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

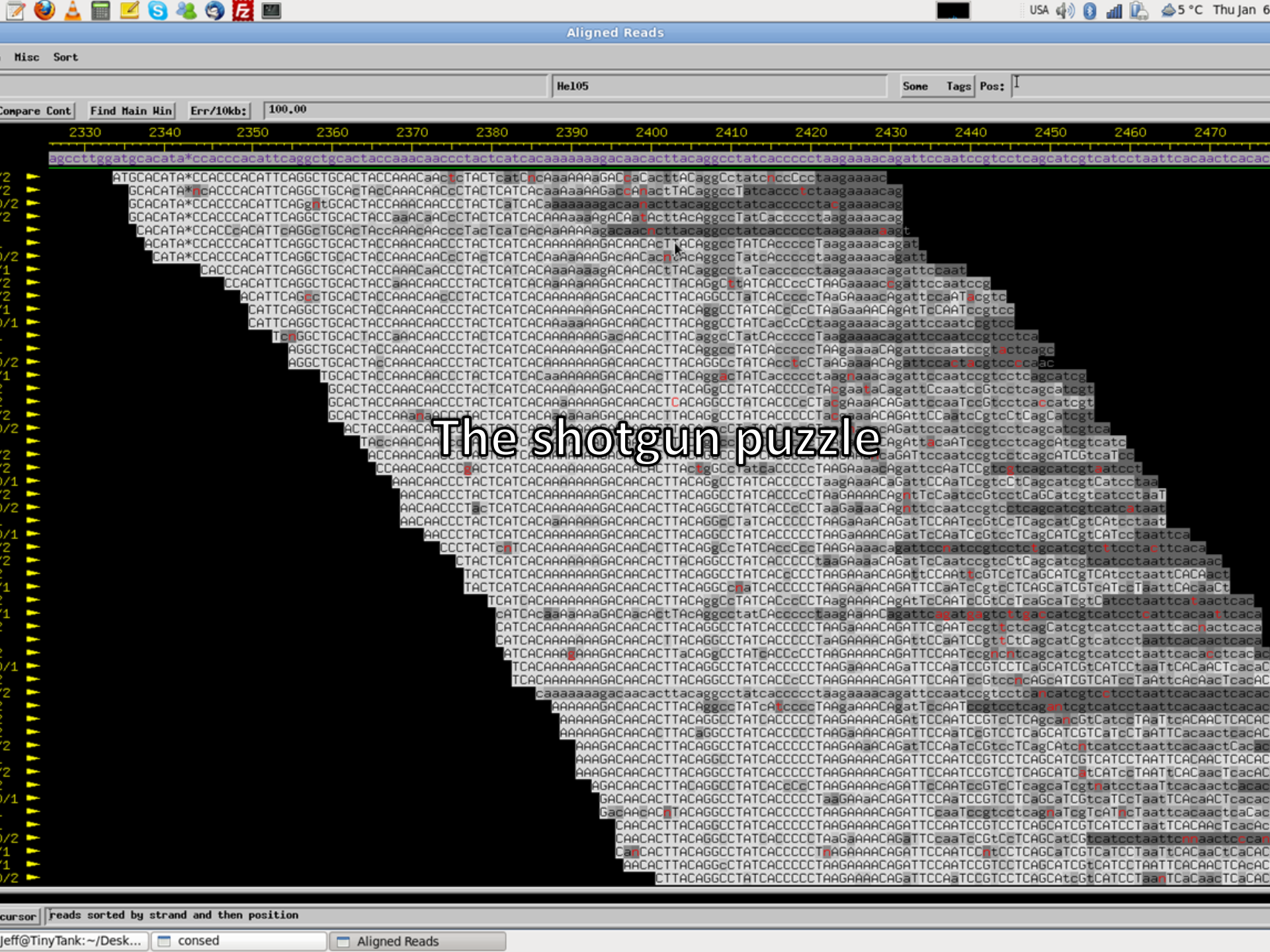
VIII. Phylogeny

Day 5

VIII. Phylogeny

WIFI Network code
SquawkinGood

V. Genomic/EST assembly


Aligned Reads
Misc Sort
Compare Cont
Find Main Win
Err/10kb:
100.00
2330 2340 2350 2360 2370 2380 2390 2400 2410 2420 2430 2440 2450 2460 2470
The shotgun puzzle
reads sorted by strand and then position
Jeff@TinyTank:~/Desk...
consed
Aligned Reads

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 != 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

