Linkage Disequilibrium and Inversion-Typing of the *Drosophila melanogaster* Genome Reference Panel

David Houle[1]

Eladio J. Márquez[2]

Department of Biological Science

Florida State University

Tallahassee, FL 32308

RUNNING TITLE: Disequilibrium in the DGRP

KEYWORDS: Linkage disequilibrium, inversion

[1] Corresponding author. Department of Biological Science, Florida State University, Tallahassee, FL 32306-4295, USA. E-mail: dhoule@bio.fsu.edu

[2] Current address: The Jackson Laboratory for Genomic Medicine, Farmington, CT 06030, USA

ABSTRACT

We calculated the linkage disequilibrium between all pairs of variants in the Drosophila Genome Reference Panel, and make available the list of all highly correlated SNPs for use in association studies. Seventy-three percent of variant SNPs are correlated at $r^2>0.5$ with at least one other SNP, and the mean number of correlated SNPs per variant over the whole genome is 64.9. Disequilibrium between distant SNPs is also common when minor allele frequency (MAF) is low: 24% of SNPs with MAF<0.1 are highly correlated with SNPs more than 100kb distant. While SNPs within regions with polymorphic inversions are highly correlated with somewhat larger numbers of SNPs, and these correlated SNPs are on average farther away, the probability that a SNP in such regions is highly correlated with at least one other SNP is very similar to SNPs outside inversions. Previous karyotyping of the DGRP lines has been inconsistent, and we used LD and genotype to investigate these discrepancies. When previous studies agreed on inversion karyotype, our analysis was almost perfectly concordant with those assignments. In discordant cases, and for inversion heterozygotes, our results suggest errors in two previous analyses, or discordance between genotype and karyotype. Heterozygosities of chromosome arms are in many cases surprisingly highly correlated, suggesting strong epsistatic selection during the inbreeding and maintenance of the DGRP lines.

INTRODUCTION

The Drosophila Genome Reference Panel (DGRP; Mackay et al. 2012) is a set of sequenced inbred lines derived from a single outbred population of *Drosophila melanogaster*. The DGRP has been used for a series of genome-wide association (GWA) studies on a wide variety of

phenotypes. Linkage (gametic-phase) disequilibrium (LD) is a challenge to all GWA studies, as it confounds the signal from variant sites (we call these SNPs for brevity) that cause phenotypic variation with those that are genetically correlated with the causal variant but that do not have effects on the phenotype. The nature of the DGRP in many ways minimizes the presence of LD relative to vertebrates or to line-cross-derived mapping populations. The DGRP lines are drawn from a natural population with large effective size, as shown by the low level of structure within the population. Mackay et al. (2012) confirmed that the average LD drops very rapidly with distance between SNPs, to an average squared correlation $r^2<0.2$ at just 10 base pairs on the autosomes. This result might suggest that the overall impact of LD on GWAS results will be low.

More detailed analysis (Huang et al. 2014) shows that there is nevertheless substantial LD within the DGRP for two major reasons. First a total of 16 alternate chromosomal inversion karyotypes are present in the DGRP (Corbett-Detig and Hartl 2012; Huang et al. 2014; Langley et al. 2012). Huang et al. (2014)'s more complete analysis suggests that three of these are fixed in seven or more lines. These more common inversion types are substantially differentiated from the Standard karyotypes, and cause LD (Corbett-Detig and Hartl 2012; Langley et al. 2012; Huang et al. 2014). Second, rare SNPs have a substantial likelihood of being highly correlated with SNPs that are more than 1 kb distant.

The presentation of Huang et al. (2014) documents the problem of LD, but to interpret associations between SNPs and phenotypes in the DGRP we need to know whether particular SNPs that are implicated are correlated with other SNPs or inversions, and where those correlated sites are. We calculated the LD between all pairs of SNPs in the 205 Freeze2 DGRP lines, and provide a comprehensive list of polymorphic sites in substantial LD with inversions

and with other sites throughout the genome. The cytogenetic karyotype assignments in Huang et al. (2014) do not always agree with other PCR-based or sequence-based assignments in two other papers (Corbett-Detig and Hartl 2012; Langley et al. 2012), and we use genotypic data to investigate why these assignments differ.

