BMEG3105 (24/25 fall) - Data analytics for personalized genomics and precision medicine | Week 9 Long Lecture: Cancer genomics overview and genomics analysis

Lecturer: Yu LI (李煜) from CSE | liyu95.com, liyu@cse.cuhk.edu.hk | Wednesday, Oct 30th 2024

Scribe: Yunwen ZHANG (1155173743)

> ⚠️ Spoiler: this session will be intensive, practical, and much on the processing of real-life extremely imperfect data, so feel free to interrupt so that we need not waste time going through the content again after.

1. From last lecture

1-1. Underfitting and overfitting

- Your model has an `underfitting` issue when your training data show `relationships` that are `more complicate` than your model. Your model has an `overfitting` issue when the opposite is the case.
- Either case, your model performs `poorly on testing data / additional data`.
- In practice, we should overfit the data first and reduce the complexity of our model to find the best fit.

1-2. Multi-omics overview and analysis

- Multi-omics links your genome, epigenome, transcriptome, proteome, metabolome `all to the phenome`, and accounts for interactions (like the feedback modulation of gene transcription by expressed proteins) that do not fit into the intuitive perception of a unidirectional phenotype production line.
- Analysis of multi-omics data can be tedious, but the `key techniques are the same` as what we covered in module 1. An additional one is `statistical testing`: student t-test.
    - Example where statistical testing is used: differential gene expression analysis. Key point is that here `both the position and the spread` should be considered.
    - The `formula` for t-test `varies` across situations, but the procedure is the same.

2. Wrap-up for last lecture

2-1. Another example of statistical testing: `gene enrichment analysis`

- Background: a `pathway` is a series of intermolecular interactions that together `produce an effect`, be it assembly of a biomolecule, a change in the expression level of a gene, or initiation of cell movement. `A set of genes` work together in a pathway.
    - Just like there are databases of gene sequences, there are also databases of pathways. Check out KEGG PATHWAY Database.
    - Examples of pathways include metabolic pathways (eg. pentose phosphate pathway), signaling pathways (eg. adenylyl cyclase-cAMP pathway), regulatory pathways (eg. p53 pathway), cell cycle pathway, etc.
- Why is it called *gene enrichment analysis*?
    - Enrichment here means those genes are overrepresented, which happens often in diseases. For genes in the same pathway which is linked to some disease condition, they are likely overrepresented together in the patient of that illness.
- Problem: experimentally 213 genes have been identified to be associated with type II diabetes. Then how to identify the pathways related to type II diabetes?
- Solution: test the `associations` - whether the overlap between the pathway genes and the disease-associated genes is statistically significant.
    - Logic:
        - We want both [number of genes associated with both that pathway and type II diabetes] and [number of genes not associated to that pathway nor type II diabetes] be *high*;
        - We want [number of genes associated with only one of that pathway and type II diabetes not the other] be *low*.
    - Implementing: use the confusion matrix / contingency matrix:

	type II diabetes yes	type II diabetes no	total #
pathway yes	a	b	$a+b$
pathway no	c	d	$c+d$
total #	$a+c$	$b+d$	$a+b+c+d$

      where a, d should be *high*, b, c should be *low*.

    - How to determine the high and low standards? We need a quantitative approach → Fisher's exact test

2-2. Fisher's exact test

: a statistical significance test used in contingency tables.

- Why is it called *exact*? Because the `p-value` could be calculated from the table `directly`, in contrast to the *inexact* t-test where the p-value is *inferred* by a t-statistic.
- Formula: $p = \frac{\binom{a+b}{a}\binom{c+d}{c}}{\binom{a+b+c+d}{a+c}} = \frac{(a+b)!(c+d)!(a+c)!(b+d)!}{a!b!c!d!(a+b+c+d)!}$
- Same criteria: if $p > 0.05$, the association is not significant, i.e., that pathway is not related to type II diabetes.

