SUPPLEMENTARY MATERIAL

SNP discovery and Genotyping

A detailed description of SNP discovery and genotyping is provided elsewhere (Stapley et al. 2008). In brief, the SNPs were identified using the QualitySNP software pipeline (Tang et al. 2006) from normalised cDNA sequences deposited in Genbank. We used the Illumina (San Diego) Golden Gate platform to genotype 354 individuals at 876 SNPs (Stapley et al. 2008). SNP physical positions were obtained using BLAST (v2.7.1) (Altschul et al. 1997) to compare sequence containing the SNP (50-121 bp) against the zebra finch genome sequence (http://genome.wustl.edu/pub/organism/Other_Vertebrates/Taeniopygia_guttata/assembly/Taeniopygia_guttata-3.2.4/). Stand-alone BLASTn was used with default parameter settings, except the expectation value (-e) was set to 1e-10 and the word size length (-W) was set to 25. In the few cases where SNP sequences had multiple hits, the best hit (lowest expectation value) was chosen provided the predicted location was consistent with the linkage map. SNPs that hit to unassembled contigs (denoted by "_random" or ChrUn) were not included in the analysis.

Haplotype Inference

There is general agreement that it is more accurate to employ a statistical procedure to infer haplotype phase when estimating LD from genotypic data (Weir 1979; Stephens et al. 2001; Slatkin 2008). There are two ways this can be done, with pedigree information or from population data (unrelated individuals). Although there are very good methods for estimating phase from population data, it is more accurate and efficient to use pedigrees (Stephens et al. 2001; Becker and Knapp 2002; Li and Jiang 2005; Slatkin 2008). In addition, haplotypes inferred from population data are least accurate when sample sizes are modest, as is the case in our study (Becker and Knapp 2002). For this reason we chose to use a pedigree based estimate.

Phase can be inferred with pedigree information using statistical and rule based methods (e.g. Minimum Recombinant Haplotype Configuration, MRHC). Statistical methods perform very well and we chose to use SimWalk2, a Maximum Likelihood (ML) method. SimWalk is a well respected and well-used ML based statistical program and performs as well as more recently developed programs based on MRHC (Li and Jiang 2005). The main disadvantage of statistical procedures is that they are time consuming to run because of the large number of possible haplotype configurations that need to be considered. One way to reduce the time required is to split the pedigree into smaller sub families. Splitting the pedigree also helps to deal with marriage loops, which are present in our pedigree. Splitting the pedigree and duplicating individuals to create unrelated families is a common procedure employed in several programs (e.g. LINKAGE, FASTLINK, PedPhase). To split the pedigree into separate unrelated families we used CRIGEN implemented in CriMap.

CRIGEN includes some individuals in more than one family, artificially inflating the size of the pedigree to 468 individuals and 153 founders compared to the true pedigree of 354 individuals and 60 founders. To ensure that this inflation did not bias the results, estimates of LD obtained from the phased haplotypes with 153 founders were compared to those obtained from the unphased genotypes using the founders of the original pedigree (n=60). The correlation coefficient for pair wise r^2 calculated from the two approaches was high (r = 0.95, Fig S1). LDmaps built using founder diplotype data are also in close agreement with the LDmaps constructed from phased haplotypes (Fig S2).

Modelling Linkage Disequilibrium

Calculation, representation and interpretation of LD is a complex topic, which has been reviewed elsewhere (Devlin and Risch 1995; Pritchard and Przeworski 2001; Ardlie et al. 2002; Zhang et al. 2002; Zhao et al. 2007; Slatkin 2008). We have adopted an approach to modelling LD that will facilitate comparison with previous studies and make useful comparison between chromosomes within the zebra finch genome. To model the decline in LD, pair wise estimates of LD such as r^2 and D′ are commonly used. In this study we avoid the use of D′ because this is sensitive to small sample sizes and r^2 is generally considered the best statistic for SNP data (Pritchard and Przeworski 2001; Ardlie et al. 2002; Weiss and Clark 2002). The r^2 statistic is the most useful in the context of mapping studies and it can be used to calculate the extent of useful LD to detect an association (Ardlie et al. 2002). The decline of r^2 was modelled using Sved's equation as described in the body of the manuscript.

