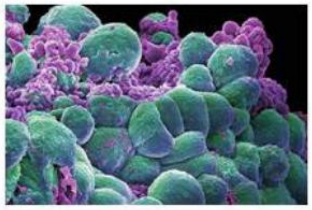
A computational method for estimating the PCR duplication rate in DNA and RNA-seq experiments

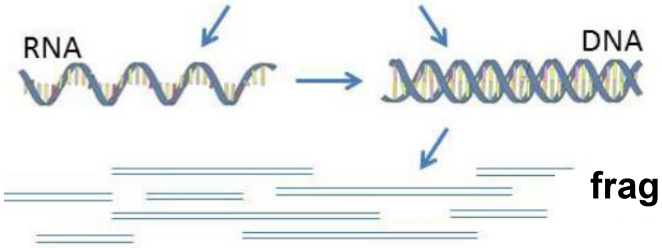
Vikas Bansal, Ph.D.

School of Medicine
University of California, San Diego

Presented at APBC, China, 2017

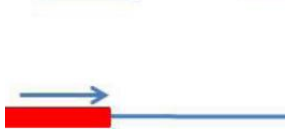
Illumina library preparation



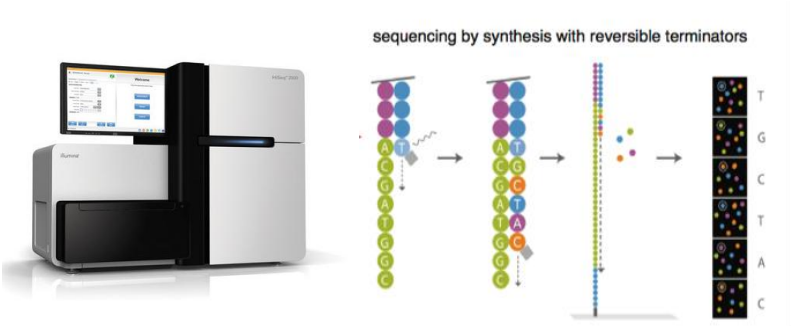


fragmentation

Ligate adapters



PCR amplification



ACGTAGCTAGCGCAC
GCATCGACGAGTATA
