Bioinformatics, Lecture 2

Genome assembly: quality assessment and annotation
Scaffolds

Unlike contigs, “scaffolds” contain unknown nucleotides, denoted as “N”.

> scaffold1  5.0
AAACTATACATTATATACGTACATAAAATATGAATTACATCAAAAATATATTATATTATATTATATTGAGTGAATAT
ATGAAATTACATCAAATATAGATTATATATACGTACATAAAATATCAAAGTACCCCA
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
Outline of this lecture

1. Assessing the quality of your genome assembly
   1. General statistics
   2. Homology based approaches
   3. Read mapping inconsistencies

2. Some strategies to improve your assembly
   1. General things to try
   2. Reference-based assemblies
   3. Long insert libraries and long reads

3. Genome annotation
   1. Ab-initio gene prediction
   2. Homology-based gene prediction
   3. RNA-seq-based gene prediction
Two measures of the quality of a genome assembly: completeness and continuity

Measures of continuity

Measures of completeness
Two measures of the quality of a genome assembly: completeness and continuity

<table>
<thead>
<tr>
<th>Measures of continuity</th>
<th>Measures of completeness</th>
</tr>
</thead>
<tbody>
<tr>
<td>Scaffold size</td>
<td>% of reference genome contained</td>
</tr>
<tr>
<td>Scaffold number</td>
<td>% of expected genes found in assembly</td>
</tr>
<tr>
<td>Genes contained in single scaffold</td>
<td></td>
</tr>
</tbody>
</table>

NB: When coverage is not very high, large K-mers tend to produce a continuous but incomplete assembly, and small K-mers a complete but fragmented assembly!
**Scaffold size: some measures**

| 5 | 5 | 8 | 9 | 9 | 10 | 50 | 50 |

**Mean:** \( \frac{5+5+8+9+9+10+50+50}{8} = 18.25 \)

→ Is very sensitive to outliers: the mean is larger than all scaffolds but 2!

**Median:** 8 scaffolds, length at which 4 scaffolds are smaller and 4 are larger = 9

→ Is generally more representative, but **does not distinguish between a good assembly with many tiny scaffolds and a bad assembly with few tiny scaffolds!**
Scaffold size: N50

N50 overcomes this issue by calculating the scaffold size $x$ at which:

- 50% of the genome is contained in scaffolds smaller than $x$
- 50% of the genome is contained in scaffolds larger than $x$
Scaffold size: N50

Total genome size: 146, 50% of genome is 73

What is the N50?
Two measures of the quality of a genome assembly: completeness and continuity

- **Genome**
- **Assembly 1** *High N50!*
- **Assembly 2** *Low N50!*

If you know the expected genome size, you can estimate

**NG50:** 50% of the known or estimated genome size that must be of the NG50 length or longer
<table>
<thead>
<tr>
<th>Statistic</th>
<th>In a good assembly will be</th>
</tr>
</thead>
<tbody>
<tr>
<td>Length of largest scaffold</td>
<td>Larger</td>
</tr>
<tr>
<td>Number of scaffolds</td>
<td>Smaller (but get suspicious if too small)</td>
</tr>
<tr>
<td>N90, N75, etc...</td>
<td>Larger</td>
</tr>
</tbody>
</table>
Two measures of the quality of a genome assembly: completeness and continuity

<table>
<thead>
<tr>
<th>Measures of continuity</th>
<th>Measures of completeness</th>
</tr>
</thead>
<tbody>
<tr>
<td>Scaffold size</td>
<td>% of reference genome contained</td>
</tr>
<tr>
<td>Scaffold number</td>
<td>% of expected genes found in assembly</td>
</tr>
<tr>
<td>Genes contained in single scaffold</td>
<td></td>
</tr>
</tbody>
</table>
### Homology-based approaches

<table>
<thead>
<tr>
<th>Genome</th>
<th>Assembly 1</th>
<th>Assembly 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Present+Complete</td>
<td>Present+Fragmented</td>
</tr>
<tr>
<td>2</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>3</td>
<td>3</td>
<td>7</td>
</tr>
<tr>
<td>4</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>5</td>
<td></td>
<td></td>
</tr>
<tr>
<td>6</td>
<td></td>
<td></td>
</tr>
<tr>
<td>7</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
What genes to use? The core eukaryotic gene set

>300 genes that are consistently found in single copy in any eukaryote
1. Locate gene in the genome (Blast)

