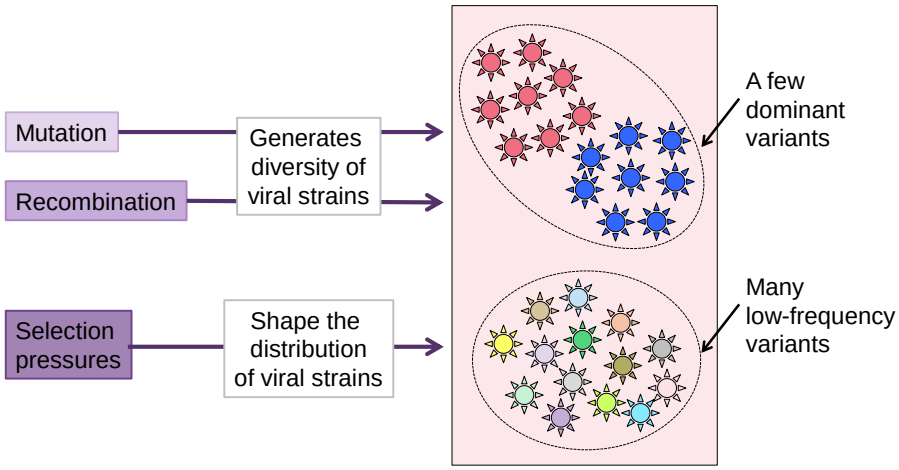
Kirby Institute

Using sequencing to understand HIV diversity

Vanessa Venturi | Infection Analytics Program

HIV quasispecies



Basic workflow

sample collection
↓
choose sequencing platform
↓
sample preparation
↓
sequencing
↓
pre-processing
↓
quality assessment
↓
read alignment
↓
identification of variants
↓
subsequent analyses

Experimental considerations

sample collection
↓
choose sequencing platform
↓
sample preparation
↓
sequencing
↓
pre-processing
↓
quality assessment
↓
read alignment
↓
identification of variants
↓
subsequent analyses

Sequencing

- Data generated:
  - Read lengths
  - Number of reads
- Sequencing errors:
  - Type of predominant errors (error profile)
  - Interpretation of quality scores
- Cost

Sample preparation

- Amount of virus available.
- Sample preparation errors (eg. associated with RNA extraction, RT, PCR amplification biases).

Comparison of sequencing technologies

	1st-generation: Conventional	2nd-generation: Next-generation	3rd-generation: Single-molecule
Technologies	Sanger	454 (Roche), Illumina, ABI SOLiD, Ion Torrent	Single-molecule real time (PacBio), Nanopore (Oxford)
Primary distinction	Gold standard	High-throughput via mass parallelisation of sequencing reactions	
			Sequence single molecules, requiring no DNA amplification
Cost	High	Medium	Low
Read length	Long	Short	Very long
Depth	Low	High	Medium
Error rates	Low	Medium	High

Illumina sequencing

- Sequencing-by-synthesis approach.
- Data output (MiSeq reagent kit v3):
  - Read lengths: 2x300bp
  - Number of reads: 44-50 million paired-end reads
- Sequencing error profile:
  - Predominant errors are substitution errors.
  - Average sequencing error rate: ~0.1% per base.
  - Higher error rates (~1% per base) towards sequence ends.
  - Higher error rates on one strand of the pair end reads.

Experimental considerations

sample collection
↓
choose sequencing platform
↓
sample preparation
↓
sequencing
↓
pre-processing
↓
quality assessment
↓
read alignment
↓
identification of variants
↓
subsequent analyses

Major challenge: the entanglement of
- technical variation (i.e. sample preparation + sequencing errors)

and
- biological variation

Using sequencing to understand HIV diversity

Experimental considerations

sample collection
↓
choose sequencing platform
↓
sample preparation
↓
sequencing
↓
pre-processing
↓
quality assessment
↓
read alignment
↓
identification of variants
↓
subsequent analyses

Major challenge: the entanglement of
- technical variation (i.e. sample preparation + sequencing errors)

and
- biological variation

↓

Requires well-designed experiments with protocols that enable
- Assessment of the overall level of technical noise (eg. control/replicate samples, spike-ins)
- Identification and potentially correction of sample preparation and sequencing errors (eg. control samples, primerIDs, BAsE-Seq, CirSeq,)

