Genomics for evolutionary inference

Emily Jane McTavish

Life and Environmental Sciences
University of California, Merced
emctavish@ucmerced.edu, twitter: snacktavish
How do you get from
You’ve seen a lot about how to get from
I’m going to talk about going from to and how those choices can affect to.
I’m going to talk about going from

to

to

and how those choices can affect
The quantity of available sequence data for inferring evolutionary relationships is increasing rapidly.

http://genome.wellcome.ac.uk/
“With the advent of modern molecular biology, the ability to collect biological sequence data has out-paced the ability to adequately analyze these data”

– Jeff Thorne (Evolutionary biologist)
“With the advent of modern molecular biology, the ability to collect biological sequence data has out-paced the ability to adequately analyze these data”

– Jeff Thorne (Evolutionary biologist)

Thorne et al., Journal of Molecular Evolution. 1991
http://genome.wellcome.ac.uk/
There are a lot of choices to make!
Biological questions

What do you want to know?
What do you already know?
Biological questions
What do you want to know?
What do you already know?

Technical questions
What data is right for our questions?
Is a closely related reference genome available?
How should we process and analyze our data?
What biases may be affecting our inferences?
General approach

- Decide what to sequence (TREE to CHAIN)
- Consensus sequence, alignment, locus selection (CHAIN to PAINT)
- Evolutionary analyses (PAINT to EVOLUTION)
- Success!
There are a lot of possible paths you can take!

Flowchart capturing genomic analysis pipelines used by participants at Trees in the Desert workshop, May 2019
What to sequence?
Different sequencing approaches enrich the samples for different components of the genome

Enrichment (smallest to largest proportion of genome)
- Directed PCR
- Targeted enrichment, Rad-tag etc
- Transcriptome
- Whole genome

Depending on your questions, any of these could be the best option!
Directed PCR

Simple and cheap for a small number of genes
Doesn’t scale so well to many genes
Doesn’t sound fancy
**Targeted enrichment** (e.g. Ultra-conserved elements, probes for orthologous single copy genes, etc.)

- Use hybridization to enrich particular regions
- Works well even on degraded DNA
- Need to synthesize probes specific to each region
  - need data to get data!
- Data sets can be combined across projects if same probe set applied
Non-targeted enrichment (RAD-tag, ddRAD etc.)

Select randomly distributed, but consistent, genome regions
Comparable across closely related taxa, but not more distant taxa
Each locus has very few variable sites (not good for generating gene trees)
Whole transcriptome

Enriched for expressed protein coding genes

Content will vary based on cell type, environment, etc.

Provides expression level data
Whole genome sequencing

Capture all the data

In a phylogenetic context, currently only cost effective for small genomes

Annotation is hard! Often need transcriptome to get genes

Mapping or assembly can be slow
Need to put the pieces back together!
Genomic sequencing

You have all the data! 👍
You have to deal with all of the data. 👎
**De novo assembly**

1. Fragment DNA and sequence

2. Find overlaps between reads

   ...AGCCTAGACCTACAGGATGCGCGACACGT
   GGATGCGCGACACGTGCATATCCGTT...

3. Assemble overlaps into contigs

4. Assemble contigs into scaffolds

(Baker, 2012)
Mapping to a reference genome

Reference Sequence

Layout

Consensus

Alignment
To make evolutionary statements, you need to align genomic regions across taxa.
Depending on evolutionary history this can be easy or hard!
To make evolutionary statements, you need to align genomic regions across taxa. Depending on evolutionary history this can be easy or hard!

(Darling et al., 2008)
An alignment is a statement of shared ancestry
Gene tree (Locus tree)

*The ancestry of a homologous region of the genome that has a single evolutionary history* (no recombination)

Enrichment methods focus our sequencing efforts on these regions
Gene duplication and loss

Orthology

Paralogy

Inference of homology is not incorrect! But our current models are limited. If you treat paralogs as orthologs, you can make incorrect inferences. figure from Casey Dunn
“investigation of genes with extreme support for turtle placement revealed unappreciated paralogy in a small proportion of alignments (<1%) that had an extraordinary influence on the inferred placement of turtles.”

(Brown and Thomson, 2016) (Chiari et al., 2012)
Challenge: The true (unknown) phylogenetic history is needed to assess orthology vs paralogy
“We used to be so blissful, back in the day. We just had data for one gene.”
Paul Lewis (this morning)
Integrated approaches can jointly estimate gene trees and species trees

\[ L(T, S, N|A) = \prod_{G_i \in G} L(G_i) \]

(Boussau et al., 2013) But are computationally expensive.
As we densely sample genomes and taxa, the size of a ‘locus’ or ‘un-recombined region’ will get smaller.
Jointly estimate recombination and ancestries along genomes

Currently only applied within humans! (Kelleher et al., 2018)
Species tree methods are robust to intra-locus recombination (based on analyses of simulated data)

Robinson Foulds (RF) distance: the symmetric difference between trees - the number of branches in tree 1 and not in tree 2 + the number of branches in tree 2 and not in tree 1.
Do you need a whole genome to answer your questions?

