



The background image shows a computer workstation in a dimly lit room. There are three monitors. The left monitor displays a web-based data interface. The middle monitor is a laptop showing a similar interface. The right monitor displays a code editor with green and white text. A keyboard and a mouse are on the desk. A sign with the number '601' is visible on a shelf in the background. The text 'Genome Data Management Workshop' is overlaid in the center.

Genome Data Management Workshop

January 22nd-26th 2011

Universidad de Puerto Rico en
Mayagüez

GDM Workshop

Day 1

- I. DNA sequencing
- II. Planning your project
- III. Hardware requirements
- IV. A primer on Linux

Day 2

- V. Genome assembly

Day 3

- VI. Genome annotation

Day 4

VII. Perl

VIII. Phylogeny

Day 5

VIII. Phylogeny



WIFI Network code
SquawkinGood

V. Genomic/EST assembly

The shotgun puzzle

Garbage **in**? Garbage **out**...

Got contaminants?

Know which one(s)? > **Map & Filter**

Don't know?

> **Assemble**

> **BLAST**

> **Map & Filter**

> **Reassemble**

QVs

Sanger Q score ($Q = -10 \log_{10} P$)

Illumina FASTQ \neq Sanger FASTQ

Trimming the ends?

Filtering paired-ends/mate-pairs?

454 uses flowgrams

Do you need an assembly?

Resequencing

Detect SNPs

^ Reads mapping

***de novo* or guided assemblies?**

By definition, guided assemblies **are
biased...**

... but can be **very useful**

Mapping software

MAQ	http://maq.sourceforge.net/
SOAP/SOAP2	http://soap.genomics.org.cn/
Bowtie & Crossbow	http://bowtie-bio.sourceforge.net/crossbow/
BWA	http://bio-bwa.sourceforge.net/

Commercial assembly software

Geneious

<http://www.geneious.com/>

Sequencher

<http://www.genecodes.com/>

SeqMan NGen DNASTar

<http://www.dnastar.com/>

CLC Assembly Cell

<http://www.clcbio.com/>

Open source/free assembly software

Velvet/Oases

<http://www.ebi.ac.uk/~zerbino/velvet/>

Ray

<http://denovoassembler.SourceForge.net/>

Newbler (GS De Novo Assembler)

Roche; private requests

ABYSS/Trans-ABYSS

<http://www.bcgsc.ca/platform/bioinfo/software/abyss>

Consed

www.phrap.org/consed/consed.html

Algorithms

Greedy

Speedy

Memory-optimized

J. R. Miller *et al.* *Genomics*. **95**, 315-327 (2010)

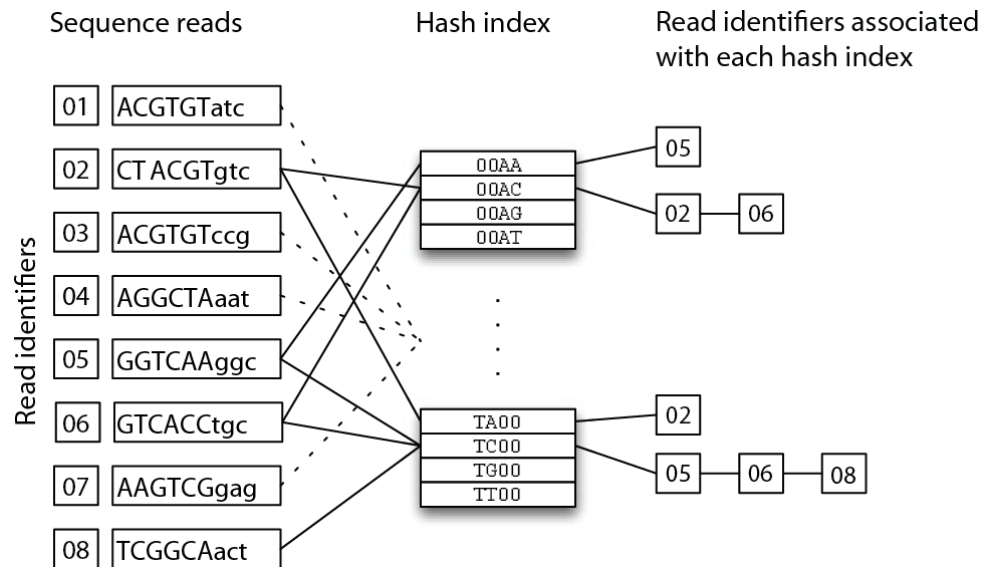
P. Flicek & E. Birney. *Nature Methods Supplement*. **6**:11, S6-S12 (2009)