3. Today's topics: cancer brief overview and relevant data analysis techniques

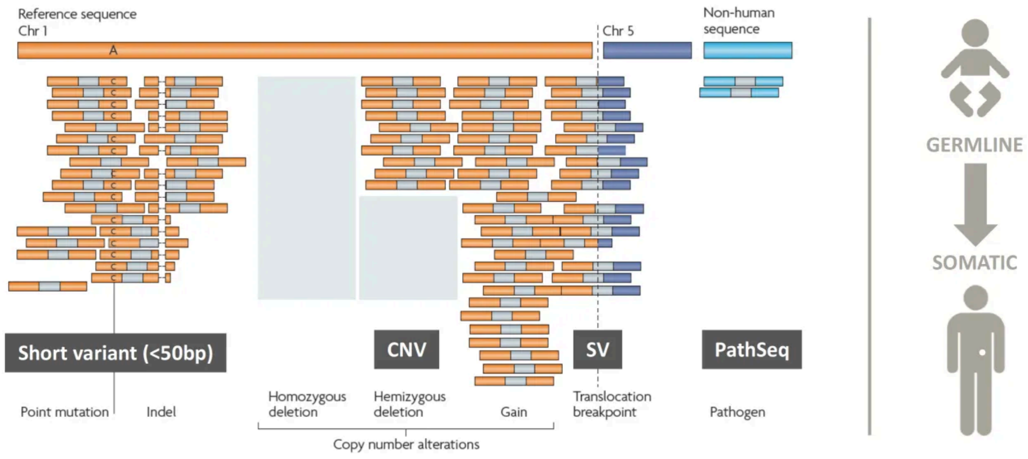
3-1. Cancer overview

- Cancer = `uncontrolled cell growth` + `spread` of such cells to other parts of the body.
- The cause of cancer is still unclear, but generally it is not external invasion; cancer is assumed to be a genomic disease. Studying a `genomic disease` requires studying not only the genome but also its expression.
- `Epigenome`: the variabilities about the chemical `modifications` on the DNA or histone that modulate the `binding and releasing of histone` and hence whether the gene gets transcribed - thus epigenome is something between genome and transcriptome.

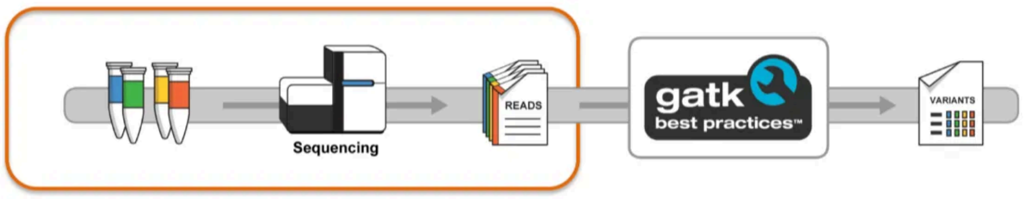
3-2. Studying cancer multi-omics

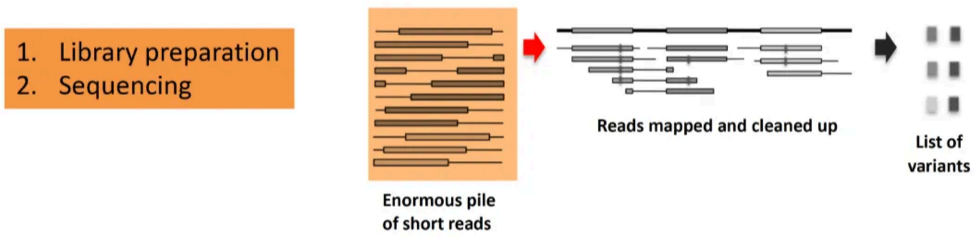
3-2-1. Studying cancer genome || variant calling and GWAS

- Variant calling = identifying genetic variations / mutations in cancer genomes compared to normal genomes.
  - Why care about variants? As any 2 humans share 99.5% of their DNA sequence, (the genome of) a human can be described efficiently in their genome variance from a reference.
  - Different types of genomic variants

