Introduction

Estudios in silico en Biomedicina

(Máster en Bioinformática, Universidad de Valencia)

David Montaner

2014-10-07

Course Overview

NGS Data Analysis | Microarrays? | Functional Genomics . . .

- 12 sessions
- computer room
- practical (toy examples)

Course Assessment

- course work (practical)
- final exam (concepts)

Most common NGS applications

RNA

- RNA-seq / Transcriptomics
  - Quantitative: genes, miRNAs, small RNA, transcription factors ...
  - Descriptive: presence / absence
  - Alternative splicing
  - *variant calling*
- Metatranscriptomics

DNA

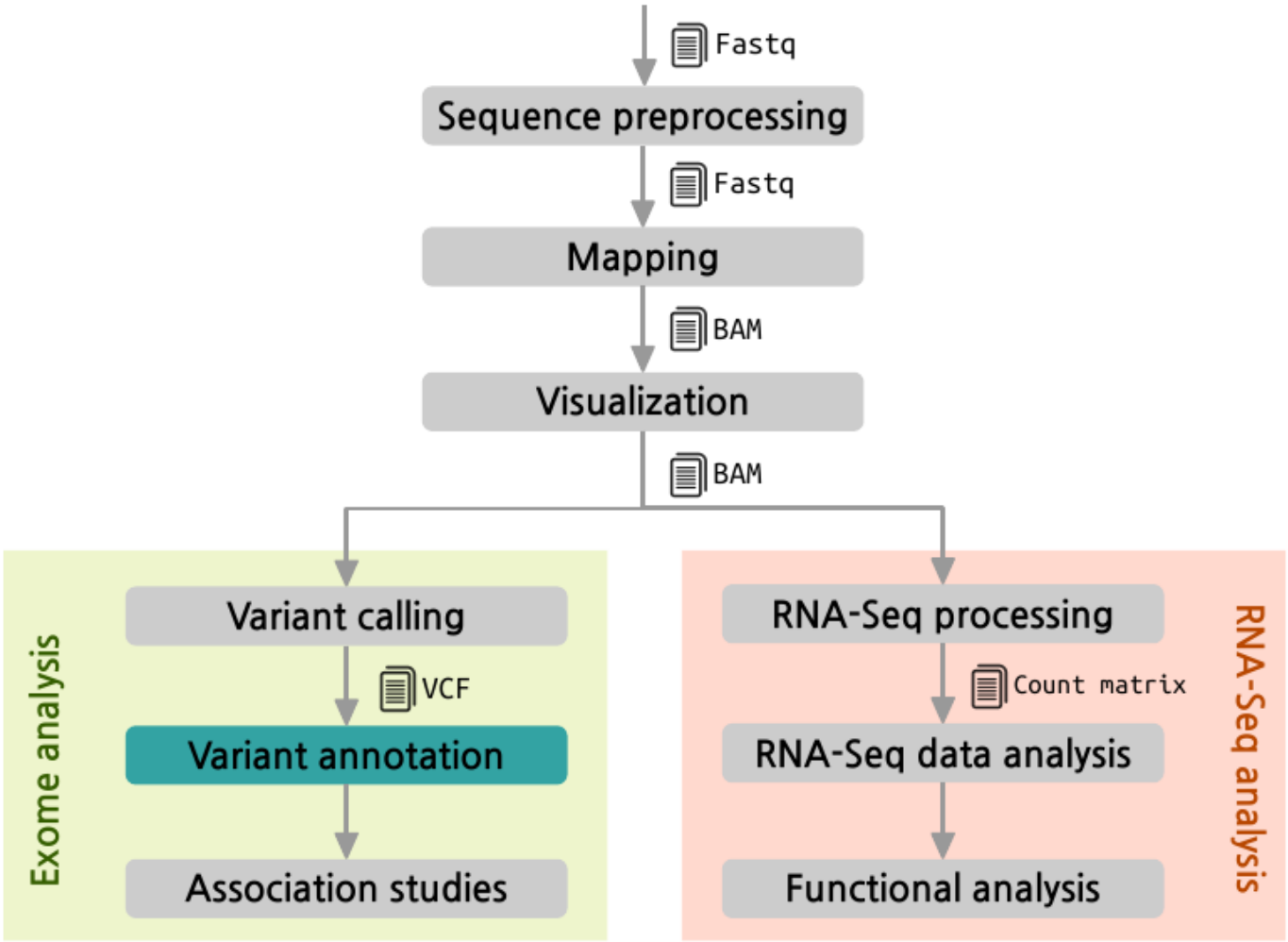
- *De novo* sequencing
- Resequencing : Mutation or variant calling
  - Whole genome
  - Exome
  - Targeted sequencing
- Copy number alterations
- ChIP-seq / Epigenomics
  - Protein-DNA interactions
  - Active transcription factor binding sites
  - Histone methylation
  - CpG island methylation
- Metagenomics
  - 16S, 18S, viruses
  - full bacterial genomes

