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], Isabela Gerdes Gyuricza[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.
Multi-parent populations (MPPs), genetically segregating model systems derived from two or more inbred founder strains, are widely used in biomedical and agricultural research. Gene expression profiling by direct RNA sequencing (RNA-Seq) is commonly applied to MPPs to investigate gene expression regulation and to identify candidate genes. In genetically diverse populations, including most MPPs, quantification of gene expression is improved when the RNA-Seq reads are aligned to individualized transcriptomes that incorporate known polymorphic loci. However, the process of constructing and analyzing individual genomes can be computationally demanding and error prone. We propose a new approach, genome reconstruction by RNA-Seq (GBRS), that relies on simultaneous alignment of RNA-Seq reads to the founder strain transcriptomes. GBRS can reconstruct the diploid genome of each individual and quantify both total and allele-specific gene expression. We demonstrate that GBRS performs as well as methods that rely on high-density genotyping arrays to reconstruct the founder haplotype mosaic of MPP individuals. Using GBRS in addition to other genotyping methods provides quality control for detecting sample mix-ups and improves 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 quantifying transcript abundance, RNA-Seq data can also identify single nucleotide polymorphisms (SNPs) and small insertion and deletions (indels) in the transcribed portion of the genome [Piskol et al., 2013]. Recent developments in analysis of RNA-Seq data have enabled detection of spontaneous mutations [Miller et al., 2013], RNA editing events [Gu et al., 2016], and allele-specific expression [Wittkopp et al., 2004]. Here we examine the potential for using RNA-Seq data for genotype reconstruction and improved quantification of gene expression in multi-parent populations (MPPs).

In typical RNA-Seq analysis, transcript abundance is quantified by counting reads that align to a reference genome-based transcriptome index [Ferragina and Manzini, 2000]. Alignment algorithms make allowance for mismatches due to sequencing errors and polymorphisms that result in differences between the sample RNA and the reference transcript sequences. Nonetheless, reliance on a reference genome can lead to biases in gene expression quantification [Degner et al., 2009]. Reference bias can be minimized when RNA-Seq reads are aligned to a transcriptome index that incorporates genetic variants that are expected to be present in the sample RNA. We have previously shown that alignment to individualized transcriptomes substantially improves expression quantitative trait locus (eQTL) mapping [Munger et al., 2014]. This approach requires prior knowledge of the individual genomes, obtained either by whole genome sequencing or by genotyping and imputation; it requires the construction of alignment indices for every sample; and it may be error prone when knowledge of sequence variants in individuals is uncertain.

MPPs are model organism genetic reference populations descendant from two or more inbred founder strains [de Koning and McIntyre, 2017]. The genome of each individual in a MPP is composed of a mosaic of founder strain genome segments. For many MPPs, whole genome sequences of the founder strains have been assembled. Individuals from a MPP can be genetically characterized with genotyping arrays [Morgan et al., 2016] or short-read DNA sequencing [Parker et al., 2016] to detect known founder strain variants and the genome mosaic can be reconstructed using a Markov model (HMM) to process the variant call data [Gatti et al., 2014, Broman et al., 2019]. The full (diploid) genome sequence of a MPP individual can then be inferred by imputation of variants onto the founder haplotype blocks of the mosaic MPP genome [Munger et al., 2014].

In any given tissue, thousands of expressed genes are distributed across the genome. Therefore, variants that are detected by RNA-Seq could replace genotyping arrays. However, it is not immedi-

ately clear if RNA-Seq data capture sufficient information to support accurate genome reconstruction. Direct approaches based on variant calling from RNA-Seq reads require deep sequencing coverage (Lopez-Maestre et al., 2016 NAR) and may not be reliable, especially for genes with low expression. Also, the distribution of expressed genes may not be sufficiently dense or uniform to accurately reconstruct haplotypes in some regions.

We propose a novel solution to the challenges of quantifying RNA-Seq data from MPP samples. Instead of building a diploid alignment index for each individual in a MPP, we use a single *multi-way* alignment index that represents the combined transcriptomes of the MPP founder strains. We align RNA-Seq reads to the multi-way index and apply a generalized version of our allele-specific quantification algorithm EMASE [Raghupathy et al., 2018] to allocate read counts to each of the founder haplotypes. We avoid direct variant calling and instead use information in the pattern of haplotype-specific read counts. Individual genes may not have sufficient sequence variation to uniquely identify a haplotype. Therefore, we apply a Hidden Markov Model (HMM) that combines information across neighboring genes to reconstruct the diploid mosaic genome for each MPP individual. We demonstrate that the information in RNA-Seq data is sufficient to compute accurate haplotype reconstructions. The procedure is implemented in an open-source Python package, `GBRS` (Genome reconstruction By RNA-Seq), available at https://churchill-lab.github.io/gbrs/.

METHODS

Overview of 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 construction of individual alignment indices is computationally demanding and potentially error prone. Instead, GBRS employs a single alignment index built with the combined predicted transcript sequences of the MPP founder strains. We hypothesized that simultaneous alignment of RNA-Seq reads to the full set of founder transcripts would provide enough information to resolve local haplotypes. The variants present in any single gene may not uniquely resolve the full set of founder haplotypes but GBRS combines information across neighboring genes using a Hidden Markov Model (HMM) to achieve full resolution and accurate estimates of the haplotype mosaic structure of individual MPP samples.

The GBRS process starts by *training* the HMM using RNA-Seq data obtained from the founder strains. For each founder strain sample, we align RNA-Seq reads to a multi-way alignment index including the full set of predicted founder transcripts. Many reads will align identically to multiple founder transcripts depending on which polymorphic loci are spanned by the read. We then apply an extension of the allele-specific weighted allocation algorithm EMASE [Raghupathy et al., 2018] to reapportion the ambiguous alignments across the founder haplotypes and obtain estimated expected read counts for each gene. RNA-Seq reads from each founder strain sample will display a characteristic distribution of expected read counts that we refer to as the *founder profile* of the gene (Figure 1). For the DO mice, which have eight founder strains (here denoted as A, B, C, $\cdots$, H), the founder profiles can be represented as a matrix with eight rows, each corresponding to RNA-Seq data from a founder strain, and eight columns, each corresponding to a founder strain-specific gene transcript. The rows of the matrix are proportions that sum to one. For a gene with polymorphisms that uniquely identify a given founder strain, the founder profile will take a value approaching 1.0 on the corresponding diagonal element. For a gene that lacks sufficient polymor-

phisms to uniquely identify a founder strain, the weighted allocation of reads from that founder may be distributed across multiple founder transcripts.

The founder profiles depend only on the predicted transcript sequences and the RNA-Seq reads obtained from the founder strain samples. Thus, the same profile will be reproduced in weighted allocation of reads from an MPP individual that is homozygous for the founder haplotype at a gene. This will be true even if the founder strain transcript is not accurate. In the case of heterozygosity, the weighted allocation of reads will represent a mixture of two founder haplotypes. This is how GBRS identifies the founder haplotype of origin at a gene in an uncharacterized sample from an MPP individual. The use of read alignment proportions distinguishes `GBRS` from other sequencing-based genotyping methods that rely on accurate variant calling.

[Figure 1 about here.]

The first step in `GBRS` analysis of an individual MPP sample is to align the RNA-Seq reads to the multi-way alignment index and compute *the sample profiles* — the proportions of reads that are allocated to each founder haplotype at each gene (Figure 2a). The sample profiles are the input data for the HMM. We then estimate the genotype states using the forward-backward algorithm [Rabiner, 1989]. The emission model of the HMM compares the MPP sample profile to each of the founder profiles including the heterozygote profiles, which we assume to be an equal mixture of two founder profiles. The transition matrix of the HMM is derived from genetic map distances between genes [Broman, 2012] (Figure 2b). The HMM combines information across neighboring genes to resolve genes for which the founder profiles are not fully informative. The forward-backward algorithm computes the marginal posterior probabilities for the hidden genotype states at each gene, which accounts for uncertainty in the estimated genotypes. For the DO mice,

there are 36 possible genotype states. The 36-state *genotype probabilities* can be collapsed to 8-state *haplotype dosages* [Gatti et al., 2014] for quantitative trait locus (QTL) mapping and other applications.

