Genotyping germline copy number variants in large-scale studies

Rob Scharpf

August 9, 2018

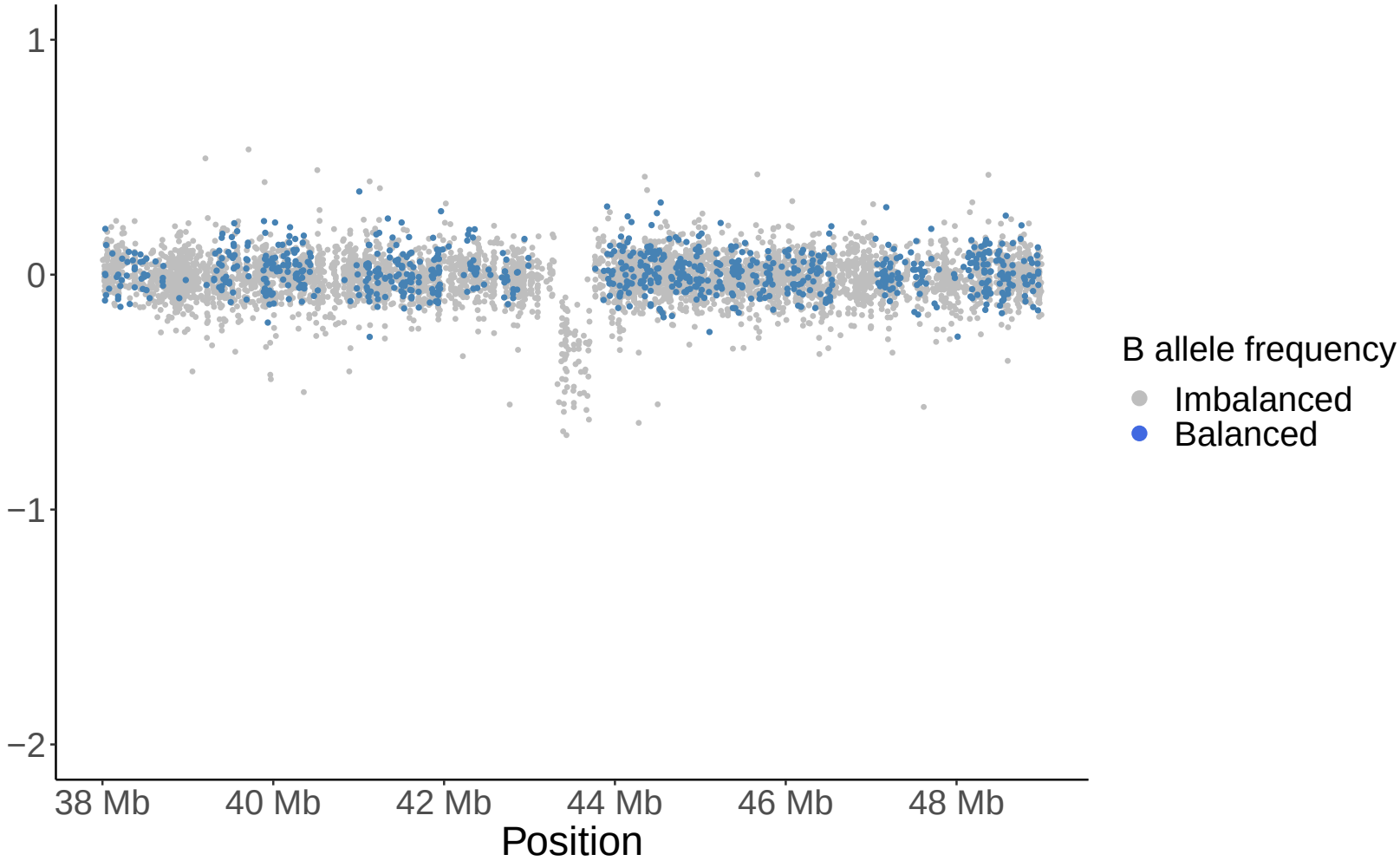
Goal

- To improve copy number calling at copy number polymorphic (CNP) regions in large-scale studies
- To extend these methods to trio-based study designs

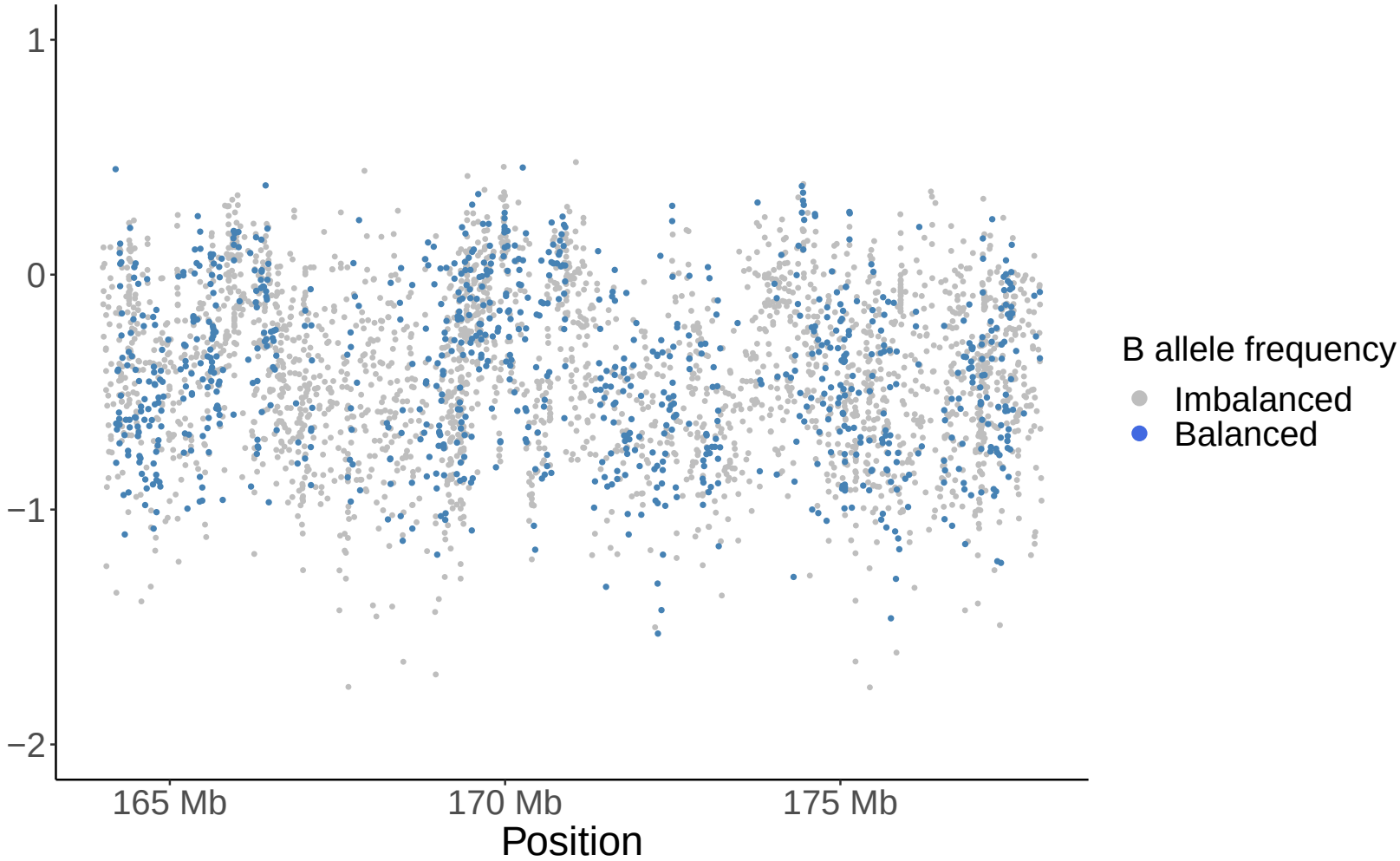
Pancreatic cancer case-control consortium

- $\approx$ 8000 participants genotyped on the Illumina Omni-Exome array
- Inherited variants in *ATM*, *BRCA2*, and *PALB2* known to increase risk
- 80% of familial clustering for this disease is unexplained

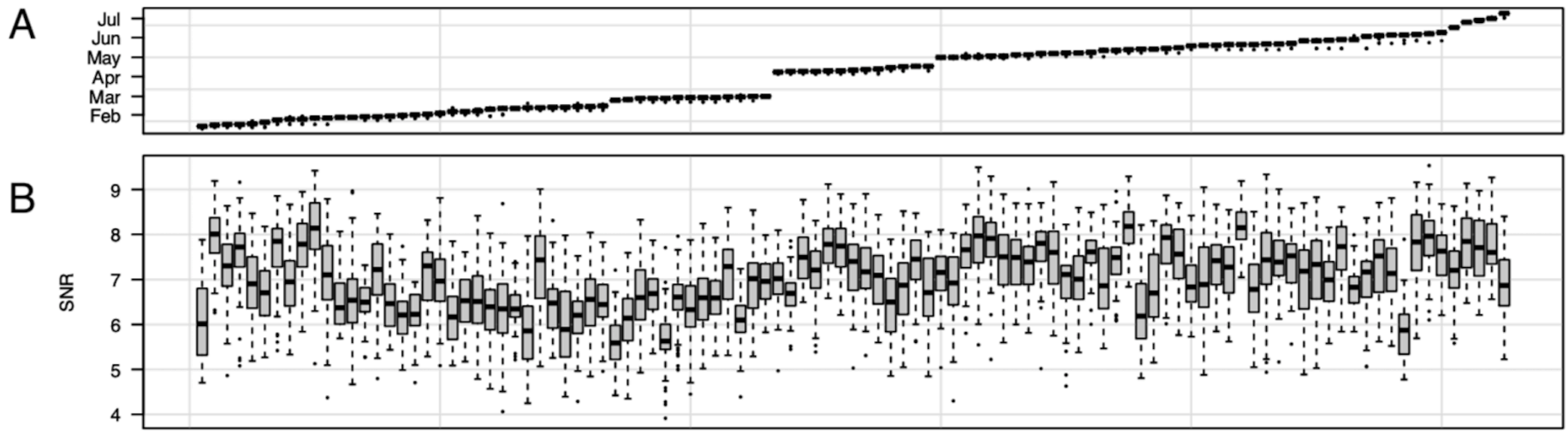
Arrays and capture-based sequencing (one genome)



