Pseudogenes: Implications in Disease and Diagnostics

Divya Pasumarthi, Ashwin Dalal

Diagnostics Division, Centre for DNA Fingerprinting and Diagnostics, Hyderabad, Telangana, India

Correspondence to: Dr Ashwin Dalal *Email:* ashwindalal@gmail.com

The genome is the complete set of deoxyribonucleic acid (DNA) in an organism. Human cells contain two copies of the haploid genome consisting of 3×10^9 base pairs of DNA. Only 1-2% of the mammalian genome codes for the protein function and remaining codes for repetitive elements and non-coding regions. Repetitive elements include transposons and Alu elements. The gene is a hereditary unit of DNA. About 25,000 genes are encoded in the DNA, which is organized into structures known as chromosomes. Unique non-coding sequences known as pseudogenes are present in 15% of the human genome.

Organization of the human genome

Human genome consists of nuclear and mitochondrial genome. The nuclear genome consists of protein coding (1%) and non-coding regions. The protein coding part consists of exons and the non-coding part includes pseudogenes, gene fragments, introns, untranslated regions, repetitive DNA etc. (Figure 1).

What are Pseudogenes and why are they important?

Pseudogenes are also known as junk DNA. These are present in the noncoding part of the human genome. Pseudogenes have a lot of resemblance to functional genes but have accumulated mutations which have inactivated them during the course of evolution. Pseudogenes can arise due to tandem duplication of normal genes and accumulation of several mutations in them, which leads to loss of the ability of gene expression. Pseudogenes

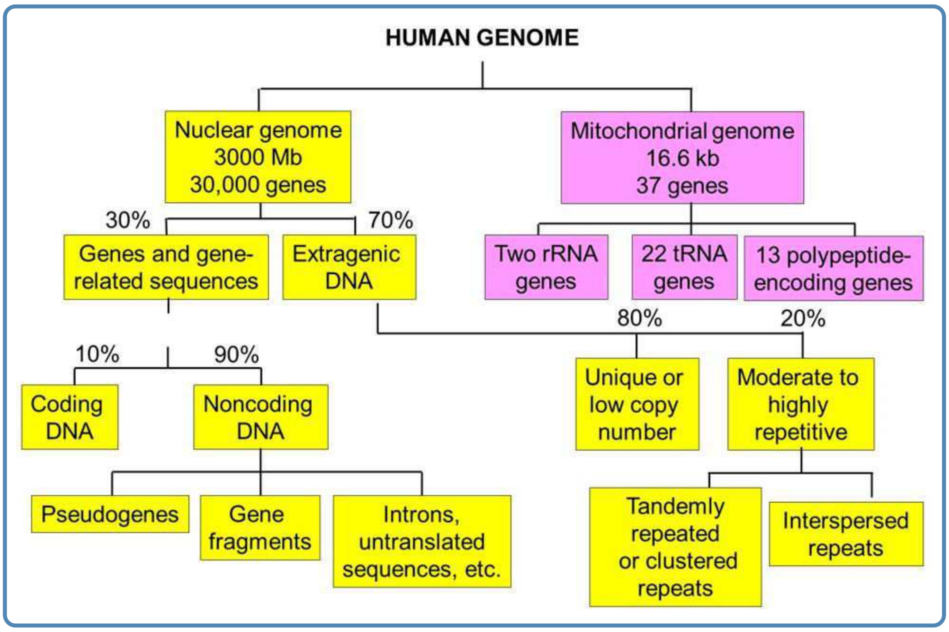
HUMAN GENOME
Nuclear genome 3000 Mb 30,000 genes
Mitochondrial genome 16.6 kb 37 genes
30%
70%
Genes and gene-related sequences
Extragenic DNA
Two rRNA genes
22 tRNA genes
13 polypeptide-encoding genes
80%
20%
10%
90%
Coding DNA
Noncoding DNA
Unique or low copy number
Moderate to highly repetitive
Pseudogenes
Gene fragments
Introns, untranslated sequences, etc.
Tandemly repeated or clustered repeats
Interspersed repeats

Figure 1 Organization of the human genome.

