Bioinformatics Consortium of Taiwan

dbDNV: Database for duplicated-gene nucleotide variants

http://bct.binfo.org.tw

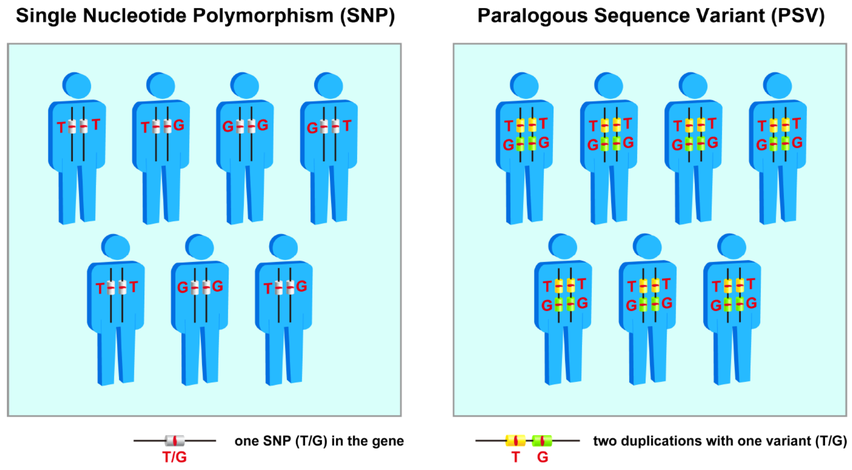
National Core Facility Program for Biotechnology- Bioinformatics Consortium of Taiwan

Abstract

Gene duplication, which is believed to play a major role in evolution, may occur as the consequence of homologous recombination, of retrotransposition events, or of the duplication of an entire chromosome. These duplicates are scattered widely in the human genome. In addition, single nucleotide polymorphisms (SNPs) are annotated by their flanking sequences, which vary in length from tens to hundreds of bases. Nearly identical paralogous sequences with single-base variants can be misjudged as sequences from individuals with SNPs. This causes tighter SNP clusters in duplicons represented in SNP-density analysis.

Focusing on the human reference genome, we systematically interrogated gene-related duplications to obtain gene-related paralogous sequence variants (rPSVs). In this study, we identified 2,965 transcripts associated with multiple duplicated gene loci (DGL). More than 10% of all human genes possessed multiple genomic loci that share a high degree of sequence similarity (> 95%). We identified 1,866,189 rPSVs via the alignment of DGL. We queried dbSNP further using the genomic location of rPSVs. There were 148,897 reference SNPs scattered identically on 291,939 genomic rPSV positions. They were named PSV-like SNPs. About 50% of exonic SNPs in DGL are PSV-like SNPs. This reflects the fact that significant PSV/SNP confounding exists in the functional regions of duplicated genes. The DGL and rPSVs derived from our pipeline were comprehensive and sufficient to reveal the gene-related duplicating structure and their sequence variation in the reference genome. We believe that they can promote the annotation of duplicating SNPs in the human genome and benefit the in-depth characterization of PSV-like SNPs, especially gene-related ones.

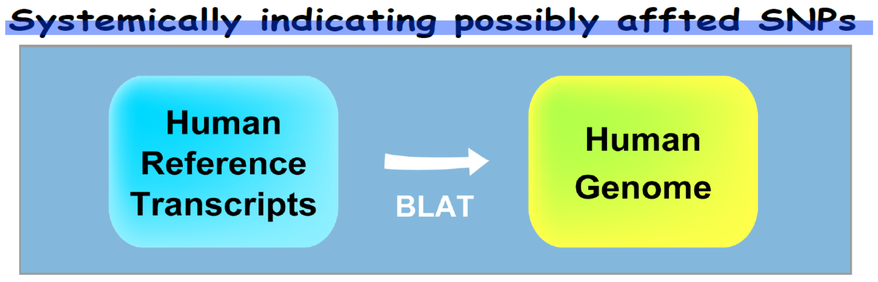
Who are they?