Another sample

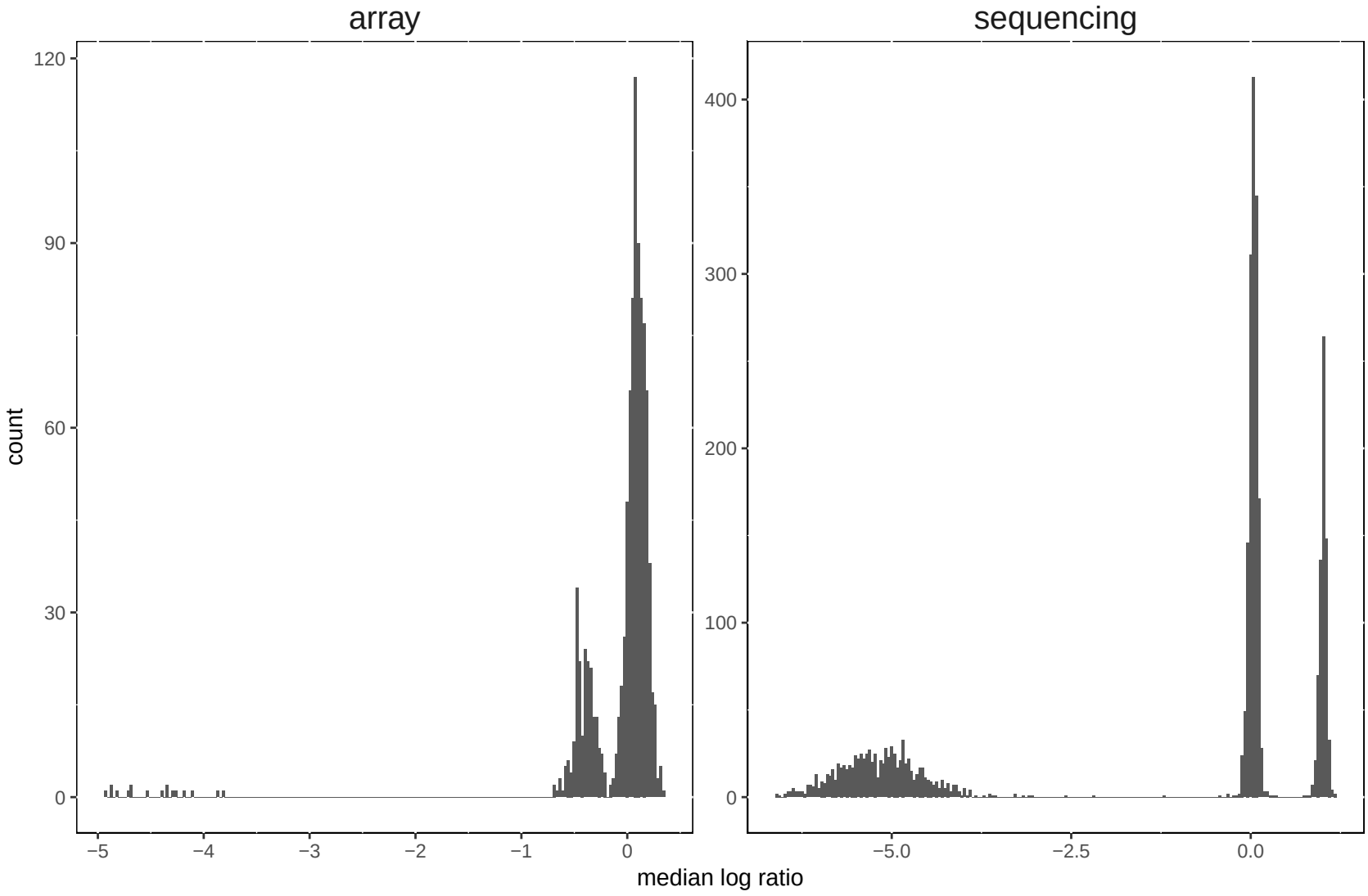


Dependency of data quality on batch



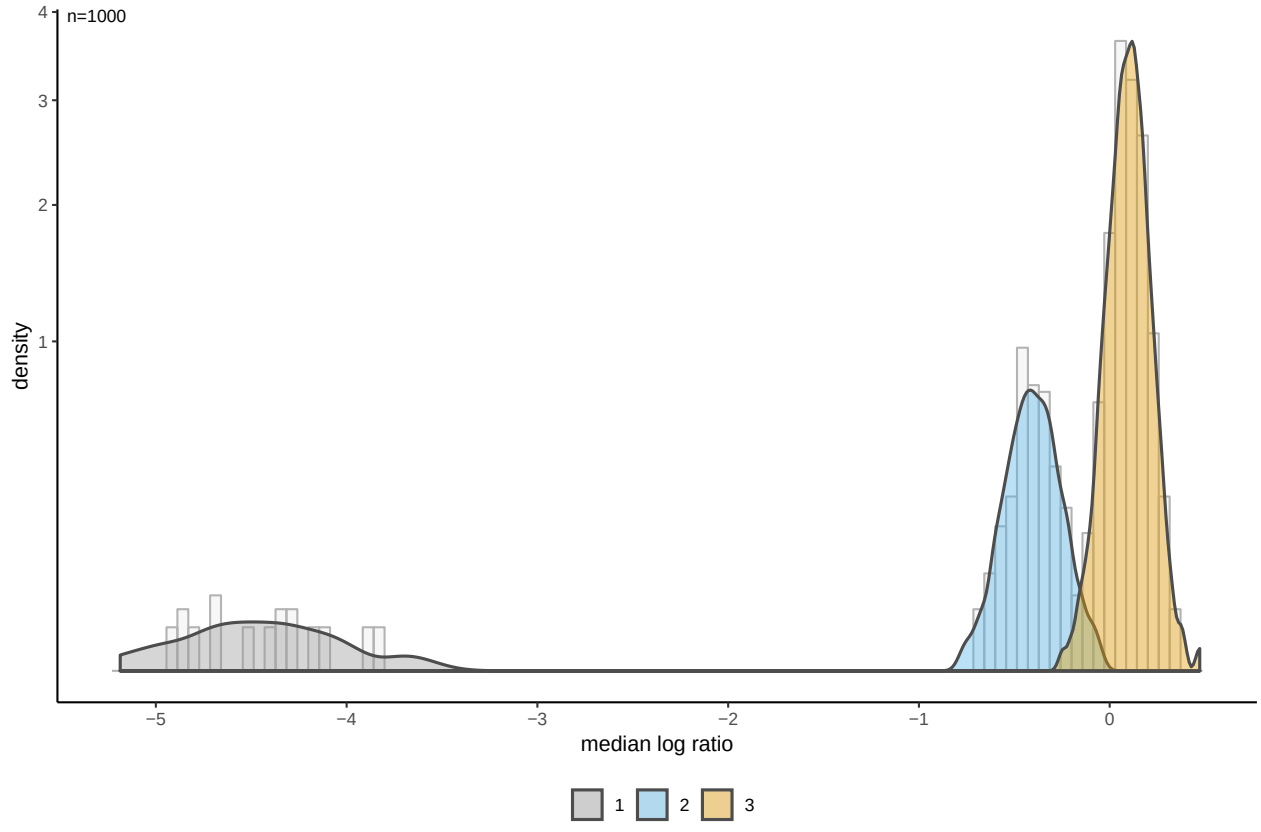
Li et al, 2014 (BMC Genomics)

Arrays and capture-based sequencing (one region)



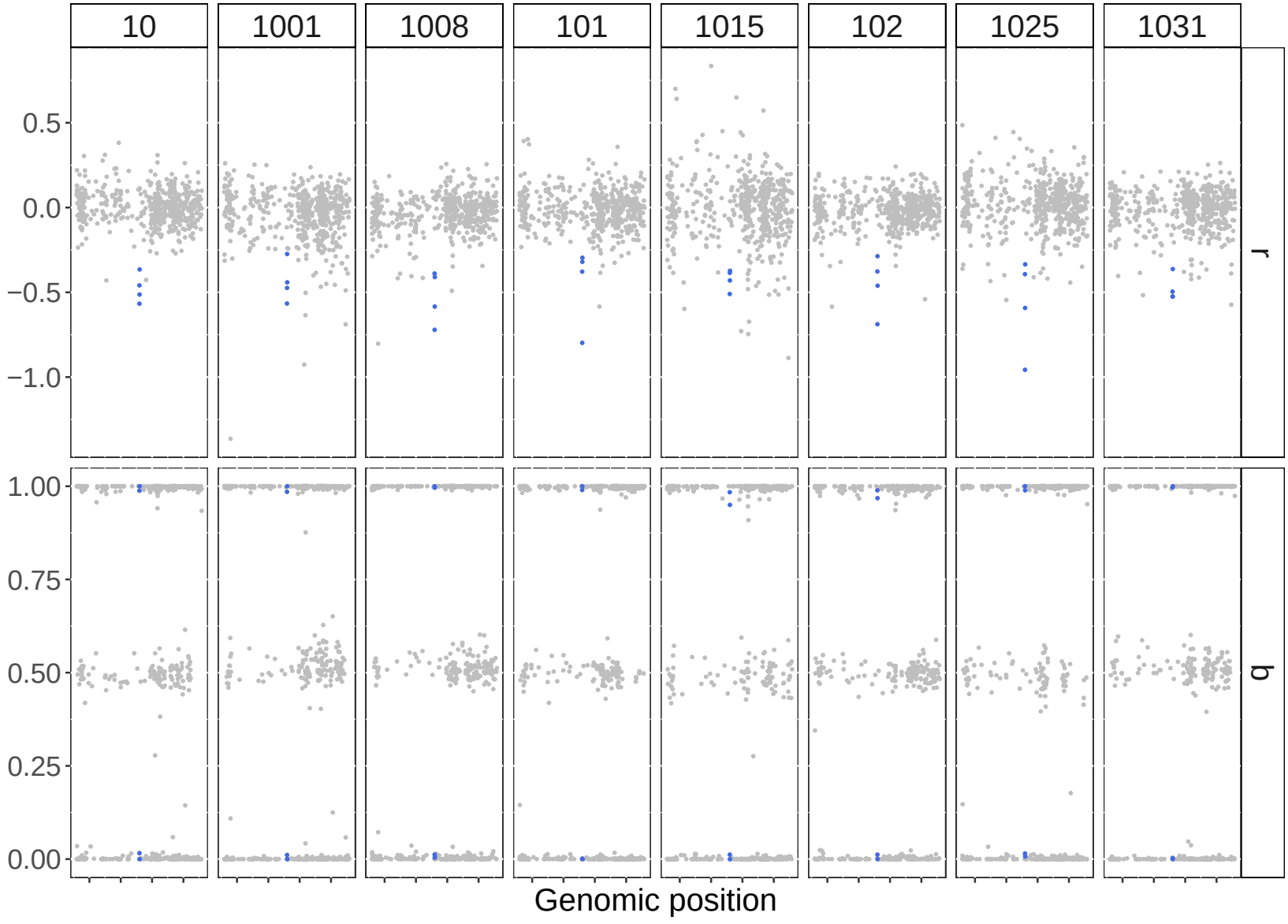
- different studies / different regions

A known CNP region with 4 SNPs



Arrays: Cardin et al., 2011 (Genetic Epidemiology)
Sequencing: XHMM, Conifer, CLAMMS, and others