In order to quantify the allele-specific and total gene expression of each gene in a MPP sample, we apply a Viterbi algorithm [Viterbi, 1967] to the genotype probabilities to obtain a maximum-probability reconstruction of the diploid founder mosaic that constitutes the individual MPP genome (Figure 2c). The Viterbi algorithm selects a single 'best' genotype state at each gene. We use the Viterbi reconstruction to reduce the allowable set of read alignments from all founder transcripts to just two and repeat the weighted allocation of the RNA-Seq reads to obtain estimated expected counts of allele-specific expression [Raghupathy et al., 2018] (Figure 2d). This step does not require re-alignment of the RNA-Seq reads, so it is very fast.

[Figure 2 about here.]

Data

We obtained RNA-Seq data on liver tissue samples from 16 mice (8 male and 8 female) from each of the eight Diversity Outbred (DO) founder strains, 128 mice in total. We also obtained RNA-Seq data on liver tissue samples from 482 DO mice including equal numbers of male and female animals from breeding generations 4 to 11 (G4∼G11) [Svenson et al., 2012]. We obtained single-ended sequence data on a HiSeq2000 [Illumina] producing 100 bp reads at a depth of 20M reads per sample. In addition, we obtained RNA-Seq data on striatum tissue samples from 416 DO mice including mice of both sexes spanning generations G21, G22 and G23 [Philip et al, in preparation]. Libraries were pooled and sequenced 100 bp paired end to a depth of 50M reads each on a HiSeq 2500 [Illumina].

The DO mice were genotyped with three different versions of the mouse universal genotyping array (MUGA) [Geneseek, Lincoln, NE]. We obtained genotypes for 282 mice from the liver study on the original MUGA array with ∼7,500 marker loci. The remaining 200 mice were genotyped on the MegaMUGA array with ∼77,800 markers. All 416 mice from the striatum study were genotyped on the GigaMUGA array with ∼150,000 markers.

Building a multi-way alignment index

We create custom genomes of the 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 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 append the founder strain code (e.g., A, B, C, $\cdots$, H for A/J, C57BL/6J, 129S1/SvImJ, NOD/ShiLtJ, NZO/HILtJ, CAST/EiJ, PWK/PhJ, and WSB/EiJ, respectively) to each transcript ID; collate the founder transcript sequences in a single fasta file; add 99 bp-long poly-A tail; and then run `bowtie-build` command to build a multi-way alignment index representing transcript sequences from the eight founder strains of the DO.

Estimating an alignment profile for each founder and MPP individual

We align RNA-seq data from each of the founders and from the MPP individuals to the multi-way alignment index using `bowtie1` v1.0.0 with '`all`', '`best`', and '`strata`' options. Many reads

map to multiple haplotypes of a gene and they may also align to multiple genes. For the DO mouse data, we find that only $\sim$2% of the reads uniquely align to a single founder transcript. In order to count the reads, we transform the raw alignment counts to a vector of estimated expected counts using a straightforward extension of our expectation maximization for allele-specific expression (EMASE) algorithm [Raghupathy et al., 2018].

From the multi-way alignment of the founder liver RNA-Seq reads, we detected expression of 12,415 genes after filtering out genes that had mean abundance below 1 TPM. Then we convert the expected counts to proportions — *the founder profile* — which indicates how specifically reads from a founder strain align to their own predicted transcript and to transcripts of the other founder strains. The more the variants that distinguish a founder haplotype, the more specific the founder profile will be. We construct alignment profiles for heterozygous genotypes, assuming there is no imbalance in allelic expression, as the equally weighted average of the corresponding founder profiles.

Similarly, we obtain *sample profiles* for every expressed gene from a multi-way alignment of an MPP individual. In the liver samples we detected expression of 12297.9 ± 423.0 genes and in the striatum samples we detected 18812.7 ± 328.4 genes after filtering genes with average expression below 1 TPM. There were 11293.1 ± 177.7 and 10517.4 ± 68.2 genes that overlap with the founder profile data in the liver and striatum samples, respectively. Although the striatum data was generated in higher depth of coverage, it had less number of genes that are common with the founder profile data because they are different tissue type.

Estimating genotype probabilities using Hidden Markov Model

GBRS uses a Hidden Markov Model (HMM) to estimate the genotype of an individual MPP sample given the sample profiles across the expressed genes, $Y = \{y_1, \cdots, y_t, \cdots, y_T\}$, as observation. The HMM has two components: *emission* model and *transition* model. The emission model describes the probability distribution of a sample profile (y_t) for a given genotype state (s_t) at a gene locus t. The founder profile ($x_{t,g}$) defines the center of the emission probability distribution. We have implemented a simple *emission probability* model based on multivariate normal distribution.

$$P\big(y_t \big| s_t = g\big) \propto \exp\left\{ -\frac{1}{2\sigma^2}\big(y_t - x_{t,g}\big)\big(y_t - x_{t,g}\big)^T \right\} \tag{1}$$

where σ^2 is a tuning parameter that we set to 0.12, a value that approximates the theoretically expected number of recombination events [Gatti et al., 2014].

The transition model describes how the genotype states can change across the intervals between each pair of genes on a chromosome. The *transition probability*, $P\big(s_t \big| s_{t-1}\big)$, is a function of the breeding generation and the distance between neighboring gene loci t-1 and t. We compute the theoretical transition probability according to [Broman, 2006, Broman, 2012] using genetic distances between the start sites of neighboring genes. We obtained the transition probabilities for each breeding generation of DO mice using DOQTL [Gatti et al., 2014] version 1.0.0.

We use these pre-computed emission and transition probabilities in our HMM. This enables us to process each MPP individual independently. In addition, we can compute the posterior probability of genotype state on each gene locus, $P\big(s_t \big| Y\big)$, with a one-time execution of the forward-backward algorithm [Rabiner, 1989]. The probabilistic representation of genotypes enables us to incorporate uncertainty in our genotype calls into the downstream analyses, such as QTL mapping.

Estimating diploid allele-specific expression

Once we have the genotype probabilities at every expressed gene, we can compute the the most probable path for $s_1, s_2, \cdots, s_T$ using a Viterbi algorithm [Viterbi, 1967].

In order to estimate allele-specific expression in each MPP individual, we mask elements of the multi-way alignment incidence profile that are inconsistent with the Viterbi genotype calls and reallocate read counts by running EMASE algorithm once more on this diploid alignment incidence matrix. This ensures that the allele-specific counts reflect the most probable diploid MPP genome.

Combining genotype probabilities estimated with GBRS and genotyping arrays

Genotype probabilities obtained with `GBRS` are consistent with and contain additional information compared to genotyping arrays. `GBRS` contains information about the haplotypes at expressed genes whereas genotyping arrays provide information in gene sparse regions. It is straightforward to combine posterior genotype probabilities, P_1 and P_2, with the following formula:

$$P\big(s_\tau = g\big|Y\big) = \frac{P_1\big(s_\tau = g\big|Y\big)P_2\big(s_\tau = g\big|Y\big)}{\sum_{g'} P_1\big(s_\tau = g'\big|Y\big)P_2\big(s_\tau = g'\big|Y\big)} \tag{2}$$

where τ is any locus on the genome. Note $P_1\big(s_\tau\big|Y\big)$ and $P_2\big(s_\tau\big|Y\big)$ are interpolated genotype probabilities between two markers flanking τ.

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 with a high percent missing genotypes were removed, leaving 275, 184 and 358 DO mice with MUGA, MegaMUGA and GigaMUGA genotypes, respectively. The founder haplotypes of these DO mice were inferred using a HMM [K. W. Broman, 2009] implemented in the R/qtl2 package (https://doi.org/10.1534/genetics.118.301595). To facilitate comparisons across genotyping platforms and with GBRS, we interpolated the genotytpe probabilities onto an evenly-spaced grid of 69,005 markers.

We examined the agreement of haplotype reconstructions between the genotyping array and `GBRS` using the Pearson correlation between each pair of samples. We assumed that a sample was mismatched if the correlation of the array sample with the same sample ID in the RNA-seq data fell below 0.6. For each mismatched sample, we then searched for the correct match in the RNA-seq samples with higher correlation ($r > 0.6$).