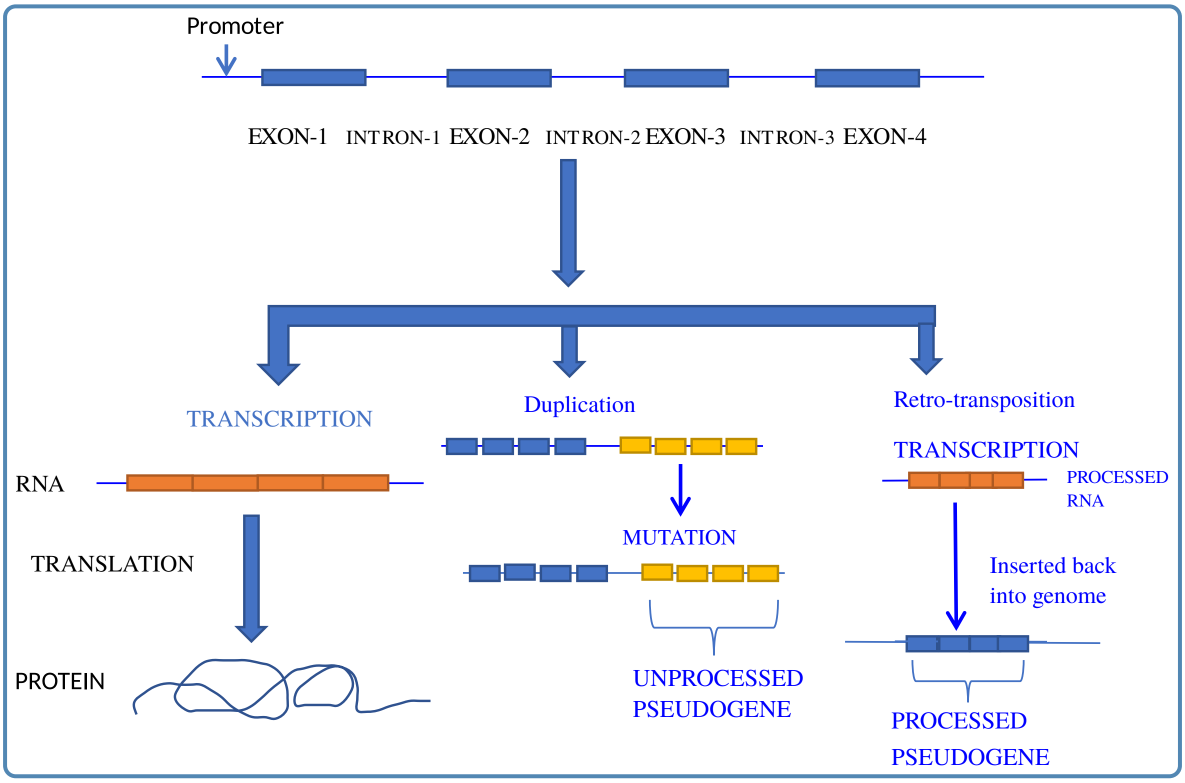

Figure 2 Schematic representation showing how functional protein is formed from gene and how pseudogenes have formed during evolution.

may not have introns and other necessary factors that are required for protein function. Most of the pseudogenes are transcribed into RNA and these transcripts may be processed into small interfering RNA (siRNA) to regulate the coding gene function through RNAi (RNA interference) pathway.

Types of Pseudogenes

There are two different types of pseudogenes:

i. Processed pseudogenes

ii. Unprocessed pseudogenes

- Processed pseudogenes: Processed pseudogenes are also called as retro-transposed pseudogenes. In this process, double stranded DNA is transcribed into mRNA and further re-transcribed into cDNA and integrated into the chromosomal DNA at a new location. Processed pseudogenes have poly-A tail but do not have introns and upstream promoters (Figure 2).

- Unprocessed genes: Non-processed pseudogenes are also known as duplicated pseudogenes and they developed during the evolution of the genome. Duplication of pseudogenes occurs due to homologous recombination between misaligned chromosomes. These pseudogenes have complete intron-exon structure and regulatory sequences. However, these pseudogenes accumulate mutations during evolution and this leads to loss of gene function, transcription and translation (Figure 2).

The human genome has numerous pseudogenes, of which approximately 10,000 to 20,000 are known. Majority of human pseudogenes are processed into mRNA while some pseudogenes are able to regulate tumor suppressor genes and oncogenes by acting as miRNA decoys [Tutar et al., 2018].

