WASP: allele-specific software for robust molecular quantitative trait locus discovery

T. Rhyker Ranallo-Benavidez

Johns Hopkins Genomics Journal Club

June 8, 2017

WASP: unbiased allele-specific read mapping and discovery of QTLs

Quantitative Trait Locus:

- A section of DNA (locus) that correlates with variation in a phenotype (quantitative trait)

Allelic Imbalance:

- $p = \frac{\alpha_h}{\alpha_h + \beta_h}$

WASP corrects for:

- biased mapping of allele-specific reads
- homozygous sites incorrectly called as heterozygous
- overdispersion of allele specific read counts

Read Mapping Bias and Correction

Mapping bias:

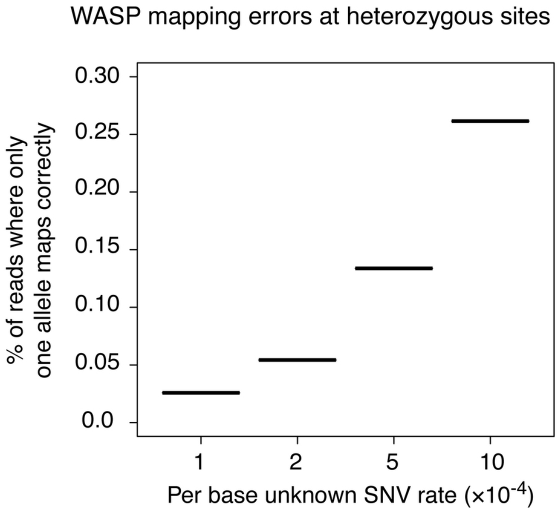
- Reads containing a nonreference allele may not map uniquely or might map to an incorrect location in the genome.

Bias correction:

- Map to a personalized genome (reference sequence is replaced with nonreference alleles)
- Map to a genome with N-masked SNPs
- Filter reads that do not map to exactly the same location when the alleles at SNPs are swapped

WASP only considers known heterozygous sites, since they find that unknown SNPs insignificantly affect proportion of sites with biased mapping.

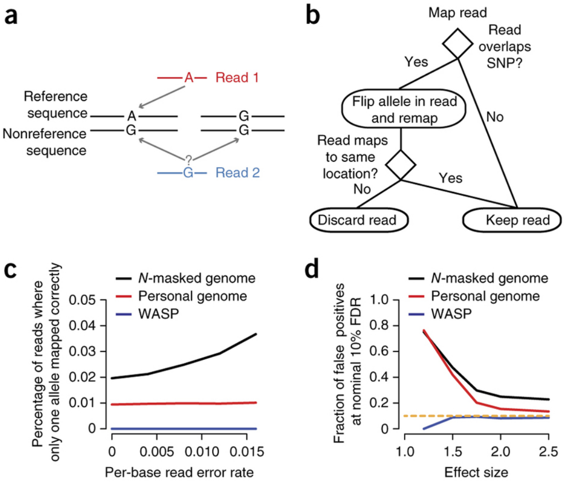
Supplementary Figure 1: Unknown SNPs



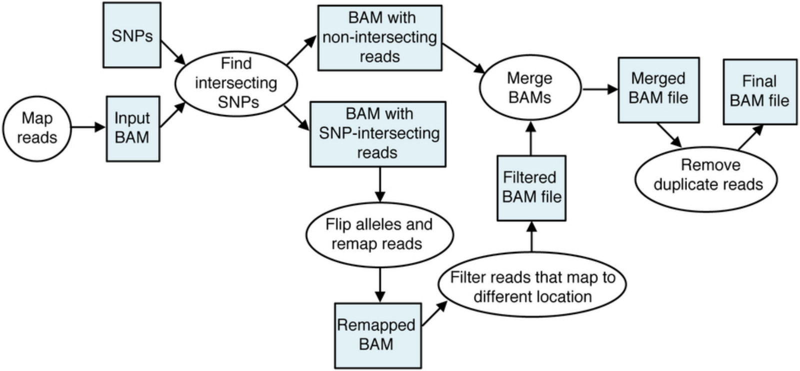
Methods

- Simulated reads at heterozygous sites in a lymphoblastoid cell line (LCL) with a predefined sequencing error rate
- Simulated all possible overlapping reads from both haplotypes
- Mapped using three approaches:
    - Masking with Ns
    - Personalized genome using AlleleSeq
    - Filtering with WASP

