Influenza Bioinformatics: Next Generation Sequencing (NGS) I

Dimitar Kenanov
Vithiagaran Gunalan
Sebastian Maurer-Stroh

Bioinformatics Institute, Singapore

Bioinformatics Institute
A*STAR

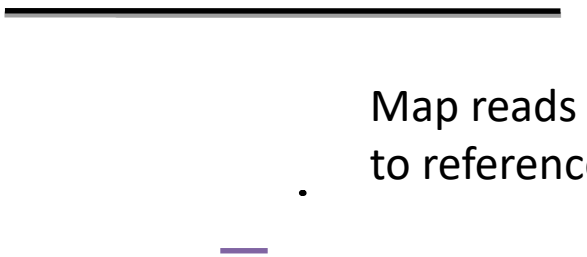
Section I

Influenza Sequence READMAPPING

Bioinformatics Institute
A*STAR

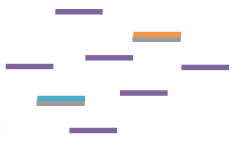
Readmapping

Reference genome

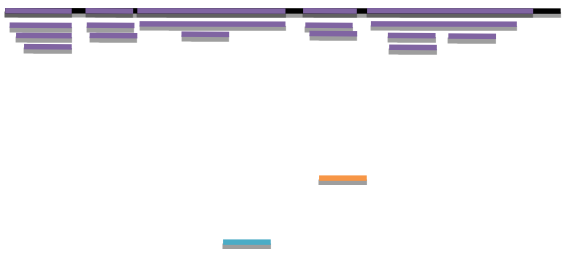


Map reads
to reference

Short reads
(fastq format)



"New" genome as consensus of mapped reads



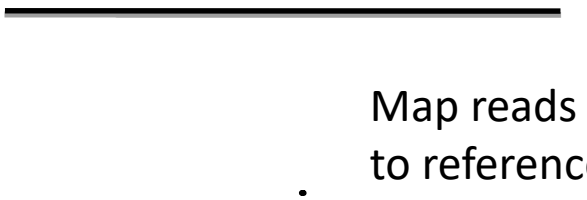
- Very fast
- Good for resequencing of individuals from species with known genome when no structural variation expected
- May not cover all areas equally
- May leave out divergent regions as unmappable reads

Example tools: bwa, bowtie, smalt…

Bioinformatics Institute
A*STAR

Readmapping

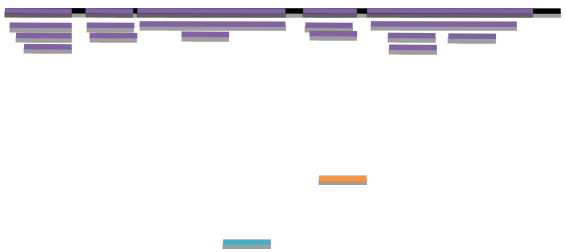
Reference genome



Map reads
to reference

Short reads

"New" genome as consensus of mapped reads



- Very fast
- Good for resequencing of individuals from species with known genome when no structural variation expected
- May not cover all areas equally
- May leave out divergent regions as unmappable reads

