Contribution of pseudogenes to sequence diversity.

Vihinen, Mauno

Published in:
Pseudogenes. Functions and Protocols.

DOI:
10.1007/978-1-4939-0835-6_2

2014

CONTRIBUTION OF PSEUDOGENES TO SEQUENCE DIVERSITY

Mauno Vihinen

Department of Experimental Medical Science, Lund University, BMC D10, SE-22184 Lund, Sweden

Keywords: pseudogene, genetic diversity, immunoglobulin, B cell receptor, T cell receptor, gene conversion, immunome, antigen variation

ABSTRACT

Pseudogenes are very common in the genomes of a wide range of organisms and, although they were originally considered as genetic junk, now several functions have been attributed to them. One important function of pseudogenes, as discussed in this chapter, is to provide material for genetic diversity. This is most prominent in the case of immunological recognition molecules such as immunoglobulins and B- and T-cell receptors, as well as in the case of antigenic variation in intracellular pathogens. Other examples discussed are olfactory receptors, ribosomal proteins, cytochrome P450s, and pseudokinases.

1. INTRODUCTION

Pseudogenes are defined as nonfunctional copies of genes. They have been identified in all types of organisms ranging from bacteria to mammals. In particular, pseudogene.org database lists 16,735 pseudogenes in the human genome [1], while GENCODE lists 13,447 pseudogenes (http://www.gencodegenes.org/stats.html).

Pseudogenes originate via two major processes. Duplicated pseudogenes are created via tandem duplication or uneven crossing over. Processed pseudogenes arise by retrotransposition, when an mRNA sequence or part of it is reverse transcribed to DNA and inserted into the genome. About 10 % of human genes have processed pseudogenes [2]. A third form of pseudogenization exists, although it is extremely rare. It involves losing gene function without having a functional version left in the genome. Only very few cases of these disabled unitary pseudogenes have been reported, whereas the other two forms are quite common.

The function and importance of pseudogenes have long been enigmatic. They were historically considered as useless junk. However, several functions have been recently linked to them. Quite many pseudogenes are expressed, even in a tissue-specific manner. Furthermore, although their coding sequence is nonfunctional, still many pseudogenes do not accumulate variations at random rate, instead they display evolutionary selection. The functions of pseudogenes include RNAi and competition for microRNA binding with other endogenous RNAs, as discussed in detail in refs. 3 – 5. Additionally, pseudogenes can represent a reservoir of sequences that contribute to increase genetic diversity.

2. GENETIC DIVERSITY

Genetic diversity means genetic variability within species and it is essential for the evolution of all species. Since the genomes of higher organisms are very scarce in genes (for example, in the human genome the protein coding sequences comprise less than 2 % of the genome length), the use of genome parts outside gene coding regions is an effective and economical way to increase the diversity of genes.

Pseudogenes contribute to increase the genetic diversity of a plethora of genes and they do so through recombination and gene conversion (Fig. 1).

Antibodies, B- and T-cell receptors, and Major Histocompatibility Complex (MHC) type I and II are a prime example of molecules that increase their sequence diversity using pseudogenic sequences. Another example of the use of pseudogenes to increase significantly genetic variation is the production of antigen variation in surface molecules by intracellular pathogens, as discussed in details in the following chapters.

Additional examples involve ribosomal proteins, olfactory receptors, and cytochrome P450 enzymes. Ribosomal proteins (RPs) have the largest number of pseudogenes: in the human

genome there are 1,822 processed ribosomal protein pseudogenes, while the number of ribosomal protein genes is just 80 [6]. It is not yet sure how much pseudogenes contribute towards the genetic diversity of RPs. However, RNA-seq analyses have indicated that several RP pseudogenes are highly expressed and in a tissue-specific manner [7], while ribosomal proteins are typically expressed in an ubiquitous manner, suggesting that the pseudogenes may contribute to tissue- specific features of RPs.