METHODS

We used Freeze 2 genotype calls for the DGRP lines obtained from ftp://ftp.hgsc.bcm.edu/DGRP/freeze2_Feb_2013/. We used only sites with quality score >-1, at sites with exactly two alternative types. For LD calculations, heterozygous calls were treated as missing data. We excluded sites where the minor allele count is 4 or fewer, or the number of missing calls was greater than 85. This left 2,640,422 sites for analysis. The median allele frequency was MAF=0.132, and the median number of lines scored at each SNP was 194 out of 205 possible.

We parameterized linkage disequilibrium (LD) as the product-moment correlation r^2 (Hill and Robertson 1966). We decided to identify all correlations between SNPs with $r^2>0.5$. We realized that only SNPs with similar minor allele frequencies (MAF) could be highly correlated, and used this fact to minimize the number of SNP pairs whose correlations were calculated. We first calculated the maximum possible correlation between two SNPs with differing minor allele frequencies over all possible combinations of MAF among 205 lines. For each possible value of MAF, we identified the largest MAF at a second SNP for which r^2 could be greater than 0.5. We then fit a quadratic equation to that upper limit, and adjusted the intercept of that equation so that all values MAF that could give values of $r^2>0.5$ were below this limit. To scan for LD over the whole genome, we then binned SNPs into frequency classes to the nearest 0.01. Starting with

the lowest frequency bin, whose rounded frequency is p_b, we calculated correlations between all SNPs in the focal bin and those SNPs with *MAF* below the empirically determined limit $p_b + 0.005 < \mathrm{MAF} < 0.01 + 1.875 p_b - 1.17 (p_b)^2$. This process was repeated for each bin. We retained a list of all pairs of SNPs with r^2>0.5. This algorithm will miss a small number of highly correlated SNPs that have missing genotype information in many lines.

Huang et al. (2014) reported that three rare inversion karyotypes were fixed in seven or more DGRP lines (In(2L)t, In(2R)NS, In(3R)Mo), while no other karyotype was fixed in more than four lines. These karyotype assignments are sometimes in disagreement with the sequence-based assignments of karyotype reported by Corbett-Detig and Hartl (2012) and Langley et al. (2012). We checked these characterizations statistically using the following approach. We assembled genotypic data from Freeze 2 as above, but including heterozygous assignments, then excluded SNPs with five or more missing genotype assignments. Missing assignments in the remaining SNPs were assigned as the common allele to provide complete genotypic data. Using the results from the genome-wide LD results, we then obtained a list of the SNPs that are inside the inversion breakpoints of the three common alternative karyotypes (Corbett-Detig and Hartl 2012; Corbett-Detig et al. 2012), and that had LD r^2>0.5 with at least 200 other SNPs more than 100k sites distant. These SNPs are likely to be characteristic of inversion-types. This provided a sample of 16,001 SNPs on chromosome 2L, 6,520 on 2R and 4,697 on 3R. We conducted separate principal components analyses (PCA) of the genotypes for each inversion, and used the scores on PC1 to diagnose which genotypes are characteristic of each karyotype. We also calculated the proportion of SNPs that were scored as heterozygous for chromosomal regions defined by the inversion breakpoints.

Calculations were carried out in SAS version 9.3 for Windows and Unix (SAS Institute 2011).

RESULTS

To characterize linkage disequilibrium (LD), we generated a list of all pairs of variable sites (SNPs for brevity) whose allelic identities are correlated with $r^2>0.5$. We use r^2 as our measure of disequilibrium (Hill and Robertson 1966) because this is the most appropriate indicator of the likelihood that analyses of pairs of SNPs will yield similar results. The full lists of SNPs with $r^2>0.5$ by chromosome arm is available at http://bio.fsu.edu/~dhoule/Downloads/Freeze2205LD.zip. (The relatively small file with just the correlations for the 4th chromosome is also available, bio.fsu.edu/~dhoule/Downloads/HiRsq 205 chrom4.csv, to provide a readily downloadable example.) We refer to pairs of SNPs correlated at $r^2>0.5$ as "highly correlated," and SNPs more than 100kb apart as "distant". Seventy-three percent of all SNPs were highly correlated with at least one other SNP. Nine and a half percent of all SNPs are highly correlated with at least one distant SNP, and 1% are highly correlated with a SNP on another chromosome.