De novo assembly

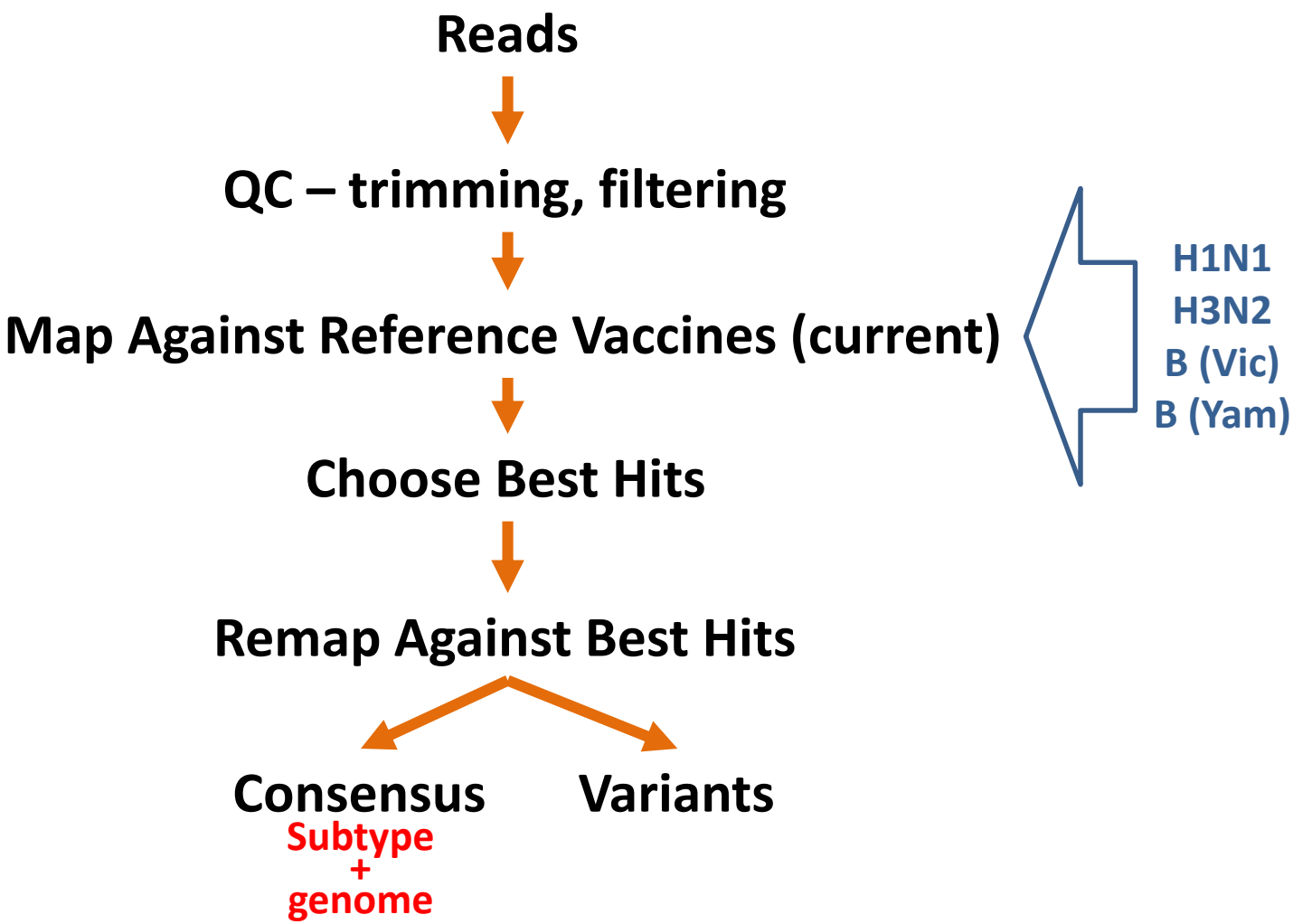
Short reads

Combine/assemble
overlapping reads

- Slow
- Only option if no reference genome available
- Can capture different genome structure
- Typically creates large genome fragments (contigs) but not complete genomes

Bioinformatics Institute
A*STAR

NGS Workflow



Bioinformatics Institute
A*STAR

Software Needed

- QC for raw reads
  - **FASTQC** (www.bioinformatics.babraham.ac.uk/projects/fastqc/)

- Raw read preprocessing
  - **fqtrim** (ccb.jhu.edu/software/fqtrim/)
  - **Trimmomatic** (www.usadellab.org/cms/?page=trimmomatic)

- Assembler
  - **IDBA** (code.google.com/p/hku-idba/downloads/list)
  - **SPAdes** (bioinf.spbau.ru/spades)

- Sequence Aligner
  - **Bowtie2** (github.com/BenLangmead/bowtie2) – read aligner
  - **BWA** (sourceforge.net/projects/bio-bwa/files/) – read aligner
  - **Samtools** (www.htslib.org/doc/samtools.html) – processing alignments
  - **Bedtools** (github.com/arq5x/bedtools) – genome coverage (best hits)

- Alignment Viewer
  - **igv** (software.broadinstitute.org/software/igv/)

Your USB Stick Contains:

- WSHOP2019.zip
- For today's exercise, all preinstalled

3 Directories

FLU_DATA – NGS data

bin – scripts

lib – required modules

2 files:

Install.sh – install

rakudo.deb – installer for Perl 6

Read the README.txt file before installing on your own!!

