Genotype-free individual genome reconstruction of Multiparental Population Models by RNA sequencing data

Kwangbom Choi[1], Hao He[1], Daniel M. Gatti[2], Vivek M. Philip[1], Narayanan Raghupathy[1], Steven C. Munger[1], Elissa J. Chesler[1], Gary A. Churchill[1,*]
[1]The Jackson Laboratory, 600 Main Street, Bar Harbor, ME, 04609
[2]Computer Science, College of the Atlantic, 105 Eden Street, Bar Harbor, ME, 04609

Abstract. Transcriptome profiling by RNA sequencing (RNA-Seq) of genetically segregating populations is widely used to investigate the regulatory programme of gene expression and its downstream phenotypic consequences. We address analytical challenges of RNA-Seq experiments in multi-parent populations (MPPs) that are derived from two or more inbred founder strains. Genotyping arrays or low-coverage DNA sequencing are commonly used to detect founder strain variants and to reconstruct the founder haplotype mosaic of MPP individuals. We propose and evaluate a new method, Genome reconstruction by RNA-Seq (`GBRS`), that simultaneously reconstructs individual diploid genomes and quantifies total and allele-specific expression directly from RNA-Seq data. We demonstrate that there is sufficient information in the genetic variants revealed in RNA-Seq data to reconstruct MPP haplotypes. Unlike existing RNA-seq-based genotyping methods, `GBRS` exploits multi-way allele-specific expression and jointly use closely neighboring variants. `GBRS` haplotype reconstructions have accuracy comparable to array-based reconstructions. `GBRS` provides RNA-Seq quantification that is tailored to individual genomes and does not suffer from biases that can arise when using reference genome alignment. `GBRS` also provides a quality control for detecting sample mix-ups and can improve power to detect expression quantitative trait loci. `GBRS` software is freely available at https://github.com/churchill-lab/gbrs.

Keywords: genome reconstruction, genotyping arrays, RNA-Seq, expression quantitative trait loci, allele-specific expression.

* Gary.Churchill@jax.org

INTRODUCTION

RNA sequencing (RNA-Seq) has revolutionized our understanding of gene expression in whole tissues and in single cells [Stark et al., 2019]. While generally used for profiling gene expression, RNA-Seq data can identify single nucleotide polymorphisms (SNPs) and small insertion and deletions (indels) in the transcribed portion of the sample genome [Piskol et al., 2013]. In addition, RNA-Seq data can reveal spontaneous mutations [Miller et al., 2013], RNA editing events [Gu et al., 2016], and allele-specific expression [Wittkopp et al., 2004] in recent development of computational approaches.

In typical RNA-Seq analysis pipelines, transcript abundance is quantified by counting reads that align to a reference transcriptome. Alignment algorithms can make allowance for mismatches due to sequencing errors or polymorphisms between the sample RNA and the reference transcript sequences. Nonetheless, reference alignment can lead to biases in RNA-Seq quantification that can be minimized when the alignment target is tailored to reflect the genetic makeup of the sample [Degner et al., 2009, Munger et al., 2014]. This requires that the individual sample genomes are known *a priori* or inferred from additional data such as genotyping arrays [Yang et al., 2009] or low-pass DNA sequencing [Gileta et al., 2020].

Multi-parent populations (MPPs) are model-organism genetic reference populations that are descendant from two or more inbred founder strains [de Koning and McIntyre, 2017]. The genome of each individual in an MPP represents a mosaic of the founder strain genomes, which for many MPPs have been fully sequenced. Genotyping is an important step in genetic analysis of MPPs and founder strain variants detected with genotyping arrays [Morgan et al., 2016] or short-read DNA sequencing [Parker et al., 2016] can be used to reconstruct the founder haplotype mosaic of individuals [Gatti et al., 2014]. We have previously demonstrated that alignment of RNA-Seq reads to individual imputed genomes or transcriptomes can substantially improve the results of expression quantitative trait locus (eQTL) mapping [Munger et al., 2014]. Likewise, if the MPP individuals are heterozygous, we can quantify allele-specific expression (ASE) of genes using an individual-specific transcriptome with representation of the two allelic copies of each predicted transcript, including the known sequence variants that distinguish the alleles. For the quantification of ASE, it is critical to use a weighted allocation read counting method to resolve reads that align equally well to be both allelic copies of a gene and potentially to other transcripts in the transcriptome [Raghupathy et al., 2018].

In any given tissue, thousands of expressed genes are distributed across the genome. Therefore, variants in expressed genes that are detected by RNA-Seq could replace other methods of variant detection such as genotyping arrays. We examine how variant information in the RNA-Seq reads can be used to construct an individual sample alignment index. Some challenges are to be expected. Constructing diploid alignment indices of each individual in an eQTL mapping study as recommended by [Munger et al., 2014] can be a computationally demanding and error-prone task. It may be difficult to uniquely determine the founder origin of RNA-Seq reads due to limited information in the observable variants. Knowledge of the founder variants may be imperfect and as a results reads may align to the wrong founder haplotypes.

We propose a novel solution to these challenges that employs a fast ASE quantification algorithm [Raghupathy et al., 2018]. Instead of building a diploid alignment index for every individual in an MPP, we resort to a novel *multi-way* alignment strategy in which we align reads to a pooled transcriptome of all MPP founder strains incorporating the known genetic variants that define founder haplotypes. This approach requires RNA-Seq data from founder strains in order to learn how reads align with equal quality to multiple founder haplotypes. We utilize the frequency of read alignments from founder-derived RNA-Seq data to define the statistical distribution of read alignments for a given haplotype. We then align the MPP sample data to the same multi-way index and use the alignment frequencies to identify the local haplotypes of the sample genomes. Individual genes may have insufficient sequence variation to uniquely identify a haplotype. Therefore, we formulated a Hidden Markov Model (HMM) that combines information across neighboring genes for haplotype reconstruction. We then convert the multi-way read alignment to a diploid haplotype reconstruction and estimate allele-specific read counts for each individual genome. In this work we demonstrate that the information in RNA-Seq read align-

ments is sufficient to compute accurate haplotype reconstructions. The process is implemented in an open-source Python package, GBRS (Genome reconstruction By RNA-Seq), available at https://churchill-lab.github.io/gbrs/.

RESULTS

Overview of the GBRS algorithm

The objectives of GBRS are 1) to reconstruct the founder haplotype mosaic of MPP individuals directly from RNA-Seq data and 2) to quantify total and allele-specific gene expression based on individual MPP genomes. The importance of using an individual genome or transcriptome to quantify RNA-Seq data has been demonstrated [Munger et al., 2014] but the method proposed requires the construction of individual alignment indices which is computationally burdensome and potentially error prone. Instead, GBRS algorithm employs a single alignment index built with haplotype-specific transcript sequences for the MPP founder strains. We hypothesized that simultaneous alignment of RNA-Seq reads to all of the founder haplotypes would provide enough information to identify the haplotype(s) present in the target sample for most genes. For genes where the founder origin is not fully resolved, we could combine information across neighboring genes using an HMM.

The GBRS algorithm requires *training* based on RNA-Seq data obtained on a set of samples including each of the founder strains. For each founder strain sample, we quantify the distribution of reads that best align to the multi-way alignment index. As many reads align to multiple founder alleles depending on the density of polymorphisms, we use weighted allocation [Raghupathy et al., 2018]. The pattern of read alignments from the founder-derived RNA-Seq data will be used to identify the haplotype origin of genes in MPP individuals (Figure 1). This

unique feature of GBRS distinguishes it from other sequencing-based genotyping methods that rely on accurate variant calling.

[Figure 1 about here.]