Olfactory receptors (ORs) vary largely in numbers between organisms [8]. In mammals, 4–5 % of the genes belong to this group and in *H. sapiens* OR genes form one of the largest gene families (415 OR pseudogenes (51.7 %) in addition to 387 OR genes) [8]. Many different OR proteins are needed to detect different odors. Each OR is encoded by a single gene and gene conversion events allow to produce new genes with new specificities. The high number of OR pseudogenes present in mammals has been explained with the development of the color vision: due to better vision, the need for odor sensing has dropped and thus many OR genes have become pseudogenized. The OR gene to pseudogene ratio has in fact been suggested to refl ect the essentiality of olfaction for the organism [9]. For example, the ratio is small in humans, while it is high in rodents that rely on their olfaction for finding their food. On the other hand, since the olfaction is based on the combination of several different receptors, the higher is the number of OR genes and the higher is the sequence exchange among them, the higher becomes the number of smells that can be detected. Therefore, it seems economical to use pseudogenic sequences in order to modify the sequence of OR genes and hence widen the spectrum of odorants that can be smelled. OR pseudogenes contribute to sequence diversity also because their sequence can vary among individuals and possibly show intact reading frames [10 , 11].

A third gene family with a large proportion of pseudogenes is that of cytochrome P450 enzymes. These enzymes have multiple functions in the synthesis, metabolism and catabolism of biological molecules, as well as of drugs and foreign chemicals. Humans have 57 genes and 58 pseudogenes, while mouse has 102 genes and 88 pseudogenes [12] and plants have substantially more, e.g., rice (*Oryza sativa*) has 356 genes and 99 pseudogenes [13]. The cytochrome P450 pseudogenes are thought to increase diversity by providing genetic material for gene rearrangements, also considering the fact that their sequence is subjected to a higher mutation rate than that of the corresponding protein coding genes.

3. PSEUDOGENES IN THE IMMUNOME

Genome-wide studies of pseudogenes are available for a number of genomes and protein families in Pseudogene.org and Pseudofam (http://pseudofam.pseudogene.org) [14]. Functional pseudogenes have been investigated in three species, human, mouse, and chimpanzee [15]. However, there are no system-wide studies, with the exception of the human immunome [16].

For the human immunome, we collected genes and proteins that are essential for human immunity [16]. The immunology related genes and the corresponding proteins were systematically collected from articles, textbooks, and databases. All these genes and proteins are directly involved in immunological processes.

Genes were included when essential for immunity. Immunodeficiency-related genes were from the ImmunoDeficiency Resource (IDR) [17] and IDbases [18]. Proteins with wide expression pattern outside immunological cell types were excluded. Only full- length genes were included. Signaling molecules were included only when involved in immunity-related cascades.

The analysis revealed 847 essential immunome genes and proteins [16] as reported in the Immunome Knowledge Base (IKB, http://bioinf.uta.fi /IKB) [19]. A pipeline was developed for the identification of pseudogenes of immunome genes. Such a pipeline was subsequently developed into PseudoGeneQuest service for the analysis and search of human pseudogenes [20]. PseudoGeneQuest is freely available at http://bioinf.uta.fi/PseudoGeneQuest .

4,816 pseudogenes or fragments were found for 313 human immunome genes [16]. Pseudogene fragments are predominant: 3,736 fragments for 229 genes. One hundred and thirty genes have duplicated pseudogenes, while the rest of the genes have processed pseudogenes. The expression of the pseudogenes varies substantially, as indicated by EST data.

4. DIVERSIFICATION OF IMMUNOLOGICAL RECOGNITION MOLECULES USING PSEUDOGENES

Recognition of the enormous range of nonself-substances is the basis of adaptive immunity and is achieved by mechanisms that produce heterogeneous receptors, namely antibodies, B- and T-cell receptors (BCRs and TCRs), and the components of the MHC.

Antibodies (immunoglobulins, Igs) are free serum proteins or form part of the BCRs. Igs are produced by B cells [21] and their main function is to recognize foreign substances and facilitate their destruction.

The number of different Ig molecules produced by the body is very big (10^{11}). It would be impossible to have that many genes in the genome, especially as the human genome is only 3×10^9 bases long. Such a high number of different molecules is in fact produced through a process called V(D)J recombination.

