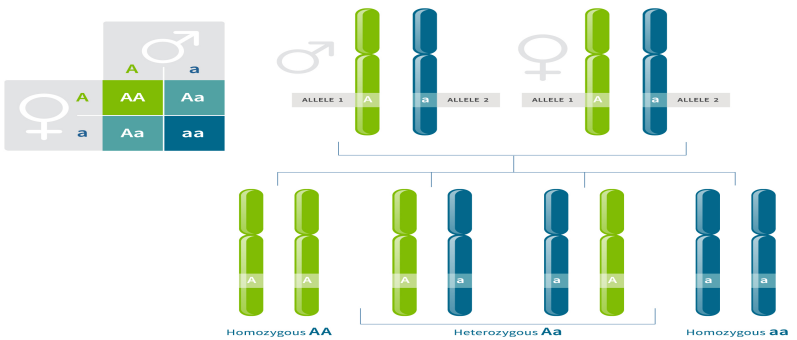
Variational Methods on Genotyping Polyploids

Tianqi Luo
Quantitative Analysis
CAS/MS'19
American University
With Professor David Gerard

Genotyping

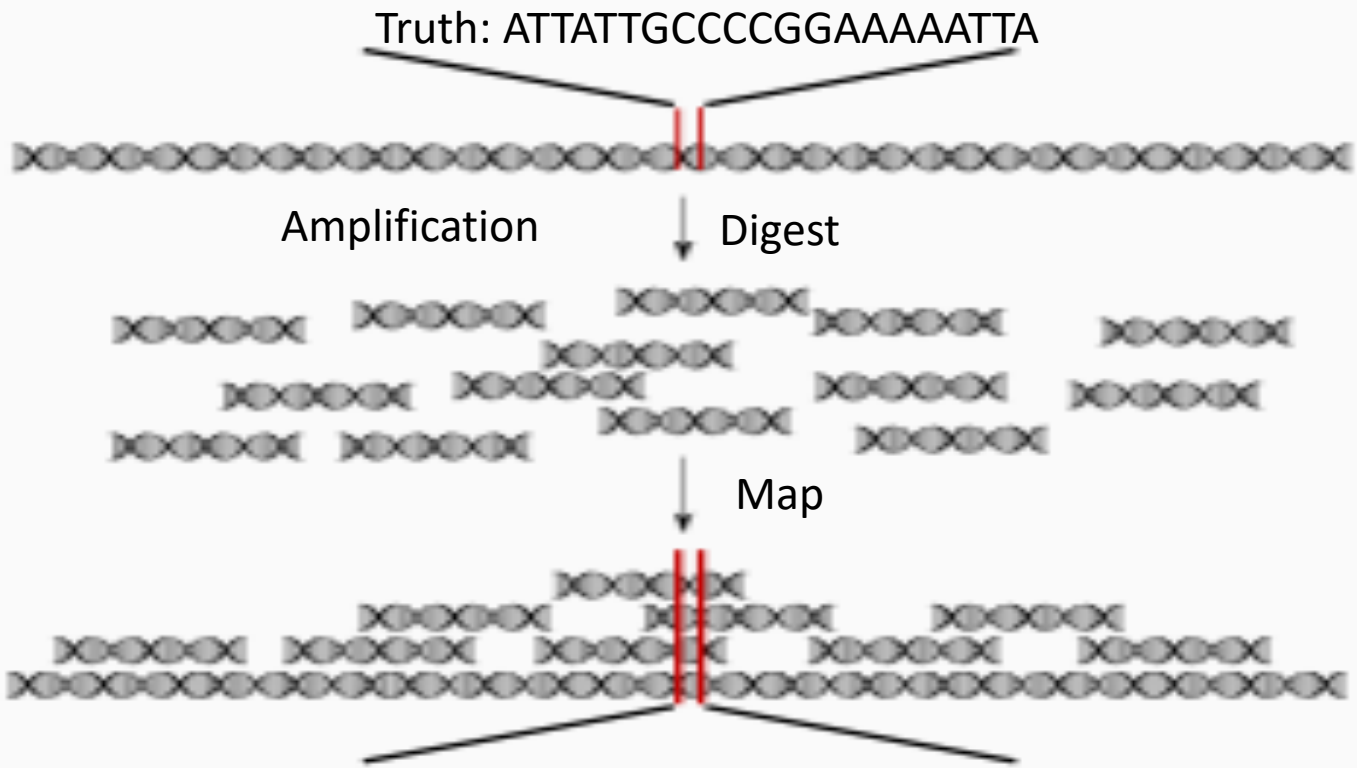
- The process of determining differences in DNA between individuals.
- Examine the individual's DNA sequence using biological analytical methods.
- Compare it to another individual's sequence or a reference sequence.