General considerations

- Genomic material (DNA/RNA) extraction:
  - amplification
  - copy or replication (complementary reverse)
  - multiplexing
  - adapters
  - capture using primers
- NGS *reads* are (ideally) a **random** selection of the purified DNA/RNA molecules prepared in laboratory ...
  Think about your experimental context

- Reference "genome" is always useful (if exists)
  - map reads against the reference "genome" (transcriptome or database)
  - assembly reads into a reference "set" of sequences
  - assembly based in a *close* reference
- Adapters or primers may be present in our data
- Paired-end or single-end
- Advantages and disadvantages over DNA **microarrays**

General Analysis Pipeline



Data processing steps: common

- File parsing: proprietary formats (`sff`, `fasta + qual...`) to `fastq`
- Split multiplexed samples
- Quality Control of the raw data
- Adapter trimming
- Filtering and trimming reads by quality
- Quality Control of the trimmed and filtered reads
- Prepare a reference
  - download genome, miRNAs... form database
  - assemble
  - **index the reference**
- Alignment / Map against the reference
- Quality Control of the mapping
- Visualization of the mapping

Data processing steps: specific

Transcriptomics

- Gene, transcript, isoform ... quantification or detection
- Gene, transcript, isoform ... discovery

Genomics

- Variant calling: SNPs, InDels
- Copy number estimation
- Annotation

Analysis

- Statistical analysis
- Functional interpretation
  - GSA
  - Pathways analysis
  - Protein networks

Statistical analysis

Transcriptomics

- Starts with a data matrix of *continuous* expression levels
- Usually in a tab delimited file
- Gene filtering
- Differential expression analysis
- Clustering ...

Genomics

- Starts with a data matrix of **discrete** variant calls
- Usually in a **VCF** file format
- Variant filtering or prioritization: SIFT / PolyPhen, data bases ...
- Association analysis
- Principal component analysis ...

File formats:

- Fasta : sequence (reference)
- FastQ : sequence + quality (**raw reads**)
- SAM / BAM : Sequence Alignment Map format
- VCF : Variant Call Format
- PED & MAP : pedigree files (variants + phenotype
- Tab separated files : matrices / data frames
- Annotation formats (annotation of the reference)
  - BED : annotation for genomic regions
  - GFF : General Feature Format
  - GTF : Gene Transfer Format

Some Remarks about formats

- Different types of compression and *indexing* may be necessary or useful to handle data files.
- Text file in tabular formats are always used
- All standards are not consensus ones they are **de facto** accepted

See more file formats at
`http://genome.ucsc.edu/FAQ/FAQformat.html`

File formats: Fasta & FastQ

Fasta

```
>SEQ_ID
GATTTGGGGTTCAAAGCAGTATCGATCAAATAGTAAATCC
```

FastQ

```
@SEQ_ID
GATTTGGGGTTCAAAGCAGTATCGATCAAATAGTAAATCC
+
!''*((((***+))%%%++)(%%%%).1***-+*''))**
```

File formats: SAM (BAM)

TAB-delimited file with an optional **header section** and an **alignment section**

```
@HD VN:1.5 SO:coordinate
@SQ SN:ref LN:45
r001  163 ref  7 30 8M2I4M1D3M = 37  39 TTAGATAAAGGATACTG *
r002    0 ref  9 30 3S6M1P1I4M *  0   0 AAAAGATAAGGATA    *
r003    0 ref  9 30 5S6M       *  0   0 GCCTAAGCTAA       * SA:Z:ref,29,-,6H5M,17,0;
r004    0 ref 16 30 6M14N5M    *  0   0 ATAGCTTCAGC       *
r003 2064 ref 29 17 6H5M       *  0   0 TAGGC             * SA:Z:ref,9,+,5S6M,30,1;
r001   83 ref 37 30 9M         =  7 -39 CAGCGGCAT         * NM:i:1
```