Antibodies consist of two light and two heavy chains. Both the heavy and light chains contain constant and variable regions, which form surface loops that recognize foreign molecules. For each part of the antibody gene, the genome contains a number (up to 100) of different segments originated by segmental duplications during evolution [22]. During V(D)J recombination, also called somatic recombination, one diversity region (D) segment is combined first with one joining (J) segment and then with one variable region (V) segment. Then, one of the constant regions (C) segments, which determine the class of the antibody, is added to produce the complete V(D)JC gene (Fig. 2a). A big number of enzymes is required for V(D)J recombination. For example, V, D, J, and C segments are flanked by recombination signal sequences (RSSs) or switch regions, which are recognized by Recombination activating Genes (RAG) 1 and 2 that produce single-strand cleavage.

Each B cell produces a unique form of Ig, which is eliminated if it does not express a functional receptor. Furthermore, the cells do not survive if there is an inappropriate heavy chain gene rearrangement (positive selection). Analogously, immature B cells enter apoptosis if their Ig (IgM) recognizes self-antigens (negative selection). In this way the immune system finds a balance between two needs: on one hand foreign antigens should elicit a specific immune response, but on the other hand self-reacting antibodies have to be eliminated to avoid adverse autoimmune reactions.

V(D)J recombination alone accounts for a large degree of genetic variation. However, three additional processes—somatic hypermutation (SHM), gene conversion with pseudogenic sequences, and class switch conversion or recombination (CSR) of the constant region—contribute to increase the repertoire even further (Fig. 2b–d). Activation-induced deaminase (AID) is essential for all the three processes.

SHM randomly adds point variations via the process of error- prone DNA repair (Fig. 2b). In SHM, AID deaminates cytosine to uracil and error prone DNA polymerases complete the mutagenesis by introducing substitutions either to the modified site or next to it.

V gene conversion introduces changes that can be extensive to the V segments and utilizes pseudogenes as sequence reservoir (Fig. 2c). The details of the mechanism are still elusive, however it is known that high sequence similarity between the gene and the pseudogene is a fundamental requisite and that the role played by AID is crucial [23].

In *H. sapiens* there are about 130 V gene segments on chromosome 14, only 38–46 of which are functional depending on the haplotype, while the rest are pseudogenes [24]. Many of the pseudogenes could be corrected into functional genes by no more than three substitutions. Interestingly, such a big number of V genes and pseudogenes is quite rare. For example, domestic animals have a much smaller repertoire [25]. Still, the pseudogene-utilizing gene conversion is the major factor for diversification in chicken. On the contrary, in other animals a similar output is obtained by alternative mechanisms. For example, *D. melanogaster* uses alternative splicing, which leads to over 30,000 gene variants [26].

Unlike the other Ig maturation processes that increase the sequence diversity of the variable region, class switch recombination alters the constant region of the heavy chain, so that antibodies of different classes (isotypes) are produced (Fig. 2d). CSR appears only after BCR activation and is regulated by cytokines released by helper T cells: activated BCRs engage naïve lymphocytes, which in turn secrete cytokines that induce the class switching.

The heavy chain exons are tandemly repeated and separated by switch regions. During CSR, some of the exons of the heavy chain locus are excised through double-strand breaks generated at the switch regions, while the remaining exons are ligated together by nonhomologous end-joining. The different Ig classes are therefore produced by the deletion of different exons (Fig. 2d).

Combinatorial effects of SHM, gene conversion with pseudogenic sequences and CSR produce the vast diversity in the immunological molecules and allow to generate recognition molecules for all possible foreign organisms and substances starting from a limited number of gene segments.

5. ANTIGEN VARIATION USING PSEUDOGENES

Numerous intracellular pathogens, both prokaryotic and eukaryotic, have developed an effective strategy to evade the immune response by varying the antigenic region of the proteins expressed on their cell surface. This is achieved by recombination of pseudogenes with a small number (often just one) of functional genes (*see* ref. 27 for a review).

