# Germinative retrotransposition of mature mRNA creates novelty across the human genomic and transcriptomic landscape.

**Fábio C. P. Navarro1,2,§, Andrei Rozanski2 and Pedro A. F. Galante2,§.**

1 Departamento de Bioquímica, Universidade de São Paulo, São Paulo, Brazil   
2 Centro de Oncologia Molecular, Hospital Sírio-Libanês, São Paulo, Brazil  
§ Corresponding authors

## Abstract

In primates, including humans, repetitive elements account for approximately 40% of nuclear DNA. Retrotransposition of these, and other transcripts, greatly influence genetic novelty among these organisms. Analysis of hundreds of human genomes suggests that protein-coding transcripts retrotransposition creates presence/absence retrocopy dimorphism (retroCNVs) across human populations. To further estimate their frequency and their influence on the human transcriptome we analyzed 2.3 thousands genomes from the 1000 Genomes Project sequenced at low coverage and hundreds of transcriptomes from the same individuals using Amazon Cloud Computing. This paper presents the most extensive and comprehensive catalog of human retroCNVs and describes evidence of retroCNVs influencing the human transcriptome.

We observed 109 retroCNVs absent in the human reference genome therefore estimating one retroCNV is generated every 3,088[N] births. Most of the retroCNVs have low allele frequency and confirmed human population structure. Finally, these events revealed competent chimeric transcripts involving host gene and retroCNVs as well as putative impact on parental gene expression. In conclusion, retroCNVs are more frequent in human populations than previously expected, and contribute to the variability of the human transcriptome. Further and functional studies are needed to confirm the functional rule of these retroCNVs; these events will need to be considered in future structural variation and transcriptome studies.

## Introduction

Structural variation (SV) is suggested as one of the major factors contributing to genotype and phenotype variation in human and other mammals genomes {Weischenfeldt:2013fm}. Their extent and impact are still elusive {Conrad:2010ja, Stankiewicz:2010fg, Weischenfeldt:2013fm} and, in contrast to Single Nucleotide Polymorphisms (SNPs), there are only a handful studies describing SVs on genomic and population scale in humans {Stewart:2011bt, GenomesProjectConsortium:2012co, Schrider:2013di, Ewing:2013cs, Abyzov:2013hk, Dayama:2014da}. Among SVs, those related to retrotransposable elements are especially interesting because they are informative to forensic science {Ray:2007ef}, population genetics and genomics {Stewart:2011bt} and to the study of human diseases {Hancks:2012ij, Cooke:2014ib, Tubio:2014gm}.

Retrotransposition is not only responsible for mobilizing their own repetitive sequences {Goodier:2008ky, Tubio:2014gm, Stewart:2011bt} but, also for the mobilization of mature protein coding messenger RNAs (mRNA) within the host genome. This complex process was initially observed by Vanin et al. {Vanin:1985gw}, and creates gene duplicates {Esnault:2000dy}, also termed “retrocopies”. The human genome presents a large set of fixed retrocopies (7.5 to 10k events) from two to three thousand parental genes {Navarro:2013fx, Harrow:2006ee, Lam:2009dr}. Most of these retrocopies are annotated as processed pseudogenes {Podlaha:2009kc}, however, approximately ten percent are expected to be functional {Kaessmann:2009fc}. Some retrocopies were extensively studied and are now described as “known genes” presenting important biological activities {Frontini:2005eo, Vemuganti:2010cf, Fairbanks:2012gb, Herrera:2014bg}.

Three groups have identified a set of unfixed retrocopies, presenting presence/absence dimorphism across human populations (retroCNVs) {Schrider:2013di, Ewing:2013cs, Abyzov:2013hk}. In humans these retroCNVs were identified using 1000 Genomes phase 1 data and locally constructed algorithms. Events were analyzed regarding selective pressure, putative function, association with pathways and presence on other species. However, the real extent of retroCNVs, their impact on individuals is still need to be further addressed.

In order to further describe retroCNVs and assess their transcriptional impact in humans, we developed a novel method using the Amazon Cloud Computing to integrate and analyze publicly available data from 2,535 human genomes {GenomesProjectConsortium:2012co}[R phase 3] and 462 paired RNA-seq {tHoen:2013kk}. We manually curate every candidate across the 26 populations to generate the most comprehensive and extensive list of retroCNVs in the human genome and how they may influence the human transcriptome.