Fig. 1 shows the probability that a SNP is highly correlated with at least one other SNP, thus complicating its interpretation in an association study. SNPs were classified as inside or outside the breakpoint of the three inversions fixed in more than four DGRP lines (In(2l)t, In(2R)NS, and In(3R)Mo). For local disequilibrium, it makes little difference whether a SNP is in an inversion or not, but SNPs in inversions are more likely to be in high LD with a SNP that is

distant (in the Standard gene order). The inversion karyotypes themselves have frequencies of 0.1 or less (see below), precluding SNPs characteristic of inversions from being highly correlated with high MAF variants. More importantly, simply excluding SNPs within inversions does not appreciably reduce the likelihood that a variant will be highly correlated with at least some other SNPs, nor preclude those highly correlated SNPs from being distant.

Combinatoric considerations suggest that disequilibrium will be particularly common for SNPs with low minor allele frequencies (MAF). We investigate this in Figure 1 for all pairs of SNPs, and all distant pairs. The numbers of highly correlated pairs are indeed substantially higher at low MAF. The difference is particularly large for distant SNPs. When MAF<0.2, 19% of SNPs are correlated with a distant variant; when MAF<0.1, 24% of SNPs are. On the other hand, the mean number of highly correlated SNPs per SNP is still substantial at all allele frequencies. This is consistent with random disequilibrium due to the very large number of SNPs with low MAF (Mackay et al. 2012), and to the smaller number of permutations that can lead to a low MAF. We refer to this as rarity disequilibrium. Of the 2.4 million SNPs in this analysis, 50% have MAF<0.13, and 25% have MAF<0.057. Figure S1 suggests that inversions may compound the effects of linkage and rarity disequilibrium, as the mean number of highly correlated SNPs inside inversions is substantially higher when MAF is less than 0.2. This is particularly so for SNPs distant from the focal variant.

The variance in the number of correlated SNPs is high and skewed towards smaller numbers, so that the mean number of correlated SNPs is quite a bit higher than the median, as shown in Figure 2. The median number of correlated SNPs across the genome nevertheless ranges from 9 for most of the lowest MAF classes to 3 for the high MAF SNPs. The median number of SNPs greater than 10^5 bp away that are in high LD is 0. The median number of

variants correlated at $r^2>0.5$ is virtually identical between SNPs inside and outside of inversions (not shown).

The cytogenetic analyses of Huang et al. (2014) found at least seven DGRP lines fixed for each of the common cosmopolitan inversions In(2L)t, In(2R)NS, and In(3R)Mo, and many more lines that were heterozygous for these karyotypes. The remaining inversion karyotypes were fixed in four or fewer lines. Two previous studies (Corbett-Detig and Hartl 2012; Langley et al. 2012) also inversion-typed a subset of the DGRP lines using PCR- and/or next-generation-sequence-based assay. These two studies were consistent in their assignments, so we refer to them collectively as CD-L. CD-L and Huang et al. both scored 501 chromosome arms for homozygotes of the three common inversions In(2L)t, In(2R)NS, and In(3R)Mo. Ten of these assignments are in conflict, and 459 are in agreement. The remaining 32 were scored as heterozygotes by Huang et al. but were not examined by Langley et al. (2012). Corbett-Detig and Hartl (2012) state they did detect heterozygotes for inversions, but only reported lines positively identified as inversion homozygotes.

Because each of these studies makes clear that inversion karyotypes are usually substantially differentiated from each other, we reasoned that SNPs within the inversion breakpoints that show high LD with a large number of distant SNPs will tend to be diagnostic for inversion type, and identify the likely source of discrepancies between previous karyotype assignments. After selecting SNPs with nearly complete genotypic data that are also in high LD with many other SNPs, we performed principal components analyses on those SNPs located within the breakpoints of each of the three common inversions.

The full list of previous karyotype inferences, scores on the first two PC eigenvectors for inversion-diagnostic genotypes for each chromosome, and heterozygosities of all SNPs within

inversion breakpoints are given in Table S1. We plot the PC1 scores vs. the average heterozygosity (H) of each inversion region in Figure 3. In all but one of the 458 cases where Huang et al. and CD-L both reported a homozygous karyotype, scores on PC1 predict inversion type. The exception is line RAL332 for chromosome 3R, where both Huang et al. and CD-L infer the Standard arrangement, while PC1 score and the observed heterozygosity predict a Standard/In(3R)Mo heterozygote. Intermediate scores on PC1 are found in chromosome arms identified as inversion heterozygotes by Huang et al, with two exceptions. In addition to the exception mentioned above, PC1 score for line RAL325, chromosome 2R indicates a Standard/In(2R)NS heterozygote, while Huang et al. reported two different inversions as heterozygotes, In(2R)Y6 and In(2R)Y7.