We mapped eQTL, using gene-level expected read counts estimated using `GBRS`. Genes with median of count value > 1 were included in the eQTL analysis. Raw counts in each sample were normalized with the variance-stabilizing transformation (VST) in the `DESeq2` R package [Love et al., 2014]. A linear mixed model with sex, diet and generation as additive covariates and a random polygenic term to account for genetic relatedness was fit at each genotype locus using `qtl2` R package. Significance thresholds were established by performing 1,000 permutations and fitting an extreme value distribution to the maximum LOD scores [Dudbridge and Koeleman, 2004]. Permutation derived P-values were then converted to q-values with the `qvalue` R package [Storey et al., 2020], using the bootstrap method to estimate π_0 and the default λ tuning parameters [Storey et al., 2004]. The significance threshold for declaring a QTL was set at a genome-wide significance level of FDR = 5%.

RESULTS

GBRS produces accurate haplotype reconstructions

The genotype probabilities (with dimensions samples$\times$markers$\times$genotypes) are used for QTL mapping, variant imputation, and other genetic analyses of MPP populations [Broman et al., 2019]. The genotype probabilities, which sum to one across genotypes for each sample and marker, capture uncertainty in our estimation of the haplotype mosaic of MPP genomes. We compared `GBRS` genotype probabilities to the array-based genotype probabilities for our DO datasets. We first collapsed the 36-state genotype probabilities to 8-state haplotype dosages as described by [Gatti et al., 2014]. In order to make direct comparison between `GBRS` and array platforms with different marker densities, we interpolated the genotype probabilities onto a common grid of 69,005 pseudo-marker locations with approximately equal spacing in genetic map units across the mouse genome. We then computed Pearson correlations between `GBRS` haplotype probabilities and the corresponding array-based probabilities. We found that $\sim$90% of samples had Pearson correlation $r > 0.8$ (Figure 3). The median correlation coefficient was 0.876, 0.862 and 0.834 between `GBRS` and the MUGA, MegaMUGA, and GigaMUGA arrays respectively. Importantly, we identified 29 MUGA, 21 MegaMUGA, and 39 GigaMUGA samples that have correlations near zero.

[Figure 3 about here.]

GBRS can identify and correct sample mix-ups

In studies that generate large numbers of tissue samples for multiple assays, one should always keep in mind the possibility of sample mix-ups [Broman et al., 2015]. In our comparison of GBRS haplotype reconstructions, derived from tissue samples collected for RNA analysis, and the

MUGA reconstructions, derived from tail-tip samples collected for DNA analysis, we identified 89 individuals (∼10% of animals in these studies) with discordant results. While disconcerting, these errors are easy to identify and to resolve. For each of the 89 samples, we compared the GBRS genotypes to each of the the array-based genotypes within the same experimental cohorts (Table 1). This comparison revealed a one-to-one correspondence for 63 of the samples that were involved in pairwise sample swaps. An additional 5 samples were resolved as 3- or 4-way sample swaps. We found 17 of the RNA samples to be duplicates and 4 samples could not be identified. Some additional work was required to determine if the handling errors occurred among the RNA samples or the DNA samples. All of the sample mix-ups reported here were determined to be due to plating errors in the DNA samples and were corrected.

[Table 1 about here.]

GBRS accurately detects recombination events

Haplotype reconstruction of MPP individuals identifies the locations of recombination events that have accumulated since derivation of the population from the founder inbred strains as well as the founder strain origins of the flanking intervals (Figure 4). In an outbreeding MPP such as the DO, recombination breakpoints accumulate at a predictable linear rate with each breeding generation [Gatti et al., 2014].

[Figure 4 about here.]

We assessed whether GBRS could detect recombination breakpoints with the same sensitivity as the MUGA genotyping arrays (Figure 5). We estimated the number of recombination breakpoints for each sample using Viterbi paths through the HMM for GBRS and MUGA data (Meth-

ods). Compared to the low-density MUGA array, applied to DO mice from generations G4, G5, and G7, we found that GBRS is more sensitive, detecting on average 36.5% more breakpoints. Compared to the MegaMUGA array on mice from generation G4, G7, and G11, GBRS detected 3.8% more recombination breakpoints and, in comparison to the highest density GigaMUGA array on striatum samples from generations G21, G22, and G23, GBRS detected on average, 0.06% fewer breakpoints. We conclude that GBRS reconstructions have sensitivity comparable to the high density (150k markers) GigaMUGA platform. This implies that the number and positional distribution of expressed genes are sufficient to detect most of the recombination events in DO mice from these outbreeding generations.

[Figure 5 about here.]

GBRS increases the power of expression QTL mapping

Thousands of genes have expression levels that are influenced by genetic variation at or near the location of their coding sequences [Chick et al., 2016, Aguet et al., 2017]. These associations can be identified by mapping local gene expression QTL (eQTL), which are prevalent and generally stronger than eQTL that map to distant loci. Incorrect genotypes are likely to reduce the apparent effect of a local eQTL. Therefore, if the genotype probabilities that we use for eQTL mapping are more accurate, the association between eQTL genotype and gene expression should be stronger. In order to assess the mapping quality of GBRS genotypes relative to array-based genotypes, we first performed a full eQTL analysis with genotype probabilities obtained from GBRS and genotyping arrays. We then compared the magnitude of LOD scores from GBRS with MUGA, MegaMUGA, and GigaMUGA arrays for all eQTL peaks, local and distal (Figure 6, 7, and 8). When we use the original sample labels without fixing the sample mix-ups, we find that 6998, 5257, and 9682

genes have higher LOD scores and that 465, 520, and 856 genes have lower LOD scores using GBRS in comparison to MUGA, MegaMUGA, and GigaMUGA, respectively for the local eQTL (Figure 6a, 7a, and 8a). Similarly, for the distal eQTL, we find that 3158, 3605, and 3581 genes have higher LOD scores and that 997, 836, and 1015 genes have lower scores using GBRS compared to MUGA, MegaMUGA, and GigaMUGA, respectively (Figure 6b, 7b, and 8b). After we identify and correct sample mix-ups, we find that 5910, 3673, and 8051 genes have higher LOD scores and that 1662, 2238, and 2680 genes have lower LOD scores using GBRS in comparison to MUGA, MegaMUGA, and GigaMUGA, respectively for the local eQTL (Figure 6c, 7c, and 8c). For the distal eQTL, we find that 2843, 3341, and 3796 genes have higher LOD scores and that 2152, 3435, and 2016 genes have lower scores using GBRS compared to MUGA, MegaMUGA, and GigaMUGA, respectively (Figure 6d, 7d, and 8d). Correcting sample mix-ups improved the LOD score 23.1%, 25.3%, 19.5% for local and 11.1%, 11.1%, 9.5% for distal eQTL in MUGA, MegaMUGA and GigaMUGA, respectively. This shows the importance of identifying and correcting sample mix-ups before mapping analysis. We still find the GBRS genotype probabilities improve both the local and distal eQTL even after the removal of sample swaps. Overall, the local eQTL show greater improvements but distal eQTL results are also as good or better with the GBRS genotype probabilities. We note that the overall improvement in LOD scores of both local and distal QTL is greater in the MUGA and GigaMUGA comparisons. MegaMUGA delivered LOD scores similar to GBRS for the dataset we examined.

[Figure 6 about here.]

[Figure 7 about here.]

[Figure 8 about here.]

Combined estimation of genotype probabilities

Genotyping arrays have the advantages of a relatively even distribution of markers across the genome and higher density of coverage. GBRS uses genotyping information at expressed genes which are unevenly distributed and less dense than the most recent versions of the MUGA array. On the other hand GBRS markers are concentrated in some of the most important functional regions of the genome - coding regions. We hypothesized that combining genotype probabilities would draw on the strengths of both approaches. We evaluated the mapping results from genotype probability that combines GBRS and each genotyping array by comparing its LOD scores with GBRS-based results. For the local eQTL, we find that 5499, 3871, and 7068 genes have higher LOD scores and that 1990, 1925, and 3228 genes have lower LOD scores using GBRS in comparison to GBRS combined with MUGA, MegaMUGA, and GigaMUGA, respectively (Figure 6e, 7e, and 8e). For the distal eQTL, we find that 3164, 3258, and 3259 genes have higher LOD scores and that 2077, 1779, and 2894 genes have lower scores using GBRS compared to GBRS combined with MUGA, MegaMUGA, and GigaMUGA, respectively 6f, 7f, and 8f). On average, GBRS performed better than the combined genotype probabilities. It appears that GBRS offers genome coverage sufficient to identify most recombination events for the current outbreeding generations of the DO. It is puzzling that combining genotype probabilities did yield improved eQTL mapping, especially for distal eQTL. It would be worthwhile to look at the power of combined genotypes in other MPPs, especially those with a higher density of recombination events.