A known CNP region with 4 SNPs



Recap

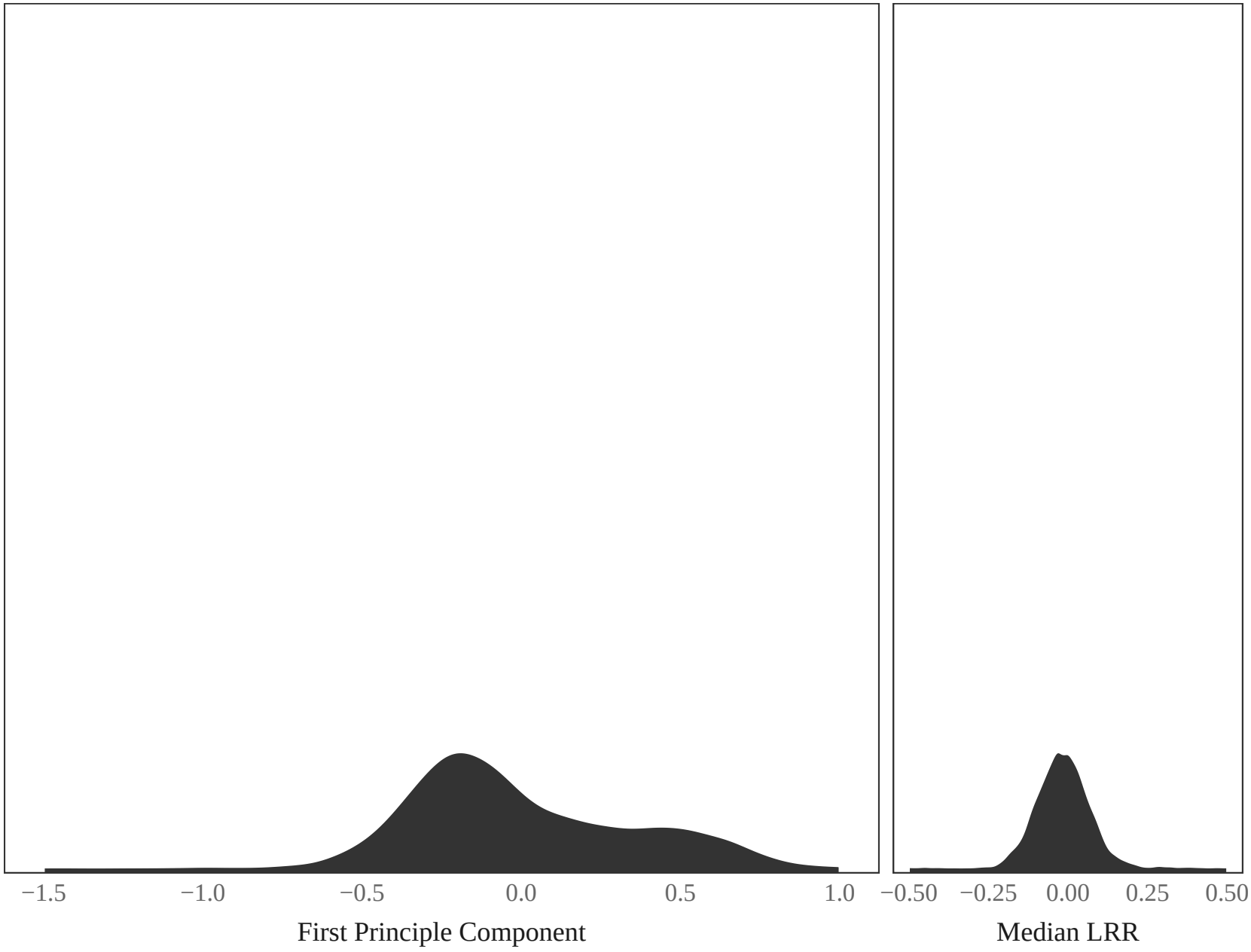
By genome:

- Bin-to-bin (or probe-to-probe) technical variation within a sample greatly limits resolution
  - GC content and other unmodeled sequence characteristics that influence PCR and measured abundances
- Latent factors that cause groups of sample to appear very different (batch effects) are completely ignored

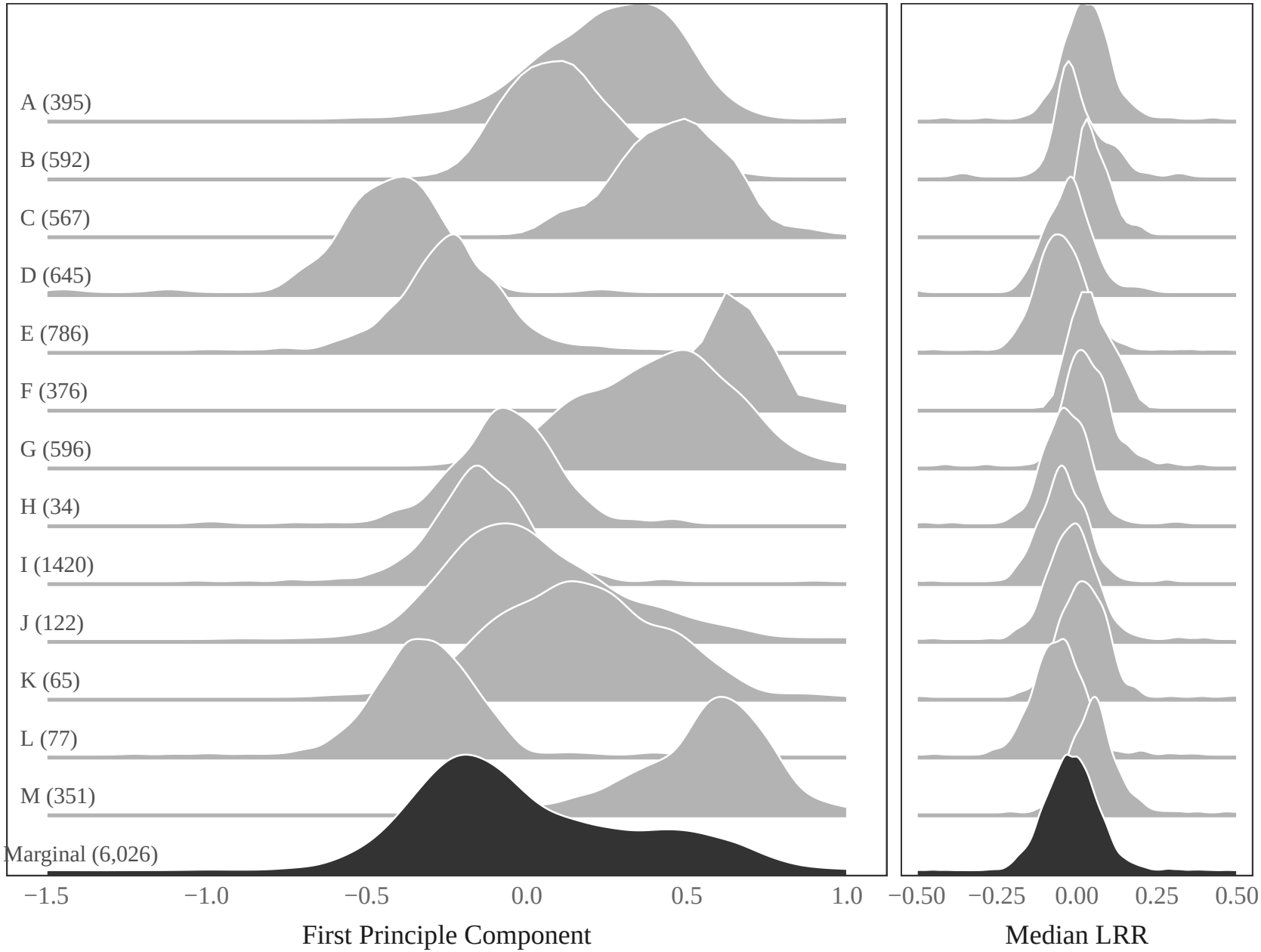
By region:

- Sequence-induced variation of abundances is less critical
- Batch effects can be estimated and modeled
- Not great for rare CNVs

