

# Processing NGS Data – assembly strategies

What to do with all this data?

# Where we are

- 13:30-14:00 – Primer Design to Amplify Microbial Genomes for Sequencing
- 14:00-14:15 – Primer Design Exercise
- 14:15-14:45 – Molecular Barcoding to Allow Multiplexed NGS
- **14:45-15:15 – Processing NGS Data – de novo and mapping assembly**
- 15:15-15:30 – Break
- 15:30-15:45 – Assembly Exercise
- 15:45-16:15 – Annotation
- 16:15-16:30 – Annotation Exercise
- 16:30-17:00 – Submitting Data to GenBank

# Filtering NGS Data

- Once you have the reads for a particular sample (say after they have been sorted by barcode) it is important to use high quality data that is also free of sequence that was not an artifact of your protocols
- Low quality data and sequencing artifacts can break de novo assemblies or cause false variations to appear in mapping assemblies

# Quality Trimming

- When sequencing, in addition to the bases (A, C, G, T, and maybe N), there are associated quality values (qv)
- The qv is usually defined as

$$qv = -10 \log_{10} p_{\text{error}}$$

- qv 10 -> 1 in 10 chance of being wrong
- qv 20 -> 1 in 100
- qv 30 -> 1 in 1000

# Software to filter low quality

- Tools exist in most bioinformatics packages
- I tend to use a Perl program named TrimBWAstyle.pl
- I also tend to use a qv cutoff of 20

# Sequence Artifacts

- Major sources of sequence artifacts
  - PCR primers – they can still work when they are an imperfect match to the template, and they usually “win” in the final data
  - Adaptors from the NGS vendor – often a problem at ends of reads, especially if the input DNA had short fragments
  - Untrimmed barcodes

# Sequence artifacts

- For adaptors and barcodes remaining at ends of reads, multiple software again depending on tools you use
- For PCR primers, might be more tricky...
  - If the amplicon was sequenced by itself, you can trim the primer sequences after assembly
  - If amplicons were pooled and then sequenced, need to examine assemblies more closely, looking at the priming sites, and not trusting reads that end in the priming site





# Mapping/Reference Assembly

- As there are many de novo assemblers, there are also many mapping and aligning tools available
- At JCVI, I use CLC Bio command line tools on a Linux cluster
- At the next session, I'll show the use of Newbler and hopefully BWA/SAMTOOLS

# Reference Assembly with CLC (Linux)

- CLC command line tools != GUI equivalents
  - **clc\_ref\_assemble\_long** -o assembly.cas -q unpaired1.fasta -p fb ss 180 250 -i paired\_1.qf paired\_2.qf -d reference1.gb
    - Requires pre-processed seq data for correct mate placement
    - No restriction of maximum read length
    - long: read length > 36 bp, short: read length <= 36
  - **Method:**
    - Mismatch score-based, local, gapped alignment results in list of optimal match locations
    - When non-specific matches are found, choose randomly (fast but may cause chimeras)
    - For mates, placement choices are decided upon using input insert length bounds but may be placed 'as fragments' if such requirements are not met.
    - Output is in 'cas' format, a proprietary CLC format analyzed by other NGS Cell tools...\*

# SNP/Indel Detection Using **CLC** Output

- **Command Line (NGS Cell)**

- Use tool **'find\_variations'** for SNP and DIP calls
  - Different calculation method than that used by the GUI platform!
- Input format = 'cas' output of CLC Reference Assembler
  - (requires prior CLC NGS cell reference assembly)
- Method :
- **WHEN** a base has sufficient coverage for comparisons (usr input)...
  - **IF** difference between the reference assembly consensus and reference
  - **OR** fraction of variant calls > fractional cutoff (optional usr input)
  - **OR** number of variant calls > numerical cutoff (optional usr input)
    - Call Variant in this location

