was used to convert YAC clones into BACs [44]. Details are described in MATERIALS AND METHODS and in Supplementary Figure 2B. Several approaches were taken to characterize the TAR-cloned material. Firstly, to prove the presence of the predicted genomic sequence in YAC isolates, yeast clones were examined by PCR with overlapping pairs of primers (Supplementary Table 1) that cover the entire CAD interval. As seen, this region is present in the TAR clones (Figure 2A). Secondly, to check the size of the cloned material, yeast chromosomal-size DNAs were exposed to a low dose of γ-rays to linearize circular YAC molecules, separated by clamped homogeneous electrical field electrophoresis (CHEF), blotted and hybridized with the yeast CEN6 probe (Supplementary Table 1). Unexpectedly, the size of the YACs was much bigger than predicted (~230 kb versus ~67/69 kb) (Figure 2B). After conversion of YACs into BACs, we analyzed the TAR/BAC clones. To check the size of BACs, they were digested by NotI, separated by CHEF and stained with ethidium bromide. The size of BACs corresponded to the size of YAC isolates (~230 kb) (Figure 2C). To demonstrate the identity of sequences in BAC clones, DNA isolated from two randomly selected BACs was digested by HindIII. The restriction profiles of the BACs were indistinguishable from each other (Figure 2F).

The larger size of TAR isolates may be explained either by the presence of duplications within the CAD interval or by the presence of non-annotated sequences in this region. To clarify this, one of the BAC clones (A218) was used for deep paired-end sequencing. This identified 82 SNPs that differed from the human genome sequence (version hg19) (Supplementary Table 2) but no evidence for large insertions of novel sequence. Surprisingly, analysis of depth of coverage as well as discordant read pair alignments suggested copy number gain of an approximately 50 kb internal region (Figure 2D) arranged in a tandem head-to-tail orientation. To confirm that the CAD-carrying clones contain segmental duplications, DNA from the BAC A218 was digested either by AgeI or PmeI or FspI endonucleases that cuts once within the proposed 50 kb repeat (Figure 2E). After digestions, it was clearly seen that the BAC insert has approximately 50 kb duplicated region (Figure 2G). For example, digestion by AgeI (Figure 2G; lane 1) produces an intense ~50 kb band and two other minor expected bands, i.e. a 16.6 kb band derived from the 5' end of the CAD interval and a smaller band of 12.5 kb derived from the vector part and the region between the hook and the first AgeI site (Figure 2E). This suggests that a 50 kb sequence is repeated four times (50 kb x 4 plus 16.6 kb plus 12.5 kb = 229.1 kb).

To prove the presence of the predicted junction from the deep sequencing (Figure 2D), we designed a specific pair of forward and reverse primers, B586/B578, corresponding to the very beginning and end of the proposed 50 kb repeat (Supplementary Table 1). These primers amplified a 684 bp product from BAC DNA. Sequencing of these PCR products revealed the following structure, i.e. 124 bp (positions 22126447 to 22126570 in hg19) corresponding to the end of the proposed duplication and 259 bp (positions 22076352 to 22076610 in hg19) that corresponds to the beginning of the repeat. Within the PCR products, these sequences are separated by the 301 bp Alu element (Figure 3A and Supplementary Table 3). SNPs polymorphism within the junction sequences allowed us to distinguish two types of junctions, Type 1 and Type 2 (Figure 3B and 3C) that were found at an approximate ratio of 2:1. In addition, we designed other pairs of primers further upstream and downstream of B586/B578 primers (see "Extended PCR junction" in Supplementary Table 1). These primers also gave the junction products (Supplementary Figure 3). Their sequencing confirmed the proposed junction between repeats.

To summarize, our results suggest that the C087 TAR isolates contain a segmental duplication (SD) of 50,635bp in size (positions 22076220 to 22126855 in hg19).

The structure of the CAD interval derived from a patient with the coronary artery disease

Based on the data described above, we predicted the physical structure of the CAD interval within the TAR isolates (Supplementary Figure 4). To prove this structure, we sub-cloned the restriction digest fragments of BAC A218 (Figure 4A) and sequenced them.

Firstly, if the proposed map is correct, complete digestion of BAC DNA either by PmeI or AatII followed

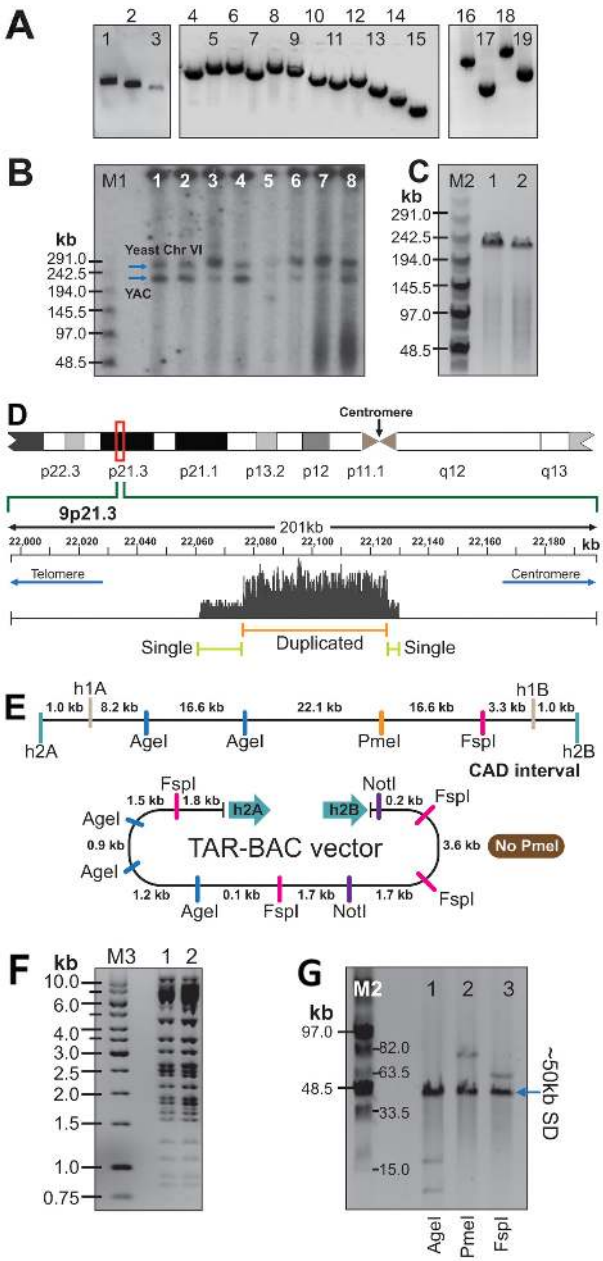

Figure 2: Physical analyses of the CAD-containing TAR/YAC and BAC clones isolated from the C087 patient cell line. **(A)** PCR analysis of a YAC clone by a set of the overlapping primers (Supplementary Table 1) covering the entire CAD region (positions 22,061,474 to 22,130,330 on the chromosome 9 in hg19). Lane 1 corresponds to New41F/New41R; lane 2 – New41F/41R; lane 3 – 41F/New41R; lane 4 - 17F/20R; lane 5 - 20F/27R; lane 6 – 33R/27F; lane 7- 46R/42F; lane 8 - 51R/46F; lane 9 – 51F/59R; lane 10 – 33F/35R; lane 11 – 33F/38R; lane 12 – 38F/40R; lane 13 – 40F/41R; lane 14 – 42F/42R; lane 15 – New41F/New41R; lane 16 – 6500F/13R; lane 17 – 13F/17R; lane 18 -6500F/17R; lane 19 – 38F/41R. **(B)** Size of the TAR/YAC-cloned material. NotI-digested DNA isolated from eight independent TAR/YAC clones was separated by CHEF gel electrophoresis and hybridized with the CEN6 probe. Arrows indicate an yeast centromere from chromosome 6 and linearized YAC molecules. **(C)** Size of the TAR/BAC clones. Two BAC DNAs (lane 1 - 9p21-2 vector; lane 2 – 9p21.1 vector) were digested by NotI, separated by CHEF, and visualized with ethidium bromide. The size of the bands is ~230 kb. **(D)** Snapshot of the region from the Integrative Genomics Viewer (IGV) showing the coverage of reads obtained from sequencing of the C087 BAC A218. As seen, the duplicated region corresponds to positions 22076300 to 22126800 on chromosome 9 (hg19). **(E)** Schemes of the CAD region at 9p21.3 and the TAR vector after retrofitting by BRV1 vector. **(F)** HindIII digestion profiles of two BACs. **(G)** Analysis of the repeated sequence in the BAC clone A218 containing the CAD region. BAC DNA were digested either AgeI (lane 1) or PmeI (lane 2) or FspI (lane 3). M1 - CHEF DNA Size Lambda Ladder (BIO-RAD); M2 – Midrange 1 PFG Marker (NEB). M3 – GeneRuler 1 kb DNA Ladder (Fermentes).

by re-ligation should produce a BAC containing a hybrid ~50 kb unit plus flanking regions containing hooks (Figure 4A). So, BAC A218 was digested, re-ligated and electroporated into *E. coli* cells. The Cm^R colonies were screened by PCR with the primers specific for the yeast HIS3 gene (a positive control), for the sequences around the PmeI site (a positive control) and for the predicted junction sequence that should not present in these BACs (a negative control) (Supplementary Table 1). BAC DNA from appropriate colonies (Figure 4B) was isolated, digested with NotI and run on CHEF. All the BACs have the expected size of ~82 kb (Figure 4D; lanes 1, 2, 3). Two BACs, A226 after PmeI digestion and A227 after AatII digestion, were chosen for deep sequencing.

