Lies, damn lies, and .... genomics
you, your data, your perceptions and reality
Christopher West Wheat

Goal of this lecture
• Present a critical view of ecological genomics
• Make you uncomfortable by sharing my nightmares
• Encourage you to critically assess findings and your expectations in light of publication biases
Disclaimer

I’m a positive person

I love my job and the work we all do

I’m just sharing scrumptious food for thought

What if ..... 50% of your favorite studies had conclusions that were just wrong?

How would that affect your expectations and work?
If the biomedical science has the most money and oversight, then ....

Their findings should be robust:

- Repeatable effect sizes
- The same across different labs
- The same across years

Publication replication failures

- Biomedical studies
  - Of 49 most cited clinical studies, 45 showed intervention was effective
  - Most were randomized control studies (robust design)

- Mouse cocaine effect study, replicated in three cities
  - Highly standardized study

Ioannidis 2005 JAMA; Lehrer 2010
Assessing reality using funnel plots

Small sample sizes affect measurement accuracy
Each dot = a study and has error
Study estimates are randomly distributed about the real value
Your study is just a random estimate of some idealized value

Publication bias increases effect size

If all studies on same question were published
Reality, low effect sizes not published

What if there is no replication?

What is most likely to publish first & where?

What publishes late?

Why Most Published Research Findings Are False

A research finding is less likely to be true when:

✓ the studies conducted in a field have a small sample size
✓ when effect sizes are small
✓ when there are many tested relationships using tests without *a priori* selection
✓ where there is greater flexibility in designs, definitions, outcomes, and analytical modes
✓ when there is greater financial and other interest and prejudice
✓ when more teams are involved in a scientific field, all chasing after statistical significance by using different tests

Ioannidis 2005 Plos Med.
But surely, this doesn’t apply to genomics….

Or does it?

8 topics first reported with $P < 0.05$

There are lies, damn lies, and ....

But wait, is that fair?

Are these really lies?

Where does this bias come from?

- Population heterogeneity
  - Space and time
- Publication bias
  - Large & significant effects publish fast and with high impact
  - Small & non-significant effects publish slow with low impact
Where does this bias come from?

And me .... All of us

YOU!!

It arises from humans doing science

The way we think

The way our institutions work

Apophenia

A universal human tendency to seek patterns in random information and view this as important

Story telling of Type 1 errors

Celebration of the false positives
Outline

• Are there biases understanding the genomic architecture of adaptations?

• What is the power of molecular tests of selection?

• What does the dissection of some classic comparative genomics study reveal?

Metabolic Pathways

How do we find the genes that matter?

Publications using molecular tests demonstrate we can sequence our way to answers

Current paradigm:
Sequence, map, find sig. patterns, make causal story, move on

......
What is the architecture of a causal variant?

What type of variant?
- SNP, indel, TE, inversion, CNV?

How predictable are adaptations?

<table>
<thead>
<tr>
<th></th>
<th>Plants</th>
<th>Animals</th>
</tr>
</thead>
<tbody>
<tr>
<td>Coding</td>
<td>71</td>
<td>163</td>
</tr>
<tr>
<td>Cis-regulatory</td>
<td>26</td>
<td>48</td>
</tr>
<tr>
<td>Other</td>
<td>16</td>
<td>7</td>
</tr>
<tr>
<td>Total</td>
<td>113</td>
<td>218</td>
</tr>
<tr>
<td>Null</td>
<td>67</td>
<td>32</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>Morphology</th>
<th>Physiology</th>
<th>Behavior</th>
</tr>
</thead>
<tbody>
<tr>
<td>Coding</td>
<td>62</td>
<td>170</td>
<td>2</td>
</tr>
<tr>
<td>Cis-regulatory</td>
<td>43</td>
<td>29</td>
<td>2</td>
</tr>
<tr>
<td>Other</td>
<td>3</td>
<td>20</td>
<td>0</td>
</tr>
<tr>
<td>Total</td>
<td>108</td>
<td>219</td>
<td>4</td>
</tr>
<tr>
<td>Null</td>
<td>41</td>
<td>58</td>
<td>0</td>
</tr>
</tbody>
</table>

Stern & Orgogozo 2008 Evolution
Individual genome sequencing: powerful insights

2-5 X per individual, sliding 2500 bp window, 500 bp step

Jones et al. 2012 Nature

Which regions are more important? Coding or expression?