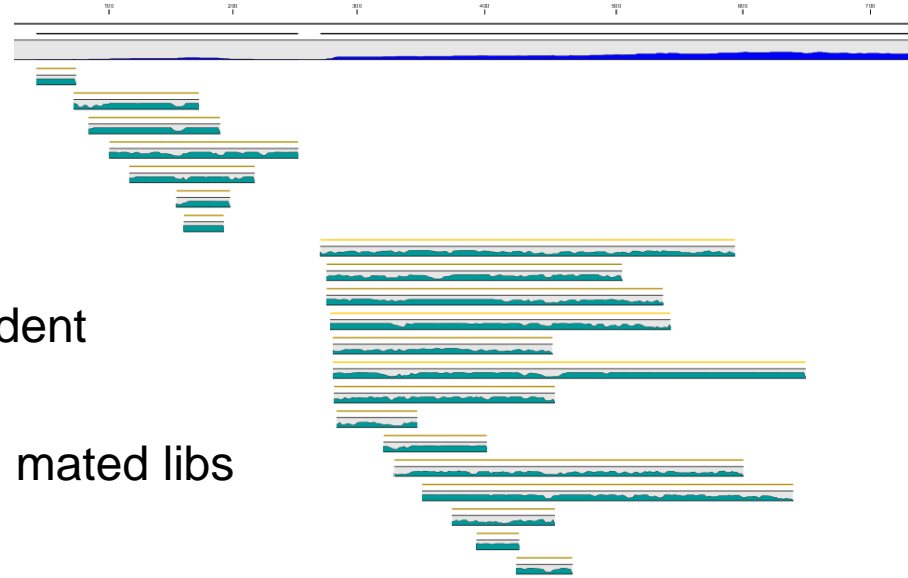
- **Output**

```
ctg1129645366355:
```

```
      9362   Deletion   A  ->  -  
     33298   Difference C  ->  T  
     34024   Insert    -  ->  A
```

# Reference Assembly with CLC (Genomics Workbench)

- Same basic algorithm, but with different capabilities regarding use of sequencing info
- **Cons:**
  - Must import data, More user-dependent
  - Higher chance of software failure
  - Can only use **ONE** insert size for all mated libs
- **Pros:**
  - Option for Global Alignment
  - Can utilize annotation information, can annotate problem regions
  - Gives output compatible with RNA-seq tools, better SNP/DIP detection and post-assembly analysis with GUI tools
  - Output also includes reference assembly table and ability to generate 'ACE'-type alignment output.
  - Post-assembly analysis includes quality information



# SNP/Indel Detection Using CLC Output

- **GUI SNP Detection**

- Based on 'Neighborhood Quality Standard' algorithm [[Altshuler et al., 2000](#)]
- More variables available for this tool than the CLC command line equivalent


- **Benefits of Workbench Version**

- Can annotate reference with SNPs
- Tabular output with more SNP info
- Output can be easily converted for DBSNP submission

Mapping	Reference ...	Consensus...	Variation T...	Reference	Variants	Allele Varia...	Frequencies	Counts	Coverage	Overlappin...	Amino Acid...
scf7180000...	984	984	SNP	G	1	A	100.0	14	14	Conflict: Co...	
scf7180000...	985	985	SNP	A	1	T	100.0	14	14	Conflict: Co...	
scf7180000...	841	839	SNP	T	1	C	100.0	12	12	Conflict: Co...	
scf7180000...	2663	2661	SNP	G	1	A	71.4	5	7	Conflict: Co...	
scf7180000...	2831	2829	Complex SNP	C	2	A/C	50.0/50.0	2/2	4	Conflict: Co...	
scf7180000...	525	462	Complex SNP	G	2	G/T	64.7/35.3	11/6	17	Conflict: Co...	

# SNP/Indel Detection Using CLC Output

- **GUI DIP Detection**

- More variables available for this tool  than cmd line version

Significance

Non-specific matches are ignored during DIP detection.

Minimum coverage

Minimum variant frequency (%)