## Results [15 para]

In order to identify retrotransposition of protein coding transcripts (retroCNVs) not present in the human reference genome, we used publicly available whole genome sequencing at 8x coverage using Illumina paired-end sequencing from the 1000 Genomes Project [R]. Genome sequencing reads were aligned by 1000 Genomes project to the human reference genome hg19 (GRCh37.p13) using BWA to identify confident alignments. We used these alignments in order to identify abnormal paired-end fragments to call putative insertion points and genotype each individual. All insertion points were manually curated in order to avoid misalignments and other putative artifacts. Event description and genotype here are made available in VCF file format at [N].

### Section 1: Overall description.

We identified 109 insertion points corresponding to mRNA retroCNVs absent on the reference genome (Figure 2), including sixty two (56.88%) novel candidates not described by studies based on phase1 data from 1000 Genomes {Schrider:2013di, Abyzov:2013hk, Ewing:2013cs}. In addition, we detected all previously validated retroCNVs with exception of CACNA1B and 77% of the previously described retroCNVs (Figure S1, Table S1). Among the undetected retroCNVs, there are 14 candidates described in one publication (9 specific to Abyzov et al and 5 specific to Ewing et al) of which none were validated and most are supported by less than ten reads on the insertion point across all individuals from 1000 Genomes and Table S2). Furthermore, a couple of these loci overlap to annotated as retrocopy or pseudogene from the same parental gene on the reference genome suggesting artifacts contamination (Table S1). We also compared our set of retroCNVs to events previously described by exon-exon junctions. We found that only 24%[N] of the parental genes with evidence of exon-exon junction gDNA have a defined insertion point (Table S3).

In contrast to previous works, we expanded or candidate set by including solo-3’UTR retroCNVs. These events correspond to retrocopied fragments of 3’UTR (from 89 to 1600nt) inserted in quasi-random genomic coordinates. In total, these events correspond to 30% and 50% of all and novel retroCNVs, respectively. Again, in order to avoid miss-annotated retroduplications, every candidate was manually curated and cases resembling genomic duplications were removed from further analysis. Therefore, using the set of retroCNVs, we could estimate the rate of retroCNVs generated in humans based on the population genetics theories of Watterson and Tajima {Watterson:1975ur, Tajima:1989un}. We estimate the θ parameter using the number of segregating retroCNVs (109 insertions absent to the reference genome) and then use this to estimate the rate of retroCNVs per generation. Assuming an effective population size of 10,000, we estimate this rate as 1 retrocopy per 3,088 live births per generation (95% confidence interval: 1/[N] and 1/[N]). Our mean rate is only slightly higher than the rate of 5,177 events per meiosis derived from comparison of the Phase1 1000 Genomes {Ewing:2013cs}.

\* Mother/Father origin? \*

In order to genotype each individual from 1000 Genomes as homozygous absence, heterozygous and homozygous presence we defined each insertion point with relative precision and made the genotype file available at (VCF file with genotypes). We also calculate the allele frequency of each retroCNV in each of the 26 analyzed populations (Figure 1). Patterns such as the high frequency of retroCNV-TDG on Africans compared to Europeans or the higher allele frequency of retroCNV-PPIA on Europeans intrigued us and led us to investigate general population genetics properties of retroCNVs.

### Section 2: Population analysis

We investigate the population genetics properties of retroCNVs across sequenced individuals from the 26 populations available at 1000 Genomes project. Similar to Single Nucleotide Polymorphisms, we found an enrichment of rare variants (45% are represented by less than 10 alleles) {GenomesProjectConsortium:2012co}, therefore, retroCNVs are predominantly found in a single super population (Figure 3A). In contrast, when evaluating the retroCNVs present in more than 100 alleles (or allele frequency higher than 2%) we found that most of the super populations and, in consequence populations, contains the insertion (Figure S2). We also sought to better describe how retroCNVs are dispersed across super populations. We initially removed the American super population to avoid admixture artifacts. We found that Africans are the super population with more variants (18 “specific retroCNVs”) followed by East Asians (EAS), South Asians (SAS), and finally Europeans (EUR) with fourteen, twelve and eight specific insertions respectively (Figure 3B). In contrast, we found only one retroCNVs specific to Native Americans. Interestingly, Native Americans shared most of the retroCNVs with Africans (10 retroCNVs), followed by Europeans (3 retroCNVs) and Asians (1 retroCNV each – Figure S3). Finally, we describe that all super populations share 19 retroCNVs.