The left panel displays one SNP (T/G) detected in the population. The base type of a specific genomic location varied among individuals; some may be homozygous and some may be heterozygous.

The right panel illustrates the notion of paralogous sequence variants (PSVs). Two copies of a sequence, duplications, exist in the genome, and these two separate copies have a different but invariant base (T/G) at the position of interest. The genotyping of each individual revealed that the PSV data were not distinguishable from the normal SNP heterozygous data.

Systemically indicating possibly affted SNPs



1. Consider > 95% identical aligned results only.
2. Keep results that show that multiple genomic loci associate with the same transcript.

Duplicated Gene Loci (DGL):

Genomic loci that have a high degree of sequence similarity with each other (in the human reference genome) and associate with the same transcript.

BLAT between DGL to obtain variant information.

Reference Paralogous Sequence Variants (rPSVs):

rPSVs that are variant between DGL.

Query dbSNP using the genomic location of rPSVs.

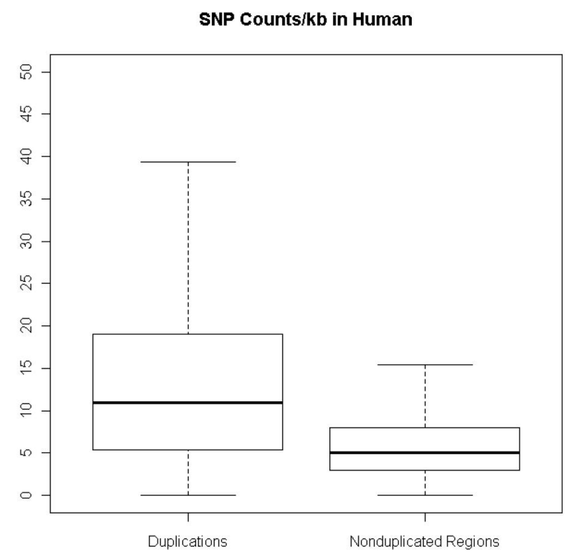
PSV-like SNPs:

Genomic locations of rPSVs are also reported as positions of reference SNPs in dbSNP.

Human gene, FCGR1

FCGR1A, FCGR1B, and FCGR1C are three related gene family members located on chromosome 1. They share a high degree of sequence similarity. The upper figure shows the multiple sequence alignment of these three genes. The marked bases are reported as both variants among three loci and reference SNPs in dbSNP.

The SNP density in duplications



The right figure is the boxplot of SNP counts/kb in two groups (duplications and nonduplicated contigs). The annotated duplications were downloaded from the Human Paralogy Server (http://humanparalogy.gene.cwru.edu). The Y-axis indexes the number of the human reference SNPs/kb in dbSNP.

Results

1. Although the variance was larger in the duplication group compared with the nonduplicated group, the difference in SNP counts in the two groups was significant.
2. More than 10% of human reference genes were associated with duplications that exhibited > 95% sequence similarity.
3. SNP density in dbSNP decreases in conserved regions, such as exons. This may just be a reflection of the fact that functional regions are more conserved under the pressure of selection during evolution and are not prone to maintaining diversity.
4. The density of both exonic and intronic SNPs in DGL was higher than that observed in solitary genes, especially in intronic ones. The increase of SNP density in DGL suggests that SNP records are enlarged by PSVs.
5. Close to 30% of DGL are tagged as exonic regions; however, only 8% of solitary gene loci are tagged as exonic regions. The increase in the exonic proportion from 8% to 30% supports the contention that DGL cover a particular amount of retrogenes.
6. The proportion of PSV-like SNPs in the SNP group was influential, in both exons and introns. Notably, the exonic proportion was close to 50%.

How serious it could be?