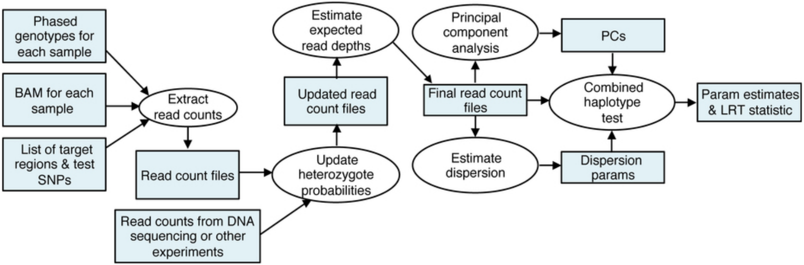
Figure 1


a
Read 1
Reference sequence
Nonreference sequence
A
A
G
G
G
?
G
Read 2
b
Map read
Read overlaps SNP?
Yes
No
Flip allele in read and remap
Read maps to same location?
Yes
No
Discard read
Keep read
c
N-masked genome
Personal genome
WASP
Percentage of reads where only one allele mapped correctly
0.05
0.04
0.03
0.02
0.01
0
0
0.005
0.010
0.015
Per-base read error rate
d
N-masked genome
Personal genome
WASP
Fraction of false positives at nominal 10% FDR
1.0
0.8
0.6
0.4
0.2
0
1.0
1.5
2.0
2.5
Effect size

Supplementary Figure 4



Supplementary Figure 5


Phased genotypes for each sample
BAM for each sample
List of target regions & test SNPs
Extract read counts
Read count files
Read counts from DNA sequencing or other experiments
Update heterozygote probabilities
Updated read count files
Estimate expected read depths
Final read count files
Principal component analysis
Estimate dispersion
PCs
Combined haplotype test
Dispersion params
Param estimates & LRT statistic

Amplification and GC Bias

Amplification Bias Correction:

- Typically remove duplicate reads with low mapping quality
- WASP removes reads at random

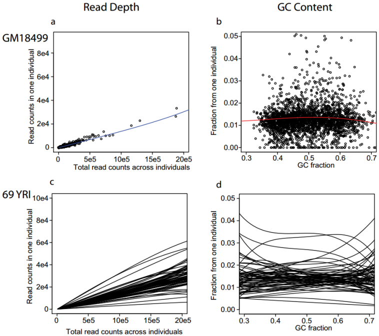
GC content:

- Affects read depth inconsistently between sequencing experiments
- Distribution of read-depths across the genome differs from experiment to experiment
- WASP corrects for these issues by fitting polynomials to the genome-wide read counts and calculating a correct read depth for each region

Correcting for GC content and other effects on expected read depth

$$\mathrm{L}\left(a_{0i}, a_{1i}, \ldots, b_{4i} \,|D\right) = \prod_{j} \Pr_{\mathrm{Pois}}\left(X_{ij} = x_{ij} \,|T^{*}_{ij}\right) \tag{1}$$

$$T^{*}_{ij} = \exp\left(a_{0i} + a_{1i}w_j + a_{2i}w_j^2 + a_{3i}w_j^3 + a_{4i}w_j^4\right)\left(b_{1i}v_j + b_{2i}v_j^2 + b_{3i}v_j^3 + b_{4i}v_j^4\right) \tag{2}$$



Estimating overdispersion parameters of allele-specific and total read-depth counts

$$\mathrm{L}\left(\phi_j \,\middle|\, D\right) = \prod_i \left[\Pr_{\mathrm{BNB}} \left(X = x_{ij} \,\middle|\, \lambda = T^*_{ij}, \Omega_i, \phi_j \right) \right]$$

$$\mathrm{L}\left(\Omega_i \,\middle|\, D\right) = \prod_j \left[\Pr_{\mathrm{BNB}} \left(X = x_{ij} \,\middle|\, \lambda = T^*_{ij}, \Omega_i, \phi_j \right) \right]$$

$$\mathrm{L}\left(\Upsilon_i \,\middle|\, D\right) = \prod_k \Pr_{\mathrm{BB-mix}} \left(Y = y_{ik} \,\middle|\, n_{ik}, p = 0.5, \Upsilon_i, \hat{H}_{ik} \right)$$

Correcting for unknown covariates using principal components

Both known and unknown covariates can confound QTL analysis

- Time of experiment
- Age of sample
- etc.