Secondly, to isolate the internal repeats, BAC A218 was digested by AgeI and run on CHEF. The fragments of ~50 kb in size were isolated from the gel and ligated into the vector V231 (Supplementary Figure 5). The Cm^R colonies were screened by PCR using the primers specific for the junction between proposed SDs (Supplementary Table 1). BACs with Type 1 (A233) and Type 2 (A234) junction were identified (Figure 4C). Two BACs of each junction type were isolated, digested with AsiSI that has a unique site in the V231 vector and run on CHEF. The predicted ~57 kb single bands were observed (Figure 4E).

Thirdly, to confirm the identical structure of the re-cloned fragments, BACs A226, A227, A233 and A234 were digested either by Taqα1 or AseI or HindIII (Figure 4F). The restriction profiles of BACs A233 and A234 were identical to each other (Figure 4F; lanes 3, 4, 5, 6). The restriction profiles of BACs A226 (PmeI digested) and A227 (AatII digested) were also identical to each other (Figure 4F; lanes 1, 2) and shared similarities with the banding profiles of BACs A233 and A234. BAC A233 (Type 1) and BAC A234 (Type 2) were chosen for deep sequencing.

Sequence analysis of the BACs confirmed that the CAD-containing TAR-cloned region consists of almost identical ~50 kb repeat units. Notably that each unit contains three characteristic SNPs, rs10757278, rs1333049

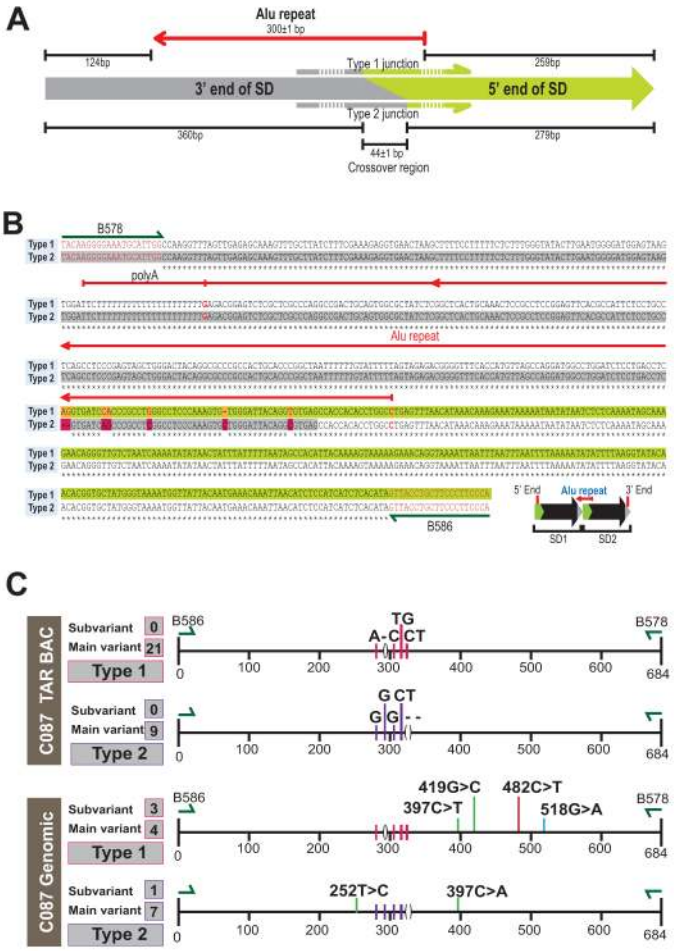

Figure 3: Junction between segmental duplications. (A) Scheme of SD junction depicting the 3' and 5' ends of the SD, the relative position of the Alu repeat and the crossover region found in Type 1 and Type 2 junction sequences. **(B, C)** Sequence alignment of two major types of junction amplified by B578/B586 primers (nucleotides in red). Red arrow is the Alu repeat. Nucleotide polymorphisms unique to each type of junction are depicted by colored boxes.

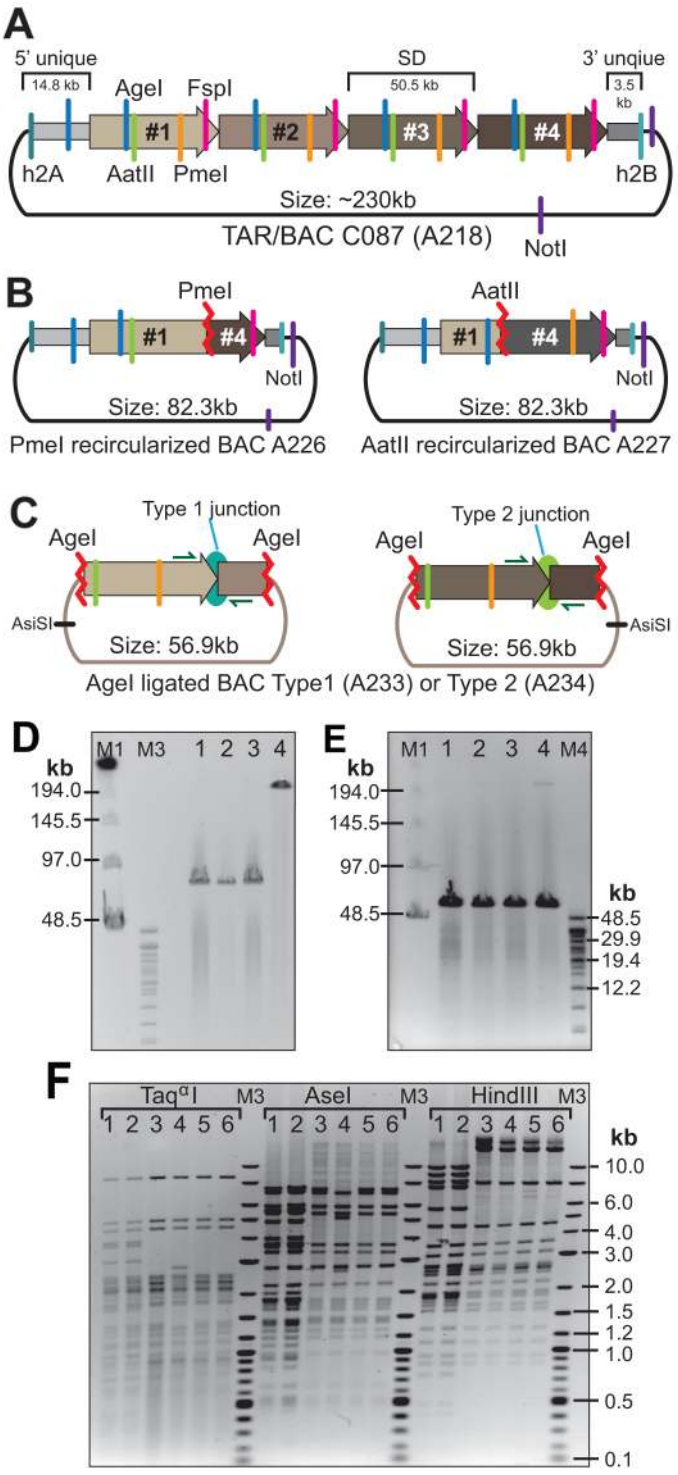

Figure 4: Confirmation of the presence of SDs in the CAD-containing C087 TAR/BAC clone. (A) A predicted structure of BAC A218. **(B)** The structure of A226 and A227 BACs obtained after PmeI or AatII digestion of BAC A218 and followed by re-ligation. **(C)** The structure of A233 and A234 BACs obtained after AgeI digestion of BAC A218 and followed by capture into V231 vector. **(D)** CHEF analysis of one isolate of NotI-digested BAC A226 (lane 1) and two isolates of BAC A227 (lanes 2, 3). Lane 4 corresponds to NotI-digested BAC A218. **(E)** Lanes 1, 2 correspond to two isolates of AsiSI-digested BAC A233. Lanes 3, 4 correspond to two isolates of AsiSI-digested BAC A234. **(F)** Restriction profiles of A226, A227, A233 and A234 BACs after digestion either by Taqα1 or AseI or HindIII. Lane 1 corresponds to BAC A226. Lane 2 corresponds to BAC A227. Lanes 3, 4 correspond to two isolates of BAC A233. Lanes 5, 6 correspond to two isolates of BAC A234.

and rs2383207, that are specific to the risk allele of the CO87 patient (Supplementary Figure 6).