For phylogenetic and population genetic questions, not necessarily! Most phylogenetic methods cannot directly handle whole genome data, but from whole genome sequencing you can get homologous loci, as well as a bunch of other stu!
Do you need a whole genome to answer your questions?

**For phylogenetic and population genetic questions, not necessarily!**

Most phylogenetic methods cannot directly handle whole genome data, but from whole genome sequencing you can get homologous loci, as well as a bunch of other stuff!
How do the choices we make in

to

to
How do the choices we make in to affect to ?
Ascertainment bias
A bias in parameter estimation or testing caused by non-random sampling of the data. (also sometimes overlapping with ‘selection bias’ or ‘acquisition bias’)

Ascertainment bias is ubiquitous!

- Surveying volunteers
- Studying undergraduates
- Sampling across ‘species’
- Discarding rare outliers
Sampling across the tree of life

(Hug et al., 2016)
Ascertainment bias in genomic data
It is important to consider what models of evolution are appropriate for your data types.

(Tagliacollo and Lanfear, 2018)
It is important to consider what models of evolution are appropriate for your data types

Entropy (rate proxy), GC content, Multinomial likelihood

Extreme rate heterogeneity in Ultra Conserved Elements, can be handled with appropriate partitioning (Tagliacollo and Lanfear, 2018)
Analyzing only variable sites (e.g. Single Nucleotide Polymorphism (SNP) analyses)

Instead of ‘Anything Can Happen Now’
Analyzing only variable sites (e.g. Single Nucleotide Polymorphism (SNP) analyses)

Instead of ‘Anything Can Happen Now’
‘Something Definitely Happened Once!’
Analyzing only variable sites (e.g. Single Nucleotide Polymorphism (SNP) analyses)

Instead of ‘Anything Can Happen Now’
‘Something Definitely Happened Once!’

This affects our ability to estimate branch lengths using likelihood
Intuitively, will increase inferred branch lengths
can also affect tree topology
Short Tree
Short Tree

Long Tree
How surprised should we be to see no invariant sites?
Very surprising, unless branches are very long
How surprised should we be to see no invariant sites?
Very surprising, unless branches are very long
but only if we looked for them!
How surprised should we be to see no invariant sites?
Very surprising, unless branches are very long 
but only if we looked for them!

Can correct by applying Lewis (2001) model for analysis of only 
variable sites implemented inference software (Leaché et al., 2015)
“it is possible in many cases to correct the ascertainment bias relatively easily, if reliable information is available regarding the details of the ascertainment scheme.” (Nielsen, 2004)
“it is possible in many cases to correct the ascertainment bias relatively easily, if reliable information is available regarding the details of the ascertainment scheme.” (Nielsen, 2004)

This information is not always available. Bias can be driven by the true, evolutionary history you are attempting to estimate!
Despite the large volume of data in genomic studies, ascertainment bias is still an issue.
Despite **because of** the large volume of data in genomic studies, ascertainment bias is still an issue.
Two case studies:
Phylogenetics of Penstemon
Tracing gonorrhea outbreaks
Phylogenetics of Penstemon using RADseq data

Question: How often have transitions between hummingbird and bee pollination occurred in Penstemon?
Data:

Restriction site-associated DNA sequencing (RADSeq)
83 species, two samples per species
No closely related reference genome
RADseq
Uses restriction enzymes to fragment DNA
Targets sequencing to the same regions across taxa

(figures from floragenex.com)
In the absence of a reference genome, you need to cluster reads. A ’cluster’ is an inference of homology.

Clustered using Stacks (Catchen et al., 2011)
Several factors can cause drop-out of alleles in RAD-seq data (i.e. not observing homologous alleles)
- Mutations at restriction digest sites
- Clustering parameters exclude homologous regions
- Low coverage
There have been many conflicting studies on the importance of missing data in phylogenetic analyses, broadly, as long as missing data is random, it shouldn’t be very problematic, but phylogenetically-biased missing data is likely to be. (Roure et al., 2013; Lemmon et al., 2009)
Missing data in RADseq can mislead inference

Figure 4: Properties of simulated RAD loci with different amounts of missing data. Loci that contain more missing data tend to result in discordant topologies (a), increased branch length errors (b), and lower bootstrap support (c). Loci that contain less missing data provide higher bootstrap support for shorter branches (d).