To quantify total and allele-specific gene expression in each MPP individual, we first reconstruct its imputed diploid genome using the same multi-way alignment index. We estimate the proportions of reads that align to the founder transcript sequences (Figure 2a) just like we did for founder-derived RNA-Seq data in the training stage. The HMM treats these proportions as observed data. GBRS then estimate the emission probabilities of the HMM by comparing the multi-way read distributions with the founder-derived RNA-Seq alignment data. We assume heterozygotes to be an equal mixture of the two founders that make up the local diplotype. The transition matrix of the HMM can be estimated from approximate genetic map distances between genes [Broman, 2012] in the target MPP population (Figure 2b). For genes that are not fully informative, the HMM can correct the missing information by smoothing across neighboring genes. GBRS runs the forward-backward algorithm and estimate genotype probabilities given multi-way read alignment observations across all the expressed genes. Then GBRS executes a Viterbi algorithm to call the best diplotype state on each gene including the ones that are not expressed (Figure 2c). We can also pull the marginal state probabilities from the HMM to compute genotype probabilities that represent the uncertainty in the genotype classification of each gene. Finally we apply the Viterbi genotype calls to the multi-way alignment incidence profile in order to convert it to a diploid alignment incidence, for which we perform weighted allocation [Raghupathy et al., 2018] once more to quantify ASE (Figure 2d). Because we don't realign our reads to the individual imputed transcriptome, GBRS can save the processing time.

[Figure 2 about here.]

An application of the GBRS algorithm

We obtained RNA-Seq data on liver tissue from 16 mice (both sexes) from each of the eight Diversity Outbred (DO) founder strains. We constructed an eight-way transcriptome alignment index with the transcript sequences from each haplotype (see Methods). We trained the emission probability model of the HMM using eight-way read distributions estimated with EMASE [Raghupathy et al., 2018] after aligning the reads with `bowtie1` (see Methods). We obtained the transition probabilities of the DO for every breeding generation using `DOQTL` [Gatti et al., 2014] version 1.0.0.

Next, we applied `GBRS` to RNA-Seq data from liver tissue samples obtained from 482 Diversity Outbred mice from generations 4 to 11 (G4∼G11) [Svenson et al., 2012]. An equal number of females and males were included in the study cohorts. The liver RNA-Seq data is archived at the Short Read Archive under project number PRJNA35625. The samples were single-ended data on a HiSeq2000 producing 100bp reads at a depth of 20M reads per sample. We genotyped 282 samples with MUGA and 200 samples with MegaMUGA (Geneseek, Lincoln, NE).

We also applied `GBRS` to striatum tissue samples from 416 DO mice of both sexes spanning generations G21, G22, and G23 [Philip et al, in preparation]. Briefly, RNA samples were ribo-depleted with the RNA RiboErase Stranded cDNA library (Kappa Biosystems). The samples were paired-end sequenced on a HiSeq2000 producing paired 100bp reads at a depth of 50M read pairs per sample with the External RNA Controls Consortium (ERCC, Ambion) spike-ins added. All 416 samples were genotyped with GigaMUGA array.

GBRS accurately detects recombination events

Haplotype reconstruction is a critical first step for gene expression profiling with MPP models (Figure 3 for a typical DO mouse example). As it enables us to jointly use variants of neighboring loci, it helps an accurate imputation of missing markers for association mapping studies. In addition, it can correct sample mix-ups more reliably [Javed et al., 2020].

[Figure 3 about here.]

We first assessed whether it is viable to use expressed genes as markers for detecting recombination events as implemented in `GBRS`. We compared the number of recombinations detected by `GBRS` with that of MUGA, MegaMUGA, and GigaMUGA genotyping arrays. Each of these genotyping arrays have 7.5k, 77.8k, and 150k SNP probes respectively. We typically detect 10k to 12k expressed genes after filtering out genes that have abundance below 1 TPM. We find that the number of recombination events identified by `GBRS` is comparable to the counts from the haplotype reconstruction by GigaMUGA, the array that has the highest marker density (Figure 4). This implies that the number and positional distribution of expressed genes are sufficient to detect most of the recombination events for the given outbreeding generations.

[Figure 4 about here.]

GBRS-based individual imputed genome are consistent with those of genotyping arrays

We evaluated the concordance of `GBRS`-based haplotype reconstruction to the array-based results using the DO datasets. We first converted the 36-state genotype probabilities to 8-state founder haplotype probabilities, which are later also used for eQTL mapping as in [Gatti et al., 2014]. Next we interpolated 8-state haplotype probabilities onto a common 69,000 equi-distance genomic

loci. We then computed Pearson correlations between GBRS-based haplotype probabilities and the result from each of the genotyping array on this common grid. We find that ∼90% of samples have Pearson correlation $r > 0.8$ (Figure 5a). The median correlation coefficient was 0.876, 0.862 and 0.834 between GBRS and the MUGA, MegaMUGA, and GigaMUGA arrays respectively. We identified 29 MUGA, 21 MegaMUGA, and 39 GigaMUGA samples that have no correlation ($r \approx 0$). We find that there are sample swaps: RNA-Seq and genotyping array data are incorrectly paired among these samples. We corrected the sample mix-ups easily by repairing the data that have the highest correlation on haplotype probabilities. For three of these samples in total, the GBRS haplotype probability did not match any array.

[Figure 5 about here.]

GBRS improves the accuracy of genotyping at the loci of expressed genes

If the genotype calls (or genotype probabilities) are more accurate, the original signal of association would be less obscured by mis-classified (or mis-genotyped) outcome values, which then leads to a higher LOD score in QTL analysis. Thousands of genes are under local mode of regulatory control [Chick et al., 2016, Aguet et al., 2017], and therefore, we can evaluate which method reconstructs individual imputed genomes more accurately by comparing the LOD scores of local eQTLs. We performed a likelihood ratio test (LRT), for each genotyping method, to assess how much the goodness of fit improves from the null model using the estimated genotype probability at the start site of each gene. We then compared from top to the 70th percentile of LODs, genes whose expressions are all highly associated with the genotype probability at their own loci. We find that GBRS consistently exhibits higher LODs on all of these local eQTLs and outperforms all three genotyping arrays (Figure 5b).

When dealing with a large number of samples, it seems that sample mix-ups are inevitable. For example, Broman et al. report as high as 18% of samples are involved in mix-ups [Broman et al., 2015]. Using GBRS these errors are easily corrected but it might take some additional work to determine of the mix-ups occurred in the RNA samples or in the DNA samples used for genotyping. We find that fixing sample mix-ups substantially improves the power of eQTL mapping. For example, the median LOD of top 30% of local eQTLs increased from 14.82 to 17.98 after correcting sample mix-ups for the DO mice measured with MUGA array. We observe similar improvement for the mice measured with MegaMUGA or GigaMUGA arrays: increasing from 11.06 to 13.63 or from 16.79 to 19.81 respectively (Figure 5b). We also find that, even though the correction of sample mix-up improves the intensity of association signals in all three genotyping arrays, GBRS still outperforms them. Therefore we recommend running GBRS to check for and correct sample mix-ups between genotyping array and RNA-Seq data [Broman et al., 2019, Javed et al., 2020] even when both RNA-Seq and genotyping array data are available.

The quality of genomes reconstructed by GBRS is comparable to that of high-density genotyping arrays

In order to compare the accuracy of imputed genomes more directly between genotyping methods, we also performed a full eQTL analysis after correcting sample mix-ups in the array datasets (Methods). We compared the LOD scores of both local and distal eQTLs (Figure 6) identified with GBRS haplotype probabilities to those of MUGA (top row), MegaMUGA (middle row), and GigaMUGA (bottom row) arrays. Panels on the left and right columns are comparisons of local and distal eQTLs respectively. For the local eQTLs, we find 5,910, 3,673, and 8,051 genes have higher LODs using GBRS, whereas MUGA, MegaMUGA, and GigaMUGA give higher LODs

in 1,662, 2,238, and 2,680 genes respectively. The results reassure the conclusion in the previous section (Figure 5b) that GBRS provides overall larger LODs for local eQTLs than all the genotyping arrays. Similarly, for the distal eQTLs, we find 2,850, 3,347, and 3,807 genes have higher LODs with GBRS, while MUGA, MegaMUGA, and GigaMUGA show higher LODs in 2,153, 3,447, and 2,027 genes respectively. We speculate that strong local signals of founder origins reflected in the estimated allelic expression are highly effective on detecting most recombination events, which then leads to more accurate genome reconstruction.

[Figure 6 about here.]

