# CANCER GENOME WORKFLOW

This document aims to bring a structure to efforts to design and implementation of a software package workflow that will handle whole genome sequence data obtained from tumor - normal sample pairs.

Different platforms may use the workflow:

- NGI Stockholm

- NGI Uppsala

- Clinical Genomics Facility

- WABI

- BILS

I tried to compartmentalize the tasks, listed here. This document will evolve during the process.

## WORKFLOW DESIGN

This can be quite cumbersome at this point since NGI is considering purchasing a commercial software package equipped with a fully functional cancer workflow. However the clinical annotation of the variants are not included in all commercial packages. Therefore we would benefit designing a modular workflow allowing **multiple start points**.

The workflow should be able to handle tumor-normal and tumor only samples as we plan to sequence FFPE tumor samples. The **modularity** here is also quite relevant since many procedures are common to these options.

One or a set of software packages can be used for each step in the workflow. The selection procedure of these packages will require careful consideration and maybe benchmarking.

An integrative approach where calls from several packaged are merged seems to be accepted well (http://bcb.io/2015/03/05/cancerval/).

## Pipelining tools

ClusterFlow (<http://clusterflow.io/>) is worth looking into for constructing a workflow. Phil Ewels (NGI Stockholm) has written it and actively using and developing it. The current version does not have the modules needed for this project but they can be added fairly easily.

Please suggest other tools

## WORKFLOW CONTENT

### Quality Control

It is essential to perform a quality check of fastq files before mapping. Even though we might choose not to eliminate samples based on QC, they can be used to resolve issues encountered after processing.

FASTQC, QC3 (<http://www.sciencedirect.com/science/article/pii/S0888754314000354> ?sceptical but it could be useful..)

### Read Mapping

Bowtie 2 using decoy sequencing seems to be the most sensible option.

Mapping to which genome assembly is still open to question.

What's new in GRCh38? (taken from <http://hgdownload.cse.ucsc.edu/gbdb/hg38/html/description.html> )

**Alternate sequences** - Several human chromosomal regions exhibit sufficient variability to prevent adequate representation by a single sequence. To address this, the GRCh38 assembly provides alternate sequence for selected variant regions through the inclusion of alternate loci scaffolds (or alt loci). Alt loci are separate accessioned sequences that are aligned to reference chromosomes. This assembly contains 261 alt loci, many of which are associated with the LRC/KIR area of chr19 and the MHC region on chr6.

**Centromere representation** - Debuting in this release, the large megabase-sized gaps that were previously used to represent centromeric regions in human assemblies have been replaced by sequences from centromere models created by Karen Miga et al. using centromere databases developed during her work in the Willard lab at Duke University and analysis software developed while working in the Kent lab at UCSC. The models, which provide the approximate repeat number and order for each centromere, will be useful for read mapping and variation studies.

**Mitochondrial genome** - The mitochondrial reference sequence included in the GRCh38 assembly and hg38 Genome Browser (termed "chrM" in the browser) is the Revised Cambridge Reference Sequence (rCRS) from MITOMAP with GenBank accession number J01415.2 and RefSeq accession number NC\_012920.1. This differs from the chrM sequence (RefSeq accession number NC\_001907) used by the previous hg19 Genome Browser, which was not updated when the GRCh37 assembly later transitioned to the new version.

**Sequence updates** - Several erroneous bases and misassembled regions in GRCh37 have been corrected in the GRCh38 assembly, and more than 100 gaps have been filled or reduced. Much of the data used to improve the reference sequence was obtained from other genome sequencing and analysis projects, such as the 1000 Genomes Project.

**Analysis set** - The GRCh38 assembly offers an "analysis set" that was created to accommodate next generation sequencing read alignment pipelines. Several GRCh38 regions have been eliminated from this set to improve read mapping. The analysis set may be downloaded from the Genome Browser downloads page.

(<http://hgdownload.cse.ucsc.edu/goldenPath/hg38/bigZips/analysisSet/> )

All in all, it just makes sense to map to hg38 since it is a much improved reference. But when we annotate our variants against datasets created using hg19, this will require lift-over.

### Variant Calling

Since there is no perfect caller for any type of variant, it seems like the accepted practice is to use multiple tools and merge the variants.

* Choice of callers
* How to merge variants? For example, do we also filter variants found only by one or should we develop a scoring system??
* Version control can be time-consuming..
* Output format compatibilities – hopefully that shouldn’t be a big problem for SNVs and INDELs, but might create problems for SV and CNV callers since they tend to report variants in any way they like…
* Computing resources and temporary storage could be a problem…
* Implementation can be tricky as well, possibly requiring complex queuing schemes…
* The variables (e.g. coverage thresholds for tumor and normal) for each software should be comparable

For a list of structural variant callers, please see

<http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4394692/>

WHAM is not included in above.

<https://www.biostars.org/p/128140/>

### Variant Annotation

To be discussed…

## WORKFLOW IMPLEMENTATION

Here is a rough schematic of a possible workflow for matched tumor samples. Timeplan of the implementation is to be discussed to be able to synchronize the efforts.

There are also additional responsibilities and tasks such as

* GitHub and branching modules…
* Keeping track of project process, Trello could be a good platform for this and regular-ish meetings
* Managing the pipeline implementation. Pipelining tool, scheduling, identifying possible deadlocks…



## ADDITIONAL CONSIDERATIONS

### Samples

To be able to make our lives easier at the analysis step, we could consider preparing a document to be given to the labs providing us the samples

* Requirements on FFPE sample quality
* Narrow insert size distribution (vital for better calls)
* Tumor heterogeneity measure
* Procedures that might introduce biases (such as the one Oslo team was talking about..)
* ??

### Possible additions

This is a sample based analysis workflow. We could have higher level of analyses where multiple samples of same type could be analysed together. An example is

<http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3570658/>