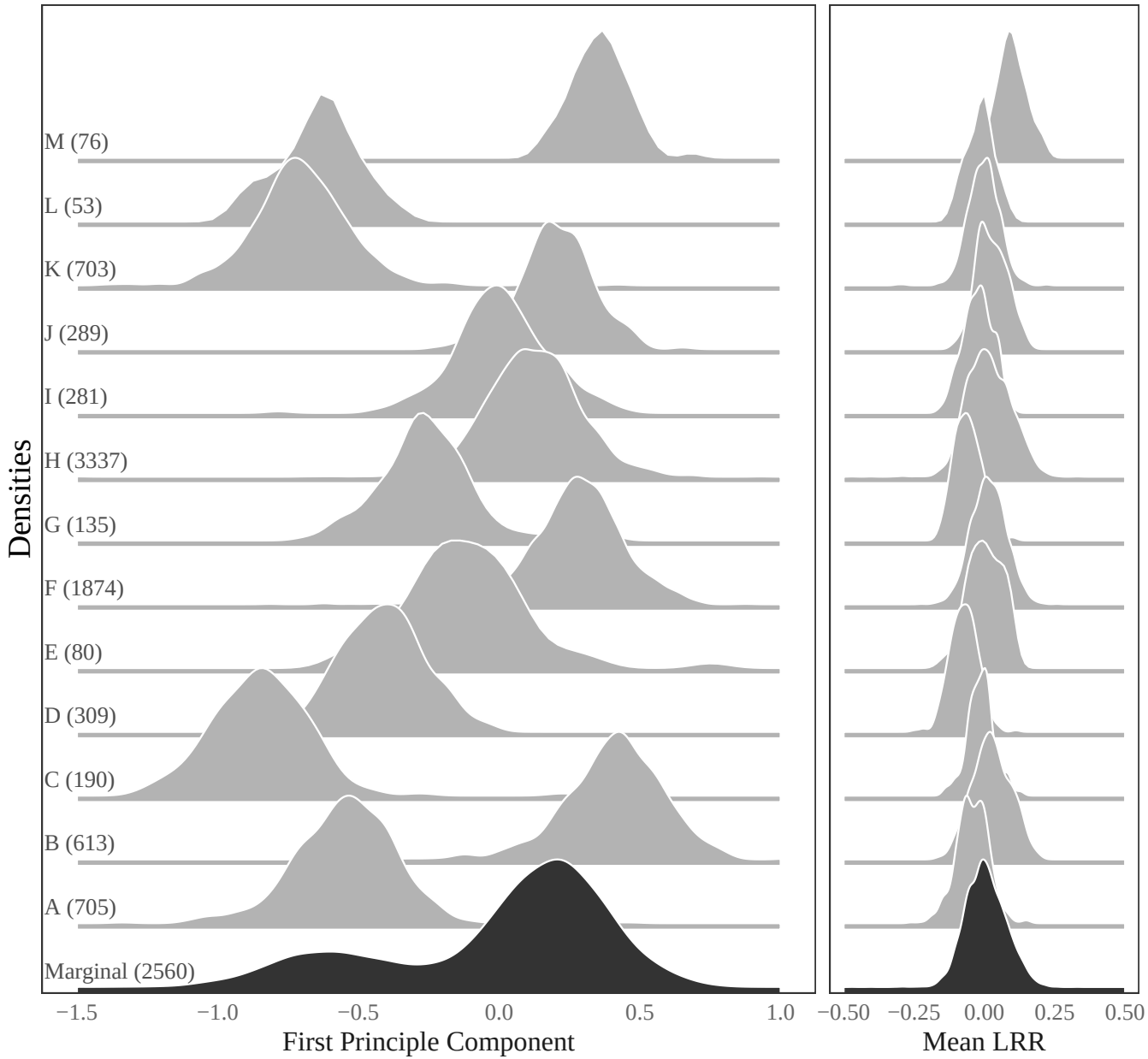
Marginal distribution



Marginal distribution



Marginal distribution

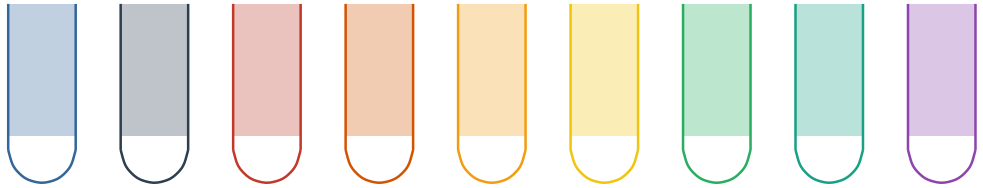


Challenges

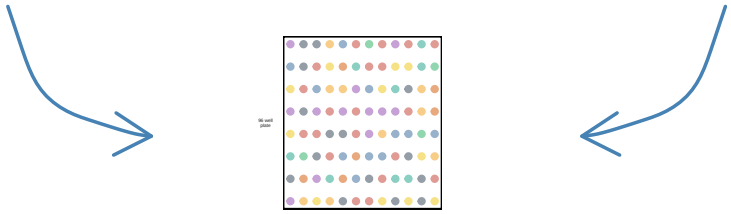
- Consequences of batch effects similar to copy number
- We do not know the batches
  - Time is often a surrogate for the unknown batch effects
  - Samples are processed on hundreds of chemistry plates in large samples

Data processing in the Pancreatic Cancer Consortium

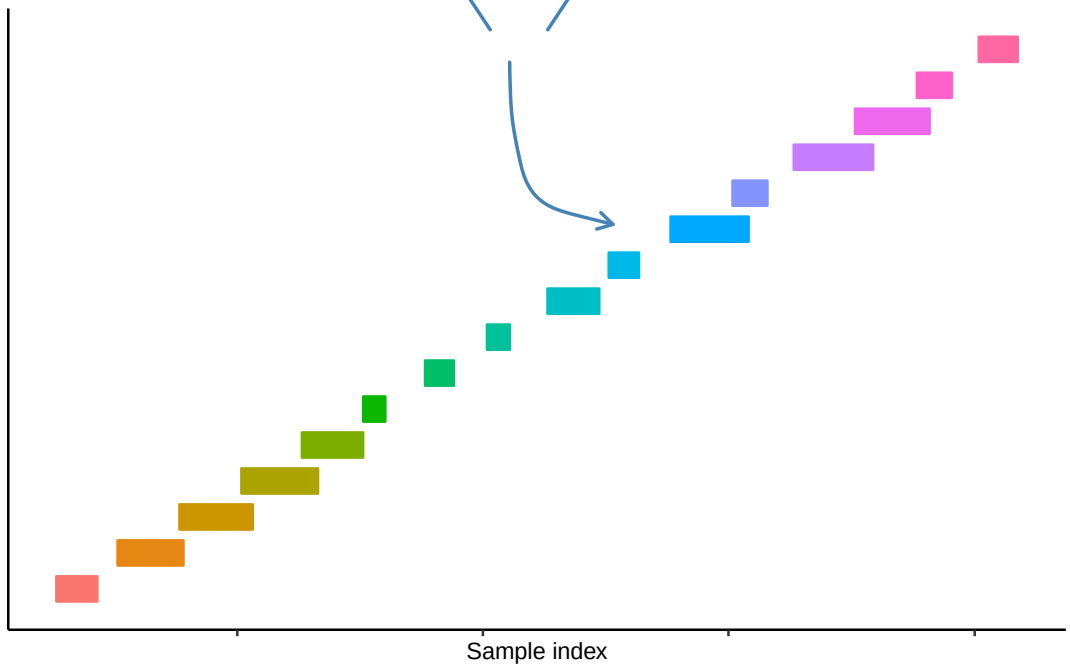
DNA extracted
from 9 centers



Randomization to
plate by
case status
and study center



Scan date
of array



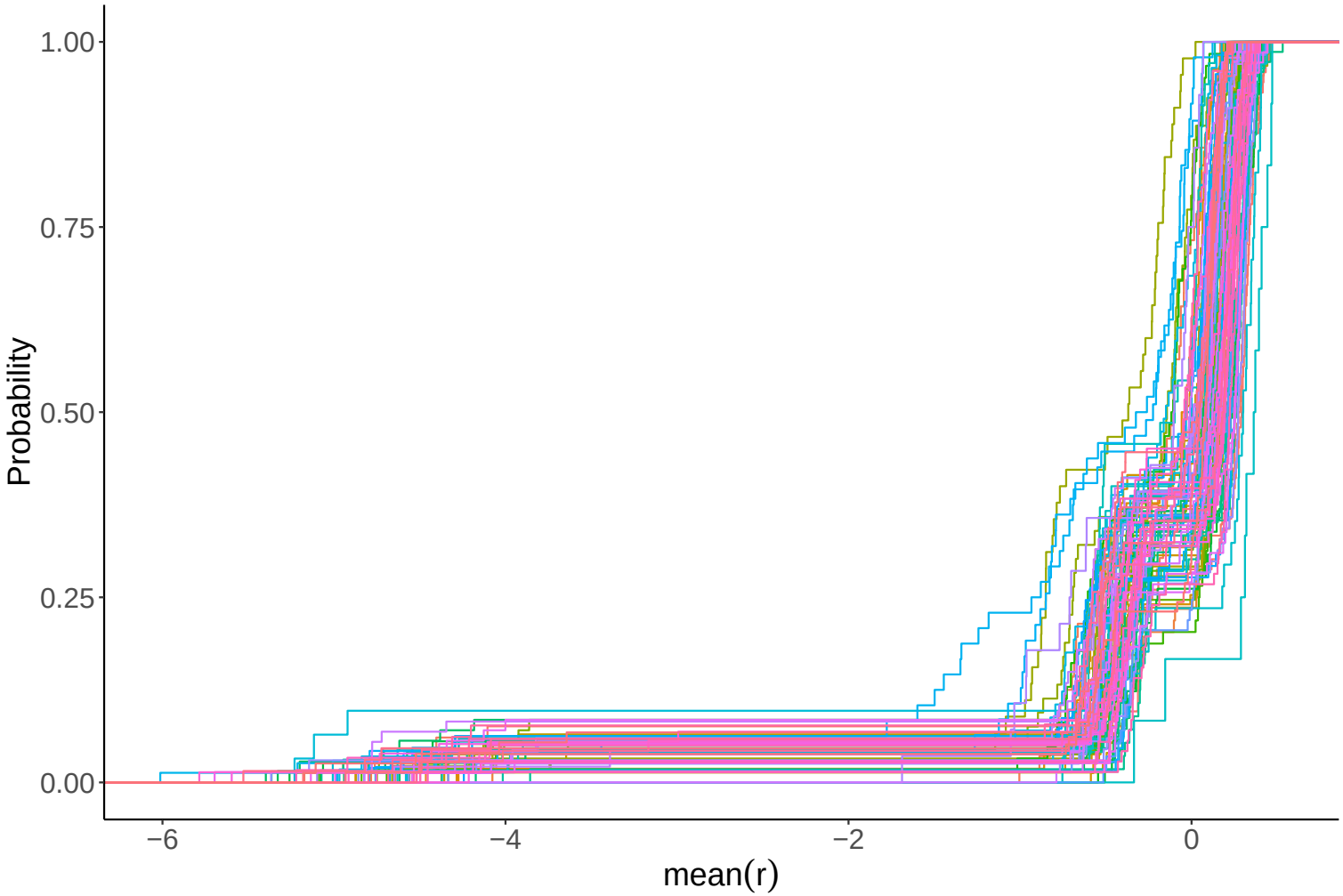
Surrogate variable analysis (SVA) for latent batch effects