Hash-based algorithms

Reads or genome
can be hashed

Tolerate high
polymorphism

Mapping algorithms



Burrow-Wheelers transform methods

Faster

Require less RAM

Struggle with high polymorphism

Mapping algorithms

1. All possible rotations

^TAGTCGAGGCTTTA\$



2. Sort

^TAGTCGAGGCTTTA\$
TAGTCGAGGCTTTA\$^
AGTCGAGGCTTTA\$^T
GTCGAGGCTTTA\$^TA
TCGAGGCTTTA\$^TAG
CGAGGCTTTA\$^TAGT
GAGGCTTTA\$^TAGTC
AGGCTTTA\$^TAGTCG
GGCTTTA\$^TAGTCGA
GCTTTA\$^TAGTCGAG
CTTTA\$^TAGTCGAGG
TTTA\$^TAGTCGAGGC
TTA\$^TAGTCGAGGCT
TA\$^TAGTCGAGGCTT
A\$^TAGTCGAGGCTTT
\$^TAGTCGAGGCTTTA



GTTTGCAGAA^TGTC\$A



3. Select final column

AGGCTTTA\$^TAGTCG
AGTCGAGGCTTTA\$^T
A\$^TAGTCGAGGCTTT
CGAGGCTTTA\$^TAGT
CTTTA\$^TAGTCGAGG
GAGGCTTTA\$^TAGTC
GCTTTA\$^TAGTCGAG
GGCTTTA\$^TAGTCGA
GTCGAGGCTTTA\$^TA
TAGTCGAGGCTTTA\$^
TA\$^TAGTCGAGGCTT
TCGAGGCTTTA\$^TAG
TTA\$^TAGTCGAGGCT
TTTA\$^TAGTCGAGGC
^TAGTCGAGGCTTTA\$
\$^TAGTCGAGGCTTTA

The Shortest Common Superstring (SCS)

Greedy

All-against-all

Pairwise alignment scores

Used in old Sanger assemblies

de novo algorithms

What is a K-mer?

A subset of a sequence

Think of sliding windows

TATTTGTAGCTGACGCTAGCTAGCTGTACGTG

TATTTGTAGCTGACGCTAG

ATTTGTAGCTGACGCTAGC

TTTGTAGCTGACGCTAGCT

TTGTAGCTGACGCTAGCTA

TGTAGCTGACGCTAGCTAG

GTAGCTGACGCTAGCTAGC

TAGCTGACGCTAGCTAGCT

AGCTGACGCTAGCTAGCTG

GCTGACGCTAGCTAGCTGT

CTGACGCTAGCTAGCTGTA

TGACGCTAGCTAGCTGTAC

GACGCTAGCTAGCTGTACG

ACGCTAGCTAGCTGTACGT

CGCTAGCTAGCTGTACGTG

Overlap-Layout-Consensus (OLC)

Implicitly use K-mers as heuristic

But rely on overlap graphs

Progressive pair-wise alignments

Unitigs & Contigs

de novo algorithms

de Bruijn graphs (DBG)

Explicitely rely on K-mer graphs

Memory intensive

All reads must be the same length

Do not use QVs

de novo algorithms

Repeated elements **break** algorithms

Dispersed vs. local

Inverted repeats

Tandem repeats

Palindromes (trick: use odd k-mers)

Now imagine telomeres...

***“If there is one message to remember [...],
it is to **not fully trust any** assembly.”***

Verify your assemblies

Try different methods

Map reads on your assemblies

Got repeats?

Watch out for chimeras

Look for scaffolding errors

About library insert sizes...

An 80 bases insert

25 bases

30 bases

25 bases

In theory

[illegible]

25 bases

20 bases

25 bases

In practice

[illegible]

25 bases

40 bases

25 bases

or

[illegible]

Standard Deviation

The usual scaffolding errors

Perfect

```
AGCTGTCTGTTTTCTGTAGCTCGTATTACATATCGATGGA
AGCTGTCTGTTTTCTGTAGCTCGTA
AGCTGTCTGTTTTCTGTAGCTCG
      TGTAGCTCGTATTACATATCGATGGA
      AGCTCGTATTACATATCGATGGA
```

Overestimation

```
AGCTGTCTGTTTTCTGTAGCNNNNNNNCGTATTACATATCGATGGA
AGCTGTCTGTTTTCTGTAGCTCGTA
AGCTGTCTGTTTTCTGTAGCTCG
      TGTAGCTCGTATTACATATCGATGGA
      AGCTCGTATTACATATCGATGGA
```



Underestimation

```
AGCTGTCTGTTTTCTGTTATTACATATCGATGGA
AGCTGTCTGTTTTCTGTAGCTCGTA
AGCTGTCTGTTTTCTGTAGCTCG
      TGTAGCTCGTATTACATATCGATGGA
      AGCTCGTATTACATATCGATGGA
```