A number of inversion regions have highly heterozygous sequence data (H>0.15) but no evidence of similarity to common inversion genotypes. In four cases (shown in green in Fig. 3), these were identified as heterozygotes for rare inversion karyotypes by Huang et al. Line RAL303 was scored as an inversion heterozygote for both In(2L)t and In(2R)NS by Huang et al., but does not have a genotype characteristic of either heterozygote. Fourteen lines for which neither PC1 scores nor Huang et al. suggest inversion heterozygosity are highly heterozygous in the region of In(3R)Mo, which may suggest the presence of balanced polymorphism not associated with in inversion.

There are a total of six cases where CD-L and Huang et al. assigned different homozygous karyotypes to the some lines; in three cases PC1 scores are consistent with CD-L, while PC1 scores and Huang et al. are in agreement for the other three. Huang et al. reported an additional five cases of karyotypic heterozygosity that do not have elevated sequence heterozygosity.

Heterozygosities of segments of chromosome arms defined by common inversion breakpoints are correlated, as shown in Table 1. Segments of the same arm always have correlations above 0.87. Segments of different arms of the same chromosome also remain highly correlated (average r=0.51). These results suggest that there is strong selection against recombinants and segregants in some of the DGRP lines. It is particularly striking that X-chromosome heterozygosity is significantly correlated with the heterozygosity of all other chromosome segments, particularly with chromosome 3, where the average r=0.37.

DISCUSSION

Our calculations have identified all the variant (SNP) pairs in 205 lines in the Freeze 2 data set that have linkage disequilibrium (LD) above r^2>0.5. This reveals both the overall patterns of LD, and will facilitate analyses that attempt to disentangle which nucleotides cause phenotypic effects.

Most SNPs in the Drosophila Genome Reference Project are in strong linkage disequilibrium (LD) with at least one other variant, and some with many other variants. More surprisingly, many SNPs in the full DGRP with minor allele frequency less than 0.2 are highly correlated with at least one SNP more than 100kbp distant. Thus, while it is true that the DGRP population has low LD relative to other eukaryotes, disequilibrium is still a major element of these data, and careful consideration should be given to its impact at all stages of an association analysis.

Linkage disequilibrium in the DGRP seems to reflect three primary causes (Huang et al. 2014). First, population-wide LD persists among closely linked SNPs because the recombination events that break down LD are insufficiently rare to counteract the weak processes of mutation, drift and potential natural selection. The signature of these events in the DGRP is that local LD remains appreciable throughout the range of SNP frequencies. This suggests that local LD would also be found in larger samples of genotypes. Second, variants with low minor allele frequencies (MAF) are on average highly correlated with multiple SNPs throughout the genome. The likely cause of this is random sampling of the very large number of low MAF variants in a relatively small sample of lines. We call this rarity disequilibrium. GWAS analyses often presume that only local LD needs to be considered, but this is not true for variants with MAF less than approximately 0.2 for the DGRP. Third, the presence of inversions allows differentiation of genotypes carried by each inversion, in turn creating additional LD. The rarity of alternative karyotypes ensures that this source of LD merely intensifies the degree of LD already present due to rarity disequilibrium in the DGRP.

Somewhat more speculatively, the pattern of correlations in heterozygosity among chromosomal regions we observe (Table 1) suggests that some of the long-distance or inter-chromosomal disequilibrium that we have detected may reflect epistatic selection. Corbett-Detig, et al. (2013) observed that some genotypic combinations in regions distant from each other are observed less frequently than expected in recombinant inbred lines in *D. melanogaster*, as well as other species. The correlations of heterozygosities we detected are the converse of this pattern, but are consistent with selection creating long-distance disequilibrium during the process of inbreeding.