Despite the usefulness of the r^2 statistic in the context of mapping, pair wise estimates of LD have some shortcomings. First, pair wise estimates between all markers are not independent, and as a result it is unclear how to combine these in a meaningful way and make inference (Pritchard and Przeworski 2001). Second, all pair wise metrics are, to varying degrees, confounded by either allele frequencies or the difference in allele frequencies between two markers (Hill and Robertson 1968) and/or differences in sample size (Slate and Pemberton 2007). This introduces potential problems when making comparisons between studies or between genomic regions. Therefore, in addition to presenting analysis of r^2, LD was modelled using population genetics theory (Morton et al. 2001), and the Malécot equation (Malécot 1948).

Estimation of Heterozygosity, GC content and Number of Genes

Total LDU, number of genes, GC content and mean heterozygosity was calculated per megabase (Mb). The number of genes, their start stop positions and the GC content were obtained from Ensembl BioMart (http://www.ensembl.org/biomart/martview/fd0d38a6a0dcc351ca2e08912f50fbc8) using database Ensembl 56, dataset *Taeniopygia guttata* genes (taeGut3.2.4). SNP heterozygosity (h_i) was calculated for autosomal markers using

$h_i = Nh_i/N_i$

where Nh_i is the number of founder individuals that were heterozygous at ith loci and N_i is the number of individuals typed at that loci.

CpG Motifs

Previous studies have identified that particular sequence motifs (CCTCCT, CTCTCCC, CCCCCCC, CTCF Consensus - CCNCCNGGNGG) are correlated with recombination rate (Shifman et al. 2006; Groenen et al. 2009). The position of each motif was estimated using EMBOSS (Rice et al. 2000) and the number of motifs per megabase was calculated. These measures are highly correlated with GC content (Fig S3) so for simplicity we only used GC content in the analysis.

References:

Altschul S, Madden T, Schaffer A, Zhang JH, Zhang Z, Miller W, Lipman D. 1997. Gapped BLAST and PSI-BLAST: A new generation of protein database search programs. *Nucl. Acids Res.* **25**: 3389-3402.

Ardlie KG, Kruglyak L, Seielstad M. 2002. Patterns of linkage disequilibrium in the human genome. *Nat. Rev. Genet.* **3**: 299-309.

Becker T, Knapp M. 2002. Efficiency of haplotype frequency estimation when nuclear familiy information is included. *Hum. Hered.* **54**: 45-53.

Devlin B, Risch N. 1995. A comparison of linkage disequilibrium measures for fine-scale mapping. *Genomics* **29**: 311-322.

Groenen MAM, Wahlberg P, Foglio M, Cheng HH, Megens H-J, Crooijmans RPMA, Besnier F, Lathrop M, Muir WM, Wong GK-S et al. 2009. A high-density SNP-based linkage map of the chicken genome reveals sequence features correlated with recombination rate. *Genome Res.* **19**: 510-519.

Hill WG, Robertson A. 1968. Linkage disequilibrium in finite populaitons. *Theoretical and Applied Genetics* **38**: 226-231.

Li J, Jiang T. 2005. Computing the Minimum Recombinant Haplotype Configuration from incomplete genotype data on a pedigree by integer linear programming. . *J. Comput. Biol.* **12**: 719-739.

Malécot G. 1948. Les Mathematiques de l'Heredite. *Maison et Cie, Paris*.

Morton NE, Zhang W, Taillon-Miller P, Ennis S, Kwok PY, Collins A. 2001. The optimal measure of allelic association. *Proc. Natl. Acad. Sci. USA* **98**: 5217 - 5221.

Pritchard JK, Przeworski M. 2001. Linkage disequilibrium in humans: Models and data. *Am. J. Hum. Genet.* **69**: 1-14.

Rice P, Longden I, Bleasby A. 2000. EMBOSS: The European Molecular Biology Open Software Suite. *Trends Genet.* **16**: 276-277.

Shifman S, Bell JT, Copley RR, Taylor MS, Williams RW, Mott R, Flint J. 2006. A high-resolution single nucleotide polymorphism genetic map of the mouse genome. *PLoS Biology* **4**: e395.

Slate J, Pemberton JM. 2007. Admixture and patterns of linkage disequilibrium in a free-living vertebrate population. *J. Evol. Biol.* **20**: 1415-1427.

Slatkin M. 2008. Linkage disequilibrium understanding the evolutionary past and mapping the medical future. *Nat. Rev. Genet.* **9**: 477-485.

Stapley J, Birkhead TR, Burke T, Slate J. 2008. A linkage map of the Zebra Finch *Taeniopygia guttata* provides new insights into avian genome evolution. *Genetics* **179**: 651-667.