TCGATGGATGATCCA

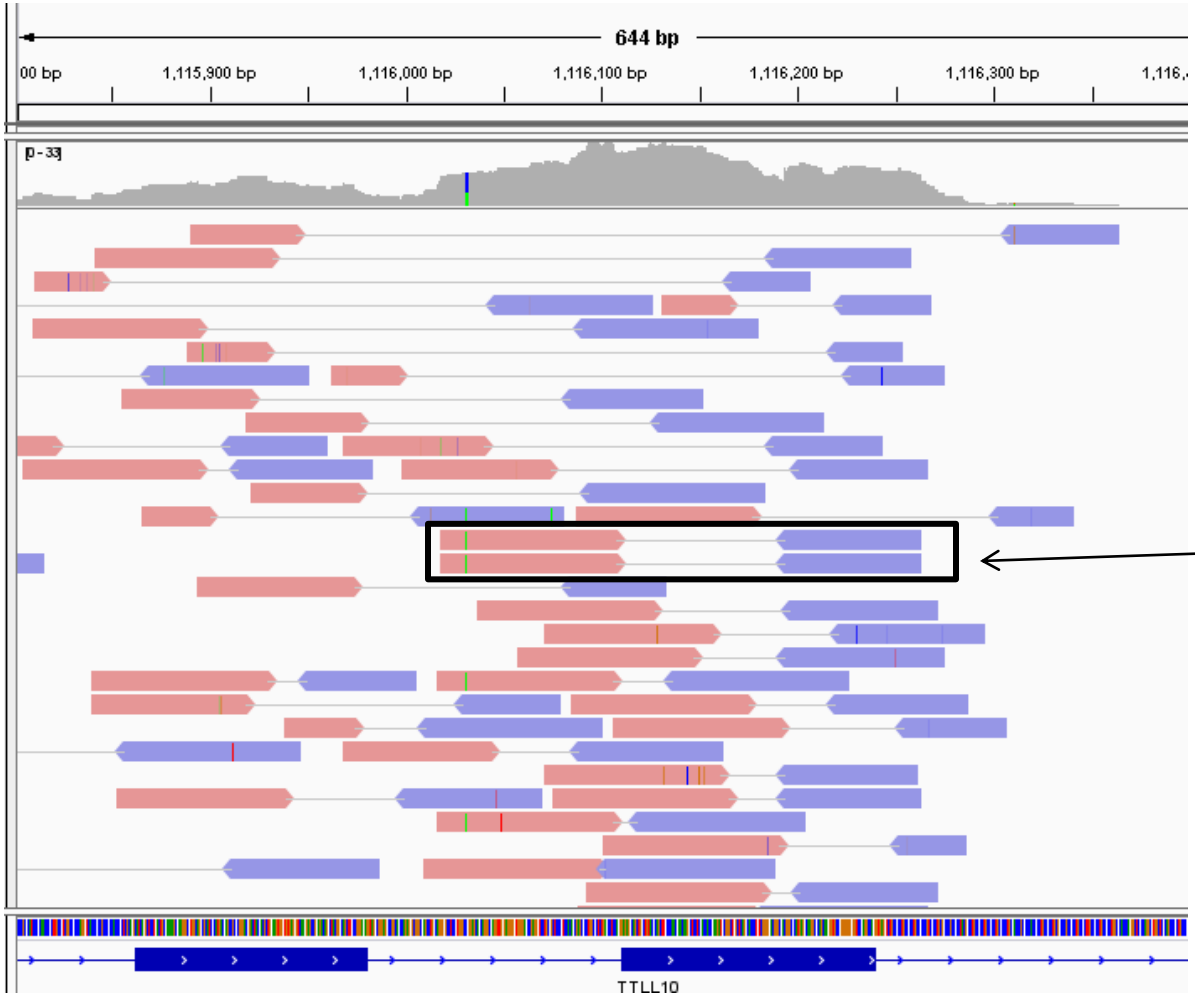
Raw sequence reads

Read mapping to reference genome

Figure adapted from Head et al., Biotechniques 2014

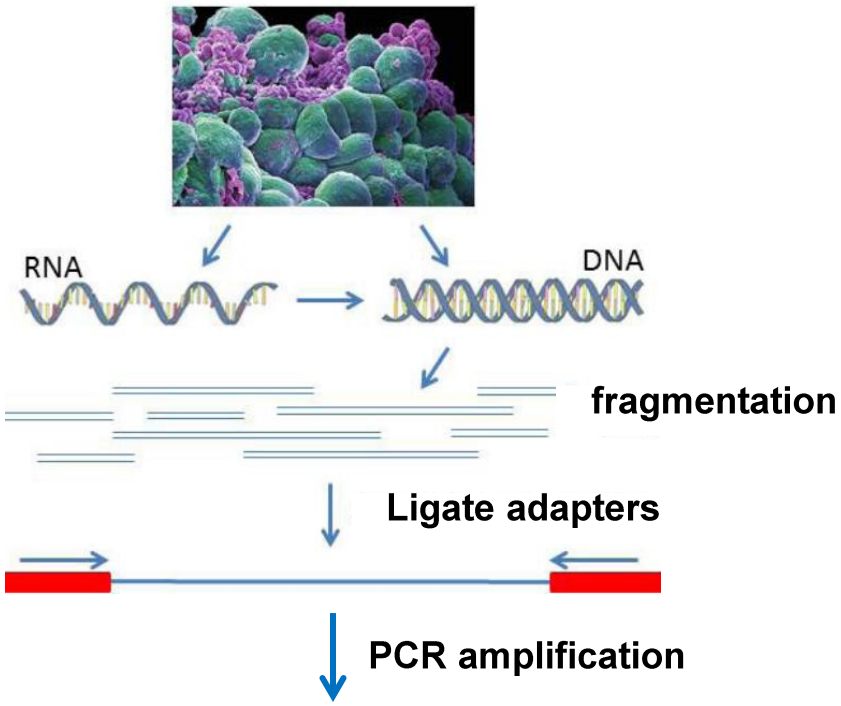
Read duplicates

Reference genome

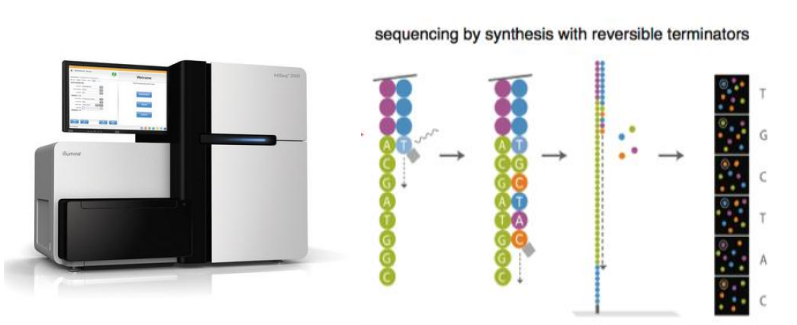


group of reads with identical alignment coordinates

Three sources of read duplicates



1. **DNA fragmentation (natural duplicates)**
2. **PCR amplification**
3. **Cluster identification (optical duplicates)**



