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

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
  - 
- Publication bias
  - impact
  - Small & non-significant effects publish slow with low impact
Where does this bias come from?

It arises from humans doing science
The way we think
The way our institutions work

And me .... All of us

YOU!!

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?

A

B

C Coding mutations: affect the mature RNA or protein

D Cis-regulatory mutations: affect gene expression

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

How predictable are adaptations?

<table>
<thead>
<tr>
<th>Types</th>
<th>Plants</th>
<th>Animals</th>
</tr>
</thead>
<tbody>
<tr>
<td>Coding¹</td>
<td>71</td>
<td>163</td>
</tr>
<tr>
<td>Cis-regulatory</td>
<td>26</td>
<td>48</td>
</tr>
<tr>
<td>Other²</td>
<td>16</td>
<td>7</td>
</tr>
<tr>
<td>Total</td>
<td>113</td>
<td>218</td>
</tr>
<tr>
<td>Null³</td>
<td>67</td>
<td>32</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Types</th>
<th>Morphology</th>
<th>Physiology</th>
<th>Behavior</th>
</tr>
</thead>
<tbody>
<tr>
<td>Coding²</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>Other⁴</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>Null⁵</td>
<td>41</td>
<td>58</td>
<td>0</td>
</tr>
</tbody>
</table>

Stern & Orgogozo 2008 Evolution
Individual genome sequencing: powerful insights

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 large are they?
    - How long do the last?

---

Finding the genes: a decision tree

Most publications each use many such tests, but report only a subset and argue findings are robust.
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
- $N = 25 / \text{breed}$
- 7 tests of selection
  - $iHS, nSL, H, \text{TajD}$, 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

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

• False-discovery rate is higher 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 size 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

The lab?

- “Signatures of selection ... [are] not associated with ‘classic’ sweeps ... More parsimonious explanations include [selection on standing variation]”
  
  Burke et al. 2010 Nature

How common were hard sweeps in our history?

- “classic sweeps were not a dominant mode of human adaptation over the past 250,000 years”
  
- “much local adaptation has occurred by selection acting on existing variation rather than new mutation”

  1000 Genomes PC 2010 Science
  Hernandez et al. 2011 Science

Certainly not everyone agrees ....

On the unfounded enthusiasm for soft selective sweeps

Jeffrey D. Jensen

- 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\textsuperscript{n} 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?

- Phylogenetic relationships
- Genome evolution
- Functional insights into genes and genomic features (e.g. regulation and inheritance)

Classic study: Evolution of genes and genomes on the \textit{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

| Species            | Total no. of protein-coding genes (per cent with D. melanogaster homologs) | Coding sequence Intron (M)
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>D. melanogaster</td>
<td>13,733 (100%)</td>
<td>38.9/21.8</td>
</tr>
<tr>
<td>D. simulans</td>
<td>11,963 (86.2%)</td>
<td>45.8/23.6</td>
</tr>
<tr>
<td>D. yakuba</td>
<td>16,888 (81.2%)</td>
<td>47.9/21.9</td>
</tr>
<tr>
<td>D. yeksha</td>
<td>16,423 (82.5%)</td>
<td>50.8/22.9</td>
</tr>
<tr>
<td>D. erecta</td>
<td>15,324 (86.4%)</td>
<td>49.1/22.0</td>
</tr>
<tr>
<td>D. ananassae</td>
<td>15,326 (83.0%)</td>
<td>57.3/22.3</td>
</tr>
<tr>
<td>D. pseudoobscura</td>
<td>16,363 (78.2%)</td>
<td>49.7/24.0</td>
</tr>
<tr>
<td>D. persimilis</td>
<td>17,321 (72.6%)</td>
<td>54.0/21.9</td>
</tr>
<tr>
<td>D. willistoni</td>
<td>15,835 (78.8%)</td>
<td>61.4/23.5</td>
</tr>
<tr>
<td>D. virilis</td>
<td>14,680 (62.7%)</td>
<td>57.6/21.7</td>
</tr>
<tr>
<td>D. mojavesensis</td>
<td>14,809 (69.8%)</td>
<td>57.8/21.9</td>
</tr>
<tr>
<td>D. grimshawi</td>
<td>15,270 (61.3%)</td>
<td>54.9/22.5</td>
</tr>
</tbody>
</table>

- Single-copy orthologues
- Conserved homologues
- Patchy homologues (with mel.)
- Patchy homologues (no mel.)
- Lineage specific

Number of gene models

0  5,000  10,000  15,000  20,000  25,000
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

\[ \frac{d_N}{d_S} \]

- > 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 tool has a larger effect than biology**

<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 386</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>258 42</td>
<td>581 70</td>
</tr>
</tbody>
</table>

Number of significant genes in common across 1, 2, 3, 4, or all 5 of the alignment methods

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?


Deficient in:
- Alignment
- Coverage
- Annotation

Data≈3000 orthologs
Positive Selected Genes
Revised PSG

Temporal inference:

fact or fiction?

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

Wheat and Wahlberg 2013 TREE
Evolution of genes and genomes on the *Drosophila* phylogeny

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)
Episodic radiations in the fly tree of life (Wiegmann et al. 2011 PNAS)

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

What is reality?

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

Thus, the age of this clade is fiction

Drosophila 12 Genomes Consortium 2007 Nature
Post-genomics challenge

“What we can measure is by definition uninteresting and what we are interested in is by definition unmeasureable”

- 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?

What type of selection?

Are you chasing a good P-value?

Is method mismatched to mechanism?

What does a bad P-value really tell you?
**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