Genome Reconstruction of Diversity Outbred Mice

Daniel M. Gatti

03 October 2013

1 Introduction

Diversity Outbred (DO) mice are dervied from eight inbred founders: A/J, C57BL/6J, 129S1/SvImJ, NOD/ShiLtJ, NZO/HlLtJ, CAST/EiJ, PWK/PhJ, WSB/EiJ. DO genomes are a heterozygous mixture of these eight founders. Their genomes can be reconstructed using a hidden Markov model (HMM) in which the hidden states are the 36 possible unphased genotypes (8 homozygotes and 28 heterozygotes). In order to estimate the genotype probabilities, the HMM requires input data that allows it to estimate which founders contributed to each sample at each marker.

Note that you will need an internet connection to run this vignette.

2 Data Setup

Below, we load in data from an example data package. The files contain X and Y intensities and allele calls.

```
> library(DOQTL)
```

```
R/QTLRel is loaded
```

```
> wd = tempdir()
> library(MUGAExampleData)
> data(x)
> data(y)
> data(geno)
> data(pheno)
```

3 Intensity Based Genome Reconstruction

Genome reconstruction can be performed using the X and Y allele intensities from the MUGA or MegaMUGA. We have found that, using the MUGA or MegaMUGA in the DO, array intensities produce better genome reconstructions than the allele calls. The reason for this is that the intensities can contain more than three genotype groups and this aids in narrowing down the genotype categories.

When you have allele intensities, the overall X and Y chromosome intensities may be used to assign the sex of each mouse. This is a good quality control check and these results should be compared with the sex recorded during your experiment.

```
> load(url("ftp://ftp.jax.org/MUGA/muga_snps.Rdata"))
> sex = predict.sex(x = x, y = y, snps = muga_snps, plot = T)
```

