VCU Bioinformatics and Bioengineering Summer Institute

The Origin of Human Genetic Variability from Cross-species Comparisons

Notes: Introduction to Single Nucleotide Polymorphisms (SNPs)

I. Overview

The Scenario begins with a grand idea of discovering the source of human variability. By the end, it gets bogged down in a technical problem, but let's leave that aside for the moment and go back to the big picture.

Human variability – at least its genetic component – is the result of differences in only a tiny fraction of the human genome. In contrast, two bacteria, *Escherichia coli* and *Salmonella typhimurium*, that you would be hard pressed to tell apart by their looks or metabolic capabilities differ in as much as x%. One lesson we might take away, of course, is that we are all genetic siblings under the skin, but the question remains: Why is it that so little difference in DNA sequence makes so large a difference in at least appearance?

The most abundant genetic variants in human genomes are single basepair differences, also called single nucleotide polymorphisms (SNPs). Understanding how SNPs arise from might therefore help us understand the major source of human variability.

II. Technical notes on sequences

SNPs by definition have positions where more than one nucleotide appear in the population. This poses a problem in representation. How do you convey the idea that the nucleotide at a given position in a sequence might be either an A ***or*** a G? A system of nomenclature has grown up to deal with this problem:

Nuc	Symbol	Nuc	Symbol	Nuc	Symbol
A	A	A or C	M *(aMino)*	A or C or G	V (*not T*)
C	C	A or G	R (*puRine*)	A or C or T	H (*not G*)
G	G	A or T	W (*Weak*)	A or G or T	D (*not C*)
T	T	C or G	S (*Strong*)	C or G or T	B (*not A*)
		C or T	Y (*pYrimidine*)	A or C or G or T	N (*aNy*)
		G or T	K (*Keto*)		

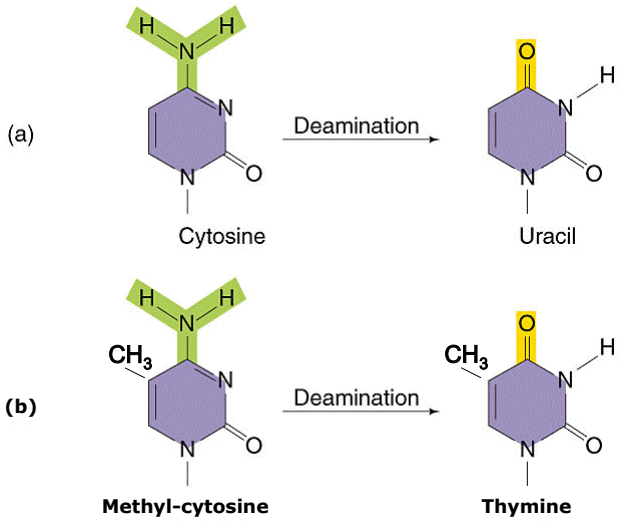
Thus the Scenario's first SNP sequence, GGGCGGGGARCCTGGCCACCA, represents a sequence that permits either an A or a G in the 10th position.

III. Mutation at CG-dinucleotides

Mammals modify many cytosines that are on the 5' side of guanines (5'-CG-3'), perhaps for purposes of regulating chromosomal architecture or gene expression. These modified cytosines become hot spots for mutation because of the chemical properties of cytosine. Cytosine spontaneously deaminates to become uracil (Fig. 1). This is true in all organisms, since it is an intrinsic property of cytosine in water, and all organisms have devised methods to cope with this source of mutations. It's not difficult to see a way out of the problem. The product of cytosine deamination is uracil, a base not naturally found in DNA. Cells have enzymes (uracil

glycosylases) that remove the uracil from the DNA. Subsequent DNA repair replication uses the other DNA strand to put the proper cytosine back where it belongs.

Look how different is the situation if the cytosine is methylated. Then deamination produces not uracil but thymine, a natural component of DNA. The cell no longer has an obvious way to detect that a mutation has taken place. There are less certain ways of correcting the error, but nonetheless, mutations of C^{me} are more common than at C. Hence, since C^{me} occurs predominantly in CG dinucleotides, mutations at CG are relatively common in mammalian DNA .



What about the nucleotides ***nearby*** CG dinucleotides? Conceivably, the effort of the cell to detect and repair CG-induced mutations may lead to errors in nearby bases. This would be the case if repair involved a process that required the extra DNA replication in the region of the dinucleotide. Some types of DNA replication are particularly prone to error and would increase the mutation rate in the entire region. It is of therefore considerable interest to determine if SNPs are more common near CG dinucleotides.

IV. SNP sequence comparisons and the direction of mutation

Key to determining the ***source*** of genomic variability is to assess the ***type*** of genomic variability. Which nucleotides are more likely to mutate to which other nucleotides and in what contexts? Unfortunately, merely collecting SNPs gives ambiguous information. If some portion of the population has the sequence GGACCT while the rest has GGGCCT, how can you tell whether the historical mutation was A in the third position mutating to G rather than G mutating to A? Without the ability to go back in time, the next best solution is to examine the sequences of other organisms who are closely related to us, with the idea that their genomes were not likely to suffer the same mutation.

Chimpanzee DNA is the obvious choice, since chimps are closer in DNA sequence to humans than any other living species. Finding sequences in the chimp genome similar to human SNPs may decide the question of mutational directionality.

How does one find similar regions of short DNA sequences in the genome of another organism? Since the amount of DNA in a primate genome is so large, one must take care to guard against spurious similarities. In this regard, the following guidelines may be useful to assess the utility of a match with an SNP sequence:

1. The SNP position must be within the aligned region
2. The nucleotide where the SNP site matches must be one of the two alleles permitted by the SNP
3. The alignment must be at least 100 bp in length
4. There must be at least 95% similarity between the sequences flanking the polymorphic site