Genomic Variants

FAIR Data Practices for Omics Analysis Workshop

University of Delaware

April 21 (Day 4)

An NIH sponsored program
Chemistry
Biology
Interface
at the University of Delaware

Center for
Bioinformatics &
Computational
Biology

Why do mutations occur?

Sources of mutations

- Environmental Causes
  - UV light
  - Radiation
  - Chemicals

- Cellular sources of damage and Enzyme errors
  - DNA polymerase is not perfect
  - 1.1 x 10^{-8} error rate per base. Proofreading capabilities correct this to about 10^{-10}. That's an error every three replications.
  - Our cells divide about 10^{16} times in a lifetime . . . That's about 1 quadrillion uncorrected mutations per person! Is that possible?!?

Do all the mutations matter?

- Yes, its possible and probably not that far from reality . . .
- But keep in mind that those mutations are each in one of 37 trillion cells in your body, most of which are Somatic
- A mutation in one of those cells is unlikely to cause you harm
- Importantly, the only cells that might be passed to your offspring are sperm or egg (Germline cells) . . .
    - Sperm average 300-400 cell divisions from the zygote (~100-130 mutations each . . . Depending on Dad's age)
    - Eggs are only about 30 cell divisions from the zygote (10 mutations)

Somatic vs Germline Mutations

- So when we think about mutations the differentiation of whether it is Germline (inheritable) or Somatic (only going to affect you) is an important one

- Somatic mutations are of concern because they have the potential to cause:
  - Cancer if they happen to hit a gene that affects proliferation
  - Other disorders, but usually only when they occur early enough in development to be passed to a significant number of cells
  - Evolution

- Germline mutations are of concern because they can be passed to every cell in an offspring's body

Are mutations always deleterious?

- No, most are seemingly neutral

- The majority of the genome is intergenic. While intergenic DNA has function, single mutations have a high chance of landing in a portion that will have little to no impact.

- Even within protein coding genes, many mutations will have little to no impact.
  - Third base wobble in codons
  - Synonymous substitutions
  - Intron vs. Exon
  - Even non-synonymous substitutions don’t always cause an effect
  - Other complications: ploidy, polygenic traits

Ploidy

- Remember that in humans there are two copies of every chromosome in somatic cells, so most genes have two copies (alleles)
- For any given trait the allelic balance can be homozygous (both the same) or heterozygous (different)
- A mutation in one allele will probably not be matched by a mutation in the other allele
- What effect does this cause:
  - Wildtype: both alleles are the same as the reference (0/0)
  - Recessive: phenotype only if both alleles have the mutation (1/1)
  - Dominant: phenotype if either or both has the mutation (0/1, 1/1)
  - Partial Dominance: homozygous mutation (1/1) has more severe effect than heterozygous (0/1)
- Some organisms are more than diploid (plants can be tetraploid, hexaploid)

Do these rules always apply?

- No
- In single celled organisms things are simpler
- Each mutation has the potential to cause immediate impact organism-wide
- A mutation in a single celled organism will always be heritable
- Most single cell organisms have more coding DNA than non-coding

What's a SNP?

SNP vs SNV

- When people think about detecting genome mutations, they almost always talk about SNPs

