Allele specific expression: How George Casella made me a Bayesian

Lauren McIntyre

University of Florida

Acknowledgments



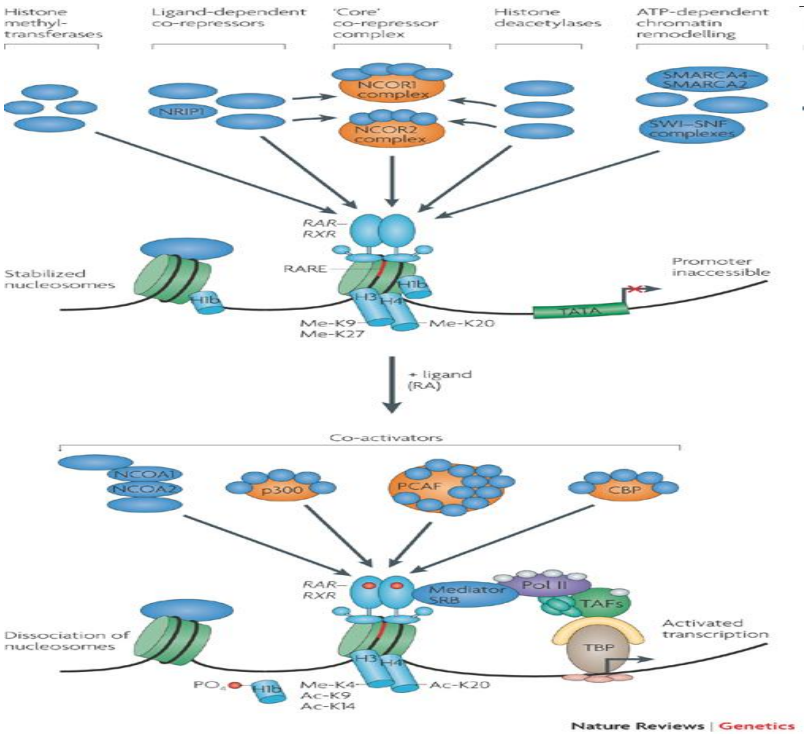


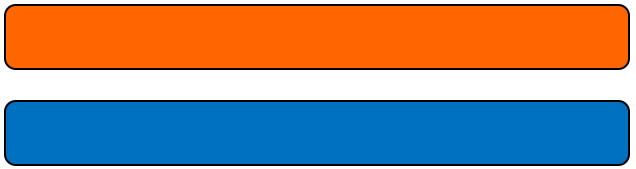


NIH ,NSF, UF EPI, UF Opportunity Fund

Allele specific expression: what is it?

- The unequal expression of alleles



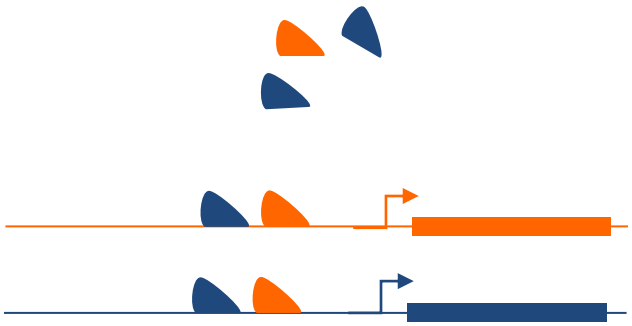


There is no genetic variation in this picture

Nature Reviews Genetics 9, 541-553 (July 2008)

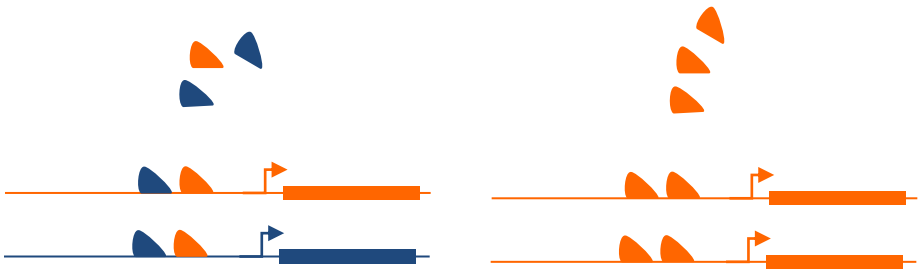
Allele specific expression: How does it happen?