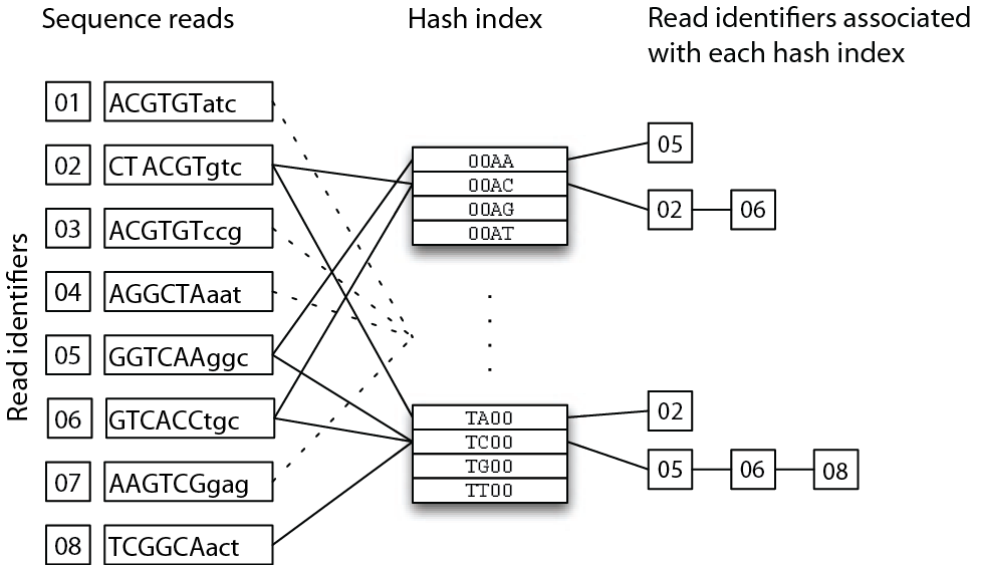
J. R. Miller *et al*. *Genomics*. **95**, 315-327 (2010)
P. Flicek & E. Birney. *Nature Methods Supplement*. **6:**11, S6-S12 (2009)

Hash-based algorithms

Reads or genome can be hashed

Tolerate high polymorphism

Mapping algorithms



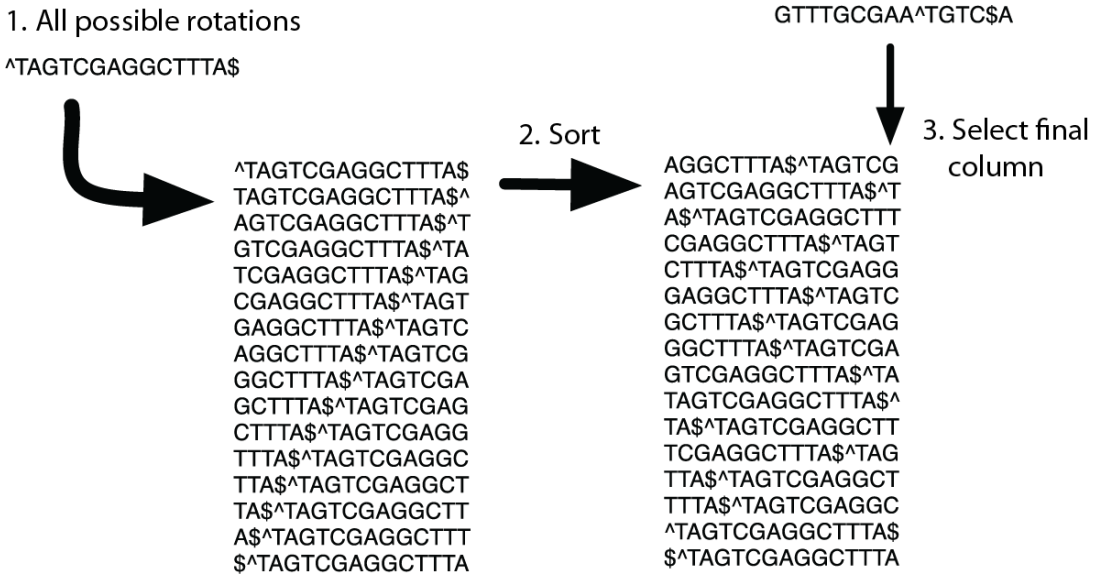
Burrow-Wheelers transform methods

Faster

Require less RAM

Struggle with high polymorphism

Mapping algorithms



The Shortest Common Superstring (SCS)

Greedy

All-against-all

Pairwise alignment scores

Used in old Sanger assemblies

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

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."