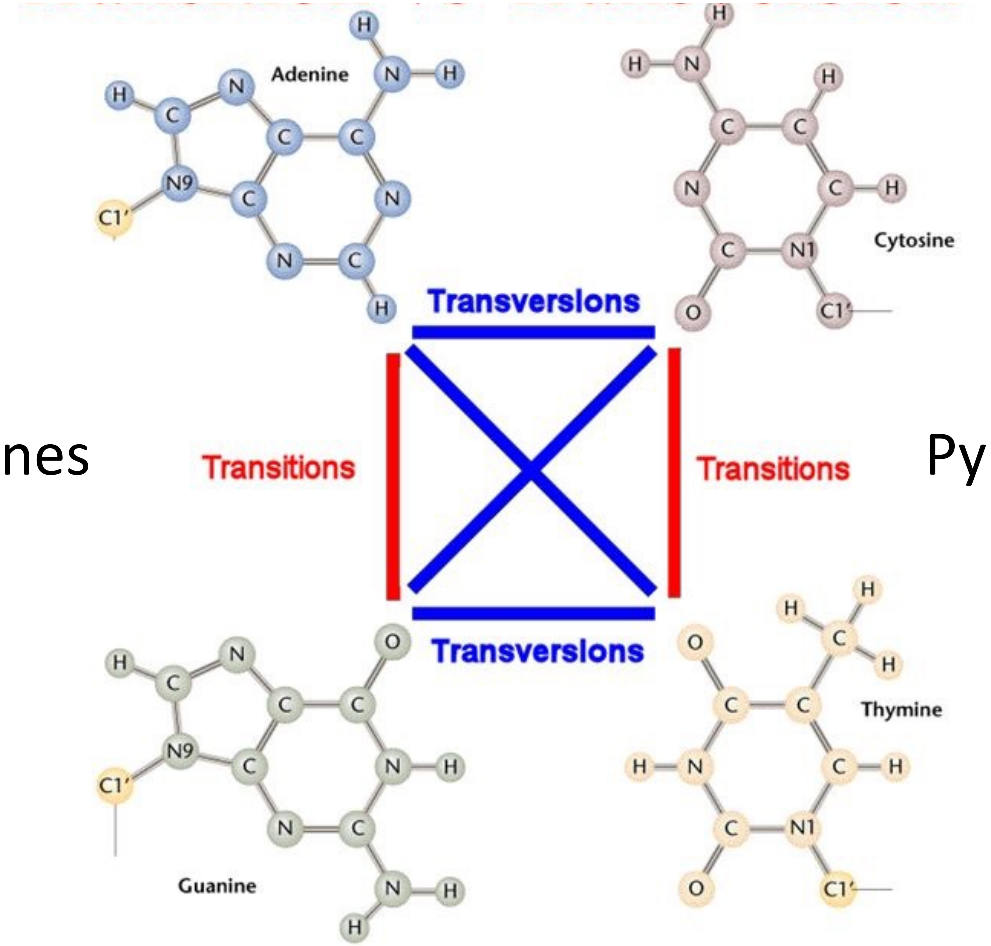
- A Single Nucleotide Polymorphism (SNP) is a single base variant (point mutation) in the genome that is found in >1% of the population
(>1% minor allele frequency – MAF)

- When you detect a variant in an individual, it is more correctly called a Single Nucleotide Variant (SNV)

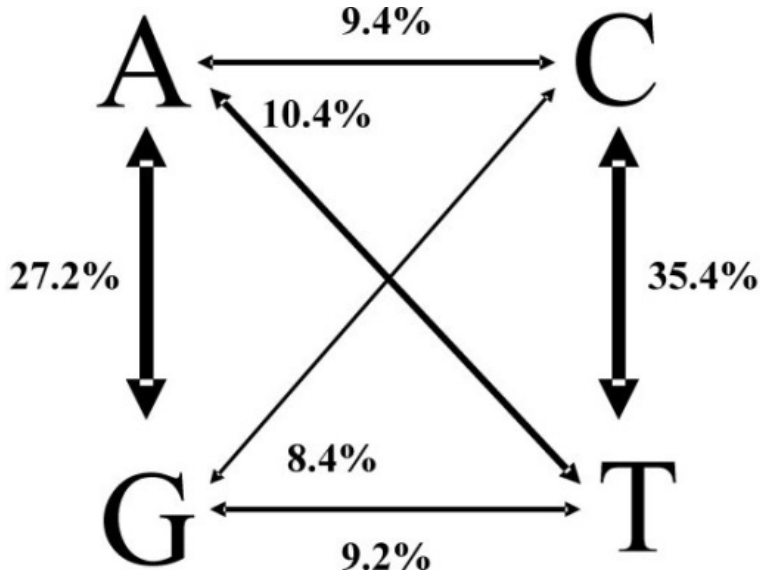
- All SNP's are SNV's, not all SNV's are SNP's

