Hybrid enrichment – Anchored phylogenomics
Collection of High-Throughput Data

• Targeted Amplicon Sequencing or Parallel Tagged Sequencing
• Multiplex PCR
• Massively Parallel Uniplex PCR
• Reduced Representation Library or RAD Sequencing
• Transcriptome Sequencing
• Hybrid Enrichment

• BUT: more data = better data ??
<table>
<thead>
<tr>
<th>Method</th>
<th>Other names or variants</th>
<th>Literature method</th>
<th>Literature examples</th>
<th>Benefits</th>
<th>Drawbacks</th>
<th>Best application</th>
</tr>
</thead>
<tbody>
<tr>
<td>Amplicon sequencing</td>
<td>Multiplex PCR, parallel tagged sequencing</td>
<td>Binladen et al. (2007), Meyer et al. (2008), Tewhey et al. (2009b)</td>
<td>Chan et al. (2010), Griffin et al. (2011), Gunnarsdóttir et al. (2011), Morin et al. (2010), Parks et al. (2009)</td>
<td>Highly targeted. Results in nearly complete data matrices. Needed coverage easy to calculate. Circumvents individual sequencing reactions and phasing nuclear loci compared to Sanger sequencing Broad, random genomic sampling of thousands of independent genomic regions. Requires no prior genomic resources whatsoever</td>
<td>Requires PCR of each individual at each locus</td>
<td>Small- to medium-scale projects targeting a limited number of genes</td>
</tr>
<tr>
<td>Restriction-digest</td>
<td>Double-digest genome reduction, RAD sequencing (RAD-seq), complexity reduction of multilocus sequences (CRoPS), Genotyping by Sequencing (GBS)</td>
<td>Baird et al. (2008), Davey et al. (2011)</td>
<td>Andolfatto et al. (2011), Amaral et al. (2009), Bers et al. (2010), Emerson et al. (2010), Gompert et al. (2010), Hohenlohe et al. (2011), Hyten et al. (2010a,b), Kerstens et al. (2009), Ramos et al. (2009), Sanchez et al. (2009), Van Oorsouw et al. (2007); Van Tassell et al. (2008), Wiedmann et al. (2008); Williams et al. (2010)</td>
<td>Reproducibility and throughput may be limited by gel extraction step. Not targeted, thus coverage can be difficult to estimate. Null alleles could skew diversity estimates</td>
<td>Intraspecific studies of recent divergence</td>
<td></td>
</tr>
<tr>
<td>Target enrichment</td>
<td>Sequence capture, targeted resequencing, primer extension capture (PEC)</td>
<td>Albert et al. (2007), Gnirke et al. (2009), Hodges et al. (2007), Okou et al. (2007), Tewhey et al. (2009a), Maricic et al. (2010) Morin et al. (2008); Mariño et al. (2008)</td>
<td>Briggs et al. (2009); Faircloth et al. (in press)</td>
<td>Rapid collection of thousands of loci without individual PCR</td>
<td>Requires some prior genomic knowledge, but not necessarily a sequenced genome</td>
<td>Phylolgeomics at taxonomic levels above at and above the species level</td>
</tr>
<tr>
<td>Transcriptome</td>
<td>RNA-seq</td>
<td>Barbazuk and Schnable (2011), Canovas et al. (2010), Chepelev et al. (2009); Geraldes et al. (2011), Hittinger et al. (2010)</td>
<td>Morin et al. (2008); Mariño et al. (2008)</td>
<td>Can leverage data from expression studies</td>
<td>Skewed read distributions can outstrip coverage, making it difficult to find orthologous loci</td>
<td>Leveraging existing cDNA libraries</td>
</tr>
</tbody>
</table>
Comparison of Methods

- Hybrid Enrichment
- Whole Genome
- Transcriptome
- RRL & RAD
- Multiplex PCR
- Targeted Amplicon
- Massively Uniplex PCR
Ultraconserved Elements Anchor Thousands of Genetic Markers Spanning Multiple Evolutionary Timescales

BRANT C. FAIRCLOTH1,*, JOHN E. MCCORMACK2, NICHOLAS G. CRAWFORD3, MICHAEL G. HARVEY2,4, ROBB T. BRUMFIELD2,4, AND TRAVIS C. GLENN5