DISCUSSION

GBRS pipeline utilizes multi-way ASE profiles instead of SNPs as markers to reconstruct individual imputed genomes. What makes GBRS unique is that it leverages multi-way ASE profiles from founder-derived RNA-Seq data in order to manage ambiguity due to variants shared by multiple founders. We build and refer to a catalogue of multi-way ASE profiles of founder genes when we determine the genotype of an expressed gene in each MPP individual. The catalogue immediately provides crucial information to evaluate whether a gene is homozygous for a particular founder haplotype. We also show it is reasonable to approximate multi-way ASE of heterozygous genes by combining founder ASE profiles, admitted that it would be the best to use additional sequencing data from F1 hybrids between founders. Multi-way ASE catalogue is dependent only upon transcript sequences of founder haplotypes, and therefore, largely agnostic of experimental settings such as tissue type. This makes the catalogue reusable for any RNA-Seq data from the same target MPP model. We provide the catalogue for the DO at JAX ftp site (See Software Availability) so users don't have to train the HMM model by themselves again.

GBRS is fast, substantially reducing the computational procedure and saving processing time as well as storage burden. The core analysis step of GBRS is built around a fast ASE quantification method [Raghupathy et al., 2018] which is what we would perform anyway with RNA-Seq data to evaluate gene expression or identify eQTLs. GBRS only requires a one time execution of multi-way alignment for individual genome reconstruction and ASE estimation. If trained HMM parameters are available like the DO, we can process each sample independently in parallel and benefit from high performance computing resource such as cluster systems. Starting from multi-way alignment incidence matrix (in emase format), it took $\sim$7 minutes to process all the 478 DO liver samples on our PBS/Torque cluster (total of 904 cores composed of 22 HP Proliant SL/XL Series servers with 2.6GHz CPUs and 256 or 512gb RAMs). GBRS uses a same single multi-way alignment index for all the individuals in an MPP model instead of building a large number of individualized read alignment indices as in Munger et al. [Munger et al., 2014]. The size of multi-way index for the DO is only 3.2gb while that of individualized indices is 787gb in total for the 478 DO samples.

A possible caveat in GBRS analysis would be that, depending on tissues, there could be genomic region that are sparse in expressed genes. This could lead to missing recombination events or missing distal eQTLs although we did not observe any remarkable impact in the data sets we examined (Figure 4 and Figure 6b, 6d, and 6f). This should not affect the local-eQTLs since GBRS relies on direct signals from coding regions. Therefore, if precise reconstruction of genomes including gene sparse regions matters, we advise to use genotyping arrays. Even in case array data is available, GBRS could be still very useful as we can use it to fix sample mix-ups, or to combine genotype probabilities to supplement each other.

There is still a room for further improvement. We would be able to improve the quality of multi-way ASE catalogue using more founder data from different tissues. As each tissue often has

distinct set of genes expressed, more genes would be added to the founder ASE catalog. We could also add ASE profiles from F1 hybrid mice derived from the founder strains. In case no founder data are available, we would be able to derive ASE profiles directly from MPP data using Baum-Welch algorithm once we get data from a large number of MPP individuals and obtain sufficient sample size for all the possible genotypes. We are also going to adapt `GBRS` algorithm to different type of sequencing data, e.g. low-coverage DNA-Seq, ChIP-Seq, ATAC-Seq, or Hi-C. Combining genotype probabilities of multiple sequencing modalities (Methods) would not only improve the quality of reconstructed genomes but also enable allele-aware integrative analysis. As there is no guarantee that all recombination events happen at the intergenic regions, we could perform `GBRS` analysis on exons as markers instead of on genes. This would increase the resolution of detecting recombination events and offer an opportunity to identify genotypically chimeric genes and investigate their functional consequences.

METHODS

Our goal is to predict the diplotype state of each gene in each sample based on how reads align when they are exposed to all the alleles of founder strains simultaneously. We call this alignment strategy as multi-way alignment. Creating individualized diploid alignment indices [Seqnature paper] for hundreds of samples requires a large amount of tedious data processing. In addition, it does not represent uncertainty in genotypes when aligning reads. multi-way alignment strategy gets around these weaknesses as we can use a single alignment index that refers to all the alleles from the founder strains. RNA-seq produces tens of millions of reads per sample, and, although they are usually short, many of them contain information of their original founder strains. Ideally, we should be still able to tell the genotype of each genes by profiling how many reads align most

to a single strain if homozygous or two if heterozygous with multi-way alignment strategy. The challenge is that, in reality, not all the genes are enriched with variants that segregate one allele from another, and the alignment profile of a single gene may not be sufficient to determine the diplotype. In these cases, we want to refer to the neighboring genes and help our prediction because in a close neighborhood it is less likely that multiple recombination events occur. HMM applies very well to this situation as it allows us to predict diplotype state of genes by combining our observations at each gene with information from neighboring genes.

Building read alignment index

We create custom genomes of the eight founder strains by introducing strain-specific SNPs and short indels (Mouse Genome Project [Keane et al., 2011] release v5 available at `ftp://ftp-mouse.sanger.ac.uk/REL-1505-SNPs_Indels/`) to the reference genome (Genome Reference Consortium Mouse Build 38 available at `http://ftp.ensembl.org/pub/release-84/fasta/mus_musculus/dna/`) using `g2gtools` (`http://churchill-lab.github.io/g2gtools`). We also adjust the coordinates of the reference gene annotation (Ensembl Release 84 [Yates et al., 2015] available at `http://ftp.ensembl.org/pub/release-84/gtf/mus_musculus/`) and extract strain-specific transcripts with `g2gtools`. We collate these transcript sequences in a single fasta file, add 99 bp-long poly-A tail to each transcript, and then run `bowtie − build` command to build a bowtie v1.0.0 alignment index of the pooled transcriptome.

Profiling multi-way allele expression for each sample

We align RNA-seq data from each MPP individual to the multi-way alignment index using `bowtie1` with '`all`', '`best`', '`strata`' options. Our raw observation on a transcript t is an alignment profile

for each read i where $1 \leq i \leq N$. We find that only ∼2% of the reads were uniquely aligning to a single haplotype. Many reads map to multiple haplotypes and/or multiple genes. In some extreme cases, a read will align to all the founder transcripts when a read is sampled from a region of a gene that has no variant.

Suppose $x_{t,i}$ denotes the true haplotype origin of a read i that aligns to a gene t where $1 \leq t \leq M$. $x_{t,i}$ is an incidence vector of length H, the number of founder strains.

$$\begin{aligned} P(x_t|s_t) &= \prod_{i}^{N} P(x_{t,i}|s_t) \\ &= \prod_{i}^{N} \sum_{h} P(x_{t,i}|h)P(h|s_t) \\ &\cdots \\ &= \sum_{h} P(x_t|h)P(h|s_t) \end{aligned} \tag{1}$$

Now we transform our raw observation $x_t = \{x_{t,1}, x_{t,2}, \cdots, x_{t,N}\}$ to $y_t = E(x_t)$ using EMASE [Raghupathy et al., 2018] and we will use $Y_1^T = \{y_1, y_2, \cdots, y_T\}$ as an input for our HMM model.

Cataloguing alignment specificity of founder strain data

We also profiled how specifically reads of founder strains align back to their original haplotypes if they are exposed to all the founder alleles simultaneously. The more the variants among alleles, the more specific the alignment would be to its own genetic background. Cataloguing alignment specificity of the parental strain data is important as we expect the equivalent multi-way alignment specificity for the alleles of the randomized outbred population. We can also derive the alignment specificity for the reads generated from heterozygous genes when no imbalance in allelic expression is assumed. We aligned RNA-seq data from the liver tissue of each founder strain that are

composed of 16 biological replicates per strain to the pooled transcriptomes of all eight founder strains. As reads are short (100 bp) and many reads align to multiple haplotypes, we estimated the expected allelic read counts using EMASE.

HMM model

HMM is frequently used when 1) our observations are ordered, for example, along the time of measurement or along the genes in a chromosome; 2) our observations are generated by one of the finite numbers of possible states in the system; and 3) we are more interested in predicting the sequence of states given our observations rather than summarizing our observation. As we are trying to identify unknown genotype state of each gene in the chromosomes based on RNA-seq alignment profile over all the possible alleles, we formulated it as an HMM. In the following sections, we describe the full HMM and our reduced model that enabled us to bypass time-consuming parameter estimation process involving hundreds of samples in a population.

Emission probability model

The success of HMM partly depends upon how we formulate emission probability, since it theoretically links our observation to hidden state we are trying to identify. The role of emission probability in HMM model is to quantify the possibility that a certain observation could be generated by each of the possible states in the system. Probability of alignment incidence at read level or of the read proportions from a post EMASE result.