Prokaryotes using this mechanism include *Borrelia burgdorferi* (major membrane protein, VlsE), *Neisseria gonorrhea* , and *Neisseria meningitidiis* (pilin proteins), *Anaplasma marginale* (major surface protein, msp2). Examples of eukaryotic pathogens include *Pneumocystis carinii* (major surface glycoprotein, msg), and *Trypanosoma brucei* (variable surface glycoprotein, vsg). These organisms cause a wide variety of diseases: *A. marginale* is a tick borne livestock pathogen causing significant mortality and morbidity. *B. burgdorferi* is also tick borne and causes human Lyme disease or borreliosis, while *N. gonorrhea* causes gonorrhea and *N. meningitidis* is behind meningitis and other meningococcal diseases. *P. carinii*, a fungus, is a human-specific pathogen that causes opportunistic infections and pneumonia. Protozoan *T. brucei* infections lead to African trypanosomiasis in human, and nagana, a related disease in animals (mainly in cattle and horses). These organisms are thus widely different and responsible for different diseases. Common to all of them is to hide from the host immune system by “camouflaging” or concealing the pathogens by expressing variable antigens on their surface.

One of the best-studied examples is the pilin antigen variation in *Neisseria*. Pilins are fibrous proteins in hair-like pili on the surface of many bacteria. Pili are common structures,

especially on Gram-positive bacteria. They are involved in the exchange of genetic material during bacterial conjugation and they are used to adhere to the surface of host cells.

The principle of antigen variation is also utilized in the phage display technology to generate peptide libraries to study protein interactions. In this technology, pili of viruses are modifi ed to carry proteins of interest and generate variations for screening [28].

Both the natural and the artificial variations of the pili exploit their structure and localization, since modifications and insertions are easily accommodated in these surface proteins.

The bacterial pathogens discussed above have just one active gene and typically tens of pseudogenes. In the case of the eukaryotic organisms the number of functional genes is higher, up to 150 in *Babesia bovis*, which has 24 pseudogenes. The largest number of pseudogenes appears on *T. brucei* (up to 2,000) [29]. GC base pairs have been shown to be crucial for the gene-pseudogene recombination, by forming special 16 base pair guanine-rich four stranded DNA structures. These structures, which have a special topology that is much wider than the normal B-DNA double helix, can act as recombination initiation sites [30].

The intracellular pathogens have also another interesting feature related to pseudogenes. Due to the environmental change from free living organisms to intracellular parasites, these species do not need all their genes anymore, as the host cells already have machineries for many essential functions such as metabolism, energy production, and catabolism. Thus, reductive evolution has led to the pseudogenization of large parts of their genome. For example, only about half of the genes of *Mycobacterium leprae*, which causes leprosy, are still active [31].

By modifying the proteins present on the surface of their membrane, the pathogens evade from the immune system and can remain persistently in the infected hosts. Viruses, especially lentiviruses, utilize the same strategy, except that their genome does not contain the variable

genetic material and the sequence stretches for antigen variation are taken from the genome of the host organism. HIV has the greatest antigenic variation among viruses, comparable only to hepatitis C virus, which however shows a much slower variation rate [32].

6. PSEUDOKINASES

Protein kinases form one of the largest protein families. They are involved in phosphorylation of target molecules and therefore are among the most crucial cellular signaling molecules. *H. sapiens* has over 500 protein kinases, out of which 106 were originally suggested to be inactive kinases (pseudokinases) due to modifications in crucial conserved sites [33]. Zeqiraj et al. [34] extensively discuss about ways to inactivate kinases. They also group pseudokinases into four classes. Some pseudokinases have been shown to be active: low activity pseudokinases have some residual activity, although it is not always clear if that is sufficient for biological function. Furthermore, examples of truly active pseudokinases, an oxymoron for a name, do exist. Despite variations at crucial sites, they can have high activity because they have acquired compensatory structural changes.