- SVA would also remove variation from the latent biological subclasses (here, the latent copy number states)

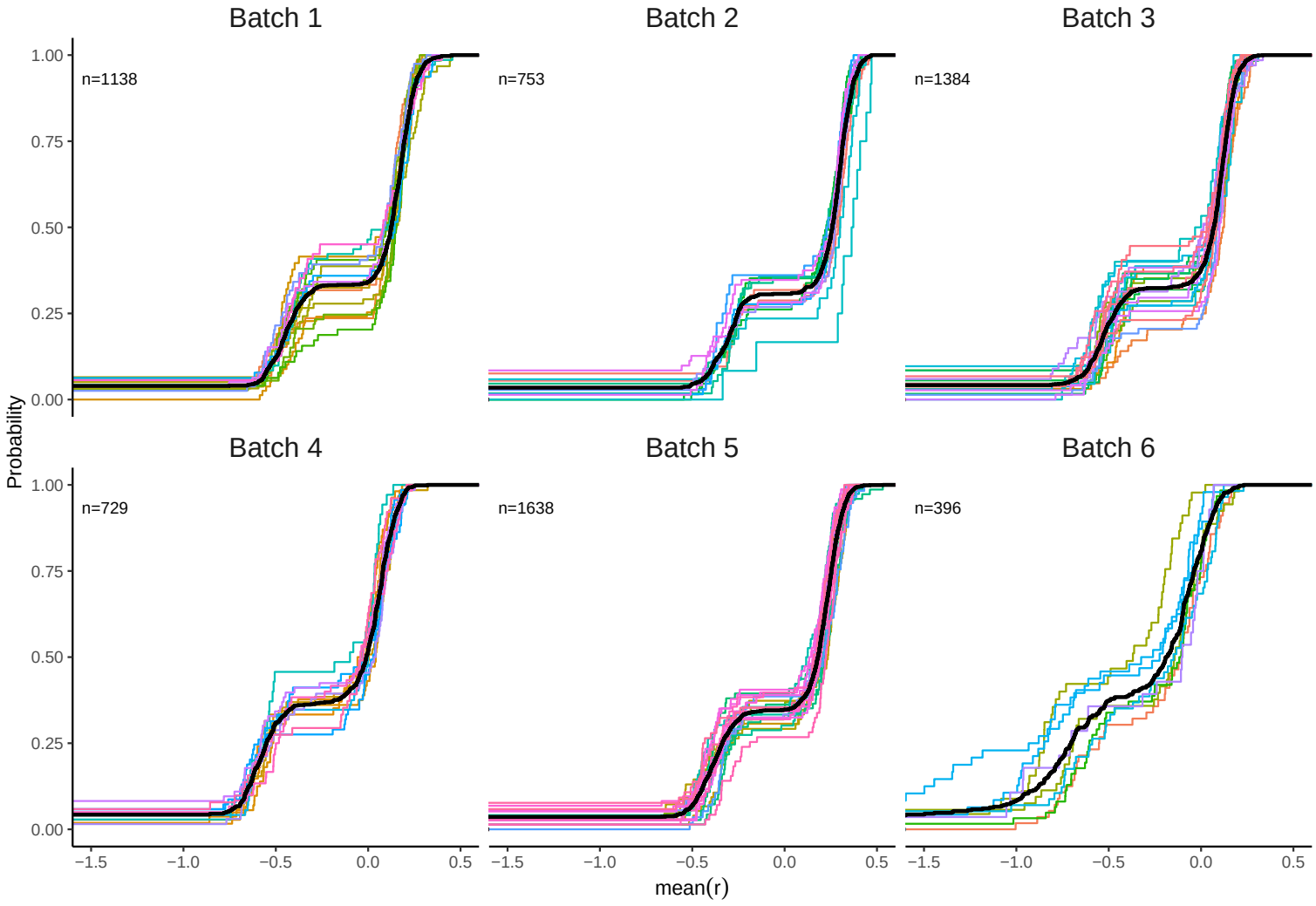
Simple approach

- Provisionally define batch using commonly available metadata available on the samples in a study
- This information is too granular for mixture models
  - hundreds of chemistry plates
  - scan / sequencing date

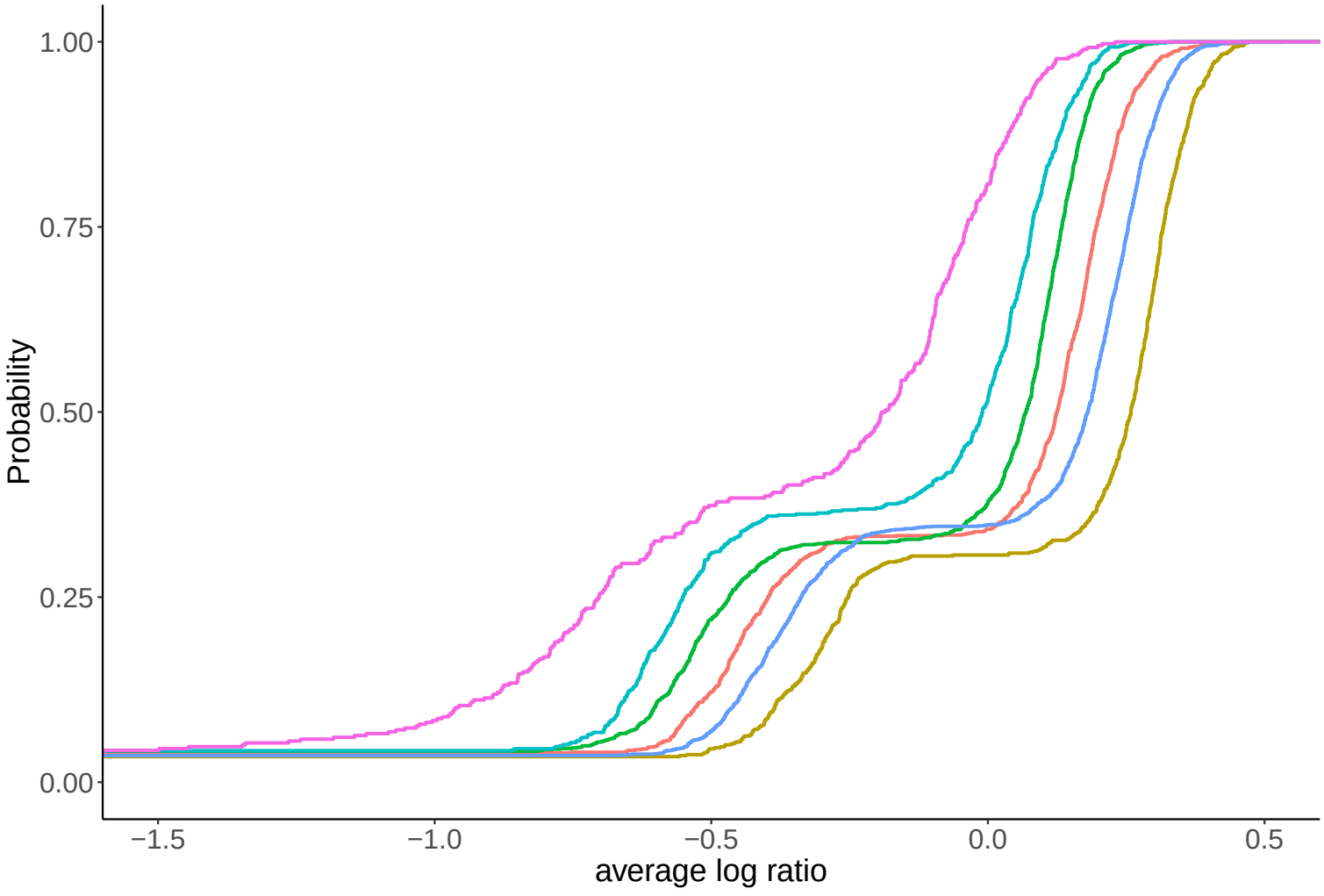
eCDFs of the individual chemistry plates



