Sequencing and assembly of small eukaryotic genomes

Dr. Tom Hsiang
University of Guelph
Ontario, Canada
email: thsiang@uoguelph.ca

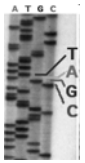
Presentation Outline

- Past & NextGen (next generation) sequencing
  - before & now
- Research Objective
  - sequence fungal genomes
- Methods
  - timeline of research
- Results
  - output of several assembly programs
- Conclusions
  - can it be done? now? later?
- Flowchart for genome sequencing & assembly

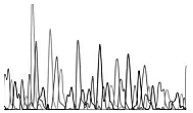
Sequencing Technologies

- 1970s: Gilbert (PNAS74:560), Sanger (PNAS74:5463), Sanger chain termination method
- 1987, ABI first automated sequencer
  - ABI 377 gel based, ABI 3700 capillary based
  - now up to 700 bp reads

autoradiograph



chromatogram



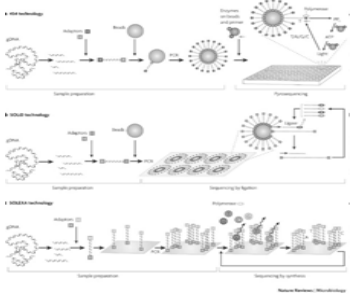
Sequencing Technologies

Next Generation Sequencing (NGS) Technologies

- 2005 - 454 pyrosequencing
  - » 150-200bp reads (now > 400 bp)
- 2006 - Solexa seq by synthesis
  - » 35bp reads, now > 100 bp
- 2008 - SOLiD = seq by oligo ligation & detection
  - » 50 bp reads

Next Gen Sequencing (nat.rev.micro6:419)

- 454
- SOLiD
- SOLEXA



Genome assembly statistics for plant, animals, fungus and bacterium and $

	Maize	Horse	Panda	*Grosmannia clavigera*	*Pseudomonas syringae*
Genome length	2.3 Gb	2.6 Gb	2.5 Gb	32.5 Mb	6.1 Mb
Sequencing technology	Sanger	Sanger	Illumina	Sanger/454/Illumina	Illumina
Number of contigs	125,325	55,316	198,274	3,361	1,346
Contig N50	40 kb	112 kb	40 kb	32 kb	11 kb
Number of scaffolds	61,161	9,687	81,469	2,322	71
Scaffold N50	76 kb	46 Mb	1.3 Mb	782 kb	317 kb
~ Sequencing cost	$30 million	$15 million	$0.6 million	$100,000	$4,000

Jackman & Birol (2010, Genome Biology 11:202) in US dollars $

"Contig or scaffold N50 is a weighted median statistic such that 50% of the entire assembly (bases) is contained in contigs or scaffolds equal to or larger than this value"

Research Objective

- use Illumina-Solexa technology to obtain the complete genomic sequences for 6 ascomycetes, and assemble with various programs

Hypotheses

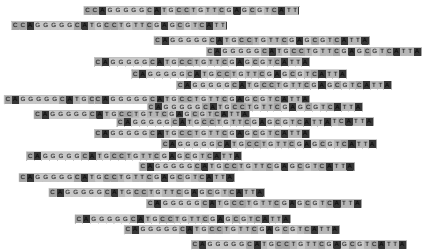
- Small genomes can be sequenced and assembled without a reference sequenced genome using Illumina technology & genome coverages of 100x and paired-end sequencing
- Genes such as mating type genes can be found in such assemblies, to make inferences about life cycles in nature

What is Genome Coverage?

$$\frac{\text{the total number of bases sequenced}}{\text{number of bases in genome}}$$

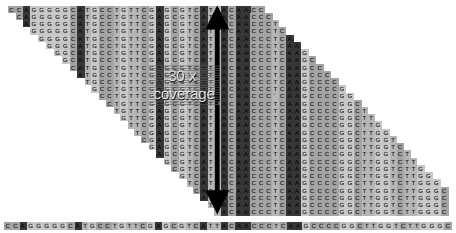
e.g. 25 million reads of 100bp length of a 50 Mb genome gives 50 x coverage

Assembly of 35 bp reads (millions)



Assembly of 35 bp reads (millions)

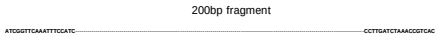
Assembly program aligns reads



Creates contigs (consensus sequences)

What is Paired-end sequencing?