When the karyotypic assignments from three previous studies are in agreement (Huang et al. 2014; Corbett-Detig and Hartl 2012; Langley et al. 2012), our genotype-based assignments of inversion type are concordant, except for one chromosome arm. That arm (3R in line RAL332, Fig. 3) is scored as an inversion heterozygote on the basis of our analysis, and homozygous Standard by Huang et al. and Corbett-Detig and Hartl. This could be due to the loss of the In(3R)Mo after sequencing. The remaining discrepancies between our results and the other scorings cannot be explained on this basis, and are likely either errors in these previous assignments, or perhaps recombination events that have separated karyotype and genotype. Several arms have combinations of PC scores and heterozygosity suggestive of recent double recombination events (e.g. 2R from RAL409 and 377). Regardless of whether the discrepancies between LD-based genotyping and karyotypic assignments are caused by errors or recombination, our genotypic typing is in most cases the most relevant for those performing association studies, as it summarizes similarity of genotype, and therefore phenotypic effects, across the large span of the genome covered by each inversion. Lines with highly heterozygous regions are a possible exception, as loss of heterozygosity between sequencing and phenotyping cannot be ruled out without additional analysis (as suggested for 3R in line RAL332).

Our LD calculations will not apply precisely to most association studies based on the DGRP, as each study is likely to use a different subset of lines for phenotyping. We have also calculated the correlations for the 184 lines that we have data for in our own association study (Márquez et al. unpublished). These results show that differences in the genotypes chosen can substantially change the inferred LD structure for rarer SNPs. Nevertheless the correlations that we have calculated will be useful as the basis for analyses of multi-SNP associations. For example, after identifying a set of SNPs with significant associations, one could reanalyze those

in multi-SNP analyses that include the most highly correlated SNPs to diagnose which SNPs are most likely to represent the variants that cause phenotypic differences, and which have their signal confounded with those from other SNPs. If such SNPs are all nearby, the inference of that genomic region as causal can be strong, even if the precise nucleotide responsible remains unknown. Follow-up studies of such regions are likely to be worthwhile. In contrast, SNPs for which the addition of distant SNPs renders effects ambiguous would be poor candidates for follow-up studies.

ACKNOWLEDGEMENTS

This work was supported by NIH 1R01GM094424-01. Computing support was provided by the Research Computing Center at Florida State University.

LITERATURE CITED

Corbett-Detig, R.B., C. Cardeno, and C.H. Langley, 2012 Sequence-Based Detection and Breakpoint Assembly of Polymorphic Inversions. Genetics 192: 131-137.

Corbett-Detig, R.B., and D.L. Hartl, 2012 Population Genomics of Inversion Polymorphisms in Drosophila melanogaster. PLoS Genetics 8: e1003056.

Corbett-Detig, R.B., J. Zhou, A.G. Clark, D.L. Hartl, and J.F. Ayroles, 2013 Genetic incompatibilities are widespread within species. Nature 504: 135-137.

Hill, W.G., and A. Robertson, 1966 The effect of linkage on limits to artificial selection. Genetical Research 8: 269-294.

Huang, W., A. Massouras, Y. Inoue, J. Peiffer, M. Rámia *et al.*, 2014 Natural variation in genome architecture among 205 Drosophila melanogaster Genetic Reference Panel lines. Genome Research 24: 1193-1208.

Langley, C.H., K. Stevens, C. Cardeno, Y.C.G. Lee, D.R. Schrider *et al.*, 2012 Genomic Variation in Natural Populations of Drosophila melanogaster. Genetics 192: 533-598.

Mackay, T.F.C., S. Richards, E.A. Stone, A. Barbadilla, J.F. Ayroles *et al.*, 2012 The *Drosophila melanogaster* Genetic Reference Panel. Nature 482: 173-178.

SAS Institute, Inc., 2011 The SAS System for Windows, Release 9.3. SAS Institute, Cary, NC.

Table 1. Pearson correlations of heterozygosities for regions of chromosome arms.