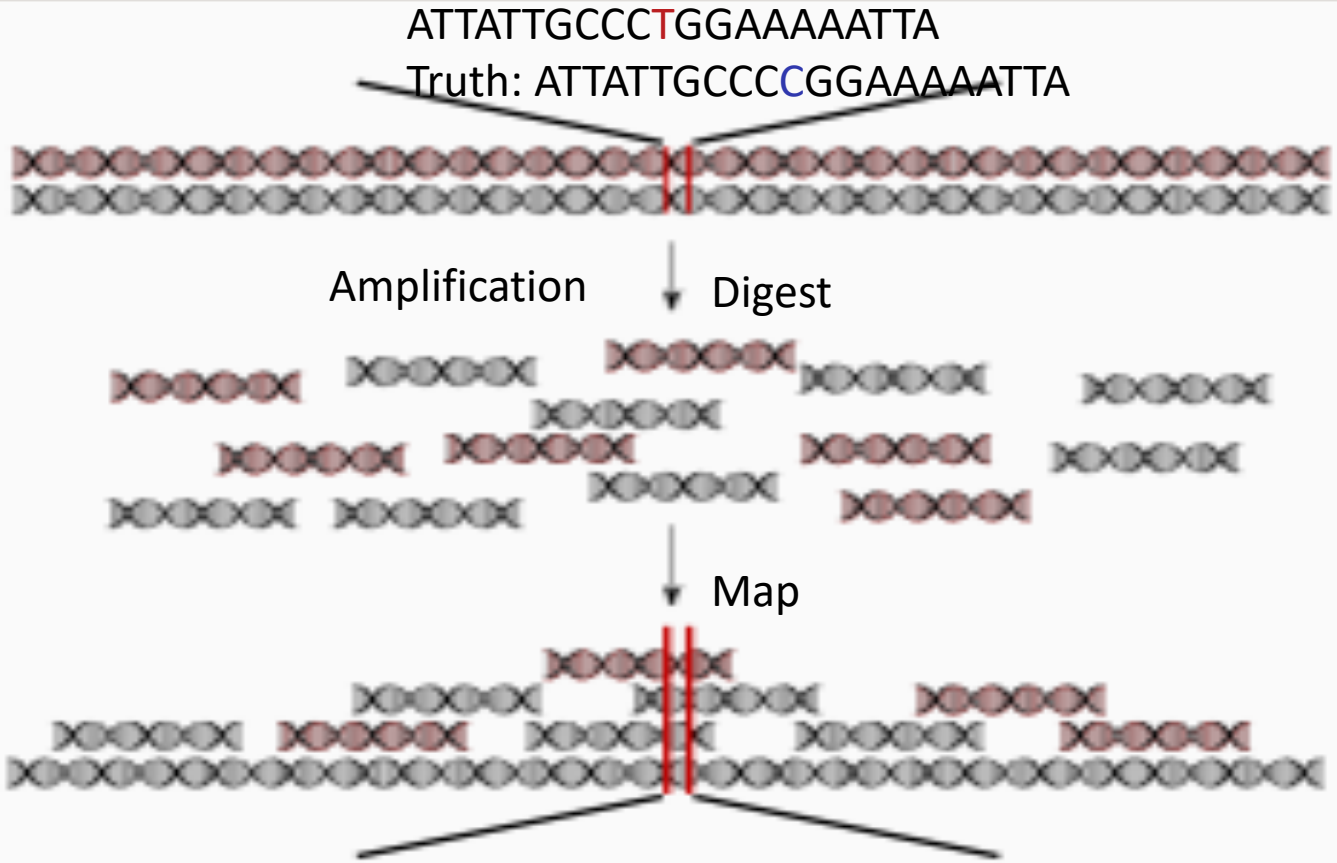
Research Topic: Genotyping By Sequencing (GBS)

- A method to discover the single nucleotide polymorphisms (SNP).
- SNPs are the most common type of genetic variation among people. Each SNP represents a difference in a single DNA building block, called a nucleotide.
- Digestion: Uses restriction enzymes to chop up the DNA.
- Copying: PCR is performed to increase fragments pool and GBS libraries are sequenced using technologies.
- Mapping: Find where the small fragments go.
- The goal is to determine whether an individual has one type of difference (allele), say “A”, or the other type of allele, say “a”.