DISCUSSION

GBRS reconstructs the individual genomes of MPP individuals directly from RNA-Seq data and simultaneously quantifies total and allele-specific gene expression. It uses the weighted allocation

of reads aligned to a multi-way index of predicted founder transcripts to characterize the genotype at each expressed gene locus. Unlike other genotyping strategies that use sequencing data, GBRS does not rely on variant calling. GBRS can deliver the advantages of aligning RNA-Seq reads to individual genomes [Munger et al., 2014] without the need to construct a large number of individual alignment indices.

The HMM component of GBRS, can be trained using RNA-Seq data obtained from founder strain tissues to estimate a founder profile at each gene. The founder profile is compared to sample profiles of MPP individuals to determine the founder strain genotypes. Variants in the predicted transcript sequences of any given gene may not fully resolve the founder haplotypes but application of a forward-backward algorithm effectively borrows information from nearby genes to improve resolution. The founder profiles depend only on the predicted transcript sequences and therefore, are agnostic to experimental conditions and tissue type(s) in the training data — although it is desirable to obtain the highest possible representation of genes. Highly expressed genes will provide the most precise founder profiles, but deep sequencing of multiple tissues will ensure broad and accurate profiles for most genes. In the work presented here, we trained a GBRS HMM for DO mice with a single training set from liver tissue samples. Recently we have expanded the training set to include multiple tissues (see Data and Software Availability).

We obtained training data from inbred founder strain tissue samples, but we were able to apply GBRS to outbred samples by approximating the profiles for heterozygous genotypes as an equal mixture of the corresponding (homozygous) founder profiles. While this does not account for the possibility of allelic imbalance or non-allelic expression, we demonstrated that for DO mice, which are $\sim$85% heterozygous with respect to founder strain haplotypes, this approximation works well. It would be possible to include F1 hybrids in the training data but the number of combinations

required may be prohibitive. Alternatively, the HMM could be trained using the Baum-Welch algorithm [Baum, 1972] directly on the individual MPP sample data. We have not tested this strategy but expect that, in the absence of other training data, with carefully chosen initial values of the emission distribution parameters, this would be an effective method to train the HMM.

In our past experience, HMMs for genotype reconstruction are fairly robust to the magnitude of the transition probabilities. However, high transition rates can result in 'choppy' genome reconstructions with clusters of presumably false positive recombination breakpoints. Low transition rates could result in 'smoothing over' of small haplotype segments. On the whole results are best when the emission models are accurate and the gene/marker loci are dense relative to the expected rate of recombination breakpoints. GBRS may not perform well across genomic regions in which expressed genes are sparse. This could result in missing recombination events especially in distal regions of chromosome that lack support from neighboring gene-rich regions on one side.

Our evaluation of GBRS using eQTL mapping examines traits (gene expression) that are associated with the same expressed genes that we use as marker loci. This choice of evaluation data may be biased in favor of GBRS, and therefore, improvement in local eQTL LOD scores might be expected when using GBRS because the local genotypes are inferred directly from nearby gene expression. However, GBRS also outperforms the genotyping arrays at distal eQTL.

The current implementation of GBRS treats whole genes as units and the model does not account for recombination breakpoints within a gene. One solution to alleviate this would be to estimate genotype probabilities at the level of individual exons. In cases where recombination breakpoints do occur within a gene (or exon) the marginal genotype probabilities from the HMM will reflect uncertainty and will be more influenced by the genotypes at flanking genes.

We have demonstrated that GBRS as a stand-alone genotyping strategy can perform well in

comparison to genotyping arrays. However, GBRS precision is limited by the numbers of expressed genes in both the training and target samples — which is usually on the order of ∼10,000 — and the distribution of genes can be uneven across the genome. As the DO mice, and other outbred MPPs evolve, the density of recombination breakpoints will increase but the gene density and distribution does not. Perhaps, the most compelling use for GBRS is for quality control to detect and correct sample mix-ups. In cases where array or sequence-based genotyping has failed, GBRS provides a suitable replacement for the missing sample genotypes. When GBRS is applied in conjunction with another method of estimating genotype probabilities the results can be combined to improve the precision of either method alone. For these reasons we still advise the use of genotyping arrays or DNA sequenced based genotyping.

We have recently adapted GBRS to work with other types of sequencing data and see potential for applications to low-coverage DNA-Seq, ChIP-Seq, ATAC-Seq, and Hi-C. Of note, we have applied GBRS successfully to single-cell RNA-Seq deconvolute cells from sample mixtures and to reconstruct the individual genomes of each sample. The current implementation of GBRS is tailored to work with DO, Collaborative Cross and related mouse MPPs. We want to encourage and support the development of GBRS for applications to MPPs derived in other model systems through open-source software development platforms.

DATA AVAILABILITY

The founder mice liver RNA-Seq data is available at the Gene Expression Omnibus (GEO) with the accession ID GSE45684. DO mice liver RNA-Seq data is archived at the Short Read Archive under project number PRJNA35625. The striatum data is archived at Sequence Read Archive (Accession number will be provided as sonn as it is available).

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, required data files are available at ftp://churchill-lab.jax.org/software/GBRS/. R scripts for eQTL mapping is available at https://thejacksonlaboratory.github.io/Workflowr_Array_GBRS/index.html.

ACKNOWLEDGEMENT

We gratefully acknowledge the contribution of the Gene Expression Service at The Jackson Laboratory for expert assistance with the work described in this publication.

FUNDING

This study was funded by NIH R01 GM070686 to GAC. The striatum dataset was funded by NIH P50 DA039841 (Center for Systems Neurogenetics of Addiction) to EJC and VMP.

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.

[Baum, 1972] Baum, L. E., 1972. An inequality and associated maximization technique in statistical estimation for probabilistic functions of Markov processes. In Shisha, O., editor, *Inequalities III: Proceedings of the Third Symposium on Inequalities*, pages 1–8, University of California, Los Angeles. Academic Press.

[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., Simecek, P., Furlotte, N. A., Prins, P., Sen, Ś., Yandell, B. S., and Churchill, G. A., 2019. R/qtl2: Software for mapping quantitative trait loci with high-dimensional data and multiparent populations. *Genetics*, **211**(2):495–502.

[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.

[Dudbridge and Koeleman, 2004] Dudbridge, F. and Koeleman, B. P. C., 2004. Efficient computation of significance levels for multiple associations in large studies of correlated data, including genomewide association studies. *Am J Hum Genet*, **75**(3):424–435.

[Ferragina and Manzini, 2000] Ferragina, P. and Manzini, G., 2000. Opportunistic data structures with applications. In *Proceedings 41st Annual Symposium on Foundations of Computer Science*, pages 390–398.

[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.

[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.

[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.

[Love et al., 2014] Love, M. I., Huber, W., and Anders, S., 2014. Moderated estimation of fold change and dispersion for rna-seq data with deseq2. *Genome Biol*, **15**(12):550.

[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.

[Rabiner, 1989] Rabiner, L. R., 1989. A tutorial on hidden markov models and selected applications in speech recognition. *Proceedings of the IEEE*, **77**(2):257–286.

[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.

[Storey et al., 2020] Storey, J. D., Bass, A. J., Dabney, A., and Robinson, D., 2020. *qvalue: Q-value estimation for false discovery rate control*. R package version 2.22.0.

