Notes

Jinliang Yang

July 23, 2015

Infer mom's genotype by JRI

We have obs. mom and obs. (selfed) kids. We want to know $P(G|\theta)$, and $P(G|\theta) \propto P(\theta|G) \times P(G)$, where θ is observed data. This consists of observed genotypes (G') of both mom and kids. So: $P(G|\theta) \propto \left(\prod_{i=1}^{k} P(G'_k|G)\right) \times P(G'_{mom}|G) \times P(G)$ This function is to impute mom's genotype from a progeny array of k kids at a single locus. inferred_mom=1 -> 00, 2->01, 3->11

Imputing Founder Genotypes

$P(G|\theta) \propto P(\theta|G) \times P(G)$

$$P(G|\theta) \propto \left(\prod_{i=1}^{k} P(G'_i|G)\right) \times \left(\sum_{n=1}^{mom} P(G'_{mom}|G)\right) \times P(G)$$

This function is to impute mom's genotype by finding the maximum likelihood of $P(G|\theta)$ from a progeny array of k kids at a single locus. - Where θ denotes observed data. It consists of observed genotypes (G') of both mom and kids.
- $P(G)$ is the Hardy-Weinberg equilibrium estimated from the population.
- $P(G'_{mom}|G)$ is the error matrix estimated from the data, i.e. homozygote error = 0.02 and heterozygote error =0.6.
- $P(G'_i|G)$ is the error matrix times Mendelian segregation rate.

Phasing Founder Genotypes

$P(H|\theta) \propto P(\theta|H) \times P(H)$
$P(H|\theta) \propto \left(\prod_{i=1}^{k} P(H'_k|H)\right) \times P(H)$
$P(H|\theta) \propto \left(\prod_{i=1}^{k} \prod_{l=1}^{n} P(G'_{i,l}|H)\right) \times P(H)$

- Where θ denotes observed data.
- $P(H)$ is the probability of the haplotype for a given window size of n.
- $P(G'_{i,l}|H)$ is the probability of kid i at locus l for a given haplotype H.
- We assume all the possible haplotypes of a given window size are equally likely.

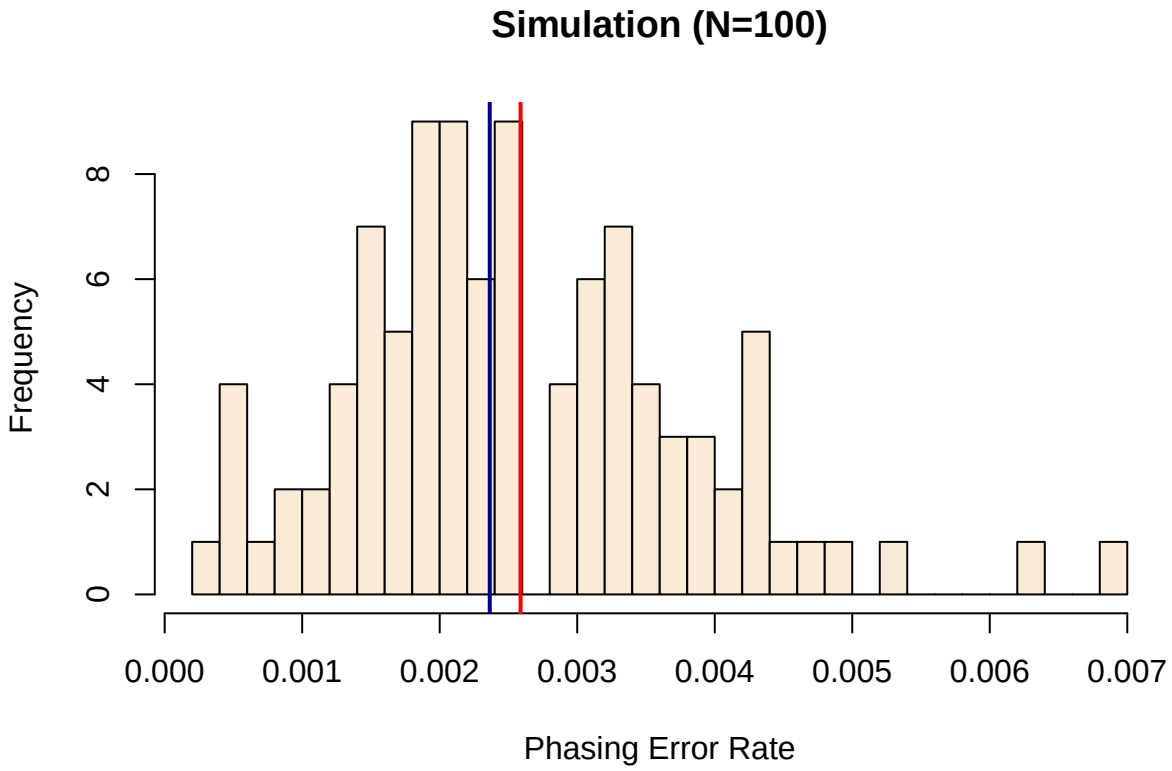
Imputing and Phasing Kids

$$P(H_k|\theta) \propto P(\theta|H_k) \times P(H_k)$$
$$P(H_k|\theta) \propto \left(\prod_{i=k} P(H'_k|H_k)\right) \times P(H_k)$$
$$P(H_k|\theta) \propto \left(\prod_{i=k} \prod_{l=1}^{n} P(G'_{i,l}|H_k)\right) \times P(H_k)$$

- Where θ denotes observed data.
- $P(H)$ is the probability of the haplotype for a given window size of n.
- $P(G'_{i,l}|H)$ is the probability of kid i at locus l for a given haplotype H.
- We assume all the possible haplotypes of a given window size are equally likely.

```
phase <- read.csv("../data/sim_phasing_res.csv")

hist(phase$er, breaks=30, main="Simulation (N=100)",col="#faebd7", xlab="Phasing Error Rate")
abline(v=mean(phase$er), col="red", lwd=2)
abline(v=median(phase$er), col="darkblue", lwd=2)
```



Phasing Dad of outcrossing progeny array

$$P(H_d|\theta) \propto P(\theta|H_d) \times P(H_d)$$
$$P(H_d|\theta) \propto \left(\prod_{i=1}^{k} P(H'_k|H_d, H_m)\right) \times P(H_m) \times P(H_d)$$

$$P(H|\theta) \propto \left(\prod_{i=1}^{k} \prod_{l=1}^{n} P(G'_{i,l}|H) \right) \times P(H)$$

- Where θ denotes observed data.
- $P(H_d)$ is the probability of the dad's haplotype for a given window size of n.
- $P(G'_{i,l}|H)$ is the probability of kid i at locus l for a given haplotype H.
- We assume all the possible haplotypes of a given window size are equally likely.