- Genetic variation– polymorphism
- Polymorphisms in sequences in areas of regulatory importance at the locus itself (*cis*)
- Differences among alleles at other loci which have a regulatory role in transcription (*trans*)



Not
Equal

Cis variation



Trans variation

Not
Equal

Allele specific expression:

Genetic variation in regulatory regions of the genome

Allele specific expression: why is it important?

- Complex diseases have been shown to have regulatory polymorphisms associated with trait variation
  - autoimmune disease (Nature, 423, 506–511)
  - rheumatoid arthritis (Nat. Genet., 34, 395–402)
  - myocardial infarction and stroke (Nat. Genet., 36, 233–239)
  - diabetes (Nat. Genet.,26, 163–175)
  - inflammatory bowel disease (Nat. Genet.,29, 223–228)
  - schizophrenia (Am. J. Hum. Genet., 71, 877–892)
  - asthma (Nat. Genet., 34, 181–186)
- Genes (Human) show evidence of allele specific expression
  - Yan *et al.* 2002; Bray *et al.* 2003; Lo *et al.* 2003; Pastinen and Hudson 2004
- We have very little understanding of this paradigm

Why the fly?

- Flies are cheap
- Flies are easy
- We can get lots of the same ones again and again
- They have complex behaviors
- They are a perfect genetic system
- There are links to other systems

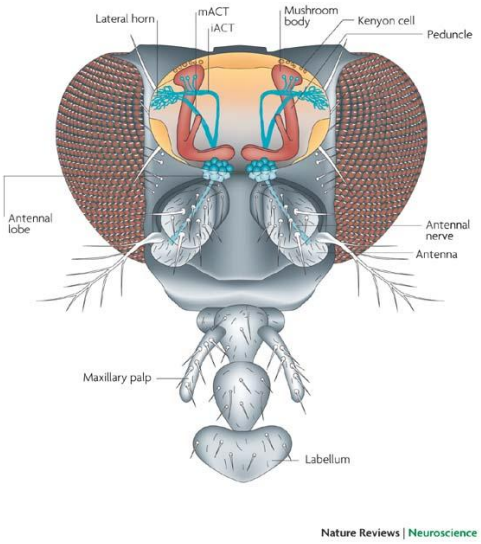
Why heads?

Olfaction, hearing and thermosensation:
Antennal segments and arista

Sight: Eyes and ocelli

Taste: Labial palps

Olfaction: Labial palps



Brain:

-reception, integration and response to sensory inputs.

-complex behaviors: mating and aggression.

-modulation of these behaviors based on environment and/or internal state.

•Many studies indicate the importance of tissue specificity in gene regulation: Isolating heads from bodies reduces complexity of the sample and focuses these studies on genes expressed in the brain and sensory organs.

•Theses tissues play a central role in the way flies sense and respond to environmental cues and enact appropriate behaviors.

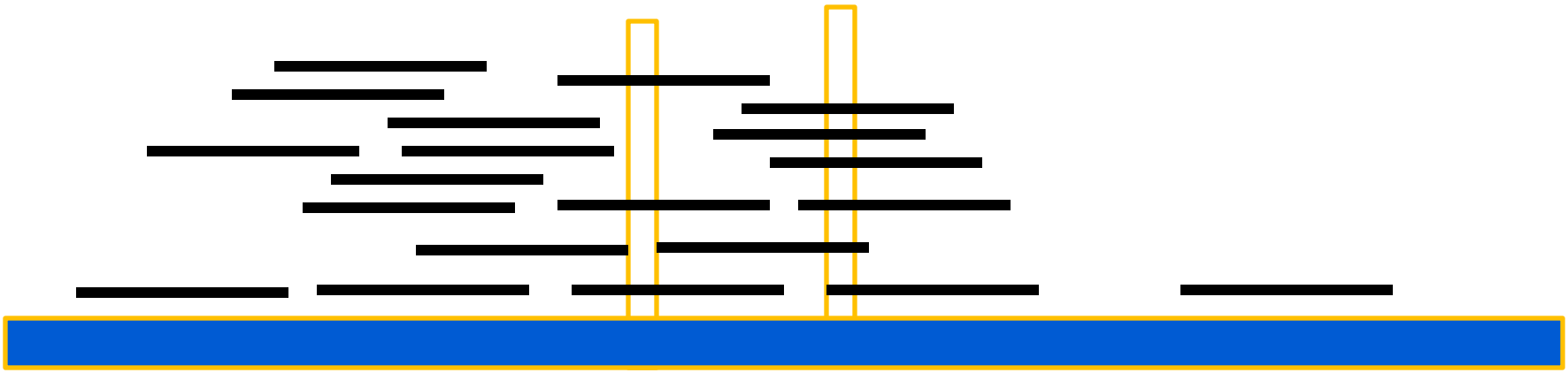
•Regulatory divergence of brain, eye and antennal genes among species may be linked to **adaptive phenotypes.**

