So .. how do we avoid Apophenia?

- Well ... let's ask ChatGPT

—how can humans overcome biases from apophenia
Test your hypotheses in independent ways

• Genomic datasets:
  – These are really observational data where patterns we observe have been created by things we barely understand
  – This is similar to all studies using observational data
    • Very susceptible to false positives
Genomic analyses easily find spurious correlations

Number of people who drowned by falling into a pool correlates with Films Nicolas Cage appeared in

Correlation: 66.6% (r=0.666004)

- Nicholas Cage
- Swimming pool drownings

https://www.tylervigen.com/spurious-correlations
Genomic analyses easily find spurious correlations

Number of people who drowned by falling into a pool correlates with Films Nicolas Cage appeared in

Correlation: 66.6% (r=0.666004)

- [https://www.tylervigen.com/spurious-correlations](https://www.tylervigen.com/spurious-correlations)
Test your hypotheses in independent ways

• Genomic datasets:
  – These are really observational data where patterns we observe have been created by things we can barely envision
  – This is similar to all studies using observational data
    • Very susceptible to false positives

• Manipulation: functional validation via manipulation of genes, pathways, environments ... real hypothesis testing!!
  – Experimental evolution, CRISPR KOs, environmental perturbations

• If you can’t manipulate, at least triangulate!
Triangulation for building evidence

Tree Height = \( \tan \alpha \times \text{Distance} \)
Triangulation for building evidence

• Combine insights from independent axes of insight
  – biological replicates, test RNA patterns using proteins, etc.

• Challenge is maintaining genomic scale
  – Genome wide SNP scan for outliers, QTL mapping, RNA-Seq, knockouts, manipulations, etc.

Triangulate rather than justifying your P-value based on one dataset

Knockout affects phenotype
These genes are DE
Outlier SNPs follow trait in F2 cross

What was ancestral state?
Is there any clinal variation?
Phenotype respond to chemical manipulation?
Response to selection experiment?
Three examples of triangulation in non-model species

- Population Genomics investigation of an adaptive phenotype
  - Independent genomic datasets
  - Orthogonal analyses

- Bioinformatic analysis of miRNA targets
  - Comparison across bioinformatic tools to assess consistency
  - Developing novel metric for biological signal in results
  - Comparative analysis for general insights and cross-check

- Functional genomic study of phenotypic plasticity
  - Experimental evolution
  - GWAS, RNAseq
  - CRISPR-Cas gene KO
Local adaptation

• Genomic scans may not be related to the trait you are focused upon
  – Large effect alleles at few loci
    • hard sweeps easy to detect via Fst, many tests
  – Many small effect alleles at many loci
    • Soft sweeps very, very difficult to reliably detect

• What is the genomic architecture of your trait of interest?
Population Genomics investigation of an adaptive phenotype

- Why does this dark morph exist?
  - Why female limited?
  - Is this an adaptive phenotype?
  - How and why did it evolve?

- Goal: find the genes, study their function
  - Connect genes to ecology

- Natural history
  - Common butterfly across Eurasia
  - Subspecies with female only dark morph in northern range limits (Sweden, Norway, Finland)
Genomic scans for local adaptation

Population re-sequencing using Pool-Seq across Europe (n=24 each, thorax)
What is the genomic architecture of your trait of interest?

Outliers may not be related to your focal trait

- Large effect alleles at few loci
  - hard sweeps easy to detect via Fst, many tests

- Many small effect alleles at many loci
  - Likely to have no outliers using genomic scans for selection
Test hypothesis using independent method: crosses
Fst of each population compared to dark morph population
Baysian analysis of all crossing data

Fst for 10 kb windows

Fst for each bp

Fst for each bp

Cortex
Cortex

Has a common role in wing pigmentation and patterning across Lepidoptera, likely via scale cell developmental processes

Tunström et al. (in prep)
Population Genomics investigation of an adaptive phenotype

• Outliers may have nothing to do with your view of how things work

• Intersection with orthogonal analysis is critical to gain deeper causal insights
  — Without validation steps, naked conclusions make weak contributions to the literature

• Here: intersection between genomic scan and crosses localized adaptation signal to single genomic region
Bioinformatic analysis of miRNA targets

Does miRNA play a role in diapause progression in Pieris napi
The role of miRNA in sculpting the transcriptome
The role of miRNA in sculpting the transcriptome

Transcription factor

Genome targets

miRNA

miRNA targeting

AGO

miRNA binding mRNA

post-transcriptional repression

mRNA degradation
translational repression

proteins
Regulatory network view of miRNA impacts

miRNA family

Functionally related genes

Unrelated genes

Functionally related genes

Protein

Protein

Protein

Protein

Protein

Protein

Protein
miRNA expression changes
OK, so some miRNAs are changing through time..