Analysis of TAR isolates from normal human cells did not reveal the presence of duplicated region within the CAD interval

We used the TAR 9p21-1 vector (Supplementary Figure 1) to TAR clone the CAD region from the normal human DNAs (Promega). Three CAD-positive TAR/YAC clones were identified and used for further analysis. The presence of the predicted genomic sequences in YAC isolates was confirmed by PCR with pairs of overlapping primers (Supplementary Table 1 and Supplementary Figure 7). To check the size of the cloned material in yeast cells, three YACs were linearized by NotI, separated by CHEF and blot-hybridized with a specific probe (Supplementary Table 1). The size of the YAC inserts was as expected (Figure 5A). After conversion of YACs into a BAC form (Supplementary Figure 2B), the BAC molecules were moved into bacterial cells and analyzed. To check the size of BACs, they were digested by NotI, separated by CHEF and stained with ethidium bromide. The size of BACs corresponded to the size of YAC isolates (Figure 5B and 5C). To demonstrate the identity of sequences in the BAC clones from a patient and normal individuals, DNA isolated from BACs was digested by HindIII (Figure 5D). As expected, the restriction profiles of the BACs were almost indistinguishable from each other.

To confirm that duplication observed in TAR clones isolated from the C087 cells was not due to amplification of the cloned material in yeast cells, we performed a control experiment. A BAC A226 containing a single copy of the CAD region was re-transformed into yeast cells. After that, genomic DNA from 10 His^+ transformants were isolated, NotI digested, separated by CHEF and blot-hybridized with a specific probe of 862 bp in size (positions on chromosome 9: 22084459 to 2085320 in hg19) (Supplementary Table 1). As seen from Supplementary Figure 8, all transformants have the expected size, which proves an absence of amplification of this region in yeast cells. As expected, the TAR isolates did not contain SNPs, rs10757278, rs1333049 and rs2383207, which are specific to the risk allele of the CAD patient (data not shown).

Population analysis of the CAD interval at 9p21.3 in the human genomes did not reveal repeated sequences

To check whether the copy number of *SD at 9p21.3* is variable in human population, we applied quantitative real-time PCR to analyze 181 DNA samples from normal individuals. Copy number of two unique sequences, 144 bp and 137 bp in size, at the beginning and the end of SD (positions on chromosome 9: 22078776 to 22078919 and 22117414 to 22117550 in hg19) was determined using two pairs of specific primers (Supplementary Table 1). As a control for a single copy region, we used a unique 87 bp sequence outside of SDs (positions on chromosome 9: 22062099 to 22062185 in human genome sequence version hg19) and the *RNAaseP* gene that allowed the reactions to be calibrated and reproducible results to be obtained. Table 1 summarizes the results of analysis. As seen, all samples contain one copy of the CAD interval sequence. In addition, Droplet digital PCR re-confirmed these results (Supplementary Table 4). Thus, more than one copy of SD is not very common in human population (less than 1%).

DISCUSSION

In the past years, an emerging group of human genetic diseases have been described that result from DNA rearrangements rather than from single nucleotide changes. Such conditions have been referred to as genomic disorders. The predominant molecular mechanism underlying some of such rearrangements [(also known as segmental duplications (SDs)] is nonallelic homologous recombination (NAHR) utilizing the repeats as substrates. These higher-order genomic architectural features usually span from ~10 kb up to hundreds of kilobases of genomic DNA where repeats/duplications share >90% sequence identity. Notably, 52% of the remaining gaps in the reference haploid human genome, refractory regions to all techniques available at the moment, are sequences consisting of SDs. In humans, copy-number variations (CNVs) of such repeats have been implicated in common traits such as neuropathy, hypertension, color blindness, infertility, and behavioral traits, including autism and schizophrenia, as well as disease susceptibility to HIV, lupus nephritis, psoriasis and Parkinson's disease [46-48].

In this study, comprehensive analysis of the CAD interval at 9p21.3 locus, where SNPs associated with coronary artery disease (CAD) were mapped [1-4], revealed a previously un-annotated duplication of ~50 kb in size in the human peripheral blood mononuclear cell line derived from a patient with CAD. Based on the sequence analysis of TAR isolates, the CAD region in this cell line contains four 50 kb units that are organized as tandem repeats. An Alu repeat is present in the junction sequence between the units. It is worth noting that the role of Alu in formation of duplications has been proposed for this type of SINE elements [49]. Quantitative real-time PCR and Droplet digital PCR analyses showed no amplification of this region in the limited number of normal human individuals (approximately 200 DNA samples were screened that means that such duplication may represents less than 1% in human population).

This CAD duplication has also not been previously detected in human genome studies. This may be explained by several reasons. Firstly, although computational

methods have been developed to identify duplicated sequences independently on the genome assembly and other experimental methods like FISH and array comparative genomic hybridization (array-CGH) have been used to validate and explore the distribution and organization of such sequences [20], these methods work