Assuming there is no dependence between two alleles on transcription event, it is reasonable to model emission probability, $P(y_t|s_t)$, as a mixture of H multivariate Gaussian components.

Suppose a genotype is $g = \{h_M, h_P\}$, then,

$$\begin{aligned} P(y_t|s_t = g) &= \sum_{h \in g} P(y_t|h)P(h|s_t) \\ &= \sum_{h \in g} \omega_h N(\mu_{t,h}, \Sigma_{t,h}) \end{aligned} \tag{2}$$

In GBRS, we obtain $\mu_{t,g}$ and $\Sigma_{t,g}$, from external information, i.e., the founder RNA-seq data. Therefore, unlike the standard HMM process, we do not estimate the model parameters from the data during the fitting of HMM model.

Transition probability model

Transition probability, $P(s_t|s_{t-1})$, is dependent on the breeding generation and the distance between markers. As GBRS uses expressed genes as marker, we compute the theoretical transition probability according to [Broman, 2006, Broman, 2012] using genetic distances (in centi-morgan) between the start sites of neighboring genes. Unlike the typical HMM process, we use the computed transition probability fixed in our HMM without fitting from the data.

Posterior genotype probability

Suppose s_t is the genotype of t-th gene ($1 \leq t \leq T$) in a chromosome C, and $Y_1^t = \{y_1, y_2, \cdots, y_t\}$ denotes our sequential observations on the first gene upto t-th gene in C. Our first goal is to estimate the posterior probability of genotype s_t given the entire sequence of observations Y_1^T. Before we describe the entire process, we need to define two probabilities: forward and backward probabilities. Forward probability, $\alpha_t(g)$, is defined as a joint probability of Y_1^t and s_t, and, based

on Markov property, we can build the following recurrence relation.

$$\begin{aligned}
\alpha_t(g) &\equiv P(Y_1^t, s_t = g) \\
&= \sum_{g'} \left\{ P(Y_1^{t-1}, s_{t-1} = g') P(s_t = g | s_{t-1} = g') \right\} P(y_t | s_t = g) \\
&= \sum_{g'} \left\{ \alpha_{t-1}(g') P(s_t = g | s_{t-1} = g') \right\} P(y_t | s_t = g)
\end{aligned} \tag{3}$$

Backward probability, $\beta_t(g)$, is defined as the conditional probability of Y_{t+1}^T given the genotype s_t. It describes how likely the following sequence of observations are, assuming a certain genotype state at t-th gene. We can also build a recurrence relation for backward probability similar to that of forward probability.

$$\begin{aligned}
\beta_t(g) &\equiv P(Y_{t+1}^T | s_t = g) \\
&= \sum_{g'} P(Y_{t+2}^T | s_{t+1} = g') P(s_t = g | s_{t+1} = g') P(y_{t+1} | s_{t+1} = g') \\
&= \sum_{g'} \beta_{t+1}(g') P(s_t = g | s_{t+1} = g') P(y_{t+1} | s_{t+1} = g')
\end{aligned} \tag{4}$$

The product of forward and backward probabilities, $\alpha_t(g)\beta_t(g)$, is the joint probability of the entire observation, Y_1^T, and the genotype state, s_t, at any gene t.

$$\alpha_t(g)\beta_t(g) = P(Y_1^t, s_t) P(Y_{t+1}^T | s_t) = P(Y_1^T, s_t) \tag{5}$$

This joint probability is crucial for computing the posterior genotype probability given the entire

sequence of observations.

$$\begin{aligned}
\gamma_t(g) &\equiv P(s_t|Y_1^T) = \frac{P(Y_1^T, s_t)}{P(Y_1^T)} \\
&= \frac{P(Y_1^T, s_t)}{\sum_{s'_t} P(Y_1^T, s'_t)} = \frac{P(Y_1^t, s_t)P(Y_{t+1}^T|s_t)}{\sum_{s'_t} P(Y_1^t, s'_t)P(Y_{t+1}^T|s'_t)} \\
&= \frac{\alpha_t(g)\beta_t(g)}{\sum_{g'} \alpha_t(g')\beta_t(g')}
\end{aligned} \tag{6}$$

Genotype calling

Once we have the posterior probability of genotype states over all genes, $P(s_t|Y_1^T)$, we can find the most likely path of states across C: $\{\widehat{s_1}, \widehat{s_2}, \cdots, \widehat{s_T}\}$ as our final genotype calls. There are prohibitively large number of possible paths to explore but, in fact, we can solve it easily using Viterbi algorithm where we build the following recurrence relation quite similar to that of forward probability.

$$\begin{aligned}
\delta_t(g) &\equiv \max_{s_1,\cdots,s_{t-1}} P(Y_1^t, s_1, \cdots, s_{t-1}, s_t = g) \\
&= \max_{g'} \left\{ \delta_{t-1}(g') \cdot P(s_t = g|s_{t-1} = g') \right\} \cdot P(y_t|s_t = g)
\end{aligned} \tag{7}$$

After solving the recurrence relation, genotype can be determined by backtracking the maximum values of $\delta_t(g)$.

$$\widehat{s}_t = \operatorname*{argmax}_{g'} \left\{ \delta_t(g') \cdot P(s_{t+1} = \widehat{s}_{t+1}|s_t = g') \right\} \tag{8}$$

Parameter estimation

For the DO, emission and transition probabilities are well defined theoretically or attainable from external resource, and therefore, do not require the estimation of model parameters. If the probability model or external data are not available, we can still estimate them from the data using the maximum likelihood estimator of $\mu_{t,g}$ and $\Sigma_{t,g}$, and the posterior genotype probabilities across all the samples.

Suppose k denotes each sample ($1 \leq k \leq K$) in our population,

$$\hat{\mu}_{t,g}^{MLE} = \frac{\sum_k \gamma_{t,k}(g) \cdot y_{t,k}}{\sum_k \gamma_{t,k}(g)} \tag{9}$$

$$\widehat{\Sigma}_{t,g}^{MLE} = \frac{\sum_k \gamma_{t,k}(g) \cdot (y_{t,k} - \hat{\mu}_{t,g})(y_{t,k} - \hat{\mu}_{t,g})^T}{\sum_k \gamma_{t,k}(g)} \tag{10}$$

Combined estimation of genotype probabilities using GBRS and genotyping arrays

We can combine posterior genotype probabilities, P_1 and P_2, with the following formula.

$$P(s_\tau = g|Y_1^T) = \frac{P_1(s_\tau = g|Y_1^T)P_2(s_\tau = g|Y_1^T)}{\sum_{g'} P_1(s_\tau = g'|Y_1^T)P_2(s_\tau = g'|Y_1^T)} \tag{11}$$

where τ is any locus on the genome. Note $P_1(s_\tau|Y_1^T)$ and $P_2(s_\tau|Y_1^T)$ are interpolated genotype probabilities between two markers flanking τ. Combining genotype probabilities of `GBRS` and genotyping arrays could be complementary to each other because `GBRS` has strength in detecting haplotype signal from coding regions while genotyping arrays can cover gene sparse region.

Estimation of diploid allele-specific expression

For each sample, we filter out alignments in the 8-way alignment incidence matrix inconsistent to the diplotype call on each gene. This process ensures that the final alignment incidence matrix is truly diploid. Then we get allele-specific and total gene expression by running emase algorithm again on this diploid alignment incidence matrix.

Software Availability

We have implemented the GBRS algorithm in an open-source python package available at http://churchill-lab.github.io/gbrs/ with MIT license. The package is dockerized and available for pulling and executing at the docker hub (https://hub.docker.com/r/kbchoi/gbrs). A tutorial that describe the whole analysis pipeline is also available on the front page of the github and dockerhub repositories. For the DO and Collaborative Cross (CC) mice, or F1 hybrids of the DO x transgenic C57BL/6J (DO-F1) or F1 hybrids of CC's (CCRIX), required data files are available at ftp://churchill-lab.jax.org/software/GBRS/.

eQTL analysis

We applied GBRS to RNA-seq data from liver samples of 482 Diversity Outbred (DO) mice, as well as striatum samples of 369 DO mice. We genotyped 282 DO liver samples with the Mouse Universal Geotyping Array (MUGA) and 200 DO liver samples on the MegaMUGA (GeneSeek, Lincoln, NE). All DO striatum samples were genotyped on the GigaMUGA. Low-quality samples which have high percent missing genotypes were removed, and then 275, 184 and 358 DO mice were left in MUGA, MegaMUGA and GigaMUGA respectively. The founder haplotypes of these DO mice were inferred using an HMM [K. W. Broman, 2009] implemented in the R/qtl2 package

(https://doi.org/10.1534/genetics.118.301595) and then used to interpolate to an evenly-spaced grid of 69,000 markers.

