Lies, damn lies, and .... genomics

you, your data, your perceptions and reality

Christopher West Wheat

Career trajectory

• 1995 – 2001 PhD California
• 2002 – 2005 Postdoc Germany
• 2005 – 2008 Postdoc Finland
• 2009 – unemployed 4 month, spent all savings
  — > 50 job applications, 1 grant application
• 2009 – visiting scientist Germany
  — 1 job offer UK
  — 1 grant Finland
• 2012 – started tenure track Sweden

What was important?
• Being able to move
• Chasing the money & skills
• Learning how to:
  — Write publications, grants
  — Believe in my ideas/skills

Needed to put science first, while having lots of fun along the way
Ecological & Evolutionary Functional Genomics

Alternative life history switches  Diapause physiology and switches  CRISPR-Cas9

Butterfly immunity  Plant / insect coevolution  Comparative genomics

CHRISTOPHER WHEAT LAB

https://christopherwheatlab.net/

Now .... Who are you?

<table>
<thead>
<tr>
<th>Topic</th>
<th>Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>Comparative genomics</td>
<td></td>
</tr>
<tr>
<td>Evolutionary genomics</td>
<td></td>
</tr>
<tr>
<td>Metagenomics</td>
<td></td>
</tr>
<tr>
<td>Population genomics</td>
<td></td>
</tr>
<tr>
<td>Single cell genomics</td>
<td></td>
</tr>
<tr>
<td>Structural variation</td>
<td></td>
</tr>
<tr>
<td>Transcription</td>
<td></td>
</tr>
<tr>
<td>Viral [ ]</td>
<td></td>
</tr>
</tbody>
</table>
What do you study?

Rough totals:
- Invertebrates: 7
- Fish: 8
- Mammals: 5
- Microbes: 16
- Plants: 4
- Humans: 4

What are your goals?

- Finding and study genomic regions that matter
- Investigating ecological processes
  - metagenomics
- Investigating physiology
  - RNAseq

Goal of this lecture

- Present a critical view of things genomic
- Make you uncomfortable by sharing my nightmares
- Encourage you to critically assess findings and expectations in light of easy errors and 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?
Adaptive protein evolution at the Adh locus in Drosophila

John H. McDonald & Martin Kreitman

Department of Ecology and Evolutionary Biology, Princeton University
Princeton, New Jersey 08544, USA

Nature 1991

McDonald Kreitman test

Lazzaro 2018 Genetics
Adaptive protein evolution at the Adh locus in Drosophila

John H. McDonald & Martin Kreitman

Department of Ecology and Evolutionary Biology, Princeton University, Princeton, New Jersey 08544, USA

TABLE 2 Number of replacement and synonymous substitutions for fixed differences between species and polymorphisms within species

<table>
<thead>
<tr>
<th></th>
<th>Fixed</th>
<th>Polymorphic</th>
</tr>
</thead>
<tbody>
<tr>
<td>Replacement</td>
<td>7</td>
<td>2</td>
</tr>
<tr>
<td>Synonymous</td>
<td>17</td>
<td>42</td>
</tr>
</tbody>
</table>

A G-test of independence (with the Williams correction for continuity)\(^1\) was used to test the null hypothesis, that the proportion of replacement substitutions is independent of whether the substitutions are fixed or polymorphic. G = 7.43, P = 0.006.

We suggest that these excess replacement substitutions result from adaptive fixation of selectively advantageous mutations.

Adaptive protein evolution at the Adh locus in Drosophila

John H. McDonald & Martin Kreitman

Department of Ecology and Evolutionary Biology, Princeton University, Princeton, New Jersey 08544, USA

TABLE 2 Number of replacement and synonymous substitutions for fixed differences between species and polymorphisms within species

<table>
<thead>
<tr>
<th></th>
<th>Fixed</th>
<th>Polymorphic</th>
</tr>
</thead>
<tbody>
<tr>
<td>Replacement</td>
<td>7</td>
<td>2</td>
</tr>
<tr>
<td>Synonymous</td>
<td>17</td>
<td>42</td>
</tr>
</tbody>
</table>

A G-test of independence (with the Williams correction for continuity)\(^1\) was used to test the null hypothesis, that the proportion of replacement substitutions is independent of whether the substitutions are fixed or polymorphic. G = 7.43, P = 0.006.
But... this was never rigorously tested
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

Sex ratio in birds

P-value = 0.05

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, non-sig are not published

What if there is no replication?