Jones et al. 2012 Nature
How do we identify the genes that matter?

• Molecular tests of selection are popular, but ...
  — What are their assumptions and power?

• What are these tests detecting?
  — What is a footprint of selection?
    • How are they formed?
    • How long do the last?
    • How large are they?

Finding the genes: a decision tree

Most publications each use many such tests, but report only a subset and argue findings are robust.

Hohenlohe et al. 2010 Int. J. Plant Science
What power do we have to detect evolution by natural selection?

Power is the probability that the test will reject the null hypothesis when the alternative hypothesis is TRUE.

Using a t-test, you would want power > 90% at reasonable sample size, right?

Breed specific morphologies

Test set of Schlamp et al. 2016:
- 25 breeds
- 12 causal loci identified by QTLs
- $N = 25$ / breed
- 7 tests of selection
  - $iHS, nSL, H, TajD$, etc.

How accurate are molecular tests of selection detect?
French Bulldog sample: low power (high type II error)

Why don’t these tests have much power?

Biological reality vs. theoretical population genetics?
Directional selection: an example of the expectations of hard selection

Population genomics has been dominated by developing methods to detect hard sweeps for past two decades

— But a proper ‘null model’ continues to be elusive, resulting in a high false positive rate since their inception

Storz 2005 Mol. Ecology

\[
\begin{align*}
\text{ATGTAAGTCATATGGGATCAAGGTTGAAATGCTAGAGCGTA} \\
\text{ATGTAAGTCATATGGGATCAAGGTTGAAATGCTAGAGCGTA} \\
\text{ATGTAAGTCATATGGGATCAAGGTTGAAATGCTAGAGCGTA} \\
\text{ATGTAAGTCATATGGGATCAAGGTTGAAATGCTAGAGCGTA} \\
\text{ATGTAAGTCATATGGGATCAAGGTTGAAATGCTAGAGCGTA}
\end{align*}
\]

Estimate of error rates using Tajima’s D, and haplotype homozygosity under the models for a human population

A) False Discovery

B) False Negative

Teshima et al. 2006 Genome Research
Simulation conclusions

• Simulations suggest
  – empirical approaches will identify several interesting candidates
  – But will also miss many—in some cases, most—loci of interest

• Power is lower when
  – directional selection involves a recessive rather than a co-dominant allele
  – when it acts on a previously neutral rather than a new allele
  – Demographic changes rather than constant population size

  Genomic scans yield an unrepresentative subset of loci that contribute to adaptations

Molecular tests ...

  BASED ON 20 YEARS OF PUBLICATIONS

• Are still chasing an elusive null model ....
  – Each performs better than previous ones under a specific set of conditions, all have poor null model

• But ... under realistic biological conditions, they all
  – Have very low power (high type II error rates)
  – Have high false positive rates
How common are hard sweeps in nature?

• “we argue that soft sweeps might be the dominant mode of adaptation in many species”  
  Messer and Petrov 2013 TREE

Certainly not everyone agrees ....

• This is an important read, critical of
  – assumptions underlying soft sweep (selection on standing variation)
  – the low power of molecular tests to detect hard & soft sweeps
How likely does natural selection use standing variation in your species?

Thought experiment:
- What fraction of species respond to selection in the lab?
- Why?
- If populations have variation, how likely is selection to use it?
- What’s likelihood of selection on standing variation in wild?

We have not been studying the dominant form of selection in the wild & cannot reliably detect it

Age and type of selection matters

- Novel mutation, large effect, hard sweep that goes to fixation
  - Probability of detection 20 - 90%, depending on demography, etc.