Role of pseudogenes in disease

Pseudogenes can be involved in disease causation due to gene conversion (Figure 3) or recombination with functional genes. Gene conversion is an event by which part of DNA is replaced by homologous sequences of another gene. Mutations in the pseudogene are transferred to the functional gene by the gene conversion mechanism and this leads to disease due to perturbation of the functional gene. Genes that are close to pseudogenes

Table 1 List of diseases known to be caused by gene conversion.

S. No.	Gene	Processed /Non processed pseudogene	Disease
1	*IDS*	Non processed	Hunter syndrome
2	*GBA*	Non processed	Gaucher disease
3	*NCF1*	Non processed	Chronic granulomatous disease
4	*PKD1*	Non processed	Autosomal dominant polycystic kidney disease
5	*IGLL1*	Non processed	B-cell deficiency
6	*ABCC6*	Non processed	Pseudoxanthoma elasticum
7	*CYP21A2*	Non processed	Congenital adrenal hyperplasia
8	*FOLR1*	Non processed	Neural tube defects
9	*SBDS*	Non processed	Shwachman-Diamond syndrome
10	*VWF*	Non processed	Type 3 von Willebrand disease
11	*CRYBB2*	Non processed	Congenital adrenal hyperplasia

are candidate genes for gene conversion because greater similarity between gene and pseudogene leads to more chances of recombination (Table 1).

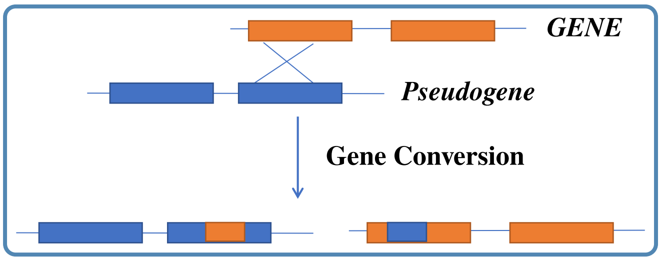

Figure 3 Gene conversion events between gene and pseudogene shown as blue and orange boxes.

Mechanisms of disease

Hunter syndrome (MPS II) is caused by mutations in the *IDS* gene [Timms et al., 1995]. Pseudogene of IDS known as IDS2 or IDSP1 shows 80kb similarity with the functional IDS gene. Pseudogene IDS2 is 96% homologous in exon 2, and intron 2, 3 and 7 of the transcribed *IDS* gene and 100% similar to exon 3 of *IDS* gene. The presence of pseudogene can result in recombination with the *IDS* gene. *IDS-IDS2* undergoes homologous recombination [Lualdi et al., 2005] and thus is involved in the formation of large complex genomic/genetic rearrangements, deletions etc. which comprise about ~13% of the patients with Hunter syndrome. Similar type of gene conversion events are known to occur in patients with Gaucher disease [Tayebi et al., 2003]. Recombinant alleles of *GBA* gene for Gaucher disease arise due to presence of the pseudogene GBAP 16 kb downstream of the functional gene. This pseudogene has 96% exonic sequence homology with *GBA*, with the region between intron 8 and 3′ UTR being >98% homologous. This promotes gene conversion and gene fusion events by non-reciprocal and reciprocal recombination. These events are most commonly seen in the exon 9–11 region. At least 10 GBA recombinants have been reported which comprise as many as 20–30% of alleles in some populations [Koprivica et al., 2000].

Problems in Molecular Diagnostics

Case Study: Sanger sequencing of the *IDS* gene in a male patient with Hunter syndrome (detected by deficient enzyme assay) revealed a heterozygous variant p.Ala85Thr (alanine to threonine at amino acid 85) caused by a G to A substitution at nucleotide position c.253 in exon 3 of *IDS* (Figure 4). Since a male has only one X chromosome and the *IDS* gene is present on the X chromosome, we expect the patient to be hemizygous for the variant with a single peak on Sanger sequencing. The heterozygous peak in this chromatogram was seen because the mutant allele came from gene and the normal allele came from the pseudogene. Thus, it is important to design primers for polymerase chain reaction (PCR) in such a way that only the gene is amplified and the pseudogene is not amplified for sequencing of genes where pseudogenes are known to be present in the genome. Similar problem is faced in interpretation of results in next generation sequencing assays like exome sequencing and hence this should be kept in mind while interpreting the results.