What is most likely to publish first & where?

What publishes late, if at all?

- studies with a low N, 0.5 effect size
- a bias in publishing sig. results (colored)
- a bias against being able to publish null results

Real effect size distribution
Biased effect size distribution

Result: inflation of true effect size

https://bids.berkeley.edu/news/visualizing-publication-bias-case-funnel-plots
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?
Outline

• Are these biases inherent in genomic studies?
• Why is this happening?
• How can we try and overcome these problems?

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 culture
  - 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's arises from humans doing science
The way we think
The way our institutions work

Apophenia
The tendency to seek and see patterns in random information and view this as important

Story telling of Type 1 errors
Celebration of the false positives
Genomics is too big to fail

• Making errors is extremely easy
• Results will very likely be significant, and sometimes dramatically so
• In non-model systems, rarely have replication studies
• You must always question your bioinformatics before falling in love with your results

When results are better than you could have dreamed,

Comparison of the transcriptional landscapes between human and mouse tissues

“the expression for many sets of genes was found to be more similar in different tissues within the same species than between species”

Time of the most recent common ancestor:

Human and Mouse
Snyder mouse controversy

“the expression for many sets of genes was found to be more similar in different tissues within the same species than between species” Lin et al. 2014 PNAS

“[after accounting] for the batch effect, … human and mouse tend to cluster by tissue, not by species” Gilad and Mizrahi-Man 2015. F1000 Research

**Correlation**

**Why? this was a technical artifact called a batch effect. confounded sequencing grouping with biological grouping**

<table>
<thead>
<tr>
<th>D87PMIN1 (run 253, flow cell D2GUACXX, lane 7)</th>
<th>D87PMIN1 (run 253, flow cell D2GUACXX, lane 8)</th>
<th>D4LHBFN1 (run 276, flow cell C2HKJACCX, lane 4)</th>
<th>MONK (run 312, flow cell C2GR3ACXX, lane 6)</th>
<th>HWI-ST373 (run 375, flow cell C3172ACXX, lane 7)</th>
</tr>
</thead>
<tbody>
<tr>
<td>heart</td>
<td>adipose</td>
<td>heart</td>
<td>brain</td>
<td>brain</td>
</tr>
<tr>
<td>kidney</td>
<td>adrenal</td>
<td>renal</td>
<td>kidney</td>
<td>pancreas</td>
</tr>
<tr>
<td>liver</td>
<td>sigmoid colon</td>
<td>sigmoid colon</td>
<td>liver</td>
<td>brain</td>
</tr>
<tr>
<td>small bowel</td>
<td>lung</td>
<td>lung</td>
<td>small bowel</td>
<td>spleen</td>
</tr>
<tr>
<td>spleen</td>
<td>ovary</td>
<td>ovary</td>
<td>testis</td>
<td>Human</td>
</tr>
<tr>
<td>testis</td>
<td>pancreas</td>
<td></td>
<td></td>
<td>Mouse</td>
</tr>
</tbody>
</table>

**Solution** = Keep technical effects orthogonal to biological

- Mouse & Human in same lane, same tissues in same lane
- Will your Core facility know to do this for you?
Personalized medicine: via an excel error

• Searching for gene expression signatures predicting sensitivity to specific cancer drugs, as patients show highly variable response to drug called cisplatin
  — treatment for advanced non-small-cell lung cancer

• Found strong signature in transcriptome between resistant vs. responsive cells to cisplatin

• Led to additional funding
  — Planned clinical trials with drugs

Hsu et al. 2007

FORENSIC BIOINFORMATICS AND REPRODUCIBLE RESEARCH IN HIGH-THROUGHPUT BIOLOGY

“Data processing, however, is often not described well enough to allow for exact reproduction of the results,

Thanks: Malachi Griffith

Baggerly and Coombes 2009
Digging revealed:

- Instances of repeated sampled data
- Only 84/122 test samples were distinct
- Some repeated samples labeled both sensitive and resistant
- Row offset in data table
The trouble with retractions: Nature News 2011

A surge in withdrawn papers is highlighting weaknesses in the system for handling them.
“the frequency of retraction varies among journals and shows a strong correlation with the journal impact factor”


Website shows retraction

PubMed

RETRACTED ARTICLE

Pharmacogenomic strategies provide a rational approach to the treatment of cisplatin-resistant patients with advanced cancer.