An iterative process

N50s

Contig lengths
Sum = 4199
Sum/2 = 2099.5

853	747	646	599	533	494	327
1600	2246	2845	3378	3872	4199	

N50 = 646

The higher the N50 the better

Joining the contigs

Chromosome walking

Look for singled
paired-ends/mate-pairs

Ye good olde PCR!

```
Aligned Reads
File Navigate Info Color Dim Misc Sort Help
core ace.1 Hel03 Some Tags Pos: clear
Search for String Compl Cont Compare Cont Find Main Win Err/10kb: 100.00

182,510 182,520 182,530 182,540 182,550 182,560 182,570 182,580
CONSENSUS attatttctagacacgtatgataaagtacgggtgtggaagtgtggaagagatagtt
HWI-EAS269:1:89:1012:1876#0/1 ATTATTTCTAGACACGTATgataAagtacGGtGtgaagtgtggaagagatagttaggcagatg
HWI-EAS269:1:26:1581:1958#0/2 attatttctagaCACGtatGataaagtacGGgtgtgaagtgtggaagagatgtagagcagtagctg
HWI-EAS269:1:36:1774:1328#0/2 ATTATTTCTAGACACGTATGATAAGTACGGgtGtgaagtgtggaagagatagttaggcagatgatgat
HWI-EAS269:1:63:457:1101#0/2 ATTATTTCTAGACACGTATGATAAGTACGGgtGtgaagtgtggaagagatagttaggcagatgatgatggg
HWI-EAS269:1:59:19:1229#0/2 ATTATTTCTAGACACGTATGATAAGTACGGgtGtgaagtgtggaagagatagttaggcagatgatgatggg
HWI-EAS269:1:26:1087:1343#0/1 ATTATTTCTAGACACGTATGATAAGTACGGgtGtgaagtgtggaagagatagttaggcagatgatgatggg
HWI-EAS269:1:12:1197:1941#0/2 attatttctagaCACGTATGATAAGTACGGgtGtgaagtgtggaagagatagttaggcagatgatgatggg
HWI-EAS269:1:10:900:1452#0/2 ATTATTTCTAGACACGTATGATAAGTACGGgtGtgaagtgtggaagagatagttaggcagatgatgatggg
HWI-EAS269:1:41:620:1574#0/2 ATTATCCTAGACACGTATGATAAGTACGGgtGtgaagtgtggaagagatagttaggcagatgatgatggg
HWI-EAS269:1:69:953:571#0/2 ATTATTT
HWI-EAS269:1:61:671:794#0/1 ATTATTTCT
HWI-EAS269:1:63:941:175#0/2 ATTATTTCTAGAC
HWI-EAS269:1:48:1723:1088#0/1 ATTATTTCTAGACACGTATGATAAGTAC
HWI-EAS269:1:58:1:982#0/2 ATTATTTCTAGACACGTATGATAAGTAC
HWI-EAS269:1:53:410:426#0/1 ATTATTTCTAGACACGTATGATAAGTACGGGTGT
HWI-EAS269:1:50:441:669#0/2 ATTATTTCTAGACACGTATGATAAGTACGGGTGTGAAGTTG
HWI-EAS269:1:10:975:1333#0/1 ATTATTTCTAGACACGTATGATAAGTACGGGTGTGAAGTTG
HWI-EAS269:1:58:860:791#0/2 attatttctagacacGTATGATAAGTACGGGTGTGAAGTTG
HWI-EAS269:1:61:452:946#0/1 ATTATTTCTAGACACGTATGATAAGTACGGGTGTGAAGTTG
HWI-EAS269:1:61:1529:957#0/2 ATTATTTCTAGACACGTATGATAAGTACGGGTGTGAAGTTG
HWI-EAS269:1:40:723:1421#0/2 AttatttctagacacGTATGATAAGTACGGGTGTGAAGTTG
HWI-EAS269:1:63:1204:1784#0/1 AttatttctagacacGTATGATAAGTACGGGTGTGAAGTTG
HWI-EAS269:1:61:424:675#0/1 AttatttctagacacGTATGATAAGTACGGGTGTGAAGTTG
HWI-EAS269:1:53:276:1104#0/2 AttatttctagacacGTATGATAAGTACGGGTGTGAAGTTG
Reads sorted by strand and then position
dismiss
```

EST assemblies **differ** from genomic ones

Coverage is not uniformly distributed

Poly(A) mRNA tails

mRNAs can be short

The software is not yet ready

Getting help!

SEQanswers

<http://seqanswers.com/>

How to use SSH in MS Windows?

Look inside *GDMW_Using_X_from_Windows.pdf*

Download and install the software

Follow the steps

MacOSX? You'll only need an SFTP program

How to connect to our server?

Look inside *GDMW_Server.pdf*

Form 8 teams

Get a unique username for your team

Connect via SSH

Server name: *bigdaddy.zoology.ubc.ca*

Port: *22*