Stephens M, Smith NJ, Donnelly P. 2001. A new statistical method for haplotype reconstruction from population data. *Am. J. Hum. Genet.* **68**: 978-989.

Tang JF, Vosman B, Voorrips RE, Van der Linden CG, Leunissen JAM. 2006. QualitySNP: a pipeline for detecting single nucleotide polymorphisms and insertions/deletions in EST data from diploid and polyploid species. *BMC Bioinformatics* **7**.

Weir BS. 1979. Inferences about Linkage Disequilibrium. *Biometrics* **35**: 235-254.

Weiss KM, Clark AG. 2002. Linkage disequilibrium and the mapping of complex human traits. *Trends Genet.* **18**: 19-24.

Zhang W, Collins A, Maniatis N, Tapper W, Morton NE. 2002. Properties of linkage disequilibrium (LD) maps. *Proc. Natl. Acad. Sci. USA* **99**: 17004-17007.

Zhao H, Nettleton D, Dekkers JCM. 2007. Evaluation of linkage disequilibrium measures between multi-allelic markers as predictors of linkage disequilibrium between single nucleotide polymorphisms. *Genet. Res.* **89**: 1-6.

Figure S1. Pair wise LD (r^2) estimated from phased haplotype data and unphased diplotype data (correlation coefficient = 0.95).

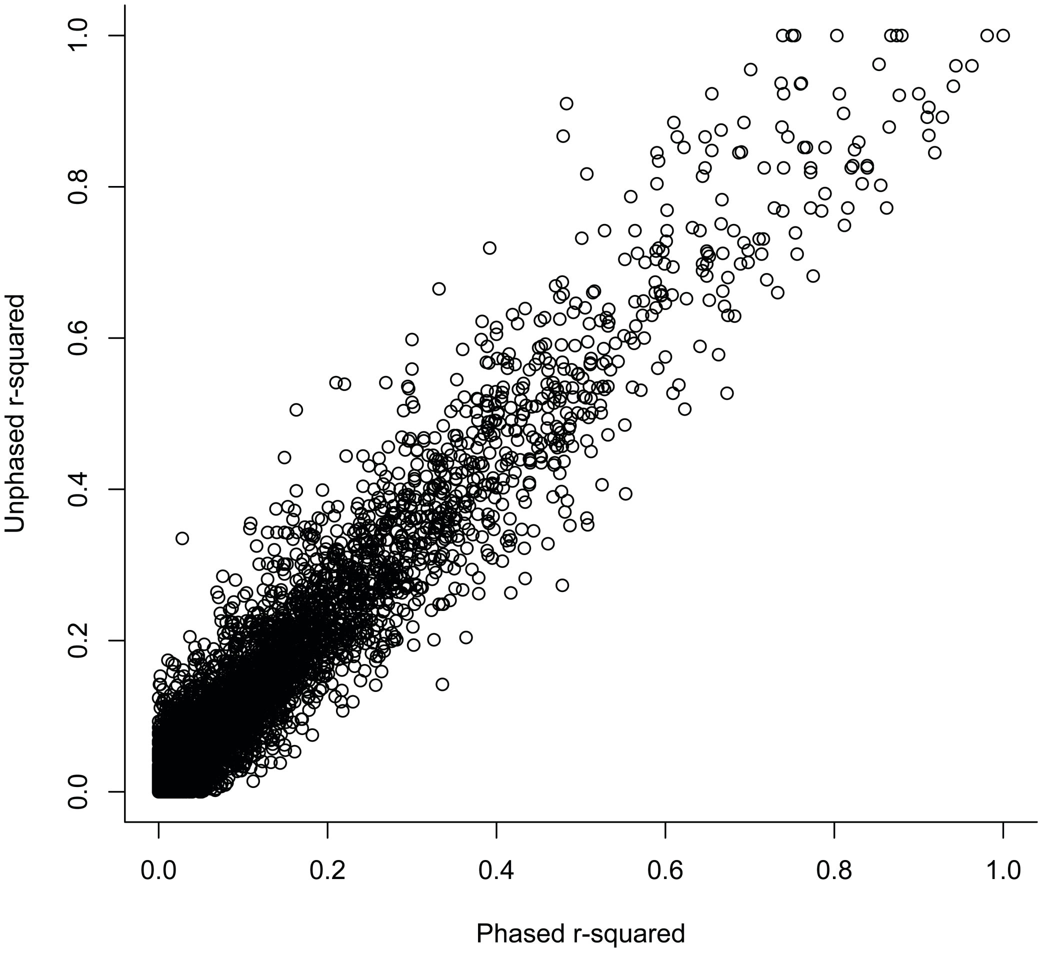


Figure S2.1. LDmaps for chromosomes constructed using phased haplotypes (black and unphased genotypes (red).