chromosome region	In(2L)t	prox. 2L	distal 2R	In(2R)NS	prox. 2R	3L	prox. 3R	In(3R)Mo	distal 3R	X	H ± S.D.
distal 2L	0.93**	0.91**	0.48**	0.44**	0.69**	0.11	0.05	0.08	-0.03	0.26**	0.046 ± 0.060
In(2L)t		0.98**	0.41**	0.39**	0.60**	0.06	0.00	0.02	-0.07	0.20*	0.046 ± 0.072
proximal 2L			0.46**	0.41**	0.61**	0.05	-0.01	0.02	-0.08	0.20*	0.045 ± 0.070
distal 2R				0.91**	0.88**	0.09	-0.02	0.00	-0.04	0.22*	0.036 ± 0.052
In(2R)NS					0.87**	0.06	-0.01	0.03	-0.02	0.20*	0.035 ± 0.056
proximal 2R						0.11	0.05	0.08	-0.03	0.29**	0.041 ± 0.049
3L							0.59**	0.55**	0.48**	0.44**	0.031 ± 0.035
proximal 3R								0.94**	0.91**	0.35**	0.047 ± 0.065
In(3R)Mo									0.92**	0.38**	0.047 ± 0.072
distal 3R										0.31**	0.044 ± 0.067
X											0.035 ± 0.016

* 0.0001<P<0.05

** P<0.0001

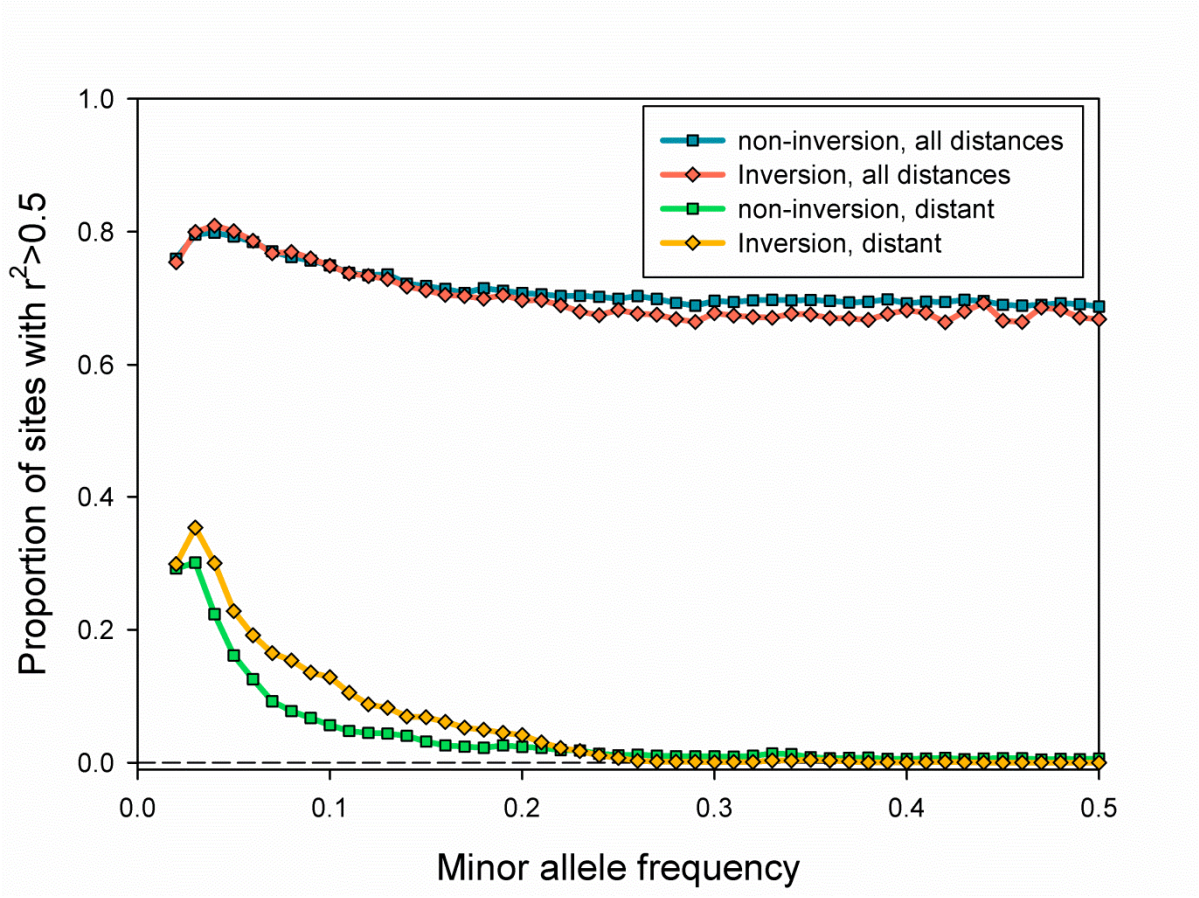

Figure 1. Probability that a focal site is correlated at $r^2>0.5$ with at least one other site in the genome, as a function of location relative to inversions. We treated the distal segment of chromosome 3R as part of In(3R)Mo (Corbett-Detig and Hartl 2012).