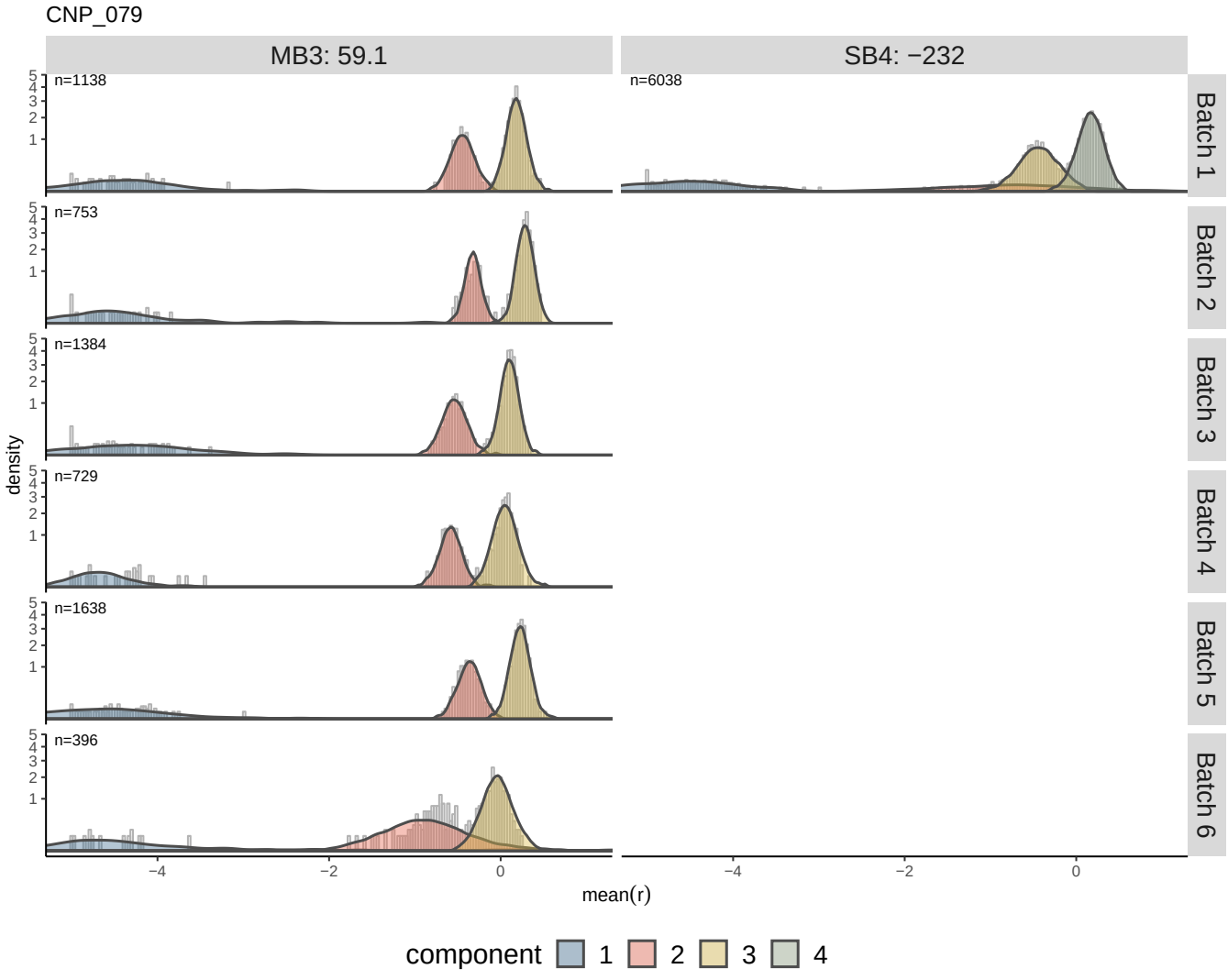
Combine plates with similar eCDFs



Batches are mostly location shifts

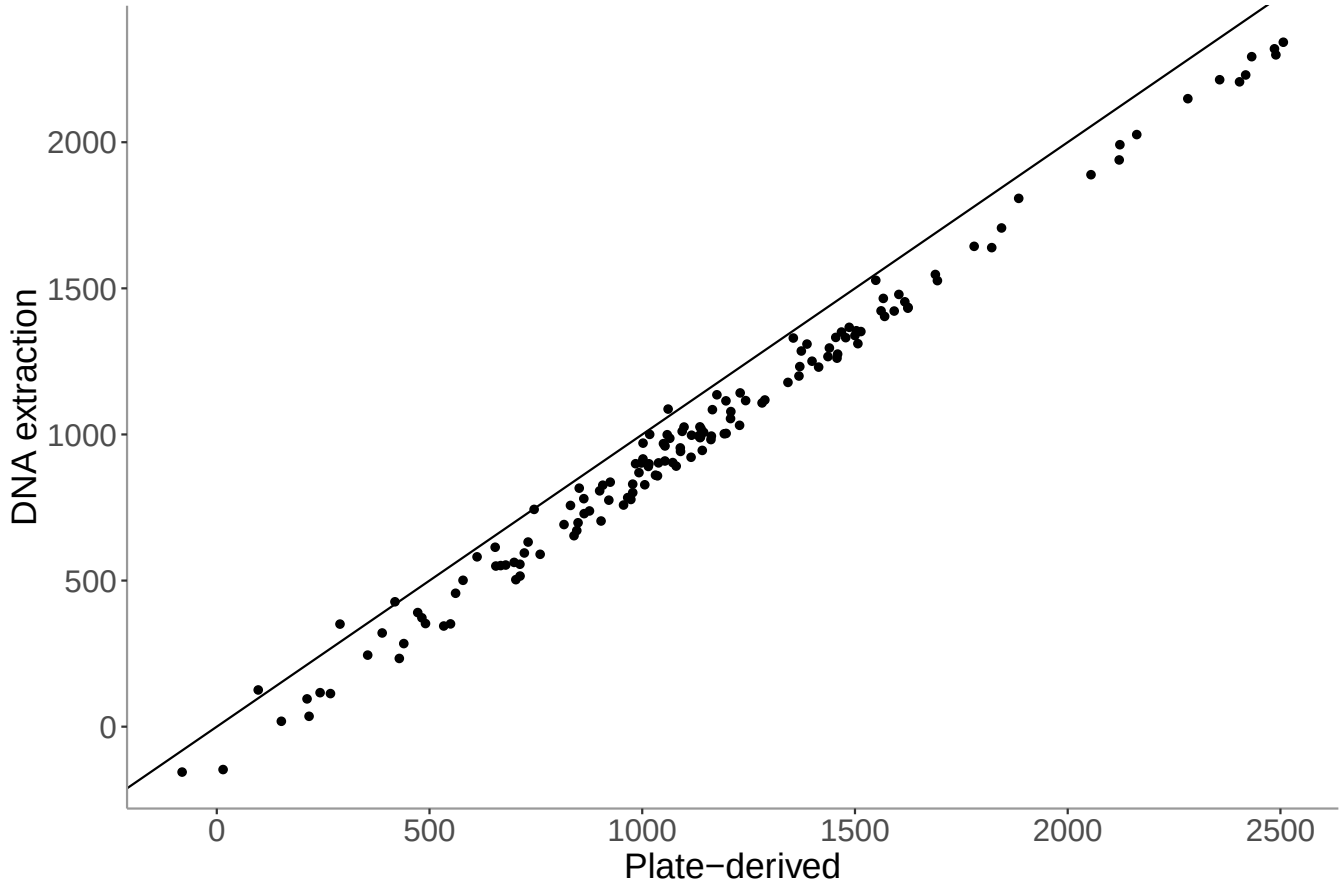


Model abundances hierarchically as a mixture of t distributions



- top 2 models by marginal likelihood

Chemistry plate was just a guess

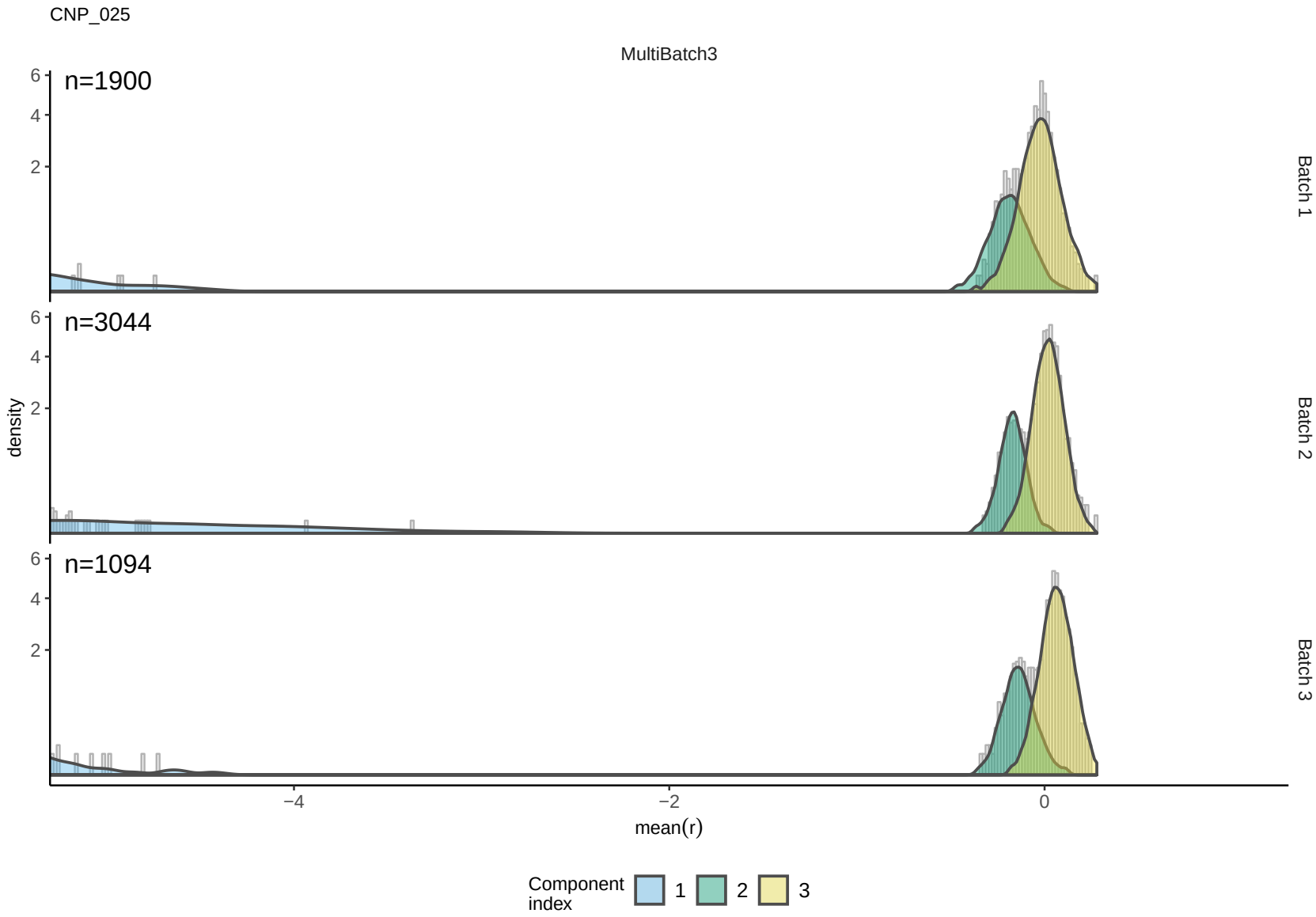


- Marginal likelihoods for 300 CNP regions
- Suggests timing explained more of the technical variation
- Or, study center / DNA extraction method was too coarse

