Dissertation Meeting Question:

Metrics for Correlating the Evolution of Binary Characters

Participants:

- Alex Dornburg
- Daniel Janies
- David Brown

Overall Dissertation Theme

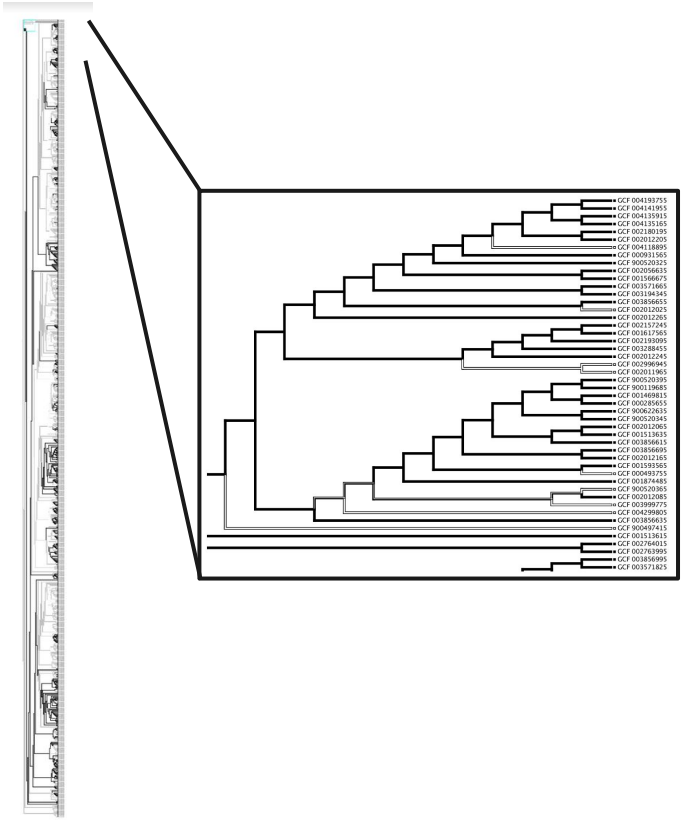
- **Introduction**
  - Bacteria with deficient mismatch repair mechanisms have been described as having a "hypermutable phenotype". This highly adaptive phenotype is suggested to drive the acquisition of antimicrobial resistance. A main component of the methyl-directed mismatch repair system in bacteria is Mutator S.
  - Taken from LeClerc et al. (1996), hypermutable phenotype *Escherichia coli* exist (1.4-6.7% incidence among pathogenic strains) and that all hypermutable phenotype *Escherichia coli* in that study had a Mutator S (MutS) "defect" described as a large deletion that "extends 212 bp into the 3' end of *mutS* [MutS]".
- **Research Questions**
  - What Mutator S variants are present in the data set?
  - Which of those Mutator S variants are "hypermutable", as shown by LeClerc et al (1996)?
  - Are there correlations with hypermutable phenotypes and multidrug-resistance phenotypes?

Specific Chapter 2 Hypothesis and Methods

- **Hypothesis**
  - Certain Mutator S variants (MutS) are correlated with multidrug resistance (MDR) in *Escherichia coli*.
- **Methods**
  - Build a phylogenetic tree on the core genome using "RAxML".
  - Identify MutS variants in the data set using "snp-sites".
  - Overlay characters for MutS variants and MDR on the tree.
  - Apply a metric to test character correlations.

Workflow - Part 1 - Build a Tree

- Start with ~10,000 *Escherichia coli* isolates.
- Implement quality controls to identify high-confidence, -purity, and -quality whole genome sequences, assembled to the "Chromosome" level and stored in RefSeq.
- Establish the **core genome** by annotating the 817 selected isolates.
- Build a tree in RAxML on the **core genome**.



Workflow - Part 2 - Establish Characters

- Represent multidrug-resistance (MDR) as a **binary character**, according to current CDC/NARMS definitions (results in a 816 x 1 matrix).
- Extract the annotated MutS gene sequences (816), translate, align, and call amino acid (AA) variants.
- Represent MutS variant positions as a **binary character** for per position agreement with the consensus AA base (results in a 816 x 141 matrix).
- ***PROBLEM*:** Superimposing characters onto the tree is not satisfying or useful without a metric.

	0	1
MDR (y)	*< 3 classes of resisted compounds*	*>=3 classes of resisted compounds*
Variant Positions (x)	*any other AA residue*	*matches consensus residue*

Table of definitions for chosen characters (rows) and their associated potential states (columns) .

Workflow - Part 3 - Implement Pagel (1994)

- Use Pagel (1994) for correlated evolution of binary characters.
- Borrow an R implementation of Pagel (1994) written by Liam Revell in the R package "phytools".
- The function "fitPagel()" tests binary character dependence (x-vs-y, y-vs-x, etc.) against a null hypothesis of independence.
- Apply "fitPagel()" to test 4 models per position:
  - dependent x,
  - dependent y,
  - both dependent, or
  - both independent.
- Collect **p-values** and **likelihood ratios** to evaluate support for each model's hypothesis test.
- Refer to the **Akaike information criterion (AIC)** to select between the four models.
- Link to inputs.
- Link to results.

My Questions - Your Feedback

- Is Pagel (1994) still an acceptable approach for correlating binary characters?
- Does Pagel (1994) work for this chapter of my dissertation?
- Are there known issues with Liam Revell's implementation of Pagel (1994) as "fitPagel()"?
- In light of the above, should I consider **changing my workflow** to another metric/approach?
- In light of the above, are there other metrics/approaches I should **mention in my writing**?

- Do I need to apply a p-value correction technique like Bonferroni or Benjamin-Hochberg?
- Refer to table on Slide 5. Should I change these so that 1 is consistent with the "majority" case?
- Describe question of pruning, distances, and "fitPagel()".

WIP Results

- This line of research is an extension of our genetic capitalism work. If the rich are getting richer, is it because they mutate more?
- No hard evidence of any of LeClerc's defect MutS. One potential sequence is deleted, but to a much larger degree than LeClerc described.
- Footnote 9 in LeClerc et al. (1996) notes that mutator vs nonmutator is an arbitrary distinction. They were considering >50-fold mutation rates, up to 1000-fold increases in mutation rate.
    - More recent research has noted that mutation rates in *E. coli* are malleable in response to stress. It is my opinion that the extreme hypermutable phenotypes described by LeClerc are exceptions, especially in the context of this more recent high-confidence, high-quality data generated from improved techniques over the past 25-30 years.
- Only a single variant position (7) supports a correlation between that AA location and a MDR phenotype.
    - Presence of phenylalanine (F) instead of the alternate, minority leucine (L) residue.
- Position is found in the mismatch binding domain.
    - Hard to crystallize without the DNA ligand.
    - Unlikely that this residue interacts directly with the ligand, but uncertain to know for sure.
- Prokaryotic MutS is a homodimer (from a single copy gene), so any mutation is expressed twice. Little to no risk of an asymmetrical dimer. (One side will not have F at 7 and the other an L.)
- Every other identified variant position fails to reject the null hypothesis of independence.
- My hypothesis: Certain Mutator S variants are correlated with multidrug resistance (MDR) in Escherichia coli.
    - Falsified in every case except one.
    - Continues research questions into next chapter: if hypermutation is not correlated with MDR, then what is?

Thank you for your time!

Workflow - Continuing the Research

Continuing correlation analysis through use of guided machine learning as a statistical technique