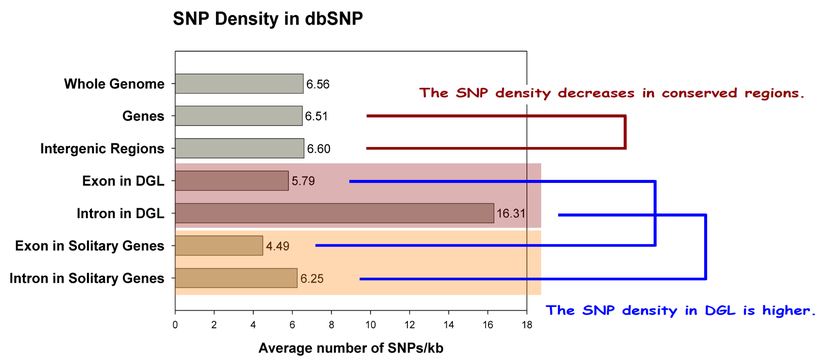
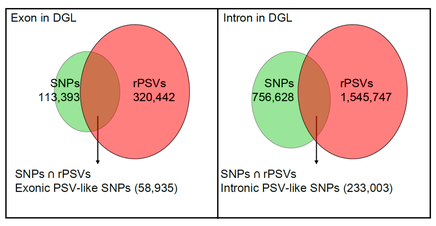


Table 1. Counts of SNPs and rPSVs in Duplicated Gene Loci (DGL)

	reference SNPs	rPSVs[1]	PSV-like SNPs[2]
Exons	113,393 (13%)	320,442 (17%)	58,936 (20%)
Introns	756,628 (87%)	1,545,747 (83%)	233,003 (80%)
DGL	870,021	1,866,189	291,939

[1]rPSV: reference paralogous sequence variant in DGL
[2]PSV-like SNP: intersection of reference SNP and rPSV



Conclusions and Discussion

1. The inference mainly focuses on gene-related loci, with a particular interest in coding regions. Duplicates are no longer considered as merely repeat segments; rather, they are seen as genes undergoing evolutionary processes.
   i. We used transcripts as repeated units, instead of performing alignments among chromosomes.
2. Documentation of the PSV-like SNPs requires short low-copy repeats that are imitative of flanking sequences, to identify possible rPSVs. Our method can be used to identify intronless DNA copies caused by retrotransposition and single-gene duplications caused by segmental duplication.
   i. BLAT is designed to efficiently find DNA sequences with more than 95% similarity. The smallest length of aligned fragments is 25 bases, and sometimes this can be as small as 20 bases.
   ii. BLAT results could include the duplicated loci as long as they possess high conservation of the exonic region.
3. More than 10% of all human genes possessed multiple genomic loci that share a high degree of sequence similarity (> 95%). About 50% of exonic SNPs in DGL are PSV-like SNPs. When the research topic of interest is related to the duplicated regions, such as duplicated genes, the number of PSV-like SNPs would be influential.
   i. We provided a gene-related PSV collection in the human reference genome that can be used by researchers to perform in-depth characterization.
   ii. We strongly suggest that the information regarding DGL and rPSVs should be annotated in the human reference genome in public SNP databases.
4. Alternative splicing (AS) is also a major evolutionary mechanism that can increase functional variation by promoting gene diversification. Once mutations hit the original AS sites, splicing mechanisms might be altered. As variants on AS sites take place within CNVs, individual transcripts may be a new type of genetic variation.
5. In the absence of whole-genome sequencing data for each individual in the population, it will be difficult to determine these ambiguous SNPs.

Abbreviations

PSVs: paralogous sequences variants
DGL: duplicated gene loci in the human reference genome
rPSVs: reference paralogous sequences variants in DGL
SNPs: single nucleotide polymorphisms
MSVs: multisite variants
AS: alternative splicing
CNVs: copy-number variations
BLAT: the BLAST-Like Alignment Tool

Welcome to contact with BCT E-mail: bct@ym.edu.tw Tel: 02-28267359 http://bct.binfo.org.tw

NCFPB 生技類核心設施平台維運計畫 National Core Facility Program for Biotechnology