Today's Exercise

- 3 different samples
  - **Cell-Culture Flu A** (illumina PE)
    - 181_S6_L001_R*_001.fastq
  - **Severe Influenza in Elderly Patient** (illumina PE)
    - A51-INFTT-17-0683_S34_L001_R*_001.fastq
  - **IonTorrent** (IonTorrent SE)
    - IonCode_NS16May2019_AWGS_25pM.fastq

Subtype

+

Consensus Sequences

FluSurver check

+

GISAID Upload

Today's Exercise (con't)

- Open Ubuntu in Windows:

- Navigate to the FLU_DATA directory

  cd /mnt/c/Users/User/Workshop_Flu/FLU_DATA

- Inspect the files:

  ls –ltrh

- To see just FastQ sample files:

  ls –ltrh *.fastq

Today's Exercise (con't)

- Make a sample file

nano samp.txt

- Type in the sample IDs:

181_S6 181S6

A51-INFTT A51

tab-separated!

- Save the sample file

ctrl-o, Enter, then crtl-x

Sample File will be used to configure the script automatically
Pipeline has to be run separately for SE and PE

Today's Exercise (con't)

- Create links for the pipeline, edit config file:

config.pl –s samp.txt –d `pwd` -c FLUAB_VACCREFMIX_PE.conf

- Run the pipeline for the paired-end samples:

run_fluAB.pl FLUAB_VACCREFMIX_PE

Today's Exercise (con't)

- IonTorrent SE

FLU_DATA directory

nano samp.txt

IonCode_0282 IonCode

ctrl-o, Enter, then crtl-x

config.pl –s samp.txt –d `pwd` -c FLUAB_VACCREFMIX_SE.conf

run_fluAB.pl FLUAB_VACCREFMIX_SE

Today's Exercise (con't)

- Result Files:

  181S6_MIXED_GENOME_CONSENSUS.fa

  &

  181S6_FINAL_STATS.txt

- Open these files in Windows (Notepad, Wordpad)

- Look at each segment, which subtype?

- Are all segments accounted for?

Influenza Bioinformatics: Next Generation Sequencing (NGS) II

Co-Infection

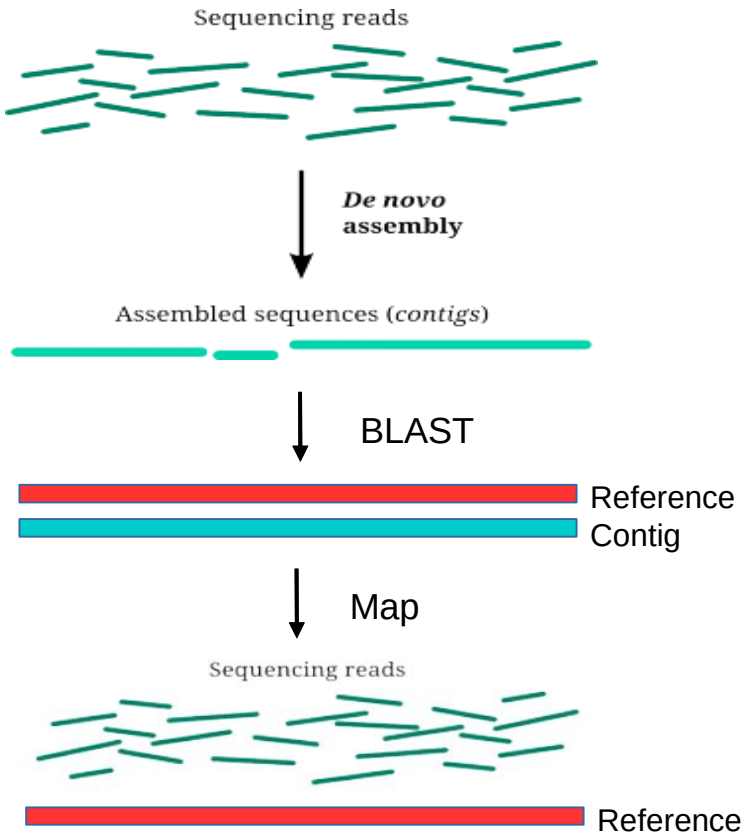
Dimitar Kenanov
Vithiagaran Gunalan
Sebastian Maurer-Stroh