(Leaché et al., 2015)
But excluding sites with high levels of missing data doesn’t solve the problem.

(Huang and Knowles, 2014)
But excluding sites with high levels of missing data doesn’t solve the problem. It biases rate estimation downwards by preferentially removing high rate loci (Huang and Knowles, 2014)

Gray shading is simulated rates, dashed line is shift due to loss of RAD sites, black line is shift due to loss of cut sites, black line shift due to loss of cut sites + post sequencing processing.

(Huang and Knowles, 2014)
Advice?
Advice?

“Given that the data matrix reflects complex interactions between aspects of library construction and processing with the divergence history itself, our results also suggest that general rules-of-thumb are unlikely.”

(Huang and Knowles, 2014)
Advice?

“Given that the data matrix reflects complex interactions between aspects of library construction and processing with the divergence history itself, our results also suggest that general rules-of-thumb are unlikely.”

(Huang and Knowles, 2014)
Tradeoffs:

Decreasing similarity cutoff captures more loci shared across the tree, at risk of incorrect homology.

Decreasing taxon representation threshold allows you to capture more loci, but representing fewer individuals.
Investigate a range of parameters

(Wessinger et al., 2016)
Missing data is phylogenetically biased
Across full dataset, many loci are only found in one of the major clades (Wessinger et al., 2016).
Variation within clades is better captured by dividing the data set and clustering separately.
Build (and report!) multiple trees using different filtering parameters

Trees from separate clade analyses (Wessinger et al., 2016)
Summary:

Bias:
Clustering parameters drive non-random missing data

Potential effect on inference:
No topological resolution
Tip branch lengths are shortened
Non-homologous regions align

Mitigation:
Estimate relationships under a range of filtering parameters

Conclusions:
Branch lengths and bootstrap support differ across filtering parameters
Different data sets may be appropriate at different phylogenetic scales
Evolutionary inferences about pollinator shifts need to be robust to this uncertainty
Case study - tracing gonorrhea outbreaks
Rapid phylogenetic updating to trace gonorrhea outbreaks

Collaboration with
Jack Cartee, Dr. Jeanine Abrams-McLean, and Jasper Toscani Field (PhD student, UC Merced)
**Neisseria gonorrhoeae**

- Gram-negative, diplococci bacteria
- Responsible for the sexually transmitted infection known as gonorrhea
- One of two pathogenic *Neisseria* species known to infect humans
- WHO estimated 78 million new cases among adults worldwide in 2012
Recent increase in rates of gonorrhea infections
*Neisseria gonorrhoeae* has progressively developed resistance to each single dose antibiotic.

![Diagram showing the year of introduction and the year resistance was reported for various antibiotics used to treat gonorrhea.](image-url)
*Neisseria gonorrhoeae* has progressively developed resistance to each single dose antibiotic.

Only remaining recommended treatment option is dual therapy with a ceftriaxone plus azithromycin.
“It is widely recognised that few antimicrobials remain effective in the treatment of *Neisseria gonorrhoeae* infection and that gonorrhoea could become untreatable in the future.” (Chisholm et al. Sex Transm Infect 2015)
To track and control outbreaks, the CDC is tracing evolutionary history of gonorrhea, across the US and globally.
Approach:

Whole genomic sequencing of *Neisseria gonorrhoea* isolates - up to thousands of lineages

Phylogenetic inference to track geographic spread and horizontal gene transfer of resistance genes
Combining geographic and evolutionary information can trace transmission, and transfer of resistance alleles across lineages

Maximum-likelihood phylogeny of Neisseria gonorrhoeae samples (N = 62) collected in Hawaii during February–May 2016.

- Resistance to azithromycin
- Resistance to azithromycin and reduced susceptibility to ceftriaxone
- Isolate sampled in the UK

Challenges:

Thousands of samples; new isolates sequenced every day
Speed from sampling → phylogeny important
Need to rely on phylogenies for public health action (requires high confidence)
Often very little nucleotide variability, but horizontal gene transfer is common.
Potential issues:

- Sequencing error
- Effect of choice of reference genome
Sequencing error
Potentially problematic when real variable sites are rare

Sequencing errors are likely to be singletons

Will overestimate tip branch lengths

Kuhner and McGill (2014) developed a correction for sequencing error in maximum likelihood phylogenetic inference.