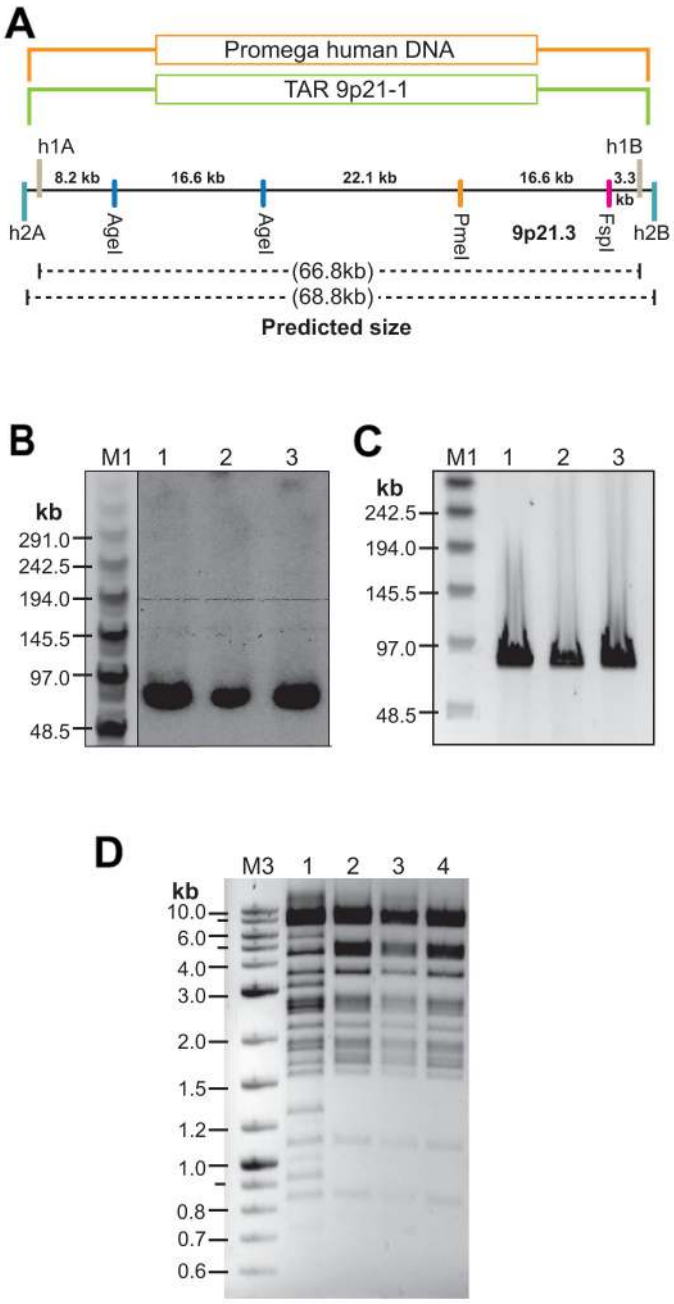

Figure 5: TAR cloning and physical analysis of the CAD-containing TAR/YAC clones isolated from genomic DNA of normal individuals. (A) A scheme of TAR cloning of the CAD region from DNA isolated from Promega genomic DNA by 9p21-1 TAR vector. h1A and h1B are targeting hooks for 9p21-1 TAR vector. h2A and h2B are targeting hooks for 9p21-2 TAR vector. **(B)** Southern analysis of the TAR/YAC isolates. Genomic DNA from three independent YAC clones was digested by NotI, separated by CHEF, and blot-hybridized with a 862 bp specific probe (positions on chromosome 9: 22084459-22085320 in human genome sequence; version hg19). The size of the predicted cloned material is ~66.8 kb. **(C)** CHEF analysis of NotI-digested three BAC isolates (lanes 1, 2 correspond to BAC clones obtained from a YAC clone #1; lane 3 corresponds to a BAC obtained from a YAC clone #2). **(D)** HindIII digestion profiles of BACs from the C087 patient and normal individuals. Lane 1 corresponds to genomic DNA isolated from the A218 BAC clone of the C087 patient containing four 50 kb repeat units; Lanes 2 – 4 correspond to genomic DNA isolated from BAC clones of normal genomic DNA (Promega). M1 - CHEF DNA Size Lambda Ladder (BIO-RAD); M2 - Quick-Load 2-Log DNA Marker (BioLabs).

Table 1: Analysis of the number of SD at the CAD region in normal individuals by qPCR*

Population	Total analyzed number	Samples with 1 copy	Samples with 2 copies	Samples with 4 copies
Korean	52	52	0	0
Italian	46	46	0	0
Caucasian	83	83	0	0
Patient C087	1	0	0	4

*The number is given per haploid genome.

well for relatively short SDs (size between 1 kb and 10 kb). Identification of large SDs by these methods remains a challenge. Secondly, the CAD region does not contain any open reading frames that are easily identified by Next-Generation Sequencing. Thirdly, there is a relatively high density of SINE and LINE repeats within the region. This makes it a problem to assemble the duplicated sequences using the multitude short reads obtained from Next-Generation Genome Sequencing. The identification of SDs within the SPANX-B locus at Xq27 is a good example of such problems. It is worth noting that the presence of SDs within the SPANX-B locus at Xq27 was determined only by direct TAR cloning of the SPANX-B region from human genomes [50, 51]. Fourthly, some duplicated regions are poorly represented in BAC libraries and subsequently become underrepresented if these libraries are used to build genome maps. One potential reason for the poor representation of some duplicated region in BAC libraries is the absence of appropriate restriction site for digestion of genomic DNA before ligation into a BAC vector. Moreover, some DNA fragments carrying large, homologous repeats are sometimes structurally unstable in yeast and bacteria, for example the SPANX-B region [50, 51]. Finally, it is a matter of data analysis. The 150 kb TAR/BAC obtained from the patient in this study was sequenced using Next-Generation Sequencing technology. Initial analysis reported and assembled a single 50 kb sequence corresponding to the known human genome sequence (version hg19). This 50 kb reported length did not match the physical length that we knew the TAR/BAC to be (150 kb). The duplication was only detected within the data when we specifically called for a comparison between the number of sequence reads in the duplicated area versus the BAC backbone.

Nevertheless, some indirect hints for the presence of the CAD duplication came from the recent work describing novel linear and circular forms of a long non-coding RNA, ANRIL [16]. This RNA consists of 19 exons. The last exons (exons 13-19) are mapped to the 50 kb CAD duplication (Figure 1A). In some forms, the order of exons is abnormal, e.g. exons 16-17-18-19-13-14-15-16. Such exon order may be explained not only by existence of a circular form but also by a splicing of a transcript overlapping two neighboring 50 kb duplicated regions within the CAD interval.

Further studies that would integrate GWAS data with those obtained from detailed genomic analysis of multiple tissues by TAR cloning in combination with qPCR are required to understand the complex genetic regulation of the 9p21.3 region and its role in the development of cardiovascular disease. The discovery of SDs within the CAD interval in the human peripheral blood mononuclear cells derived from a patient with the coronary artery disease suggests that the number and/or divergence of duplications may be an additional factor affecting the development of this disease. Though the CAD interval does not contain protein coding genes but it does encode a long non-coding RNA (ANRIL). It is presumed that some disease risk factor acts through long non-coding RNAs. Thus, the number of duplications may affect the level of ANRIL expression or splicing of this non-coding RNA located within the duplicated region that may determine predisposition to CAD. It is worth noting that specific SNPs linked to CAD are located within the duplicated region and are presented in the spliced products. This is very similar to the 12 kb repeat at Xq27, within which the SPANX-B gene is located. The number of this repeat varies from 1 up to 7 copies in different individuals [51]. It is known that SPANX-B is expressed only in testis. So, it may be assumed that the change of SPANX-B expression due to its copy number is linked to predisposition to prostate cancer or infertility. Thus, only a direct analysis of ANRIL expression in multiple CAD patients will give a final answer of involvement of the found duplication in this particular locus in cardiac disease.

To summarize, before a conclusive link between the SDs and the cardiovascular diseases can be made, further analysis is required on the CAD interval in more patients with coronary artery disease and in the human population, using the TAR cloning technique in combination with qPCR or Droplet digital PCR developed in this work. Moreover, our finding may be interesting to other labs working with the 9p21.3 region tightly related with other diseases such as myocardial infarction, ischemic stroke and aortic aneurysm.

MATERIALS AND METHODS

Cell lines and media

The cell line C087 was derived from somatic cells of a patient with unique mutations within the CAD (coronary artery disease) interval. The cells were grown in mTeSR™1 medium (Stemcells™ Technologies). The patient C087 is known to have coronary artery disease, and is a male, with an age of 49 years old at the time of PBMC isolation. The patient carries risk alleles at 9p21.3, i.e. rs10757278, rs1333049, and rs2383207.

Genomic DNA

Human DNA purified from blood cells of normal individuals was purchased from Promega (Cat. No. G3041). Genomic DNA from normal human individuals was used for qPCR and Droplet digital PCR. An additional 66 DNA samples were purchased from Coriell Institute for Medical Research. More DNA samples were obtained from Dr. Joanna Schluetker (Institute of Medical Technology, University of Tampere, Finland). A detailed description of these samples is presented elsewhere [50, 51].

Construction of TAR cloning vectors

The transformation-associated recombination (TAR) vectors, 9p21-1 and 9p21-2, were constructed using the basic vector pVC604 containing a yeast selectable marker HIS3 and a centromere from chromosome 6 [43]. To construct TAR vector 9p21-1, the 171-bp XhoI-ClaI (hook1A) and 209-bp ClaI-SpeI (hook1B) fragments corresponding to 5′ and 3′ regions of the CAD region were inserted into the polylinker of pVC604. The 5′ and 3′ targeting sequences of the vector 9p21-1 were designed based on the available information (hg19) and correspond to positions 22,062,540 to 22,062,711 and 22,129,014 to 22,129,313 on the chromosome 9 sequence. The expected size of the targeted genomic fragment is 66,763 bp. To construct TAR vector 9p21-2, the 185-bp XhoI-ClaI (hook2A) and 242-bp ClaI-SpeI (hook2B) fragments corresponding to 5′ and 3′ regions of the CAD region were inserted into the polylinker of pVC604. The 5′ and 3′ targeting sequences of the vector 9p21-2 were designed based on the available information (hg19) and correspond to positions 22,061,474 to 22,061,658 and 22,130,089 to 22,130,330 on the chromosome 9 sequence. The expected size of the targeted genomic fragment is 68,856 bp. Hook2A is located 1,066 bp downstream of hook1A. Hook2B is located 776 bp upstream of hook1B. The targeting sequences (hooks) were cloned into vector pVC604 in orientations corresponding to their orientations in the human genome sequence (version hg19). The TAR vectors were linearized with ClaI (the site is located between the targeting sequences) before transformation to yield a molecule bounded by the desired region(s) sequences. Detailed physical maps of the TAR vectors are shown in Supplementary Figure 1.