Bioinformatics Institute, Singapore

Bioinformatics Institute
A*STAR

Detection of Influenza strain/s and/or co-infection using Next Generation Sequence data analysis

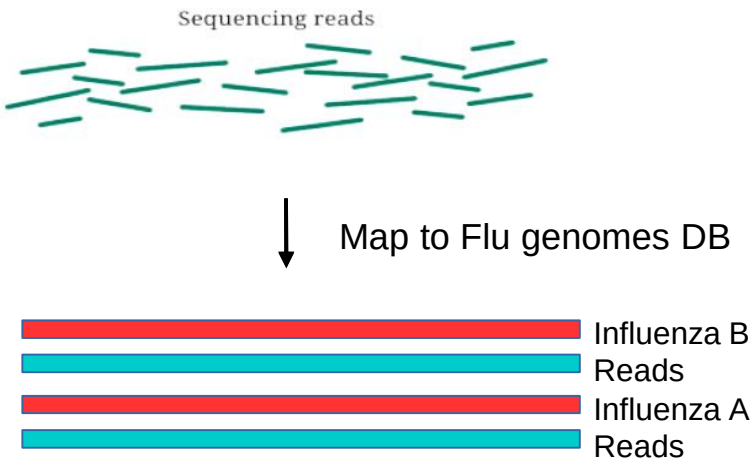
'Classical' approach



Pros:
- method which works

Cons:
- both assembly and BLAST can take long time!
- some contings might be meaningless