- Old mutation and/or polygenetic that does not sweep to fixation
  - Probability of detection close to 0

- Finding the causal mechanism
  - Coding > expression (but allele specific expression can be lightening rod for expression)
  - SNPs > more complex mutations (indel, TE, CNV)
  - Ongoing gene flow & grouping by phenotype across replicate populations helps a lot

- Demographic effects
  - Nearly all species have experienced a major demographic change in the past 10,000 generations
  - Demographic change significantly reduces power and increases false positive rates.

- What is the relative frequency of these?
  - What will be the architecture of your phenotype?
  - What does your method have the highest power to detect?
Get ready, here come the 1000<sup>th</sup> genomes

- Roughly 20 arthropods sequenced to date
  - plans to sequence 5,000 more
- Many other large scale projects coming online

An unprecedented opportunity for large scale errors?

Classic study: Evolution of genes and genomes on the *Drosophila* phylogeny

Classic study: Evolution of genes and genomes on the *Drosophila* phylogeny

Drosophila 12 Genomes Consortium 2007 Nature
Tempo and mode of chromosome evolution

- > 20 My, chromosomal order completely reshuffled in Diptera
  Drosophila 12 Genomes Consortium 2007 Nature

Genome evolution
Drosophila 12 Genomes Consortium 2007 Nature
Selection dynamics across functional categories

- 33.1% of single-copy orthologues have experienced positive selection on at least a subset of codons.

Drosophila 12 Genomes Consortium 2007 Nature

Gene Family Evolution across 12 Drosophila Genomes

- One fixed gene gain/loss across the genome every 60,000 yr
- 17 genes are estimated to be duplicated and fixed in a genome every million years

Drosophila 12 Genomes Consortium 2007 Nature
Hahn et al. 2007 Plos Genetics
Comparative Genomics: a house of cards?

• Data scale is too large to thoroughly assess errors ...  
  – Perhaps the findings are just ... wrong

• All conclusions, at some stage, rest upon  
  – Simple bioinformatics  
  – Assumptions that get incorporated into seemingly unbiased methods

Let’s exploring two pillars of these studies, their error and repercussions  
  – Gene alignments in detecting positive selection  
  – Calibrations in temporal analysis

Published studies allow ...

Follow up studies to reveal limitations  
But, must have enough details to be repeatable
Genome-wide selection dynamics:

How robust are these conclusions?

Codon based tests of selection

\[ d_N / d_s \]

- > 1 positive sel.
- = 1 neutral
- < 1 purifying sel.

IMPRS workshop,
Comparative Genomics
Evolution of genes and genomes on the *Drosophila* phylogeny

Drosophila 12 Genomes Consortium 2007 Nature

**dN/dS estimates by aligner**

- 6690 orthologs
- 5 alignment methods
- Alignment methods affect dN/dS estimates

Markova-Raina & Petrov 2011 Genome Biology
Comparing results across methods is responsible bioinformatics!!!!!

Since we can’t look at our data, we need approaches that allow 1st principal assessments

Aligner has a larger effect than biological signal

<table>
<thead>
<tr>
<th>Aligner</th>
<th>12 genomes, M7/8</th>
<th>12 genomes, M1a/2a</th>
<th>12 genomes, M7/8, with removed gaps</th>
<th>Melanogaster group, M7/8</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>95% (a) 99% (b)</td>
<td>95% (c) 99% (d)</td>
<td>95% (e) 99% (f)</td>
<td>95% (g) 99% (h)</td>
</tr>
<tr>
<td>AMAP</td>
<td>817  213</td>
<td>256  110</td>
<td>558  104</td>
<td>973  257</td>
</tr>
<tr>
<td>MUSCLE</td>
<td>1043 306</td>
<td>379  192</td>
<td>764  155</td>
<td>1134 366</td>
</tr>
<tr>
<td>ProbCons</td>
<td>1013 281</td>
<td>346  180</td>
<td>801  182</td>
<td>1128 371</td>
</tr>
<tr>
<td>T-Coffee</td>
<td>1290 479</td>
<td>612  353</td>
<td>824  173</td>
<td>1248 (909) 463 (218)</td>
</tr>
<tr>
<td>ClustalW</td>
<td>902  261</td>
<td>244  117</td>
<td>666  112</td>
<td>1269 453</td>
</tr>
<tr>
<td>Total in 5</td>
<td>1902 673</td>
<td>799  441</td>
<td>1562 384</td>
<td>1737 (1723) 652 (620)</td>
</tr>
<tr>
<td>PRANK</td>
<td>468  49</td>
<td>49   16</td>
<td>238  42</td>
<td>581  70</td>
</tr>
</tbody>
</table>