- sequence both ends of DNA fragment (>200 bp)
- gives more information on the position of the small fragments, and better mapping of ambiguous reads than for single ends



- skip over regions difficult to sequence

Methods – Ascomycota tested

Species	Class	Order
Diplodia pinea	Dothideomycetes	Botryosphaeriales
Discula destructiva	Sordariomycetes	Diapothales
Gaeumannomyces cylindrosporus	Sordariomycetes	Magnaporthales
Neonectria faginata	Sordariomycetes	Hypocreales
Sclerotinia homoeocarpa	Leotiomycetes	Helotiales
Volutella buxi	Sordariomycetes	Hypocreales

Methods

- Spring 2011 (my lab)
  - grew and harvested mycelium (500 mg) of fungal hyphae grown on cellophane over PDA
  - extracted DNA with Qiagen DNeasy kit
- June 2011 (my lab)
  - sent to Sequencing Center 1 µg of genomic DNA of each of six fungal species
- July 2011 (Sequencing Center)
  - created genomic sequencing libraries
  - 3 lanes with 2 libraries per lane of 100bp paired-end runs on an Illumina-Solexa GAIIx machine

Methods

- Sept 2011 (Sequencing Center)
  - gave us data on external hard drive (2 Tb)
- Oct & Nov 2011 (my lab)
  - assembled reads with 3 assembly programs
    - » Abyss
    - » SOAPdenovo
    - » Velvet
- Dec 2011 (my lab)
  - predicted genes with Augustus (prediction pgm)
  - searched for mating type genes in assemblies

Results

- What I Hoped for
  - assembly of a few hundred contigs
- What I got
  - millions of 100bp records (fastq)

```
@SOLEXA5:2:120:19806:18845#0/1
TACTAGAA.CTCTAATGATATTACTATACCCTCAGTAATTAACTGGGCCACTTAAAAAGTTTCCTTTAAAAGTTTA
+SOLEXA5:2:120:19806:18845#0/1
ccccа^_UBa``b``caccccccccccccccccacc^cccccb`^caacccTcccbaaac`^cbccYcbYY^`abb^
```

Results – sequence reads

Species	Reads
Diplodia pinea	27,291,386
Discula destructiva	28,036,272
Gaeumannomyces cylindrosporus	29,479,478
Neonectria faginata	30,697,374
Sclerotinia homoeocarpa	28,436,528
Volutella buxi	35,064,118

start 100 bp seq — 300 bp non-sequenced fragment — end 100 bp seq

Assembly stats: *Diplodia pinea*

Program	Contigs	N50
Abyss 1.2.7	12,818	44 kb
SOAP 1.04	1,705	74 kb
Velvet 1.1.05	17,811	5.9 kb

Contigs = number of assembled pieces from raw reads
Highest N50 calculated from various K-values, cutoffs, etc.

Results – N50 assembly stats

Species	Abyss	SOAP	Velvet
Diplodia pinea	44 kb	74 kb	5.9 kb
Discula destructiva	21 kb	35 kb	9.0 kb
Gaeumannomyces cylindrosporus	58 kb	82 kb	23 kb
Neonectria faginata	13 kb	18 kb	1.4 kb
Sclerotinia homoeocarpa	48 kb	57 kb	10 kb
Volutella buxi	44 kb	56 kb	7.1 kb

Results – assembly size (Mb)

Species	Abyss	SOAP	Velvet
Diplodia pinea	38.4	37.0	38.3
Discula destructiva	47.9	47.2	49.0
Gaeumannomyces cylindrosporus	44.5	42.5	43.4
Neonectria faginata	48.5	44.0	44.8
Sclerotinia homoeocarpa	45.9	44.7	45.6
Volutella buxi	29.0	29.3	29.5

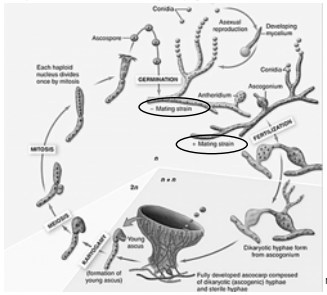
Results – assembly stats

Species	Reads	~Size	Coverage
Diplodia pinea	27,291,386	38 Mb	144 x
Discula destructiva	28,036,272	48 Mb	117 x
Gaeumannomyces cylindrosporus	29,479,478	43 Mb	137 x
Neonectria faginata	30,697,374	46 Mb	133 x
Sclerotinia homoeocarpa	28,436,528	45 Mb	126 x
Volutella buxi	35,064,118	29 Mb	263 x

Results – gene predictions with Augustus

Species	Genes
Diplodia pinea	~13,000
Discula destructiva	~12,000
Gaeumannomyces cylindrosporus	~13,000
Neonectria faginata	~14,000
Sclerotinia homoeocarpa	~ 9,000
Volutella buxi	~ 9,000