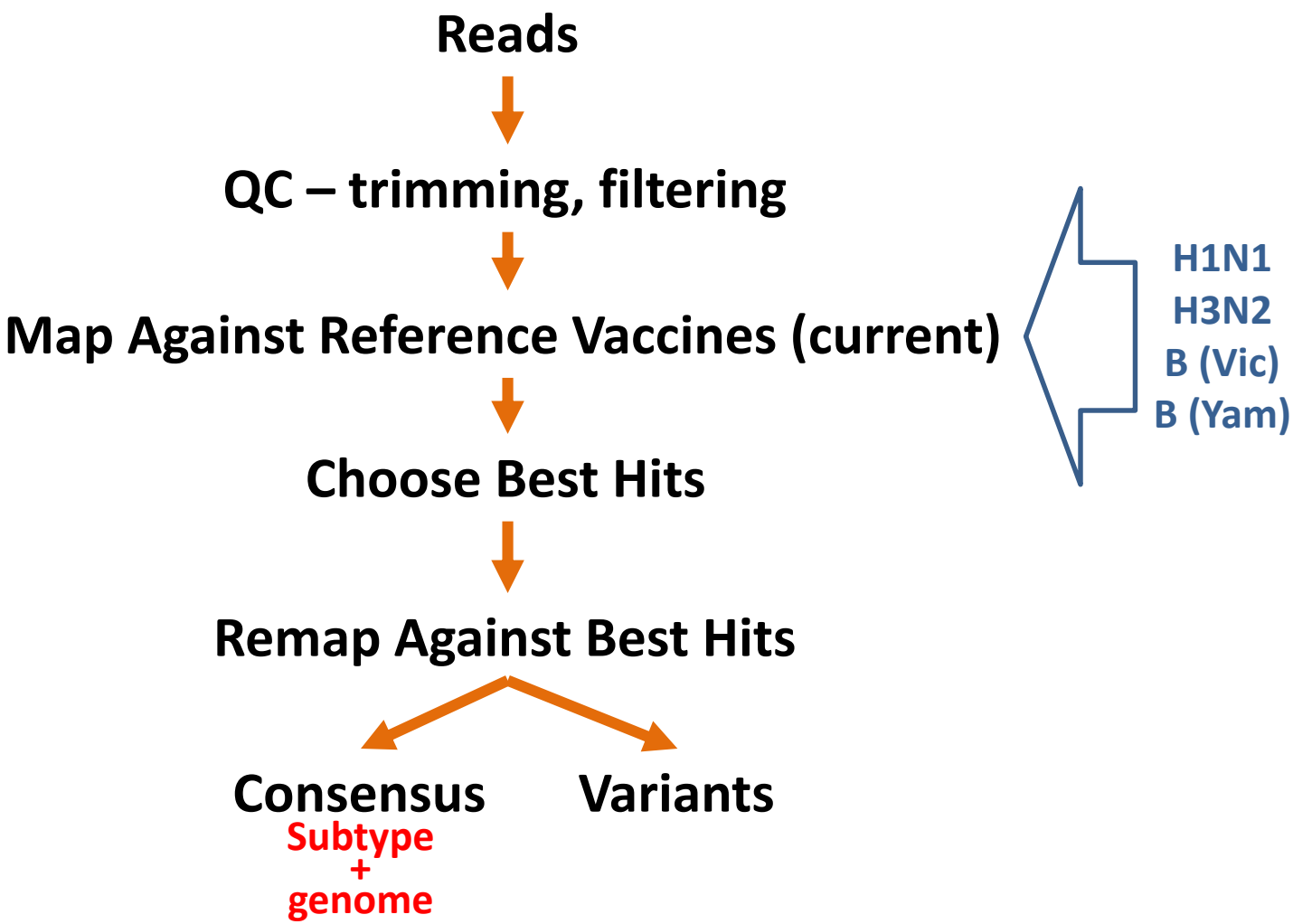
Our approach



Pros:
- method which works
- works fast, no assembly nor BLAST
- can save hours per sample
- can detect co-infection in one step

Cons:
- must prepare the DB containing Influlenza genomes beforehand

NGS Workflow



Bioinformatics Institute
A*STAR

Today's Exercise (con't)

- Open Ubuntu in Windows:

- Navigate to the FLU_DATA directory

  cd /mnt/c/Users/User/Workshop_Flu/FLU_DATA

- List FastQ sample files:

  ls –ltrh *.fastq

SRR1928163_R1.fastq

SRR1928163_R2.fastq

Clinical Sample, Thailand, 2012

(Rutvisuttinunt et al, Journal of Clinical Virology 2015)

Today's Exercise (con't)

- Remove old logfile:

rm FLUAB_VACCREFMIX_PE.log

- Similar commands as yesterday:

nano coinfect.txt

SRR1928*<tab>*SRR

ctrl-o, Enter, then crtl-x

config.pl –s coinfect.txt –d `pwd` -c FLUAB_VACCREFMIX_PE.conf

run_fluAB.pl FLUAB_VACCREFMIX_PE

Today's Exercise

- Result File: ***SRR_FINAL_STATS.txt***

FRAG	VTYPE1	VTYPE2	MREAD1	MREAD2	DMR	PCF1	PCF2
HA	H1	H0	42	24	18	91.67	75.14
MP	H1N1	H0N0	38	20	18	79.54	83.97
NA	N1	N0	19	21	2	71.04	74.28
NP	H1N1	H0N0	46	36	10	81.62	94.00
NS	H1N1	H0N0	23	28	5	86.13	91.37
PA	H1N1	H0N0	49	36	13	86.59	77.11
PB1	H1N1	H0N0	53	30	23	85.49	75.48
PB2	H1N1	H0N0	72	33	39	96.89	65.67

- ***VTYPE1 & VTYPE2 are Virus Types***
- ***H0N0 is Flu B!***
- ***PCF1 & PCF2 are percent coverage***

We have a Co-Infection

Today's Exercise

- Result File: ***SRR_FINAL_STATS.txt***

FRAG	PID1	PID2	VTEMPL1	VTEMPL2
HA	98.32	98.65	H1N1:A/Brisbane/02/2018	H0N0:B/Colorado/06/2017
MP	99.50	98.76	H1N1:A/Brisbane/02/2018	H0N0:B/Colorado/06/2017
NA	99.31	97.92	H1N1:A/Brisbane/02/2018	H0N0:B/Colorado/06/2017
NP	99.68	98.56	H1N1:A/Brisbane/02/2018	H0N0:B/Phuket/3073/2013
NS	96.34	98.37	H1N1:A/Brisbane/02/2018	H0N0:B/Colorado/06/2017
PA	99.16	99.83	H1N1:A/Brisbane/02/2018	H0N0:B/Colorado/06/2017
PB1	99.39	99.24	H1N1:A/Brisbane/02/2018	H0N0:B/Phuket/3073/2013
PB2	98.58	99.34	H1N1:A/Brisbane/02/2018	H0N0:B/Phuket/3073/2013

- ***H1N1 and Flu B coinfection***
- ***HA/NA of B virus are from Victoria lineage***
- ***PID1 and PID2 are percent identity to reference***