PubMed ID: 20877042

• Keep community updated
• Help kill zombie papers that keep getting cited when they should not
• Starting to get integrated into different websites for automatic scans
• Be sure you are never keeping zombies alive
But there are lots of errors out there ...

In most instances, this is scientific progress ...

But, you must navigate these to calibrate your expectations and approaches

Bioinformatics: get it right!

Can happen using the most basic tools / steps in genomics:

• Clustering of groups
• Mapping of reads against genome
• Comparative sequence alignment
Enterotypes of the human gut microbiome

we identify three robust clusters (referred to as enterotypes hereafter) that are not nation or continent specific ... mostly driven by species composition

Published cluster was generated by setting to generate 3 clusters.
The only robust cluster found inherent in the data is by sequencing technique

Acute myeloid leukemia (AML) is a cancer of myeloid blood cells
   — sequencing the complete genomes of primary tumor, relapsed tumor, and matched normal (skin) samples
AML relapse is associated with the addition of new mutations and clonal evolution, which is shaped, in part, by the chemotherapy
AML genome in an individual patient is clearly a ‘moving target’; eradication of the founding clone and all of its subclones will be required to achieve cures.

Ding et al. 2012 Nature
How well can we track early stages of relapse?

Intratumor genetic heterogeneity (ITGH)

- The coexistence of genetically distinct but clonally related cancer cells within the same patient

- 34%–80% of the discordant somatic variants, which could be interpreted as ITGH, were found to constitute technical noise

- Excluding mutations affecting low mappability regions or occurring in certain mutational contexts was found to reduce artifacts
Codon based tests of selection

$d_N/d_s$ ratio

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

Evolution of genes and genomes on the *Drosophila* phylogeny

Drosophila 12 Genomes Consortium 2007 Nature
Genome-wide selection dynamics:

How robust are these conclusions?

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

Number of significant genes

<table>
<thead>
<tr>
<th>Aligner</th>
<th>95% (a)</th>
<th>99% (b)</th>
</tr>
</thead>
<tbody>
<tr>
<td>AMAP</td>
<td>817</td>
<td>213</td>
</tr>
<tr>
<td>MUSCLE</td>
<td>1043</td>
<td>306</td>
</tr>
<tr>
<td>ProbCons</td>
<td>1013</td>
<td>281</td>
</tr>
<tr>
<td>T-Coffee</td>
<td>1290</td>
<td>479</td>
</tr>
<tr>
<td>ClustalW</td>
<td>902</td>
<td>261</td>
</tr>
<tr>
<td>Total in 5</td>
<td>1902</td>
<td>673</td>
</tr>
<tr>
<td>PRANK</td>
<td>468</td>
<td>49</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
- Removing these regions doesn’t fix all problems (Gblocks)

What makes us difference from chimps?
Is it really just 2%
Using better alignments, only 2 genes of original 59 remained significant!! (a huge bioinformatic effect)

- Many chimpanzee-specific divergent sites are adjacent to indels
- removing nucleotides within five positions of indels abolished most adaptive signals

**How do we avoid Apophenia?**

- Double check your tables and analyses
  - Plot your data, look at it, does it make sense?

- Test your hypotheses in an independent way
  - Test your findings using separate data and a different analysis
  - Functional Validation
Published studies allow ...

You to practice your bioinformatics

Assess their repeatability

Papers need enough details for replication

Functional Validation

Rodenburg 2017
Gatedness in horses

Andersson 2012 Nature

Horse WT LRLII.KXXXEUGAVWLVLSPPQRPQGRDSB.CRTSRA.EPEHLSJPSKGSK
Cattle GGGGGGCGCGCGCGCGCG
Human CVCVGCVCVCVC
Chimp CCVCVCVCVC
Dog CVCVCVCVC
Mouse GGGGGGC
Rat CVCVCVC
Chicken CVCVCVC
Zebrafish CVCVCVC

Scores from breeding field test

Number of horses

Nature
On the importance of negative results

There is a great need, and little incentive to publish negative results

How can we change this?
- Free publication charges
- Change the name from negative to .... ?
- ???