Bioinformatics considerations

sample collection
↓
choose sequencing platform
↓
sample preparation
↓
sequencing
↓
pre-processing
↓
quality assessment
↓
read alignment
↓
identification of variants
↓
subsequent analyses

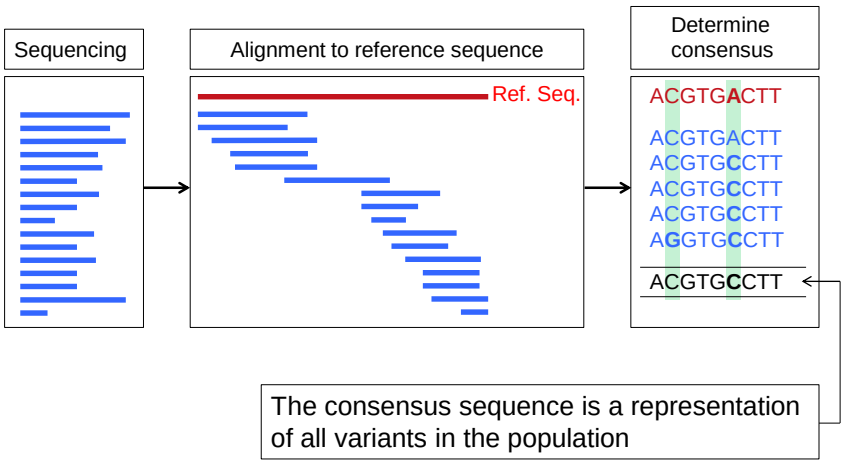
- Computational methods work best on data generated by well designed experiments.
- Quality filtering and error-correction algorithms can “improve” data quality (at the cost of data yield) but are a poor substitute for experimental protocols.
- Pipelines (or specific algorithms utilised) are often developed for platform-specific data.

Assembly / alignment of reads

sample collection
↓
choose sequencing platform
↓
sample preparation
↓
sequencing
↓
pre-processing
↓
quality assessment
↓
read alignment →
- Reference-based mapping
- *De novo* assembly

↓
identification of variants
↓
subsequent analyses

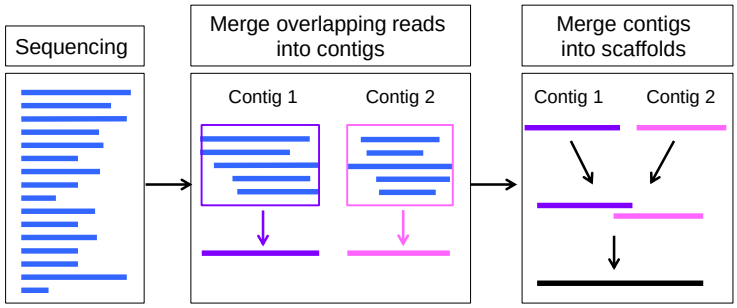
Reference-based mapping



Reference-based mapping

- Requires a reference sequence (eg. Los Alamos National Laboratory HIV database, https://www.hiv.lanl.gov/).
- Biased towards:
  - Sequences that are more similar to the reference sequence.
  - More abundant variants.
- Reads that substantially differ from the reference sequence align poorly and are often discarded in subsequent analyses.
- Advantage that small variations (eg. SNVs) are more easily positioned.

De novo assembly



De novo assembly

- No bias towards a reference sequence.
- The assembly is often more fragmented.
- Works better for medium- to large-scale variations.

Assembly / alignment of reads

- Assembly using hybrid reference mapping / de novo approaches have some advantages.

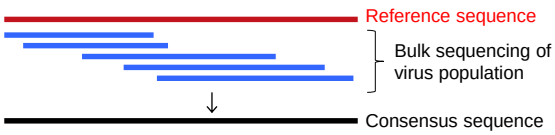
- The quality of sequence alignments depends on the strengths and weaknesses of the alignment algorithm used:
  - Computational efficiency.
  - Alignment biases.