Not all SNV's are equally common: Transitions and Transversions

Purines



Pyrimidines



Wondji CS, Hemingway J, Ranson H - BMC Genomics (2007)

Causes of SNV's

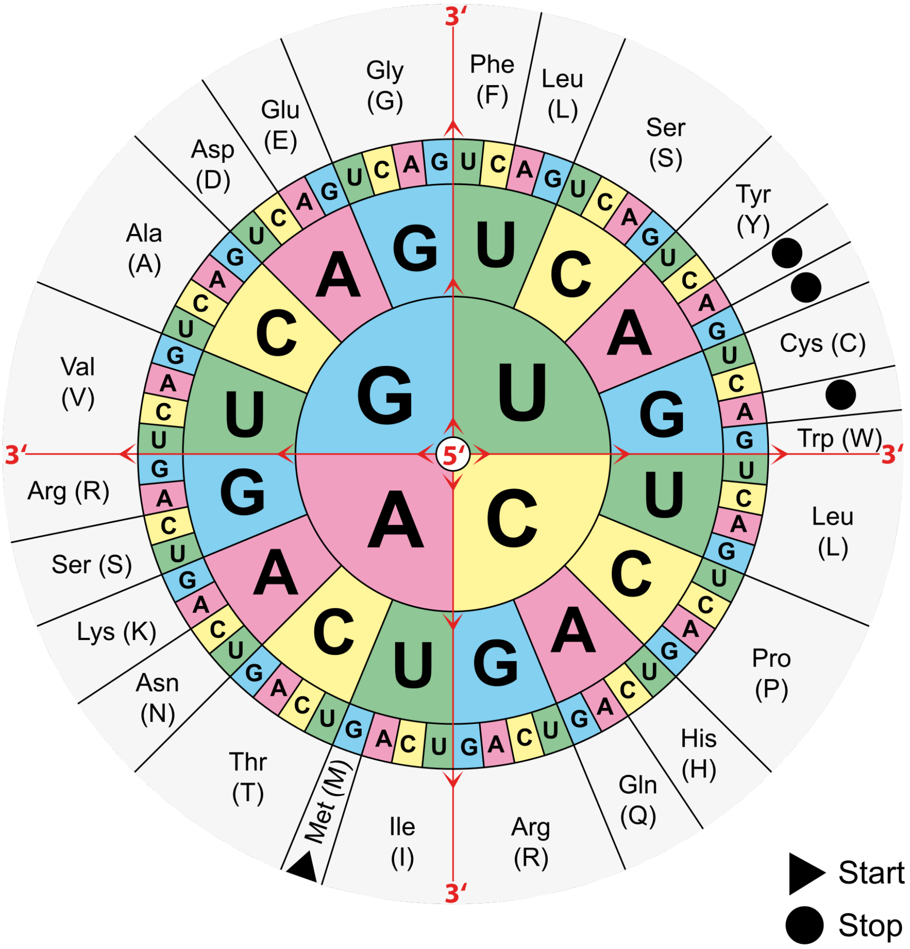
- Polymerase grabs the wrong base . . . Proofreading enzyme misses (usually transition)

- Mutagen: Chemical, UV, radiation can cause breaks or chemical modifications to bases
  - May cause polymerase error
  - More often it is DNA Damage Repair System that repair incorrectly