RetroCNVs specific and to a super population or population can recover population structure (Fig 3C). Performing Principal Component Analysis (PCA) enabled us to define the eigenvectors that better explained the variability between individuals (using genotype data) and populations (using allele frequency data). Despite the fact that we are only analyzing 109 variants, we found that PC1 can separate African individuals and populations from other populations (Fig 3C right panel in orange), while PC2 separates Europeans from East Asians (Fig 3C right panel in blue (EUR) and green (ASN)). Finally the PC3 separates South Asians from other populations. We reconstructed the human populations phylogenetic tree relaying on the retroCNVs common to a specific to each populations or super population by using the Jaccard index as distance measure (Fig 3D). We find that approximately 20% of the retroCNVs were present all super populations, of which, most (11 retroCNVs) are present in all analyzed populations.

### Session 3 – “Functional” Impact.

In order to assess the transcriptional impact of retroCNVs, we first analyzed their genomic insertion context. We used gene coordinates from GENCODE v19 and ENCODE regulatory annotation. We found that the annotation of retroCNVs insertion point is exactly what expected by chance: i) 65% of the retroCNVs we inserted into intergenic regions, ii) 43% into intronic region and two percent are insertions within coding regions. An extended annotation considering proximal promoters, distant regulatory modules and highly occupied by transcription factor were also carried, but the retroCNVs again fit the expected distribution. In concordance to a recent variant distribution and neutral selection pressure by a quasi-random insertion profile by L1 reverse transcriptase machinery {Gasior:2006dp} (Table S4 and Figure S4). Even though we did not find an enrichment or depletion of retroCNVs insertions into specific genomic context, we carefully analyzed retroCNVs searching for potentially impactful events.

Rare deletions and insertions on coding regions are known to cause loss of function [R]; therefore, we initially analyzed retroCNVs inserted into coding exons from host genes. The retroCNV-UQCR10 is an event previously described by us {Schrider:2013hb} and {Ewing:2013cs} where we found that the full mature mRNA from UQCR10 is inserted into the second exon of the host gene C1orf194 (Figure S5). This retroCNVs was found on 133 different individuals across most of populations analyzed but higher allele frequency on Europeans. Despite the fact the C1orf194 is annotated we found that this gene is expressed at relative high levels at testis (Figure S6). We also were able to identify evidence of expression of a chimeric transcript involving both host gene and retroCNV using RNA-seq data from the Geuvadis project. We also detected a second retroCNV inserted into a coding region (Figure S7). Differently from the previous case (a multi-exonic retrocopy), the retroCNV-ATRX encompasses only 250 nucleotides from the parental 3‘UTR inserted into DCDC1’s fifth exon. This retroCNV was found in 24 individuals across SAS, EAS, AFR and EUR. For this host-retroCNV pair, we did not found any evidence of expression of a chimeric transcript, however, it could be a limitation of analyzing RNA-seq from lymphoblastic cell-lines or low expression of the host gene (Figure S8).

Next, we analyzed RetroCNVs inserted into introns, which may also harbor genetic variance by creating novel chimeric transcripts {Baertsch:2008jj}. Belonging to this class, we found the retroCNV-CBX3, a previously described event {Schrider:2013di, Ewing:2013cs, Abyzov:2013hk} comprised of a full mRNA insertion into the third intron of the gene CCDC32 gene (Figure 4A). This allele is present on more than 75% percent of the analyzed genomes and higher allele frequency on EAS (Figure 2 and Figure S9). Both host and parental gene show higher expression on lymphocytes (Figure S10 and Figure S11). Therefore, we used RNA-seq RNA-seq data from the Geuvadis project (short reads) and PacBio data (long reads) {Tilgner:2014hw} to detect chimeric expression involving CBX3 (retroCNVs) and CCDC32 (host). We were able to discern between two distinct chimeric transcripts: i) a transcript contains a small intron retention creating a premature stop codon, truncating the CCDC domain (Figure 4A); and ii) a transcript not containing any intronic sequence but using a de novo canonical splicing site present in the 5’UTR retrocopy. Therefore, a small segment of the 5’UTR and two coding exons from CBX3 may be used in the chimeric transcript. Despite the fact the protein coding sequence from CBX3 is used as protein coding, the reading frame is different, creating a CDC domain with a new C-Terminal domain (Figure 4A). Interestingly, this novel transcript contains a stop codons truncating the protein CCDC32, however, since the retroCNV does no contain introns, it can not be recognized by the Non-sense Mediated Decay (NMD) machinery and could be translated to proteins (Figure 4A; Figure S10).