Where are they targeting? 
What are they doing?

What functional groups or pathways might they regulate?
miRNA target detection

- miRNAs primarily bind a very short, ±7 bp region of the 3'UTR of mRNA
- This binding ultimately leads to a decrease of translated proteins
- There are 100,000's of 7 bp motifs in genome, of which miRNAs bind small fraction
Assessing functional enrichment for targets of each predicted miRNA gene

![Graphical representation of miRNA family and target network]
Assessing functional enrichment for targets of each predicted miRNA gene
Why variation in functional enrichment in targets

- Targetscan was run using 7 species alignment of 3’UTRs, identifying 7 bp motifs that were identical
  - Under strong purifying selection, a expected when functional
- miRanda, RNAhybrid
  - Run on only 1 species, appear to have a very false positive rate
  - This is well documented in literature

So, if TargetScan is really doing better, can I find functional enrichment in other species target sets?
For some miRNAs, their targets are enriched for genes of related function.
So, why don’t more people use Targetscan with alignments?

- **Running miRanda:**
  - Download, load 3’UTR data from your species, load miRNA seed sites, run
So, why don’t more people use Targetscan with alignments?

- Running TargetScan7 with alignments
  - Download scripts, generate 3’UTR alignments for 7 species, load miRNA seed sites, run
Bioinformatic analysis of miRNA targets

• Detecting miRNA expression changes is easy, but target detection is inherently very difficult

• Intersection
  — Comparison across bioinformatic tools
    • Revealed inconsistent results, primarily because used VERY different methods (e.g. using vs. not using alignments)
  — Developed novel metric for assess biological signal in results
  — Used cross species comparisons for cross-check & generality

• Here: intersection across divergent methods, 1st principals metric, and comparative analysis using other data
This is a piece of toast
Functional genomic study of phenotypic plasticity

- Identifying the genetic basis of plastic phenotypes is very challenging

- Here researchers used
  - Experimental evolution to fix trait so they could map it
  - GWAS between the alternative lines of high vs. low trait
  - RNAseq between the alternative lines of high vs. low trait
  - CRISPR-Cas gene KO to test candidate genes
Genomic architecture of a genetically assimilated seasonal color pattern

- Made selection line having no plastic response
- Crossed back to plastic line
- GWAS on offspring for plastic response

Contig:

- No differentially expressed genes

-Log10(p)
Functional genomic study of phenotypic plasticity

• An integrated study identified several genes underlying a plastic phenotype
• Integration involved
  – Manipulation of trait using experimental evolution
  – Intersecting GWAS and RNAseq results
  – Functional validation using gene KOs
• Importantly
  – Investigated gene without annotation, found functional association, increased knowledge of phenotype for future studies
On the importance of functional validation

- P-values do not indicate effect size
- Genes likely do not function the way we image
- Organisms are gloriously more complex than we can imagine

Without functional validation, we let past glimpses of insight retard progress towards deeper understanding
Review

Functional genomic tools for emerging model species

Erik Gudmunds, 1,* Christopher W. Wheat, 2 Abderrahman Khila, 1,3 and Arild Husby 1,*
(A) [Diagram showing genetic sequences and modifications.]

(B) Removal of CRE with dual sgRNAs

(C) CDS disruption with NHEJ (null mutation)

(D) CRE replacement with HDR

(E) Codon replacement with HDR
Bioinformatic wisdom, pt. 1

• Expect errors and noise
  – Analysis results need many rounds of refinement
  – Invoke biological causes of results last

• 70% of your time will be troubleshooting
  – This is normal, keep a notebook, intermediate files

• Fear the new and shiny programs that will simplify your life
  – 80% of all new software will not be usable
    • Un-installable, no manual, no test examples, not repeatable
Cookbooking ...

- Google and AI are your friends
- Use them, but don't trust them ..
- Test what you use, then learn from it.
Keep good bioinformatic notes

• I keep a special file with commands I learned and like
  – use it to quickly find commands, refresh memory

• Use positive and negative controls to test the output of the commands you run
  – I call these sanity checks
  – Always test to make code is working correctly
    • Great reason to use > 1 method, right?

• Read up on good file structure, version control, and how to parallelize your commands (Doug’s lecture was awesome)
Publish your code, no matter how messy

Yours is without a doubt the worst code I've ever run

But it runs
Many different ways to make a pipeline

Many ways to run, performance varies across species, samples, etc.
Analysis paralysis is common

Which is the right way?

• Just get through a single pipeline
• Then try different approach to assess your first results
Bioinformatic wisdom, pt. 2

• If all publications provided all their code, science would advance faster, with more accuracy

• Provide your code with all your publications, along with all your data. Be part of the solution.