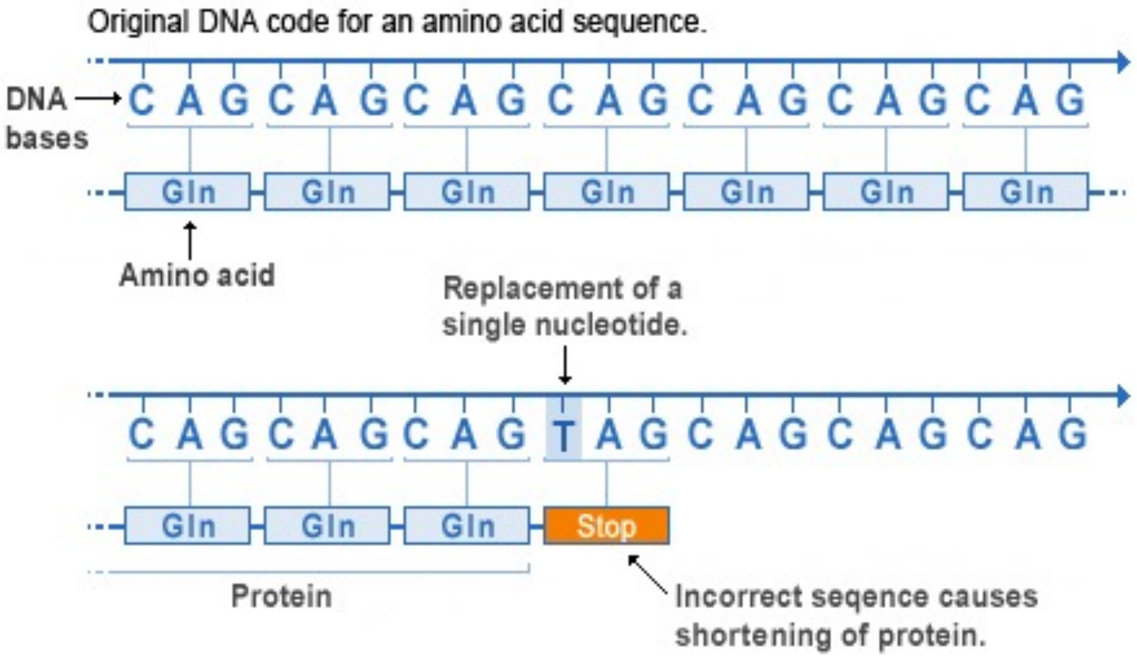
SNV Types

- Non-coding
  - May impact regulatory sites like promotors, enhancers, protein binding sites

- Coding
  - Synonymous (silent): no amino acid change →
  - Non-Synonymous: changes the coding sequence
    - Missense – changes amino acid
    - Nonsense – changes amino acid to stop codon (premature termination)
    - Run-on – Stop codon changes to an amino acid (delayed termination)
    - Splice Site – disrupts intron/exon boundary (intron may get translated)