Additionally to the origin of chimeric mRNAs, intragenic insertion of retroCNVs may also cause indirect impact on the human transcriptomic landscape. For example, we found the retroCNVs-PKRA. This allele is present on 220 individuals across all populations, with higher allele frequency on EAS and AMR (Figure 2). The insertion consists of a full mRNA from PRKRA into the 3’UTR regions from FAM136A. This insertion could create a new regulatory context for the host genes by adding novel miRNA targets {Baertsch:2008jj}. RetroCNVs inserted into transcriptional active context may also harbor for the host gene or into predicted promoters (retroCNV-NACC1, retroCNV-C14orf93) and enhancers (retroCNV-GYG1, retroCNV-RPL10, retroCNV-PSMD8) creating the opportunity of expression per-se or disruption of proximal genes.

Finally, we investigated the influence of retroCNVs alleles on parental gene expression level. Using genotyping results and RNA-seq data from Geuvadis project we are able to compare the expression parental gene expression level for individuals genotyped as homozygous retroCNV presence, heterozygous and homozygous retroCNV absence. For most of the retroCNVs (87%) we could not evaluate the retroCNV impact due to the small overlap between individuals with the retroCNVs allele and expression data. However, for the remaining 14 retroCNVs we could and found that most of the events do not influence parental expression (CBX3 included). We found two retroCNVs that significantly changed parental expression. Both are solo-3’UTR and quite frequent on human populations. The retroCNV-BAZ2A and retroCNV-PCMTD1 alleles are present on 89.66% and 94.16% of the analyzed genomes respectively. We found that these retroCNVs have opposite impact on the parental gene expression. The retroCNV-PCMTD1 comprises a long fragment (approximately 1300 nucleotides) inserted into an intergenic region. Using ENCODE transcription factor data we found a putative highly occupied by transcription factors which may create an opportunity to express the 3’UTR fragment. BAZ2A is a small fragment (approximately 400 nucleotides) inserted into the intron of FSIP1 gene. While the presence of retroCNV-BAZ2A allele significantly decreases the expression of the BAZ2A gene (Figure 4B), the presence of retroCNV-PCMTD1 allele significantly increases the expression of the parental gene (Figure 4C). It is still not clear how expression of 3’UTR fragments may influence the transcriptional landscape.

## Discussion

In the current study we have used an efficient computational method to expand the set of retroCNVs and analyze their influence on the human transcriptome. Our results, based on 26 human populations, supports that retroCNVs creates more genetic variability than previously suggested {Schrider:2013di, Ewing:2013cs, Abyzov:2013hk}. Furthermore, retroCNVs may directly and indirectly contribute to increase the human transcriptome complexity by creating novel and chimeric transcript, as well as influencing the host and parental gene expression. These results suggest that retroCNVs are relevant to the understanding of human evolution and biology.

RetroCNVs are created by the retrotransposition of abundant transcripts in the cytosol [R]. Using the expression level of parental genes in the ovary or testis, we estimated that [N] retroCNVs have higher chance of being originated at a woman (mother) genome than on a man (father genome). Furthermore, the likely of the same transcript being retrotransposed to the same genomic position is very small; therefore, retroCNVs have low homoplasy and are excellent markers for population genetics/genomics. Besides evaluating whom the retroCNVs probably originated, we also were able to estimate the rate of retroCNV creation, ~1 event every three thousand births. Compared to other variations, such as SNPs and LINEs, [N] mutations and [N] {Ewing:2010da}, respectively, retroCNVs have a very low rate, however, despite being quantitatively low, it consists of and important mechanism for gene creation.