Consensus eukaryotic core gene
Homology-based: BUSCO

1. Locate gene in the genome (Blast)

2. Predicts structure/sequence of the gene in the assembly (Augustus)

Consensus eukaryotic core gene
Homology-based: BUSCO

1. Locate gene in the genome (Blast)

2. Predicts structure/sequence of the gene in the assembly (Augustus)

3. Compares consensus and predicted (HMMER):
   Are they really the same (orthologs)? Or just part of the same gene family?
   Is the gene sequence complete or fragmented?
Homology-based: BUSCO

**Genome/transcriptome assembly** + **Core eukaryotic gene set**

### BUSCOs:

<table>
<thead>
<tr>
<th>Component</th>
<th>Consensus sequence</th>
<th>Block-profiles</th>
<th>HMMs</th>
<th>Classifier</th>
</tr>
</thead>
<tbody>
<tr>
<td>Genome assembly</td>
<td>tBLASTn</td>
<td>Augustus</td>
<td>HMMER 3</td>
<td>C [D], F, M, n</td>
</tr>
<tr>
<td>run-time:</td>
<td>15%</td>
<td>80%</td>
<td>5%</td>
<td>C: Complete [ D: Duplicated ] F: Fragmented M: Missing n: no. of genes</td>
</tr>
<tr>
<td>Transcriptome</td>
<td></td>
<td>Find ORF</td>
<td>HMMER 3</td>
<td></td>
</tr>
<tr>
<td>Gene set</td>
<td></td>
<td></td>
<td>HMMER 3</td>
<td></td>
</tr>
</tbody>
</table>
### Homology-based: BUSCO

#### Eukaryotic lineages:

<table>
<thead>
<tr>
<th>Lineage</th>
<th>Number of species Total : Selected</th>
<th>Number of BUSCO groups</th>
</tr>
</thead>
<tbody>
<tr>
<td>eukaryota_odb9*</td>
<td>90 : 65</td>
<td>303</td>
</tr>
<tr>
<td>nematoda_odb9</td>
<td>10 : 8</td>
<td>982</td>
</tr>
<tr>
<td>metazoa_odb9*</td>
<td>330 : 65</td>
<td>978</td>
</tr>
<tr>
<td>arthropoda_odb9</td>
<td>133 : 60</td>
<td>1'066</td>
</tr>
<tr>
<td>insecta_odb9</td>
<td>116 : 42</td>
<td>1'658</td>
</tr>
<tr>
<td>endopterygota_odb9</td>
<td>100 : 35</td>
<td>2'442</td>
</tr>
<tr>
<td>diptera_odb9</td>
<td>53 : 25</td>
<td>2'799</td>
</tr>
<tr>
<td>hymenoptera_odb9</td>
<td>32 : 25</td>
<td>4'415</td>
</tr>
<tr>
<td>vertebrata_odb9*</td>
<td>172 : 65</td>
<td>2'586</td>
</tr>
<tr>
<td>actinopterygii_odb9</td>
<td>23 : 20</td>
<td>4'584</td>
</tr>
<tr>
<td>tetrapoda_odb9</td>
<td>146 : 55</td>
<td>3'950</td>
</tr>
<tr>
<td>aves_odb9</td>
<td>54 : 40</td>
<td>4'915</td>
</tr>
<tr>
<td>mammalia_odb9</td>
<td>84 : 50</td>
<td>4'104</td>
</tr>
<tr>
<td>euarchontognires_odb9</td>
<td>37 : 25</td>
<td>6'192</td>
</tr>
<tr>
<td>laurasiatheria_odb9</td>
<td>33 : 25</td>
<td>6'253</td>
</tr>
</tbody>
</table>
Third measure of the quality of a genome assembly: correctness

Measures of correctness:
- sequencing errors included
- mistakes in the order of the scaffolding
- duplications collapsed

Paired-end reads can be used to detect problems in the assembly without a reference genome!
Read mapping: good assembly

- Scaffold
- Paired-end reads
- Coverage
- Inconsistent paired-end reads
- SNP density
Read mapping problems: case 1

Scaffold
Paired-end reads
SNP

Coverage
Inconsistent paired-end reads
SNP density
Read mapping problems: Collapsed duplicated region