First, we assessed the ability of GBRS in estimating recombination events. Second, for each sample, the agreement of haplotype reconstructions between the DNA genotyping array and GBRS reconstruction is calculated by the Pearson correlation between each pair of samples. We assumed that a sample was mismatched if a correlation of the array sample with the same sample ID in the RNA-seq data is below 0.6. For each mismatched sample, we then looked for the correct match in the RNA-seq samples with higher correlation ($r > 0.6$) but not assigned to other samples. Third, we mapped eQTLs, using gene-level counts estimated from alignment of reads to the reference transcriptome. Genes with median of count value > 1 were included in the eQTL analysis. Raw counts in each sample were then normalized by the variance-stabilizing transformation (VST) in DeSeq2 R package [PMID: 25516281]. A linear mixed model was performed with sex, diet and generation as additive covariates and a random polygenic term to account for genetic relatedness using qtl2 package. The significance thresholds were established by performing 1,000 permutations and fitting an extreme value distribution to the maximum LOD scores [PMID:15266393]. Permutation derived P-values were then converted to q-values with the QVALUE R package, using the bootstrap method to estimate π_0 and the default λ tuning parameters (https://doi.org/10.1111/j.1467-9868.2004.00439.x). The significance threshold for declaring a QTL was set at a genome-wide significance level of FDR = 5%.

References

[Aguet et al., 2017] Aguet, F., Brown, A. A., Castel, S. E., Davis, J. R., He, Y., Jo, B., Mohammadi, P., Park, Y., Parsana, P., Segrè, A. V., *et al.*, 2017. Genetic effects on gene expression across human tissues. *Nature*, **550**(7675):204–213.

[Broman, 2006] Broman, K. W., 2006. The genomes of recombinant inbred lines. *Genetics*, **173**(4):2419–2419.

[Broman, 2012] Broman, K. W., 2012. Haplotype probabilities in advanced intercross populations. *G3: Genes, Genomes, Genetics*, **2**(2):199–202.

[Broman et al., 2019] Broman, K. W., Gatti, D. M., Svenson, K. L., Sen, Ś., and Churchill, G. A., 2019. Cleaning genotype data from diversity outbred mice. *G3: Genes, Genomes, Genetics*, .

[Broman et al., 2015] Broman, K. W., Keller, M. P., Broman, A. T., Kendziorski, C., Yandell, B. S., Sen, Ś., and Attie, A. D., 2015. Identification and correction of sample mix-ups in expression genetic data: A case study. *G3 (Bethesda, Md.)*, **5**(10):2177–2186.

[Chick et al., 2016] Chick, J. M., Munger, S. C., Simecek, P., Huttlin, E. L., Choi, K., Gatti, D. M., Raghupathy, N., Svenson, K. L., Churchill, G. A., and Gygi, S. P., *et al.*, 2016. Defining the consequences of genetic variation on a proteome-wide scale. *Nature*, **534**(7608):500–505.

[de Koning and McIntyre, 2017] de Koning, D.-J. and McIntyre, L. M., 2017. Back to the future: Multiparent populations provide the key to unlocking the genetic basis of complex traits. *G3: Genes, Genomes, Genetics*, **7**(6):1617–1618.

[Degner et al., 2009] Degner, J. F., Marioni, J. C., Pai, A. A., Pickrell, J. K., Nkadori, E., Gilad, Y., and Pritchard, J. K., 2009. Effect of read-mapping biases on detecting allele-specific expression from RNA-sequencing data. *Bioinformatics*, **25**(24):3207–3212.

[Gatti et al., 2014] Gatti, D. M., Svenson, K. L., Shabalin, A., Wu, L.-Y., Valdar, W., Simecek, P., Goodwin, N., Cheng, R., Pomp, D., Palmer, A., *et al.*, 2014. Quantitative trait locus mapping methods for diversity outbred mice. *G3 (Bethesda)*, **4**(9):1623–1633.

[Gileta et al., 2020] Gileta, A. F., Gao, J., Chitre, A. S., Bimschleger, H. V., St. Pierre, C. L., Gopalakrishnan, S., and Palmer, A. A., 2020. Adapting genotyping-by-sequencing and variant calling for heterogeneous stock rats. *G3: Genes, Genomes, Genetics*, **10**(7):2195–2205.

[Gu et al., 2016] Gu, T., Gatti, D. M., Srivastava, A., Snyder, E. M., Raghupathy, N., Simecek, P., Svenson, K. L., Dotu, I., Chuang, J. H., Keller, M. P., *et al.*, 2016. Genetic architectures of quantitative variation in rna editing pathways. *Genetics*, **202**(2):787–798.

[Javed et al., 2020] Javed, N., Farjoun, Y., Fennell, T. J., Epstein, C. B., Bernstein, B. E., and Shoresh, N., 2020. Detecting sample swaps in diverse ngs data types using linkage disequilibrium. *Nature Communications*, **11**(1):3697.

[K. W. Broman, 2009] K. W. Broman, S. S., 2009. *Guide to qtl mapping with r/qtl*. Springer-Verlag New York.

[Keane et al., 2011] Keane, T. M., Goodstadt, L., Danecek, P., White, M. A., Wong, K., Yalcin, B., Heger, A., Agam, A., Slater, G., Goodson, M., *et al.*, 2011. Mouse genomic variation and its effect on phenotypes and gene regulation. *Nature*, **477**(7364):289–294.

[Miller et al., 2013] Miller, A. C., Obholzer, N. D., Shah, A. N., Megason, S. G., and Moens, C. B., 2013. Rna-seq-based mapping and candidate identification of mutations from forward genetic screens. *Genome Res*, **23**(4):679–686.

[Morgan et al., 2016] Morgan, A. P., Fu, C.-P., Kao, C.-Y., Welsh, C. E., Didion, J. P., Yadgary, L., Hyacinth, L., Ferris, M. T., Bell, T. A., Miller, D. R., *et al.*, 2016. The mouse universal genotyping array: From substrains to subspecies. *G3: Genes, Genomes, Genetics*, **6**(2):263–279.

[Munger et al., 2014] Munger, S. C., Raghupathy, N., Choi, K., Simons, A. K., Gatti, D. M., Hinerfeld, D. A., Svenson, K. L., Keller, M. P., Attie, A. D., Hibbs, M. A., *et al.*, 2014. Rna-seq alignment to individualized genomes improves transcript abundance estimates in multiparent populations. *Genetics*, **198**(1):59–73.

[Parker et al., 2016] Parker, C. C., Gopalakrishnan, S., Carbonetto, P., Gonzales, N. M., Leung, E., Park, Y. J., Aryee, E., Davis, J., Blizard, D. A., Ackert-Bicknell, C. L., *et al.*, 2016. Genome-wide association study of behavioral, physiological and gene expression traits in outbred cfw mice. *Nature Genetics*, **48**(8):919–926.

[Piskol et al., 2013] Piskol, R., Ramaswami, G., and Li, J. B., 2013. Reliable identification of genomic variants from rna-seq data. *American journal of human genetics*, **93**(4):641–651.

[Raghupathy et al., 2018] Raghupathy, N., Choi, K., Vincent, M. J., Beane, G. L., Sheppard, K. S., Munger, S. C., Korstanje, R., Pardo-Manual de Villena, F., and Churchill, G. A., 2018. Hierarchical analysis of RNA-seq reads improves the accuracy of allele-specific expression. *Bioinformatics*, **34**(13):2177–2184.