It is known that the origin of new genes contribute the phenotypic novelty in humans {Zhang:2014gb} and other eukaryotes {Jones:2005ki, Long:2013ja}. It has been estimated that approximately 300 new genes were create in the human lineage since chimpanzee divergence (REF). Many mechanisms to the novel gene (or gene regions) origin have been proposed, including retroposition of coding genes. For example, a retrotransposition event created a novel transcription factors {Frontini:2005eo} that give immunity to the HIV virus {Sayah:2004ca} on Owl monkeys. Since than retrocopies are absente of regulatory regions, their insertion point is essential to determine their functionality: insertions into intragenic or near promoter regions are especially interesting because they increase the probability to a retrocopy be transcribed (REF). Here, we found that retroCNVs are “evenly” distributed along the genome, showing no enrichment or depletion on any genomic context. Therefore, there are some interesting intragenic retroCNVs insertions. For example, we found an exonic insertion (retroCNV-UQCR10) present on [N]% of the studied genomes that is responsible for truncating the protein C11orf194. Additionally, we found a second case of chimeric transcript, a fusion between CBX3 and CCDC32. This transcript seems to encode a truncated version of CCDC32 with and alternative C-terminal comprised of frame shifted CBX3 amino acids. This transcript was also recently reported as associated Gastric cardia adenocarcinoma {Xu:2014ko} and mistakenly annotated recurrent cancer fusion {Chen:2013ba}. Interestingly enough, this may be a powerful mechanism to create exon shuffling, by inserting mature RNA sequences into random genes at the genome [R](Exon shuffling). We also found two expressed retroCNVs that seem to be influencing the expression of their parental genes. For the first case (BAZ2A), a gene recently associated with the progress of prostate cancer {Gu:2014ku}, we found that the presence of its retroCNV is associated with a parental gene expression decreasing. RetroCNV and parental gene are in different chromosomes, therefore, the influence is not *cis*-related [R]. The difference could be explained by different allele frequency on different populations, however, event when considering populations independently, we found a significant association. Being an intragenic insertion context, it is possible that the retroCNV coopt the promoter from the host gene and get expressed generating small RNA that binds to the parental sequence and degrade the parental mRNA, decrease the parental mRNA stability or event impede protein synthesis. In contrast, the second case, the presence of the retroCNV-PCMDT1 allele increases the expression of its parental gene. Mechanistically, the retroCNV mRNA could function as a competitive RNAs decoying small RNAs that target the parental sequence {Poliseno:2010he}. Despite these interesting examples, it is expected that insertion into intragenic, especially into coding regions, should be wiped from the genome. In contrast, retroCNVs are very recent genomic events, which may not have had enough time to be influenced by natural selection.

As stated before, here we developed and new method using Amazon Cloud computing to systematically search for retrocopies in the all individuals from the 1000 Genomes projects by analyzing paired-end read mapping. Specifically, we evaluated all 2535 individual genomes at the same time; therefore enhancing the number of reads supporting each retroCNV insertion point. This strategy is especially well suited for 1000 Genomes Project data because individually genome are sequenced at low coverage. Therefore, while this strategy performs well for variants at all spectrum of allele frequency, the performance when identifying alleles at very low frequency is lower. We also expect an under representation of retroCNVs presenting less frequent features, for example, retrotransposition of mRNAs with intron retention and transductions. Non-coding genes, and insertions on repetitive loci may are totally absent on our results. Both caveats may be overcome with new sequencing technologies that enhance either sequence coverage on a population scale or read length.

Previous studies on retroCNVs have also used 1000 Genomes data to detect these events, however, with results greatly diverged. Here, we compared the phase3 results to previous studies (using less individual genomes (REFs)), and found that we unified the calls for human retroCNVs insertion points missing only retroCNVs with few evidences supporting the insertion point. For example, for most of the putative retroCNVs previously identified using only exon-exon junction, we have not found evidence to their insertion point. Whether these events are either, truth or artifacts (RNA or cDNA contamination, for example), remains to be clarified since none of these events had their insertion point experimentally validated.

1000 Genomes Consortium sequenced genomes from 26 different populations divided across 5 super-populations (Americans [AMR], Africans [AFR], East Asians [EAS], Europeans [EUR] and South Asians [SAS]). In contrast to previous studies {Schrider:2013di, Ewing:2013cs, Abyzov:2013hk}, we were also able to detect rare variants. Most retroCNVs are super-population specific (Figure 3B) and 13 retroCNVs (11%) are population specific. Most of population specific retroCNVs are present on Asians (5 SAS and 4 EAS) suggesting a recent expansion on the rate of retrocopy generation or a lower rate of retrocopy purging. Considering super populations (Figure 3B), we found that Africans harbor more variants when removing admixture populations [R]. Also, assuming that Native Americans are the most recent population among humans, one would expect that Americans would show the least number of specific retroCNV; indeed, we found that Americans have only three specific retroCNVs (Figure S3). Being an admixture population, we expected to find more shared retroCNVs between Americans and other populations. Interestingly, we found that most of the AMR retroCNVs were shared with AFR, followed EUR in agreement to recent genetic influx from European colonization and African immigration.

Defining the human population structure is the goal of many studies. Here, using PCA and genotype data for each individual independently we were able to separate AFR from super-populations on PC1 and EUR from EAS on PC2. We also used population allele frequency of retroCNVs and performed PCA. Both results greatly overlapped (Figure 3C).  Similar results were obtained using SNPs and microsatellites {Li:2008en, Pemberton:2013ij}. Noteworthy, we only used 109 makers to perform these analyses while SNPs and microsatellites use thousands and hundreds of markers respectively. We are only able to separate populations by the joint power of few retroCNVs, and there is no retroCNVs that independently account for population variations. In other words, we did not found any event fixed in one population and absent in others therefore there is no such thing as truly population specific event.