Advanced

Minimum paired coverage

Maximum coverage

Minimum variant count required  and sufficient

---

Ploidy

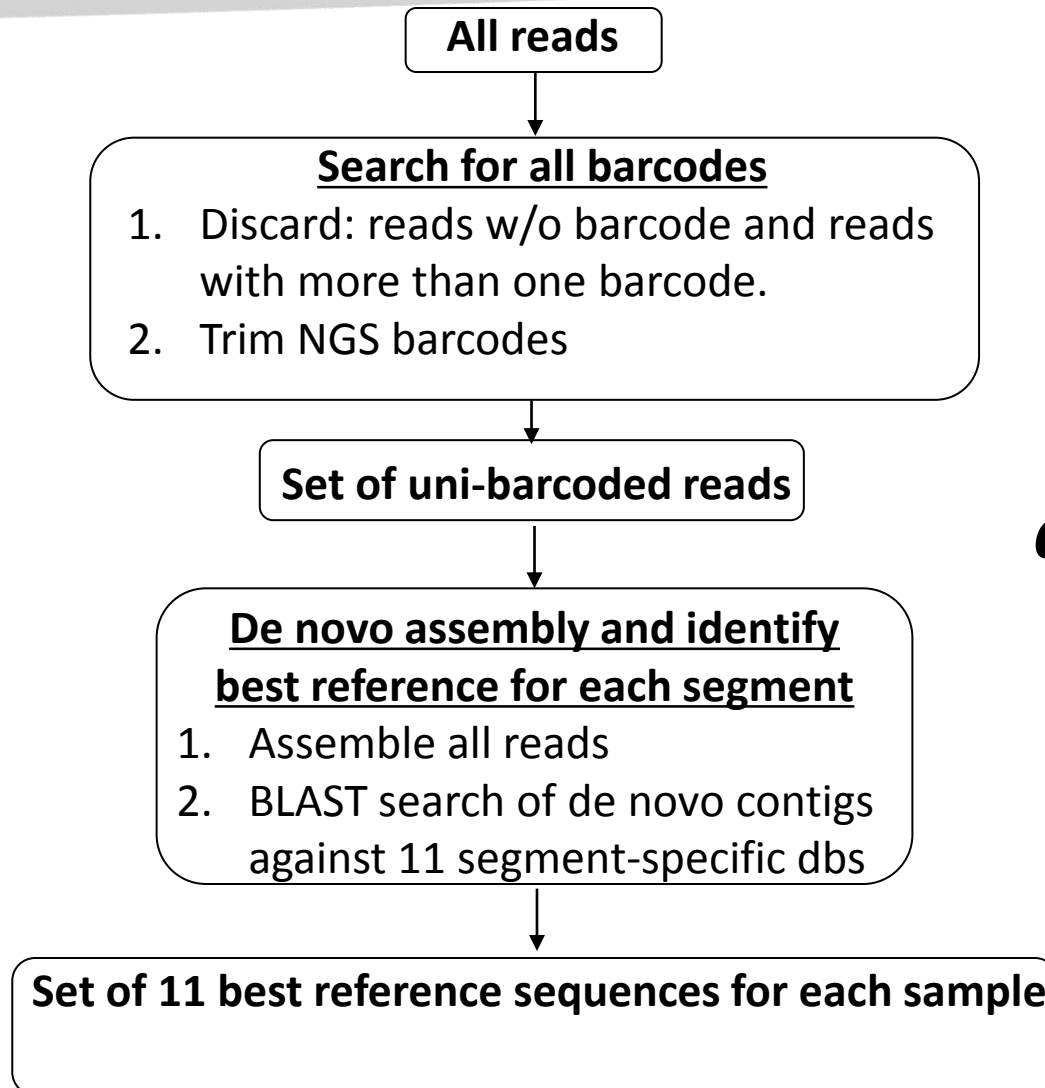
Maximum expected variations

- **Benefits of Workbench Version**

- Can annotate reference with DIPs
  - Tabular output with more DIP info
  - Used along with SNP output for DBSNP submission

Mapping	Reference ...	Consensus...	Variation T...	Length	Reference	Variants	Allele Varia...	Frequencies	Counts	Coverage
scf7180000...	2316110	2312177	DIP		1 -	1 C	1 C	60.0	3	5
scf7180000...	2560552	2556619	DIP		3 TAT	1 ---	1 ---	60.0	3	5
scf7180000...	2560850	2556917	DIP		2 AT	1 --	1 --	62.5	5	8
scf7180000...	2814885	2810948	Complex DIP		1 A	2 -/A	2 -/A	50.0/50.0	2/2	4
scf7180000...	3097595	3093664	DIP		1 -	1 T	1 T	100.0	4	4
scf7180000...	3888810	3883712	Complex DIP		1 T	2 T/-	2 T/-	50.0/50.0	2/2	4

# Processing of NGS multiplexed data (Rotavirus example)



*de novo  
assembly*

# Processing of NGS multiplexed data (Rotavirus example, conc.)

Set of 11 best reference sequences for each sample

Set of 11 best references from  
GenBank

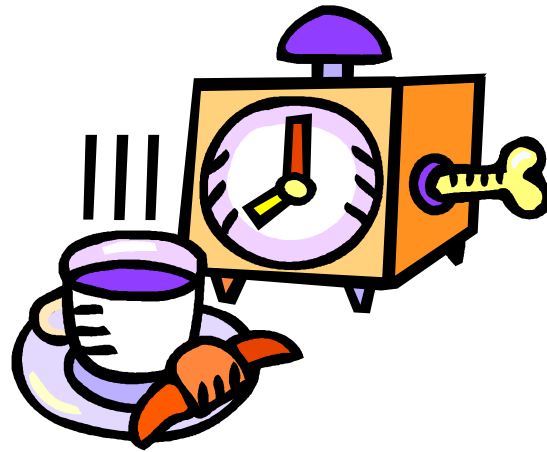
1. Reference mapping for NGS reads
2. Optional - Update references with variations identified & perform reference mapping again

Assembled genomes

*mapping  
assembly*



# Time for First Break



- See you at 15:30 when you will try assembling a viral genome yourself