Markova-Raina & Petrov 2011 Genome Biology
Alignment results highlight importance of alignment score!

- Tcoffee finds 3 selected sites indicated by arrows
- ProbCons identifies region with low alignment score, not used

What about recent genomes?

Surely they are better?

and mammals ... they have good genomes

and alignment problems rarely happen

... right?
What about recent genomes on cute mammals?

- How did I evolve to be so cute?


<table>
<thead>
<tr>
<th>Positive Selected Genes</th>
<th>Revised PSG</th>
</tr>
</thead>
</table>

Deficient in:
- Alignment
- Coverage
- Annotation

Data:
~3000 orthologs

23.3%

10.2%

17.5%

2.0%

4.2%

3.8%

2.8%

0.7%

0.9%

2.8%

11.6%
Post-genomics challenge

“What we can measure is by definition uninteresting and what we are interested in is by definition unmeasurable.”
- Lewontin 1974

“What we understand of the genome is by definition uninteresting and what we are interested in is by definition very damn difficult to sequence and assemble and annotate and analyze at genomic scale”
- Wheat 2015

For example:
- indels & inversions
- gene family dynamics
- evolutionary dynamics

What does a good P-value really tell you?

What does a bad P-value really tell you?

When did selection happen?

Are you chasing a good P-value?

What type of selection?

Is method mismatched to mechanism?
Significant P-values

Robust understanding requires validation:
- Genetic manipulation
- Field study manipulations

Goal of this lecture

- Present a non-typical view of ecological genomics
  - So you have a more complete view of the field
- Make you uncomfortable
  - Provide a context for understanding your results
- Encourage you to rethink the reality presented by publication biases
  - Overcoming this bias is a continual challenge
Temporal inference: fact or fiction?
Evolution of genes and genomes on the *Drosophila* phylogeny

Determining objective priors is challenging

Priors in Bayesian rel. clock analysis:

- Mu = lab observed mutation rate
- A1,2 = geological calibration, small Ne
- C1,2 = geological calibration, large Ne

Priors directly influence posteriors

Prior distributions matter

- Integrative science is challenging
- Discuss or collaborate with experts to evaluate your approach.

How do we gain dating confidence when we are in the dark?

- Fossils and DNA are likely to rarely agree
- How can we assess the temporal signal in the DNA in a robust manner?
  - Reducing prior biases and using lots of DNA data, while modeling likely violations of analysis models

Wheat and Wahlberg 2013 Trends Ecology & Evolution
• 1000’s of false positive SNPs (FP SNPs) result from misassembly x mapping x calling
• Genome was small (~125 Mbp) with few repeats (Arabidopsis thaliana)
• FP rates likely much higher with larger, more complex genomes

Ribeiro et al. 2015 BMC Bioinformatics
### Unlikely results

How a small proportion of false positives can prove very misleading

<table>
<thead>
<tr>
<th>False</th>
<th>True</th>
<th>False negatives</th>
<th>False positives</th>
</tr>
</thead>
</table>

1. Of hypotheses interesting enough to test, perhaps one in ten will be true. So imagine tests on 1,000 hypotheses, 100 of which are true.

2. The tests have a false positive rate of 5%. That means they produce 45 false positives (5% of 900). They have a power of 0.8, so they confirm only 80 of the true hypotheses, producing 20 false negatives.