• Look at others code:
  – Discover new ways of coding, reporting
  – Become frustrated that other published work is not repeatable

• If work is not reproducible, how much can we trust it?
Data management
- Get your raw data uploaded to ENA as soon as possible.
- It's a free backup and you can set embargo date
  - keep pushing the date on the embargo

Reproducibility is super important
- Know about Snakemake or Nextflow ... but
- Be careful of how you invest your time, as some people will try to convince you to learn their pipeline ... that you use once ...

Is the pipeline for
- you, or others
- A few, or many samples?
Here come the genomes .... and all their glorious errors ...

—Annotation
—Gene alignment
—Functional annotation
Get ready, here come the $1000^n$ genomes

An unprecedented opportunity for large scale errors?

- Phylogenetic relationships
- Genome evolution
- Functional insights into genes and genomic features (e.g. regulation and inheritance)
So ... how many of you are sequencing a genome?

- What does that mean? Have you told your mom?
- What kind of genome are you generating?
- How accurate do you need your genome to be?
  - Short term vs. long term goals?
  - Are these in conflict?
Identifying the causes and consequences of assembly gaps using a multiplatform genome assembly of a bird-of-paradise

They made lots of assemblies along the way

<table>
<thead>
<tr>
<th>Assembly</th>
<th>Technology</th>
<th>Software</th>
<th>Contig N50 (bp)</th>
<th>N contigs</th>
<th>Scaffold N50 (bp)</th>
<th>N scaffolds</th>
</tr>
</thead>
<tbody>
<tr>
<td>lycPyrIL</td>
<td>Illumina HiSeq2500 (PE + MP)(^c)</td>
<td>ALLPATHS-LG</td>
<td>620,719</td>
<td>10,766</td>
<td>4,227,710</td>
<td>3,216</td>
</tr>
<tr>
<td>lycPyrPB</td>
<td>PacBio RSII C6-P4</td>
<td>Falcon</td>
<td>6,644,420</td>
<td>3,422</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>lycPyrSN1</td>
<td>10X Genomics Chromium HiSeqX</td>
<td>Supernova2</td>
<td>144,856</td>
<td>29,791</td>
<td>4,360,585</td>
<td>13,934</td>
</tr>
<tr>
<td>lycPyrSN2</td>
<td>10X Genomics Chromium HiSeqX</td>
<td>Supernova2</td>
<td>149,640</td>
<td>27,366</td>
<td>4,748,626</td>
<td>14,217</td>
</tr>
<tr>
<td>lycPyrHiC</td>
<td>PacBio + Phase Genomics Hi-C</td>
<td>Proximo</td>
<td>6,644,420</td>
<td>3,422</td>
<td>70,588,898</td>
<td>2,927</td>
</tr>
<tr>
<td>lycPyrILPB</td>
<td>lycPyrIL + gap-filling with PacBio</td>
<td>PBJelly</td>
<td>1,982,606</td>
<td>6,895</td>
<td>4,229,628</td>
<td>3,216</td>
</tr>
<tr>
<td>lycPyr2</td>
<td>PacBio + Dovetail CHICAGO</td>
<td>HiRise</td>
<td>6,294,665</td>
<td>3,463</td>
<td>6,644,037</td>
<td>3,227</td>
</tr>
<tr>
<td>lycPyr3</td>
<td>lycPyr2 + 10X Genomics</td>
<td>ARCS + LINKS</td>
<td>6,294,665</td>
<td>3,463</td>
<td>8,009,555</td>
<td>3,121</td>
</tr>
<tr>
<td>lycPyr4</td>
<td>lycPyr3 + Phase Genomics Hi-C</td>
<td>Proximo</td>
<td>6,294,665</td>
<td>3,463</td>
<td>69,071,023</td>
<td>1,713</td>
</tr>
<tr>
<td>lycPyr5</td>
<td>lycPyr4 + manual curation with alignments + gap filling</td>
<td>PBJelly</td>
<td>7,540,011</td>
<td>3,269</td>
<td>74,173,823</td>
<td>1,700</td>
</tr>
<tr>
<td>lycPyr6</td>
<td>lycPyr5 + manual curation with Hi-C</td>
<td>Juicer</td>
<td>7,540,011</td>
<td>3,271</td>
<td>74,173,823</td>
<td>1,700</td>
</tr>
</tbody>
</table>

Errors that can happen in assemblies

Genomes are scary and messy, especially when we re-assemble them with crude tools

MHC IIB: complex tandem repeats as a case study

High diversity

Highly diverse copies are very difficult to assemble

Low diversity

All loci are from same chromosome

Very challenging to place them accurately
Post-genomics challenge

“What we can measure is by definition uninteresting and what we are interested in is by definition immeasurable”
- 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 the genomic scale”
- Wheat 2015

For example:
- structural variants
.. but revisit Evan Eichler’s talk, there is hope for the future!
• Using RNAseq and protein alignments to identify gene regions and exon boundaries
Comparative genomics commonly use annotations.

