Implementation

Are you getting what you want out of workshop?
What would you like to know about this topic ?
What do you fear with respect to NGS/informatics? 1 minute free write exercise…

CAVEAT/TAKE HOME MESSAGE: There is no "Easy Button" or permanent cookbook because the field is moving faster than Moore's law… life long learning…so I'll emphasize principles over "specific tips" and hope to address questions…

J. Chris Pires – University of Missouri

NGS Workshop Botany 2012

What is in a $100, $1,000, $10,000, and $100,000 genome?

An Illumina High-Seq lane now outputs 185-190M reads per lane
(some centers are reporting 200M+).
After quality trimming and removing "read though adapter contamination"
(typically removes ~5-10% of data),
we typically recover ~170M reads per lane.
(see where new technology is in 6 months), So:

$100 genome

The $100 genome (sequence cost only) would be ~16 libraries per lane of 1 x 100.

This is more than sufficient to sequence both organelle genomes (e.g. 3-5 plastid contigs) to recover CDS (genome skimming, GSS, Ultra-barcoding) and rDNA, some novel repeats, etc.

You can barcode 3X per lane (48 libraries) and still recover all those CDS (plastid gene space)

*NOTE: This highly depends on your input DNA and to a lesser extent on genome size.

$200 genome

The $200 genome (sequence cost only) would get paired end (PE) data which would be optimal to get "full circles" for structural evolution of plastids and mitochondria. In addition to organelles and rDNA loci, would get millions bases of unique assembled sequences from the nuclear genome … (repetitive elements have interesting natural history also…)

De novo assembly followed by "Reference-Based Scaffolding" (to orient overlapping contigs).

$1,000 genome

The $1,000 genome (sequence cost only) would be ~1/3 of a PE lane on the HiSeq.

Applications: Definitely sequence chloroplast and mitochondria and find repeats.

Resequence an *Arabidopsis thaliana* ecotype or EMS mutant and align reads to reference genome (e.g., Schneeberger et al. 2011 PNAS and SUPERLOCAS workflow)

$10,000 genome

The $10,000 genome (sequence cost only) would be ~4 PE lanes (1/2 flow cell).

Applications: Resequence a *Brassica oleracea* morphotype (have reference sequence for cabbage, now want to find SNPs for broccoli, cauliflower, kale, and kohlrabi).

…Epigenomics (bisulfite sequencing)

…metagenomics/microbiome of plant roots/etc

Gene space/light draft genome for non-model species(this is what DOE JGI often starts with for plants).

*Depends on genome size for coverage, which needs

$100,000 genome

The $100,000 genome (sequence cost only) For example; Integrate methods discussed this morning:

Follow "All-Paths" Recipe for sequencing with many insert sizes; mix in 454/nanopore (GC bias, long reads)

Do GBS based genetic map (cheap, $30 per line)

Do tissue-specific transcriptomics to annotate genome

Do a "Keygene" style-physical map ($250,000?)

Applications: Draft sequence of a non model genome! Milkweed, Venus fly trap, insert-your-favorite-organism here Can do this with a "standard" NSF grant !

***Caveat: Depends on genome size for coverage**

****But "genome browser" may be another $100,000….**

Sequence depth vs genome coverage

Note that 5x sequencing depth does not mean 5x genome coverage. An example from a recent human genome resequencing project:

When the sequencing depth is 30X, only half of the regions (51%) are covered at above 30X. While at 100X and 200X sequencing depths, a higher percentage (81% and 90%) covered.

So even 200X sequence depth results in "only" 90% of the genome being "covered".

TRANSCRIPTOMES and RNA-SEQ

For gene discovery (normalized libraries), 1/4 PE lane of Illumina Hi-Seq is perfect for de novo assembly of any organism (routine)

****** Quality of the RNA input is crucial ******

1/6 SE lane x 50bp reads is more than sufficient to quantify expression (~$180 per library).

See you later PCR!!! Results now like REAL TIME PCR for all Gene Models (plus splice variants).

Phylogenomics: Things you can look forward to thinking about …

Imagine now you have 200 to 20,000 nuclear gene phylogeny from either transcriptomes, hyb-seq, or comparing draft genomes…

You construct individual gene trees and do not agree;

However, you get one highly supported MP/ML/Bayes tree…

What does that mean ? Are you happy now ?

- Rokas et al. 2003 and letters he received
- Drosophila 12 genomes paper – have all data and still not everything resolved – back to old philosophical questions of identity/lineages and networks/trees…

As botanists, we hav got to love the gray zone :)

From Systematics to Systems Biology

Good news! Genomics is just the start, integrating all the other –omics is coming fast for non-model organisms (phenomics, metabolimcs, proteomics, etc)

My strategy is to move tools from model organisms to closely related non-model organisms

e.g. Arabidopsis to Brassica to Brassicales

Exciting because in my lifetime, if not very soon, we will be able to do "**systems biology of Venus Fly Trap**" Used to take 6-12 people to make databases for fly, yeast, human; But now informatics modules to drag into non-models; "smart phone apps"

And good because future jobs in this area…

What kind of NGS sequencing do you want to do ?

One minute free write exercise…

Think, pair share, questions ?