Yeast strain and transformation

A general scheme of TAR cloning is presented in Supplementary Figure 2A. For transformations, the highly transformable *Saccharomyces cerevisiae* strain VL6–48 (MATα, his3-Δ200, trp1-Δ1, ura3-52, lys2, ade2-101, met14) that has HIS3 deleted was used [43]. For TAR cloning, 2 to 3 μg of high molecular weight human genomic DNA were prepared, mixed with a ClaI-linearized TAR vector (1 μg), and presented to freshly prepared yeast spheroplasts. Yeast transformants were selected on synthetic complete medium plates lacking histidine. The yield of His^+ transformants per 2–3 μg of genomic DNA using 1 μg of the TAR vector and 5×10^8 spheroplasts varied between 10 and 60 colonies per plate. To identify desired CAD region-containing clones, the transformants were combined into pools and examined with the diagnostic primers (Supplementary Table 1) for the unique sequences not present in the TAR vectors but specific for the targeted genomic region. With DNA isolated from the C087 cells using the 9p21-1 and 9p21-2 TAR vectors the total number of transformants examined was 360. The transformants were organized into 12 pools, each containing 30 transformants. Eight pools were CAD-region-positive. In each pool, one individual transformant was CAD-region-positive. To identify which allele was cloned, we examined three TAR clones by PCR using pairs of primers specific for the mutated allele of the patient (three SNPs are characteristic for the mutant allele of the C087 patient, i.e. rs10757278, rs1333049, rs2383207) (Supplementary Table 1). Two transformants from the C087 patient were proven to be mutant after sequencing of PCR products. TAR cloning from normal genomic DNA (Promega) was carried out using the TAR vector 9p21-1. Three CAD-region-positive clones among 270 transformants analyzed were identified by a set of PCR reactions using pairs of diagnostic primers (Supplementary Table 1).

Conversion of TAR/YAC isolated clones into a BAC form

A general scheme of retrofitting of a TAR/YAC molecule into a BAC molecule is presented in Supplementary Figure 2B. A retrofitting vector either BRV1 [52] or JH-BRV1 [53] contains the two short targeting hooks (300 bp each), separated by the unique BamHI site, that flank the ColE1 origin of replication in the pVC604-based TAR cloning vector. The hooks are homologous to the vector sequences of pVC604. Recombination of a BamHI-linearized retrofitting vector

with a TAR/YAC vector part in yeast leads to replacement of the ColE1 origin of replication in the TAR cloning vector by a cassette containing the F' factor origin replication, the chloramphenicol acetyltransferase (Cm^R) gene, and the URA3 yeast selectable marker. A standard lithium acetate transformation procedure was used for retrofitting of YACs into BACs. YAC retrofitting was highly efficient: more than 95% of Ura^+His^+ transformants contained retrofitted YACs. The YAC/BACs were moved from yeast to *Escherichia coli* by electroporation. In brief, yeast chromosome-size DNAs were prepared in agarose plugs and, after melting and agarase treatment, the DNAs were electroporated into DH10B competent cells (Gibco/BRL) by using a Bio-Rad Gene Pulser as previously described [52].

Physical characterization of YAC/TAR clones

Several approaches were taken to characterize the cloned material in a YAC form. To prove the presence of the predicted genomic sequences in YAC isolates, DNA from the yeast clones was examined by PCR with pairs of overlapping primers (Supplementary Table 1) covering the entire CAD region. To check the size of the YAC inserts, Southern blot hybridization was performed with a ^{32}P-labeled probe. Specifically, genomic yeast DNA from CAD-positive clones was prepared in agarose plugs and exposed to a low dose of γ-rays (30 krad) to linearize circular YAC molecules [52]. The irradiated DNA was CHEF separated, and blot hybridized with a 125 bp yeast CEN6 probe. CEN6 sequence was amplified from the TAR vector using a unique pair of primers (Supplementary Table 1). Blots were incubated for 2 hrs at 65 °C in prehybridization Church's buffer (0.5 M Na-phosphate buffer containing 7% SDS and 100 μg/ml of unlabeled salmon sperm carrier DNA). The labeled probe was heat denatured in boiling water for 5 min and snap cooled on ice. The probe was added to the hybridization buffer and allowed to hybridize overnight at 65 °C. Blots were washed twice in 2 × SSC (300 mM NaCl, 30 mM sodium citrate, pH 7.0), 0.1% SDS for 30 min at room temperature, then three times in 0.1× SSC, 0.1% SDS for 30 min at 65 °C. Blots were exposed to X-ray film for 24 hrs at −70 °C.

Physical characterization of BAC clones

Several approaches were taken to characterize the TAR-cloned material in a BAC form. To check the size of the cloned inserts in the CAD-region-positive BAC clones, BAC DNA was digested with NotI that cuts only in the vector part, separated by clamped homogeneous electrical field electrophoresis (CHEF), and stained with EtBr. To prove the identity of TAR isolates, BAC DNA was digested by HindIII and run in 1% agarose gel. The ends of the BAC inserts were sequenced using specific primers (Supplementary Table 1). To demonstrate the presence of segmental duplications in the TAR isolates, the BAC DNA was digested either by AgeI or PmeI or FspI endonucleases that are present only once in the CAD interval.

Analysis of junction between duplications

The junction sequence between duplications within the CAD interval was confirmed by PCR reaction using specific pairs of primers (Supplementary Table 1). The PCR products were blunt end cloned into pBluescipt II plasmid (Stratagene) and then sequenced using vector primers (Supplementary Table 1). All sequences were aligned and categorized.

Construction of the vector V231

Construction of the V231 vector, which was used to re-clone the 50 kb repeats within the BAC A218, is illustrated in Supplementary Figure 5. Primers B095 and B102 (Supplementary Table 1) were used to amplify a subsection of the TAR-BRV-tTA^{VP64} plasmid [54], which included the BAC backbone, Chloramphenicol resistance gene (Cm^R) and the PmeI-AscI-BamHI-AsiSI polylinker. The B095/B102 PCR product was then circularized to produce the plasmid A225. The plasmid A225 was then linearized with SacI and AscI. In parallel, the primers B635 and B629 (Supplementary Table 1) were used to PCR amplify the pBlueScript II KS plasmid (Stratagene) and add a polylinker. The resulting PCR fragment was digested with SacI and AscI. Then the linearized A225 plasmid and the SacI/AscI-digested PCR product were ligated to produce the vector V231.

Sub-cloning of 50 kb repeats from the CAD-containing BAC

The general scheme of sub-cloning of the BAC regions is shown in Figures 4. First, DNA isolated from the CAD-containing BAC A218 (from a patient-derived cell line) (Figure 4A) was digested either by PmeI or AatII and then re-ligated. The ligation mixtures were electroporated into the ElectroMax DH10B bacterial cells (Life Technologies, Cat. No. 18290015). The Clm^R colonies were screened by three rounds of PCR reaction: i) by the primers specific for the yeast HIS3 gene (a positive control), ii) by the primers specific for the sequences around a PmeI site (a positive control) and iii) by the primers specific for the junction sequence, B578/B586, (a negative control) (Supplementary Table 1) to select the desirable clones for Next-Generation Sequencing (Figure 4B). Second, DNA from BAC A218 was digested by AgeI and run on CHEF. Then the fragments of ~50 kb in size corresponding to duplicated units were gel-isolated and ligated into the vector V231. The ligation

mixture was electroporated into the ElectroMax DH10B bacterial cells (Life Technologies, Cat. No.18290015). The Clm^R colonies were screened by PCR reaction using the primers specific for the junction between duplications (Supplementary Table 1). BACs negative for junction sequence were selected and used for Next-Generation Sequencing.