ACGTAGCTAGCGCAC
GCATCGACGAGTATA
TCGATGGATGATCCA

Raw sequence reads

Read mapping to reference genome

Figure adapted from Head et al., Biotechniques 2014

PCR duplicates = redundant information

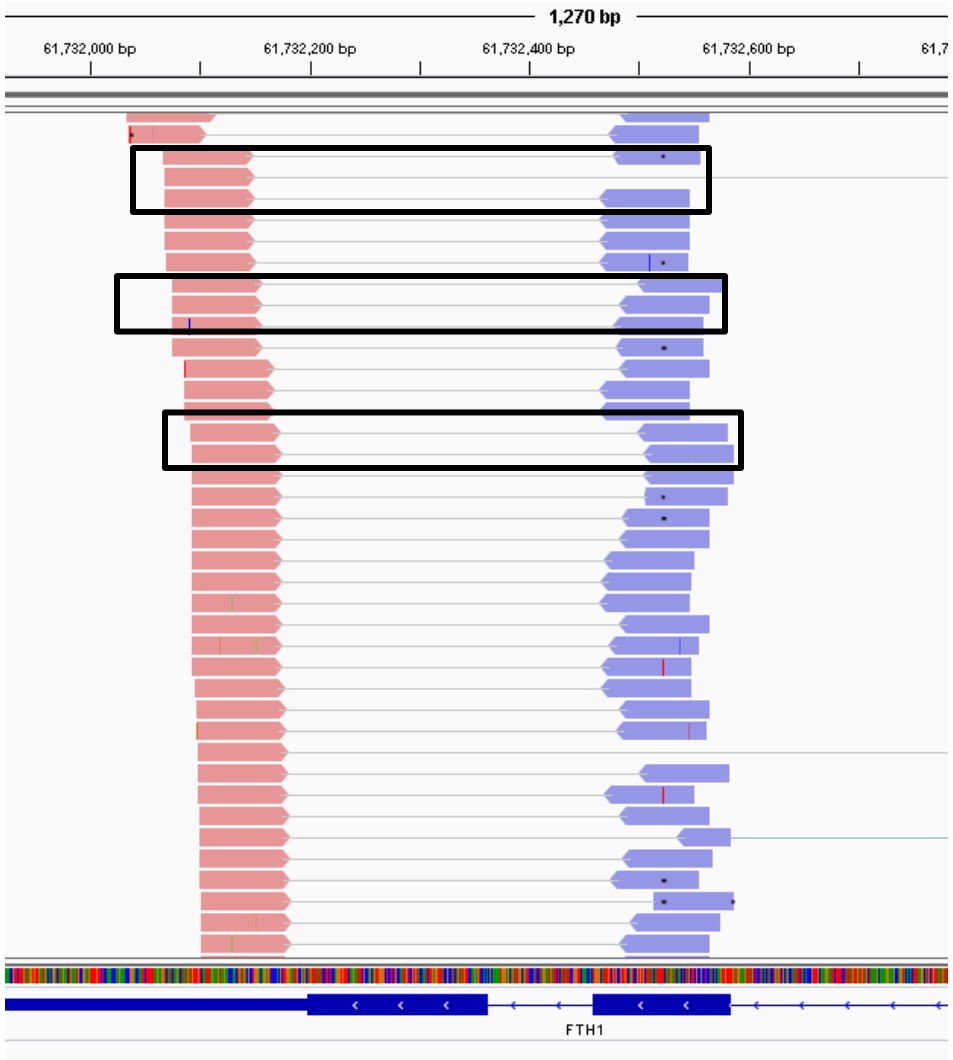
- False positives in variant calling
- Bias estimates of allele frequencies in deep sequencing experiments

- **Solution:** computationally identify read duplicates and keep only one read per group for analysis

- **Limitation:** cannot differentiate between PCR duplicates and natural duplicates

Removing natural duplicates in RNA-seq biases gene expression estimates

- Highly expressed genes have 1000's of reads mapped to small space
- Read duplication rate of 20-25% is normal



Differentiating natural duplicates from PCR duplicates using UMIs

- Random barcodes added to each DNA fragment before amplification
- Read duplication rate = $1 - \frac{3}{66} = 0.95$
- PCR duplication rate = $1 - \frac{19}{66} = 0.71$
- Costly and requires custom library preparation

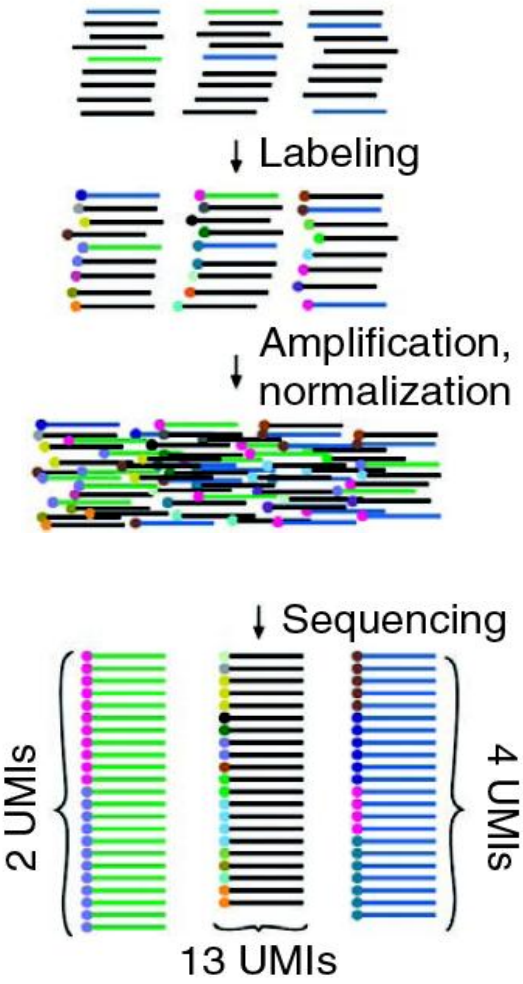
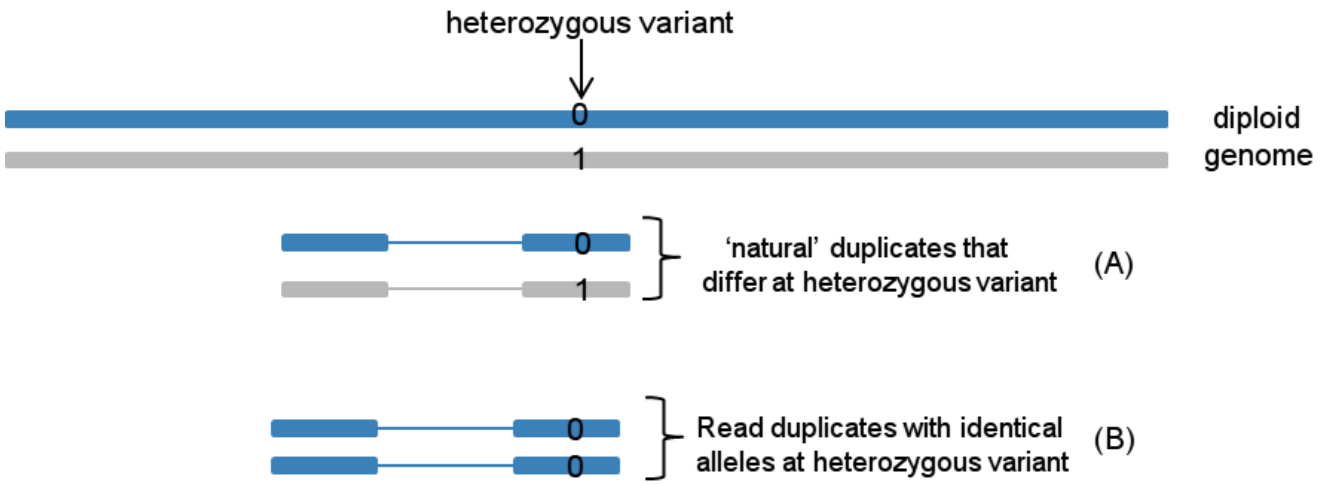