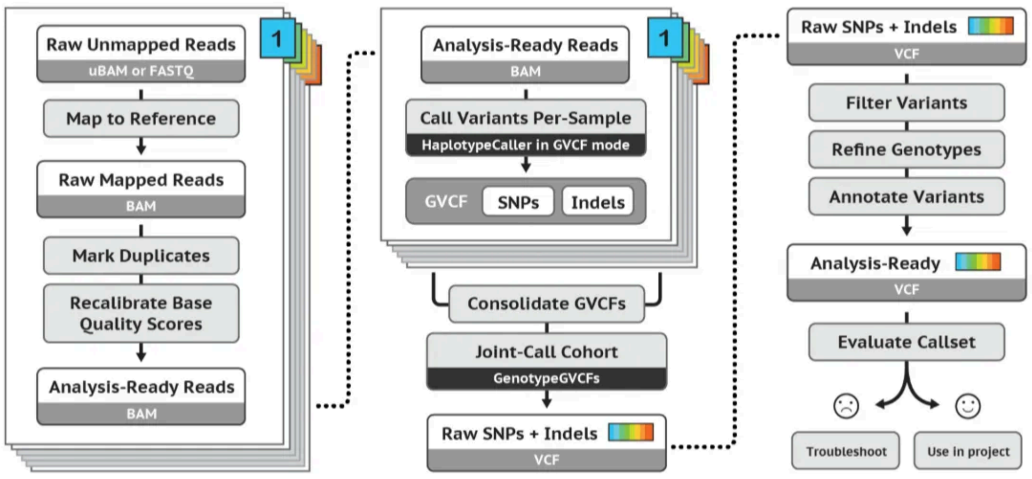


    - Copy number alteration = the number of same copy of the same gene changed within your chromosome. Larger effect than point mutation and indel.
      - Detecting a structural variance: if in gene browser, at a site many reads differ from each other, then there is possibly a structural variance.
    - Detecting point mutation: mapping the person's sequences to the reference.
    - Indel = insertion or deletion. Has larger effect than point mutation.
    - Germline vs somatic mutations: the former happen in eg. germ cells / zygote. The latter contributes to most cancer.
  - Gene variant detection pipeline



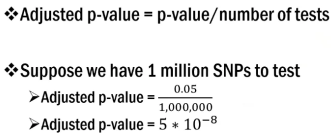


    - After getting the reads: mapping
      - Mapping algorithms can be imperfect. They may recognize a error as a variation.
      - `variants` (real change) vs. `errors` (artifacts)
        - Sources of error: PCR, sequencing (in case of long sequences), alignment (algorithm approximations to compromise accuracy for speed), variant calling (samples few), genotyping
          - `Coverage` / `depth` of a site / region = number of reads mapped to it. That is where the terms "shallow / deep" sequencing come from.
    - Data pre-processing
      - Step 1: mapping and alignment + quality control
        - Before mapping: `FASTQ` file → if low score, discard the sequence.
        - After mapping: `SAM / BAM` file,
          - The `header` contains the meta-information for all reads.
          - The `records` contains structured read information: read name, flags, position, MAPQ, CIGAR, read sequence, quality score and metadata.
            - `CIGAR` = Concise Idiosyncratic Gapped Alignment Report, summary of whether there is a match / mismatch / insertion / deletion etc., and how many if any.
      - Step 2: removing duplicates
        - `Duplicates` are *non-independent* measurements.
          - Can think of they seem to occupy a dimension while contributing nothing to the dimensions (which will affect the number of reads being matched and thus cause `over-confidence of variant callling`).
          - Examples include duplicated data when merging databases, library duplicated occurred in PCR, etc.
        - How to solve it: remove the duplicate.
  - Variant calling pipeline in detail



    - In short: mapping → quality control → per-sample variant calling → family / population / cohort variant calling → further downstream analysis and cleaning → reaching the variant calling result.
      - After per-sample the calling, the output is saved in a `VCF` file (again "header (defining the 5 last columns in "record") + records format").
      - Trough joint (-genome) analysis we obtain family / population / cohort variant calling. The output file is the `final multi-sample VCF` file.
      - The calling quality (statistical confidence) is high in joint calling than the per-sample ones.

- Further downstream analysis after variant calling
    - Linking mutation to the disease: genome-wide association study (GQAS) (statistical testing) - to check all mutations related to a disease within the whole genome.
        - The problem: we need a smaller p value as threshold here.
            - p = 0.05 means there is 5% possibility that your conclusion is incorrect.
            - With a whole genome at hand, we are dealing with 3.5 million SNPs. There are millions of *testings* to do. 0.05 probability is too large a number of incorrect results to tolerate.
                - Do the math: the probability that at least 1 result is incorrect: $1-(1-0.05)^{3\text{ million}} \approx 1$ Thus, statistical testing is no longer valid.
            - i.e. p = 0.05 doesn't work for GWAS.
        - Solution: the Bonferroni correction: adjusting p-value for GWAS



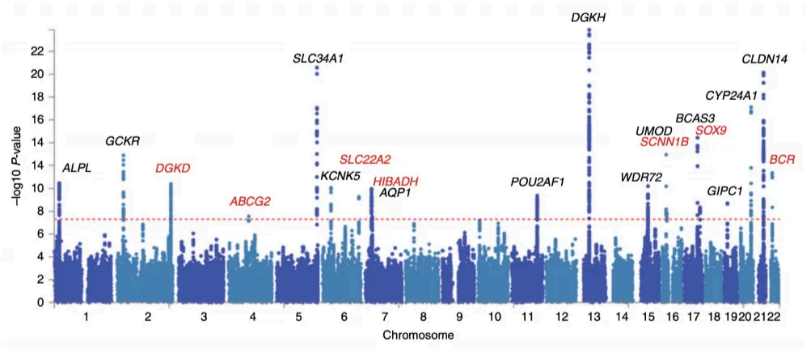
❖Adjusted p-value = p-value/number of tests