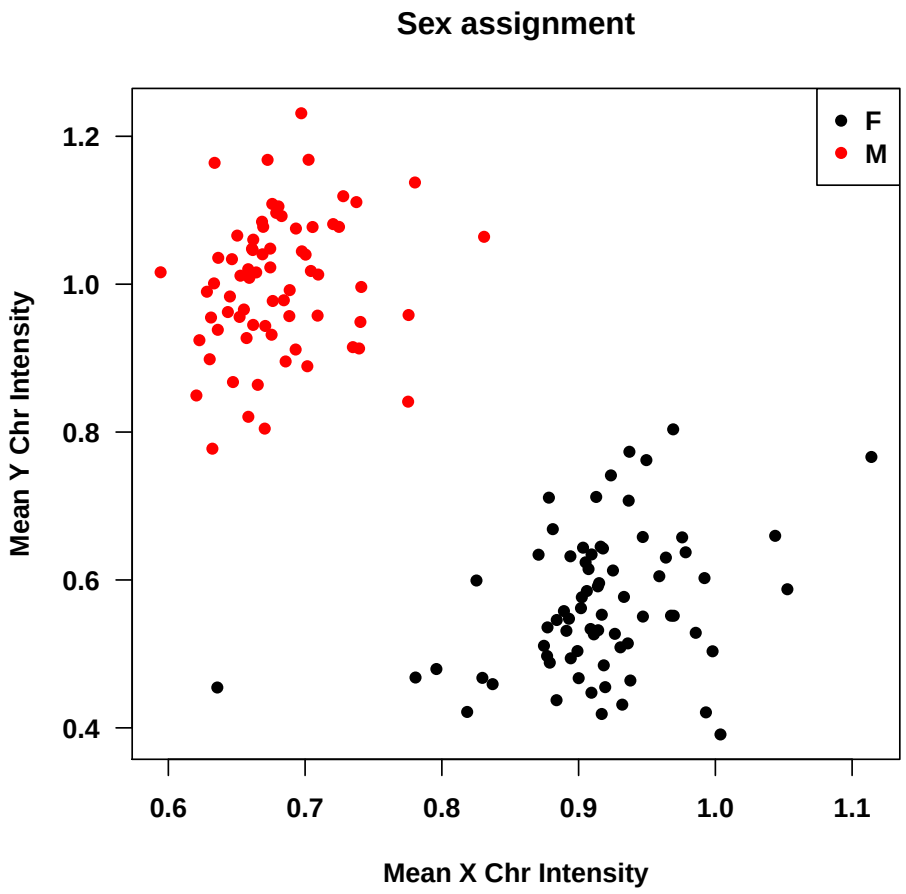

Figure 1: Mean X and Y chromosome intensity for DO samples, colored by sex. The females (black) have a higher mean X intensity than the males (red). Conversely, the males have higher mean Y chromosome intensity. The single female at (0.63, 0.43) may be an XO female.

When DO mice are genotyped using the MUGA or MegaMUGA, DOQTL has access to the SNP locations and founder genotype data. You must supply the sample intensities, sex and DO outbreeding generation in the `data` argument.

data is a named list containing: data

x: numeric matrix containing X allele intensities with samples in rows and markers in columns. Sample names in rownames and SNP IDs in colnames.

y: numeric matrix containing Y allele intensities with samples in rows and markers in columns. Sample names in rownames and SNP IDs in colnames.

sex: character vector containing either `M` or `F`. Sample names in names.

gen: character vector containing 'DO' followed by teh DO outbreeding generation, with no space between them (i.e. DO4). Sample names in names.

The DO outbreding generation should be requested when ordering mice.

```
> gen = paste("DO", gsub("^G|L[12]$", "", pheno$Gen), sep = "")
> names(gen) = rownames(pheno)
> gen = gen[names(gen) %in% names(sex)]
> gen = gen[match(names(sex), names(gen))]
> stopifnot(all(rownames(x) == names(sex)))
> stopifnot(all(rownames(x) == names(gen)))
> data = list(x = x, y = y, sex = sex, gen = gen)
> calc.genoprob(data = data, chr = "all", output.dir = wd, array = "muga")
```

`calc.genoprob` uses the X and Y intensity data to reconstruct the DO genomes in terms of the 8 founder haplotypes. It then writes out one file for each sample that contains the 36 estimated genotype probabilities at each marker. It also writes out the model parameters and the 8 founder haplotype contributions for each sample at each locus.

The output directory will now contain a single file for each sample containing the genotype probabilities at each marker. It will also contain a file with founder haplotype contributions for all samples and all markers called `model.probs.Rdata`. This is the file that will be used for QTL mapping.

```
> load(paste(wd, "F15.genotype.probs.Rdata", sep = "/"))
> plot.genotype.max(prsmth = prsmth, snps = muga_snps, main = "F15")
```

We can summarize the recombinations using `summarize.genotype.transitions()`.

```
> recomb = summarize.genotype.transitions(path = wd, snps = muga_snps)
```

This uses the maximum probability at each locus to determine the recombinations in each sample. The function summarizes the recombination locations and genotypes for each sample.

```
> head(recomb)
```

```
                                                sample prox.geno dist.geno
1 C:\\Users\\dgatti\\AppData\\Local\\Temp\\Rtmp2NKDEm/F01        DE        CD
2 C:\\Users\\dgatti\\AppData\\Local\\Temp\\Rtmp2NKDEm/F01        CD        DG
3 C:\\Users\\dgatti\\AppData\\Local\\Temp\\Rtmp2NKDEm/F01        DG        EG
4 C:\\Users\\dgatti\\AppData\\Local\\Temp\\Rtmp2NKDEm/F01        EG        GG
```