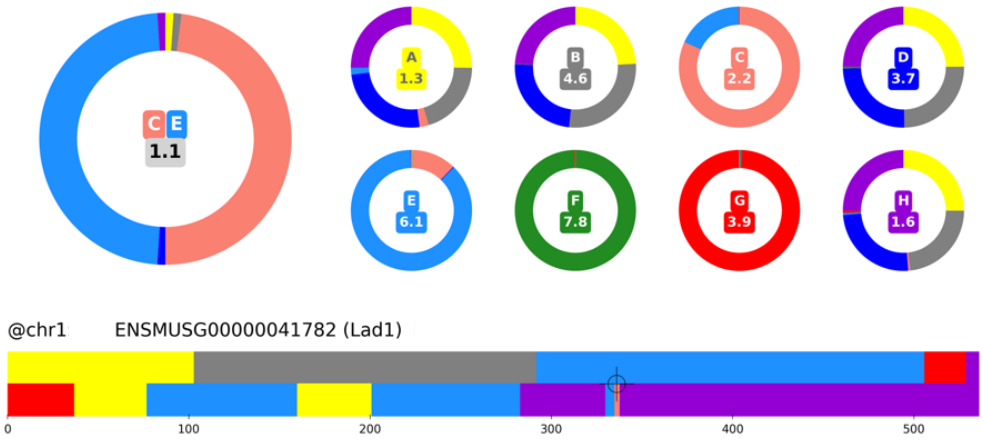
[Stark et al., 2019] Stark, R., Grzelak, M., and Hadfield, J., 2019. Rna sequencing: the teenage years. *Nature Reviews Genetics*, **20**(11):631–656.

[Svenson et al., 2012] Svenson, K. L., Gatti, D. M., Valdar, W., Welsh, C. E., Cheng, R., Chesler, E. J., Palmer, A. A., McMillan, L., and Churchill, G. A., 2012. High-resolution genetic mapping using the mouse diversity outbred population. *Genetics*, **190**(2):437–447.

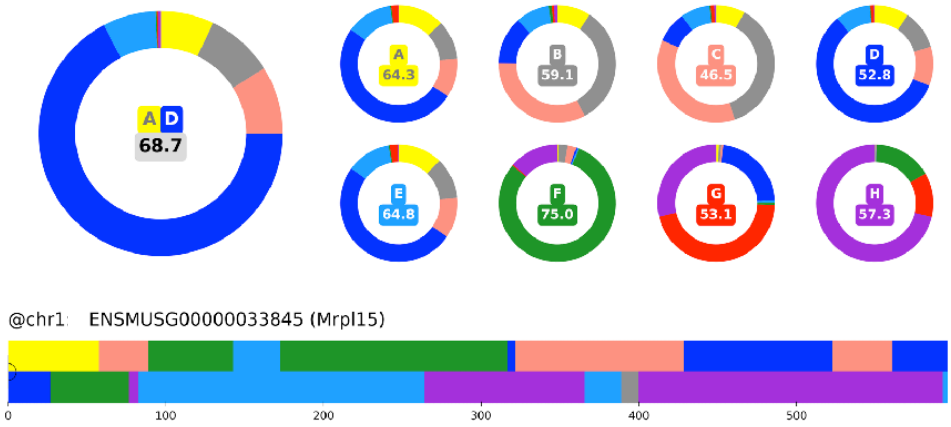
[Wittkopp et al., 2004] Wittkopp, P. J., Haerum, B. K., and Clark, A. G., 2004. Evolutionary changes in cis and trans gene regulation. *Nature*, **430**(6995):85–88.

[Yang et al., 2009] Yang, H., Ding, Y., Hutchins, L. N., Szatkiewicz, J., Bell, T. A., Paigen, B. J., Graber, J. H., de Villena, F. P.-M., and Churchill, G. A., 2009. A customized and versatile high-density genotyping array for the mouse. *Nature Methods*, **6**(9):663–666.

[Yates et al., 2015] Yates, A., Akanni, W., Amode, M. R., Barrell, D., Billis, K., Carvalho-Silva, D., Cummins, C., Clapham, P., Fitzgerald, S., Gil, L., *et al.*, 2015. Ensembl 2016. *Nucleic Acids Research*, **44**(D1):D710–D716.



(a)



(b)

Figure 1: multi-way ASE profiles of known genetic background to form a guage to infer the unknown genotypes. GBRS employs a common multi-way alignment strategy for every MPP individual instead of building hundreds or thousands of individualized alignment indices as in Munger et al [Munger et al., 2014]. GBRS determines the genotype of a gene in an MPP individual primarily by comparing its multi-way ASE profile (a bigger donut chart on the left) with a catalogue of ASE profiles of founder strains (eight smaller donut charts each estimated from separate founder RNA-seq data). (a) An example of *Lad1* gene where A, B, D, and H shares similar multi-way ASE profile. This happens in many genes as they (including C) are subspecies closer to each other. On the other hand, C, E, F, and G gene expression contain enough information to distinguish themselves among the eight. (b) An example of *Mrpl15* where