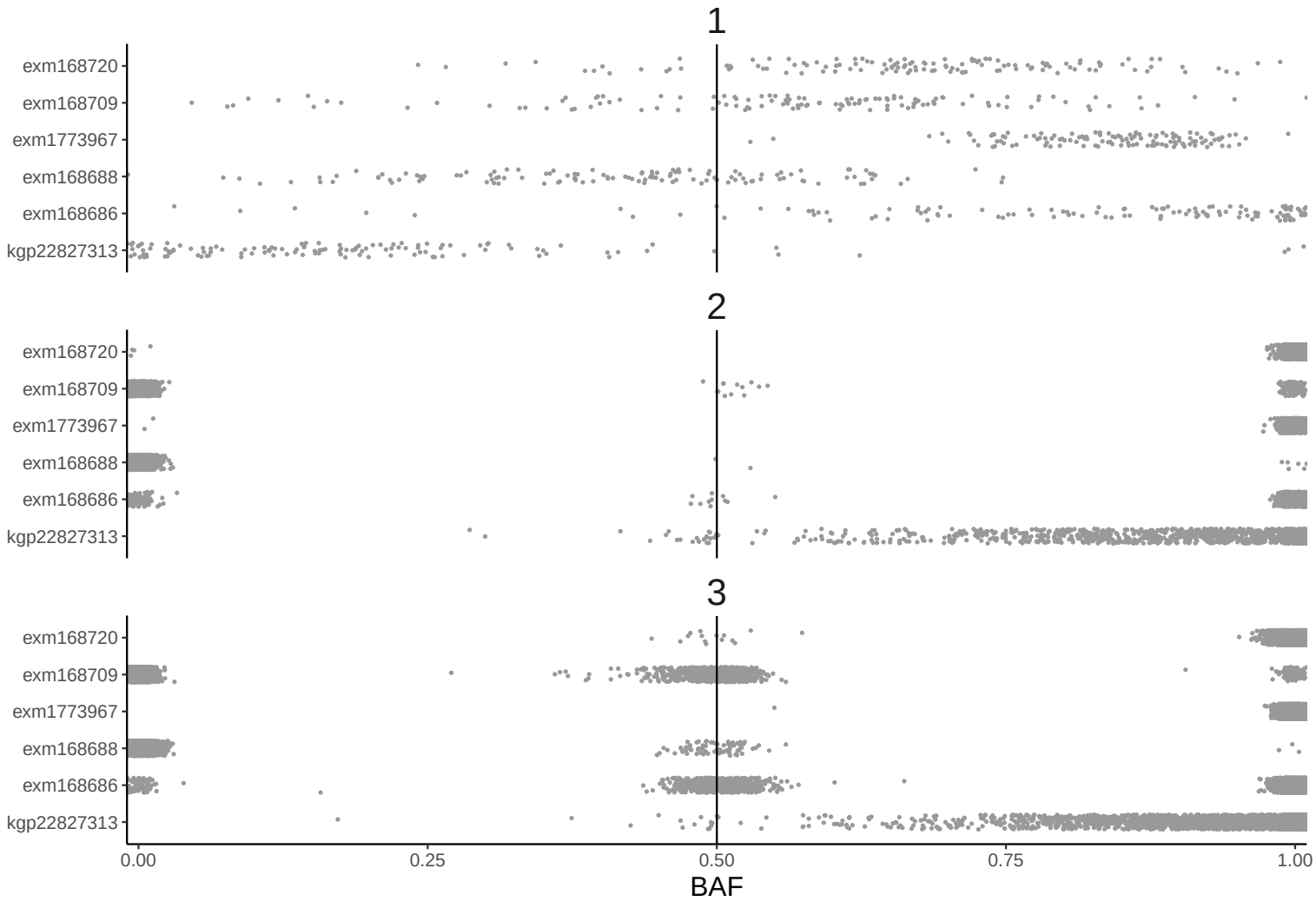
Mixture components need not correspond to differences in latent copy number (unfortunately)

- batch estimates do not always account for skewed / heavy-tailed data
- merge components by amount of overlap
  - distinct copy number states with substantial overlap
  - same copy number state with small overlap
  - merging does not genotype components
- the actual copy number is critical for improving trio-based inference

Components with substantial overlap

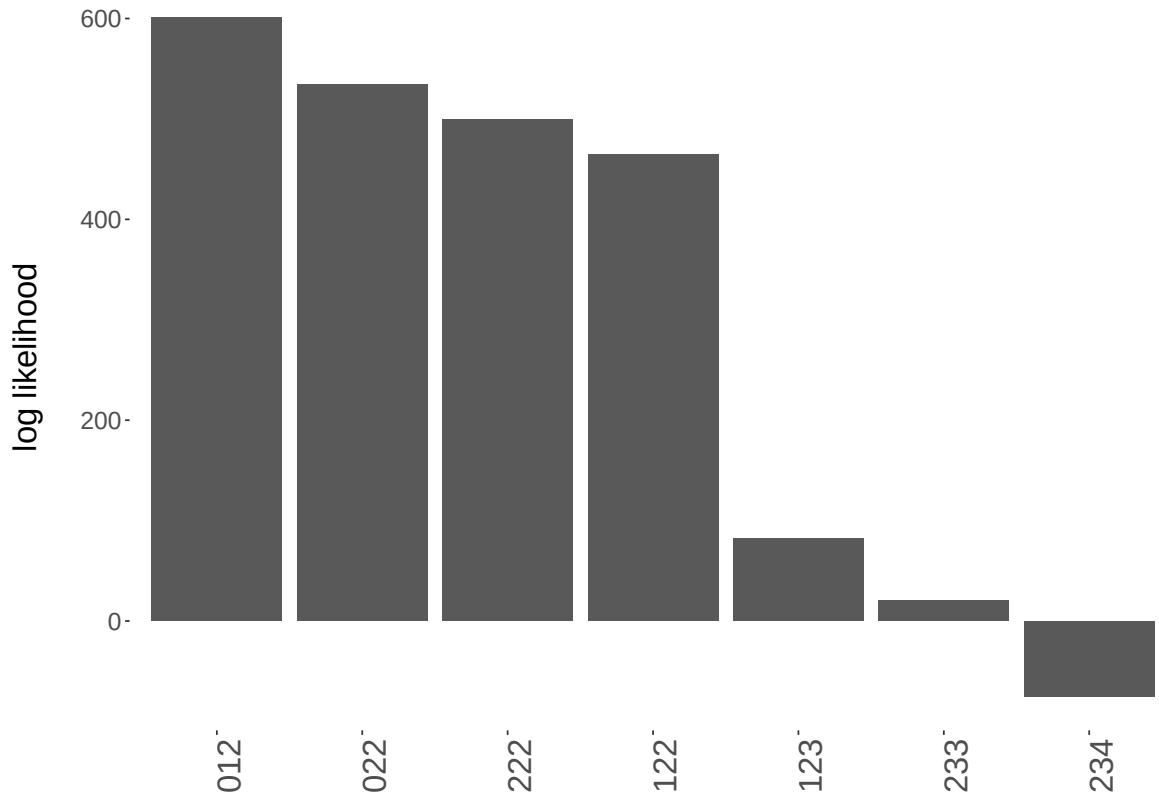


Approach: fit yet another mixture model

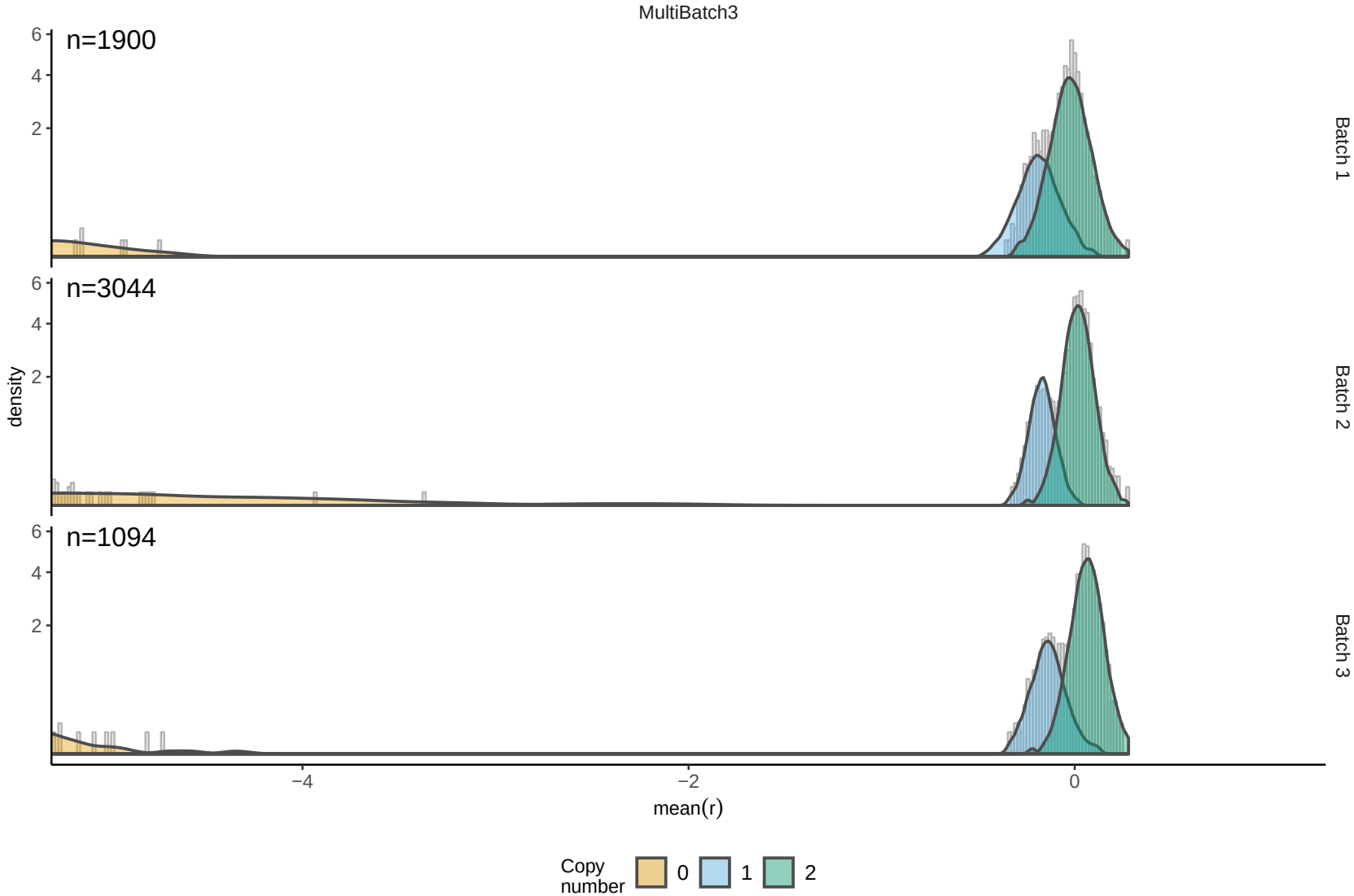


- What copy number states maximize the likelihood of the observed allele frequencies?