Uses a constant expected error per site
Currently, coverage and error information from sequence reads are discarded following # to #. We have information on confidence in individual base calls, but don’t use it.
Could use a “genotype likelihood”, capturing coverage and read quality (Nielsen et al., 2011)

Not currently implemented in phylogenetic likelihood models
At high coverage, effect of sequencing error is likely low, but newer long read data have higher error rates!
Effect of reference choice
Reference based mapping of short reads can speed up generating a consensus sequence.
BUT: Reference choice can affect evolutionary inference
BUT: Reference choice can affect evolutionary inference

In humans, in highly polymorphic regions variant calling is biased toward the reference base (Brandt et al., 2015)
BUT: Reference choice can affect evolutionary inference

In humans, in highly polymorphic regions variant calling is biased toward the reference base (Brandt et al., 2015)

In fragmented DNA samples from beetles, branch lengths change based on reference choices

(Kanda et al., 2015)
A reference mapping based approach will discard information about structural variants not found in the reference.
Reference choice can affect topology

Mapping sequencing reads to reference genomes requires similarity cutoffs that generate biased missing data (Bertels et al., 2014)
Problem: The true (unknown) phylogenetic history will affect how reads map across the genome.
How can we efficiently estimate phylogenies from genomic data, without being biased by the evolutionary history itself?
How can we efficiently estimate phylogenies from genomic data, without being biased by the evolutionary history itself?

*Take advantage of prior inferences.*
Mapping search for best reference taxon

- Tax 1
- Tax 2
- Tax 3

Read alignment: Bowtie2 ~ 4 minutes
Mapping search for best reference taxon

Read alignment: Bowtie2 ~ 4 minutes

Core genome processing: Seqtk, Samtools, Bcftools, Fastx toolkit ~ 4 minutes
Read mapping to best match

Tax 1

Tax 2

Tax 3

Read alignment: Bowtie2 ~ 50 seconds

New extended alignment

Tax 1
Tax 2
...
Tax 5
Ref

Core genome processing: Snek, Samtools, Bcftools, Fastx toolkit ~ 4 minutes
Read mapping to best match

Tax 1

Tax 2

Tax 3

Read alignment: Bowtie2 ~ 50 seconds

New extended alignment

Tax 1
Tax 2
...
Tax 5
Ref

Core genome processing: Segk, Samtools, Bcf tools, Fastx toolkit ~ 4 minutes

ML analysis from starting tree:
RAxML2 ~ 40 seconds
Phylogenetically informed phylogenomic updating approach:

- Assembles only homologous regions of interest
- Uses phylogenetic structure to select multiple references to generate consensus sequence
- Tree search speed up due to starting tree

[GitHub link] github.com/mctavishlab/phycorder
Tree from traditional method

Updated tree

Unweighted RF: 92
Results:
Ok... 😕 the tree is different! but is it better or worse?
Testing the approach using simulations:
**TreeToReads**
Takes into account:
- Phylogeny and model of evolution
- Insertions and deletions
- Distribution of mutations across the genome
- Read coverage
- Sequencing error profiles (observed or estimated)
Generates short read data with which to test assembly, alignment and inference pipelines.
Input genome for simulation is a tip on simulated tree
Can test alignment to other empirically observed genomes
(McTavish et al., 2017)

github.com/snacktavish/treetoreads

Other new approaches for generating reads from phylogenies:
NGSphy (Escalona et al., 2018), Jackalope (R package) (Nell, 2019)
Take observed outbreak tree

Simulate reads using empirical parameters
Infer trees from reads using two different reference genomes. Reference within outbreak reference Distant (1% sequence divergence)
Simulation summary

- In this example, even distant reference genome did not affect parameter of interest (monophyly of outbreak), although it did affect branch lengths
- Effects of read mapping parameters and reference genome choice are likely to be idiosyncratic
- By using empirical estimates for evolutionary model, can investigate effects on parameters of interest
- Currently applying this approach to test gonorrhea phylogenetic updating procedure
Summary

**Bias:** Sequencing error, reference choice

**Effect on inference:**
- Sequencing error can increase terminal branch lengths relative to internal branches
- Not mapping reads on lineages more distant from reference genome will decrease those branch lengths

**Mitigation:** Use multiple reference genomes, simulation based tests to assess accuracy

**Conclusions:**
- When a closely related reference is available, alternatives worsen inference
- At high (around 40x) coverage all mutations are confidently recovered
- Even at lower coverage (around 5x) high confidence in monophyly of outbreak clade
Big picture

All data sets are biased, genome scale data is no exception

Careful project planning helps

Interrogate potential biases in data sets
What to do?

- What data will answer your questions?
- Are there existing data you want to be able integrate with?
- Consider in which direction biases are likely to sway results
- Use the most appropriate available model for your data
- Re-sample your data to test if your key conclusions are robust to choices
- Simulation approaches to test if parameters of interest are affected by sampling and ascertainment schemes
Questions?