Nat Rev Neurosci, *8*(5), 341-354. doi:10.1038/nrn2098

Measure the alleles separately

- Arrays
  - Track the alleles on tiling arrays
    - (Graze et. al. 2009)
- Next generation sequencing!
  - RNA-seq
    - Track the alleles
  - Whole genome re-sequencing
    - Find the regulatory polymorphisms

Align to a *reference* genome

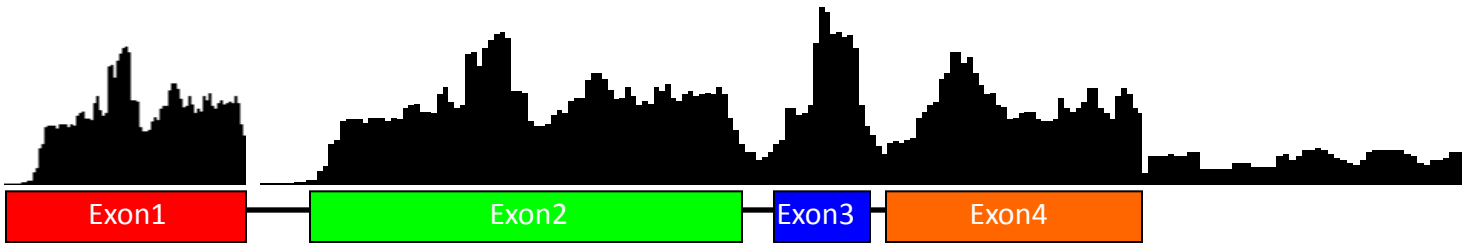


6

3

RNA-seq: The data

Gene X



Summarizing the data

- Option 1
  - Use previously identified gene models with definitions of exons/genes
  - Count how many reads (or partial reads) fall inside each exon/gene
- Option 2
  - Use the data to find boundaries of transcription
  - Count how many reads inside the boundaries

What kind of experiments will let you measure allele specific expression?

- Need a heterozygote!
  - Separate in your mind tracking the alleles from the regulatory polymorphisms that cause allelic imbalance
- F1 hybrids between species
- F1 hybrids within a population
- Chromosomal substitutions, crossed appropriately and other fun genetic designs

Experiment: F1 hybrid *D. simulans* and *D. melanogaster*:

- Divergence between these species is known to be extensive, with thousands of individual transcript level differences observed.
- 1 Sequence variant ~every 300 nt
  - Many reads on NGS will be able to be assigned allele specifically

Nat Genet, 33(2), 138-144; Science, 300(5626), 1742-1745; Mol Biol Evol, 21(7), 1308-1317; Molecular Biology and Evolution, 10(4), 804-822

Issues

- Re-sequencing relies on the reference genome
  - Reference genomes: *D. melanogaster , D simulans* assembled on a *D. melanogaster* backbone
  - Our experiment is a hybrid between D. *melanogaster* and D. *simulans*
  - Map bias can obscure allele measurements (Degner et. al. 2009)
- Technological issues with particular alleles (systematic bias)
- Structural variation Genome divergence in copy number (systematic bias)