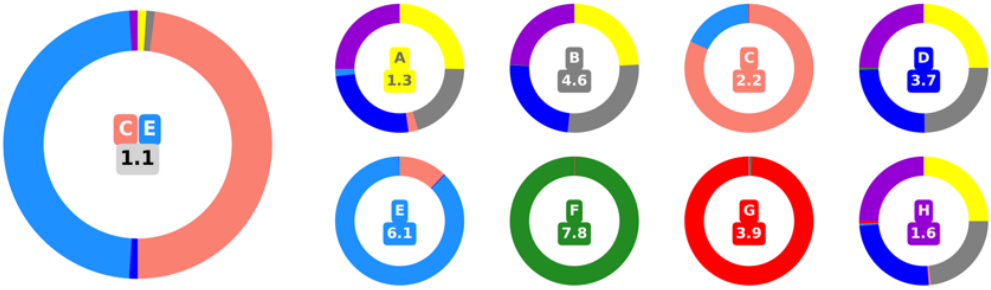
[Storey et al., 2004] Storey, J. D., Taylor, J. E., and Siegmund, D., 2004. Strong control, conservative point estimation and simultaneous conservative consistency of false discovery rates: a unified approach. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **66**(1):187–205.

[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.

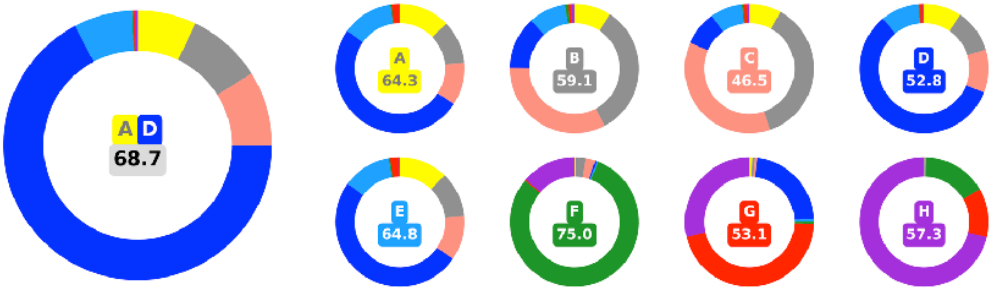
[Viterbi, 1967] Viterbi, A., 1967. Error bounds for convolutional codes and an asymptotically optimum decoding algorithm. *IEEE Transactions on Information Theory*, **13**(2):260–269.

[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.

[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: Founder profiles for genotyping diversity outbred mice. A founder profile for a gene is estimated by aligning RNA-Seq reads from the founder strain to a multi-way alignment index and applying weighted allocation to obtain the proportions of reads that align to each of the predicted founder transcripts. `GBRS` compares a sample profile from a DO individual to the founder profiles and predicts its genotype. (a) RNA-Seq reads from founder samples are allocated to the gene *Lad1* with proportions indicated by colored (smaller) circles on the right. The label at the center of each circle indicates the founder strain origin of the sample and the colors in the circular bar graphs are proportional to the weighted allocation of reads to the founder transcripts. For *Lad1*, the founder profiles B, D and H are identical because there are no polymorphisms that distinguish these strains. The founder profile A is nearly identical to B, D, and H profiles due to high sequence similarity. The founder profiles C and E are closely related and distinct from the other six founders. These two founder sequences identify themselves after weighted allocation as they are able to attract a majority of reads originating from themselves. Founder profiles F and G each uniquely identify these strains due to their high levels of divergence. According to the sample profile from a DO individual (a larger circle on the left), `GBRS` predicts its genotype to be CE heterozygote as it is an equally-weighted mixture of the C and E founder profiles. (b) At the *Mrpl15* gene locus, founder profiles A, D and E are identical as are profiles B and C. It is interesting to note that RNA-Seq reads from the A and E founder strains show a higher proportion of reads allocated to predicted transcripts from strain D presumably due to inaccurate annotation. The profile of heterozygote — AD, AE, and DE — would be identical to the A, D and E homozygotes for this gene. Given the sample profile of a DO individual (a large circle on the left), it is difficult to call its genotype by just comparing it to the founder profiles. `GBRS` gets around this challenge by combining information from the neighboring genes, and predicts the genotype, for example, AD for this example. The numbers in the sample profiles are the expression levels in TPM.

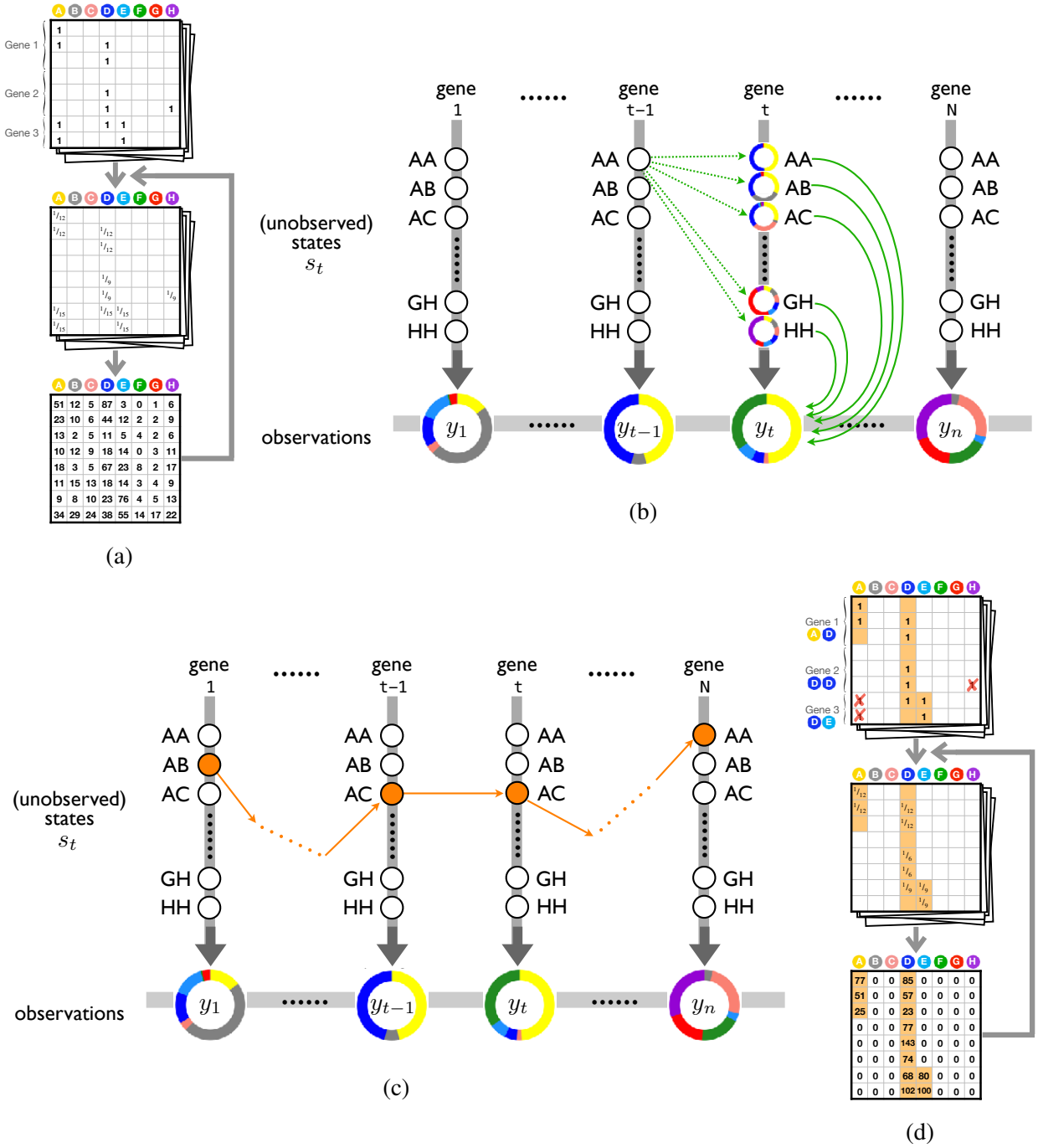


Figure 2: An overview of GBRS algorithm. (a) GBRS employs multi-way alignment to estimate the MPP sample profiles — the proportion of reads from an MPP individual that are allocated to founder transcripts at each gene. (b) A HMM forward-backward algorithm uses the sample profiles as input data and estimates genotype state probabilities at each gene. The emission probabilities (green solid arrows) of the HMM are estimated from the founder profiles. Transition probabilities (green dotted arrows) are derived theoretically [Broman et al., 2015] based on intergenic distances. (c) A Viterbi algorithm identifies the maximum-probability path across genotype states (orange arrows) to define a diploid genotype at each gene loci. (d) The diploid genome reconstruction is used to mask components of the multi-way alignment index (boxes denoted with 'X') and repeats the weighted allocation to generate diploid allele-specific expression estimates.