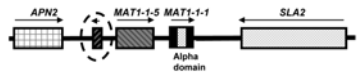
Ascomycete mating types



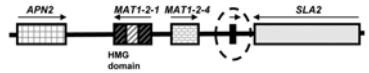
McGraw-Hill

Heterothallism

- two genotypes of different mating type needed for sexual reproduction (MAT1-1 and MAT1-2)



Heterothallic *Botrytis cinerea* MAT1-1

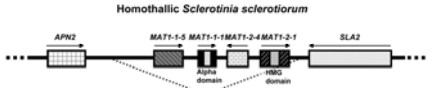


Heterothallic *Botrytis cinerea* MAT1- 2

Plos Genetics e1002230 (2011)

Homothallism

- both MAT1-1 and MAT1-2 present in a genotype, which can produce ascospores by itself



Plos Genetics e1002230 (2011)

Results: mating type genes

Contigs from the various assemblies were compared against many ascomycete MAT1-1 & MAT1-2 genes

Isolate	MAT1-1	MAT1-2
Diplodia pinea SP16a	Yes	Yes?
Discula destructiva 10115	No	Yes?
Gaeumannomyces cylindrosporus 08145	No	Yes?
Neonectria faginata 11171	Yes	Yes?
Sclerotinia homoeocarpa SH84	No	Yes?
Volutella buxi 09052	No	Yes?

Results: mating type genes (Amy Shi)

- *Volutella buxi* causes blight on boxwood
- MAT1-2 primers were designed for For *V. buxi*
- Among isolates from Ontario, 16 were found to have MAT1-2, and another 12 did not
- Contig containing the MAT1-2 sequence also had the two genes surrounding it, so primers were made from these to amplify a 12 kb fragment
- Conserved MAT1-1 primers designed from other species and used in a nested PCR to fish out the MAT1-1 gene (did not work on genomic DNA)
- sexually reproducing? (16:12, Mat1-1 : Mat1-2)

Results: mating type genes (M. Stanescu)

- *Discula destructiva* causes dogwood anthracnose
- MAT1-2 primers were designed
- all isolates tested to date (~20) have MAT1-2 including ones from Ontario and B.C.
- may be capable of sexual reproduction, but other mating type was not introduced when the fungus arrived in North America in the 1970's?

Conclusions

- many secrets in a genome
- if a small genome is already sequenced
  - assemble another isolate for < $1000
- If not sequenced yet (*de novo*)
  - need >100x coverage and different sized libraries (paired-end 200bp, mate-pair 3kb, >$2000?)
- Is a $1000 *de novo* genome possible?
  - Maybe, but only if your team learns how to put the billions of nucleotides and millions of pieces together!

Flowchart of a genome sequencing and assembly process

- Figure 8 (with text explanations) from
  - Haridas S, Breuill C, Bohlmann J, Hsiang T. 2011. A biologist's guide to de novo genome assembly using next-generation sequence data: a test with fungal genomes. Journal of Microbiological Methods 86:368-375.
  - available at www.uoguelph.ca/~thsiang/pubs
  - glossary of sequencing-related words provided

Flowchart of genome sequencing & assembly

4.1) Prepare DNA for submission
- regular DNA extraction methods with kits are sufficient

⇓

4.2) Select a depth of coverage and specify library fragment size
- a single lane of 100 bp reads may be sufficient *(GAIIx = 10 Gb/lane)*
- *HiSeq-2000 (30 Gb/lane) multiplexed (up to 12 samples in one lane)*
- typical insert sizes are 200 bp to 10 kb
- give these instructions to sequencing center to make DNA library *(typically $200 per library, and $2500-3000/lane)*

⇓

4.3) Set up or obtain access to computing facilities *(& learn to use)*
- set up Linux on a PC with >16 Gb RAM or
- obtain access to a High Performance Computing Cluster

⇓

Generation of sequence data at sequencing center (2-4 mo)

NGS genomic DNA library preparation

- get fragments of target size (e.g. 500 bp, with the two 100bp ends sequenced)
- current rate-limiting step for NGS
- Illumina library prep ~$250, or you can do it yourself with kits for < $100

Flowchart of genome sequencing & assembly

Generation of sequence data at sequencing center

⇓

4.4) Download or copy and convert the raw sequence data
- use software to download the gigabytes of data
- convert the data format for processing

⇓

4.5) Choose assembly program (e.g. Abyss, SOAP, Velvet)
- choose between several (freely) available software programs (Linux)

⇓

4.6) Choose k-mer for assembly
- lower values have higher sensitivity, while higher values have higher specificity
- scripts can be used to produce a wide range of k-mer assemblies

⇓

4.7) Accept that your Illumina assembly is incomplete
- current technological limitations cannot overcome issues of repetitive regions