Southern-blot hybridization analysis

Southern-blot hybridization was performed with a ^{32}P-labelled probe as described previously [44, 52] with minor changes. Genomic DNA was prepared in agarose plugs and restriction-digested by NotI in the buffer recommended by the manufactory. The digested DNA was CHEF (CHEF Mapper, Bio-Rad) separated (autoprogram, 5-300 kb range, 16 hrs transfer), transferred to membrane (Amersham Hybond-N+) and blot-hybridized with a 862-bp probe specific for the CAD interval sequence in the 9p21.3 region. DNA sequence for the probe was amplified by PCR using the primers indicated in Supplementary Table 1. Blots were incubated for 2 hrs at 65°C in pre-hybridization Church's buffer (0.5 M Na-phosphate buffer containing 7% SDS and 100 μg/ml of unlabelled salmon sperm carrier DNA). The labeled probe was heat denatured in boiling water for 5 min and snap cooled on ice. The probe was added to the hybridization buffer and allowed to hybridize overnight at 65°C. Blot was washed twice in 2× SSC (300 mM NaCl, 30 mM sodium citrate, pH 7.0), 0.05% SDS for 10 min at room temperature, then twice in 2× SSC, 0.05% SDS for 5 min at 60°C, twice in 0.5× SSC, 0.05% SDS for 5 min at 60°C and twice in 0.25× SSC, 0.05% SDS for 5 min at 60°C. Blot was exposed to X-ray film for 24–72 hrs at -80°C.

Determination of the copy number of SDs by quantitative real-time PCR

Three regions within the CAD interval were chosen for analysis. Two regions correspond to 5' and 3' ends of the 50 kb repeated unit. The third region is outside of the repeated unit and was used as the first control for copy number. The TaqMan probe and primers were designed using the PrimerQuest® Design Tool (IDT Inc.), following the criteria indicated in the program. The specific TaqMan 9p21_5' *27 bp* probe complementary to the analyzed region was designed (Supplementary Table 1). The 9p21_5' probe contains a fluorophore FAM as a reporter. 9p21_5' forward and reverse primers were created from the region on chromosome 9 (positions 22078776 to 22078919 in hg19). The size of the amplicon is 144 bp. The probe and primers were provided by IDT Inc. as a pre-mixed assay, which was diluted with sterile 1x TE buffer till 20x working concentration. RNAaseP kit was used as an internal reference (Applied Biosystems). The kit contains 20xRNAaseP mix with a VIC-labeled probe and specific primers for the RNAaseP gene used as the second control for copy number. As an additional control and double check, another set of probes was synthetized. The specific TaqMan 9p21_3' probe 24 bp in size complementary to the analyzed region (chr9: positions 22117414 to 22117550 in hg19) and the specific TaqMan 9p21_control probe 23 bp in size complementary the unique DNA sequence present only in one copy per haploid genome (chr9: positions 22062099 to 22062185 in hg19) were designed. The 9p21_3' probe contains a fluorophore 5′ FAM as a reporter. The size of the amplicon is 137 bp. The 9p21_control probe contains a fluorophore HEX/ZEN as a reporter. The size of the amplicon is 87 bp. Primer sequences and additional information are described in Supplementary Table 1. The samples were analyzed separately with set 1 (*9p21_5'* and RNAaseP) and set 2 (*9p21_3'* and *9p21_control*). In control experiments, amplification efficiency was close to 100% for all amplicons, signifying that both reactions proceeded with very high efficiencies. qPCR reactions were carried out using an CFX Connect (Bio-Rad) in a 96-well optical plate with a final reaction volume of 20 μl. All reactions in each plate were prepared from a single PCR Mastermix consisting of 2×iTaq Universal Probes Super Mix (Bio-Rad), 20×9p21_5' Mix, 20×RNAaseP Mix, and HPLC pure water. A total of 10 ng of DNA template (5 μl) was dispensed into each of three sample wells for triplicate reactions. Each sample was run in triplicates to quantify the CAD analyzed region compared to the control probe sequences. Thermal cycling conditions included a pre-run of 3 min at 95°C. Cycle conditions were 40 cycles at 95°C for 10 sec and 60°C for 30 sec, according to the CFX 2 Step PCR Amplification Protocol (Bio-Rad). The relative copy number of *repeated units* was calculated for each sample using the comparative Pfaffl method.

Determination of the copy number of SDs by droplet digital PCR

Droplet digital PCR was performed according to manufacturer's protocol (Bio-Rad). A specific FAM-labeled probe (dd9p21) against a region of interest was designed by Bio-Rad (Supplementary Table 1) and provider as ready-to-work 20x solution. As a control probe, we used Bio-Rad provided control against RPP30 which is present at one copy per cell. All reactions in each plate were prepared from a single PCR Mastermix consisting of ddPCR Supermix for Probes (Bio-Rad), dd9p21 Mix, RPP30 Mix, and HPLC pure water. A total of 10 ng of DNA template (5 μL) was dispensed into each of the three sample wells for triplicate reactions. Each sample was run in triplicates to quantify the 9p21.3 studied region compared with the internal RPP30 control gene. Bio-Rad provided "Quantasoft" software was used for data analysis. Reactions were performed with help of CCR Genomics Core (NIH/NCI).

Illumina sequencing and library preparation

Sequencing libraries were prepared by tagmentation using the Nextera DNA Library Preparation Kit (Illumina Inc., San Diego, CA) according to the manufacturer's instructions. In short, 50 ng BAC DNA was tagged and fragmented using the Nextera transposome. The DNA was column purified, and amplified by suppression PCR introducing P5 and P7 ends with dual 8-nucleotide index sequences. Libraries were quantified by qPCR and sequenced on either MiSeq or NextSeq sequencers.

Illumina alignment and variant calling

Alignments were performed with bwa version 0.7.12-r1039 [55] using the bwa mem algorithm against the hg19 reference genome. Duplicates were marked with picard version 1.92(1464) (https://github.com/broadinstitute/picard) using the MarkDuplicates tool. Variants from the reference genome were identified with freebayes version 0.9.21-7-g7dd41db [56] and non-SNVs were removed with vcftools version 0.1.14 [57].

Author contributions

NK, NCOL, ML, VNN performed the experiments. JJW, RLW, PSM conducted the sequencing experiments. NK, NCOL, ML designed the experiments. NK, VL, NCOL wrote the manuscript. The manuscript was prepared by NK, ML, NCOL.

ACKNOWLEDGMENTS

This work was supported by the Intramural Research Program of the NIH, National Cancer Institute, Center for Cancer Research, USA (VL and NK). The authors would like to thank Dr. Fyodor D. Urnov (Sangamo BioSciences Inc.), Dr. William Ferguson (The Scripps Research Institute), Dr. Kristin K. Baldwin (The Scripps Research Institute), and Dr. Ali Torkamani (The Scripps Research Institute) for the interest to this work. The cell line C087 was kindly provided by Dr. Eric Topol (The Scripps Translational Science Institute, The Scripps Research Institute and Scripps Health, La Jolla, CA 92037, USA).

CONFLICTS OF INTEREST

The authors declare that they have no competing interest.

REFERENCES

1. de los Campos G, Gianola D, Allison DB. Predicting genetic predisposition in humans: the promise of whole-genome markers. Nat Rev Genet. 2010; 11:880-6. https://doi.org/10.1038/nrg2898.
2. Gschwendtner A, Bevan S, Cole JW, Plourde A, Matarin M, Ross-Adams H, Meitinger T, Wichmann E, Mitchell BD, Furie K, Slowik A, Rich SS, Syme PD, et al. Sequence variants on chromosome 9p21.3 confer risk for atherosclerotic stroke. Ann Neurol. 2009; 65:531-9. https://doi.org/10.1002/ana.21590.
3. Matarin M, Brown WM, Singleton A, Hardy JA, Meschia JF; ISGS investigators. Whole genome analyses suggest ischemic stroke and heart disease share an association with polymorphisms on chromosome 9p21. Stroke. 2008; 39:1586-9. https://doi.org/10.1161/STROKEAHA.107.502963.
4. Helgadottir A, Thorleifsson G, Magnusson KP, Gretarsdottir S, Steinthorsdottir V, Manolescu A, Jones GT, Rinkel GJ, Blankensteijn JD, Ronkainen A, Jaaskelainen JE, Kyo Y, Lenk GM, et al. The same sequence variant on 9p21 associates with myocardial infarction, abdominal aortic aneurysm and intracranial aneurysm. Nat Genet. 2008; 40:217-24. https://doi.org/10.1038/ng.72.
5. Biros E, Cooper M, Palmer LJ, Walker PJ, Norman PE, Golledge J. Association of an allele on chromosome 9 and abdominal aortic aneurysm. Atherosclerosis. 2010; 212:539-42. https://doi.org/10.1016/j.atherosclerosis.2010.06.015.
6. Bown MJ, Braund PS, Thompson J, London NJ, Samani NJ, Sayers RD. Association between the coronary artery disease risk locus on chromosome 9p21.3 and abdominal aortic aneurysm. Circ Cardiovasc Genet. 2008; 1:39-42. https://doi.org/10.1161/CIRCGENETICS.108.789727.
7. Broadbent HM, Peden JF, Lorkowski S, Goel A, Ongen H, Green F, Clarke R, Collins R, Franzosi MG, Tognoni G, Seedorf U, Rust S, Eriksson P, et al. Susceptibility to coronary artery disease and diabetes is encoded by distinct, tightly linked SNPs in the ANRIL locus on chromosome 9p. Hum Mol Genet. 2008; 17:806-14. https://doi.org/10.1093/hmg/ddm352.
8. Cluett C, McDermott MM, Guralnik J, Ferrucci L, Bandinelli S, Miljkovic I, Zmuda JM, Li R, Tranah G, Harris T, Rice N, Henley W, Frayling TM, et al. The 9p21 myocardial infarction risk allele increases risk of peripheral artery disease in older people. Circ Cardiovasc Genet. 2009; 2:347-53. https://doi.org/10.1161/CIRCGENETICS.108.825935.
9. Helgadottir A, Thorleifsson G, Manolescu A, Gretarsdottir S, Blondal T, Jonasdottir A, Jonasdottir A, Sigurdsson A, Baker A, Palsson A, Masson G, Gudbjartsson DF, Magnusson KP, et al. A common variant on chromosome 9p21 affects the risk of myocardial infarction. Science. 2007; 316:1491-3. https://doi.org/10.1126/science.1142842.
10. Holdt LM, Teupser D. Recent studies of the human chromosome 9p21 locus, which is associated with atherosclerosis in human populations. Arterioscler Thromb Vasc Biol. 2012; 32:196-206. https://doi.org/10.1161/ATVBAHA.111.232678.
11. McPherson R, Pertsemlidis A, Kavaslar N, Stewart A, Roberts R, Cox DR, Hinds DA, Pennacchio LA,