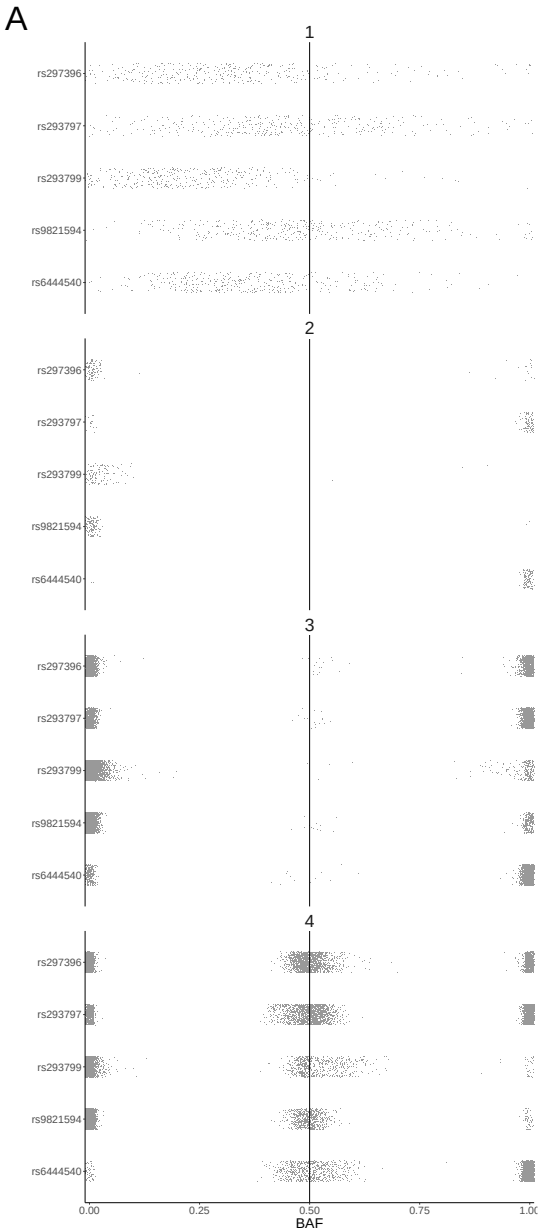
Log likelihoods for the allele frequencies

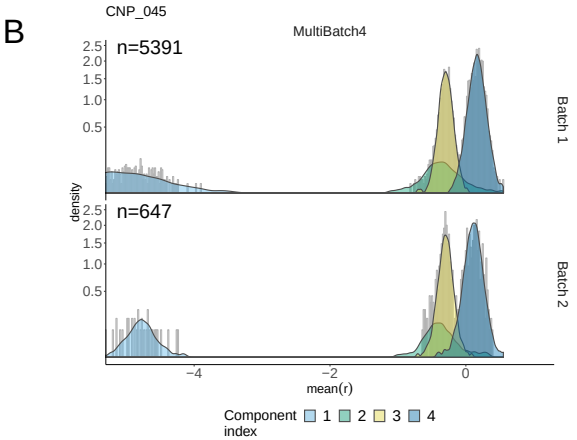


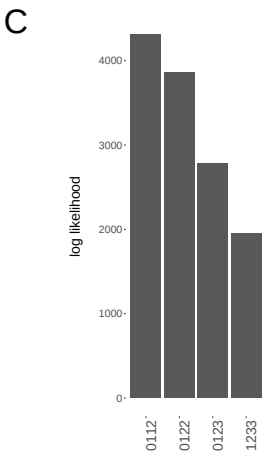
Mapped components

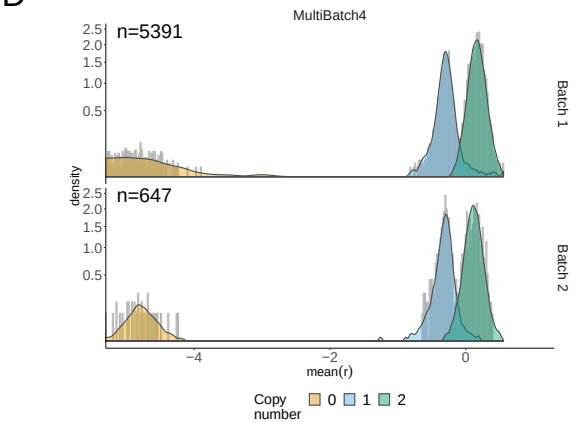


Components with substantial overlap



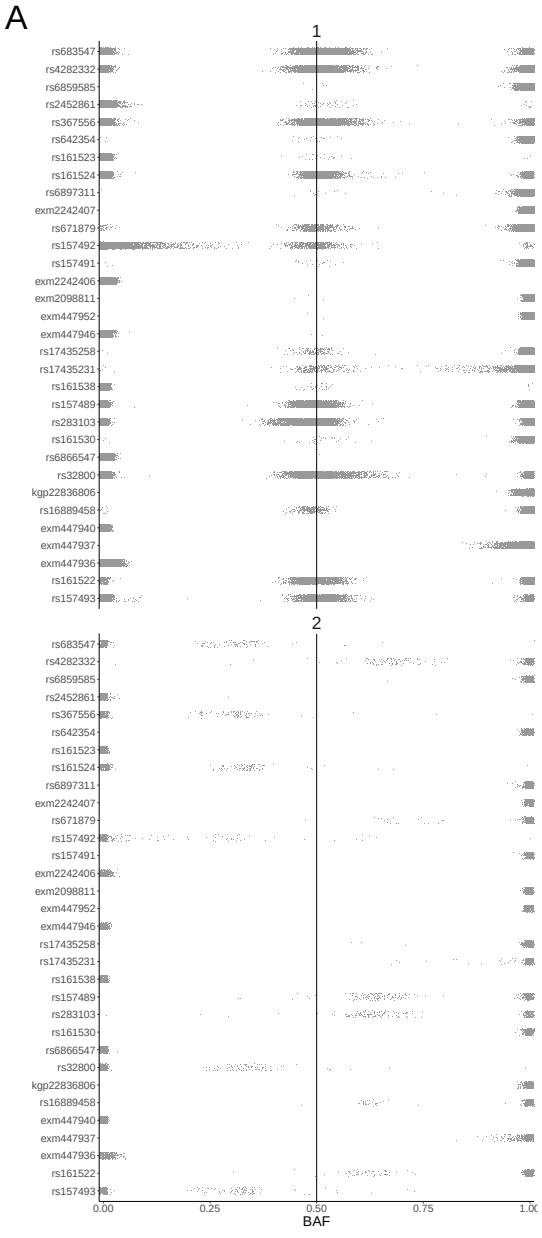


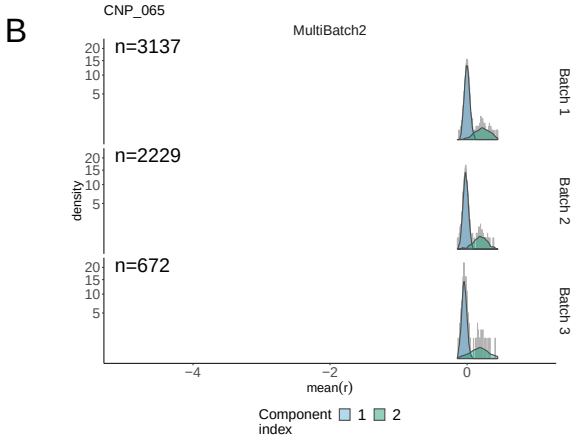


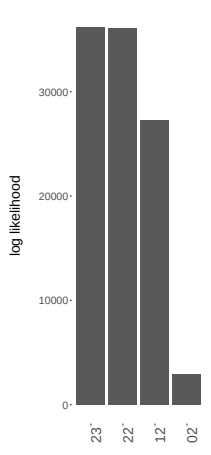


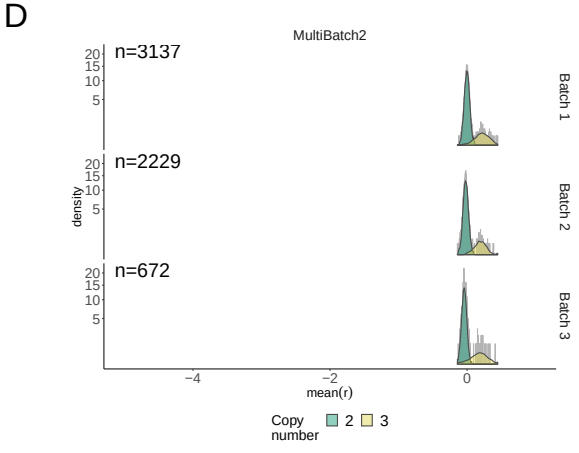
- Components 2 and 3 capture the heterozygous deletion

Duplication polymorphism









CNPBayes

Bioconductor
OPEN SOURCE SOFTWARE FOR BIOINFORMATICS

Home | Install | Help | Developers | About

Search:

Home » Bioconductor 3.7 » Software Packages » CNPBayes

CNPBayes

platforms all | downloads top 50% | posts 0 | in Bioc 3 years
build warnings

DOI: 10.18129/B9.bioc.CNPBayes

Bayesian mixture models for copy number polymorphisms

Bioconductor version: Release (3.7)

Bayesian hierarchical mixture models for batch effects and copy number.

Author: Stephen Cristiano, Robert Scharpf, and Jacob Carey

Maintainer: Jacob Carey <jcarey15 at jhu.edu>

Citation (from within R, enter `citation("CNPBayes")`):

Cristiano S, Scharpf R, Carey J (2018). *CNPBayes: Bayesian mixture models for copy number polymorphisms*. R package version 1.10.0, https://github.com/scristia/CNPBayes.

Installation

To install this package, start R and enter:

```
## try http:// if https:// URLs are not supported
source("https://bioconductor.org/biocLite.R")
biocLite("CNPBayes")
```

Documentation »

Bioconductor

- Package vignettes and manuals.
- Workflows for learning and use.
- Course and conference material.
- Videos.
- Community resources and tutorials.

R / CRAN packages and documentation

Support »

Please read the posting guide. Post questions about Bioconductor to one of the following locations:

- Support site - for questions about Bioconductor packages
- Bioc-devel mailing list - for package developers

Conclusions

- Batches are inevetible in large scale studies
- Be careful using principal components to summarize copy number
- Metadata on the samples can be used to provisionally define batch
- Copy number (not mixture component indices) critical for extension to trio-based studies

Acknowledgements

- Stephen Cristiano
- Alison Klein
- David McKean
- Jacob Carey