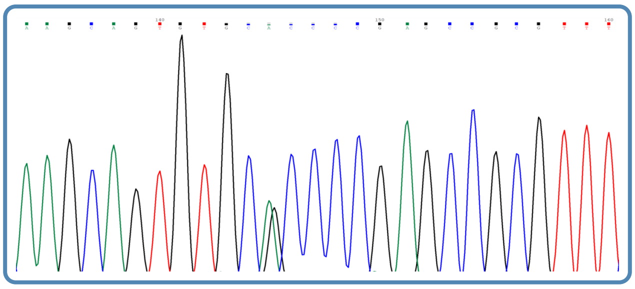

Figure 4 Electropherogram showing heterozygous peak in the c.253G>A position, due to the simultaneous amplification of gene and pseudogene.

Diagnostic assays for genes with pseudogenes

In molecular diagnostics for genetic diseases, it is important that the sequencing information should come from the gene and not from the pseudogene. Hence, modifications need to be done to identify a gene. This can be done in two ways:

1. If a pseudogene is highly homologous to the functional gene and there are no differences in intronic regions, then the strategy employed is Long Range PCR followed by nested PCR. In this method a long-range PCR assay is designed using primers designed in such a way that the 3' end of the primer falls on a single nucleotide change between gene and pseudogene found nearest to the gene. This is followed by a long PCR using special DNA Polymerase and then the PCR product is used as a template for subsequent PCR and sequencing using flanking primers for each exon. The long PCR ensures that only the gene is sequenced for interpretation of results (Figure 5).

2. If there are multiple differences in intronic regions of the gene and the pseudogene then flanking primers for each exon can be designed in such a way that the 3' end of primer falls on a single nucleotide change. Thus, there is no need of long PCR in such cases (Figure 5).

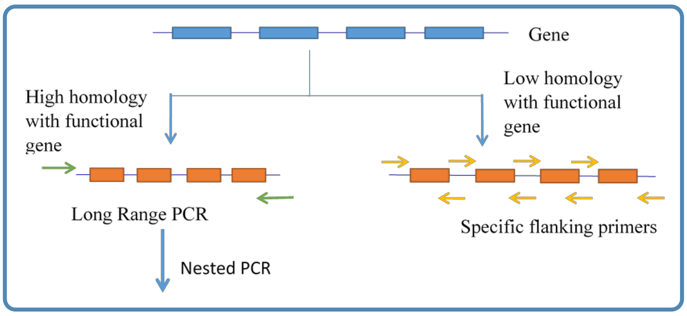
Gene
High homology with functional gene
Low homology with functional gene
Long Range PCR
Specific flanking primers
Nested PCR

Figure 5 Diagnostic strategies for genes with pseudogenes.

Conclusion

Pseudogenes are essential parts of gene regulation. Understanding the mechanism of pseudogene actions is likely to help researchers to solve several essential biochemical pathways. Pseudogenes might be functional and participate in gene expression and molecular mechanisms of gene interactions. Pseudogenes can lead to genetic diseases due to gene conversion and they also pose a problem in genetic diagnostics.

References

1. Koprivica V, et al. Analysis and classification of 304 mutant alleles in patients with type 1 and type 3 Gaucher disease Am J Hum Genet 2000; 66:1777-1786.
2. Lualdi S, et al. Characterization of iduronate-2-sulfatase gene–pseudogene recombinations in eight patients with Mucopolysaccharidosis type II revealed by a rapid PCR-based method. Hum Mutat 2005; 25: 491-497.
3. Tayebi N, et al. Reciprocal and nonreciprocal recombination at the gluco cerebrosidase gene region: implications for complexity in Gaucher disease. Am J Hum Genet 2003; 72: 519-534.
4. Timms KM, et al. 130 kb of DNA sequence reveals two new genes and a regional duplication distal to the human iduronate-2-sulfate sulfatase locus. Genome Res 1995; 5: 71-78.
5. Tutar L, et al. Involvement of miRNAs and pseudogenes in cancer. Methods Mol Biol 2018; 1699: 45-66.