Scaffold
Paired-end reads

Corrected assembly

Copy 1

Copy 2

SNP
Read mapping problems: case 2

Coverage

Inconsistent paired-end reads

SNP density
Read mapping problems: inverted part of scaffold

Coverage

Inconsistent paired-end reads

SNP density
Read mapping approach: ALE

Sequencing errors

Inserted sequence

misassembly

Collapsed duplication

# Read mapping inconsistency

<table>
<thead>
<tr>
<th>Error type</th>
<th>Transcripts</th>
<th>Assembly</th>
<th>Read evidence</th>
</tr>
</thead>
<tbody>
<tr>
<td>Family collapse</td>
<td>geneAA geneAB geneAC</td>
<td>n=3</td>
<td></td>
</tr>
<tr>
<td>Chimerism</td>
<td>geneC</td>
<td>n=2</td>
<td>coverage</td>
</tr>
<tr>
<td>Unsupported insertion</td>
<td></td>
<td>n=1</td>
<td>no reads align to insertion</td>
</tr>
<tr>
<td>Incompleteness</td>
<td></td>
<td>n=1</td>
<td>read pairs align off end of contig</td>
</tr>
<tr>
<td>Fragmentation</td>
<td></td>
<td>n=1</td>
<td>bridging read pairs</td>
</tr>
<tr>
<td>Local misassembly</td>
<td></td>
<td>n=1</td>
<td>read pairs in wrong orientation</td>
</tr>
<tr>
<td>Redundancy</td>
<td></td>
<td>n=1</td>
<td>all reads assign to best contig</td>
</tr>
</tbody>
</table>

https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4971766/
1. Assessing the quality of your genome assembly
   1. General statistics
   2. Homology based approaches
   3. Read mapping incongruence

2. Some strategies to improve your assembly
   1. General things to try
   2. Reference-based assemblies
   3. Long insert libraries and long reads

3. Genome annotation
   1. Ab-initio gene prediction
   2. Homology-based gene prediction
   3. RNA-seq-based gene prediction
The influence of Kmer size

- Try a range or use a multi-Kmer assembly!

- Use sufficient coverage that even high Kmer gives continuous assembly!
  >10x coverage, ideally >50x coverage for large genomes!
When there are large repetitive sequences, more coverage won’t help
Reference-based assembly: closely related species/strain

1. Map reads to reference
2. Assemble contigs
2. Scaffold into chromosome

“Align-then-assemble” approach
Reference-based assembly: A. thaliana strains

De novo: 12.6% of non-centromeric genome missing.

Reference-based: 3.7% of non-centromeric genome missing

http://www.pnas.org/content/108/25/10249.full
Reference-based assembly: more distantly related species

1. Assemble contigs de novo
2. Map contigs to reference
3. Scaffold

"Assemble-then-align" approach.
Reference-based assembly: black grouse

Two types of SOLID reads:

-Single end: 793 M reads of length = 75 bp

-Paired end1: ~3000 M reads of length of 60 bp

Very short for de novo assembly!

Only ~55% of reads left after filtering!

Reference-based assembly: black grouse

1. Quality and length filtering of reads (Fastx)
2. Single-end and mate-paired reads
   - De novo assembly (SOAPdenovo)
   - Contigs length ≥ 1Kb
   - Mapping to chicken genome (BWA-SW)
   - Contigs ≥ 100bp
   - Merging to black grouse scaffolds (BWA-SW)
3. Only mate-paired reads
   - Mapping to chicken genome (BWA)
   - Black grouse backbone scaffolds (SAMtools)
   - Black grouse scaffolds

Contamination check (BLAST)
Black grouse unscaffold contigs

Reference-based assembly: black grouse

Figure 3
Heatmap showing the proportion of the regions sequenced on the chromosome scaffolds.

When there are large repetitive sequences, more coverage won’t help.
When there are large repetitive sequences, more coverage won’t help.
Mate-pairs (Illumina)

Illumina Sequencing

Because of the way the illumina machines work, you can’t directly sequence huge fragments.

<1000bps

In order to get larger fragments, you need to sequence **mate-pairs** instead of paired end reads.
Mate-pairs (Illumina)

Genomic DNA (blue) is tagmented with a Mate Pair Tagment Enzyme, which attaches a biotinylated junction adapter (green) to both ends of the tagmented molecule.

The tagmented DNA molecules are then circularized and the ends of the genomic fragment are linked by two copies of the biotin junction adapter.
Circularized molecules are then re-fragmented yielding smaller fragments. Sub-fragments containing the original junction are enriched via the biotin tag (B) in the junction adapter.