Figure adapted from Kivioja et al. Nat. Methods 2011

A computational approach for estimating the fraction of read duplicates that are due to PCR amplification

Differentiating natural duplicates from PCR duplicates at heterozygous sites



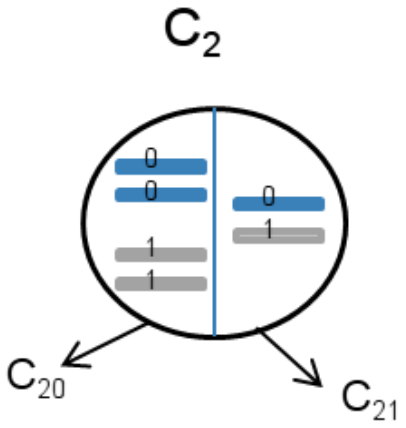
PCR duplicates: identical alleles at heterozygous site

Natural duplicates: equally likely to have identical or different alleles

Estimating fraction of PCR duplicates (clusters of size 2)

C_{20} = # of clusters with identical alleles

C_{21} = # of clusters with different alleles



1. E [# of natural duplicates] = 2 x C_{21}

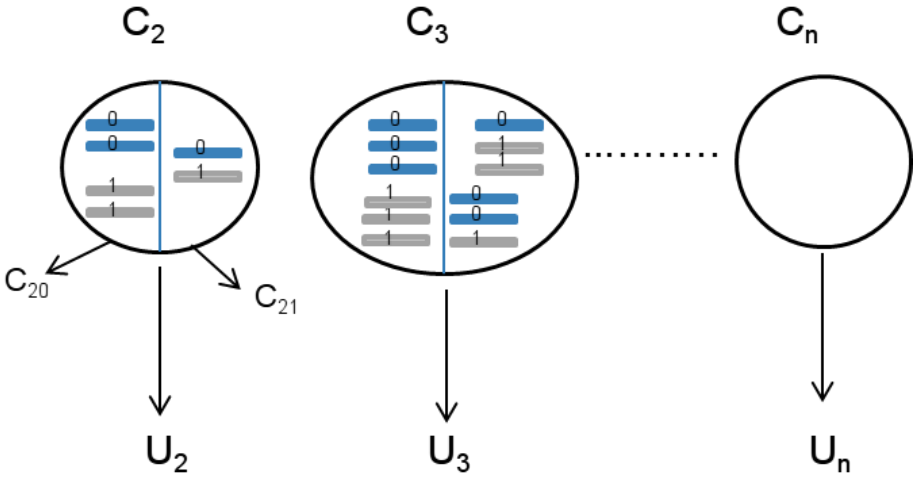
2. **E [# PCR duplicates] = C_{20} - C_{21}**