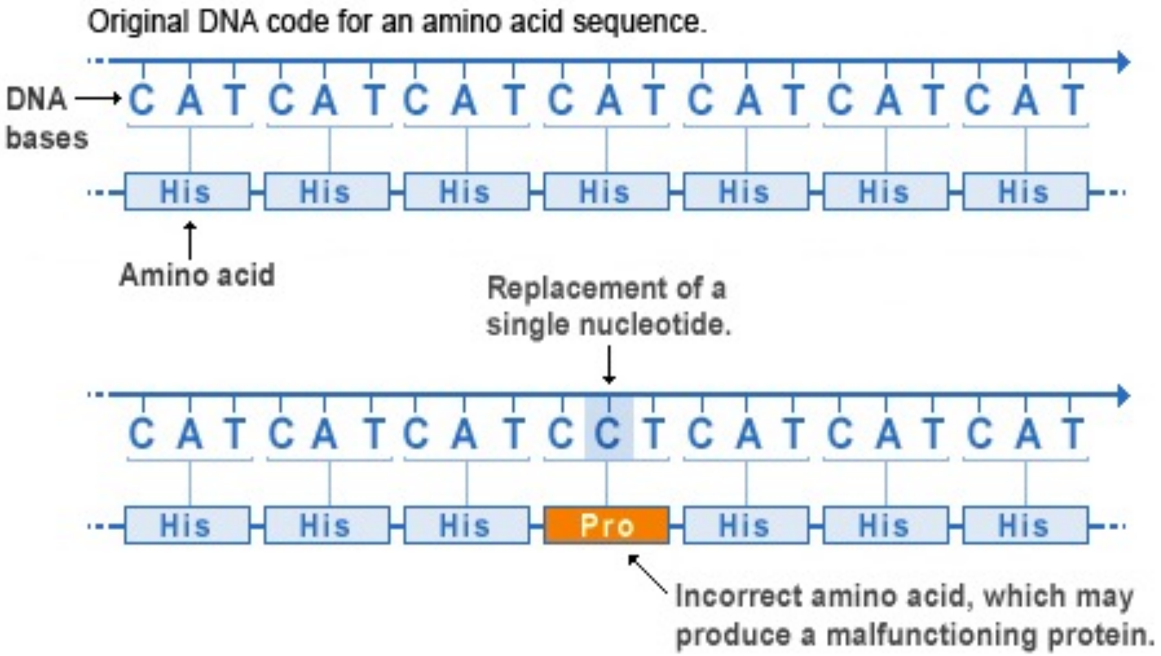


Nonsense mutation

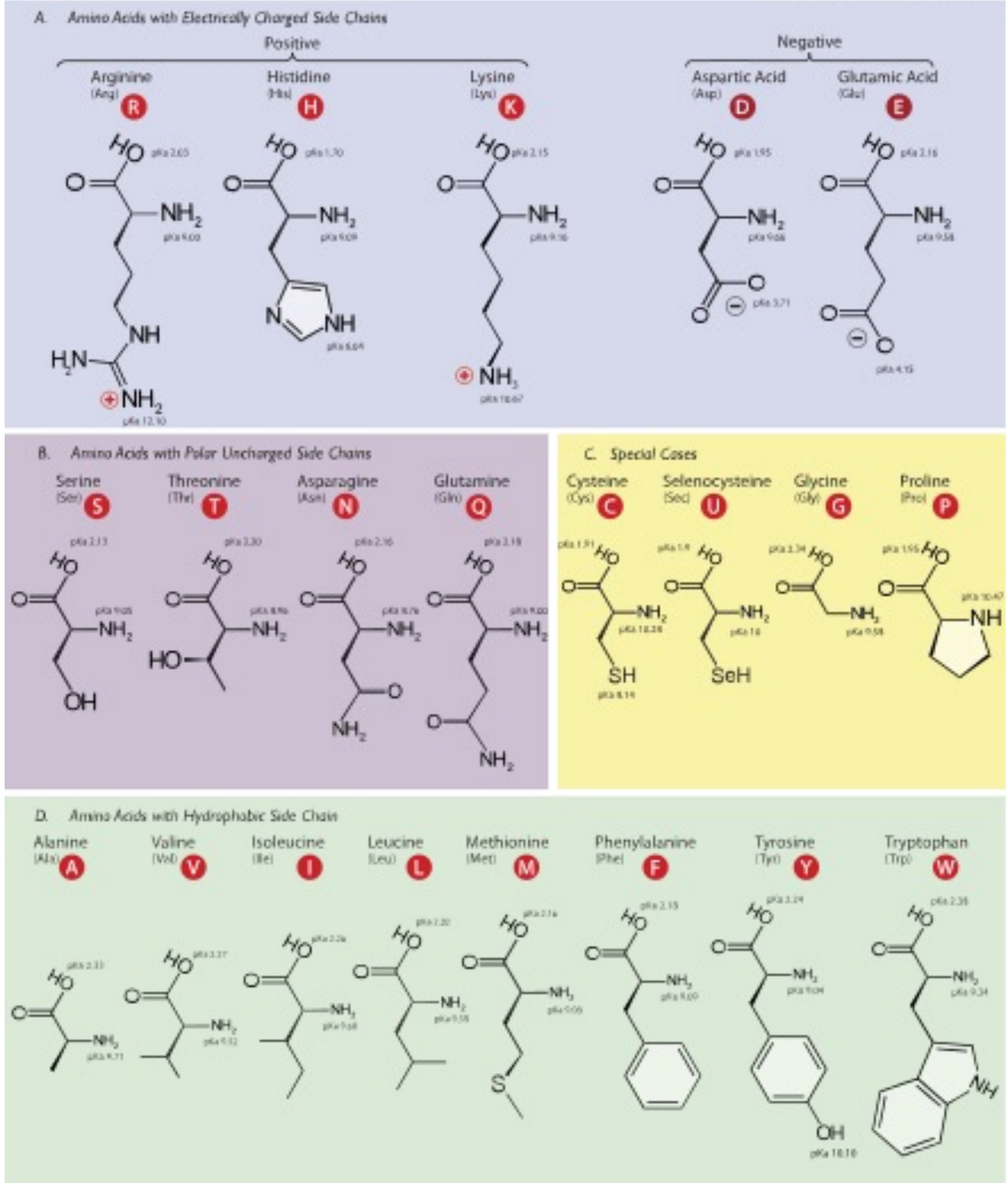

Original DNA code for an amino acid sequence.
DNA bases
C A G C A G C A G C A G C A G C A G C A G
Gln Gln Gln Gln Gln Gln Gln
Amino acid
Replacement of a single nucleotide.
C A G C A G C A G T A G C A G C A G C A G
Gln Gln Gln Stop
Protein
Incorrect seqence causes shortening of protein.

U.S. National Library of Medicine

Missense mutation


Original DNA code for an amino acid sequence.
DNA bases
C A T C A T C A T C A T C A T C A T C A T
His His His His His His His
Amino acid
Replacement of a single nucleotide.
C A T C A T C A T C C T C A T C A T C A T
His His His Pro His His His
Incorrect amino acid, which may produce a malfunctioning protein.

U.S. National Library of Medicine



Missense Substitutions may be Conservative

Ala	4																			
Arg	-1	5																		
Asn	-2	0	6																	
