Diving In!
NGS Implementation

J. Chris Pires – University of Missouri

Email: jchrispires@missouri.edu

Twitter: @jchrispires

Today: 1 Think about projects and goals before you start

2 Thinks I wish I knew when I started after Aaron Liston told me about Solexa sequencing…

NGS Workshop

New Orleans - Botany 2013

Diving In: NGS Implementation

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

…and now for many ugly text slides – all will get posted and distributed online so do not try to write down…

The Future of Species Identification

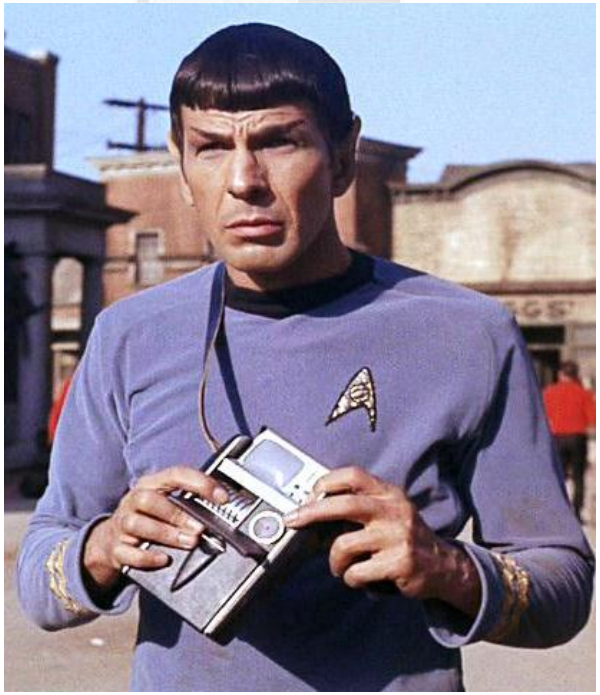
Identification Using Whole Genome Sequences

- **Systematics, Ecology, Conservation**
- **Forensics/Border patrol**
- **Restoration ecology**
- **Invasive species control**
- **Citizen Science**

Current tool: iPhone and Bench-top sequencer

Time Travel !

Microbiome !



Future tool: Hand-held field biocorder

(Steele and Pires 2011 Amer. J. Bot.)

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

An Illumina High-Seq lane now outputs 185-200M+ reads per lane

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 CDS (plastid gene space)

***NOTE: This highly depends on your input DNA**

$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, one gets millions of bases of unique assembled sequences from the nuclear genome (repetitive elements have interesting natural history)

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

See Michelle Tang Poster # PGP005

$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)

$10,000 genome

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

Applications: Resequence a Brassica oleracea (if have reference sequence for cabbage, now want 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 to be calculated

$100,000 genome

The $100,000 genome (sequence cost only) --- Follow "All-Paths" Recipe for sequencing with many insert sizes; mix in mate pairs & long reads

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

Do tissue-specific transcriptomics to annotate

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; "genome browser" may be another $100,000; physical map be another $250,000?

Where is the real “cost” ?

The “real cost” is not sequencing or even library preparation; but in time and resources spent on bioinformatics and analyses.

Do you want just 80 chloroplast genes to build a phylogeny or want actual complete circles?

Similar reagant costs, but huge difference in analysis time.

Sequence depth vs genome coverage

Note that 5x sequencing depth does not mean 5x genome coverage. An example from the 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 sequenced genome being "covered".

Even “completely sequenced” human genome still missing up to 10% and still discovering CNV, new genes, etc….

TRANSCRIPTOMES and RNA-SEQ

For gene discovery, 1/4 PE lane of Illumina Hi-Seq is perfect for de novo assembly of any organism (routine, ~$500)

****** 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).

Asparagales Phylogeny

79 cp genes ML BS values below – Steele et al. 2012 (*=100)

? = not sampled below Iridaceae

19 nuclear genes ML BS values above -McKain et al. unpublished (DRAFT!)