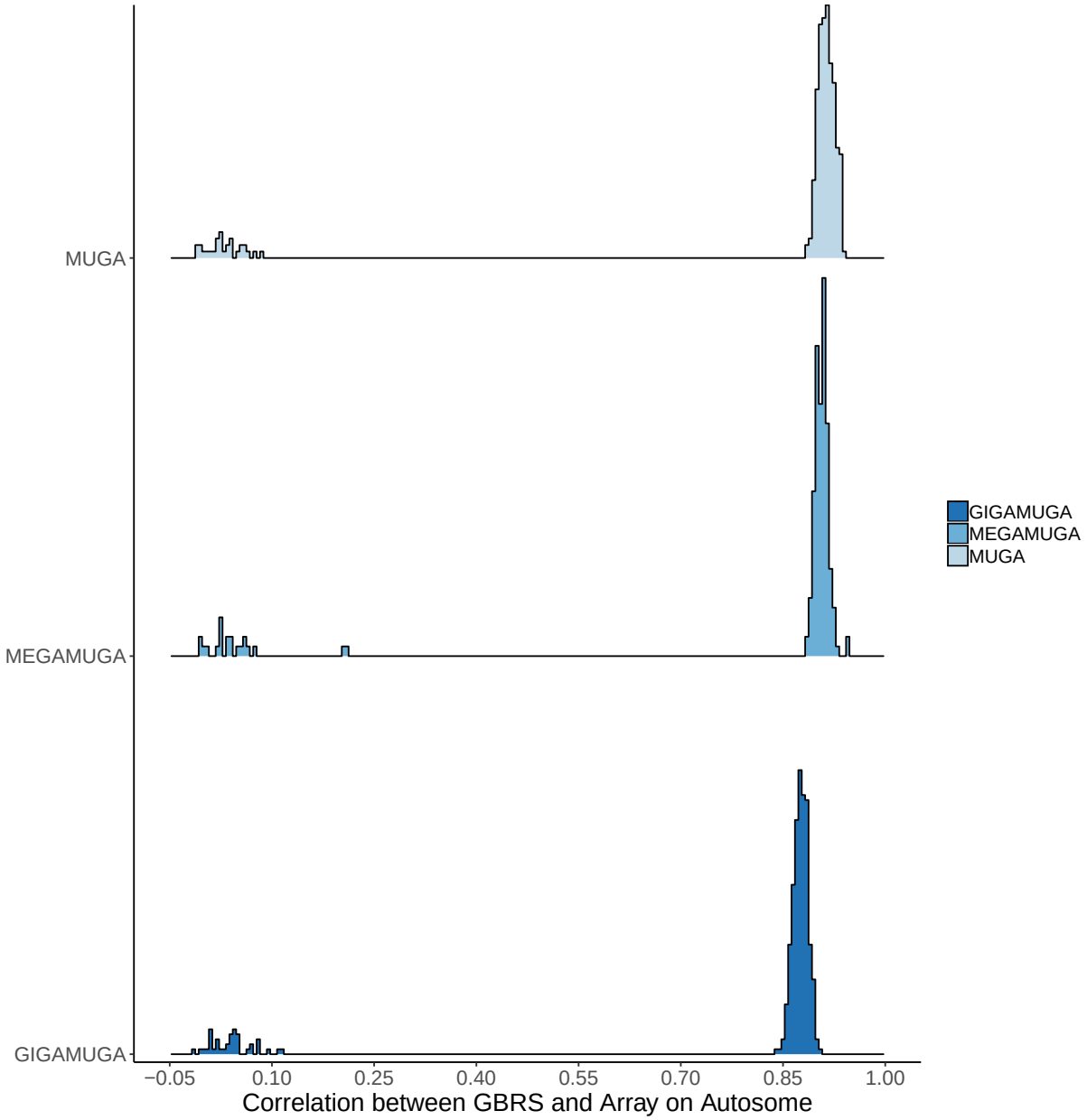
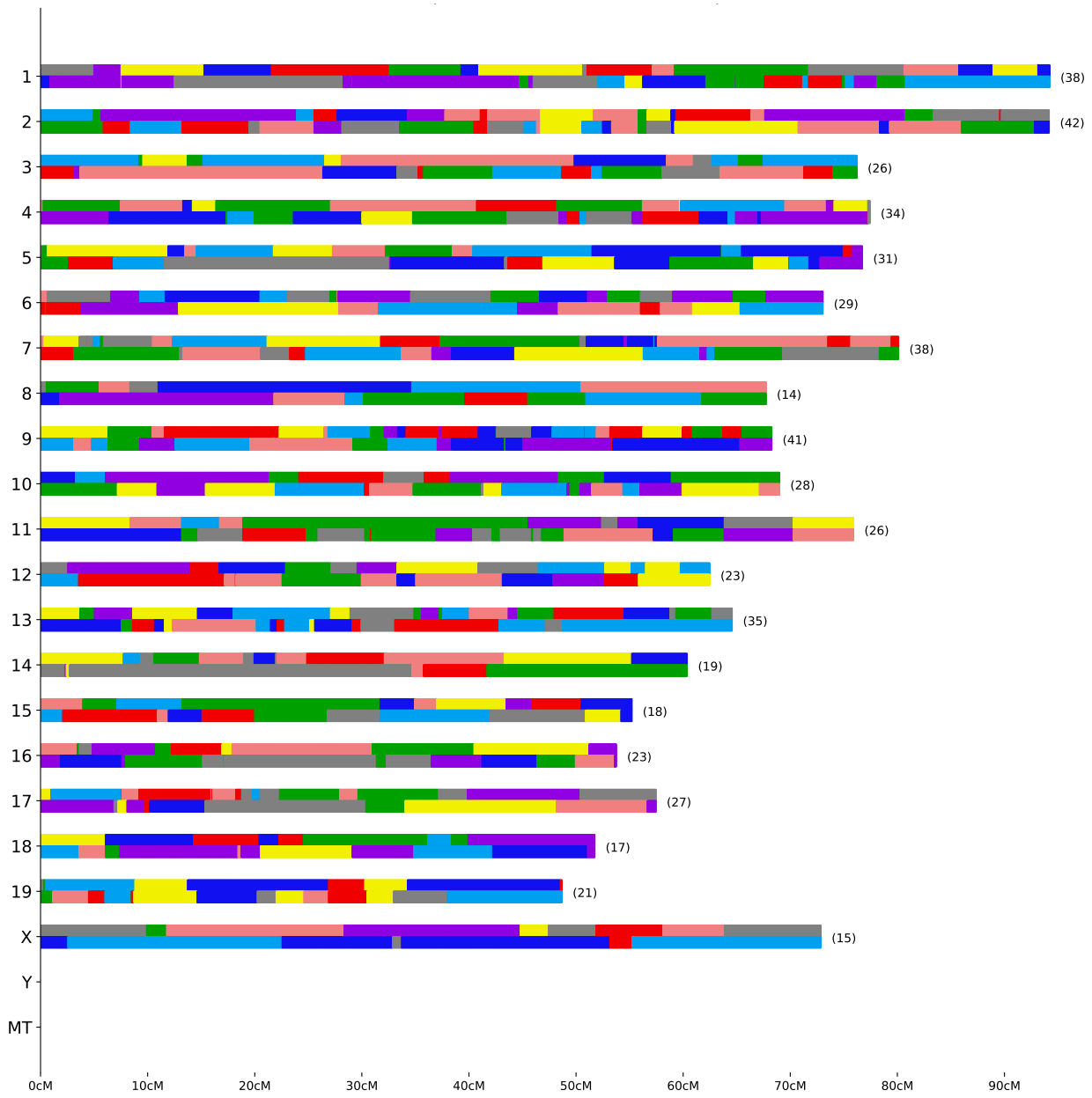
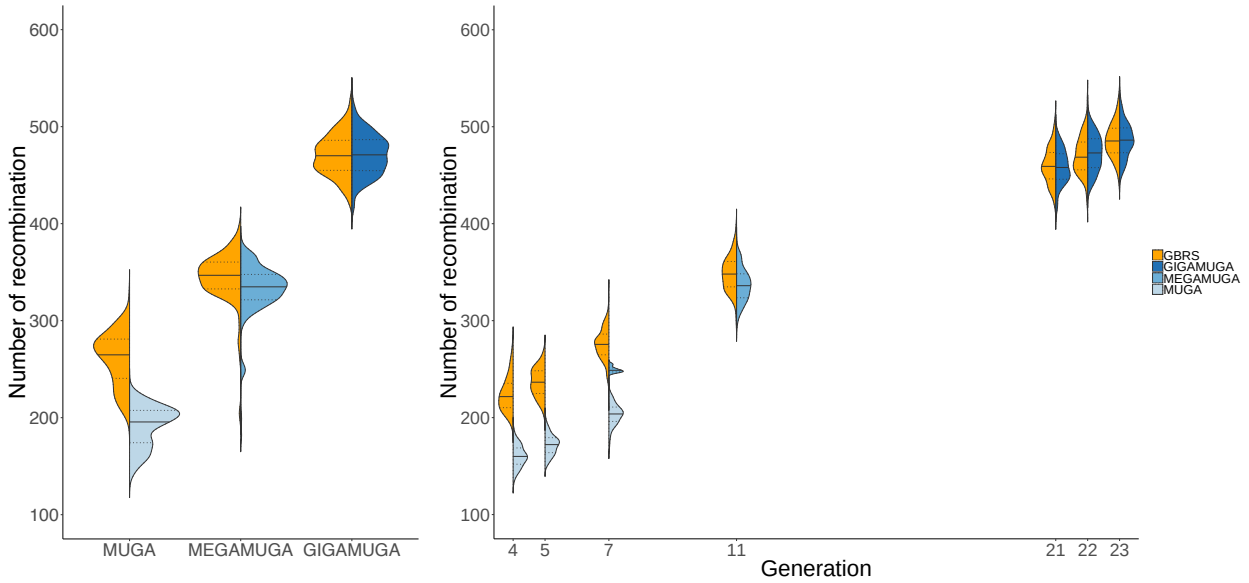