Genotyping By Sequencing: One Genome



Genotyping By Sequencing: Two Genome

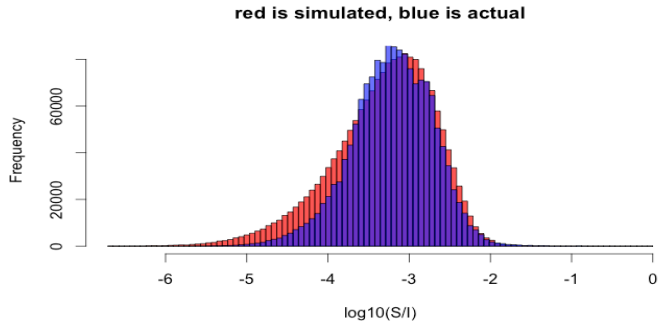


Likelihood Function From Gerard, Ferrão, Garcia, and Stephens (2018):

- $x_i \sim$Beta-Binomial $(n_i,\ \xi(p_i,\varepsilon,h),\ \tau)$,
- $\xi(p_i, \varepsilon, h) := \dfrac{f(p_i, \varepsilon)}{h\{1-f(p_i,\varepsilon)\}+f(p_i,\varepsilon)}$,
- $f(p_i, \varepsilon) := \varepsilon(1 - p_i) + (1 - \varepsilon)p_i$.

- x_i: counts of reads with reference allele for individual i.
- n_i: total counts of reads for individual i.
- $p_i \in \{0/K,1/K,...,K/K\}$: The allele dosage (proportion of individual i's genome that contains the reference allele).
- K: Ploidy of the species.
- ε: The sequencing error rate ($\sim$ 0.001).
- h: Allele bias [Pr(a after selected) / Pr(A after selected)].
- τ: Overdispersion parameter.
  - $\tau=0 \Rightarrow$Binomial; $\tau=1 \Rightarrow$get new data.

Empirical Bayes Approach



- Empirical Bayes methods are procedures for the statistical inference in which the prior distribution is estimated from the data.
- Stands in contrast to standard Bayesian methods, which the prior distribution is fixed before any data are observed.

Empirical Bayes Approach (Updog package)

- Implements empirical Bayes approach to genotype polyploids from next generation sequencing data.
- Accounts for allelic bias, overdispersion, and sequencing error.

Why Empirical Bayes could be a problem?