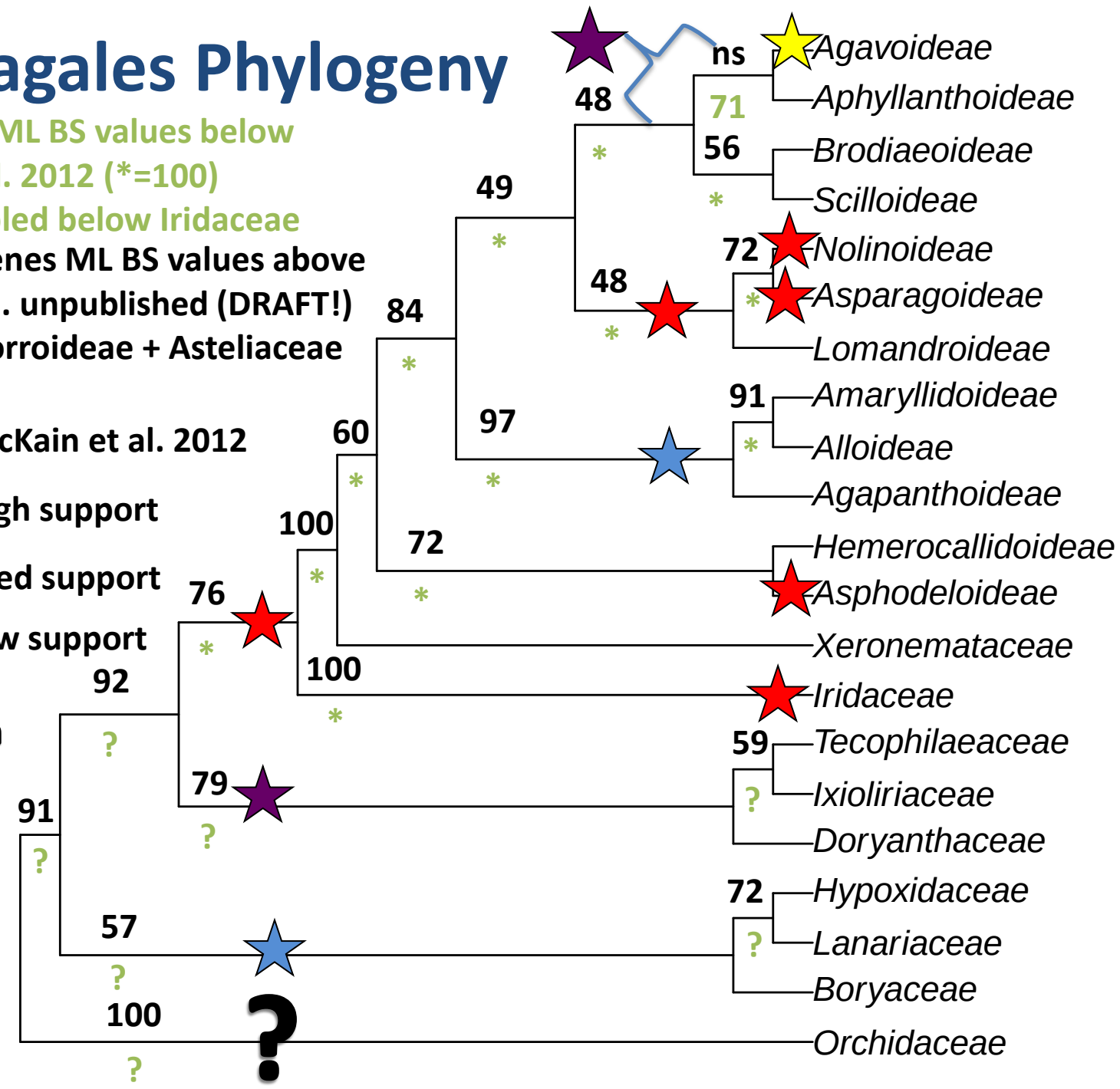
Note: Xanthorroideae + Asteliaceae Not sampled

= WGD McKain et al. 2012

= WGD high support

= WGD med support

= WGD low support



Sarah Unruh poster PSY060

Pause for Paradigm Shift

PLATE TECTONICS:
Nobody believed in "continental drift" in 1930,
(almost) everyone does today…

NUCLEAR GENES: Ten years ago we established orthology first, then sequenced nuclear genes, (would never consider using expression data for phylogenetic purposes)…
Now we do transcriptome sequencing first and then establish orthology without a problem





Phylogenomics: Things you can look forward to thinking about …

Imagine now you have 200 to 20,000 nuclear gene phylogeny from transcriptomes, hyb-seq, or draft genomes; and you construct individual gene trees and they do not agree; however, you get one highly supported MP/ML/Bayes tree…

What does that mean ? Are you happy now ?

Drosophila 12 genomes paper – have all data and still not everything resolved – back to old philosophical questions of identity/lineages and networks/trees… botanists love the gray zone :)

-Rokas et al. 2003 and letters he received; 2013

From Systematics to Systems Biology

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

My strategy is to move tools from model organisms to closely related non-model organisms; e.g. Arabidopsis to Brassica, etc. Exciting because 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"

Halfway thru talk;
What kind of NGS sequencing do you want to do given your goals?

One minute free write exercise…

- **What do you want to do with NGS given your goals or questions?**
- **What do you fear with respect to NGS/informatics?**

Think, pair share, questions ?

Now for the bad news… Informatics pace of change is fast, requires new skills/training, and often 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?

#CodeAndCoffee

So how can you keep up?

#CodeAndCoffee

If you are not doing bioinformatics,
you are not doing biology" – Twitter quote
Almost all of biology is moving to "Big Data"

Many ways to succeed-my goal is to train people

What are the new rate-limiting steps?

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

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

What are the new rate-limiting steps?

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)

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, so what are new rate limiting steps?

Things I wish I knew six 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 six 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 six 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 six years ago 4

It takes a village because field is moving fast and increasingly interdisciplinary; train students how to collaborate (we have a SKYPE call with somebody every other week – so cross-train, collaborate, send yourself or students to other labs, check each other's work (Trinity!) out source as needed... Learn how to talk to CS/Stats

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., got 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 we 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... can't quality trim to a good assembly. Spend time analyzing input quality

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 !

Make friends with people in your computer science department; co-teach a class - 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)

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

Kate Hertweck (repetitive DNA specialist, wants your garbage)

Roxi Steele (how to make conservation genetics cheaper)

Sarah Unruh, Michelle Tang, Kevin Bird / 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?