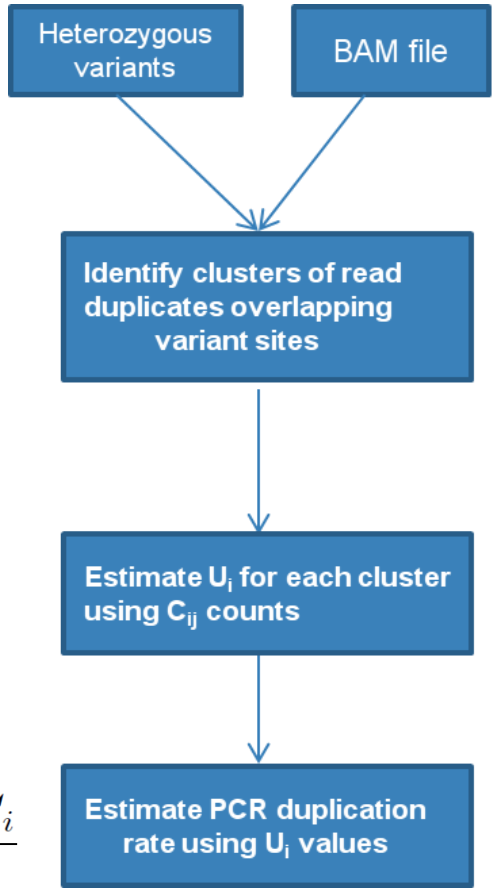
4

+

$$U_2 = \frac{C_{20} - C_{21}}{C_2} \times C_{21}$$

Overview of method

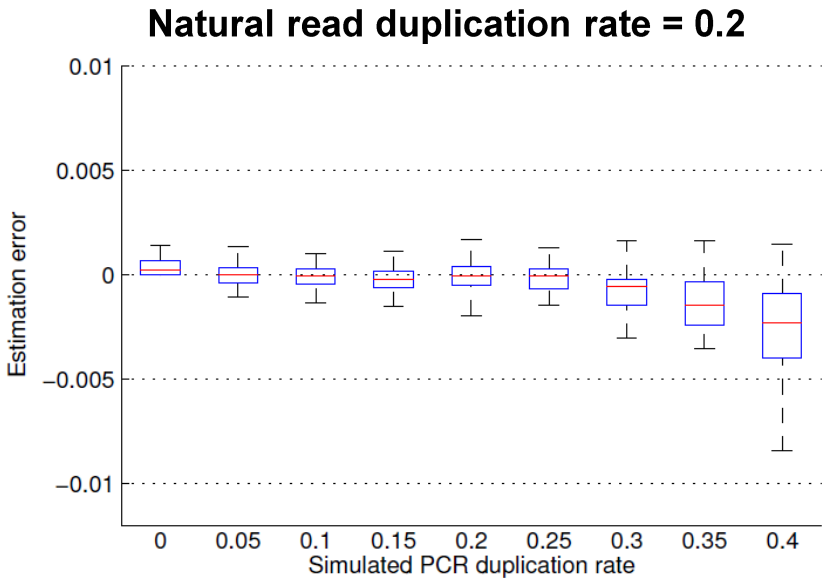


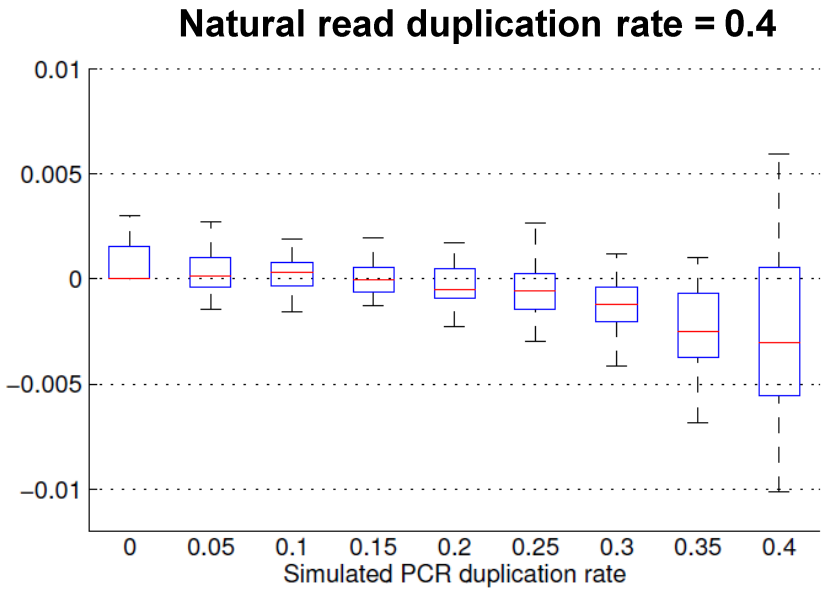


$$1.0 - \frac{\sum_{i=1}^{n} U_i \hat{C}_i}{R}$$

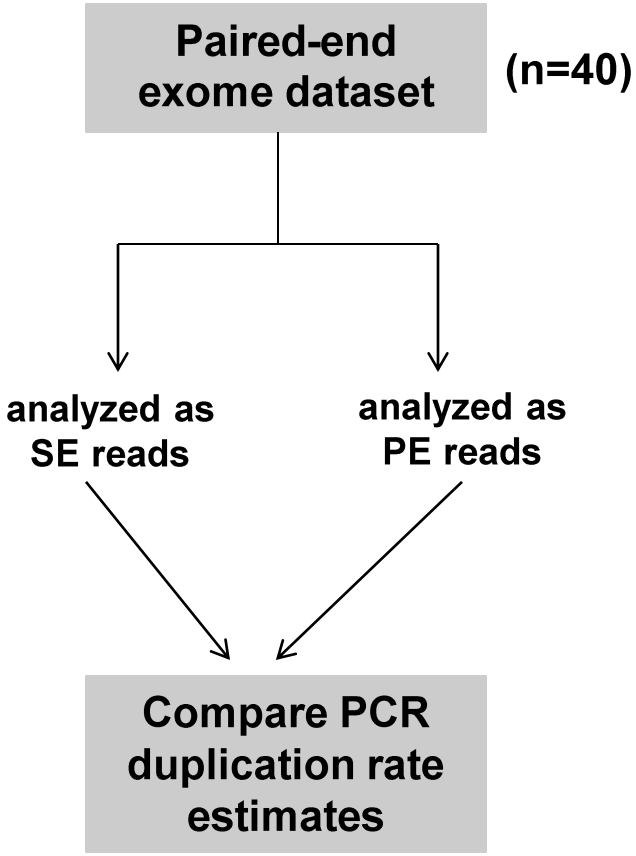
Accuracy on simulated data

