THE TROUBLE WITH RETRACTIONS

BY RICHARD VAN NOORDEN

A surge in withdrawn papers is highlighting weaknesses in the system for handling them.
The trouble with retractions: Nature News 2011

"the frequency of retraction varies among journals and shows a strong correlation with the journal impact factor"

Publications with significant human error that have not been retracted

**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”

**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

**Genome–wide signatures of convergent evolution in echolocating mammals**

More genes underwent positive selection in chimpanzee evolution than in human evolution

**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

**Human – Mouse TMRCA**

~ 90 MYA

**Correlation**

- 1
- 0.6
- 0.8
- 0.7
- 0.6
- 0.5

**Brain – Kidney TMRCA?**

“[after accounting for the batch effect, … brain and mouse were clustered by tissue, not by species” Gilad and Mizrahi-Man 2015. F1000 Research
Batch effect: confounding 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 C2HKACXX, 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></td>
</tr>
<tr>
<td>kidney</td>
<td>adrenal</td>
<td>adrenal</td>
<td>kidney</td>
<td></td>
</tr>
<tr>
<td>liver</td>
<td>sigmoid colon</td>
<td>sigmoid colon</td>
<td>liver</td>
<td></td>
</tr>
<tr>
<td>small bowel</td>
<td>lung</td>
<td>lung</td>
<td>small bowel</td>
<td></td>
</tr>
<tr>
<td>spleen</td>
<td>ovary</td>
<td>ovary</td>
<td>testis</td>
<td></td>
</tr>
<tr>
<td>testis</td>
<td>pancreas</td>
<td>pancreas</td>
<td>Human</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Mouse</td>
<td></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?

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

Result of using supervised clustering method, set to generate 3 clusters. Their clusters are not a property of dataset, as the only robust cluster is found grouping by sequencing technique

Courtesy of Paul McMurdie
Genome-wide signatures of convergent evolution in echolocating mammals

“Strong and significant support for convergence among bats and the bottlenose dolphin was seen in numerous genes linked to hearing or deafness, consistent with an involvement in echolocation.”

- 2326 orthologous genes
- Site-wise log-likelihood support (SSLS)
  - Negative values support convergence H1, H2
  - 824 mean support for H1
  - 329 mean support for H2

Parker et al. failed to conduct orthogonal ‘test’ of findings or estimate proper ‘null’ expectation

What makes us difference from chimps?

Is it really just 2%

• 201 citations since 2007
  Only 2 genes of original 59 were validated!!
  (at bioinformatic level)
  • Many chimpanzee-specific divergent sites are adjacent to indels
  • removing nucleotides within five positions of indels abolished most adaptive signals

More genes underwent positive selection in chimpanzee evolution than in human evolution
Evolutionary Inference = House of Cards?

The quality of our evolutionary inference

Is proportional to assumptions of orthology

Orthologous genes ... can their phenotypic effects drift over evolutionary time?

- RNAi phenotypes assessed for 1,300 genes in two nematodes
  - TMRA ~ 24 MYA
  - 7% had divergent phenotypic effects (in lab, etc.)
  - Likely higher in nature

Verster et al. 2014. PLoS Genet
1001 ways for your pipeline to break, or feed you sewage

An overview of genomic pipeline challenges

Informatics and Biology

• We need to make sure we put the ‘bio’ into the bioinformatics
  — Do results pass 1st principals tests
  — Always double check data from your core facility or service company
  — Use independent analyses as ‘controls’ on accuracy
    • What are your + and – controls?
    • Do independent methods converge?

• Need to re-assess our common metrics for potential bias in the genomic age
  — Bootstraps on genomic scale data
  — P-values, outlier analyses, demographic null models
Batcow says, take a break!!!!

Outline

• Transcriptome analyses in non-model species
  — Walk through pipeline and highlight issues of concern
  — What is validation?

