

Antibody Diversity, Part 1

Here's some amazing facts. It's estimated that humans can make a billion (10^9) or more different specific antibodies, but they have only about 3×10^4 distinct genes. We think that genes encode proteins, so how can we get so many different proteins from so few genes? In addition, each B lymphocyte makes one and only one kind of Ab—that is, it makes a single combination of V_L and V_H regions, but it can make several different kinds of Ig—e.g., a naive B cell expresses both IgM and IgD on its surface, and after stimulation, its descendants can produce other classes of Ig as a result of the Ig class switch.

How can we reconcile these ideas? How can there be an immune system and still leave enough genes to have the rest of the organism? Well, one early realization that reduced the size of the problem was that antibodies, and particularly the antigen-binding region, consist of both H and L chains. Suppose an organism had 3 separate H and 3 separate kinds of L chains—how many combinations could it produce? How about if it had 10^4 H and 10^4 L chains? How many combinations would be possible? Well, 10^4 H + 10^4 L is how many different genes = 20,000. Since that's less than the 30,000 genes that vertebrates are thought to possess, vertebrates have (barely) enough genes to encode all the Ab molecules.

Even 20,000 genes is a lot. How does the body get them? Before the answer was known, one school of thought was that it would be beneficial for each individual organism of a species to have all the Abs that the species could make, so each individual inherited from its parents thousands of genes that encoded H and L chains of antibodies. This was called the Germ Line Theory, because all the Ab genes were thought to be in the germ line (i.e., egg and sperm cells) and passed from generation to generation. Others thought that there were too many different Ab molecules to be encoded by germ line genes. Rather, their idea was that the germ line only contained a few H and L chain genes, and then during B cell maturation, somehow mutations cropped up in these genes—different mutations in different B cells—so that new H and L chains were generated during the development of each individual. During the lifetime of each individual, then, the maturation of B cells give rise to a very large number of different Ab molecules from what was originally only a few inherited Ab genes. This idea was called the Somatic Mutation Theory (or Somatic Diversification

or Somatic Variation). What that means is that non-germ cells (somatic cells, namely lymphocytes) generate the diversity. The germ line theory basically implies that antibody diversity was generated in a species over evolutionary time, while the somatic variation theory implies that antibody diversity is generated by individuals during their lifetime. Amazingly, both ideas have proved to be partially correct. That is, the germ line contains lots of Ab genes, but further diversity can be generated by somatic mutation.

An early ingenious effort to deal with the problem of how to encode so many different proteins with a limited number of genes was provided by Dryer and Bennett in 1965, shortly after it was realized that all H and L chains consisted of V and C regions. They suggested that an organism could save genetic space by having only one gene for the C region of a H or L chain, and having separate genes encoding the many V regions. Since the C region is half or more of the total protein, one could reduce the amount of genetic information needed in the germ line by at least 50% with this mechanism. Then, somehow, during development of B cells, either V and C genes could be stitched together, or V and C mRNA or V and C protein fragments could be joined. This violated prevailing ideas of gene organization and protein synthesis and was largely disregarded at the time. However, 11 years later, after the advent of restriction enzyme technology, Hozumi and Tonegawa, two Japanese scientists who were working in Switzerland, showed that Dryer and Bennett were essentially correct.

At about the same time that Tonegawa showed that the Ig genes could rearrange, it was becoming possible to clone and sequence large amounts of DNA. Immunoglobulin genes were among the first to be sequenced., and in a flurry of activity in the late 70s and early 80s several groups, including Tonegawa's, Honjo's, Leder's, and Hood's, sequenced the H and L chain genes from humans and mice. I'll summarize what they found in each case.

For kappa (κ) light chains they found nearly 100 different coding regions in DNA ("genes") for variable regions in mice (about half that number in humans), and one coding region for C κ . Each V κ coding sequence was preceded by a short hydrophobic "leader" or "signal" sequence, that is characteristic of secreted proteins. Each V κ encoded amino acids 1-95 of the V κ

protein. The C κ coding sequence contained the information for amino acids 109-214. What happened to 96-108? Subsequently, they found several regions between the V κ and C κ coding regions that had the genetic information for the missing amino acids, and that were named the J κ , for joining regions. There were 5 of these but only 4 of them encoded functional proteins (one had an internal stop sequence and is therefore called a pseudogene--it has a promoter, a start codon, and a stop codon, but the internal stop codon prevents it from making functional protein.)