Typical genome report comparing gene content among species:
- Rates of birth, death
- Lineage specific genes
Estimates of gene evolution rely upon good annotations.
Gene birth-death dynamics

[Diagram showing a phylogenetic tree with various species and their birth-death dynamics dates.]
Gene birth-death dynamics

• Do changes in gene numbers have physiological meaning?

• Fundamental and important evolutionary question

• Very difficult to assess accurately
  – Need good genomes, annotations
  – Then good analyses
Are all annotations equal among species?

- Do species genomes differ in:
  - When they were sequenced, thus technology?
  - The quality of their assembly (e.g. N50, haploid state)?
  - How they did their annotation (proteins only vs. lots of RNAseq)?

Then resulting annotation protein sets likely differ due to technology, not biology.

Will this impact analyses that rely upon accurate protein sets?
Non-standard annotations introduce major artifacts

- Lineage specific genes inflated by
  - 10 to 1000’s of genes, with increases up to 15 fold

Weisman et al. 2022. Current Biology
What are the ramifications?
Gene content evolution in the arthropods
"Although the majority of these gene sets were built using MAKER, variation in annotation pipelines and supporting data, introduce a potential source of technical gene content error in our analysis."
Proteins sets:
a mixed bag of isoforms and pseudo-duplicates

- Unfortunately, many studies are not isoform filtering their protein sets prior to analysis
  - Using raw protein sets from genome projects must always be filtered down to one protein per locus
  - This will have ramifications at all levels
    - Will severely impact ortholog assessments, gene birth death analysis

- Some genomes are not properly haploidified
  - Causes a pseudo-inflation of predicted genes
  - Creates artifacts in analyses
BUSCOs, when used properly, are very helpful

- Never report only complete BUSCO estimates
- Single copy and duplicated components are important
  - single copy indicates completeness
  - duplicated indicates haploid status
- If not haploid, mapping your data to it will be very problematic
Species with nearly 2x gene content has high duplicated %

<table>
<thead>
<tr>
<th>Family</th>
<th>Species</th>
<th>Predicted Genes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Nymphalidae</td>
<td><em>Heliconius melpomene</em></td>
<td>20075</td>
</tr>
<tr>
<td></td>
<td><em>Heliconius erato lativitta</em></td>
<td>14613</td>
</tr>
<tr>
<td></td>
<td><em>Heliconius erato demophoon</em></td>
<td>14517</td>
</tr>
<tr>
<td></td>
<td><em>Junonia coenia</em></td>
<td>19234</td>
</tr>
<tr>
<td></td>
<td><em>Melitaea cinxia</em></td>
<td>16667</td>
</tr>
<tr>
<td></td>
<td><em>Bicyclus anynana</em></td>
<td>22642</td>
</tr>
<tr>
<td></td>
<td><em>Maniola jurtina</em></td>
<td>36294</td>
</tr>
<tr>
<td></td>
<td><em>Danaus plexippus</em></td>
<td>15130</td>
</tr>
</tbody>
</table>

Legend:
- Non-reported pest
- Polyphagous herbivore
- Reported pest
- Monophagous herbivore
- Complete, single copy
- Complete, duplicated
- Fragmented
- Missing
Put the **BIO** in your informatics!!

Use independent analyses as ‘controls’

—What are your + and - controls?

<table>
<thead>
<tr>
<th>Analysis # 1</th>
<th>Analysis # 2</th>
<th>Analysis # 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mapper</td>
<td>HiSat2</td>
<td>Bwa-mem2</td>
</tr>
<tr>
<td>Normalization</td>
<td>none</td>
<td>TMM</td>
</tr>
<tr>
<td>Analysis</td>
<td>PCA</td>
<td>DEseq</td>
</tr>
</tbody>
</table>

Should independent methods converge?
Interrogate your results

• “you need to be in charge of the analysis”

• The more you analyze your data, your confidence will grow
  — Let your findings talk to you in different ways

• Graph your results – visualize the patterns, assess 1st principals
  — Always start with PCA or MDS plot (how do your samples cluster?)
  — Compare with your different analysis results

• If you find interesting genes or patterns, can you test this hypothesis?
  — Using independent samples?
  — At a higher level of biological organization?
  — In some manipulative, functional way?
Molecular spandrels:

Story telling vs. Causal understanding

Genomics is full of adaptive stories
Treat your findings as hypotheses
How can you test these?
Never forget your origins and biases

Find ways to test your genomic hypotheses, cause they are easy to get and believe
Come say hi if you’re in town!