The function of pseudokinases is usually related to regulation *(see* ref. 35 for a detailed discussion). An interesting case is the JAK family of kinases (JAK1, JAK2, JAK3, and TYK2), which contain a functional kinase domain and a pseudokinase domain within the same protein chain. The catalytically active kinase domain (JH1) is in charge of the phosphorylation of downstream targets, while the catalytically inactive pseudokinase domain (JH2) is a negative regulator of the kinase domain.

Pseudofam lists 68 human pseudogenes for 53 parent genes in the protein kinase (Pkinase PF00069) family. The mouse kinome contains 97 pseudogenes [36]. Some of the kinases in *H. sapiens* appear as pseudogenes in the mouse and vice versa.

Very little is known about the function of the kinase pseudogenes in any organism, although some are known to be expressed and could be functionally active (*see*, e.g., ref. 37). The action of kinases is reversed by phosphatases, which are abundant in number as well (around 200 in the human genome). There are about 12 tyrosine phosphatase pseudogenes and some of them are expressed [38]. Since they could impact on crucial signal transduction pathways, the study of the catalytic activity, if any, of such pseudogenic kinases and phosphatases should be undertaken systematically.

REFERENCES

Commented [NC1]: Please insert the references in the text using EndNotes (SpringerBasicNumber format)

1. Karro, J.E., Yan, Y., Zheng, D. et al. (2007) Pseudogene.org: a comprehensive database and comparison platform for pseudogene annotation. *Nucleic Acids Res* **35**, D55-60

2. Zhang, Z., Harrison, P. M., Liu, Y. et al. (2003) Millions of years of evolution preserved: A comprehensive catalog of the processed pseudogenes in human genome. *Genome Res.* **13**, 2541-2558

3. Balakirev, E., Ayala, F.J. (2003) Pseudogenes: Are they ”junk” or functional DNA? *Annu Rev Genet* **37**, 123-151

4. Pink, R.C., Wicks, K., Caley, D. P. et al. (2011) Pseudogenes: Pseudo-functional or key regulators in health and disease. *RNA* **17**, 792-798

5. Wen, Y.-Z, Zheng, L.-L., Hu, L.-H. et al. (2012) Pseudogenes are not pseudo any more. *RNA Biol.* **9**, 27-32

6. Balasubramaniam, S., Zheng, D., Liu, Y.-J. et al. (2009) Comparative analysis of processed ribosomal pseudogenes for four mammalian genomes. *Genome Biol* **10:R2**

7. Tonner, P., Srinivasasainagendra, V., Zhang, S. et al. (2012) Detecting transcription of ribosomal protein pseudogenes in diverse human tissues from RNA-seq data. *BMC Genomics* **13:**412

8. Niimura, Y. (2009) Evolutionary dynamics of olfactory receptor genes in chordates: Interaction between environments and genomic contenst. *Hum Genomics* **4**, 107-118

9. Olender, T., Lancet, D., Nebert, D.W. (2008) Update on the olfactory receptor (*OR*) gene superfamily. *Hum Genomics* **3**, 87-97

10. Glusman, G., Sosinksy, A., Ben-Asher, E. et al. (2000) Sequence, structure, and evolution of a complete human olfactory receptor gene cluester. *Genomics* **63**, 227-245

11. Mombaerts, P. (2001) The human repertoire of odorant receptor genes and pseudogenes. *Annu Rev Genomics Hum Genet* **2**, 493-510

12. Nelson, D.R., Zeldin, D.C., Hoffman, S.M.G. et al. (2004) Comparison of cytochrome P450 (*CYP*) genes from the mouse and human genomes, including nomenclature recommendations for genes, pseudogenes and alternative-splice variants. *Pharmacogenet* **14**, 1-18

13. Nelson, D.R., Schuler, M.A., Paquette, S.M. (2004) Comparative genomics of rice and Arabidopsis. Analysis of 727 cytochrome P450 genes and pseudogenes from monocot and dicot[1[w]]. *Plant Physiol* **135**, 756-772.