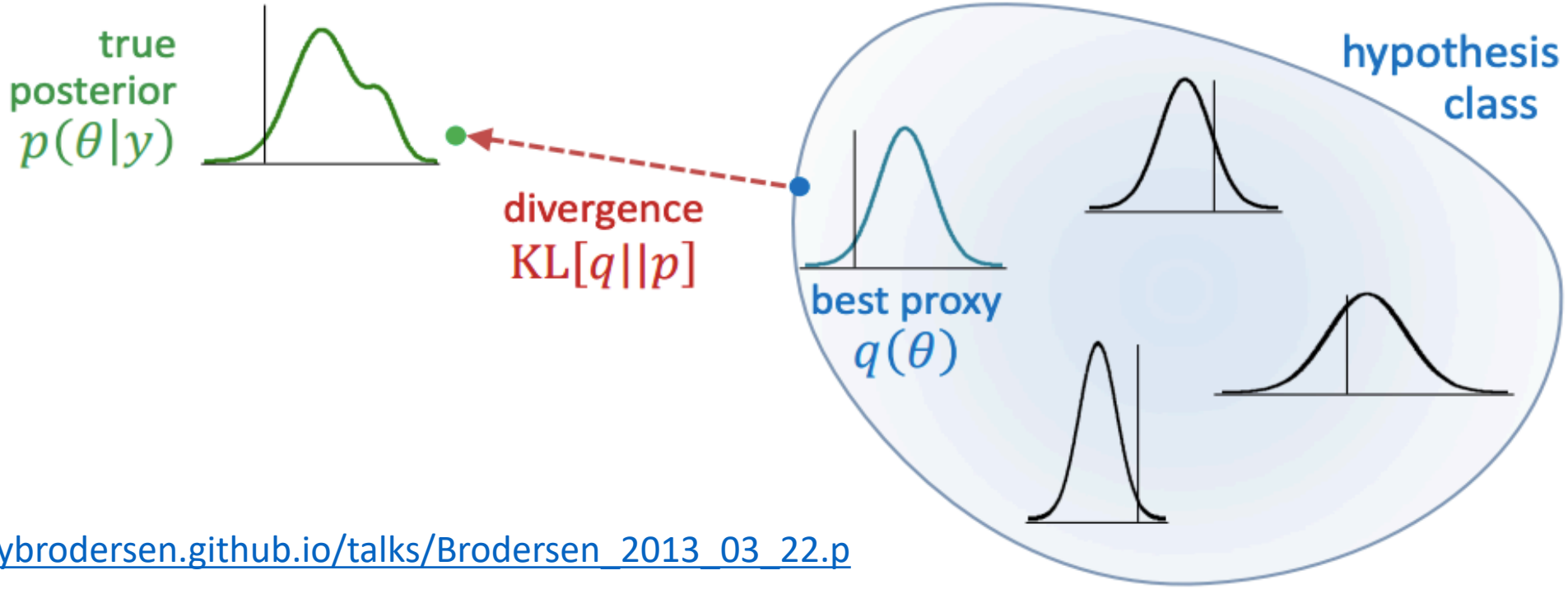
- Sometimes we don’t have enough data to observe, hence there is not enough to estimate the prior.
- Doesn’t estimate the prior well when the model is complex.

Variational Methods Approach

- Typically used in complex statistical methods.
- Derives a lower bound for the marginal likelihood of the observed data(marginal probability of the data given the model).
- Takes an approach to statistical inference over complex distributions that are difficult to estimate.
- Provides a locally-optimal, exact analytical solution to an approximation of the posterior.

Variational Bayes Methods

- Find an approximate density that is maximally similar to the true posterior by using K-L Divergence.
- K-L Divergence: Measure of how one probability distribution is different from a second, reference probability distribution.



https://kaybrodersen.github.io/talks/Brodersen_2013_03_22.pdf

Stan Language

- **Stan** is a probabilistic programming language for statistical language written in C++. The Stan language is used to specify a (Bayesian) statistical model by calculating the log probability density function.

- We'll be using Stan in our research.

Stan Implementation in R (Rstan package)

```
// saved as 8schools.stan
data {
  int<lower=0> J;         // number of schools
  real y[J];              // estimated treatment effects
  real<lower=0> sigma[J]; // standard error of effect estimates
}
parameters {
  real mu;                // population treatment effect
  real<lower=0> tau;      // standard deviation in treatment effects
  vector[J] eta;          // unscaled deviation from mu by school
}
transformed parameters {
  vector[J] theta = mu + tau * eta;        // school treatment effects
}
model {
  target += normal_lpdf(eta | 0, 1);       // prior log-density
  target += normal_lpdf(y | theta, sigma); // log-likelihood

}
```

Goals of our simulation studies

- Simulate genotypes using functions from updog package.
- Estimate genotypes from Empirical Bayes Approach using updog function flexdog().
- Estimate genotypes from Variational Bayes Approach by using vb() function in package rstan. Write a standog() function that estimates genotypes for variational Bayes methods.
- Compute the misclassification errors for both approaches by comparing the simulated genotypes with the ones estimated from both methods.
- If Variational Bayes Method works better, build an R package that estimates genotypes using Variational Bayes Method.