Tybjaerg-Hansen A, Folsom AR, Boerwinkle E, Hobbs HH, Cohen JC. A common allele on chromosome 9 associated with coronary heart disease. Science. 2007; 316:1488-91. https://doi.org/10.1126/science.1142447.

12. Seidelmann SB, Kuo C, Pleskac N, Molina J, Sayers S, Li R, Zhou J, Johnson P, Braun K, Chan C, Teupser D, Breslow JL, Wight TN, et al. Athsq1 is an atherosclerosis modifier locus with dramatic effects on lesion area and prominent accumulation of versican. Arterioscler Thromb Vasc Biol. 2008; 28:2180-6. https://doi.org/10.1161/ATVBAHA.108.176800.

13. Smith JG, Melander O, Lovkvist H, Hedblad B, Engstrom G, Nilsson P, Carlson J, Berglund G, Norrving B, Lindgren A. Common genetic variants on chromosome 9p21 confers risk of ischemic stroke: a large-scale genetic association study. Circ Cardiovasc Genet. 2009; 2:159-64. https://doi.org/10.1161/CIRCGENETICS.108.835173.

14. Thompson AR, Golledge J, Cooper JA, Hafez H, Norman PE, Humphries SE. Sequence variant on 9p21 is associated with the presence of abdominal aortic aneurysm disease but does not have an impact on aneurysmal expansion. Eur J Hum Genet. 2009; 17:391-4. https://doi.org/10.1038/ejhg.2008.196.

15. Ye S, Willeit J, Kronenberg F, Xu Q, Kiechl S. Association of genetic variation on chromosome 9p21 with susceptibility and progression of atherosclerosis: a population-based, prospective study. J Am Coll Cardiol. 2008; 52:378-84. https://doi.org/10.1016/j.jacc.2007.11.087.

16. Burd CE, Jeck WR, Liu Y, Sanoff HK, Wang Z, Sharpless NE. Expression of linear and novel circular forms of an INK4/ARF-associated non-coding RNA correlates with atherosclerosis risk. PLoS Genet. 2010; 6:e1001233. https://doi.org/10.1371/journal.pgen.1001233.

17. Bansal V, Harismendy O, Tewhey R, Murray SS, Schork NJ, Topol EJ, Frazer KA. Accurate detection and genotyping of SNPs utilizing population sequencing data. Genome Res. 2010; 20:537-45. https://doi.org/10.1101/gr.100040.109.

18. Harismendy O, Notani D, Song X, Rahim NG, Tanasa B, Heintzman N, Ren B, Fu XD, Topol EJ, Rosenfeld MG, Frazer KA. 9p21 DNA variants associated with coronary artery disease impair interferon-gamma signalling response. Nature. 2011; 470:264-8. https://doi.org/10.1038/nature09753.

19. Scott LJ, Mohlke KL, Bonnycastle LL, Willer CJ, Li Y, Duren WL, Erdos MR, Stringham HM, Chines PS, Jackson AU, Prokunina-Olsson L, Ding CJ, Swift AJ, et al. A genome-wide association study of type 2 diabetes in Finns detects multiple susceptibility variants. Science. 2007; 316:1341-5. https://doi.org/10.1126/science.1142382.

20. Zeggini E, Weedon MN, Lindgren CM, Frayling TM, Elliott KS, Lango H, Timpson NJ, Perry JR, Rayner NW, Freathy RM, Barrett JC, Shields B, Morris AP, et al. Replication of genome-wide association signals in UK samples reveals risk loci for type 2 diabetes. Science. 2007; 316:1336-41. https://doi.org/10.1126/science.1142364.

21. Bishop DT, Demenais F, Iles MM, Harland M, Taylor JC, Corda E, Randerson-Moor J, Aitken JF, Avril MF, Azizi E, Bakker B, Bianchi-Scarra G, Bressac-de Paillerets B, et al. Genome-wide association study identifies three loci associated with melanoma risk. Nat Genet. 2009; 41:920-5. https://doi.org/10.1038/ng.411.

22. Cunnington MS, Santibanez Koref M, Mayosi BM, Burn J, Keavney B. Chromosome 9p21 SNPs associated with multiple disease phenotypes correlate with ANRIL expression. PLoS Genet. 2010; 6:e1000899. https://doi.org/10.1371/journal.pgen.1000899.

23. Shete S, Hosking FJ, Robertson LB, Dobbins SE, Sanson M, Malmer B, Simon M, Marie Y, Boisselier B, Delattre JY, Hoang-Xuan K, El Hallani S, Idbaih A, et al. Genome-wide association study identifies five susceptibility loci for glioma. Nat Genet. 2009; 41:899-904. https://doi.org/10.1038/ng.407.

24. Wrensch M, Jenkins RB, Chang JS, Yeh RF, Xiao Y, Decker PA, Ballman KV, Berger M, Buckner JC, Chang S, Giannini C, Halder C, Kollmeyer TM, et al. Variants in the CDKN2B and RTEL1 regions are associated with high-grade glioma susceptibility. Nat Genet. 2009; 41:905-8. https://doi.org/10.1038/ng.408.

25. Li WQ, Pfeiffer RM, Hyland PL, Shi J, Gu F, Wang Z, Bhattacharjee S, Luo J, Xiong X, Yeager M, Deng X, Hu N, Taylor PR, et al. Genetic polymorphisms in the 9p21 region associated with risk of multiple cancers. Carcinogenesis. 2014; 35:2698-705. https://doi.org/10.1093/carcin/bgu203.

26. Burdon KP, Crawford A, Casson RJ, Hewitt AW, Landers J, Danoy P, Mackey DA, Mitchell P, Healey PR, Craig JE. Glaucoma risk alleles at CDKN2B-AS1 are associated with lower intraocular pressure, normal-tension glaucoma, and advanced glaucoma. Ophthalmology. 2012; 119:1539-45. https://doi.org/10.1016/j.ophtha.2012.02.004.

27. Nakano M, Ikeda Y, Tokuda Y, Fuwa M, Omi N, Ueno M, Imai K, Adachi H, Kageyama M, Mori K, Kinoshita S, Tashiro K. Common variants in CDKN2B-AS1 associated with optic-nerve vulnerability of glaucoma identified by genome-wide association studies in Japanese. PLoS One. 2012; 7:e33389. https://doi.org/10.1371/journal.pone.0033389.

28. Gil J, Peters G. Regulation of the INK4b-ARF-INK4a tumour suppressor locus: all for one or one for all. Nat Rev Mol Cell Biol. 2006; 7:667-77. https://doi.org/10.1038/nrm1987.