E. Birney. *Nature Methods*. **8**, 59 (2011)

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 TATTTGTAGCTGACGCTAGCTAGCTNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNGTACGTGCATCGCAGTCGATGCGTC

25 bases 20 bases 25 bases

In practice TATTTGTAGCTGACGCTAGCTAGCTNNNNNNNNNNNNNNNNNNGTACGTGCATCGCAGTCGATGCGTC

25 bases 40 bases 25 bases

or TATTTGTAGCTGACGCTAGCTAGCTNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNGTACGTGCATCGCAGTCGATGCGTC

Standard Deviation

The usual scaffolding errors

Perfect

AGCTGTCTGTTTTCTGTAGCTCGTATTACATATCGATGGA
AGCTGTCTGTTTTCTGTAGCTCGTA
AGCTGTCTGTTTTCTGTAGCTCG
TGTAGCTCGTATTACATATCGATGGA
AGCTCGTATTACATATCGATGGA

Overestimation

AGCTGTCTGTTTTCTGTAGCNNNNNNNTCGTATTACATATCGATGGA
AGCTGTCTGTTTTCTGTAGCTCGTA
AGCTGTCTGTTTTCTGTAGCTCG
TGTAGCTCGTATTACATATCGATGGA
AGCTCGTATTACATATCGATGGA

Underestimation

AGCTGTCTGTTTTCTGTTATTACATATCGATGGA
AGCTGTCTGTTTTCTGTAGCTCGTA
AGCTGTCTGTTTTCTGTAGCTCG
TGTAGCTCGTATTACATATCGATGGA
AGCTCGTATTACATATCGATGGA

An iterative process

N50s

Contig lengths **853** **747** **646** **599** **533** **494** **327**

Sum = 4199

Sum/2 = 2099.5

1600 **2246** **2845** **3378** **3872** **4199**

N50 = 646

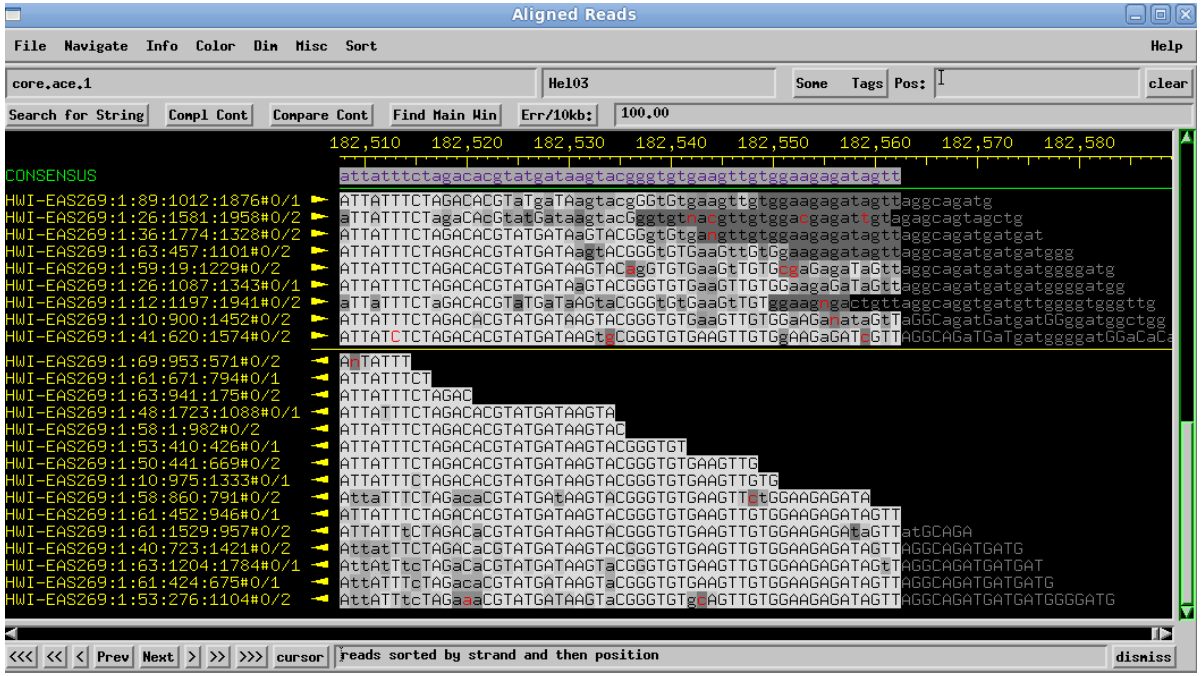
The higher the N50 the better

Joining the contigs

Chromosome walking

Look for singled paired-ends/mate-pairs

Ye good olde PCR!



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*