Here's a cartoon of the arrangement of the mouse κ chain genes. (See Figure 5-3) In the "germ line" the genes encoding V, K and C segments are physically separated, but all on the same chromosome #6 in mice and #2 in humans. Subsequently it was shown that the κ protein is encoded by a single mRNA molecule, so the genetic information is no longer separate in the mRNA. How does this happen? Is the DNA rearranged or are separate pieces of RNA stitched together into one? It appears that the genetic information (DNA) has to be physically rearranged by the B cell before it can make a functional protein (Fig. 5-4). First the V regions is rearranged next to one of the J regions, and the intervening DNA is somehow destroyed. Then a second rearrangement brings the V-J region adjacent to the C region, again with loss of DNA. Now the gene can be transcribed into RNA. This RNA contains both coding information (exons) and non-coding information (introns) and must be processed in the nucleus into a continuous messenger RNA molecule by removal of the introns. This RNA processing is different from the DNA processing I just described--different molecules involved, different enzymes, etc. The mRNA can then be transported to the cytoplasm, where it is translated into protein by ribosomes attached to the rough endoplasmic reticulum. The protein is passed into the ER lumen, the leader peptide is removed by an enzyme (signal peptidase), and the H and L chains are assembled into Ig molecules.

How many different κ chains can a mouse make? Well, if any V κ can join with any J κ chain, then the possibilities are $85 \times 4 = 340$ at least (it's actually more as we'll see in a while).

With some minor differences the same general principles appear to apply to both λ and H chain genes. The λ genes are on a different chromosome (16 in mice, and 22 in humans). In mice there is less diversity in the λ genetic arrangement, because there are only 2 V λ segments in mice,

and 3 functional $J\lambda$ segments, each with its own $C\lambda$ segment (Fig. 5-3a). Thus rearrangement can only produce about 6 different $V\lambda$ genes. Humans have more like 30 $V\lambda$ and 4 $J\lambda$ segments, so they can generate 120 different $V\lambda$ genes.

H genes are similar to kappa genes. There are a large number of V_H segments in mice and about 50 in humans; both species have 4 J_H segments. However, these plus the C segment don't encode the whole heavy chain. Another segment, called the diversity (D) segment, was discovered that encodes only 3 amino acids (95-97 of the H chain, which is, not surprisingly, in the third hypervariable region or CDR3. Roughly one dozen (mice) or two dozen (humans) of these D segments are located between the V_H and J_H s segments. (Fig. 5-3c). These appear to rearrange sequentially like the kappa chain, with the D and J segments first rearranging, and then the V segment rearranging to give the final VDJ arrangement adjacent to the $C\mu$ coding region (Fig. 5.5). Notice that coding regions for all the different C regions of H chains are lined up in a row behind the $C\mu$ coding region. One curious feature of this arrangement is that the transcription stop signal is actually located after the $C\delta$ gene segment, not after $C\mu$. So the primary transcript (Fig. 5.5) contains both coding regions, but this primary transcript can be processed into mRNA in two ways in the nucleus giving rise to either an IgM or and IgD heavy chain coding sequence. So naive B cells actually make two different classes of Ig, but both of them have exactly the same VDJ region coding sequence, and so both are specific for binding to the same antigen(s).

How many different H chains can a mouse generate? Again assuming that the different segments can recombine randomly, if there are about 130 V_H segments, 13 D segments and 4 J segments for $130 \times 13 \times 4 \sim 7000$ different H chain sequences.. For humans the numbers are $51 \times 27 \times 6 > 8000$. If there are 300+ different light chain sequences possible, then a mouse or human can generate at least 8000×300 (or 7000×300) $> 2 \times 10^6$ different Ig molecules. In other words, this *combinatorial* rearrangement of a relatively few gene segments can give rise to an enormous number of different antibody molecules.

Now, one obvious question about all this genetic rearrangement is how does it occur, and what happens to the DNA that used to be between the segments on the chromosome when they get

rearranged next to each other? One of the curious things that became apparent when people began to obtain the DNA sequence of the Ig genes is that all the segments that are rearranged during development are “flanked” by a common sequence of bases and that these “conserved flanking sequences” are actually complementary to each other. These flanking sequences consist of 7 and 9 bp regions (heptamers and nonamers) that are separated by 12 or 23 bp. The idea is that 12 or 23 bps makes up one or two turns of the double helix, and that allows the nonamer and heptamer to be aligned with each other (See Fig. 5-7). Because these flanking sequences serve to identify the sites of recombination during Ig gene rearrangements, they are often called *recombination signal sequences*.