Simulation Study (Part 1)

- Write the likelihood function model in Stan.

```
parameters {
  real<lower=0,upper=1> alpha;
  real<upper=0> logit_eps;
  real<upper=0> logit_tau;
  real log_h;
}

// logpostprobmat: logpostprobmat[i, k + 1] is the log of the *unnormalized*
//                 posterior probability that individual i has genotype k.
transformed parameters {
  matrix[N, K + 1] logpostprobmat;
  for (i in 1:N) {
    for (k in 0:K) {
      real logprobk = binomial_lpmf(k | K, alpha);
      real epsilon  = inv_logit(logit_eps);
      real tau      = inv_logit(logit_tau);
      real h        = exp(log_h);
      real p        = (k * 1.0) / K;
      real fi       = p * (1.0 - epsilon) + (1.0 - p) * epsilon;
      real xii      = fi / (h * (1.0 - fi) + fi);
      real alpha_bb = xii * (1.0 - tau) / tau;
      real beta_bb  = (1.0 - xii) * (1.0 - tau) / tau;
      logpostprobmat[i, k + 1] = logprobk + beta_binomial_lpmf(x[i] | n[i], alpha_bb,
    }
  }

}
```

```
model {
  // priors
  logit_eps ~ normal(-4.5, 1) T[, 0]; // these upper bounds help identify the model
  logit_tau ~ normal(-5.5, 1) T[, 0];
  log_h     ~ normal(0, 1);
  alpha     ~ uniform(0, 1);

  // likelihood. Integrate out mixing indicators.
  for (i in 1:N) {
    target += log_sum_exp(logpostprobmat[i]);
  }
}

generated quantities {
  matrix<lower=0,upper=1>[N, K + 1] postprobmat;

  for (i in 1:N) {
    postprobmat[i] = softmax(logpostprobmat[i]')';
  }
}
```

Simulation Study (Part 2)

- Simulate the data and the parameters.

```
itermax <- 100
bias_vec <- c(1)
seq_vec  <- 0.01
od_vec   <- c(0)
itervec  <- seq_len(itermax)
ploidy_vec <- 6
allele_vec <- c(0.5, 0.9)
nsamp_vec <- c(30)            ## Simulate dataframes with these parameters
recount_vec <- c(100)
paramdf <- expand.grid(bias = bias_vec,
                       seq = seq_vec,
                       od = od_vec,
                       iter = itervec,
                       allele = allele_vec,
                       nsamp = nsamp_vec,
                       recount = recount_vec,
                       ploidy = ploidy_vec)
```

Simulation Study. (Part 3)

- Simulate the genotypes by rgeno().
- Run the model with Empirical Bayes approach(flexdog()) and extract the estimated genotypes.
- Write a standog() function to run the Variational Bayes method, and extract the estimated genotypes.

```
genovec <- rgeno(n           = nsamp,
                 ploidy      = ploidy,
                 model       = "hw",          ## Use rgeno to simulate genovec
                 allele_freq = allele_freq)
```

```
sizevec = rep(sim_list$recount, sim_list$nsamp)   ## Simulate sizevec


refvec <- rflexdog(sizevec = sizevec,
                   geno    = genovec,           ## Simulate the refvec
                   ploidy  = ploidy,
                   seq     = seq,
                   bias    = bias,
                   od      = od)

uout <- flexdog(refvec = refvec, sizevec = sizevec, ploidy = ploidy, model = "hw")
```

```
standog = function(refvec, sizevec, ploidy) {
  vbout <- rstan::vb(object = vbmodel, data = list(K = ploidy, N = nsamp, x = refvec

  postprobmat <- lgeno(vbout)

  stan_geno <- get_maxgeno(postprobmat)

  return(list(postprobmat = postprobmat, geno = stan_geno))

}        ## Write the standog function


sout = standog(refvec = refvec, sizevec = sizevec, ploidy = ploidy)

standog_geno = sout$geno      ## Extract genotypes from updog and standog
```

Simulation Study (Part 4)

- Compare the genotypes estimated from Empirical Bayes and Variational Bayes with genotypes simulated from rgeno.
- Calculate the misclassification error rate.

```
stan_classification_error = mean(genovec != standog_geno)

updog_classification_error = mean(genovec != updog_geno)
```

Conclusion

- Successfully coded likelihood function into Stan.
- Successfully simulated the parameters and the genotypes.
- Successfully wrote a standog function that uses Variational Bayes methods to estimate genotypes.
- Successfully calculated the misclassification error rate of a particular simulated case.

Future Goals

- Summarize the overall classification error rates for both approaches to decide which is better.
- If Variational Bayes Methods is better, build an R package that uses Variational Bayes methods.

Sources

- https://en.wikipedia.org/wiki/Genotyping
- https://cran.r-project.org/web/packages/updog/index.html
- https://en.wikipedia.org/wiki/Variational_Bayesian_methods
- https://www.genetics.org/content/210/3/789

Thank you!!