Halfway thru talk…
Now for the bad news…
Informatics – pace of change is fast, requires new skills/training, and serious computational resources

As NGS technologies move rapidly with new platforms out every 6-12 months, any specific informatics skill sets have a short half life and can't "retool" every few years on sabbatical… you have to constantly keep up on new methods – so how can you keep up?

What are the new rate-limiting steps?

We are no longer (sequence) data limited, and with other 'omics datasets and even phenotyping becoming high-throughput,

what are new rate limiting steps?

One minute brainstorm…

then pair/share….

Questions?

What are the new rate-limiting steps?

We used to be data limited, for my PhD we spend most of time in lab doing Sanger Sequencing… now spend one week on sample preps and Illumina sequencing and can get enough data for a publication (and well-trained undergrad can generate data, but no $10,000 mistakes and need excellent note taking – don't want to sequence the wrong genome…e.g., grape…)

Now we are data-management and analysis limited….

Lab is empty while doing six months of bioinformatics…

great because we can ask many question with large data sets

MOST FRUSTRATING PART: Getting computational resources

DYI: make your own 1 Tb local playground?

Use or start a core campus network?

Cloud resources? (NESCent, iPlant, NSF EXCEDE, Amazon, etc)

Our lab uses all of these (our lab, campus, cloud)

Things I wish I knew five years ago 1

- Good questions and biology with right organisms trump technology every time; and with informatics, phenotyping and developing genetics resources is now the rate limiting step (tell every grad student to start selfing or making DH lines, develop mapping populations and diversity sets)
- In converse, to those who are adverse to new technology, just know that it lets you go genome wide with any biological question, and not just single-gene analyses…
- Solution: balance your enthusiasm with technology with your original passion/questions about natural history

Things I wish I knew five years ago 2

Collect high-quality DNA and RNA now (test!)

Can barcode/index libraries to see if good before doing a lot of them

(also test libraries / do more than one library)

By time finish analyses; obsolete methodologically!

We've done transcriptome assemblies four times now for Brassicales because method gets upgraded every 6 months, so get the "pipeline down" and publish ASAP!

(Horror story/problem with OneKp project!)

SOLUTION: Don't start sequencing a lot until pilot sequencing and informatics experiments done and ready to write it up!

Things I wish I knew five years ago 3

Don't believe everything you hear but test alternative methods & get multiple opinions (wasted lots of time in lab and on computer doing it how someone else did it 2 years ago; e.g., plastome isolations, not quality trimming data, reference based assembly)

False advertising rampant For example,

"Genome Hype" of lastest sequencing platform

(e.g., 454, Pac-Bio, etc) or informatics approach

(e.g., SoapDeNovo-trans, etc).

Solution: Do pilot experiments!

Things I wish I knew five years ago 4

It takes a village because field is moving fast and increasingly interdisciplinary; so train yourself and student on how to collaborate (we have a SKYPE call with somebody every other week at least) – with the pace of technology changing, as the PI or even your lab group will be challenged to stay up to speed – so cross-train, collaborate, send yourself or students to other labs, and out source as needed...

We are all in this together; few of us are computer scientists, just can't be afraid of the computer and making friends who can help…

GET NETWORKED! (e.g., YASRA pre-publication distribution)

Call people on phone, SKYPE, etc. (e.g., as our lab moves into Hyb-Seq, we'll contact a half dozen labs as get started…)

NGS Wrapup: Old and New Lessons

Old school lessons that still apply:

Have a good biological question

Collect metadata in field, greenhouse, lab (and as use your computer – keep a journal!)

Use a sound experimental design (see statistician, people forgetting lessons from array days as move into RNA-Seq….)

New lessons to consider:

Need Informatics to handle large data sets

New interdisciplinary training now required (CS)

Computational resources needed

Implementation: take home messages

Garbage in/garbage out: Lab: Quality of RNA/DNA input ("real garbage") ... can't quality trim your way into a good assembly.

Spend time analyzing input quality: Gel & Bioanalyzer.

Computer: be sure to always quality trim your data – no excuses ! ("Lazy does double…")

You can outsource the wet lab work, but hold on to your natural history and informatics

Learn to love command line/basic scripts !

No way around this, Don't be scared - you are not alone !

Go make friends with people in your computer science department and co-teach a class – it is fun!

Acknowledgements

- **NSF (MonAToL, Systematics), DOE JGI/Brassica genomics**
- **NESCent, iPlant/OneKp,**
- **Collaborators: People we SKYPE a lot with…Jim Leebens Mack, Claude dePamphilis, Liston & Cronn labs, others…**

My lab: Dustin Mayfield (de novo GSS assembly, undergrads)

Tatiana Arias (quantitative RNA-seq expression/statistics/R)

Patrick Edger (spends 20 hours per week on sequencing &informatics blogs; de novo transcritpome assembly, genomes)

Kate Hertweck (repetitive DNA specialist, wants your garbage)

Roxi Steele (how to make conservation genetics cheaper)

Computer science grads/rotating students/Undergrads: They are fearless and not afraid to try things…make us better

Questions about implementation ?

Are you getting what you want out of workshop?

What would you like to know about this topic ?

What are you terrified of with respect to NGS sequencing and phylogenomics?