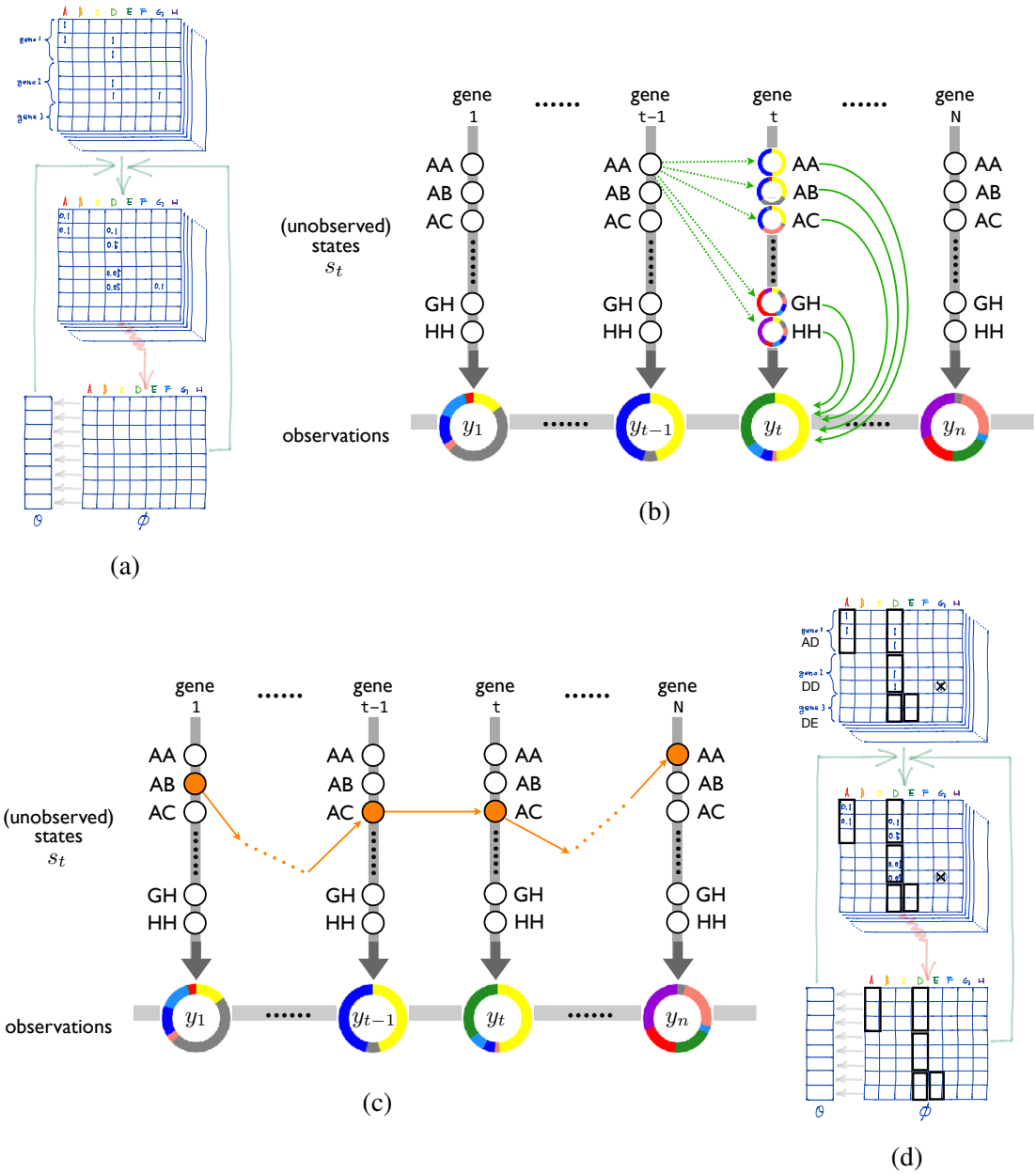
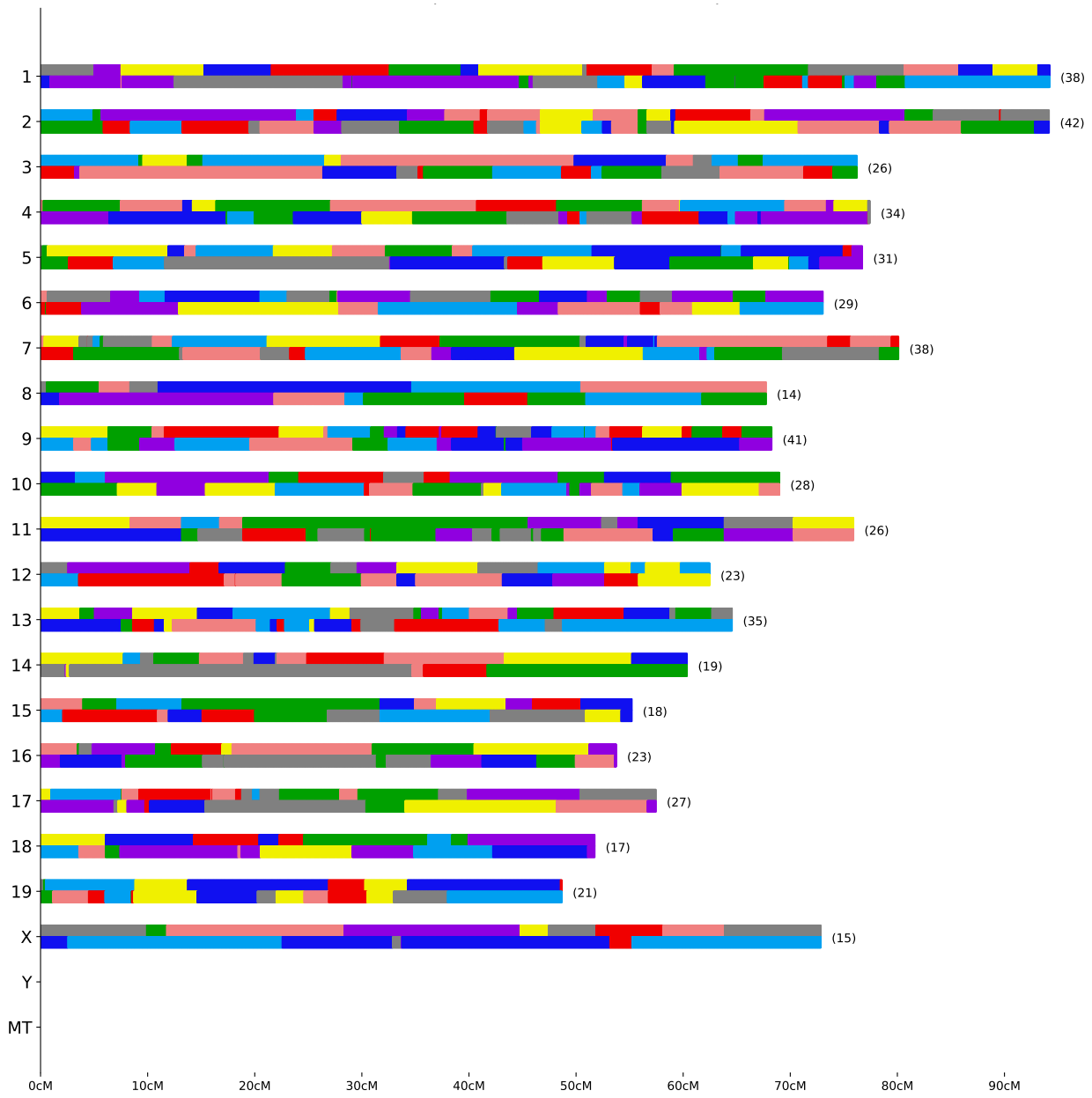
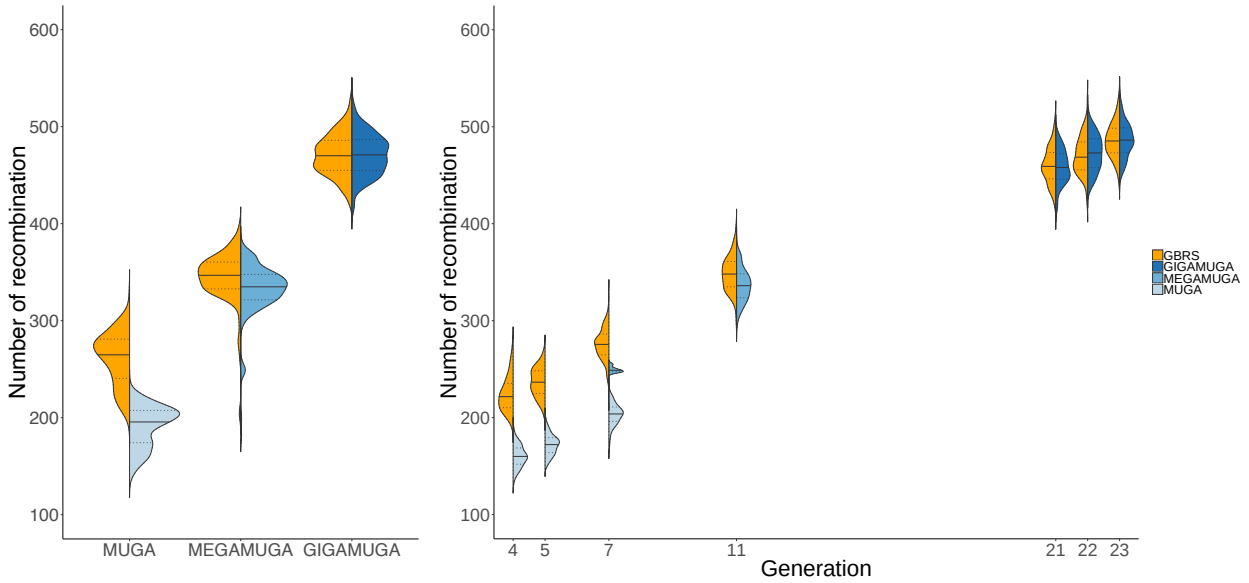


Figure 2: An overview of GBRS algorithm. GBRS employs multi-way alignment strategy, and estimates multi-way ASE as observation on each gene (a). GBRS uses forward-backward algorithm (HMM) to estimate genotype probabilities for every expressed gene. In this process, multi-way ASE from the MPP founder strains (Figure 1) is used as essential information for estimating emission probabilities (green solid arrows). Transition probabilities (green dotted arrows) are estimated theoretically [Broman et al., 2015] based on intergenic distances (b). GBRS identifies the maximum-probability path across genotype states (orange arrows) using Viterbi algorithm, and calls genotype on each gene loci (c). An example diploid genome reconstruction based on the genotype calls is shown in Figure 3. GBRS masks alignments not consistent with the genotype calls (boxes denoted with 'X') and runs EM once more to generate diploid version of ASE (d).



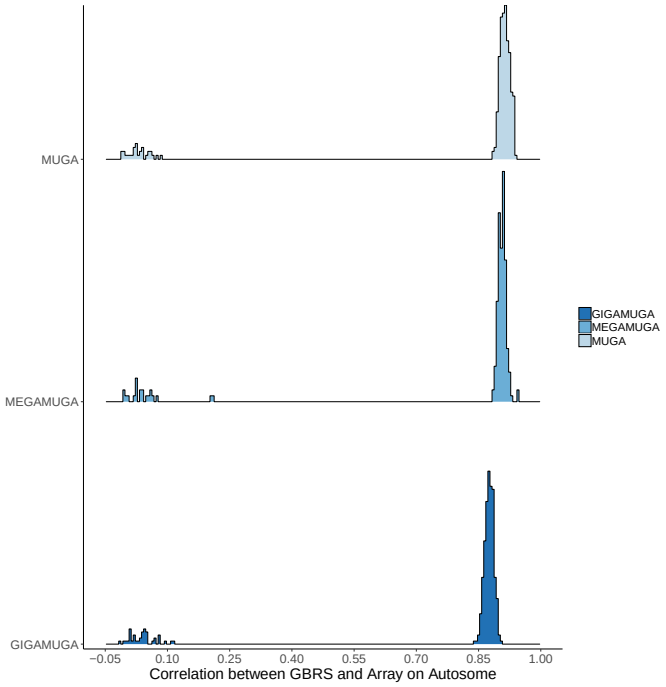
(a)

Figure 3: An example of Viterbi genotype calls and individualized genome. GBRS reconstructed the diploid genome for a female sample from the striatum data set (ID 8684) and identified 545 recombination events in total.