14. Lam, H.Y., Khurana, E., Fang, G. et al. (2009) Pseudofam: the pseudogene families database. *Nucleic Acids Res* **37**, D738-743

15. Svensson, Ö., Arvestad, L., Lagergren, J. (2006) Genome-wide survey for biologically functional pseudogenes. *PloS Comput Biol* **2(5):e46.**

16. Ortutay, C., Siermala, M., Vihinen, M. (2007) Molecular characterization of the immune system: emergence of proteins, processes and domains. *Immunogenet* **59**, 333-348.

17. Samarghitean, C., Väliaho, J., Vihinen, M. (2007) IDR knowledgebase for primary immunodeficiencies. *Immunome Res.* **3**:6.

18. Piirilä, H., Väliaho, J., Vihinen, M. (2006) Immunodeficiency mutation databases (IDbases). *Hum Mutat* **27**, 1200-1208

19. Ortutay, C. and Vihinen, M. Immunome Knowledge Base (IKB): An integrated service for immunome research. (2009) *BMC Immunol.* **10**:3.

20. Ortutay, C., Vihinen, M. (2008) PseudoGeneQuest – Identification of different pseudogene types in the human genome. *BMC Bioinf* **9**, 299

21. Ollila, J., Vihinen, M. (2005) B cells. I*nt J Biochem Cell Biol* **37**, 518-523

22. Pramanik S, Cui X, Wang H-Y et al (2011) Segmental duplication as one of the driving forces underlying the diversity of the human immunoglobulin heavy chain variable gene region. *BMC Genomics* **12**:78

23. Tang, E. S., Martin, A. (2007) Immunoglobulin gene conversion: Synthesizing antibody diversification and DNA repair. *DNA Repair* **6**, 1557-1571

24. Li, H., Cui, X., Pramanik, S. et al. (2002) Genetic diversity of the human immunoglobulin heavy chain V_H region. *Immunol Rev* **190**, 53-68

25. Sun, Y., Liu, Z., Ren, L. et al. (2012) Immunoglobulin genes and diversity: what we have learned from domestic animals. *J Animal Sci Biotechnol* **3:**18

26. Schmucker, D., Clemns, J. C., Shu, H. (2000) *Drosophila* Dscam is an axon guidance receptor exhibiting extraordinary molecular diversity. *Cell* **101**, 671-684

27. Cahoon, L.A., Seifert, H.S. (2011) Focusing homologous recombination: pilin antigenic variation in the pathogenic *Neisseria*. *Molec Microbiol* **81**, 1136-1143

28. Smith, G.P. (1985) Filamentous fusion phage - novel expression vectors that display cloned antigens on the virion surface. *Science* **228**, 1315-1317

29. Göringer, H.U. (2012) 'Gestalt', composition and function of the *Trypanosoma brucei* editosome. *Annu Rev Microbiol* **66**, 65-82

30. Cahoon, L. A., Seifert, H. S. (2009) An alternative DNA structure is necessary for pilin antigenic variation in *Neisseria gonorrhoaeae*. *Science* **325**, 764-767

31. Singh, P., Cole, S.T. *Mycobacterium leprae*: genes, pseudogenes and genetic diversity. *Future Microbiol* **6**, 57-71

32. Ndung'u, T, Weiss, R.A. (2012) On HIV diversity. *AIDS* **26**, 1255-1260

33. Manning, G., Whyte, D.B., Martinez, R. et al. (2002) The protein kinase complement of the human genome. *Science* **298**, 1912-1934

34. Zeqiraj, E., van Aalten, D.M.F. (2010) Pseudokinases-remnants of evolution or key allosteric regulators? *Curr. Opin. Struct. Biol.* **20**, 772-781

35. Boudeau, J., Miranda-Saavedra, D., Barton, G.F. et al. (2006) Emerging roles of pseudokinases. *Trends Cell Biol* **16**, 443-452

36. Caenepeel, S., Charydczak, G., Sudarsanam, S. et al. (2004) The mouse kinome: discovery and comparative genomics of all mouse protein kinases. Proc Natl Acad Sci U S A 101, 11707-11712

37. Han, Y.J., Ma, S.F., Yourek, G., Park, Y.D. et al. (2011) A transcribed pseudogene of MYLK promotes cell proliferation. *FASEB J* **25**, 2305-2312