Asp	-2	-2	1	6																
Cys	0	-3	-3	-3	9															
Gln	-1	1	0	0	-3	5														
Glu	-1	0	0	2	-4	2	5													
Gly	0	-2	0	-1	-3	-2	-2	6												
His	-2	0	1	-1	-3	0	0	-2	8											
Ile	-1	-3	-3	-3	-1	-3	-3	-4	-3	4										
Leu	-1	-2	-3	-4	-1	-2	-3	-4	-3	2	4									
Lys	-1	2	0	-1	-3	1	1	-2	-1	-3	-2	5								
Met	-1	-1	-2	-3	-1	0	-2	-3	-2	1	2	-1	5							
Phe	-2	-3	-3	-3	-2	-3	-3	-3	-1	0	0	-3	0	6						
Pro	-1	-2	-2	-1	-3	-1	-1	-2	-2	-3	-3	-1	-2	-4	7					
Ser	1	-1	1	0	-1	0	0	0	-1	-2	-2	0	-1	-2	-1	4				
Thr	0	-1	0	-1	-1	-1	-1	-2	-2	-1	-1	-1	-1	-2	-1	1	5			
Trp	-3	-3	-4	-4	-2	-2	-3	-2	-2	-3	-2	-3	-1	1	-4	-3	-2	11		
Tyr	-2	-2	-2	-3	-2	-1	-2	-3	2	-1	-1	-2	-1	3	-3	-2	-2	2	7	
Val	0	-3	-3	-3	-1	-2	-2	-3	-3	3	1	-2	1	-1	-2	-2	0	-3	-1	4
	Ala	Arg	Asn	Asp	Cys	Gln	Glu	Gly	His	Ile	Leu	Lys	Met	Phe	Pro	Ser	Thr	Trp	Tyr	Val

BLOSUM62 matrix
(higher numbers are likely to be conservative)