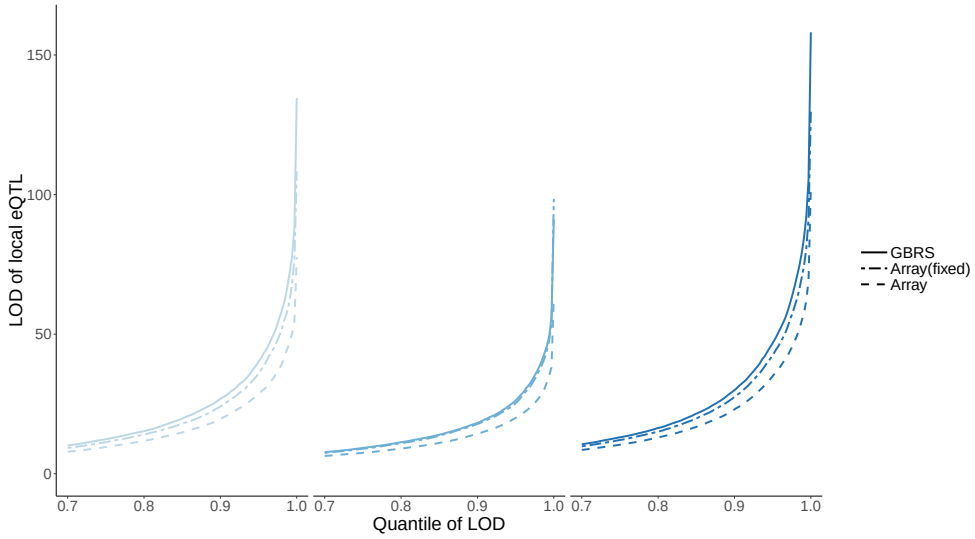


(a) (b)

Figure 4: The number of recombination events identified by GBRS (color coded in orange) in comparison to MUGA (light blue), MegaMUGA (blue), and GigaMUGA (dark blue) genotyping arrays (a), and the numbers along the outbreeding generations (b). GBRS identifies recombination events comparable to that of GigaMUGA, a genotyping array with the highest resolution (over 140,000 SNPs with average marker spacing of ∼22.5 kb).



(a)



(b)

Figure 5: The genotyping accuracy of GBRS. Individual genomes reconstructed by GBRS are concordant with those of MUGA (color coded in light blue), MegaMUGA (blue), and GigaMUGA (dark blue) arrays. The median of Pearson correlation of genotype probabilities is 0.876, 0.862 and 0.834 respectively. Using GBRS, we can also identify samples that are mixed up in the experimental procedure (correlation around 0). We realign samples by pairing ones with highest correlations (a). The genotyping accuracy is reflected in the LOD scores of local eQTLs because mis-specified genotype as a predictor has a negative impact on estimated likelihood. GBRS genotypes result in consistently higher LODs compared to MUGA, MegaMUGA, and GigaMUGA across top 30 percentile of local eQTLs. Fixing sample mix-ups substantially improved the local regulatory signals on all three array-based genotyping but accuracy of GBRS is still higher for these datasets.

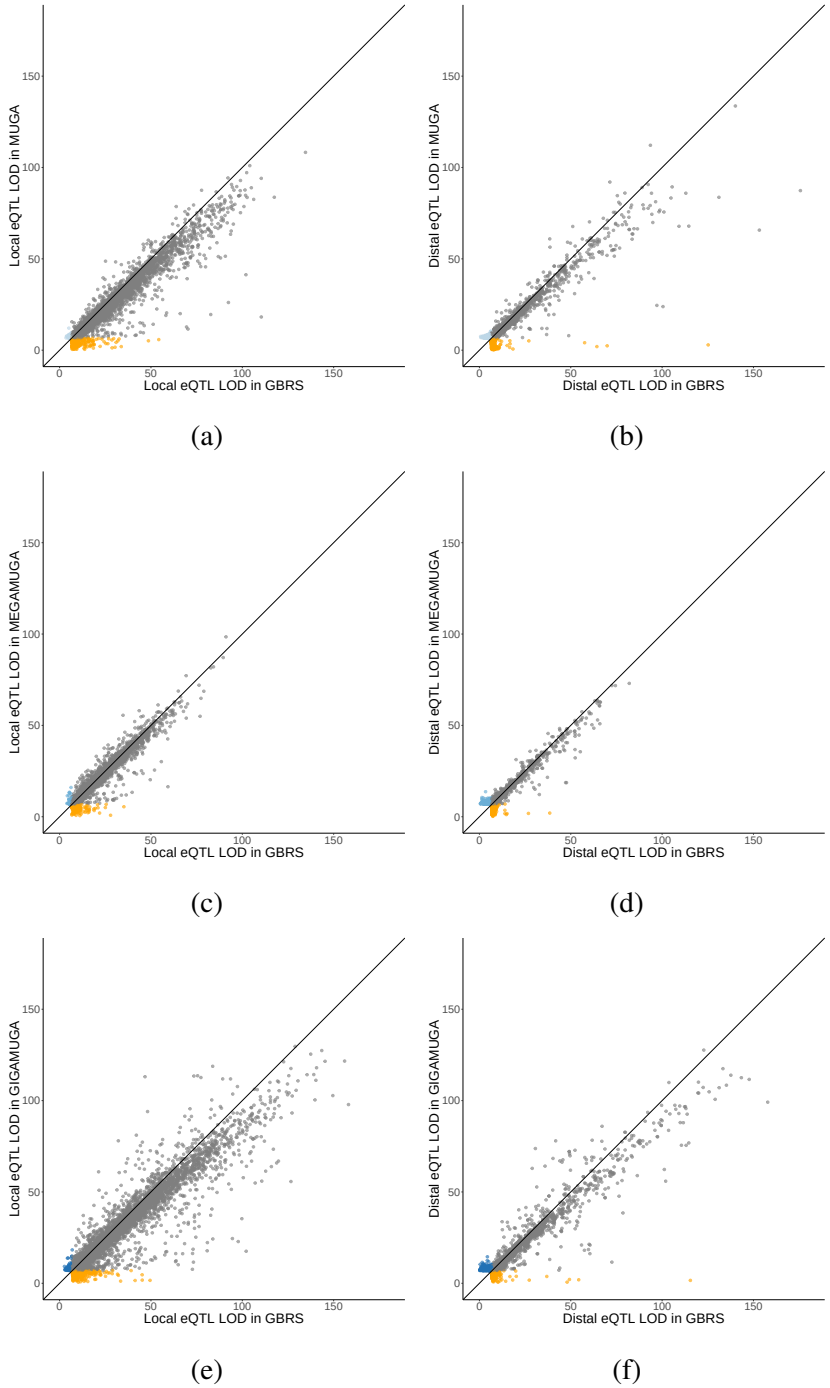


Figure 6: Genotyping accuracy of GBRS with full eQTL analysis in direct comparison with MUGA (a)(b), MegaMUGA (c)(d), and GigaMUGA (e)(f) arrays. The LODs of local and distal eQTLs are shown in (a)(c)(e) and (b)(d)(f) respectively. Most data points are distributed along the diagonal line. This implies GBRS genome reconstruction is concordant to array-based results. Overall GBRS resulted in higher LODs as we can see more data points below the diagonal line. Even though GBRS uses the expressed genes as a marker to reconstruct individual genomes, it achieves comparable accuracy as GigaMUGA, the array with over 140,000 markers.