❖Suppose we have 1 million SNPs to test

➢Adjusted p-value = $\frac{0.05}{1{,}000{,}000}$

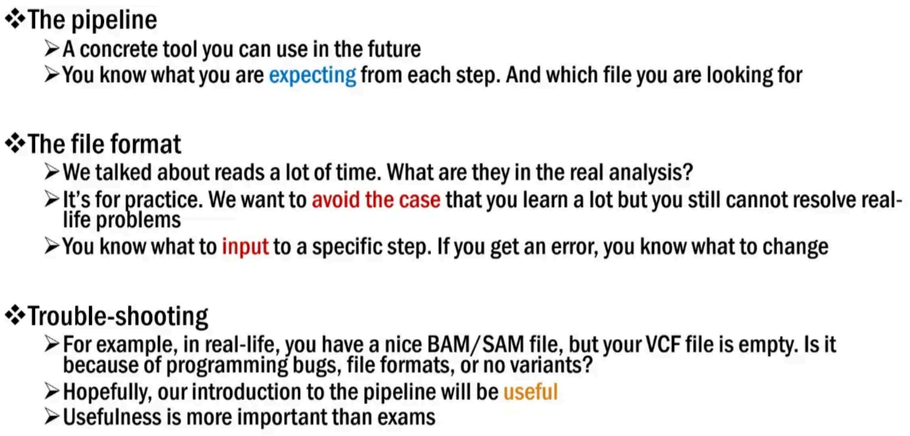
➢Adjusted p-value = $5 * 10^{-8}$

        - In actual studies: only select the results that exceed the line, i.e. their p-value are lower than that small negative power of 10.



Recap:

1. what the lecture can hopefully give you



❖The pipeline

➢A concrete tool you can use in the future

➢You know what you are expecting from each step. And which file you are looking for

❖The file format

➢We talked about reads a lot of time. What are they in the real analysis?

➢It's for practice. We want to avoid the case that you learn a lot but you still cannot resolve real-life problems

➢You know what to input to a specific step. If you get an error, you know what to change

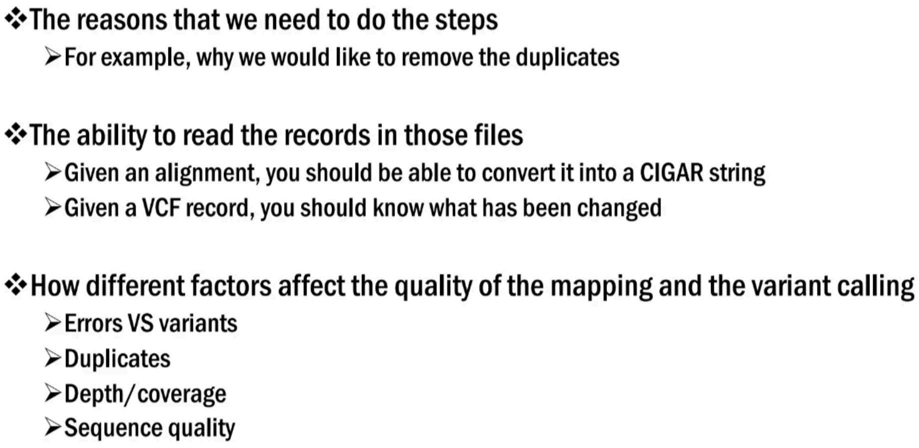
❖Trouble-shooting

➢For example, in real-life, you have a nice BAM/SAM file, but your VCF file is empty. Is it because of programming bugs, file formats, or no variants?

➢Hopefully, our introduction to the pipeline will be useful

➢Usefulness is more important than exams

2. what you are expected to know from this part (will be covered in the final)



❖The reasons that we need to do the steps

➢For example, why we would like to remove the duplicates

❖The ability to read the records in those files

➢Given an alignment, you should be able to convert it into a CIGAR string

➢Given a VCF record, you should know what has been changed

❖How different factors affect the quality of the mapping and the variant calling

➢Errors VS variants

➢Duplicates

➢Depth/coverage

➢Sequence quality

Quiz: which of the following will affect the variant calling step?

☐ A. FASTQ file format

☑ ~~B. Preprocessing quality control~~

☐ C. GWAS corrected p-value

☐ D. VCF file format

Next lecture: structural variant calling, RNA-Seq, and epigenome peak calling

All figures are adapted from Prof. Yu LI's lecture notes.