Figure 3: Correlation of GBRS and array-based genotype probabilities. Individual genomes reconstructed by GBRS are concordant with MUGA (light blue), MegaMUGA (blue), and GigaMUGA (dark blue) reconstructions ($r > 0.8$) for the majority of MPP samples. The median of Pearson correlation of genotype probabilities is 0.876, 0.862 and 0.834 respectively. Approximately 10% of samples in each group show low Pearson correlation, indicating that the DNA and RNA samples do not correspond due to possible mix ups in sample handling.



(a)

Figure 4: A Viterbi reconstruction of the haplotype mosaic of an individual MPP genome. GBRS reconstructed the diploid genome for a female sample from the striatum data set (ID 8684) and identified 545 recombination breakpoints. The founder origins of haplotype blocks in indicated by color coding.



(a) (b)

Figure 5: GBRS detects recombination breakpoints. (a) The number of recombination breakpoints identified by GBRS (orange) in comparison to MUGA (light blue), MegaMUGA (blue), and GigaMUGA (dark blue) genotyping arrays is shown for each array type. (b) The numbers of recombination breakpoints detected by GBRS and by each genotyping array is shown as a function of outbreeding generation of the DO samples.

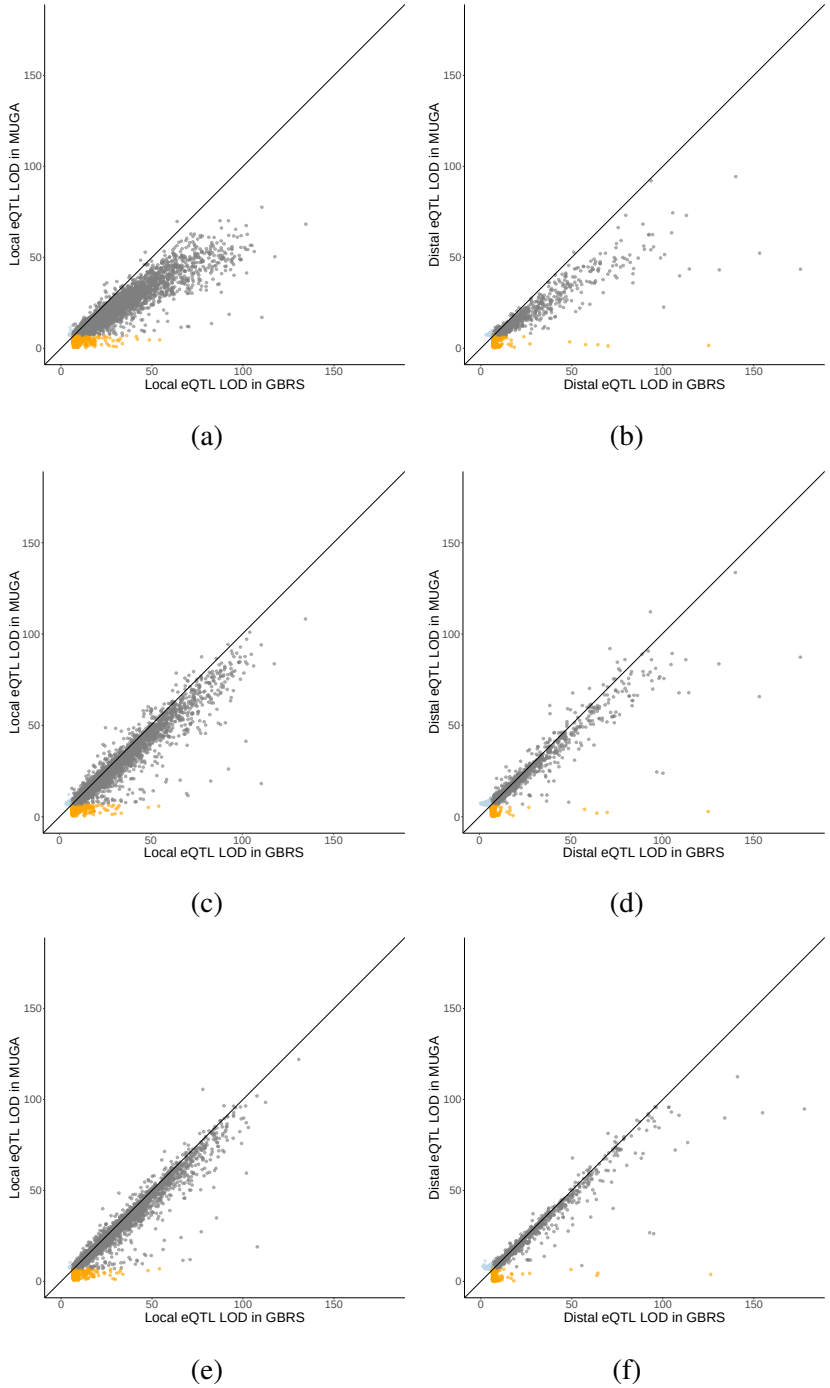


Figure 6: LOD scores for eQTL obtained with GBRS and MUGA array genotype probabilities. The LOD scores of all local (a,c,e) and distal (b,d,f) eQTL that exceed a suggestive significance threshold (LOD>6) are shown in comparison to LOD scores obtained using genotypes from before correcting sample mix-ups (a)(b), after correcting sample mix-ups (c)(d), and the MUGA genotype probability combined with GBRS genotype probability (e)(f). An identity line is drawn on each scatterplot for reference. Points to the right and below this line indicate better performance with GBRS.