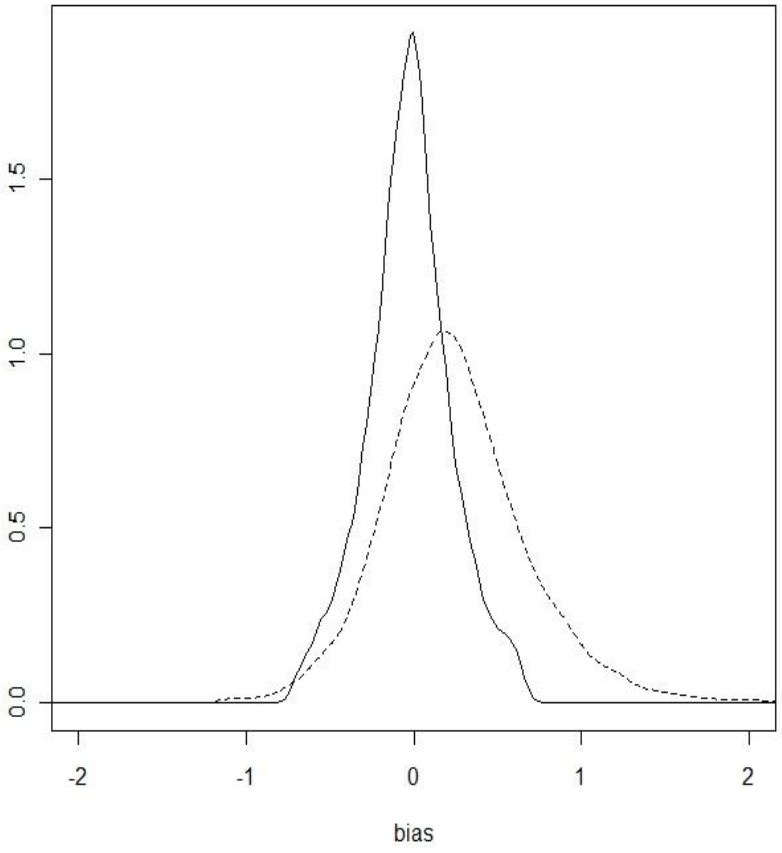
Genotype specific references

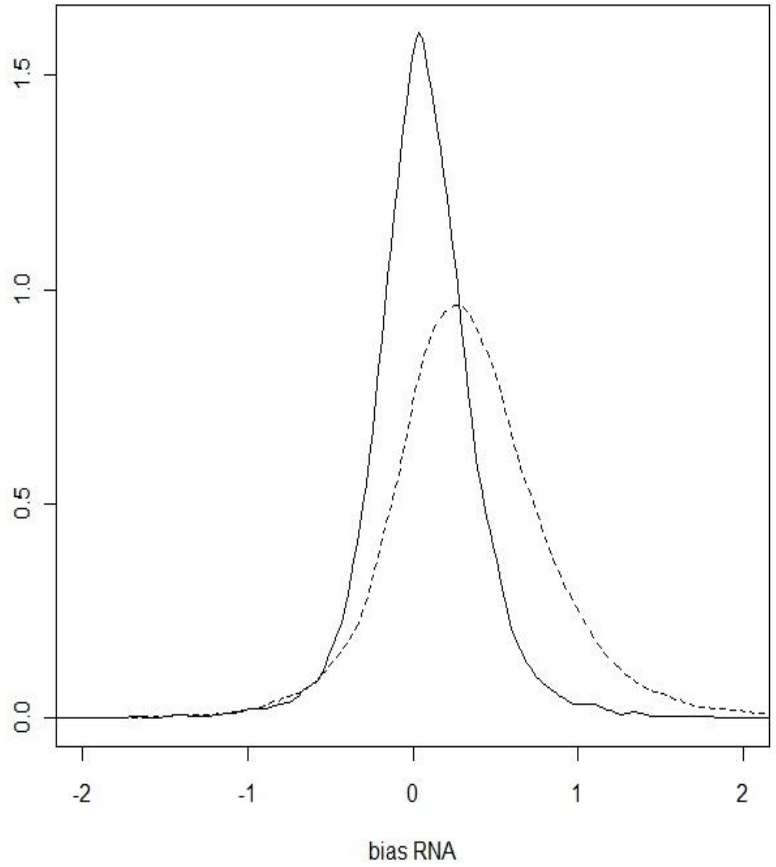
- Focus on the Exons and start with the existing reference
  - *D. melanogaster* reference genome
  - *D. simulans* DPGP sequence aligned to *D. melanogaster* reference

- Use RNA seq data from the *parents* to update the reference
  - Map reads to each reference
  - identify polymorphisms
  - Update the reference
  - Repeat until almost no polymorphisms identified

Improve alignments and reduce bias

Replicate	Total	Genome-aligned	Exon-aligned^{S}	Exon-aligned^{U}
1	40.95 M	32.0 M	25.92 M	26.4 M
2	44.81 M	34.41 M	26.44 M	26.6 M
3	42.58 M	32.78 M	28.28 M	29.0 M





Reduced error in allele-assigment

- Error in allele assignment was calculated by examining reads corresponding to exons in Mitochondrial genes (100% melanogaster)
- **initial reference**
  - **RNA:** 2.1% of the reads were erroneously assigned to *D. sim*.
  - **DNA**: 3.5%. of the reads were erroneously assigned to *D. sim*.
- **updated references**,
  - **RNA:** <1% (.09%) allele assignment error.
  - **DNA:** <1% (.45%) allele assignment error

Testing for allelic differences:

- Outstanding issues
  - Bias in technology
  - Genome duplications in one species but not the other
- DNA as a control

Bayesian Model : Reads are RANDOM

X_{ij} is the number of "A" in the RNA for biorep i and techrep j

Y_{ij} is the number of "A" in the DNA for biorep i and techrep j

i= 1,…,I and j=1,….J

RNA	DNA
$X_{ij} \mid N_i,\theta_i \sim$ Negative Binomial (N_i,θ_i)	$Y_{ij} \mid N_i,\theta_i \sim$ Negative Binomial (Y_i,p)
$\theta_i \mid p \sim$ beta $(pt,(1-p)t)$	$p \sim$ beta (v,v);

t: the strength of the prior = sum of all counts

P corrects for bias centering the prior on 1-p

θ is the proportion of reads from the M allele

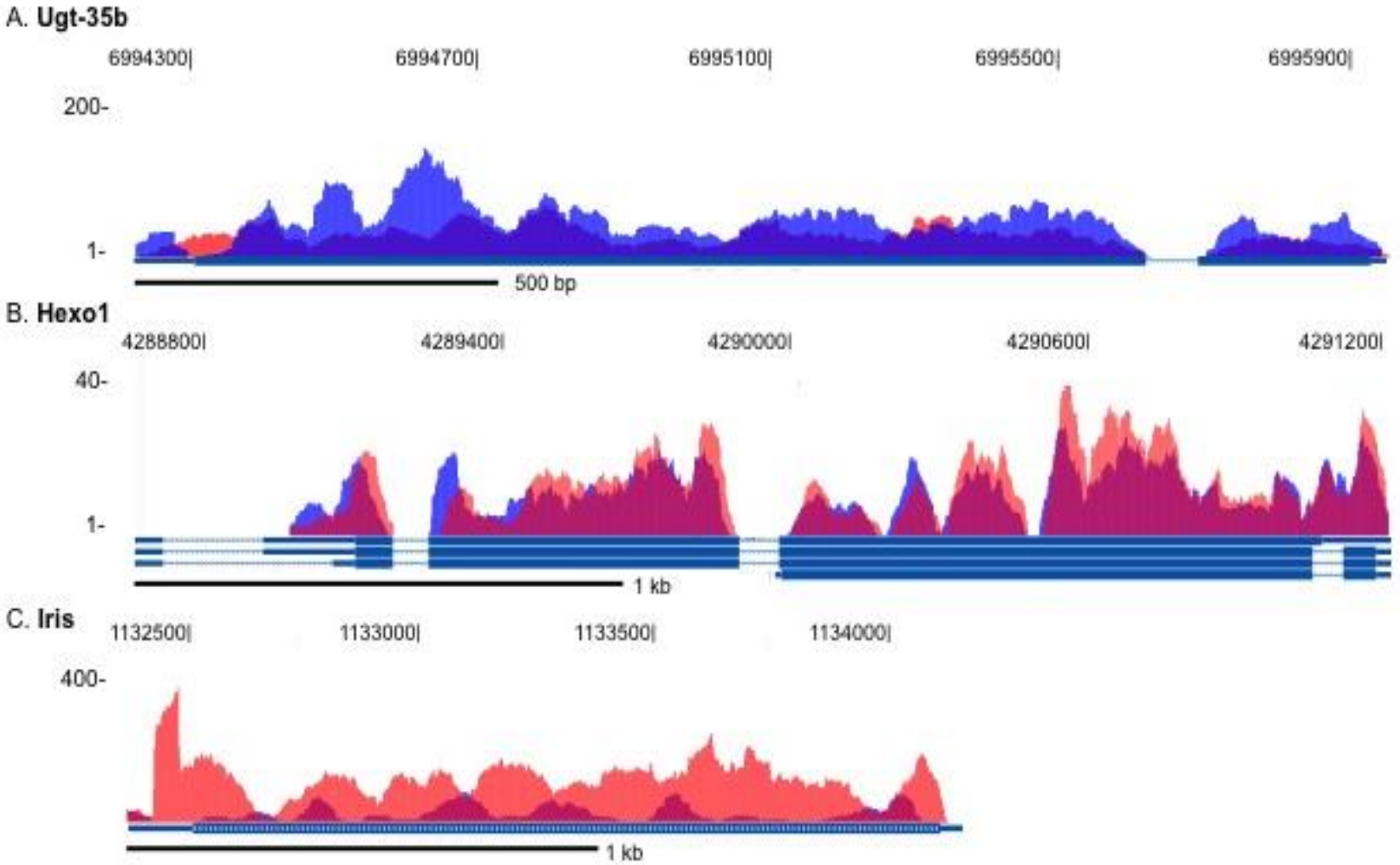
The number of counts is a RANDOM variable

Results

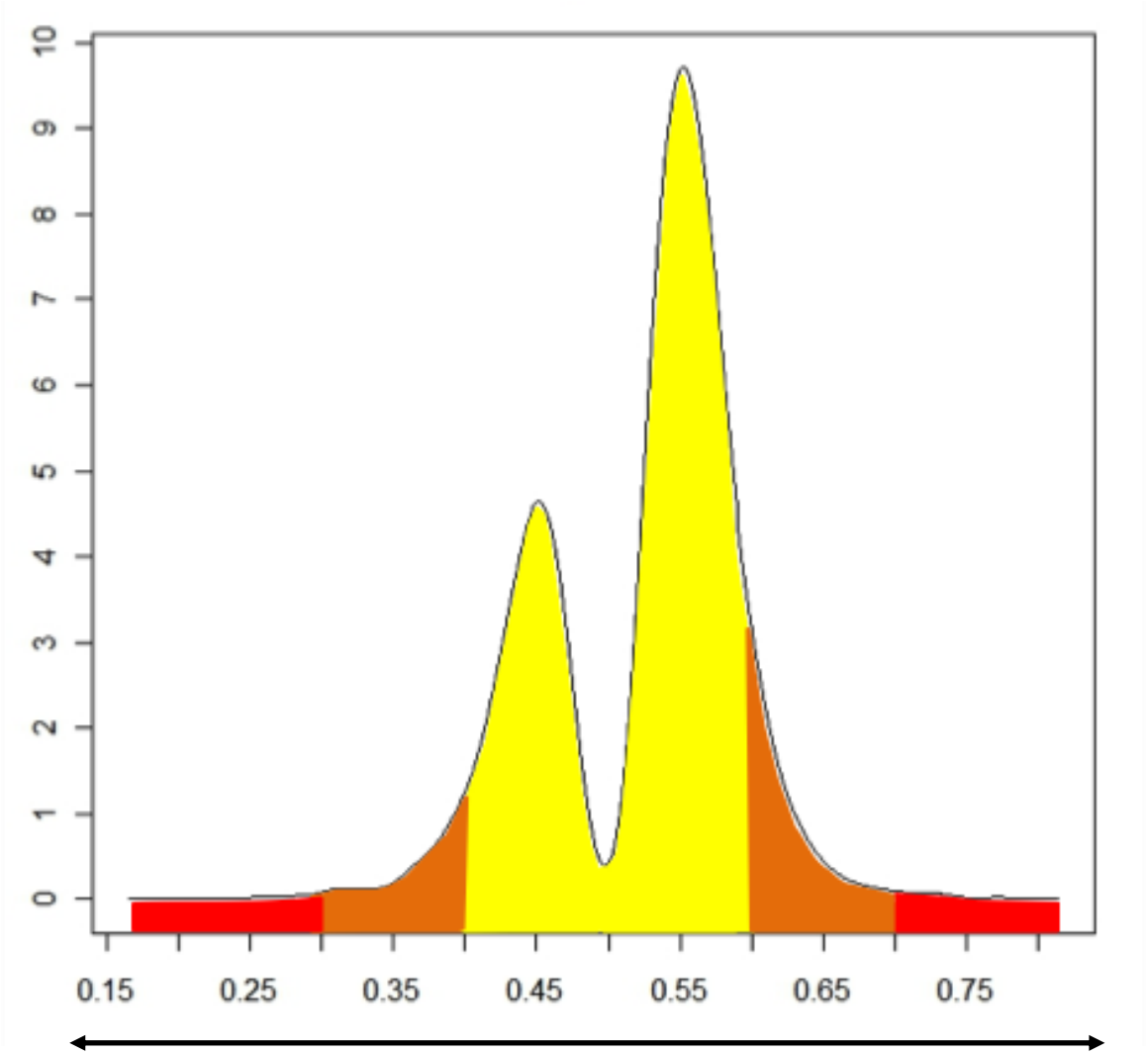
	RNA			DNA				
Genes	**Mel**	**All**	**Bias**	**Mel**	**All**	**Bias**	**θ**	**CI**
pdfr	294	369	.80	278	346	.80	.50	+/-.04
fax	168	654	.26	30	106	.28	.48	+/-.05
Iris	14048	14786	.95	1171	2572	.46	.75	+/-.01
Hexo1	541	945	.572	272	561	.49	.54	+/-.03
Ugt35b	1992	6546	.30	256	475	.54	.38	+/-.02