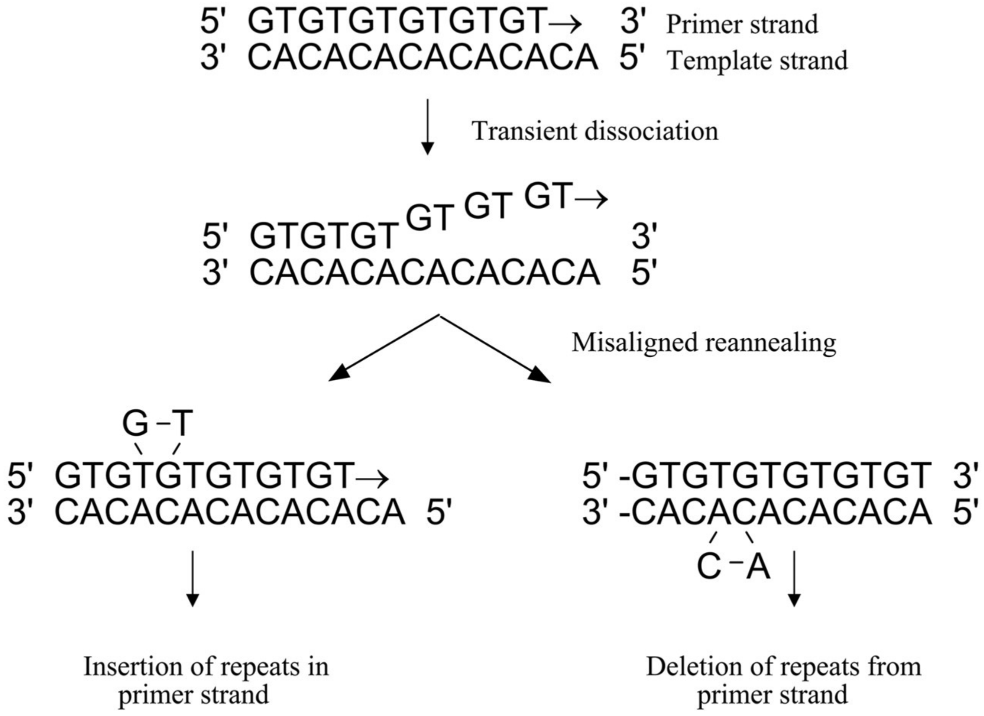
InDels or DIPs

- InDels are small Insertions or deletions in the genome (typically <~50bp in length . . . longer are structural variants)

- Sometimes called DIPs (Deletion Insertion Polymorphism), but this should probably carry the same population definition as SNP

Causes of InDels

- Primer Slippage
- Template Slippage
- Mobile Genetic Elements
  - Transposons
  - Prophage
- Misdirected enzymes
  - DNA repair enzymes gone wrong
  - CRISPRs



Impact of InDels

- Can disrupt (or delete) entire regulatory motifs – promoters, transcription factor binding sites etc

- In protein coding regions:
    - In an exon, frameshift WILL OCCUR unless the inDel is a multiple of 3
    - Disrupt intron exon boundaries
    - Disrupt start (no transcription) or stop (missense) codon
    - Insertion or deletion of sub-domain level motifs

What are structural variants?

Structural Variants

- As the name implies Structural Variants (SV's) are mutations that have a larger impact on the structure of the DNA than small variants like SNV's and InDel's
- SV's usually operate at the local to chromosomal scales (Kb to Mb distances)
- Larger SV's are called Chromosomal Abnormalities and they operate at larger scales within and across chromosomes and can usually be detected microscopically (karyotyping)

Types of Structural Variants

- Insertions and Deletions – like InDels, but bigger (minimum length definition varies, but always over 50bp)

- Duplication – Segment of DNA gets duplicated – Can be tandem (adjacent) or distant

- Inversions – Segment of DNA is reversed from what it should be

- Translocations – Segment of DNA moves from one position to another

Detecting SV's

- DNA-Seq can be performed on Illumina (paired end)
  - Map to reference chromosome
    - Identify breakpoints – Places where reads seem to stop mapping properly
    - Identify discontiguous mapping – Either pairs (or two different parts of same read) map in different places than expected
    - Identify Copy Number Variations (next slide)
  - Difficult, limited resolving power (reads are short)

- Longer Read technologies (PacBio) very promising for overcoming these issues . . . (expensive, but price is reducing)