May be compressed and indexed using `samtools`

File formats: SAM (BAM)

The alignment of each read is described in a row of the file by **11 fields**:

1. QNAME: read name
2. FLAG: bitwise flag
3. RNAME: chromosome
4. POS: leftmost genomic position
5. MAPQ: mapping quality ([Phred scale][phred-quality-score-wikipedia])
6. CIGAR: CIGAR string (gaps, clipping)
7. RNEXT: paired read name
8. PNEXT: paired read position
9. TLEN: total length of template
10. SEQ: read base sequence
11. QUAL: read base quality

File formats: VCF

Tab delimited text file with a **header section**

```
##fileformat=VCFv4.2
##fileDate=20090805
##source=myImputationProgramV3.1
##reference=file:///seq/references/1000GenomesPilot-NCBI36.fasta
##contig=<ID=20,length=62435964,assembly=B36,md5=f126cdf8a6e0c7f379d618ff66beb2da,species="Homo sapiens",taxonomy=x>
##phasing=partial
##INFO=<ID=NS,Number=1,Type=Integer,Description="Number of Samples With Data">
##INFO=<ID=DP,Number=1,Type=Integer,Description="Total Depth">
##INFO=<ID=AF,Number=A,Type=Float,Description="Allele Frequency">
##INFO=<ID=AA,Number=1,Type=String,Description="Ancestral Allele">
##INFO=<ID=DB,Number=0,Type=Flag,Description="dbSNP membership, build 129">
##INFO=<ID=H2,Number=0,Type=Flag,Description="HapMap2 membership">
##FILTER=<ID=q10,Description="Quality below 10">
##FILTER=<ID=s50,Description="Less than 50% of samples have data">
##FORMAT=<ID=GT,Number=1,Type=String,Description="Genotype">
##FORMAT=<ID=GQ,Number=1,Type=Integer,Description="Genotype Quality">
##FORMAT=<ID=DP,Number=1,Type=Integer,Description="Read Depth">
##FORMAT=<ID=HQ,Number=2,Type=Integer,Description="Haplotype Quality">
#CHROM POS     ID        REF    ALT     QUAL FILTER INFO                              FORMAT      NA00001        NA00002        NA00003
20     14370   rs6054257 G      A       29   PASS   NS=3;DP=14;AF=0.5;DB;H2           GT:GQ:DP:HQ 0|0:48:1:51,51 1|0:48:8:51,51 1/1:43:5:.,.
20     17330   .         T      A       3    q10    NS=3;DP=11;AF=0.017               GT:GQ:DP:HQ 0|0:49:3:58,50 0|1:3:5:65,3   0/0:41:3
20     1110696 rs6040355 A      G,T     67   PASS   NS=2;DP=10;AF=0.333,0.667;AA=T;DB GT:GQ:DP:HQ 1|2:21:6:23,27 2|1:2:0:18,2   2/2:35:4
20     1230237 .         T      .       47   PASS   NS=3;DP=13;AA=T                   GT:GQ:DP:HQ 0|0:54:7:56,60 0|0:48:4:51,51 0/0:61:2
20     1234567 microsat1 GTC    G,GTCT  50   PASS   NS=3;DP=9;AA=G                    GT:GQ:DP    0/1:35:4       0/2:17:2       1/1:40:3
```

May be compressed and indexed using `tabix`

File formats: VCF

Each variant is described by **8 fields**

1. CHROM: chromosome
2. POS: position
3. ID: name
4. REF: reference base(s)
5. ALT: non-reference alleles
6. QUAL: quality score of the calls (phred scale)
7. FILTER: PASS / filtering_tag
8. INFO: additional information

Genotype data for several samples may be included in a batch of additional columns (one for each sample) preceded by a FORMAT column which describes their format.