    Example: Bias introduced in the handling of gaps. Many algorithms do not support gapped alignments, try to minimise gaps at all costs, or preferentially position gaps (eg. in homopolymer regions).
  - Loss of reads.

Sequencing approaches for HIV diversity

- **Consensus Sanger sequencing**
- **Next-generation sequencing (NGS)**
- **Single genome sequencing (SGS)**

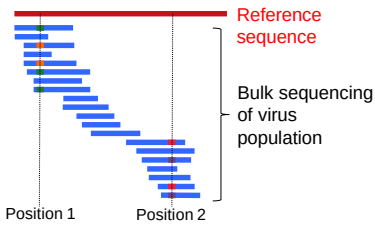
Consensus Sanger sequencing



- Has a 20% frequency detection threshold.
- Limited sensitivity to detect low-frequency variants -> underestimation of variant diversity.

Next-generation (NGS) sequencing

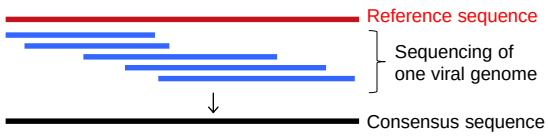
- High coverage assists in the detection of low-frequency mutations at particular positions, but difficult to link mutations at one position with mutations at another position, due to short read lengths.



- Coverage can be uneven, with potential fragmentation.
- Higher sequencing error rates -> challenge to distinguish real variation from technical noise (sample preparation + sequencing error).

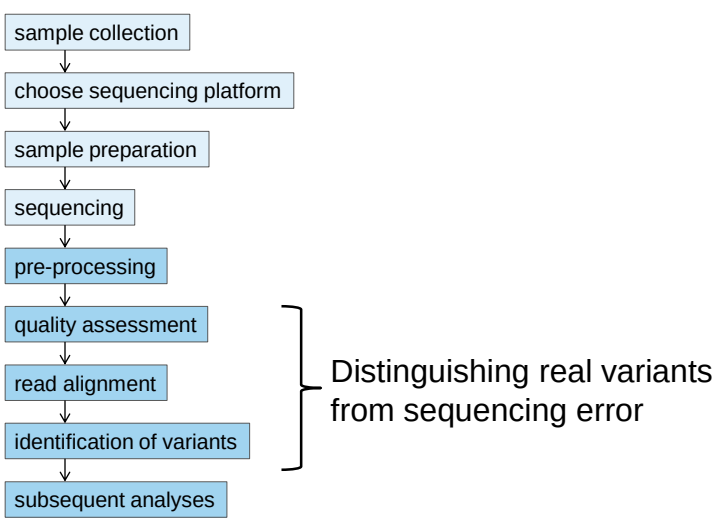
Single genome sequencing

- Isolation of single cell -> nucleic acid extraction -> single genome amplification (SGA) -> sequencing library preparation -> sequencing.

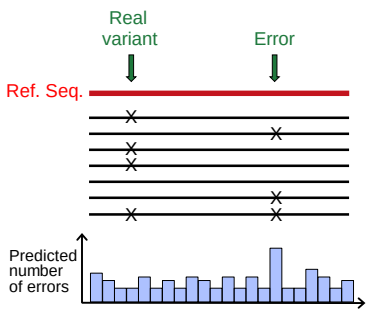


- Advantages:
  - Consensus sequence represents a single viral strain.
  - Enables detection of distant co-occurring mutations.
- Challenges:
  - Efficient physical isolation of individual cells.
  - Minimising amplification bias.

Identification of variants



Identification of variants



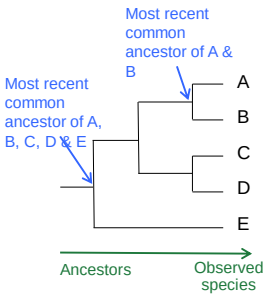
- Identify mismatches with the reference sequence that are more frequently observed than predicted by an error model.
- Various methods for modelling the distribution of errors at each sequence position, based on:
  - Quality scores.
  - Adaptive quality threshold estimated for each sample.
  - Control samples.