Formatted: Swedish (Sweden)

38. Andersen, J.N., Jansen, P.G., Echwald, S.M. (2004) A genomic perspective on protein tyrosine phosphatases: gene structure, pseudogenes, and antigenic disease linkage. *FASEB J* **18**, 8-30

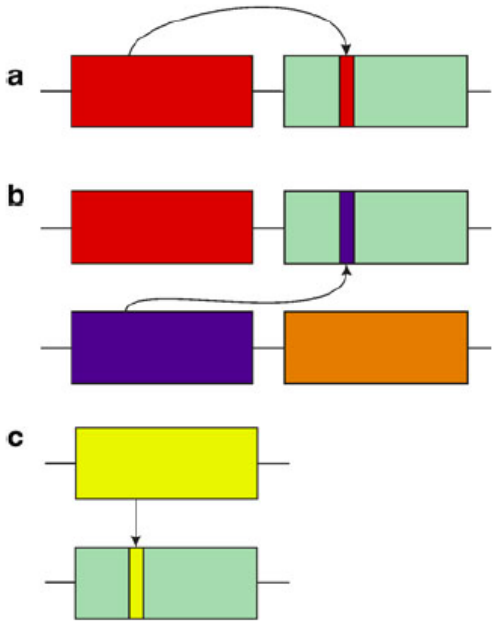

Figure 1. Schematic description of gene conversion. (a) Nonallelic gene conversion *in cis* between genes in the same chain. (b) Interlocus gene conversion *in trans* can occur either within a chromosome or between chromosomes. (c) Inter-allelic gene conversion between alleles on homologous chromosomes

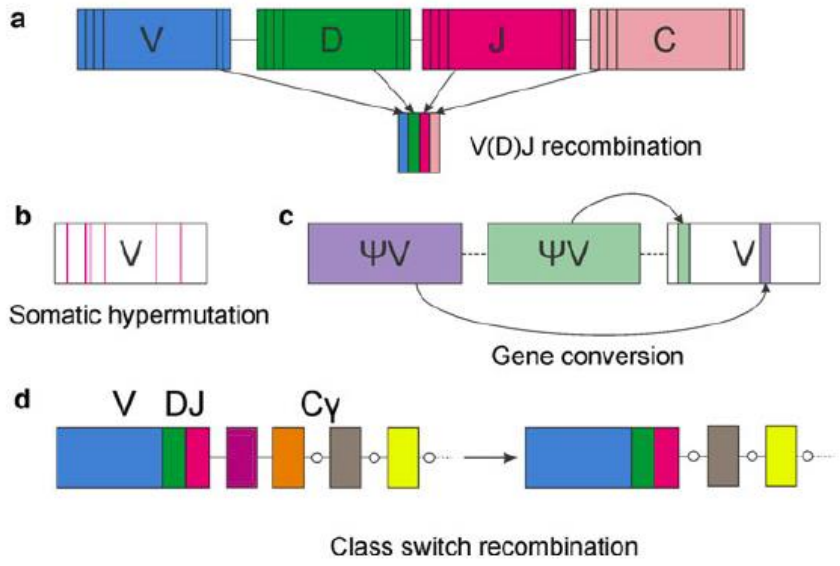

Figure 2. Schematic depiction of immunoglobulin gene diversification. (**a**) V(D)J recombination combines together 1 V, 1 D, and 1 J segment. 1 C segment is then added to obtain the complete Ig gene. (**b**) The sequence of the V segment can be further modified by somatic hypermutation that generates point mutations (indicated by *red lines*). (**c**) An additional mechanism to modify the sequence of the V segment is gene conversion, which introduces gene fragments from V region pseudogenes (ΨV). (**d**) The final modification step is the class switching in which one of the alternative exons of the constant region is attached to the variable region. This step is performed only when the BCR is activated and the Igs are needed. Highly conserved switch or recombination signal sequences (RSS) regions (depicted as *white circles*) guide the recombination process. The schema is not drawn to scale