• Insights from candidate genes
  — Can Second Gen methods get us there?
Pipeline Overview

Image from http://sfg.stanford.edu/guide.html
### Computer Infrastructure

**RNAseq dataset:**
4 conditions X 2 tissues X 3 families X 3 replicates = 72 x 10^6 reads

<table>
<thead>
<tr>
<th>Process</th>
<th>File Sizes (Gb)</th>
<th>CPUs</th>
<th>RAM (Gb)</th>
<th>Time</th>
</tr>
</thead>
<tbody>
<tr>
<td>Raw files *gz</td>
<td>(1.5 Gb)</td>
<td></td>
<td></td>
<td>~3 hours / file</td>
</tr>
<tr>
<td>Raw files expanded</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>TA assembly</td>
<td></td>
<td>64</td>
<td></td>
<td>~6 – 12 days</td>
</tr>
<tr>
<td>Mapping (BAM)</td>
<td></td>
<td>64</td>
<td></td>
<td>~5 – 20 hours / file</td>
</tr>
<tr>
<td>Annotation</td>
<td>100 Mb</td>
<td></td>
<td></td>
<td>~6 – 12 days</td>
</tr>
<tr>
<td>Analysis</td>
<td>&lt; 20 Mb</td>
<td>4</td>
<td>4</td>
<td>~&lt; 1 hour</td>
</tr>
<tr>
<td>Visualization</td>
<td>BAM files</td>
<td>≥ 4</td>
<td>≥ 8</td>
<td></td>
</tr>
</tbody>
</table>

Get ready for your data by downloading similar sized dataset from the Short Read Archive. Do not wait till it arrives.
Pipeline Overview

Core facilities and non-model species

Statements from core facilities that are not true:

- Here is your data
- You can’t do RNA-Seq without a genome
- We’ll have your data back in < 1 month
Pipeline Overview

Gene Ontology: order in the chaos

- Addresses the need for consistent descriptions of gene products in different databases in a species-independent manner

- GO project has developed three structured controlled vocabularies (ontologies) that describe gene products in terms of their associated:
  - biological processes
  - cellular components
  - molecular functions

http://www.geneontology.org/
Comparisons among annotation tools

Batch processing for GO terms

Pipeline Overview

Seqencing → Tissue collection → Library preparation → Computer setup → Quality Control processing of raw data → De novo assembly → Mapping reads to a reference → BLAST comparison & annotation
Template mismatch effects: excellent yeast study

Does alignment software matter?

Mappers don’t appear to matter

Wrong

• Genomic scale data can hide widespread biases that unless you specifically look, are hard to find

• Mapping programs differ in their settings and design
  – DNA to DNA vs. RNA to DNA
  – Are usually compared using species without much genetic variation
  – Indels, splicing, SNPs all affect mapper performance

SNP effects can be large

Insertions & deletions (indels) have large effects

15 mapping results

Dramatic differences in ability to handle a 2 bp insertion in reference compared to reads

TopHat, SpliceMap, Bowtie and Soap

— do not identify indels
— they fail to accurately align reads to these regions

Allelic bias in read mapping

- Essentially identical to allele specific PCR bias … but on a scale you can’t detect unless you care to look
- Do your genes of interest have more than 3 SNPs / 100 bp?

Sedlazeck et al. 2013 Bioinformatics
Mapping reads in outbred species

Average genome polymorphism levels (ignores indels)

Leffler et al. 2012 Plos Biol

Sig. expression differences by method

A: Stampy mapping
B: Cuffdiff analysis
C: Likely error source
Normalization matters, as it directly affects false-positive rate.

(a) Equivalent library sizes / Presence of high count genes

RNA-Seq

Real world example

2 factor analysis with family effects

Dillies 2010 Brief in Bioinfo.
Developmental plasticity in *Bicyclus anynana*

**Save energy, live long**

*long* lifespan *short*

delayed reproduction *fast*
inactive behaviour *active*

high fat reserves *low*
cryptic wing pattern *conspicuous*

**Live fast, die young**

---

**Bicyclus anynana**

Marjo Saastamoinen

Sensitive period

Environmental conditions

Alternate phenotypes
Experimental design

7 full-sib families

seasonal temperature

food stress

use 2 body parts

- 2 seasonal x 2 food stress x 2 body parts = 8 conditions
- 7 families with n = 2 - 3 per condition → 144 RNA libraries
- 10 million reads / library

<table>
<thead>
<tr>
<th>body part</th>
<th># libraries</th>
<th># clean reads (per library)</th>
<th># nucleotides (per library)</th>
<th>GC content</th>
</tr>
</thead>
<tbody>
<tr>
<td>abdomen</td>
<td>72</td>
<td>15,261,019</td>
<td>3,052,203,767</td>
<td>45%</td>
</tr>
<tr>
<td>thorax</td>
<td>72</td>
<td>15,633,416</td>
<td>3,126,683,150</td>
<td>46%</td>
</tr>
<tr>
<td>total</td>
<td>144</td>
<td>2,224,399,290</td>
<td>444,879,858,000</td>
<td>45%</td>
</tr>
</tbody>
</table>

14 samples: one from each family, thorax and abdomen

69,075 contigs

edgeR

# reads ~ season + stress + family +
season*stress + season*family + stress*family
season*stress*family
What should I be looking at first?