After End Repair and A-Tailing, TruSeq DNA adapters (grey and purple) are then added, enabling amplification and sequencing.
Mate-pairs: need to tell assemblers their orientation

Paired end: correct orientation

Mapped reads:
Mate-pairs: need to tell assemblers their orientation
Mate-pairs: need to tell assemblers their orientation
In the configuration file:

```plaintext
#maximal read length
max_rd_len=101

[LIB]
#average insert size
avg_ins=200
#if sequence needs to be reversed
reverse_seq=0
#if in which part(s) the reads are used
asm_flags=3
#if use only first 100 bps of each read
rd_len_cutoff=101
#if in which order the reads are used while scaffolding
rank=1
#if cutoff of pair number for a reliable connection (at least 3 for short insert size)
pair_num_cutoff=3
#if minimum aligned length to contigs for a reliable read location (at least 32 for short insert size)
map_len=32
#A pair of fastq file, read 1 file should always be followed by read 2 file
q1=fastq1_read_1.fq
q2=fastq1_read_2.fq
```

Change to 1
When there are large repetitive sequences, more coverage won’t help.

“Third generation sequencing”
Longer reads: PacBio

Fluorescently labeled nucleotides

Fixed polymerase

Single molecule of DNA

Excitation

Emission
Only happens if you start out with great quality DNA!
PacBio: very high error rates!

<table>
<thead>
<tr>
<th>Technology</th>
<th>Read length (bp)</th>
<th>Error rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>ABI/Solid</td>
<td>75</td>
<td>Low (~2%)</td>
</tr>
<tr>
<td>Illumina/Solexa</td>
<td>100–150</td>
<td>Low (&lt;2%)</td>
</tr>
<tr>
<td>IonTorrent</td>
<td>~200</td>
<td>Medium (~4%)*</td>
</tr>
<tr>
<td>Roche/454</td>
<td>400–600</td>
<td>Medium (~4%)*</td>
</tr>
<tr>
<td>Sanger</td>
<td>Up to ~2,000 bp</td>
<td>Low (~2%)</td>
</tr>
<tr>
<td>Pacific Biosciences</td>
<td>Up to ~15,000†</td>
<td>High (~18%)</td>
</tr>
</tbody>
</table>

http://www.nature.com/nrg/journal/v14/n3/full/nrg3367.html

→ You either get very high coverage (expensive!)
→ Or used hybrid approaches: eg correct errors using short reads, then assemble.
Longer reads: Oxford nanopore

You do the sequencing on site! Currently has high sequencing errors (accuracy ranging 65%–88%) and unreliable output (<0.1 GB to 1 GB), but should improve in the future!
Oxford nanopore: on site sequencing of ebola epidemic

Could be taken on site to Guinea

Sequences less than 24 hours after the arrival of the sample!

http://www.nature.com/nature/journal/v530/n7589/full/nature16996.html
Don’t sequence a genome unless you have to: other strategies

http://www.nature.com/nrd/journal/v12/n5/full/nrd3979.html
1. Assessing the quality of your genome assembly
   1. General statistics
   2. Homology based approaches
   3. Read mapping incongruence

2. Some strategies to improve your assembly
   1. General things to try
   2. Reference-based assemblies
   3. Long insert libraries and long reads

3. Genome annotation
   1. Ab-initio gene prediction
   2. Homology-based gene prediction
   3. RNA-seq-based gene prediction
Genome annotation: classifying the genome into structural/functional components

First step is usually finding genes & understanding their structure.
Gene prediction: ab-initio detection of large open reading frames (ORF)

In a random sequence, there should be a stop codon approximately every 20–25 codons!

→ If you find 300 codons in a row with no STOP, very likely to be part of a gene!
Gene prediction: ab-initio detection of large open reading frames (ORF)

How does it fit into the structure of the gene?
Gene prediction: ab-initio detection of large open reading frames (ORF)

START codon

ORF

No other ORFs

Gene
Gene prediction: ab-initio detection of large open reading frames (ORF)

START codon

No START codon

ORF

other ORFs

Exon 1

Exon 2

Exon 3

Gene
Gene prediction: homology-based

Genome of closely related species

Genes 1 to 7

Your genome
Gene prediction: homology-based

Problem:
- you may not find genes that evolve very fast (especially non-coding genes)
- species-specific genes will also be missing
Gene prediction: RNA-seq

Your genome

RNA-seq reads

-often contains regulatory sequences before and after the coding sequence

-can be combined with ORF information to obtain gene structure!

-also provides information about non-coding genes!
Modern genome annotation: MAKER

1 predictions