3. Not knowing what is false and what is not, the researcher sees 125 hypotheses as true, 45 of which are not. The negative results are much more reliable—but unlikely to be published.

---

... and now for pt. 2
The Fate of Mutations Surfing on the Wave of a Range Expansion

Now handling genomic data

http://www.jnr-eeb.org/index.php/jnr
Microevolution effects

Previous examples were at deep evolutionary time scales

Surely such problems don’t exist at the within genera level ..... Right?
Recombination violates dN/dS tests

- 13% of sites simulated at omega = 2.5
- Sample size = 30 sequences

False positives can increase to over 30%

Codeml inferred selection:

Anisimova 2003 Genetics

Posterior distribution estimates of substitution rates from mitochondrial control region from Beringian bison

Ho et al. 2007 Systematic Biology
Significant implications for phylogeographic studies that use fixed rates to assess demographic with environmental change.

What power do we have to detect balancing selection?

- For *Drosophila melanogaster*, power = 50% with window size of 200 bp, using 24 diploid individuals.
- For species with larger population size, power likely lower.
- Recombination and gene conversion destroy ‘footprint’ rather quickly.
Fst outlier analyses are common

Pervasive selection or is it…? why are $F_{ST}$ outliers sometimes so frequent?
Bierne et al. Molecular Ecology 2013

8 topics where first study $P > 0.05$, but became significant after meta-analyses

What is our power to detect hard sweeps?

When did selection act on your phenotype?
Zhai, Nielsen & Slatkin 2008 MBE

Hard vs. soft or incomplete sweeps in populations
Fst outlier analysis

9900 Neutral, 100 selected sites
N=1500 (20 ind. per 75 populations)

What is our power to detect hard sweeps within a population?

When did selection act on your phenotype?
What’s the demographic history of your population?

Zhai, Nielsen & Slatkin 2008 MBE

What’s a good way to assess molecular tests?

• Computer simulations of evolution
  – Across range of demographic scenarios

• What else?

• Testing them on real data where we know the targets of selection = real world validation
  – Which ones work and when
  – We could then use this to make better tests, right? (very rare)

Hard selection case example: threespine stickleback fish
**Threespine stickleback fish** (*Gasterosteus aculeatus*)

- Has body armor in the ocean
- Loses almost all armor in lakes

**Parallel adaptation in fresh water lakes via hard sweeps**

Marine population

<table>
<thead>
<tr>
<th>Marine population</th>
<th>Population 1</th>
<th>Population 2</th>
<th>Population 3</th>
<th>Population 4</th>
</tr>
</thead>
<tbody>
<tr>
<td>Variation within populations</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Variation between populations</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Invaded fresh water lake

Natural selection

Predominant form
Non – adaptive

Adaptive

disease, aging, height, etc.
salinity, color, resistance, etc.

\[ \text{generally ...} \]

1000’s of loci, each of small effect size

One or several loci of large effect

\text{Is this a publication bias?}

Will your trait have 1000’s of small effect genes, or a few genes of large effect?

Sear (2010) … Is bigger always better? Rockman (2011) … All that’s gold does not glitter

\textbf{Timing of divergence}

- Directly affects rate estimates
- Deriving unbiased dates from molecular data
  - Large field of software development
- Bayesian methods, while potentially informative and unbiased
  - Can be easily, and are routinely, abused

\textit{Wheat and Wahlberg 2013 TREE}
Calibration: Kauai age of 5.1 my for divergence of two Hawaiian species

1. No phylogeny
2. Fixed clock rate
3. Between 3 – 64 genes in pairwise comparisons

Temporal patterns in fruitflies (Tamura et al. 2004 MBE)

Drosophila clade:
– Schizophora constrained to maximum of 70 Ma
– Without constraint, goes to 115 Ma

What is reality?

Episodic radiations in the fly tree of life (Wiegmann et al. 2011 PNAS)
Thus, the age of this clade is fiction