The variation of retroCNVs across humans poses many technical difficulties when analyzing the human genome and transcriptome. RetroCNVs here mentioned are absent in the human reference genome therefore any analysis based on the reference genome is agnostic to these sequences. Consequently, single nucleotide variants created during reverse transcription may create false SNPs on the parental genes. Even further, retroCNVs events may be mistakenly as recurrent structural variations in cancer or other WGS analyses.

Another limitation is regarding the gene expression analysis, here we used Geuvadis RNA-seq results {Lappalainen:2013el}. Therefore, all expression analyses are restricted to one tissue (Epstein–Barr-virus-transformed lymphoblastic cell line). It is possible that expanding the number of tissue, we would be able to evaluate the expression of additional retroCNV-host gene interactions such as the expression of DCDC1 which is expression preferentially in the testis and pituitary gland. Similarly, our expression analyses are biased towards the number of individuals and populations. Therefore, we were restricted to analyze variants with higher allele frequency and four populations, of which, most are EUR and AFR. To overcome these limitations, it would be necessary to sequence both genome and transcriptome of individuals from different regions and populations. It would also be interesting to evaluate Native American genomes to better understand their origin and population history.

## Methods

### Rationale

[ACC]

Fig1.

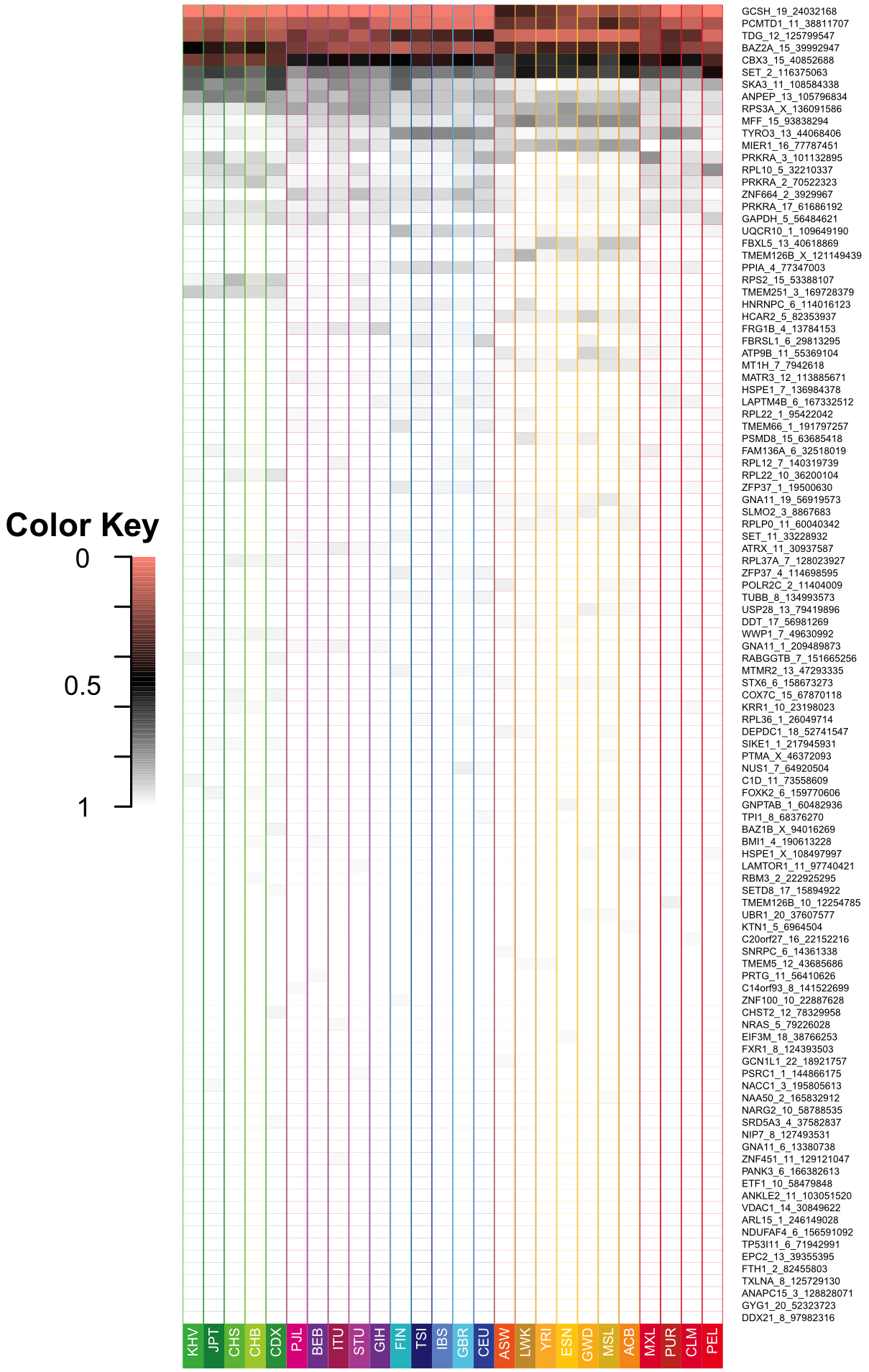


Figure 2. Allele Frequency of retroCNVs across human populations.

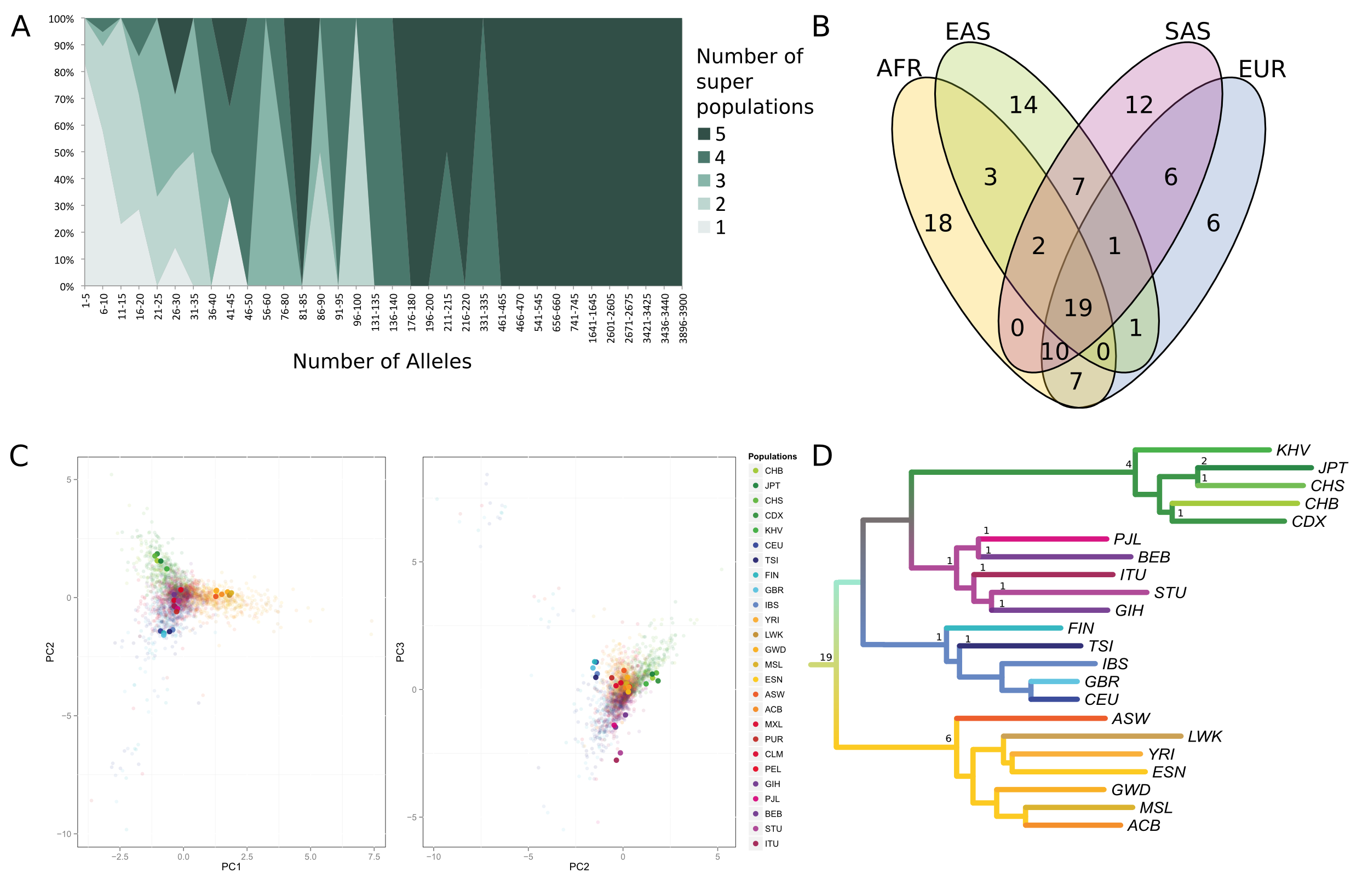


Figure 3. Population genetics analysis.