Anchored Hybrid Enrichment for Massively High-Throughput Phylogenomics

ALAN R. LEMMON1,*, SANDRA A. EMME2, AND EMILY M. MORIARTY LEMMON2
UCE workflow

a) UCEs identified in alignments of birds and lizard

b) Probes designed from UCE regions

- Probe
  - UCE < 180 bp
  - Flank added
- UCE = 180 bp
- UCE > 180 bp

RNA probes mixed with sheared genomic DNA from non-model organisms

d) Target DNA isolated, enriched, tagged, and pooled for NGS
e) Contigs assembled from NGS reads, aligned to probe, and consensus called for locus

f) Consensus loci aligned among species and gene trees estimated for all loci $j_1 \rightarrow n$

g) Species tree estimated from gene trees
Anchored Phylogenomics Workflow

1. Locus Selection
2. Probe Design
3. Data Collection
4. Assembly
5. Phylogenomics
Locus Selection & Probe Design

- **Locus Selection**
  - ~500 “single copy” loci (typically long exons)
  - Conserved element (~20% divergence required)
  - Adjacent to less conserved regions
  - Loci are selected based on broad taxonomic group (e.g., vertebrates)

- **Probe Design**
  - Incorporate sufficient number of lineages
  - Tile probes across conserved region
  - Goal is to capture ~1500bp regions
  - Probe sets are designed for project-specific clade
Phases of Hybrid Enrichment for Phylogenomics

1. Bioinformatics of locus selection
   - Choose target loci using genomes

2. Bioinformatics of probe design
   - Tile probes across target regions

3. Wet-lab sample processing
   - Genomic DNA $\rightarrow$ raw sequence reads

4. Bioinformatics of raw data analysis
   - Raw sequence reads $\rightarrow$ phased alleles, alignments, models of evolution, concordance analyses, and phylogenies
Solution-Phase Hybridization

Solution hybrid selection with ultra-long oligonucleotides for massively parallel targeted sequencing

Gnirke et al. 2009
Hybrid Enrichment

- **Disadvantages:**
  - Large equipment investment for small operations
    - Equipment required (Bioanalyzer, Covaris or other sonicator, qPCR machine, etc.)
  - Bioinformatic training required for locus selection, probe design, analysis of raw data
  - Substantial investment in reagents
    - Indexes for library preparation
    - Other library preparation reagents
    - Hybrid enrichment kit
Hybrid Enrichment

- **Advantages:**
  - Large quantity of data
  - Complete data matrices! Can trim out all missing data and still have more than enough
  - Fast data collection (DNA to phylogeny 2 weeks)
  - Works on degraded samples and ancient DNA
  - DNA starting material (RNA not needed)
  - Can use single probe design (kit) for broad taxonomic group (e.g., Vertebrata)
  - No problem for non-model systems
Own experiences: samples $\rightarrow$ trees

Requirement: > 1.5 - 2.0 micrograms of RNA-free DNA, quantified by Qubit
Output summary
nGenes = 434
Number of taxa = 9
Number of sites = 711063
Number of variable sites = 16986
Number of informative sites = 4373
Number of characters (total) = 6399567
% Missing characters (N's and -'s considered only) = 1.4503637511725402
% Missing characters (all considered) = 1.5992800762926618
Approaches to analyze phylogenomic data

Important issues to consider:

!!! amount of missing/ambiguous data
!!! alignment
! heterozygous data (phased vs. unphased!)

Concatenation versus species tree inference (coalescent analyses)

Concatenation:
- NJ – Geneious
- MP – TNT
- ML – RAxML (…), ~ FastTree
- BI – MrBayes, BEAST …
Coalescent-based Methods for Species Tree Inference

- **Summary statistic methods:** Start with estimated gene trees
  - Using estimated branch lengths:
    - STEM (Kubatko et al. 2009)
    - STAR, STEAC (Liu et al. 2009)
  - Using topology information only:
    - Minimize Deep Coalescences (PhyloNet; Than & Nakhleh 2009)
    - MP-EST (Liu et al. 2010)
    - ST-ABC (Fan and Kubatko 2011)
    - STELLS (Wu 2011)

- **Methods that utilize the full data:** Input is aligned sequences
  - BEST (Liu and Pearl 2007)
  - *BEAST* (Heled and Drummond 2010)
  - New method based on algebraic statistics (Chifman and Kubatko 2013)
Comparison of approaches:

- Summary statistic methods
  - Advantage: Quick
  - Disadvantage: Ignore information in data
  - Most current implementations do not easily allow for assessment of uncertainty

- Full data methods
  - Advantage: Fully model-based framework
  - Disadvantage: Computationally intensive, sometimes prohibitively so
  - Both BEST and *BEAST utilize a Bayesian framework and involve MCMC