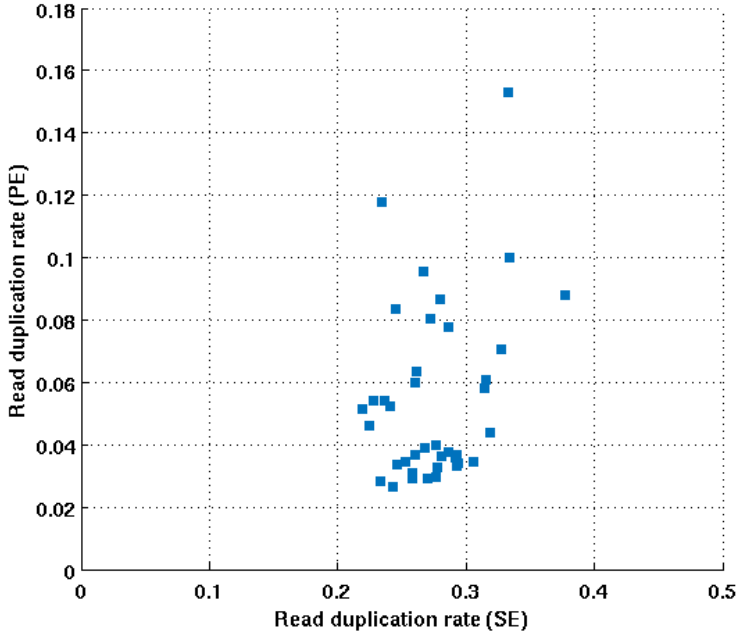
- Used real exome data to create a dataset with no read duplicates
- Simulated natural read duplicates and PCR read duplicates using ‘sampling with replacement’ (50 replicates per combination)





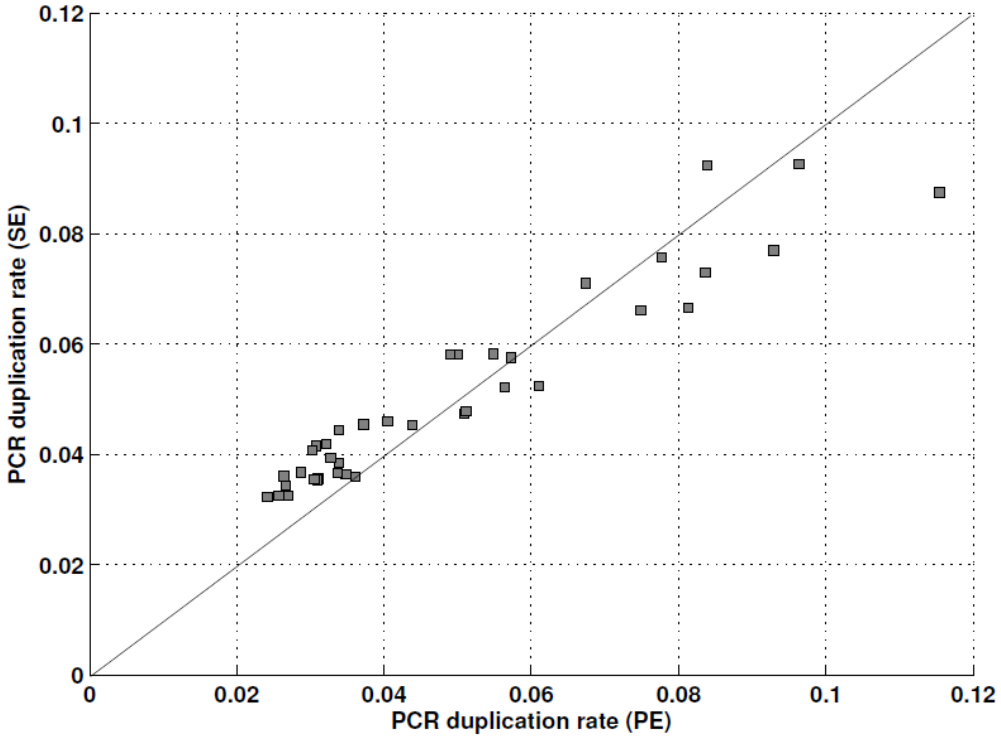
Analysis of exome datasets from 1000 Genomes project