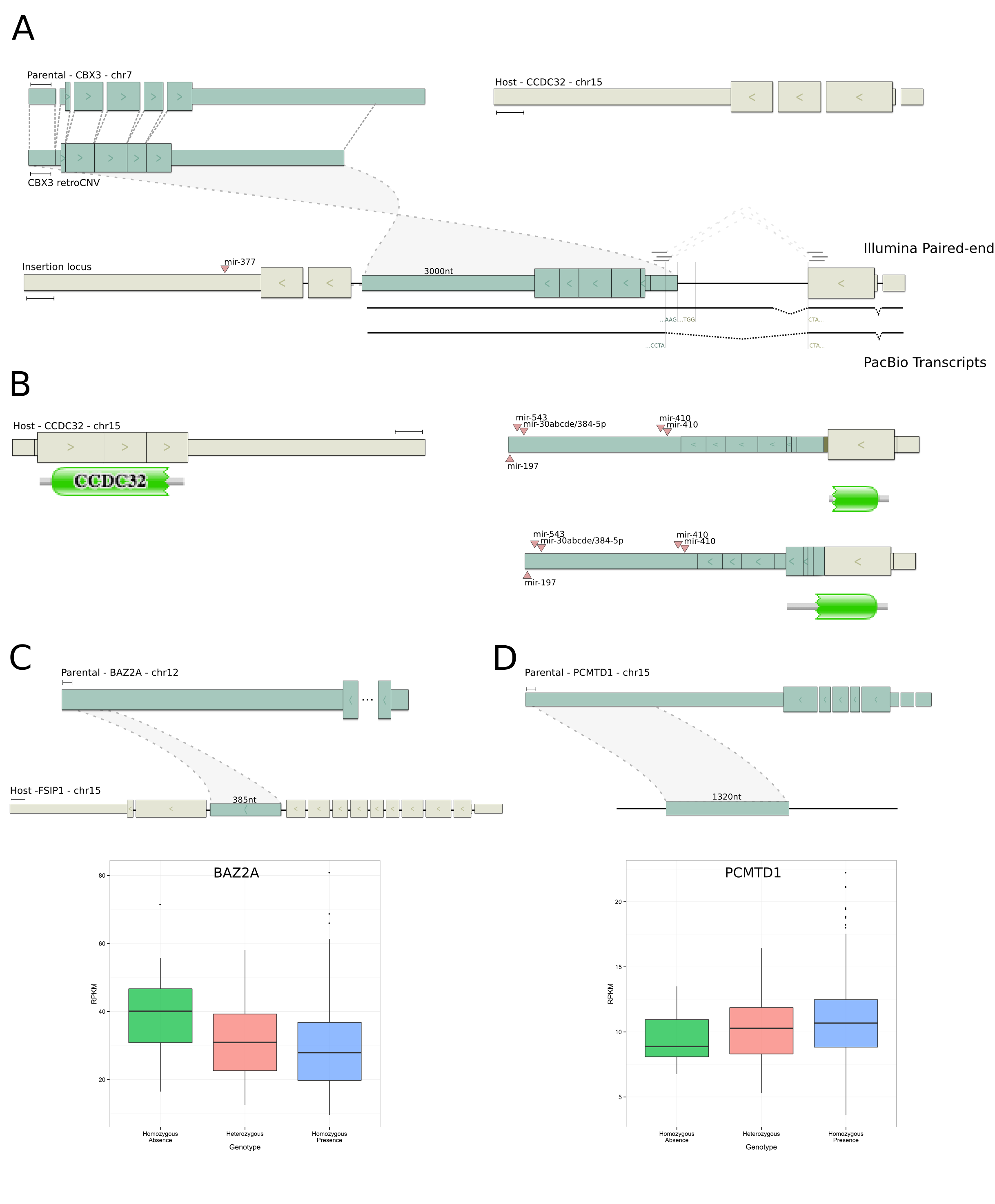


Figure 4. retroCNVs influences the human transcriptomic landscape.