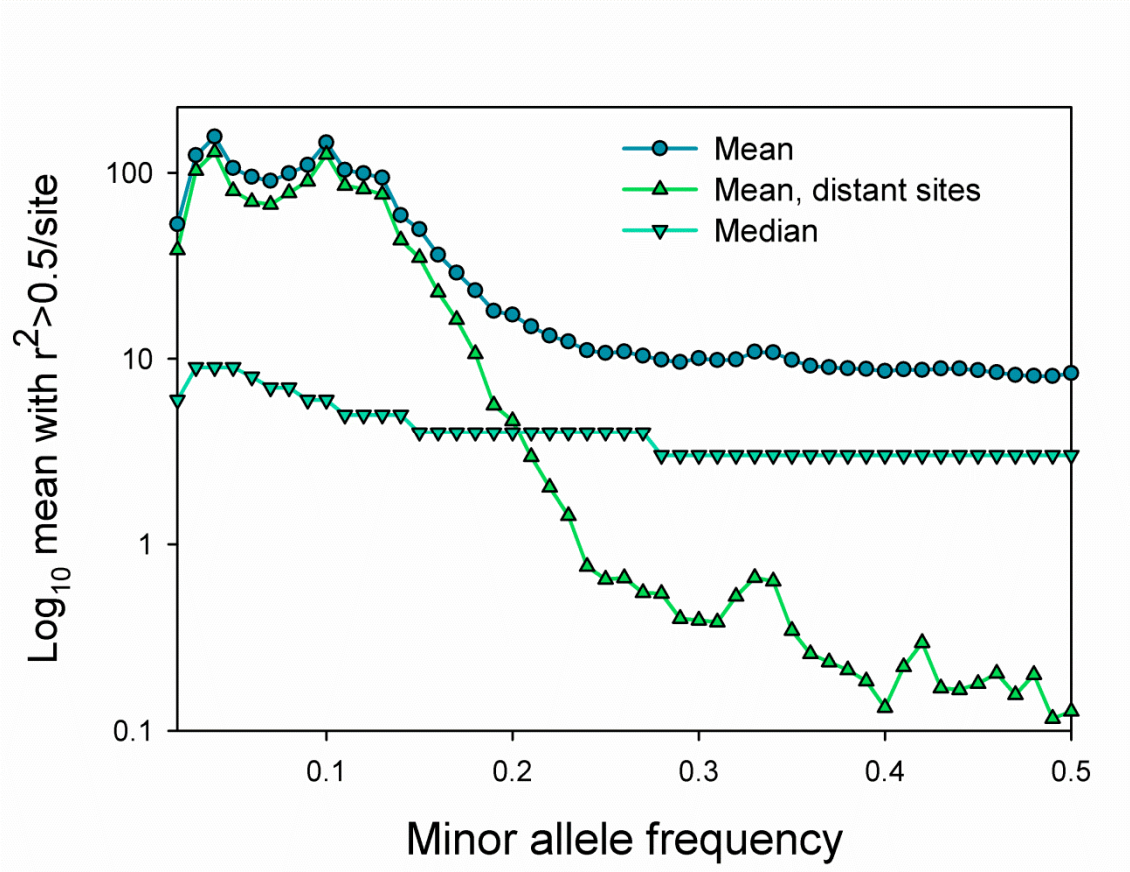

Figure 2. Mean and median number of sites correlated with variant sites at $r^2>0.5$ as a function of minor allele frequency.