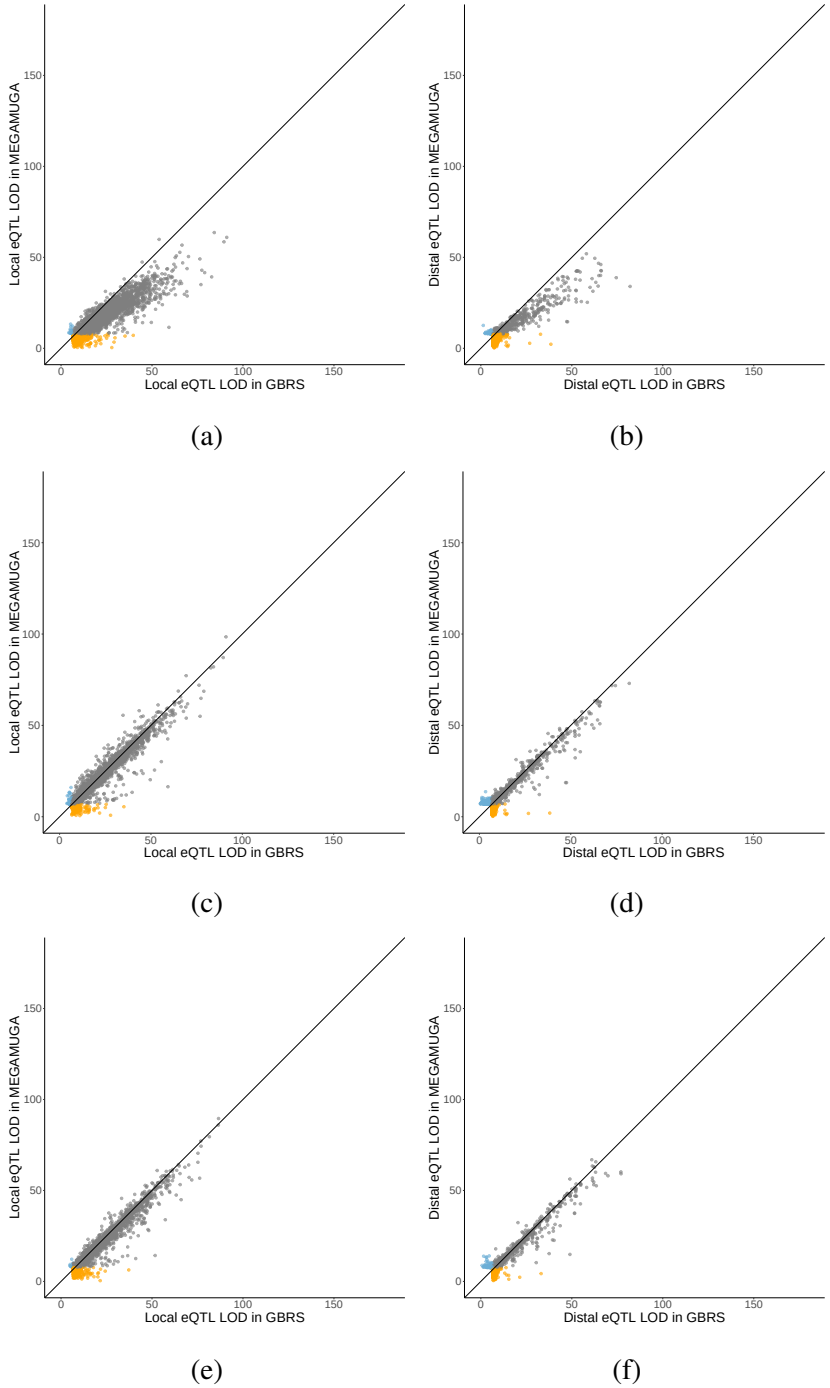


Figure 7: LOD scores for eQTL obtained with GBRS and MegaMUGA array genotype probabilities. The LOD scores of all local (a,c,e) and distal (b,d,f) eQTL that exceed a suggestive significance threshold (LOD>6) are shown in comparison to LOD scores obtained using genotypes from before correcting sample mix-ups (a)(b), after correcting sample mix-ups (c)(d), and the MegaMUGA genotype probability combined with GBRS genotype probability (e)(f). An identity line is drawn on each scatterplot for reference. Points to the right and below this line indicate better performance with GBRS.

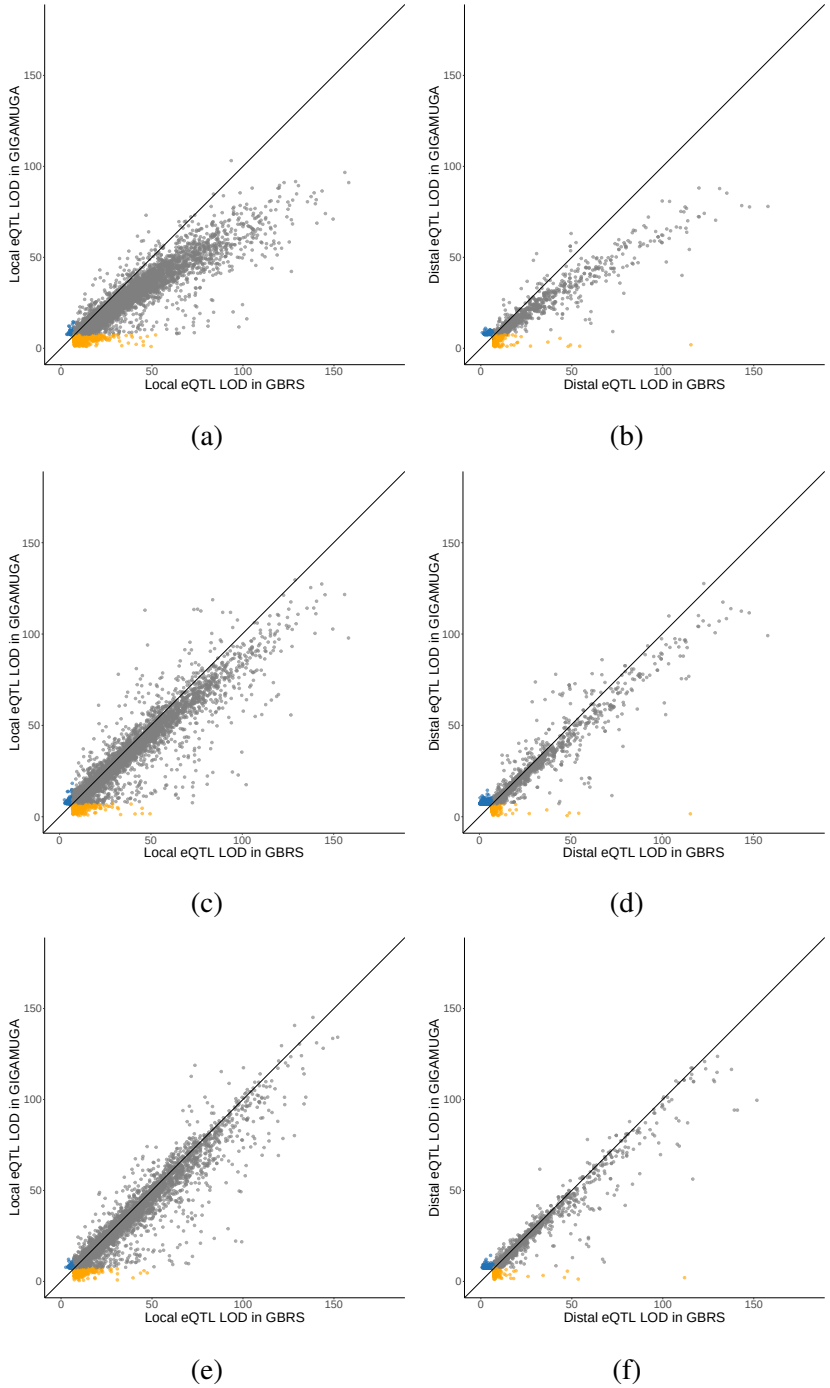


Figure 8: LOD scores for eQTL obtained with GBRS and GigaMUGA array genotype probabilities. The LOD scores of all local (a,c,e) and distal (b,d,f) eQTL that exceed a suggestive significance threshold (LOD>6) are shown in comparison to LOD scores obtained using genotypes from before correcting sample mix-ups (a)(b), after correcting sample mix-ups (c)(d), and the GigaMUGA genotype probability combined with GBRS genotype probability (e)(f). An identity line is drawn on each scatterplot for reference. Points to the right and below this line indicate better performance with GBRS.

Table 1: Sample mismatches identified.

(a)

	match	mismatch	fixed	failed
MUGA	275	29	24	5[a]
MegaMUGA	184	21	18	3[b]
GigaMUGA	358	39	26	13[c]

[a] 4 duplications and 1 new sample
[b] 1 duplication and 2 new samples
[c] 12 duplications and 1 new sample