Analysing diversity

- Phylogenetic analysis
- Measures of diversity

Phylogenetic analysis

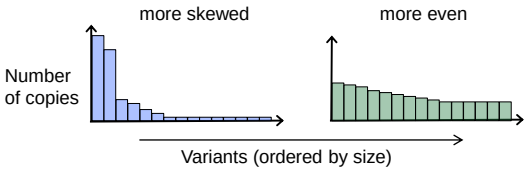
- Provides insights into the viral genetics evolution and relationships between viral variants.
- Assesses the genetic distance between sequences and identifies clusters of similar sequences.
- Various approaches (eg. neighbor joining, maximum likelihood, Bayesian) make different model assumptions and include different parameters.



Measures of diversity

Counting observed variants

- Highly sensitive to sampling effects, particularly
  - Differences in the copy number distribution



  - Differences in sample size (i.e. number of sequences)

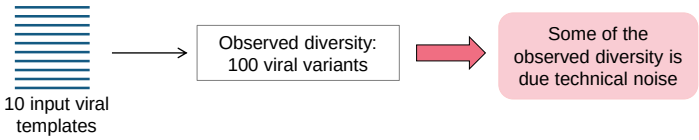
Measures of diversity

Diversity indices

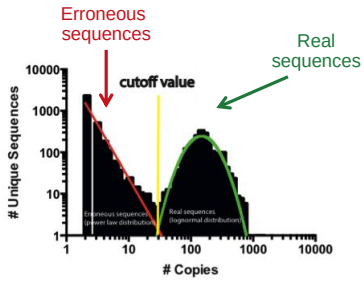
- Used in ecology and genetics.
- Account for number of unique variants and their copy number distribution, and differences in sample size.
- Compare sample diversity between groups of samples (eg. Simpson's, Shannon, Gini indices)
- Estimate total diversity of population from the samples (eg. Chao estimators)

Indicators of technical noise contribution

- Ratio of observed variants to input viral templates.



- Over-sequencing copy number distribution.

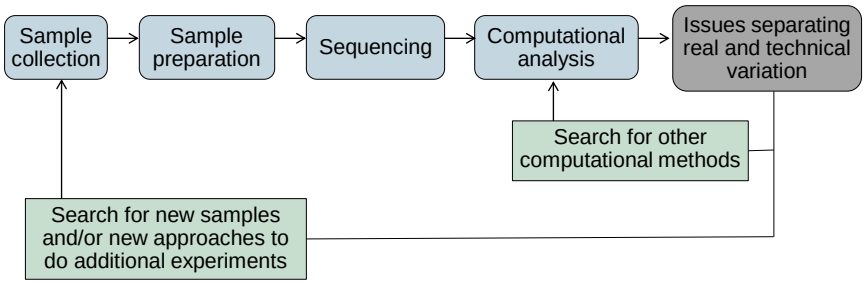


Input: 5000 virions

(Fennessey et al PLoS Path 2017)

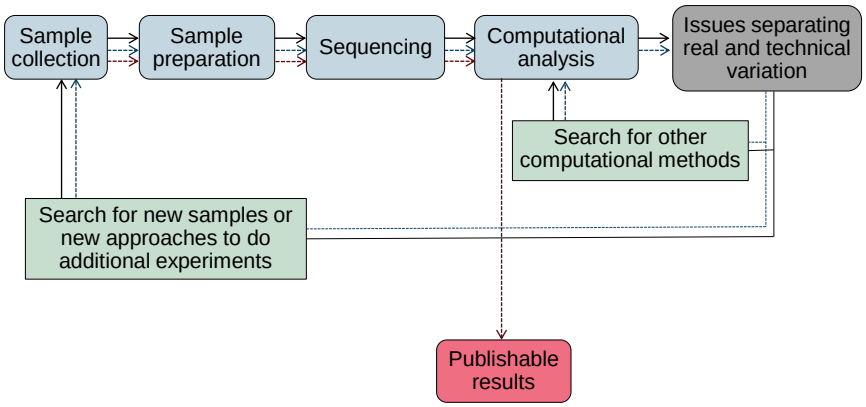
Concluding remarks

Sequencing project progression often looks like:



Concluding remarks

Sequencing project progression often looks like:



Concluding remarks

Sequencing project progression should look like: