Thomas Kono
2017-08-02

RIS Interview Seminar
Slides: https://z.umn.edu/TK_RIS

Who I Am

- Undergrad: UC Davis (Biochem. + Molecular Bio.)
- PhD: University of Minnesota (Applied Plant Sciences)
    - Advisors: Peter L. Morrell and Robert M. Stupar
- Postdoc: University of Minnesota (Agronomy/EEB)
    - Advisors: Candice N. Hirsch and Suzanne E. McGaugh

Outline

Can we predict the phenotypic effect of a mutation?

Tools:

SNPMeta: Annotation of SNPs in non-model species

BAD_Mutations: Predicting deleterious SNPs

Applications:

Deleterious variant identification in two crop genomes

A different mutation type: genome content variation

SNPMeta: Questions

- What information on SNPs is available through public data sources?

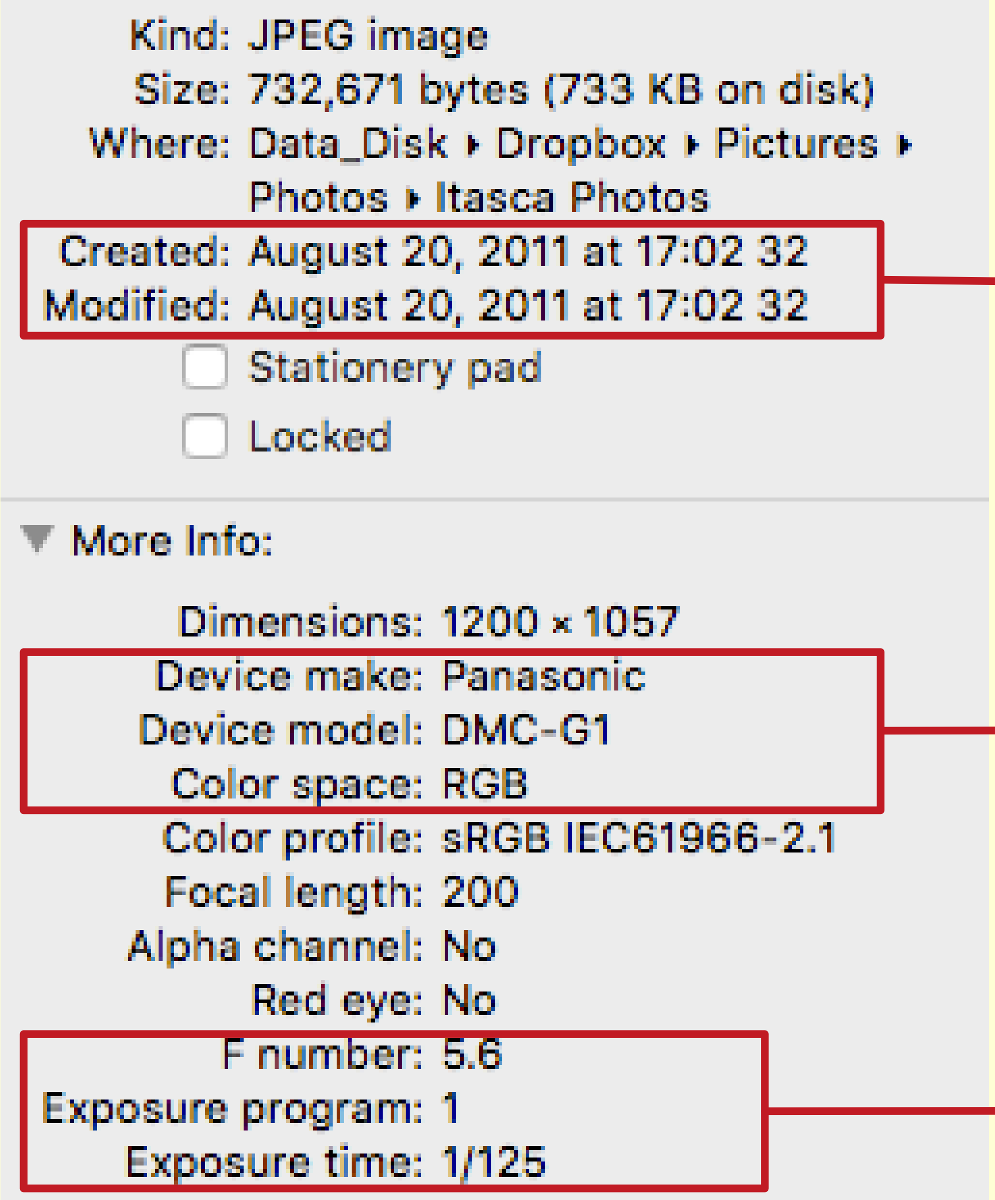
- How do annotations from GenBank entries compare to those from a reference genome?

SNPMeta: A Photo Analogy

Purpose: to collect metadata on SNPs

SNPMeta:
A Photo Analogy

Purpose: to collect metadata on SNPs



Kind: JPEG image
Size: 732,671 bytes (733 KB on disk)
Where: Data_Disk ▸ Dropbox ▸ Pictures ▸ Photos ▸ Itasca Photos
Created: August 20, 2011 at 17:02 32
Modified: August 20, 2011 at 17:02 32

□ Stationery pad

□ Locked

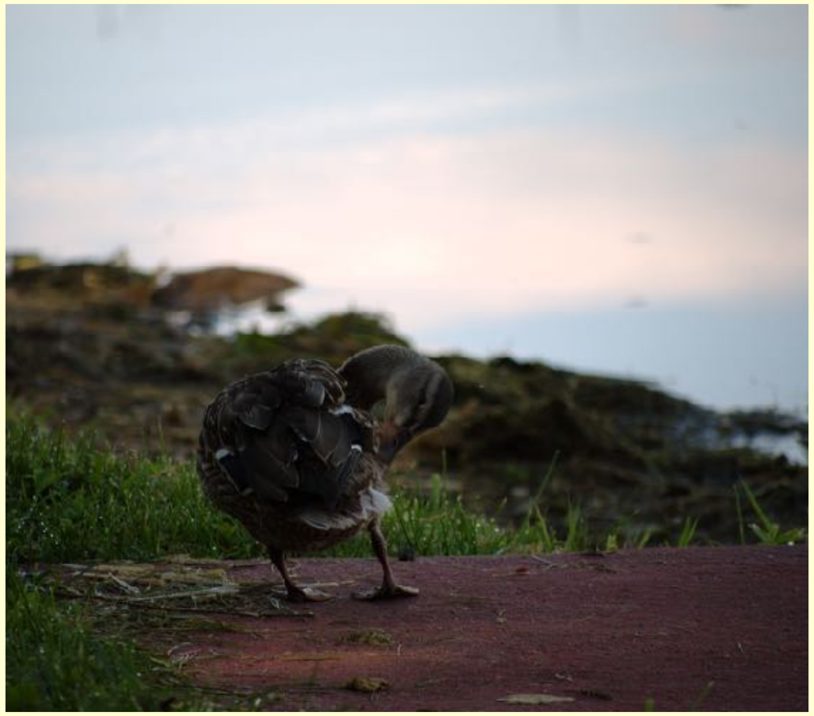
▼ More Info:

Dimensions: 1200 × 1057
Device make: Panasonic
Device model: DMC-G1
Color space: RGB
Color profile: sRGB IEC61966-2.1
Focal length: 200
Alpha channel: No
Red eye: No
F number: 5.6
Exposure program: 1
Exposure time: 1/125

Date info

Camera info

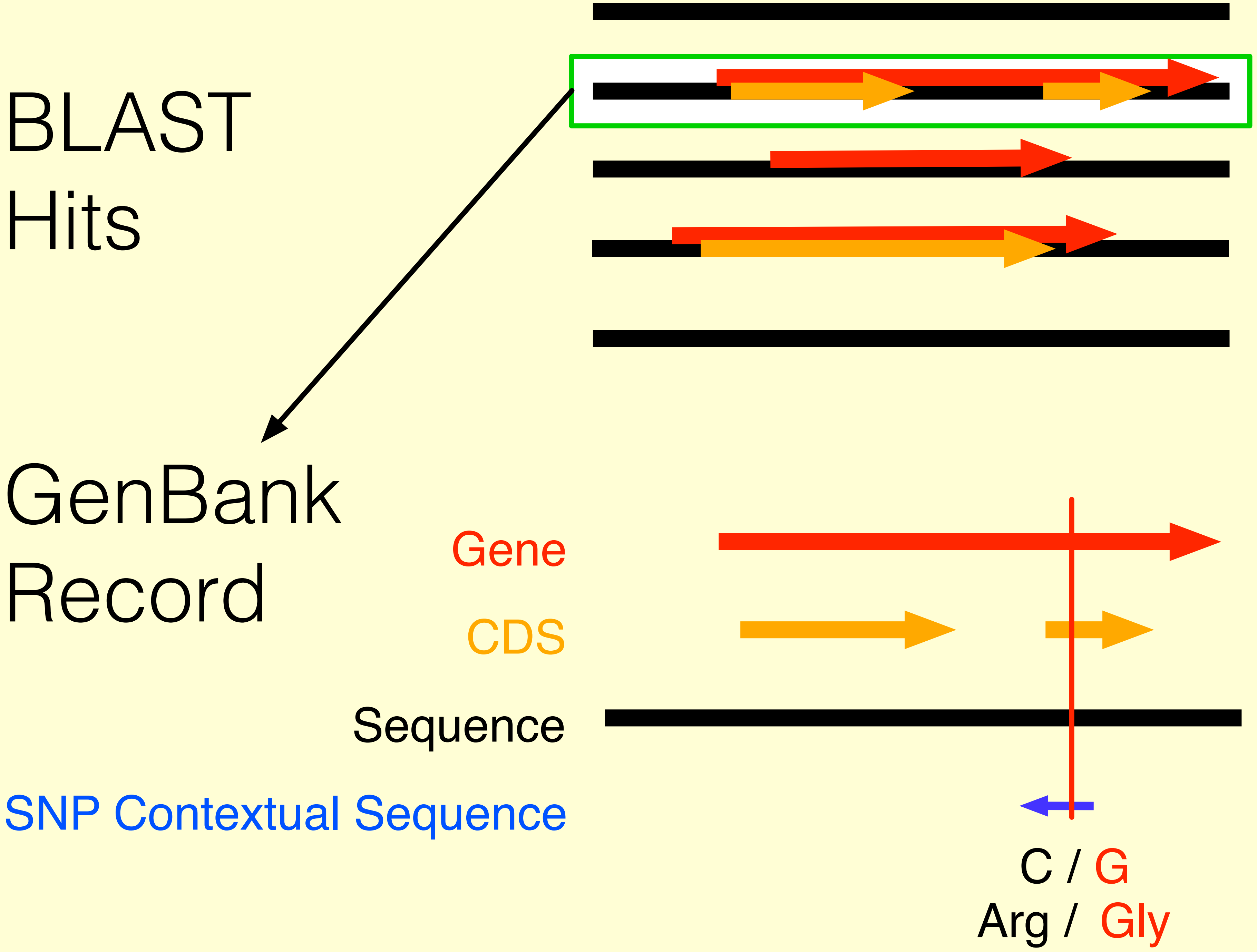
Exposure info



SNPMeta:
Pipeline Development

Amber Eule-Nashoba

Kiran Seth



BLAST Hits

GenBank Record

Gene

CDS

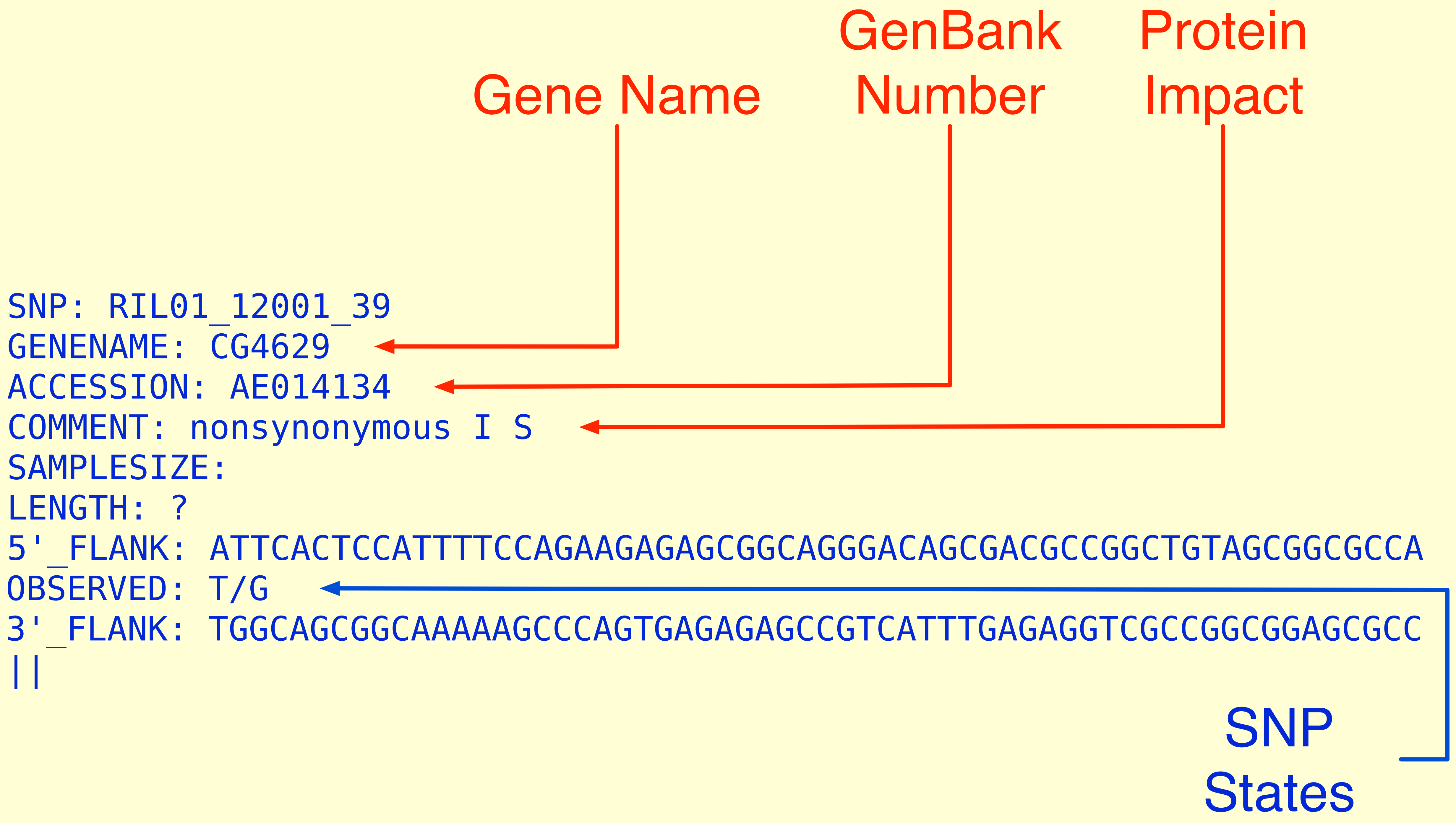
Sequence

SNP Contextual Sequence

C / G

Arg / Gly

SNPMeta:
Output to dbSNP



Gene Name

GenBank Number

Protein Impact

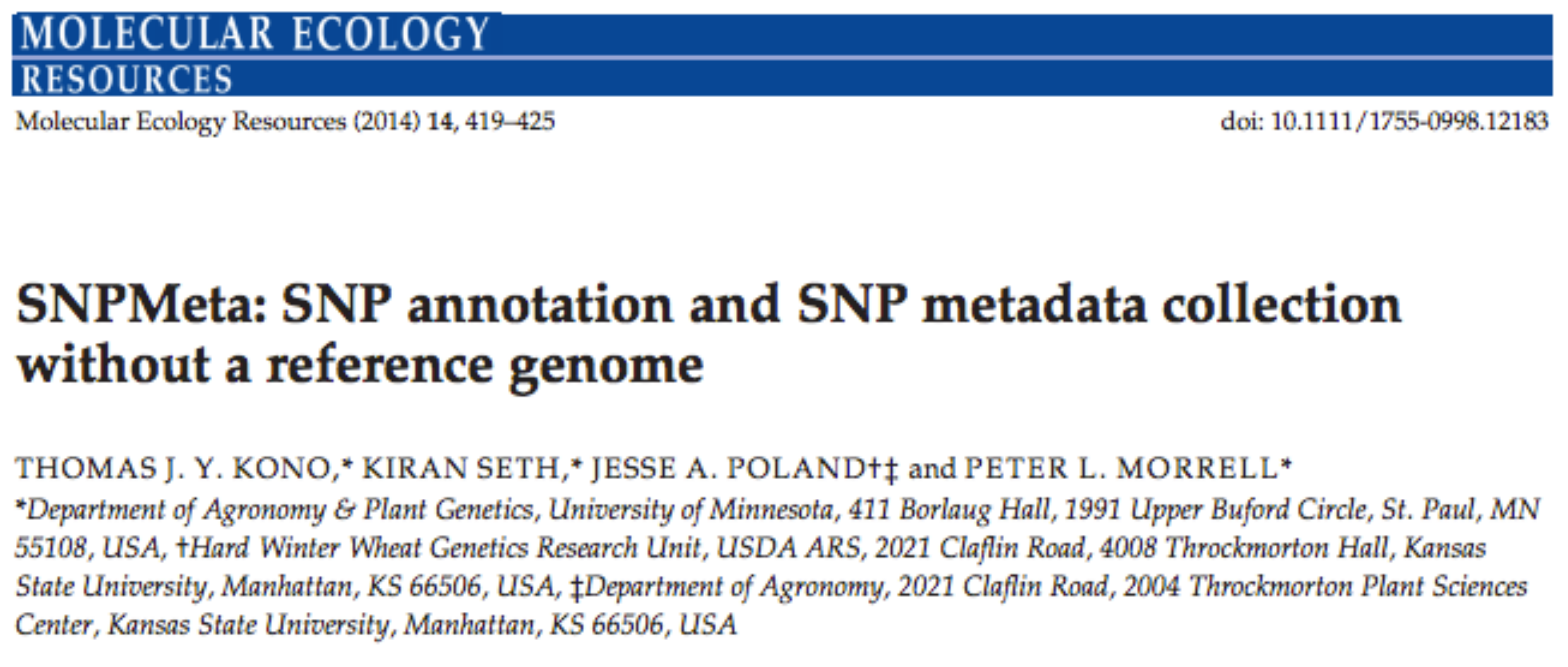
```
SNP: RIL01_12001_39
GENENAME: CG4629
ACCESSION: AE014134
COMMENT: nonsynonymous I S
SAMPLESIZE:
LENGTH: ?
5'_FLANK: ATTCACTCCATTTTCCAGAAGAGAGCGGCAGGGACAGCGACGCCGGCTGTAGCGGCGCCA
OBSERVED: T/G
3'_FLANK: TGGCAGCGGCAAAAAGCCCAGTGAGAGAGCCGTCATTTGAGAGGTCGCCGGCGGAGCGCC
||
```

SNP States

Kono et al. 2014

SNPMeta: Availability

Software available in GitHub
https://github.com/MorrellLAB/SNPMeta



MOLECULAR ECOLOGY
RESOURCES

Molecular Ecology Resources (2014) 14, 419–425 doi: 10.1111/1755-0998.12183

SNPMeta: SNP annotation and SNP metadata collection without a reference genome

THOMAS J. Y. KONO,* KIRAN SETH,* JESSE A. POLAND†‡ and PETER L. MORRELL*

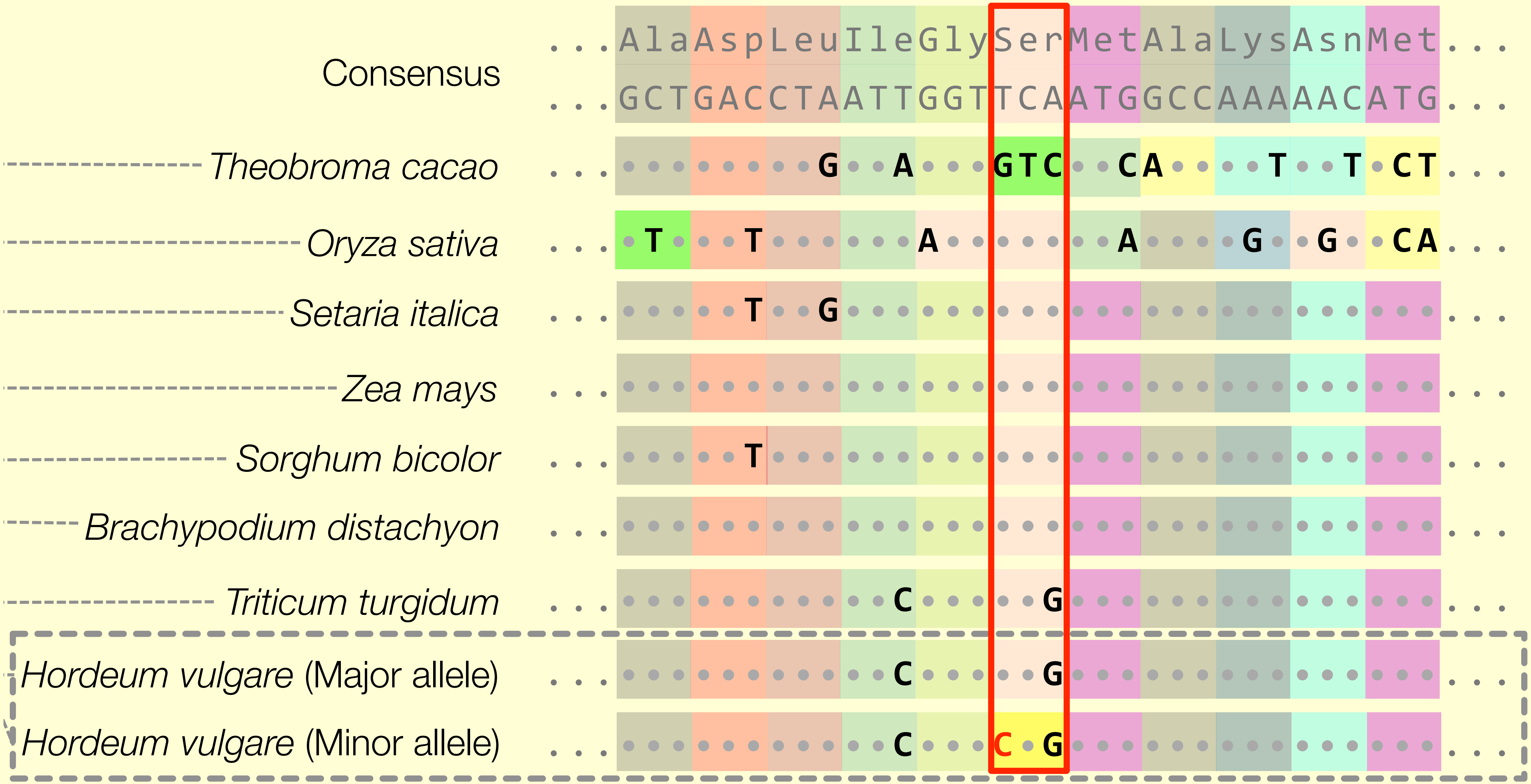
**Department of Agronomy & Plant Genetics, University of Minnesota, 411 Borlaug Hall, 1991 Upper Buford Circle, St. Paul, MN 55108, USA, †Hard Winter Wheat Genetics Research Unit, USDA ARS, 2021 Claflin Road, 4008 Throckmorton Hall, Kansas State University, Manhattan, KS 66506, USA, ‡Department of Agronomy, 2021 Claflin Road, 2004 Throckmorton Plant Sciences Center, Kansas State University, Manhattan, KS 66506, USA*

BAD_Mutations: Yet Another Deleterious Prediction Program

- Why make another deleterious variant prediction program?
  - SIFT, PolyPhen2, PROVEAN, MAPP, GERP++, etc...
- Ideal program:
  - Is species-agnostic
  - Uses consistent high-quality data for each gene
  - Is hypothesis-driven
  - Corrects for reference bias[1]

[1] 1: Simons et al. 2014

BAD_Mutations:
What a Deleterious Variant Looks Like



```
Consensus                     ...AlaAspLeuIleGlySerMetAlaLysAsnMet...
                              ...GCTGACCTAATTGGTTCAATGGCCAAAAACATG...
Theobroma cacao               ...••••••••G••A•••GTC••CA••••T••T•CT...
Oryza sativa                  ...•T•••T••••••A••••••••A••••G••G••CA...
Setaria italica               ...•••••T••G•••••••••••••••••••••••...
Zea mays                      ...•••••••••••••••••••••••••••••••••...
Sorghum bicolor               ...•••••T••••••••••••••••••••••••••...
Brachypodium distachyon       ...•••••••••••••••••••••••••••••••••...
Triticum turgidum             ...••••••••••••C••••G•••••••••••••••...
Hordeum vulgare (Major allele)...••••••••••••C••••G•••••••••••••••...
Hordeum vulgare (Minor allele)...••••••••••••C••••C•G•••••••••••••••...
```

BAD_Mutations: Hypothesis Test

Based on a likelihood ratio test (LRT) of sequence constraint from Chun and Fay (2009)

$$LLR = \log \frac{L(D|T, \theta, d_N = \hat{C} d_S)}{L(D|T, \theta, d_N = d_S)}$$

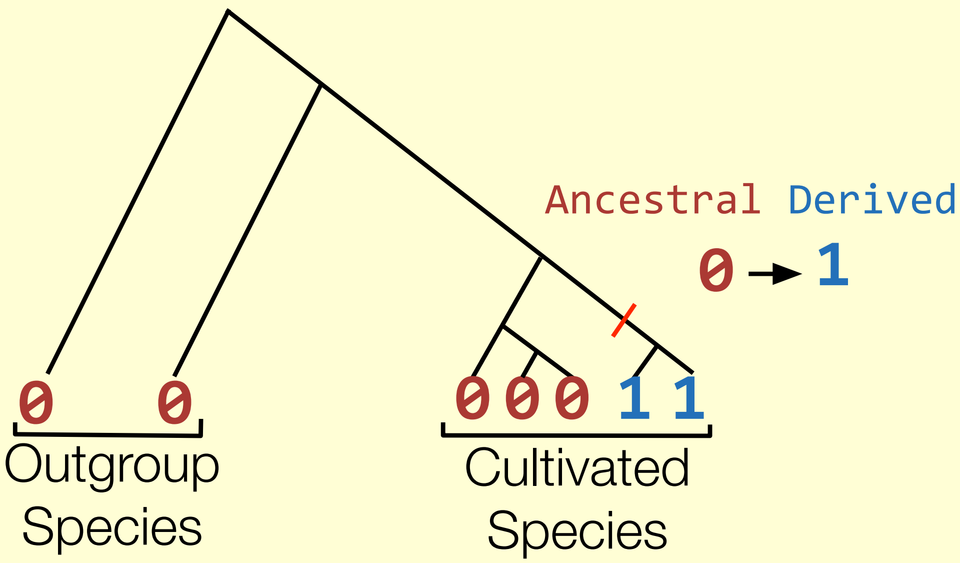
D = Codon alignment
T = Phylogeny
θ = Local substitution rate

BAD_Mutations: Addressing Reference Bias

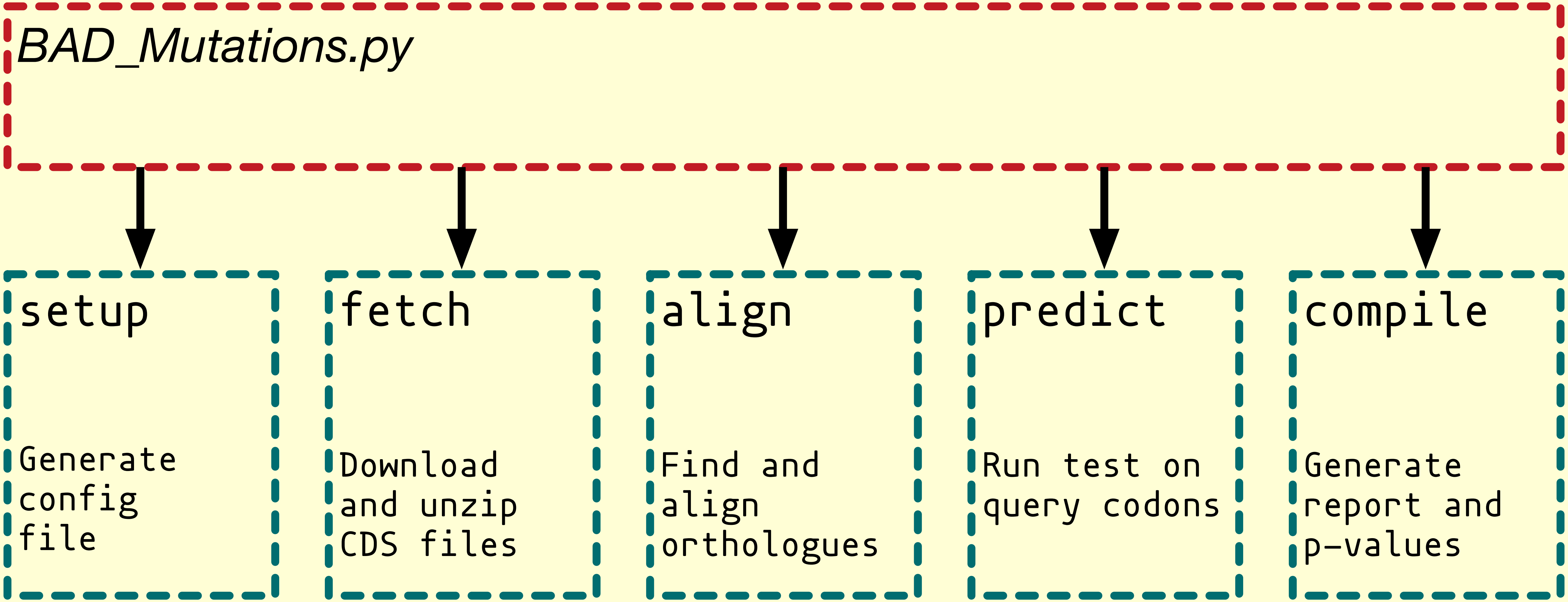
	Morex **(Reference)**	**Steptoe**	**Harrington**	**Kindred**
SNP 1	A	G	G	A
SNP 2	T	C	T	T
SNP 3	C	C	T	T
Diff. From Reference	**0**	**2**	**2**	**1**

BAD_Mutations:
Addressing Reference Bias

	Ancestral	Morex (Reference)	Steptoe	Harrington	Kindred
SNP 1	A	A	G	G	A
SNP 2	C	T	C	T	T
SNP 3	T	C	C	T	T
Diff. From Ancestral	**0**	**2**	**2**	**2**	**1**



BAD_Mutations: Modular Design



BAD_Mutations.py

`setup`

Generate config file

`fetch`

Download and unzip CDS files

`align`

Find and align orthologues

`predict`

Run test on query codons

`compile`

Generate report and p-values

Use config for controlling other subcommands

Download databases once, refresh when needed

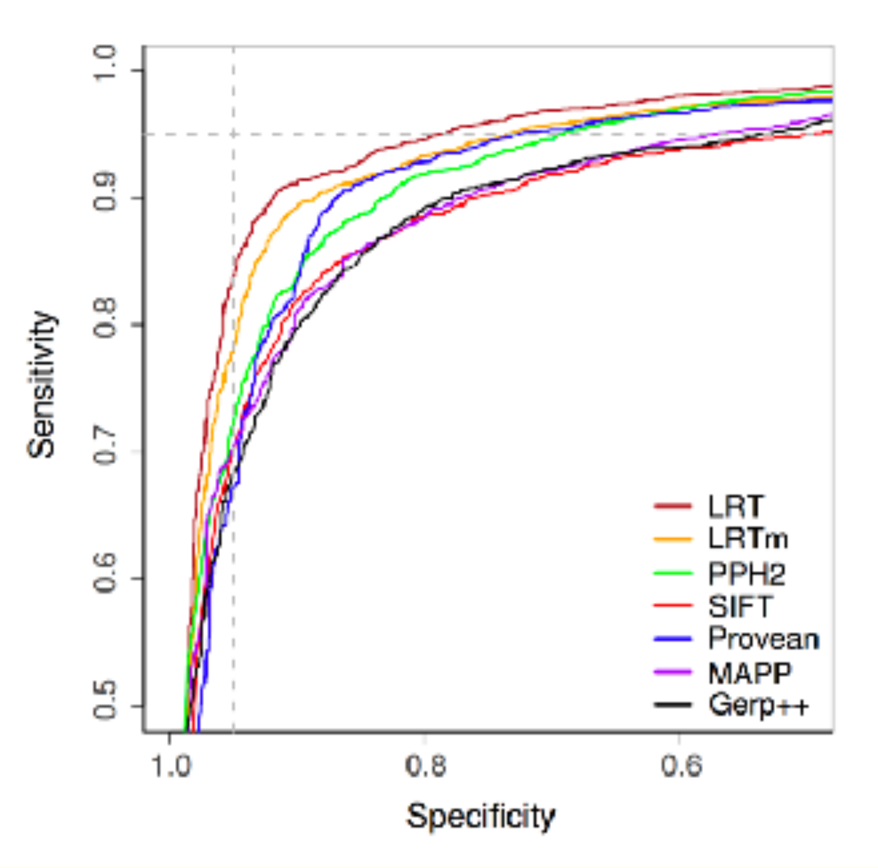
Parallelize on a gene-basis for a genome wide analysis

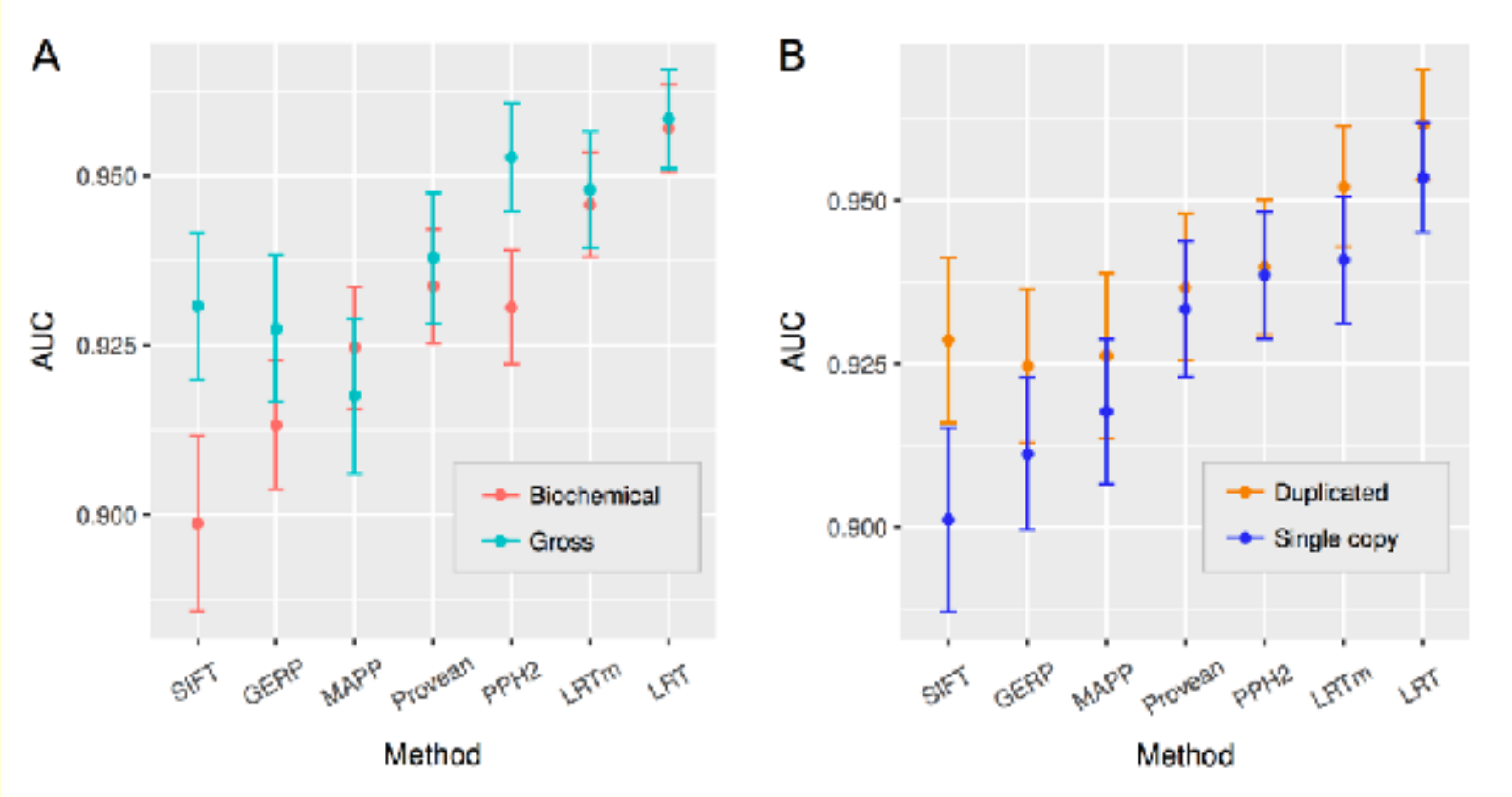
Parallelize on a gene-basis for a genome wide analysis. Slow.

Run locally, parses HYPHY text reports.

BAD_Mutations: High Accuracy

Used a curated set of 2,910 SNPs in *A. thaliana* with mutant phenotypic effects and 1,583 SNPs at high frequency (neutral)

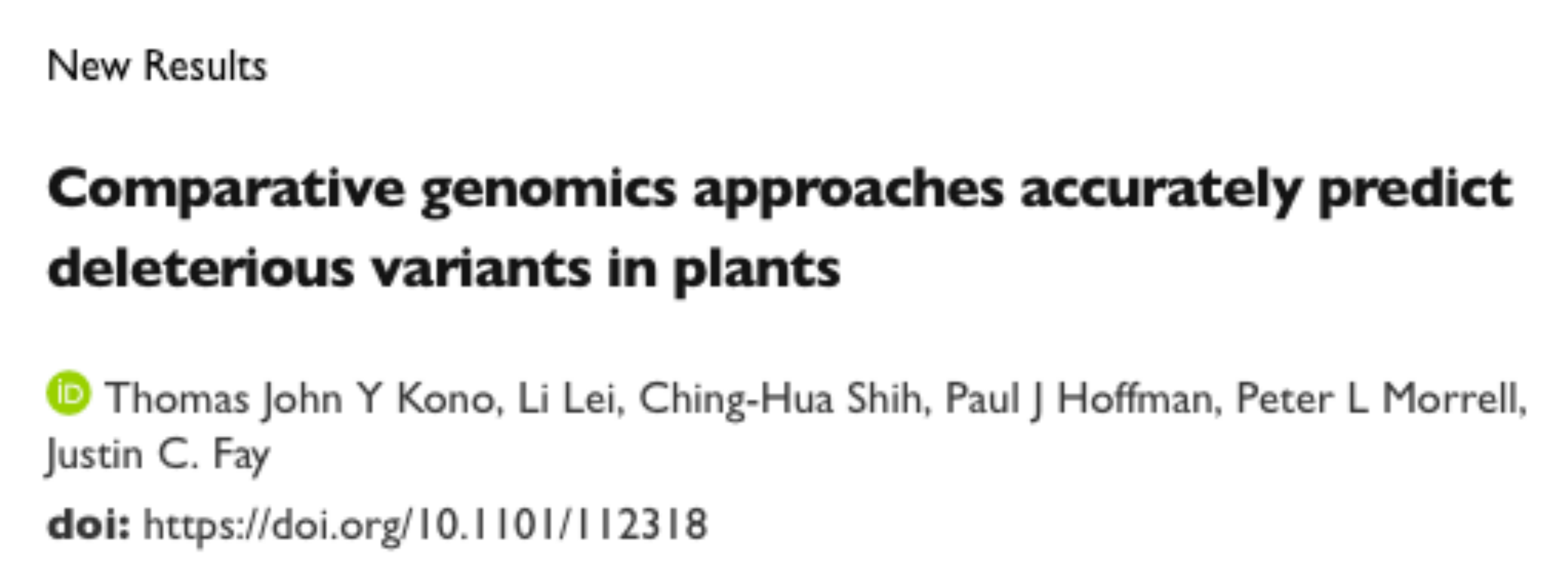




Kono et al. 2017 Preprint

BAD_Mutations: Availability

- Manuscript in BioRxiv



New Results

Comparative genomics approaches accurately predict deleterious variants in plants

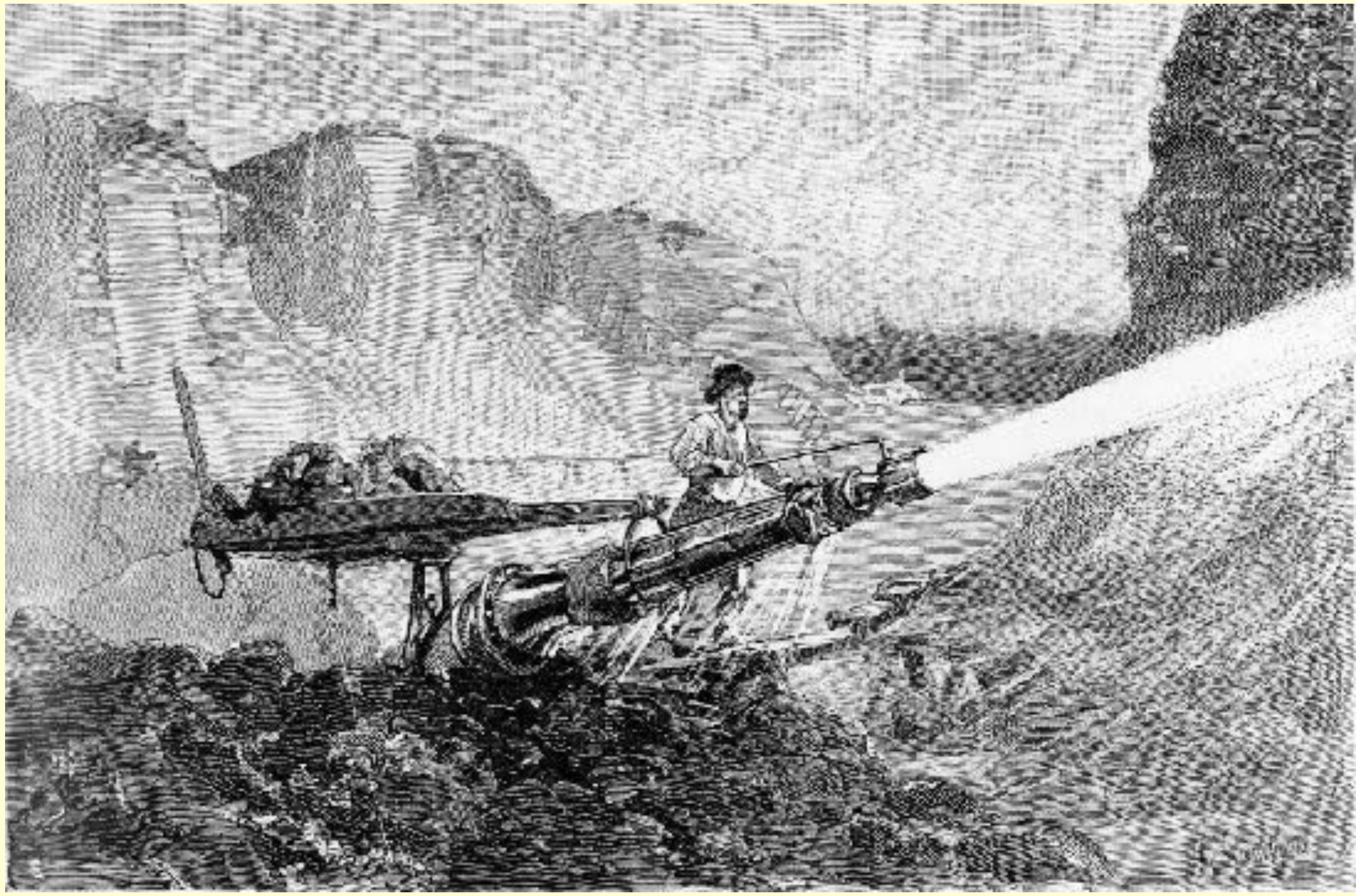
Thomas John Y Kono, Li Lei, Ching-Hua Shih, Paul J Hoffman, Peter L Morrell, Justin C. Fay

doi: https://doi.org/10.1101/112318

Software available in GitHub

- https://github.com/MorrellLAB/BAD_Mutations
- Manual: https://github.com/MorrellLAB/BAD_Mutations/blob/master/Manual/Manual_v1.0.md

Deleterious Variants:
A "Gold Rush"



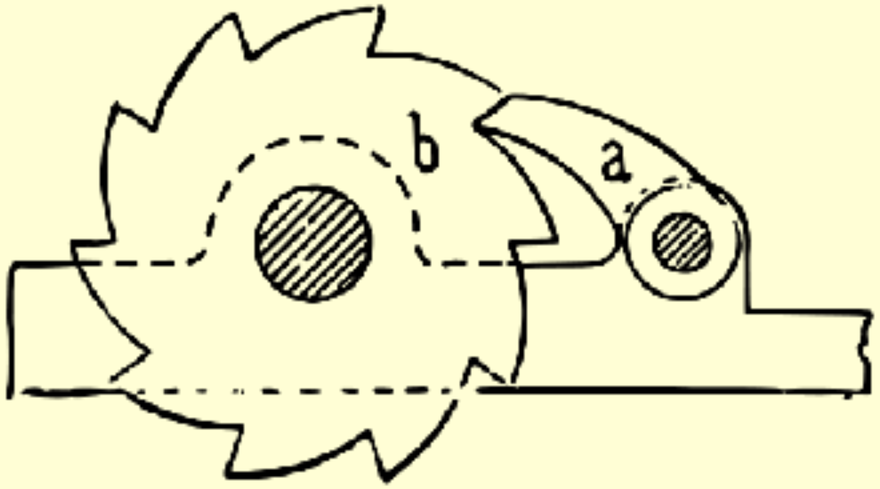
Lewontin 1974
Rockman 2012

Deleterious Variants: Theoretical Basis

- Purging effects - Finite N_e limits the effectiveness of purifying selection[1]

- "Muller's Ratchet"[2] - Deleterious mutations fix in low recombination regions

- Linked selection effects - genetic hitchhiking[3]



1:Takebayashi and Morrell 2001
2: Muller 1964
3: Hill and Robertson 1966

Deleterious Variants: Questions

- How many putatively deleterious SNPs segregate in two crop species?

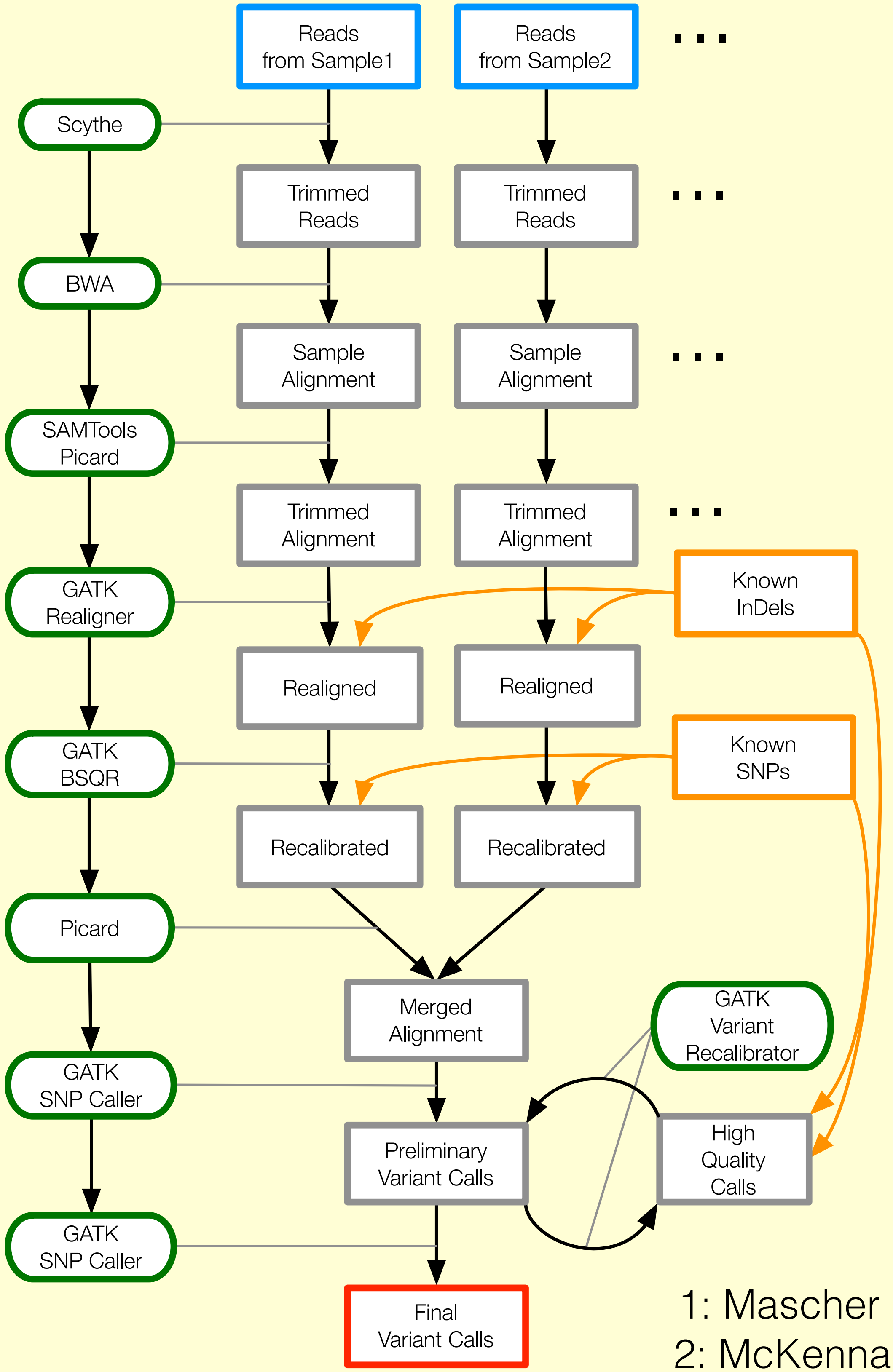
- Are SNPs that are causative for a phenotypic variant more likely to be annotated as deleterious, as compared to those without known phenotypic effects?

Deleterious Variants: Resequencing Data Analysis Pipeline

- Exome resequencing[1] of 15 barley accessions

- Whole genome resequencing of 8 soybean accessions

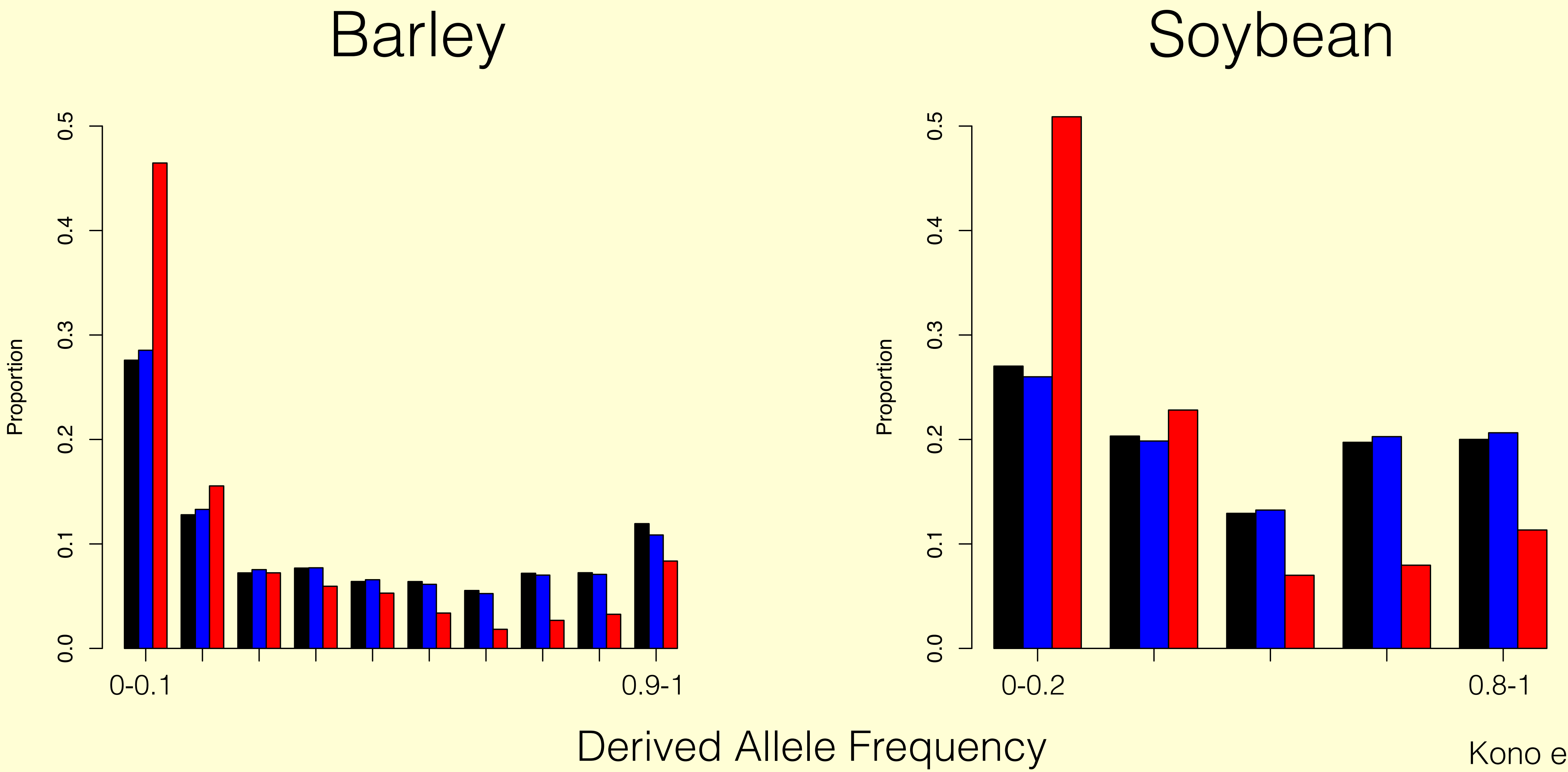
- Workflow based on GATK best practices[2,3]



1: Mascher et al. 2013
2: McKenna et al 2010
3: DePristo et al. 2011

Deleterious Variants: Enrichment for Low Frequency

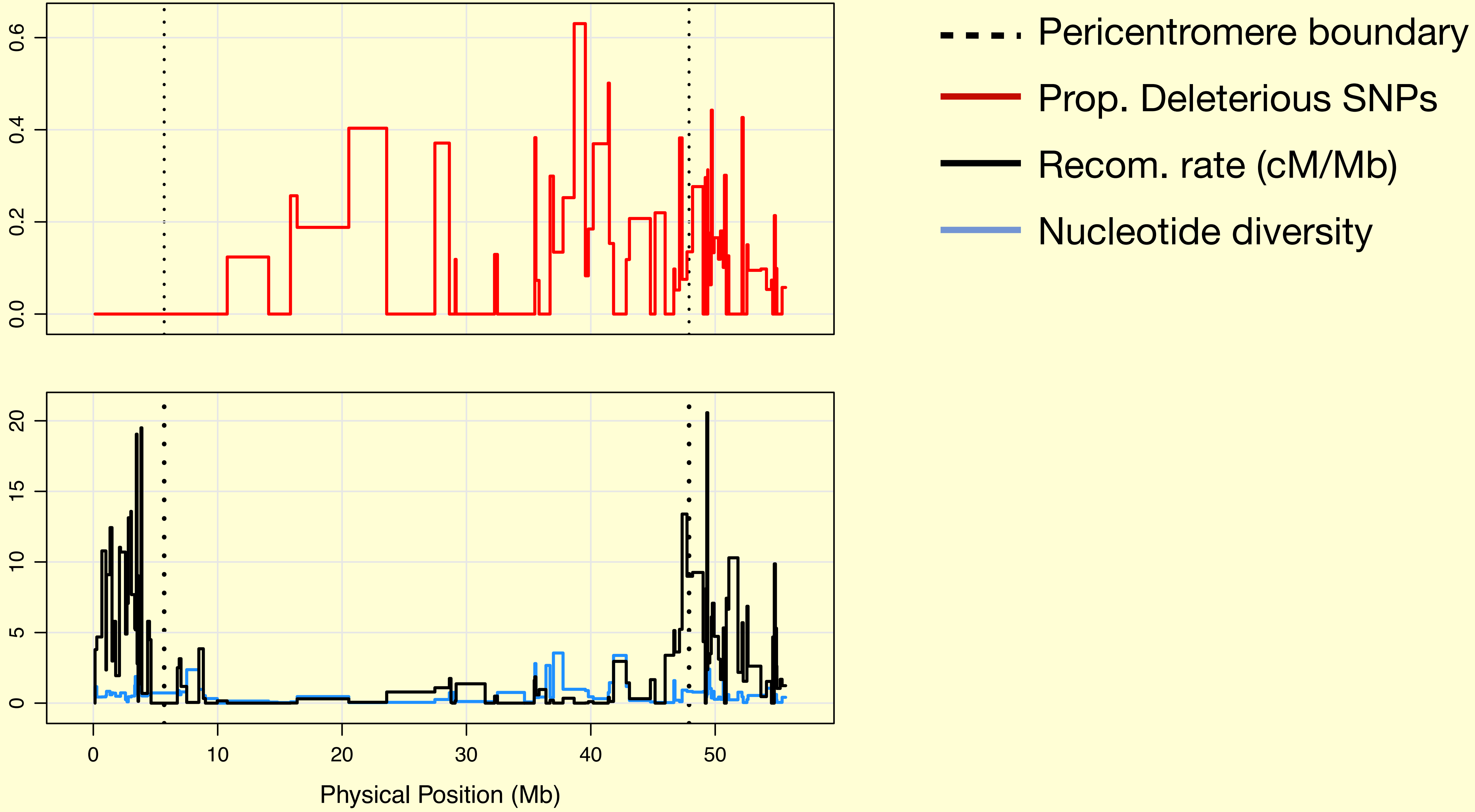
- Deleterious variants tend to be at lower derived frequency than “tolerated” variants



Kono et al. 2016

Deleterious Variants: Enrichment in Low Recombination Regions

- Tend to occur in low recombination regions



Kono et al. 2016

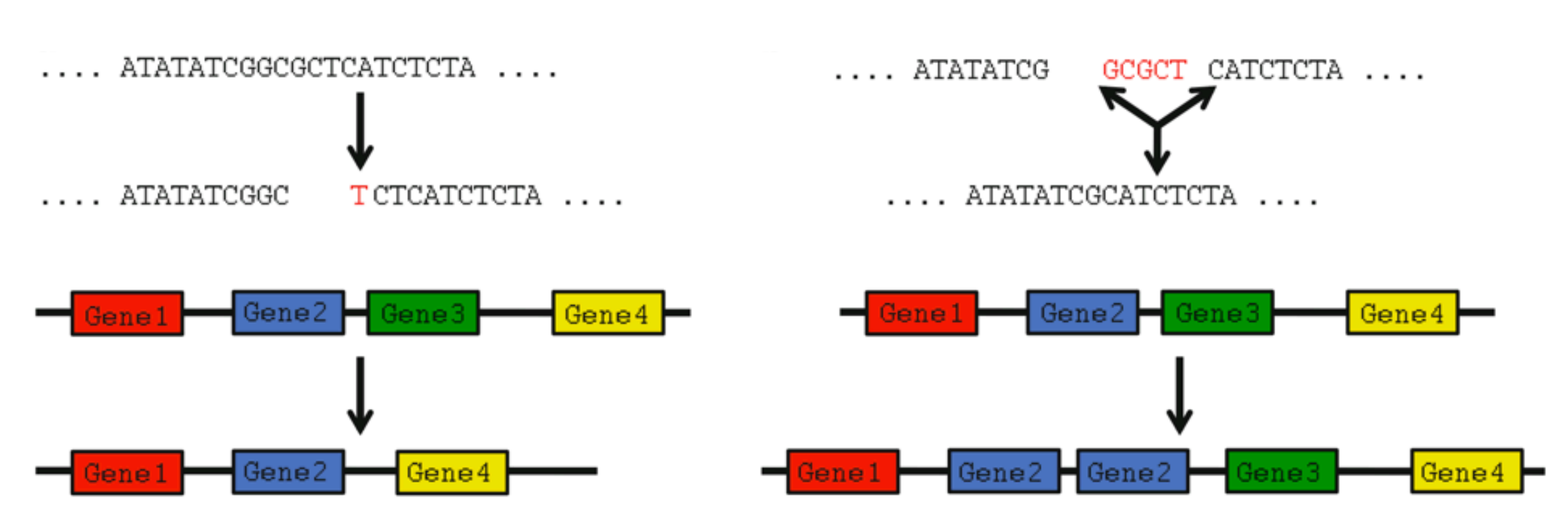
Deleterious Variants: Causative SNPs are "Deleterious"

- Causative variants tend to be called deleterious more frequently than those without *a priori* known phenotypic impacts

	Tolerated	Deleterious	Total
Causative	23 (**67.6%**)	11 (**32.4%**)	34
No Known Phenotype	29,259 (**94.2%**)	1,790 (**5.8%**)	31,046

Kono et al. 2016

A Different Mutation Type: Genome Content Variation


.... ATATATCGGCGCTCATCTCTA ....
.... ATATATCGGC T CTCATCTCTA ....
Gene1
Gene2
Gene3
Gene4
Gene1
Gene2
Gene4
.... ATATATCG GCGCT CATCTCTA ....
.... ATATATCGCATCTCTA ....
Gene1
Gene2
Gene3
Gene4
Gene1
Gene2
Gene2
Gene3
Gene4

Hirsch 2014

Tandem Duplicate Evolution: Background

- Tandem duplicates may be particularly compelling because they are segmental duplications of genes or gene fragments

- May have interesting evolutionary outcomes because they are duplications in the same genomic "neighborhood."

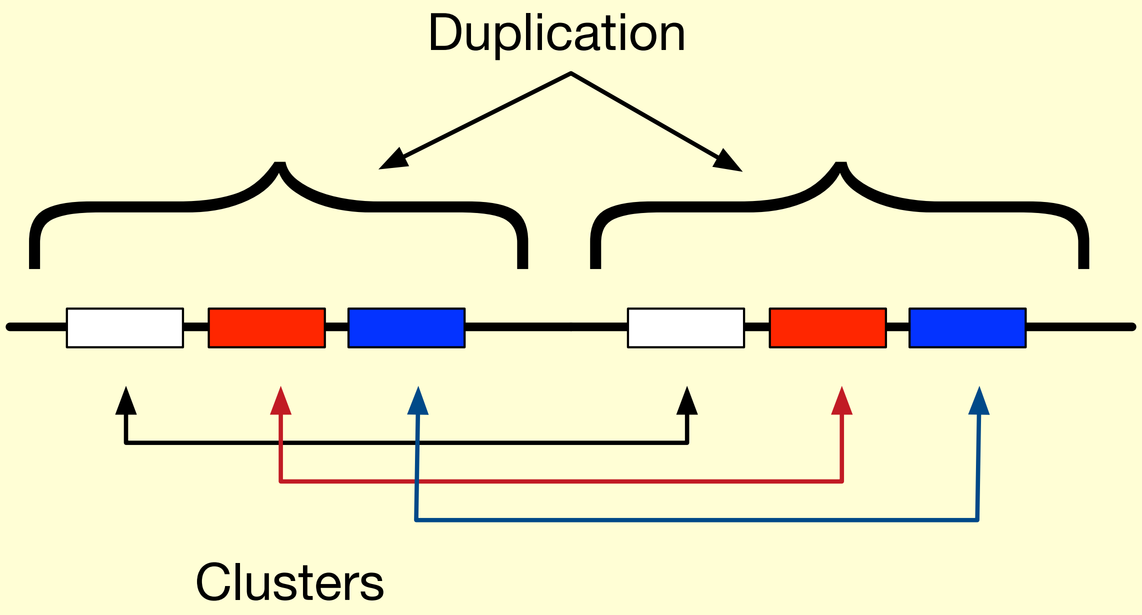
- New long read assembly in maize allows for better resolution of tandem duplicate sequences

Tandem Duplicate Evolution: Questions

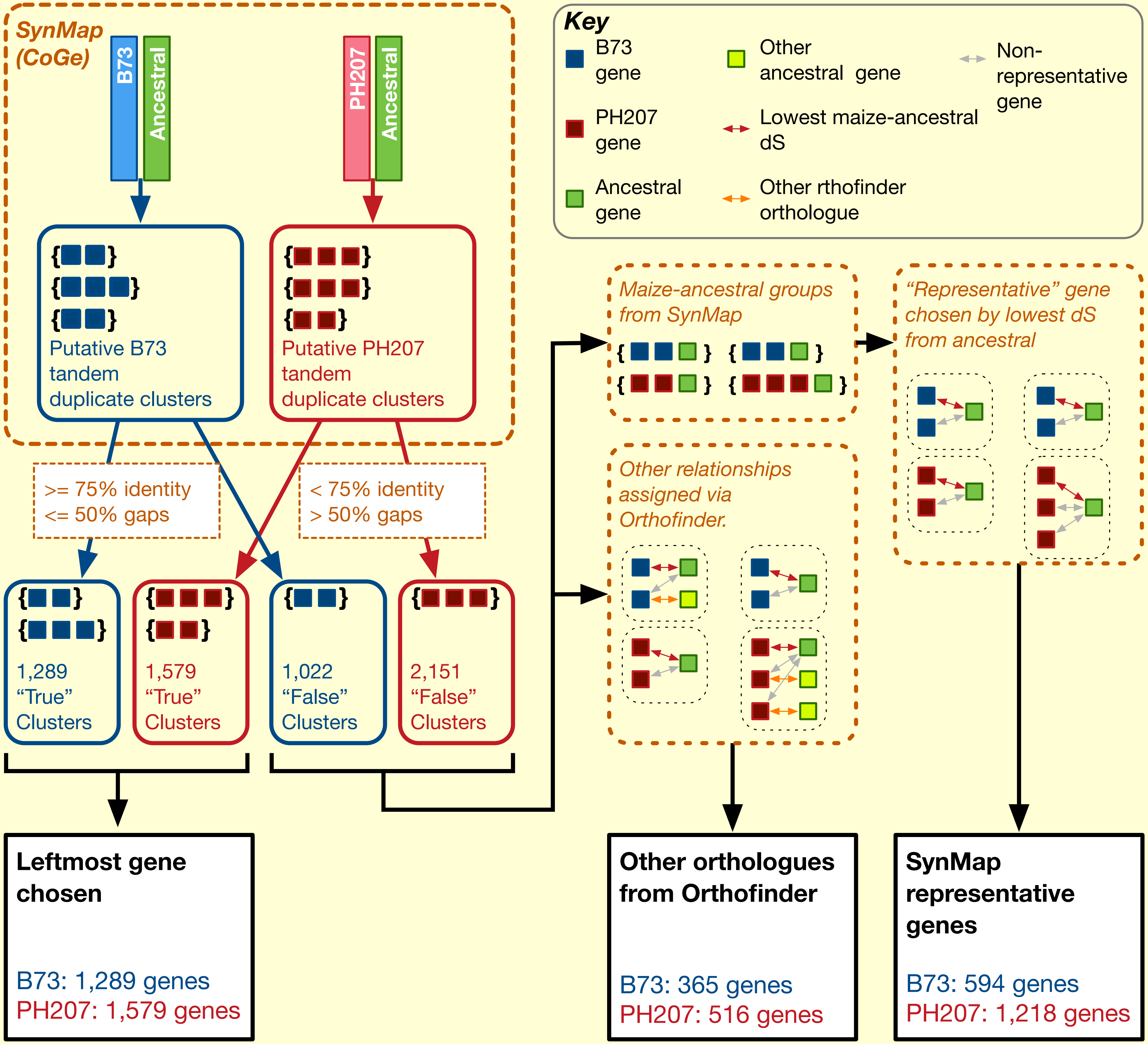
- Where are tandem duplicate genes in maize? Is there a genomic feature that explains their distribution?

- How old are tandem duplications in maize? Do they arise continuously, or happen in “bursts?”

- What are the estimated functional outcomes of tandem duplicate genes in maize? Is there a relationship between the age of a duplication and its outcome?



Tandem Duplicate Evolution: Identifying Tandem Duplicates



SynMap (CoGe)

Putative B73 tandem duplicate clusters

Putative PH207 tandem duplicate clusters

>= 75% identity
<= 50% gaps

< 75% identity
> 50% gaps

1,289 "True" Clusters

1,579 "True" Clusters

1,022 "False" Clusters

2,151 "False" Clusters

Key

- 🟦 B73 gene
- 🟥 PH207 gene
- 🟩 Ancestral gene
- 🟨 Other ancestral gene
- ↔ Lowest maize-ancestral dS
- ↔ Other rthofinder orthologue
- ↔ Non-representative gene

Maize-ancestral groups from SynMap

"Representative" gene chosen by lowest dS from ancestral

Other relationships assigned via Orthofinder.

Leftmost gene chosen

B73: 1,289 genes
PH207: 1,579 genes

Other orthologues from Orthofinder

B73: 365 genes
PH207: 516 genes

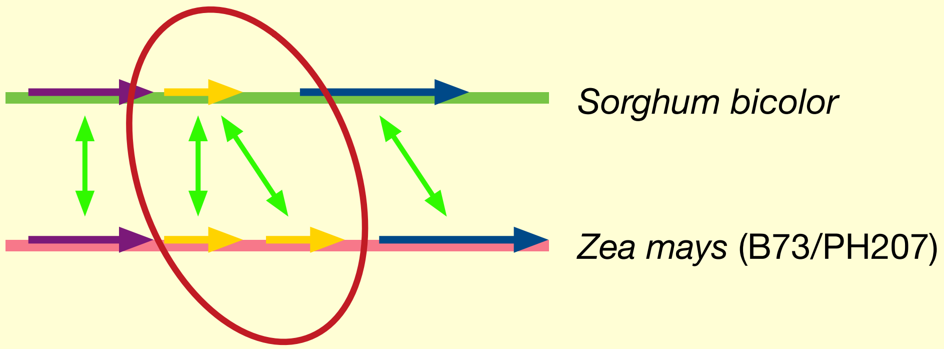
SynMap representative genes

B73: 594 genes
PH207: 1,218 genes

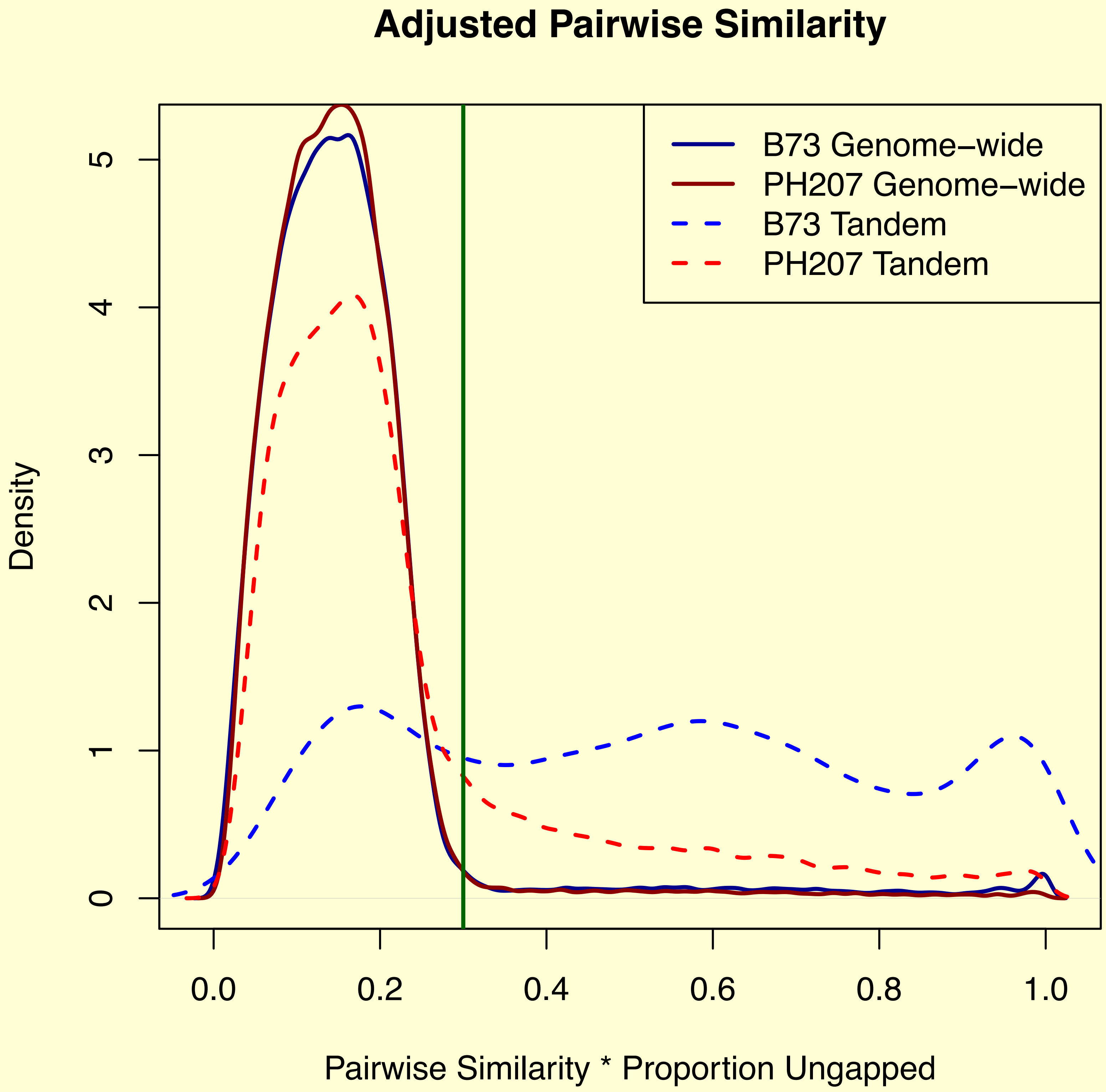
Brohammer et al. *In prep*

Tandem Duplicate Evolution:
Identifying Tandem Duplicates

- Had to filter tandem duplicates - CoGe identification procedure was over-calling



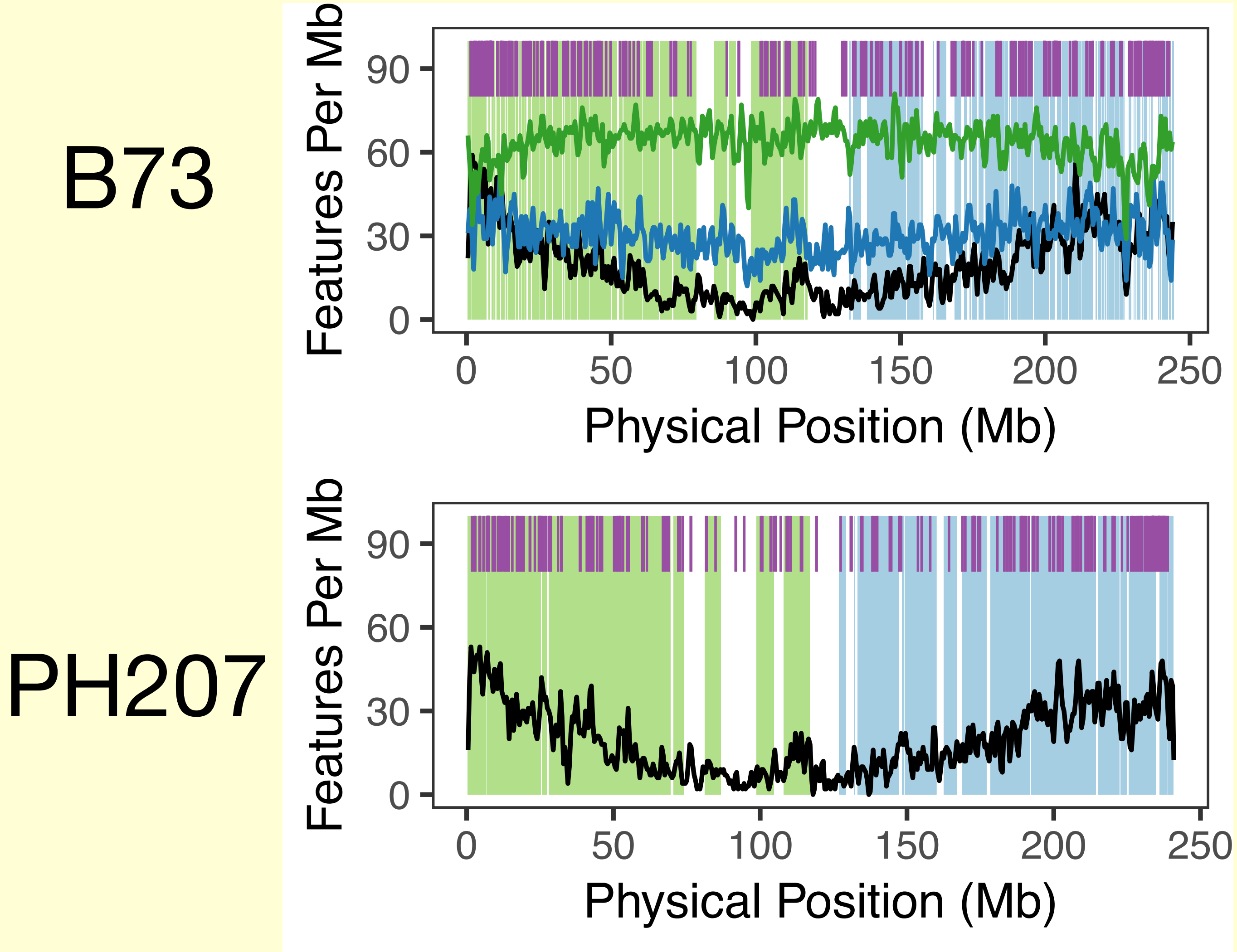
- "Adjusted pairwise similarity" = pairwise similarity, down-weighted for gapping

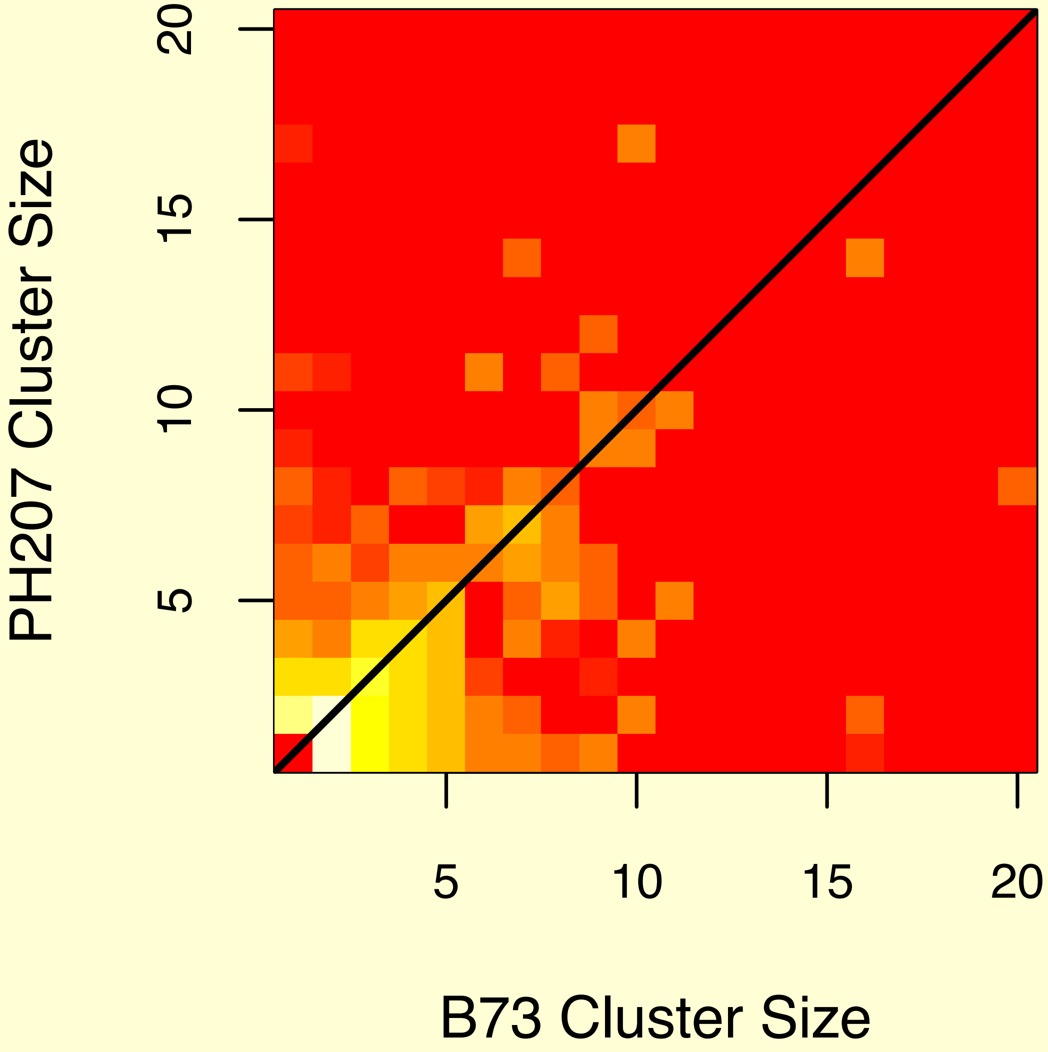


Tandem Duplicate Evolution: Most Duplicates are Shared

Tandem duplications happen where there are genes

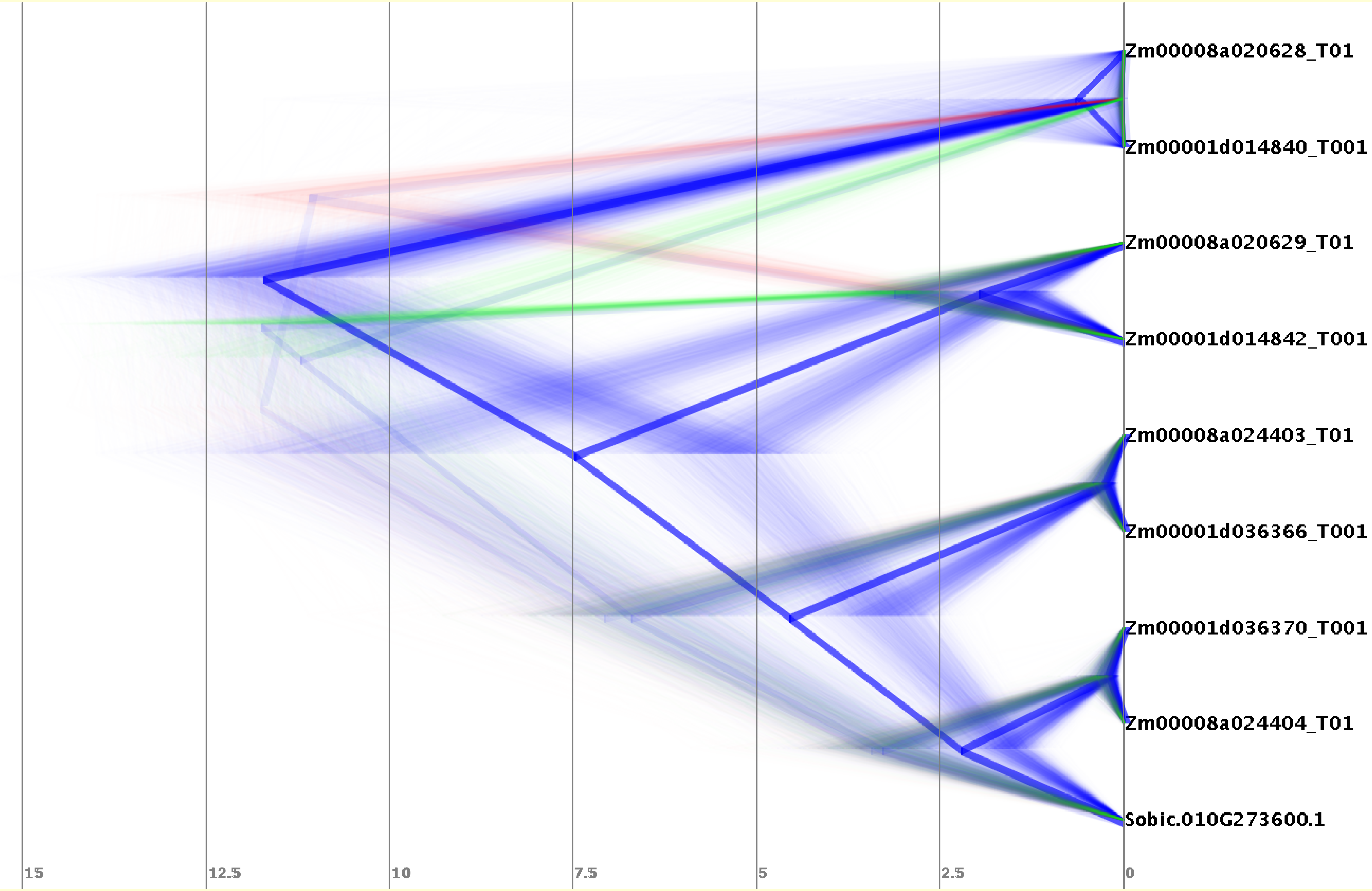
Most tandem duplicates are shared between B73 and PH207





Kono et al. *In prep*

Tandem Duplicate Evolution: Diverge Date Estimation



Genome Content Variation: Background

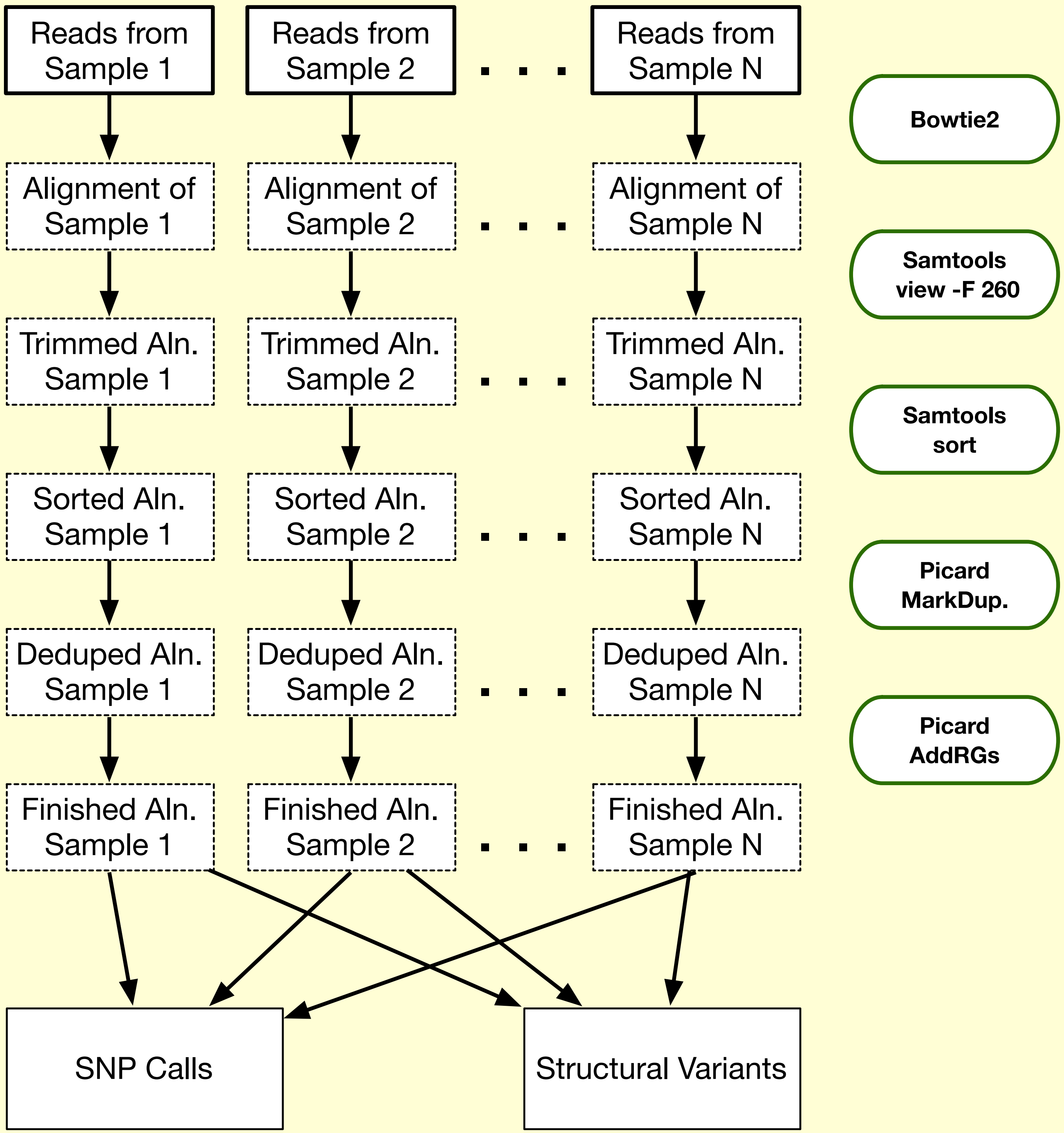
- Presence-absence variation and copy number variation may play a large role in phenotypic variation

- SNPs can only explain so much variation in GWAS, and most SNPs are thought to have neutral or very small effects on phenotypic variation

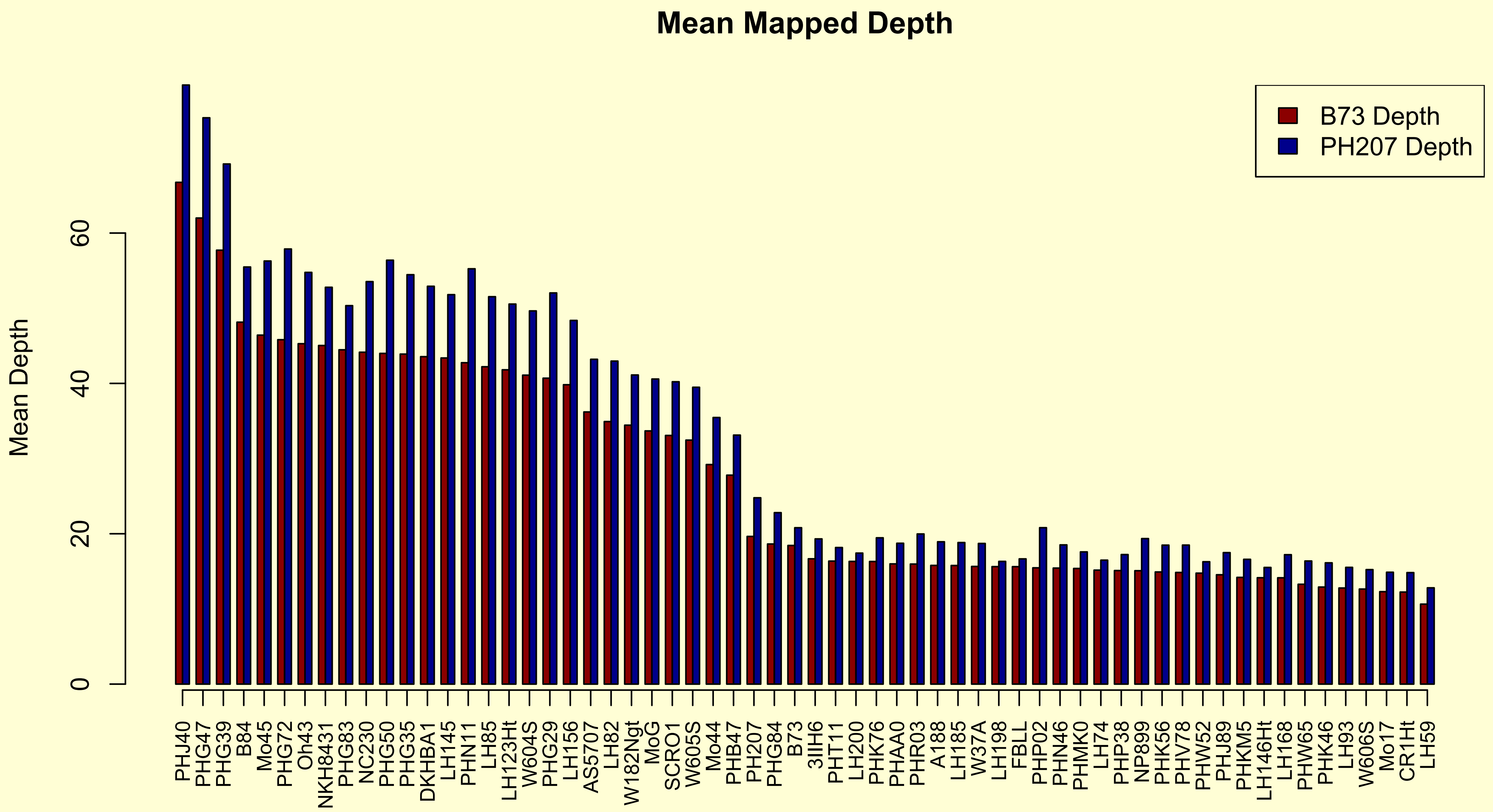
- Gene deletions/duplications, on the other hand, may have much greater impact, as they more drastically alter protein function and regulation

Genome Content Variation: Sequence Analysis Pipeline

- Whole genome resequencing of 62 maize inbred lines
- Currently have short read resequencing, but will collect long read data
- Mapped against two reference genomes

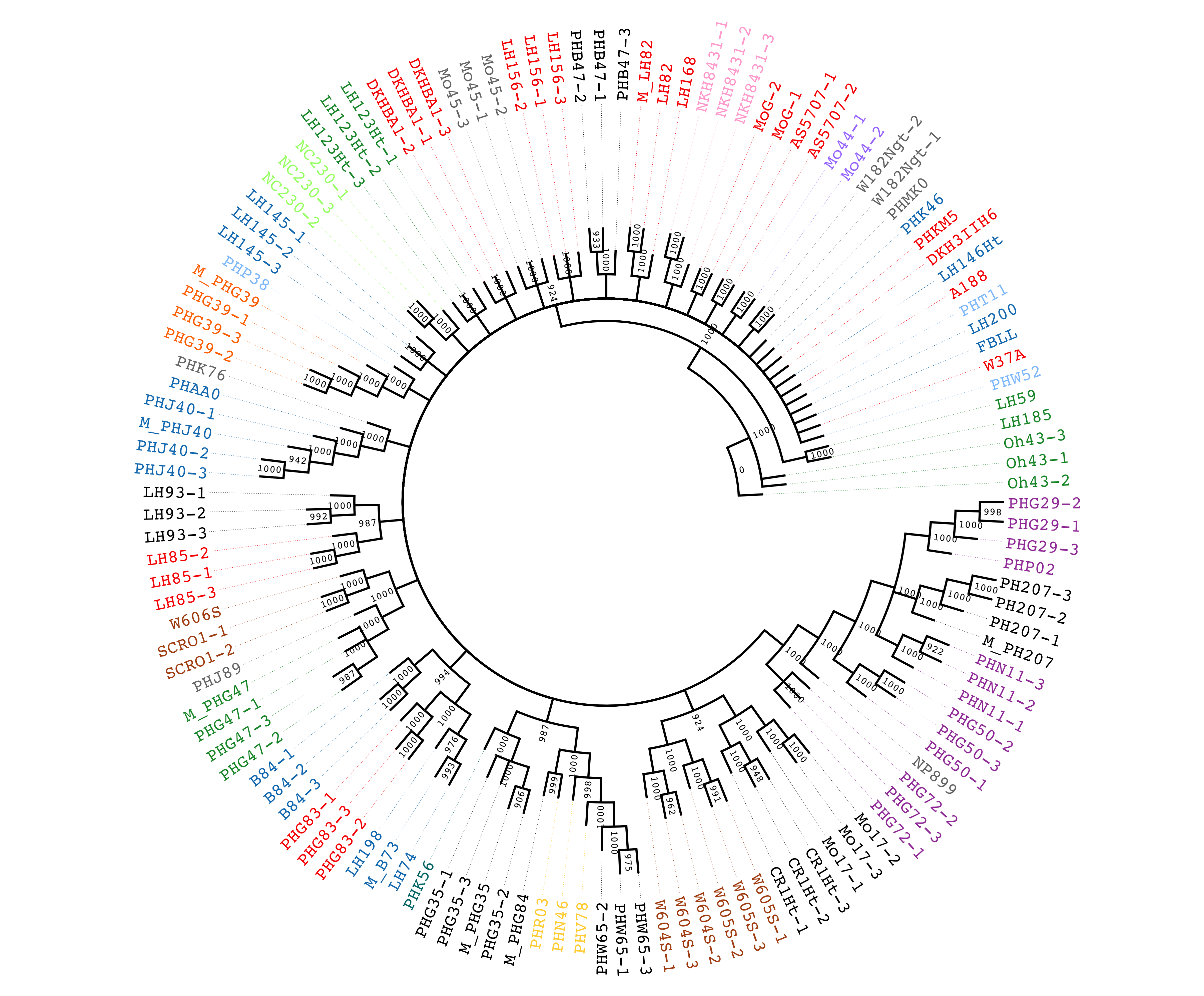


Genome Content Variation: Varying Depths of Sequencing



Genome Content Variation: A "Biological QC"

- Bootstrapped NJ tree for sample verification
- Samples from multiple libraries and data sources cluster by genotype



Genome Content Variation: Next Steps

- First: downsampling analysis with the high-coverage samples
    - Identify the nature of the tradeoff between sequencing depth and variant identification sensitivity
    - What depth gives a good balance between variant discovery and cost?
- Later: Call genome content variants, and estimate the frequencies of gene loss and duplication

Acknowledgements

Advisors:

Peter L. Morrell
Robert M. Stupar
Candice N. Hirsch
Suzanne E. McGaugh

Funding:

USDA NIFA National Needs Fellowship
MnDRIVE Global Food Ventures
UMN Doctoral Dissertation Fellowship
NSF Plant Genome Research Program

Collaborators:

Justin Fay
Michael Kantar
Mohsen Mohammadi
Jesse Poland
Amber Eule-Nashoba
Kiran Seth
Justin Anderson
Jean-Michel Michno
Alex Brohammer