$$\lambda_{hi} = \begin{cases} 2\alpha_h(1 + c_{h1}u_{i1} + c_{h2}u_{i2} + \ldots)T_i & \text{if } G_{im} = 0 \text{ (homozygous allele 1)} \\ (\alpha_h + \beta_h)\,(1 + c_{h1}u_{i1} + c_{h2}u_{i2} + \ldots)T_i & \text{if } G_{im} = 1 \text{ (heterozygous)} \\ 2\beta_h(1 + c_{h1}u_{i1} + c_{h2}u_{i2} + \ldots)T_i & \text{if } G_{im} = 2 \text{ (homozygous allele 2)} \end{cases}$$

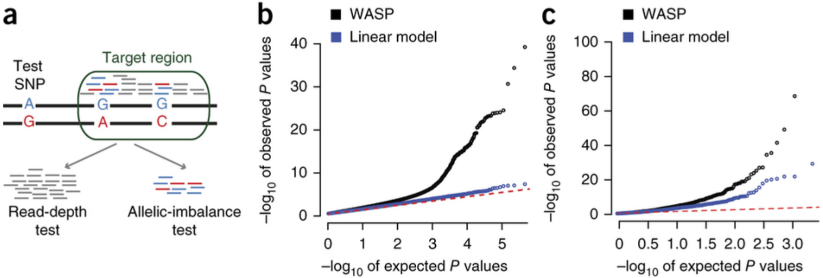
Combined Haplotype Test

Combined haplotype test (CHT) to identify cis-acting QTLs:

- allelic imbalance at phased heterozygous SNPs
- total read depth in the target region
- Combines allele-specific (beta-binomial) and read depth (beta-negative binomial) information

CHT calculates a likelihood ratio statistic: $\Lambda = \frac{L(\hat{\Theta_1}|D)}{L(\hat{\Theta_0}|D)}$ where $\hat{\Theta_0}$ and $\hat{\Theta_1}$ are the maximum-likelihood estimates of the parameters under the null hypothesis $(\alpha_h = \beta_h)$ and alternative hypothesis $(\alpha_h \neq \beta_h)$ respectively.
P-value calculated assuming $-2\log(\Lambda)$ is χ^2 distributed with one degree of freedom.

Figure 2: CHT Model and Results

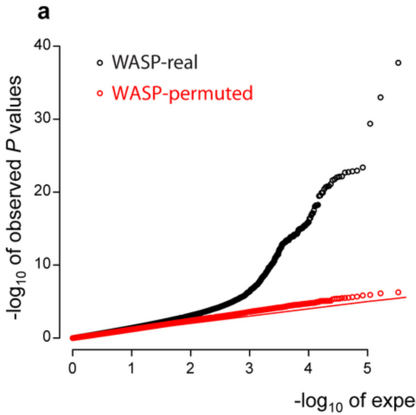


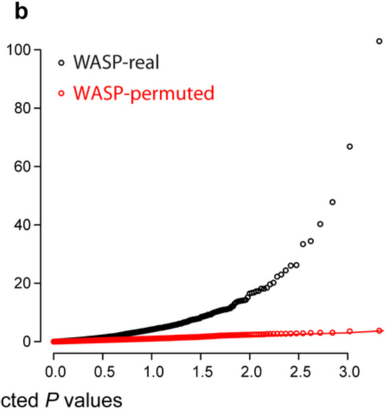
WASP identified 2,426 QTLs; linear regression did not identify any

WASP discovered 627 eQTLs; linear regression identified 446

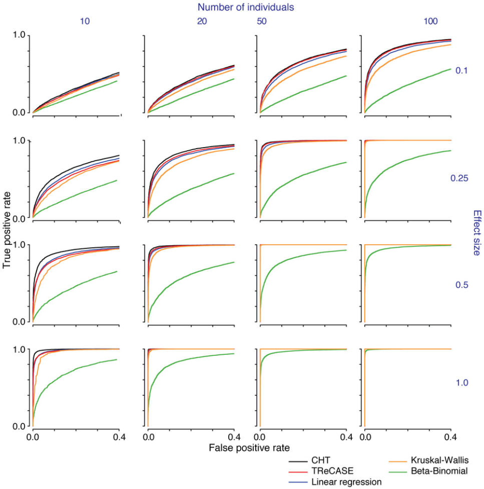
(673 vs. 617 when include five principal covariates)

Supplementary Figure 2





Supplementary Figure 3



Questions

Any questions?