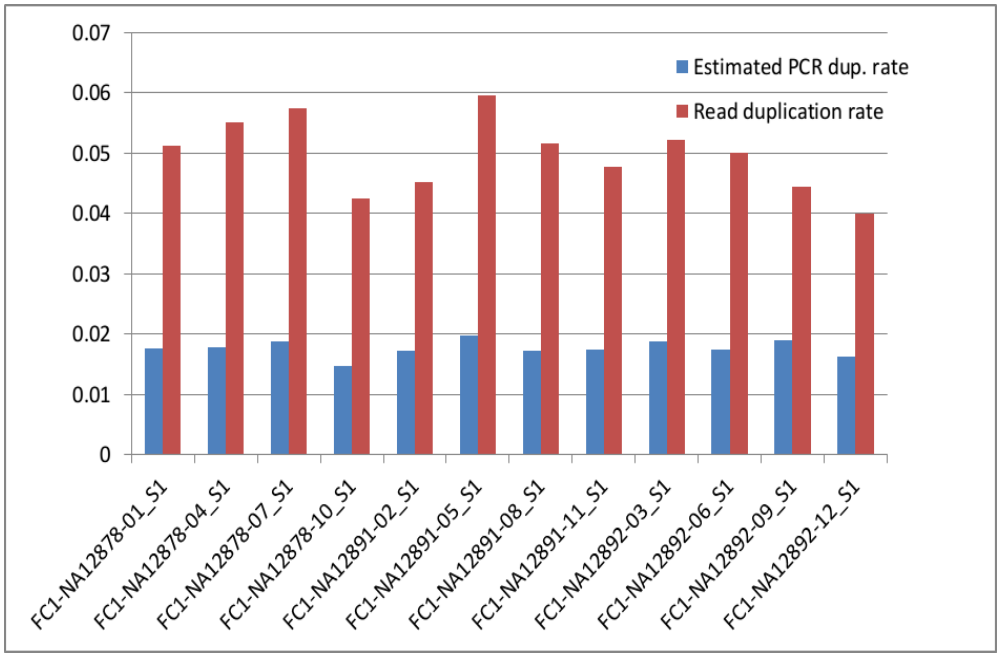


High concordance between estimated PCR duplication rates from SE and PE reads

$R^2 = 0.977$ and mean $|\Delta| = 0.0073$



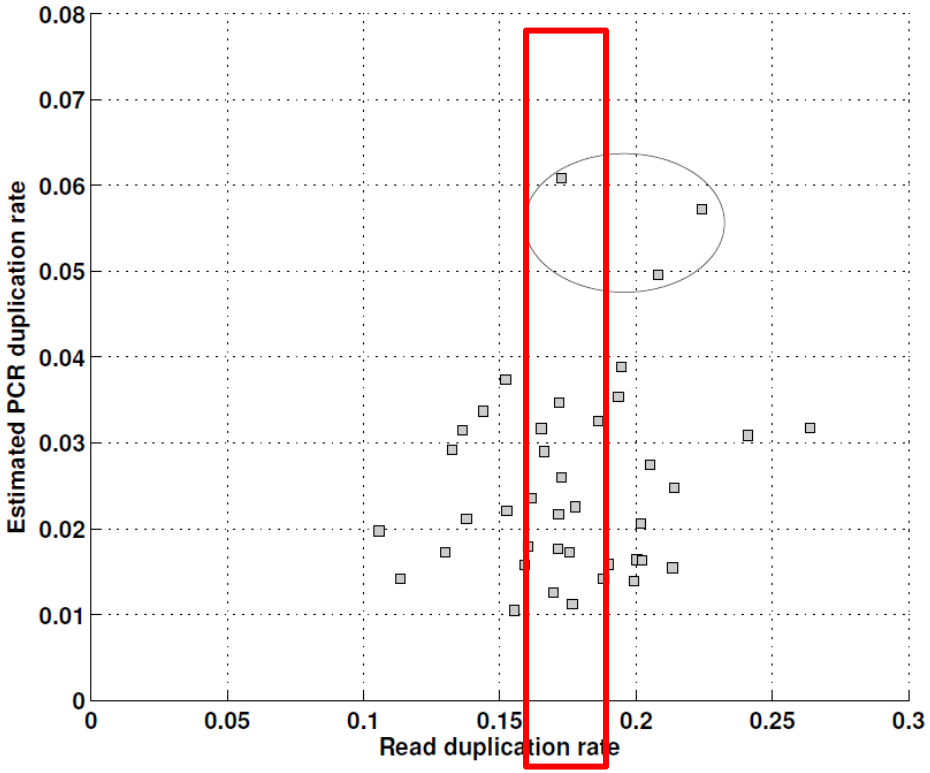
Analysis of exomes from Nextera library preparation method



45% of read duplicates correspond to natural duplicates

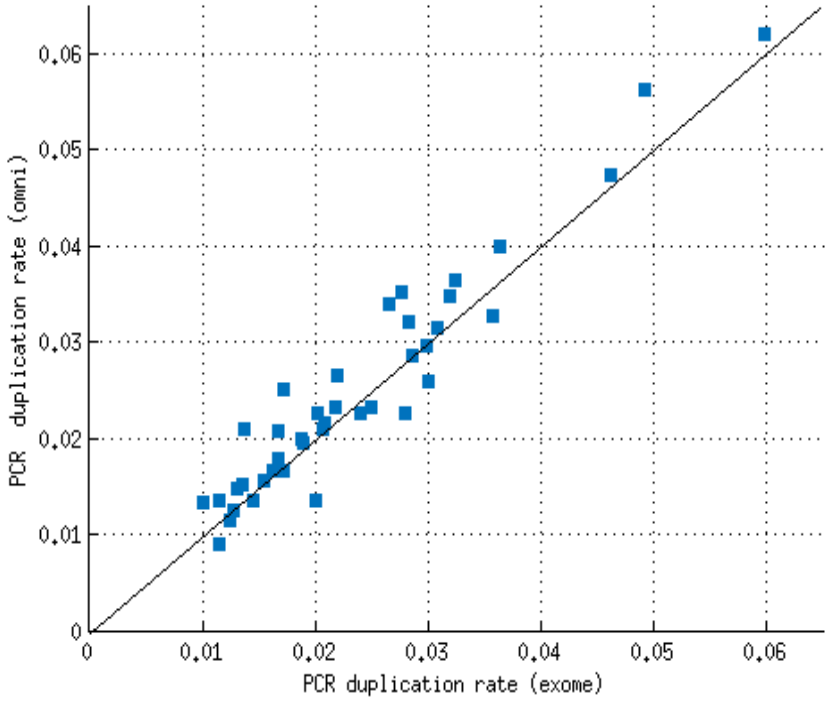
PCR duplication rates for RNA-seq data

- 40 RNA-seq datasets from the Geuvadis project
- Heterozygous variants using exome data from 1000 Genomes Project