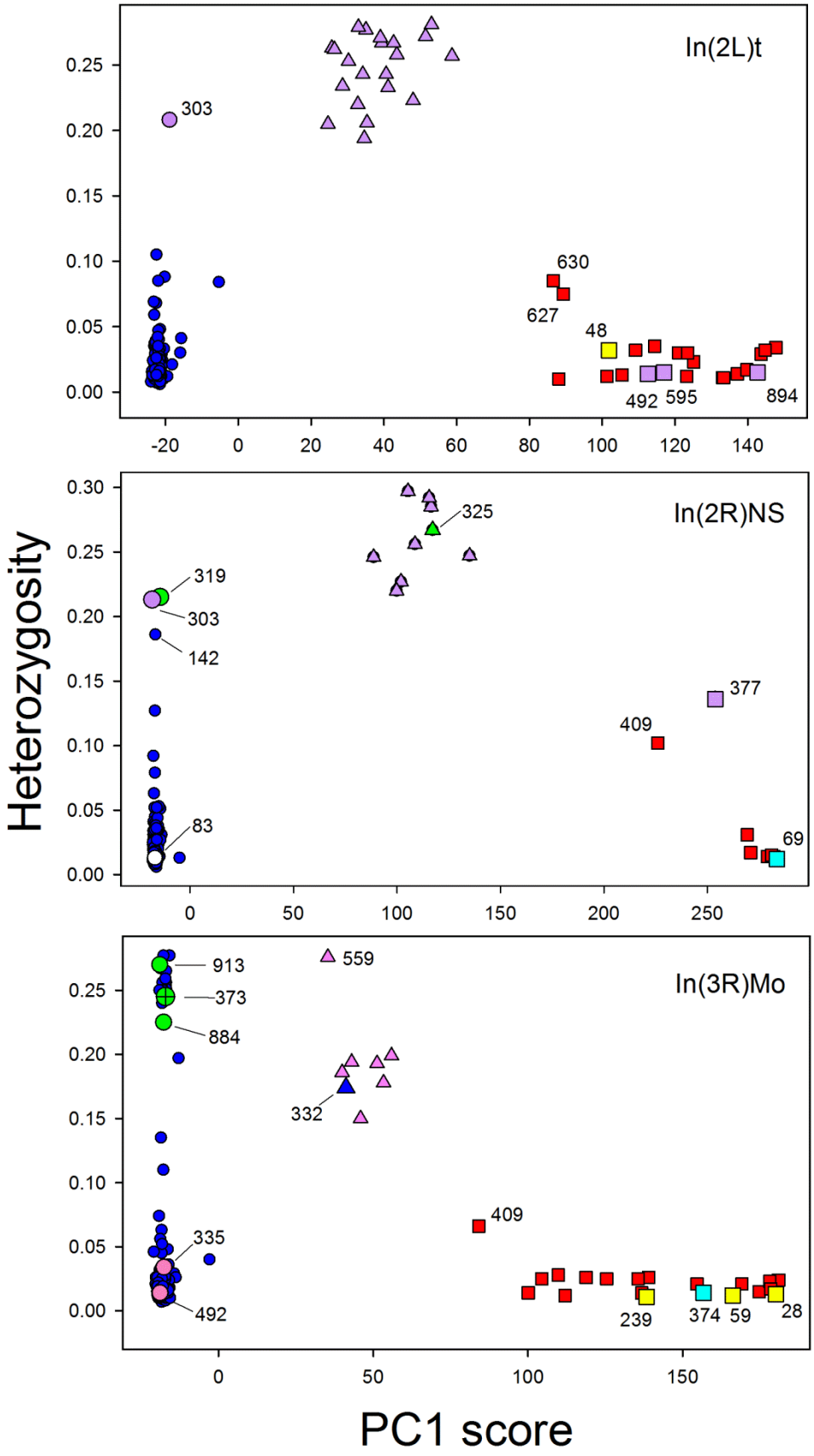

Figure 1. Inferred karyotype and heterozygosity of DGRP lines for common inversions. DGRP line numbers indicated for anomalous cases. Predictions based on our analyses are shown by symbol shapes: squares=inversion homozygotes, circles=Standard homozygotes, triangles=inversion heterozygotes. Colors reference the states inferred in Huang et al. and CD-L: Red=inversion homozygote in both; blue=standard homozygote in both; lavender=heterozygous in Huang et al. (CD-L in most cases did not scores heterozygotes); yellow=Standard homozygote in Huang et al., inversion homozygote in CD-L (exc. RAL48, not scored by CD-L); cyan=inversion homozygote in Huang et al., standard homozygote in CD-L; green=predicted heterozygote for a rare inversion (Line 373 for 3R was predicted as In(3R)Mo homozygote by CD-L); white=Standard homozygote in Huang et al., In(2R)NS homozygote by CD-L. Note that separate PC analyses were carried out for each inverted region, so the scale of PC1 scores for each chromosome arm is different.