Introduction to genome variation

Genome Variation

Genome Variation typically means variation of individual genomes within a species. Variation between species is the realm of phylogenetics and/or comparative genomics.

Variants commonly detected by NGS consist of:

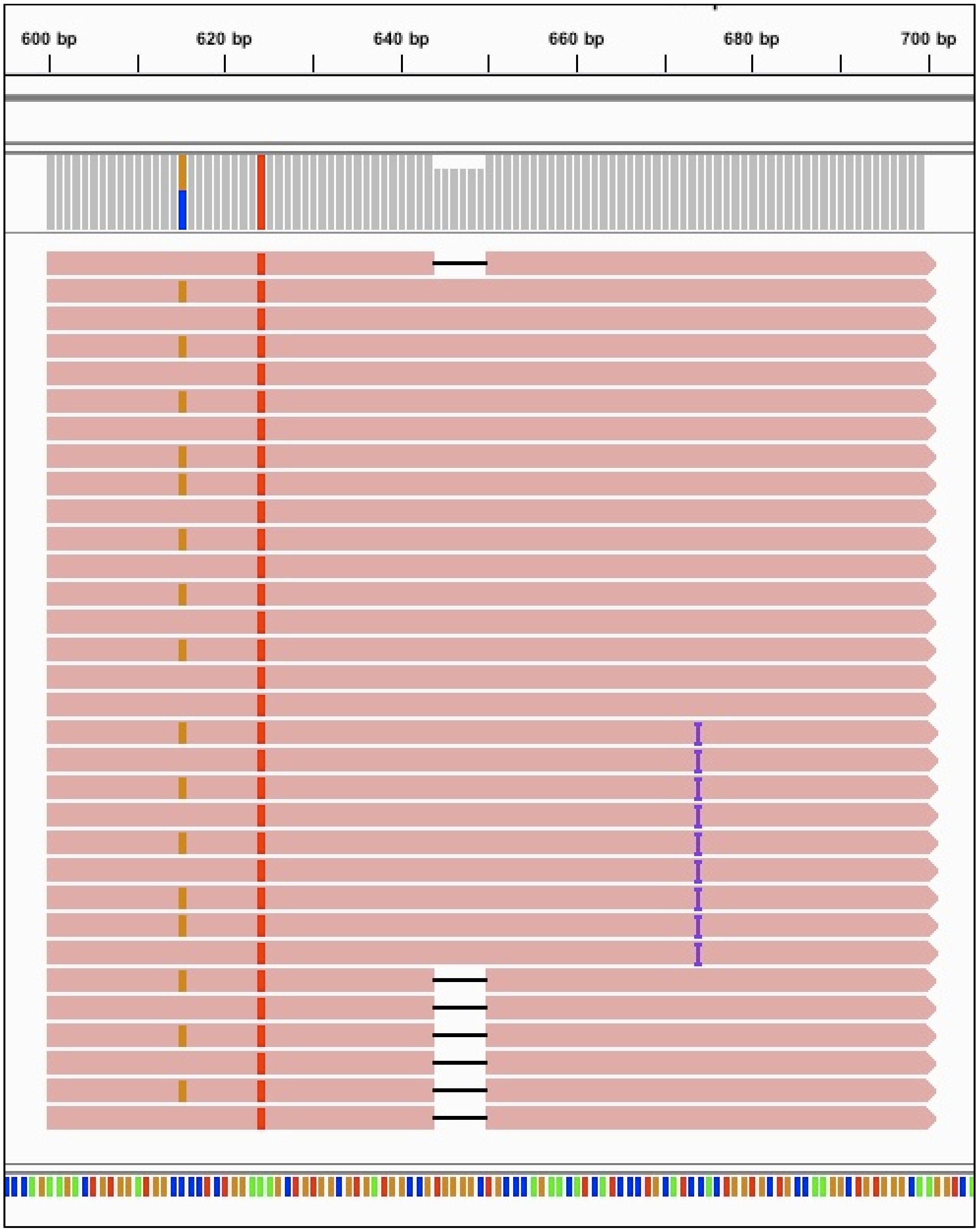
- single base base changes (SNPs)
- insertions and deletions, and
- larger scale structural changes such as large deletions, large duplications (up to and including whole chromosomes) and translocations

"Larger scale" is usually defined relative to the capabilities of the technology; for example, a "small indel" usually means "detectable within a single sequence read". In 2009, sequence reads were about 50 bp but in 2011 they were 100 bp.

More classic variants such as microsatellites, STRP's, and Alu's, can be somewhat more difficult to detect with NGS because their length is just slightly beyond typical NGS read lengths but shorter than what could be reliably detected with paired-end sequence data given the variation in fragment lengths with NGS. Fortunately, these methods were usually proxies for the functional genetic polymorphisms which can now be detected directly with NGS.

According to the Genome News Network, about 90% of genome variation occurs as single nucleotide polymorphisms. dbSNP as of June 26, 2012 contains 187,852,828 SNP submissions (ss #'s) which condense to 53,558,214 familes (refSNPs, rs #'s).

Yesterday we looked at the generalized workflow for finding variants with NGS. Here is an image displaying SNPs and indels in IGV. BP 615 is a heterozygote WT/MT (0/1); BP 624 is a homozygous SNP; BP 643-649 is a heterozygous 6 bp deletion; BP 674 is a heterozygous 3 bp deletion.



Would you believe me if I told you this data were from a normal individual mammalian genome?

> It's obviously not diploid - the various heterozygous features should associated to 2 alleles; this diagram shows 6! The SNP at 615 is associated with both forms of the deletion AND both forms of the insertion. This would not be expected from normal tissue from a single individual. But it could represent: a) a mixed population of individuals, b) a mixed population of mutated cells, e.g. cancerous tissue, or c) a hexapolid (or greater) genome.

Variant analysis is almost always comparative, whether comparing variation rates, tracking traits, or looking for causal mutations. These comparisons can start with tools to analyze BAM files to identify putative variants (the Variant Caller) and then continue into more detailed annotations (the Variant Annotator).

We will introduce you to all these stages; keep in mind that this course makes no assumption about the genome you're analyzing - it could be:

1. a mixed population of haploid genomes (bacterial cultures),
2. diploid pedigrees (human mendelian diseases),
3. cancer genomes (heterogeneous populations of dipolid genomes)
4. polyploid or even "mixiploid" (where polidy varies between individual cells of the organism)