29. Popov N, Gil J. Epigenetic regulation of the INK4b-ARF-INK4a locus: in sickness and in health. Epigenetics. 2010; 5:685-90.

30. Nobori T, Takabayashi K, Tran P, Orvis L, Batova A, Yu AL, Carson DA. Genomic cloning of methylthioadenosine phosphorylase: a purine metabolic enzyme deficient in multiple different cancers. Proc Natl Acad Sci U S A. 1996; 93:6203-8.

31. Behrmann I, Wallner S, Komyod W, Heinrich PC, Schuierer M, Buettner R, Bosserhoff AK. Characterization

of methylthioadenosin phosphorylase (MTAP) expression in malignant melanoma. Am J Pathol. 2003; 163:683-90. https://doi.org/10.1016/S0002-9440(10)63695-4.

32. Pasmant E, Sabbagh A, Vidaud M, Bieche I. ANRIL, a long, noncoding RNA, is an unexpected major hotspot in GWAS. FASEB J. 2011; 25:444-8. https://doi.org/10.1096/fj.10-172452.

33. Pasmant E, Laurendeau I, Heron D, Vidaud M, Vidaud D, Bieche I. Characterization of a germ-line deletion, including the entire INK4/ARF locus, in a melanoma-neural system tumor family: identification of ANRIL, an antisense noncoding RNA whose expression coclusters with ARF. Cancer Res. 2007; 67:3963-9. https://doi.org/10.1158/0008-5472.CAN-06-2004.

34. Yu W, Gius D, Onyango P, Muldoon-Jacobs K, Karp J, Feinberg AP, Cui H. Epigenetic silencing of tumour suppressor gene p15 by its antisense RNA. Nature. 2008; 451:202-6. https://doi.org/10.1038/nature06468.

35. Folkersen L, Kyriakou T, Goel A, Peden J, Mälarstig A, Paulsson-Berne G, Hamsten A, Hugh Watkins, Franco-Cereceda A, Gabrielsen A, Eriksson P; PROCARDIS consortia. Relationship between CAD risk genotype in the chromosome 9p21 locus and gene expression. Identification of eight new ANRIL splice variants. PLoS One. 2009; 4:e7677. https://doi.org/10.1371/journal.pone.0007677.

36. Gupta RA, Shah N, Wang KC, Kim J, Horlings HM, Wong DJ, Tsai MC, Hung T, Argani P, Rinn JL, Wang Y, Brzoska P, Kong B, et al. Long non-coding RNA HOTAIR reprograms chromatin state to promote cancer metastasis. Nature. 2010; 464:1071-6. https://doi.org/10.1038/nature08975.

37. Pandey RR, Mondal T, Mohammad F, Enroth S, Redrup L, Komorowski J, Nagano T, Mancini-Dinardo D, Kanduri C. Kcnq1ot1 antisense noncoding RNA mediates lineage-specific transcriptional silencing through chromatin-level regulation. Mol Cell. 2008; 32:232-46. https://doi.org/10.1016/j.molcel.2008.08.022.

38. Rinn JL, Kertesz M, Wang JK, Squazzo SL, Xu X, Brugmann SA, Goodnough LH, Helms JA, Farnham PJ, Segal E, Chang HY. Functional demarcation of active and silent chromatin domains in human HOX loci by noncoding RNAs. Cell. 2007; 129:1311-23. https://doi.org/10.1016/j.cell.2007.05.022.

39. Terranova R, Yokobayashi S, Stadler MB, Otte AP, van Lohuizen M, Orkin SH, Peters AH. Polycomb group proteins Ezh2 and Rnf2 direct genomic contraction and imprinted repression in early mouse embryos. Dev Cell. 2008; 15:668-79. https://doi.org/10.1016/j.devcel.2008.08.015.

40. Zhao J, Sun BK, Erwin JA, Song JJ, Lee JT. Polycomb proteins targeted by a short repeat RNA to the mouse X chromosome. Science. 2008; 322:750-6. https://doi.org/10.1126/science.1163045.

41. Liu Y, Sanoff HK, Cho H, Burd CE, Torrice C, Mohlke KL, Ibrahim JG, Thomas NE, Sharpless NE. INK4/ARF transcript expression is associated with chromosome 9p21 variants linked to atherosclerosis. PLoS One. 2009; 4:e5027. https://doi.org/10.1371/journal.pone.0005027.

42. Kouprina N, Larionov V. TAR cloning: insights into gene function, long-range haplotypes and genome structure and evolution. Nat Rev Genet. 2006; 7:805-12. https://doi.org/10.1038/nrg1943.

43. Kouprina N, Larionov V. Selective isolation of genomic loci from complex genomes by transformation-associated recombination cloning in the yeast Saccharomyces cerevisiae. Nat Protoc. 2008; 3:371-7. https://doi.org/10.1038/nprot.2008.5.

44. Larionov V, Kouprina N, Graves J, Resnick MA. Highly selective isolation of human DNAs from rodent-human hybrid cells as circular yeast artificial chromosomes by transformation-associated recombination cloning. Proc Natl Acad Sci U S A. 1996; 93:13925-30.

45. Kouprina N, Larionov V. Transformation-associated recombination (TAR) cloning for genomics studies and synthetic biology. Chromosoma. 2016; 125:621-32. https://doi.org/10.1007/s00412-016-0588-3.

46. Stankiewicz P, Lupski JR. The genomic basis of disease, mechanisms and assays for genomic disorders. Genome Dyn. 2006; 1:1-16. https://doi.org/10.1159/000092496.

47. Carvalho CM, Zhang F, Lupski JR. Evolution in health and medicine Sackler colloquium: genomic disorders: a window into human gene and genome evolution. Proc Natl Acad Sci U S A. 2010; 107:1765-71. https://doi.org/10.1073/pnas.0906222107.

48. La Cognata V, Morello G, D'Agata V, Cavallaro S. Copy number variability in Parkinson's disease: assembling the puzzle through a systems biology approach. Hum Genet. 2017; 136:13-37. https://doi.org/10.1007/s00439-016-1749-4.

49. Bailey JA, Liu G, Eichler EE. An Alu transposition model for the origin and expansion of human segmental duplications. Am J Hum Genet. 2003; 73:823-34. https://doi.org/10.1086/378594.

50. Kouprina N, Noskov VN, Solomon G, Otstot J, Isaacs W, Xu J, Schleutker J, Larionov V. Mutational analysis of SPANX genes in families with X-linked prostate cancer. Prostate. 2007; 67:820-8. https://doi.org/10.1002/pros.20561.

51. Kouprina N, Pavlicek A, Noskov VN, Solomon G, Otstot J, Isaacs W, Carpten JD, Trent JM, Schleutker J, Barrett JC, Jurka J, Larionov V. Dynamic structure of the SPANX gene cluster mapped to the prostate cancer susceptibility locus HPCX at Xq27. Genome Res. 2005; 15:1477-86. https://doi.org/10.1101/gr.4212705.

52. Kouprina N, Annab L, Graves J, Afshari C, Barrett JC, Resnick MA, Larionov V. Functional copies of a human gene can be directly isolated by transformation-associated recombination cloning with a small 3' end target sequence. Proc Natl Acad Sci U S A. 1998; 95:4469-74.

53. Kim JH, Kononenko A, Erliandri I, Kim TA, Nakano M, Iida Y, Barrett JC, Oshimura M, Masumoto H,

Earnshaw WC, Larionov V, Kouprina N. Human artificial chromosome (HAC) vector with a conditional centromere for correction of genetic deficiencies in human cells. Proc Natl Acad Sci U S A. 2011; 108:20048-53. https://doi.org/10.1073/pnas.1114483108.

54. Kononenko AV, Lee NC, Liskovykh M, Masumoto H, Earnshaw WC, Larionov V, Kouprina N. Generation of a conditionally self-eliminating HAC gene delivery vector through incorporation of a tTAVP64 expression cassette. Nucleic Acids Res. 2015; 43:e57. https://doi.org/10.1093/nar/gkv124.

55. Li H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. ArXiv e-prints. 2013.

56. Garrison E, Marth G. Haplotype-based variant detection from short-read sequencing. ArXiv e-prints. 2012.

57. Danecek P, Auton A, Abecasis G, Albers CA, Banks E, DePristo MA, Handsaker RE, Lunter G, Marth GT, Sherry ST, McVean G, Durbin R; 1000 Genomes Project Analysis Group. The variant call format and VCFtools. Bioinformatics. 2011; 27:2156-8. https://doi.org/10.1093/bioinformatics/btr330.