# Identify Variation

Bioinformaticians in laboratory genomics typically develop software and processes to enable or facilitate the analysis and interpretation of data

1. What is the role of the bioinformatics pipeline in variant interpretation?
2. Which kind of software identifies variation? Give an example of this type of software.
3. If you want to find single nucleotide variants (SNVs) and copy number variants (CNVs) in your data, do you need to use more than one variant caller?

Sometimes we might want to confirm that a variant we have detected is present in the sample using a different technology (to exclude the possibility that it is an artefact)

1. What sequencing technology is typically used in laboratory genomics to confirm the presence of an SNV in the sample originally detected by Next Generation Sequencing?

## Various sources

### Somatic variation

We have already looked at variation from a germline sample called using a germline pipeline in our laboratory (DNA Sequence third part). This was looking for inherited variation. We will now look at variation from a tumour sample. Here we are focusing on looking for acquired variation.

The somatic variant bioinformatics pipeline is on the cluster with the other pipelines.

Working on the Apple computer and using the Trainee account:

* Open the file 190109\_18M05126.vcf. This contains the variants called by the somatic enrichment pipeline within the sample 18M05126.

There are several hundred variants in the vcf. Many of these will not be related to the reason for the patient’s referral. To further tailor the variants we look at to the referral reason, we can use a panel of genes. These genes will be selected as they are known to have potential involvement in the type of cancer the patient has (e.g. biomarkers of this type of cancer)

* Open the text file 18M05126-OS\_Colorectal\_VariantReport.txt. This contains variants from the sample 18M05126 in the genomic regions known to be associated with colorectal cancer.
* Find the variant 12:25398281C>T in the file. Load the bam file 190901\_18M05126-OS.bam into IGV and find this variant.
  + What is the allele frequency of the variant? (Hint- this can be found both in the file and through IGV)
  + This is a somatic variant. What allele frequencies would you expect to see for germline variants?

Somatic variants are sometimes found at a very low frequency compared to germline variants.

1. Can you use a variant caller designed for calling germline variants to reliably detect somatic variants?
2. What might happen to low frequency variants when a germline variant caller is used?

The ACGS guidelines that are followed for germline samples are not as useful for making decisions regarding acquired variation. The US Association for Molecular Pathology published some guidelines in 2017, but there are currently no national guidelines available for interpretation of somatic variants. To aid with reporting while national guidelines are being developed, an in-house classification system and Standard Operating Procedure (SOP) has been developed for use when interpreting somatic variants.

The focus is on actionability of the variant, so is there a potential treatment (or clinical trial) available, does the variant tell us anything about the prognosis for the patient or is it diagnostic, or does the variant detected have the potential to be an inherited change (rather than only in tumour cells) and so have implications for the family as well. There are four criteria (instead of 5 as we have seen in the germline variant interpretation guidelines):

* Tier I: Variants of strong clinical significance
* Tier II: Variants of potential clinical significance
* Tier III: Variants of unknown clinical significance
* Tier IV: Benign or likely benign variants

Tier I and II class variants are considered actionable.

As with interpreting the clinical significance of germline variants, databases of information are useful evidence.

* An example of a widely used database of somatic variants is called COSMIC. Look up the variant 12:25398281C>T in the COSMIC database. (Hint- you may need to use some of the annotations in the file to get the correct input to the COSMIC database).
* Look up the variant in the MyCancerGenome database.

1. Is the variant found in the databases?
2. Is there any information on how patients with this variant might respond to drug treatments?

### Quality control

In addition to using sequencing data to identify variation, we also use it to make an assessment of the quality of the data we have obtained from our sequencing. This is important, as with low quality data we will not be confident that the variants we call from our samples are accurate.

* Follow the quality control process along and make a few notes on the main quality checks done and why they are important. This includes:
  + Checking the quality of all of the sequencing on a run
  + Checking the quality of each individual sample on the run individually

## Resources

<http://www.acgs.uk.com/media/1025075/ngs_bioinformatics_bpg_final_version_2016.pdf>

<https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5852328/>

<https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4493402/>

<https://jmd.amjpathol.org/article/S1525-1578(16)30223-9/pdf>

<https://cancer.sanger.ac.uk/cosmic>

<https://www.mycancergenome.org/>

190109\_18M05126.vcf

18M05126-OS\_Colorectal\_VariantReport.txt

190901\_18M05126-OS.bam

190901\_18M05126-OS.bai