File formats: VCF INFO column

May include several semicolon separated fields containing information about the variants coded in key value style:

`<key>=<data>[,data]`

Some reserved (but optional) keys are:

- AA ancestral allele
- AC allele count in genotypes, for each ALT allele, in the same order as listed
- AF allele frequency
- CIGAR cigar string describing how to align an alternate allele to the reference allele
- DB dbSNP membership
- MQ RMS mapping quality, e.g. MQ=52
- MQ0 Number of MAPQ == 0 reads covering this record

File formats: PED & MAP

Classic format to represent genomic variants for several individuals

<---- normal.ped ---->

```
1 1 0 0 1  1  A A  G T
2 1 0 0 1  1  A C  T G
3 1 0 0 1  1  C C  G G
4 1 0 0 1  2  A C  T T
5 1 0 0 1  2  C C  G T
6 1 0 0 1  2  C C  T T
```

<--- normal.map --->

```
1  snp1   0  5000650
1  snp2   0  5000830
```

would be represented as TPED/TFAM files:

<------------- trans.tped ------------->

```
1 snp1 0 5000650 A A A C C C A C C C C C
1 snp2 0 5000830 G T G T G G T T G T T T
```

<- trans.tfam ->

```
1  1  0  0  1  1
2  1  0  0  1  1
3  1  0  0  1  1
4  1  0  0  1  2
5  1  0  0  1  2
6  1  0  0  1  2
```

Some variants of the format are described depending on the software used to read or write them. Those variants may include *transposed* versions of the format which is closer to standard *genomic* representation of this kind of information.

File formats: PED & MAP

PED file

1. Family ID
2. Individual ID
3. Paternal ID
4. Maternal ID
5. Sex (1=male; 2=female; other=unknown)
6. Phenotype (1=unaffected; 2=affected; 0 missing; -9=missing)
7. . . . genotypes . . .

MAP file

1. chromosome (1-22, X, Y or 0 if unplaced)
2. rs. . . or SNP identifier
3. Genetic distance (Morgans)
4. Base-pair position (bp units)

NGS Data Analysis Software I

- FastQC : A quality control tool for high throughput sequence data.
- cutadapt : A tool that removes adapter sequences from DNA sequencing reads.
- [FASTX Toolkit] : A collection of command line tools for Short-Reads FASTA/FASTQ files preprocessing.

- Bowtie 2 : Tool for aligning sequencing reads to long reference sequences.
- TopHat : A fast splice junction mapper for RNA-Seq reads. Depends on Bowtie or Bowtie 2.
- BWA (Burrows-Wheeler Aligner) : A software package for mapping low-divergent sequences against a large reference genome.

NGS Data Analysis Software II

- SAMtools : Tool for aligning sequencing reads to long reference sequences.
- Picard : Java-based command-line utilities to manipulate SAM files.
- VCFtools : A package for working with VCF files: merging, comparing, annotating ...
- tabix : tabix: compress and index TAB-delimited files. Useful for handling GFF, GTF, BED and VCF files.

- Cufflinks : Transcript assembly, differential expression, and differential regulation for RNA-Seq
- GATK (Genome Analysis Toolkit): A package to analyze next-generation re-sequencing data, primary focused on variant discovery and genotyping.

NGS Data Analysis Software III

- [CuffDiff] : Transcript assembly, differential expression, and differential regulation for RNA-Seq
- PLINK : whole genome association analysis.
- IGV (Integrative Genomics Viewer): a visualization tool for interactive exploration of large, integrated genomic datasets.
- DWGSIM: simulate datasets

NGS Software Installation

Generally download binaries ...

```
./configure
```

```
make (install)
```

You may:

- Call directly to the binary or executable files
- Make them accessible via the **PATH** variable of the shell using for instance the `.profile` file in your home: export PATH=/path/to/bin/dir:$PATH

Or use **configuration files** `.profile` (better than `.bashrc`)

Some hints here:
https://github.com/biocosas/ngs_software_installation

Useful Linux commands

- cut : to get some fields from a tabular text file
- wc . to count the number of lines in a file
- grep : to find lines in a file with certain pattern
- md5 : to check that files have been properly downloaded
- head / tail : to "see" fists or last lines in a text file

Execute java programs;

```
java -jar MY_COMPILED_JAVA.jar <program options>
```

Databases

- NCBI web page for reference genomes.
- 1000 genomes project
- Sequence Read Archive (SRA)