Significant variation (1-6%) in the PCR duplication rate

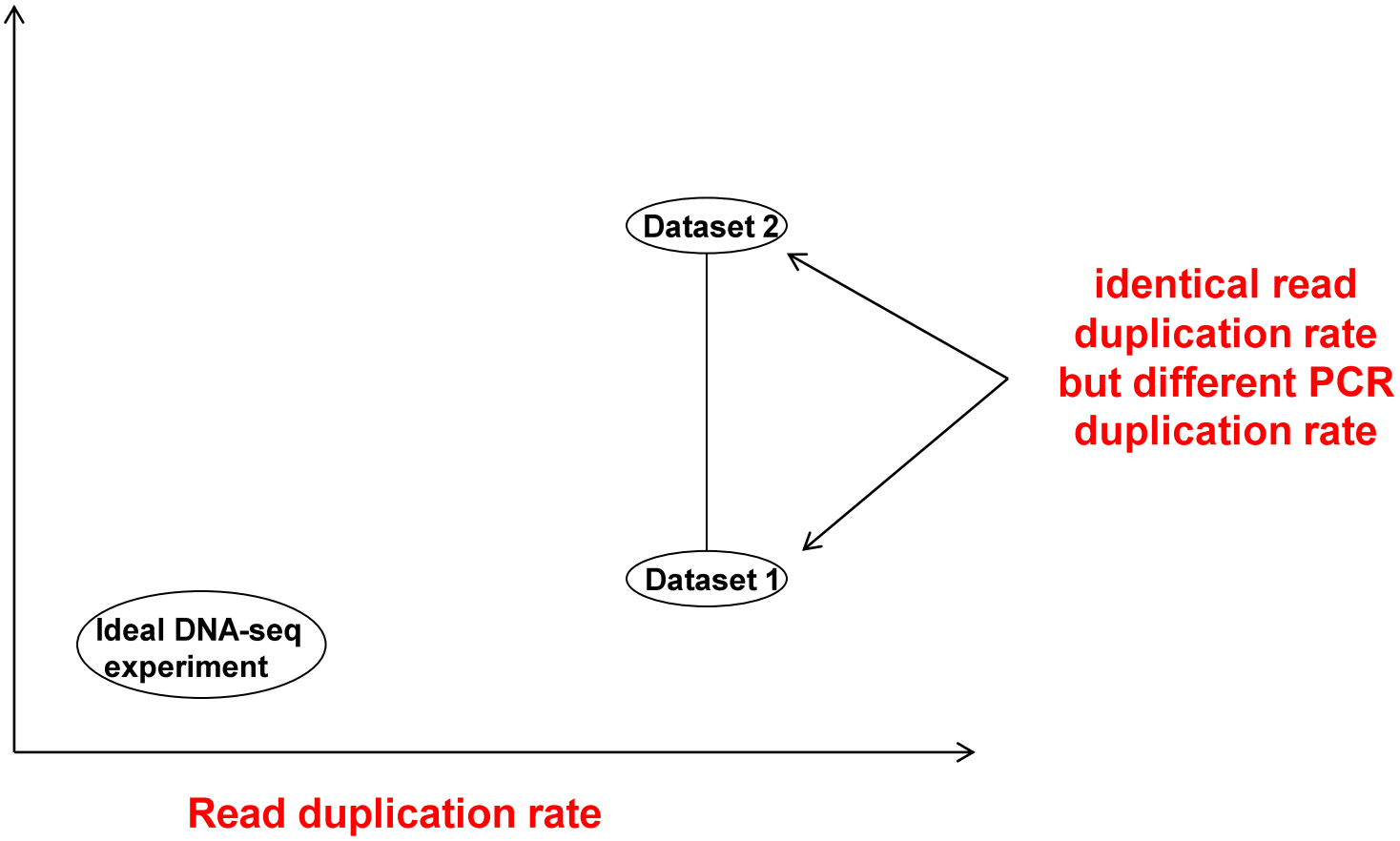
Robustness of estimates to ‘variant calls’



Two sets of variant calls: exome sequencing and Illumina Omni genotyping

R^2 = 0.96 and mean $|\Delta|$ = 0.0027

General utility of our method



PCR duplication rate estimate can be used as covariate in gene expression analysis from RNA-seq data

Summary

- **novel computational method for estimating the PCR duplication rate that accounts for natural duplicates**
  - uses reads overlapping heterozygous variants
  - Rigorous mathematical model

- **Results**
  - Validation using simulations and exome data
  - High proportion of ‘natural read duplicates’ in Nextera protocol
  - 75-90% of read duplicates in RNA-seq are NOT due to PCR amplification

Software available: https://github.com/vibansal/PCRduplicates