Once these sequences were discovered, it was assumed that some protein recognizes the nonamer/heptamer flanking sequences and catalyzes a rearrangement of the DNA. It's now known that two lymphocyte specific proteins, called RAG-1 (recombination activating gene) and RAG-2, are required to recognize and cut the DNA at these sites, and that other DNA repair proteins are involved with splicing the DNA back together. Figure 5-7 suggests a way that this might happen, with the heptamers and nonamers from different V and J genes, say, coming into alignment so that they can form base pairs with one another. The 7/9 arrangement and the 12/23 bp space means that both the heptamer and nonamer are on the same side of the helix and can pair with equivalent sequences flanking other segments of DNA. The arrangement also ensures that only D genes can pair with J genes but not other D genes, and that D can also pair with V, but V can't pair directly with J. Either the intervening piece of DNA is excised, formed into a circular structure, and degraded by DNases in the cell, or its relocated on the chromosome and retained, depending on the orientation of the stretches of DNA being spliced together. In other words, the rearrangement of Ig genes can result in loss of genetic information by lymphocytes. It's the only known process that involves ablation (destruction) of genetic information, rather than regulation of gene expression.

As you might guess, such a complex process is not perfect, and it's been found that many of these rearrangements fail to give a functional coding sequence for H or L chain proteins. That is, sometimes the joins are out of reading frame and result in either premature stop codons or nonsense (Fig. 5-9). Somehow the B cell seems to know if it messed up, because there's evidence

that B cells go through various stages during the course of differentiation. In general the Ig genes rearrange in a predictable sequence. First, the Ig H chain genes rearrange on one of the homologous chromosomes (14 in human and 16 in mouse) (see Fig. 5-11). If this fails, then the Ig H chain genes on the other homologous chromosome undergoes rearrangement. If this fails, the immature B cell self-destructs through a process of programmed cell death that's called apoptosis. If one of the H chain rearrangements succeeds in making a functional protein, this IgM chain is displayed on the cell surface along with a nonfunctional protein that looks somewhat like a light chain. If the cell makes a functional IgH chain (and sometimes even if it doesn't), then the kappa chain genes rearrange on one chromosome. If this is successful, then the cell can make both a H and L chain, and thus a functional antibody. If not, the cell tries to rearrange the other kappa genes on the other homologous chromosome, and if that fails, there are still two lambda chain clusters it can play with. So the cell gets 2 tries at making a functional H chain gene and 4 tries at a light chain gene. Despite that, it's been estimated that only about 10% of pre-B cells in the bone marrow, or Bursa, actually succeed in rearranging both H and L chain genes to make functional Ab molecules. The rest commit suicide. Even with this high mortality rate, people have estimated that there are over 10^9 B cells with different Ig molecules (i.e., different Ag binding sites) on their surfaces.

This rearrangement explains why each B cell makes one and only one kind of Ig molecule, with a single Antigen-binding site (that is, it explains why Abs are clonally distributed). The way the genes are set up at first in a progenitor cell, they can't produce functional H or L chains. Only after gene rearrangement can the B cell make H and L chains, and then it creates only one functional H chain gene and one functional L chain gene. So of course it can make only one kind of Ig (well, two kinds, an IgM and an IgD, but they both have the same variable regions).

It also turns out that each B cell makes only one allotype of Ig, even if the animal inherited different allotypes from the two parents. That is, suppose that the animal inherited A11 from mom and A12 from dad. All things being equal you might think that B cells will have some A11 H chains and some A12 H chains, so some Ig will have to A11, some will have 2 A12 and some will have a combination. This never happens. That is, a given B cell makes either 2 A11 H chains or 2 A12 H

chains, but never both and never a combination (see Fig. 5-10). This phenomenon is called allelic exclusion. It used to seem mysterious--how did the B cell manage to turn off one of its H chain genes, but it's actually a trivial consequence of how H chain genes get built. They have to be rearranged from fragments of DNA on one chromosome, which is either from mom or from dad. If that works, then the cell makes all that kind of H chain protein, containing only one allotype. If that fails, then the cell tries to rearrange the other chromosome to make an H chain protein, which would contain the other allotype. Since no cell has functional genes from both parental chromosomes, then each cell must be making only one allotype.

*Created and copyright by Gary Reiness
Last updated: Feb. 6, 2004*