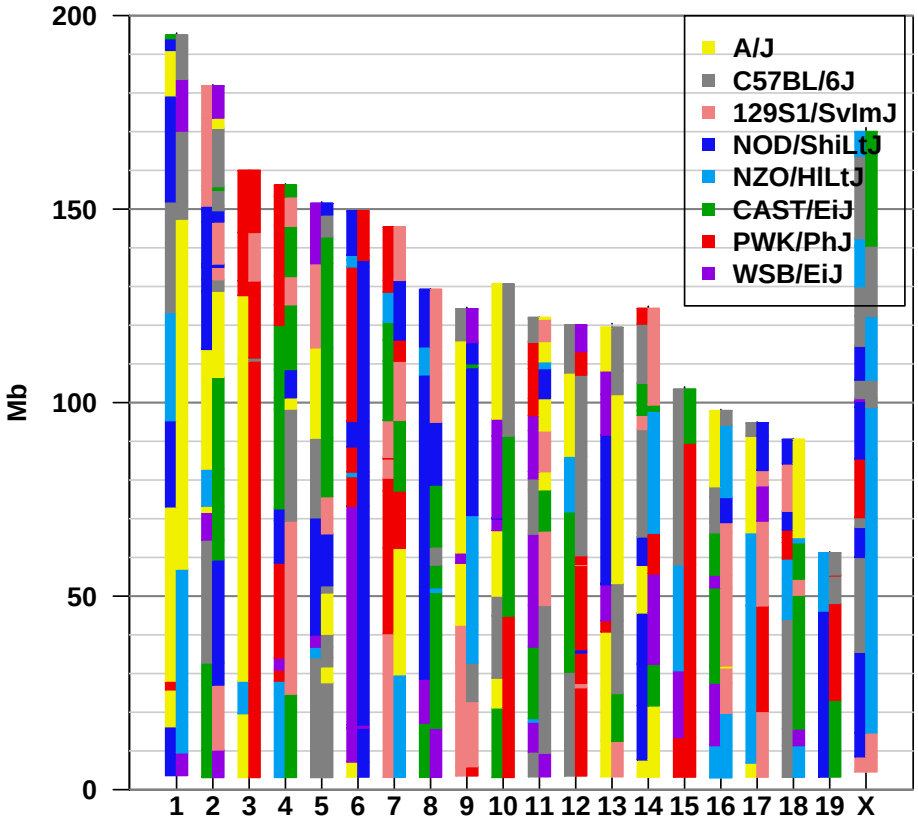

Figure 2: Genome reconstruction of DO mouse using array intensities.

```
5 C:\\Users\\dgatti\\AppData\\Local\\Temp\\Rtmp2NKDEm/F01        GG        FG
6 C:\\Users\\dgatti\\AppData\\Local\\Temp\\Rtmp2NKDEm/F01        FG        AG
      prox.snp     dist.snp chr prox.loc dist.loc
1  JAX00242845  JAX00000980   1 15.71881 16.16213
2 UNC010048083 UNC010530301   1 23.01304 23.36290
3  JAX00001692 UNC010878755   1 25.69794 25.82519
4 UNC010537386 UNC010054938   1 28.15468 28.61041
5 UNC010542162 UNC010442706   1 34.95175 35.10234
6 UNC010445227 UNC010484549   1 49.98282 50.09254
```

We can look at the average number of recombinations per sample.

```
> mean(table(recomb[,1]))

[1] 250.0284
```

4 Allele Call Based Genome Reconstruction

When DO mice are genotyped using the MUGA or MegaMUGA, DOQTL has access to the SNP locations and founder genotype data. You must supply the sample intensities, sex and DO outbreeding generation in the `data` argument.

data is a named list containing: data

geno: character matrix containing allele calls with samples in rows and markers in columns. Sample names in rownames and SNP IDs in colnames.

sex: character vector containing either `M` or `F`. Sample names in names.

gen: character vector containing 'DO' followed by the DO outbreeding generation, with no space between them (i.e. DO4). Sample names in names.