- From the posterior sample we compute the 95% Credible interval
- We need large counts to infer AI
  - small DNA counts estimates of p_t disperse
  - small RNA counts estimates of θ_t disperse

Some examples



How much *cis*?


0
1
2
3
4
5
6
7
8
9
10
0.15
0.25
0.35
0.45
0.55
0.65
0.75
0.15
.5
.85
D. simulans
D. melanogaster

Allelic Imbalance is widespread

- **41%** of exons (5,877) show differences in ASE – this is a result of *cis* regulatory divergence between species
  - mel biased (4,024) sim biased (1,853)
- Most *cis* differences observed are modest in effect
- McManus 2010 (mel/sech 78%) and Fontanillas 2010 mel/sim 454 (68%)

What about within species?

- Within population examination of regulatory variation
- ~200 genotypes of D. melanogaster
  - ~160 from TFC MacKay Raleigh
  - ~40 from SV Nuzhdin Winters
- Everyone crossed to a tester line (t) w1118

No more DNA

- With ~200 genotypes we can not afford to do DNA controls
- Poisson Gamma model
  - As the NB it can adjust for systematic bias
  - The adjustment is via the structure of the model and not the prior
- Simulation ?
  - (Degner et. al. 2009)

Poisson Gamma model

$X_i | \mu, \alpha, \beta_i, q \sim \text{Poisson}(\mu\alpha\beta_i q)$

$Y_i | \mu, \alpha, \beta_i, q \sim \text{Poisson}(\mu\alpha\beta_i(1-q))$

x_i is allele i counts in rep i

y_i is allele t counts in rep I

μ overall mean

β_i rep variation

q information about the bias

α when $\alpha \neq 1 \Rightarrow$ AI

Poisson Gamma

$$\theta = \frac{\mu\alpha\beta_i}{\mu\beta_i + \mu\alpha\beta_i} = \frac{\alpha}{1+\alpha}$$

Under the null α=1 and

$$E\left(\frac{x_i}{y_i}\right) = E\left(\frac{E(x_i|x_i+y_i)}{x_i+y_i}\right) = q$$

Standard gamma priors for the rest of the parameters

Compare the NB and PG

- Consider q random as in the NB model and use the DNA to inform the result

NB\PG	AB	AI
AB	0.57	0.07
AI	0.01	0.36

- Similar results

No DNA

- Simulated all possible reads from the two species
- Aligned them using bowtie with the same settings as the real data
- Estimate q_{sim}
- $q_{0.5}$ set q=0.50
- Compare PG q_{sim} vs PG q_{DNA}
- Compare PG $q_{0.5}$ vs PG q_{DNA}

DNA is the “gold standard”

q_{sim} \ q_{DNA}	AB	AI
AB	0.27	0.16
AI	0.12	0.45

$q_{0.5}$ \ q_{sim}	AB	AI
AB	0.04	0.01
AI	0.35	0.59

- Only exons where |qsim-0.5|>0.2 approximately 500
- Simulations help, the false positive rate is lower although false negatives are higher
  - They are not perfect, they only capture ambiguity in the genome and not unknown structural variation
  - There are more exons with a bias from the DNA that are not captured by the simulation,
    - unknown structural variation

Conclusions

- Bayesian models account for variability due to RANDOM effects from the number of reads
- The NB and PG models are very similar
- When there are no DNA controls simulations can help reduce false positives
  - At the expense of increasing false negatives
- There is structural variation between genomes that simulations can not capture
- There is potentially technical variation due to non-randomness of sequencing that simulations can not capture

Bayesians have more fun