Colored by Family
Log fold change

Log (p-value)

Stress

Log fold change
Effect of filtering the mapping to Trinity contigs

| 0 zero-read samples allowed | 32 zero-read samples allowed | 71 zero-read samples allowed |

**D. melanogaster** lacks an orthologous reproductive physiology
Most studies are annotation limited

- What is the biological meaning of the top P-value genes?
- Low P-value or expression genes are certainly important
- Gene set enrichments are key to insights
  — Thus, annotation is very important

<table>
<thead>
<tr>
<th>Description</th>
<th>Uniprot</th>
<th>-log10P</th>
</tr>
</thead>
<tbody>
<tr>
<td>Oxidoreductase.</td>
<td>Q9VMH9</td>
<td>7.087008</td>
</tr>
<tr>
<td>Hypothetical protein.</td>
<td>Q95T13</td>
<td>6.315473</td>
</tr>
<tr>
<td>S027L10p.</td>
<td>Q85X2</td>
<td>6.300667</td>
</tr>
<tr>
<td>S001790p.</td>
<td>Q95T13</td>
<td>5.316371</td>
</tr>
<tr>
<td>Electron-transfer-flavoprotein</td>
<td>Q00CH2</td>
<td>5.4125</td>
</tr>
<tr>
<td>Pseudouridylate synthase.</td>
<td>Q9VZ2U</td>
<td>4.784378</td>
</tr>
<tr>
<td>Hypothetical protein.</td>
<td>Q9VJX0</td>
<td>4.750469</td>
</tr>
<tr>
<td>CG14636-PA (RE68889a)</td>
<td>Q9VJX0</td>
<td>4.650051</td>
</tr>
<tr>
<td>Chromosome 11 SCAF14979, w/t</td>
<td>Q8TOSR</td>
<td>4.500648</td>
</tr>
</tbody>
</table>

7 of 20 (35%) no Uniprot ID

Sources of error

Transcriptome assembly can be huge source of bias:
- Fragmentation creates multiple contigs of same gene
- SNPs and alternative splicing generates more contigs
- 1 locus = frag. X SNPs X alt. splicing = many contigs

We can observe effects in expression analyses:
  — Family effect mapping bias
  — Pseudo-inflation in Gene Set Enrichment Analyses
Put the **BIO** in your informatics!!

Use independent analyses as ‘controls’ on accuracy

— What are your + and − controls?

<table>
<thead>
<tr>
<th>Mapper</th>
<th>Analysis # 1</th>
<th>Analysis # 2</th>
<th>Analysis # 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>TopHat2</td>
<td>STAR</td>
<td>?</td>
<td></td>
</tr>
<tr>
<td>none</td>
<td>TMM</td>
<td>TMM</td>
<td></td>
</tr>
<tr>
<td>PCA</td>
<td>RSEM</td>
<td>EDGER</td>
<td></td>
</tr>
</tbody>
</table>

Should independent methods converge?

**Interrogate your results**

- “you need to be in charge of the analysis” – B. Cresko

- This will give you confidence
  — Bring freedom to your findings (no waterboarding)

- Graph your results – visualize the patterns
  — PCA or MDS plot
  — P-value distributions

- Assess gene copy number in gene set enrichment analyses (GSEA)
  — Do these levels fit to 1st principals expectations?
  — Do you have extra copies due to your Transcriptome assembly?
A major challenge for Ecological Genomics

- What causes natural selection in the wild?
  - How does genetic variation at one region of the genome interact with its environment (genomic, abiotic, and biotic)

- DNA alone can’t tell us about selection dynamics in the wild
  - Molecular tests are very weak and uninformative about selection dynamics

- Research community is demanding actual demonstration of natural selection when making claims of adaptive role
  - Triangulate!!!!

Story telling vs. Causal understanding

Genomics is full of adaptive stories

Functional and field validation of SNPs effects are needed to discern facts from fiction

Storz & Wheat 2010 Evolution
Barrett & Hoekstra 2011 Nat Rev Genet
This is ongoing work

- Currently trying to write commentary on biases in field
- Please send along other examples I might have missed
  — Feedback / critique is greatly appreciated

This is all due to the Workshop on Genomics gang, thanks to your unwavering support over the years!

Captain Skoot

The Gang