The DO outbreding generation should be requested when ordering mice.

```
> sex = as.character(pheno$Sex)
> names(sex) = rownames(geno)
> gen = paste("DO", gsub("^G|L[12]$", "", pheno$Gen), sep = "")
> names(gen) = rownames(geno)
> sex = sex[names(sex) %in% rownames(x)]
> gen = gen[names(gen) %in% rownames(x)]
> data = list(geno = geno, sex = sex, gen = gen)
> calc.genoprob(data = data, chr = "all", output.dir = wd, array = "muga",
+               method = "allele")
```

The output directory will now contain a single file for each sample containing the genotype probabilities at each marker. It will also contain a file with founder haplotype contributions for all samples and all markers called `model.probs.Rdata`. This is the file that will be used for QTL mapping.

```
> load(url("ftp://ftp.jax.org/MUGA/muga_snps.Rdata"))
> load(paste(wd, "F15.genotype.probs.Rdata", sep = "/"))
> plot.genotype.max(prsmth = prsmth, snps = muga_snps, main = "F15")
```

Note that the genotype resonstructions look quite different. We can summarize the number of recombinations as above.

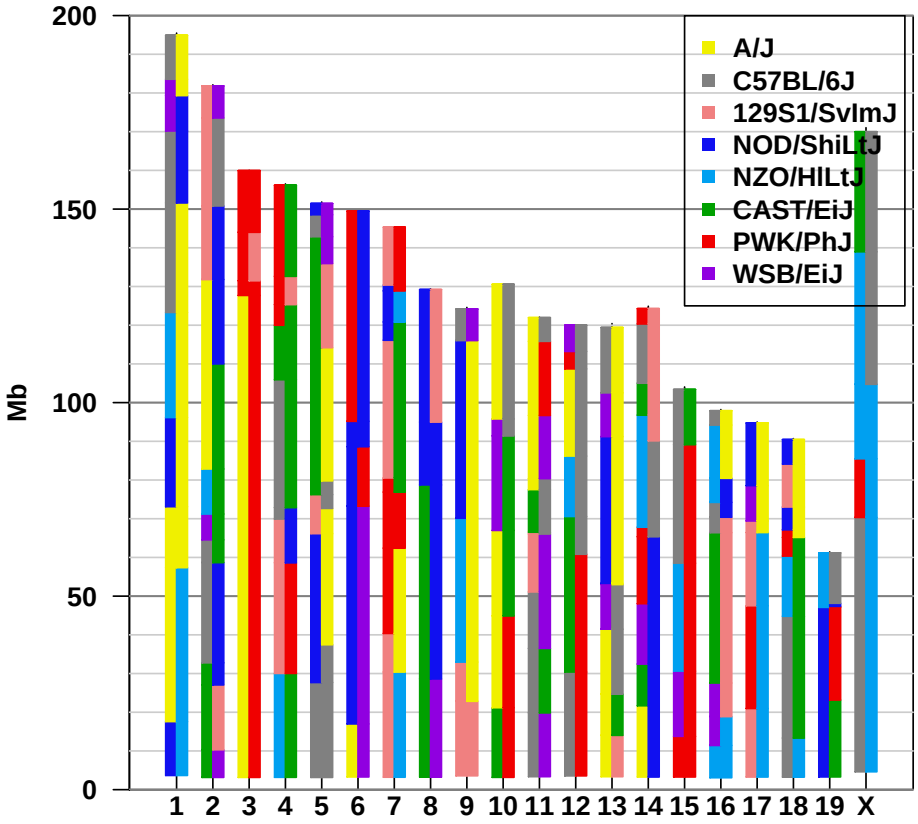

Figure 3: Genome reconstruction of DO mouse using array allele calls.

```
> recomb = summarize.genotype.transitions(path = wd, snps = muga_snps)

> mean(table(recomb[,1]))

[1] 133.227
```

Whereas there were an average of 250 recombinations per sample above, now there are 133. This is consistent with the observations of several groups that the array intensities on the MUGA provide better genome reconstructions than that allele calls.