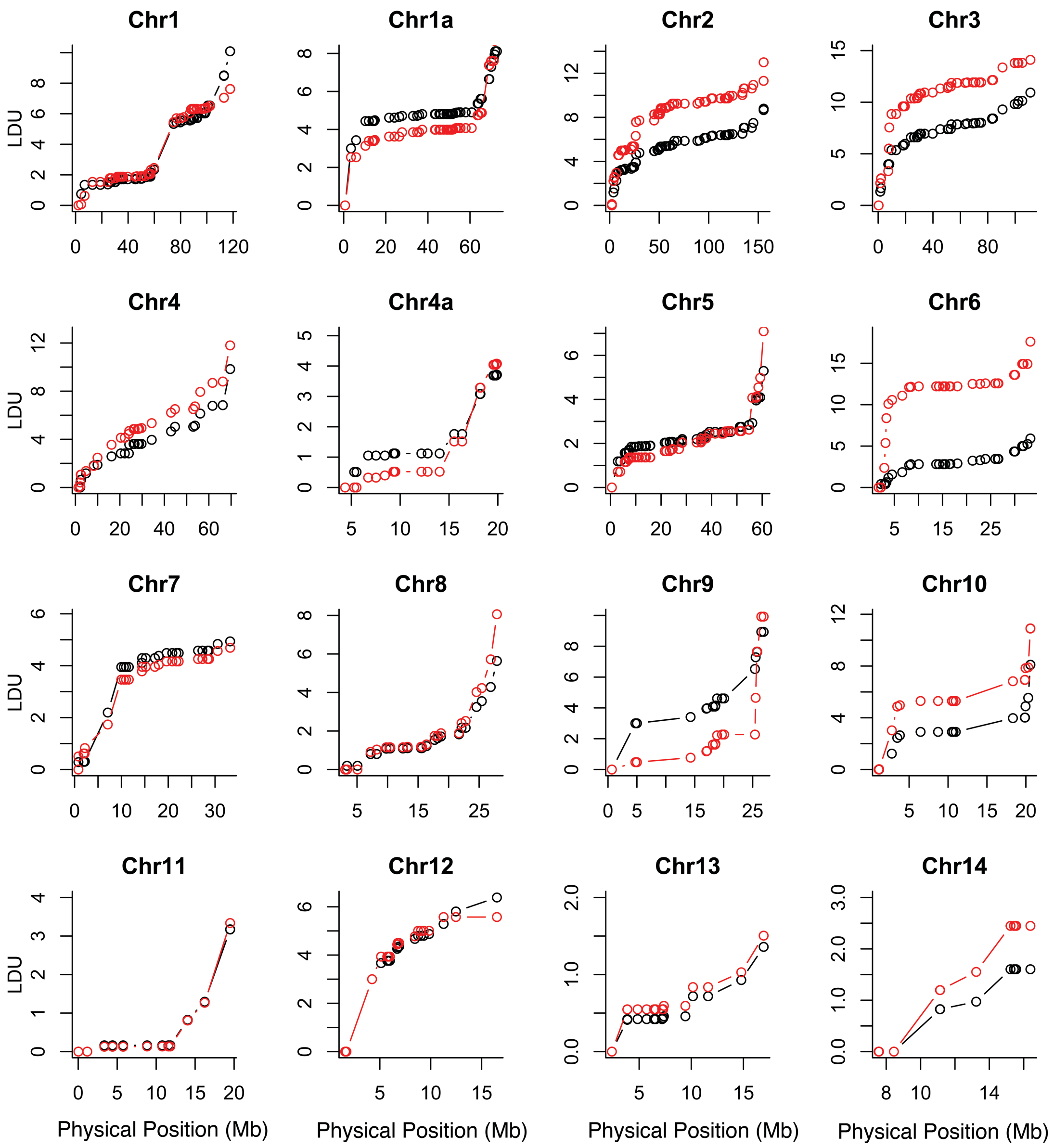


Figure S2.2. LDmaps for chromosomes constructed using phased haplotypes (black and unphased genotypes (red).

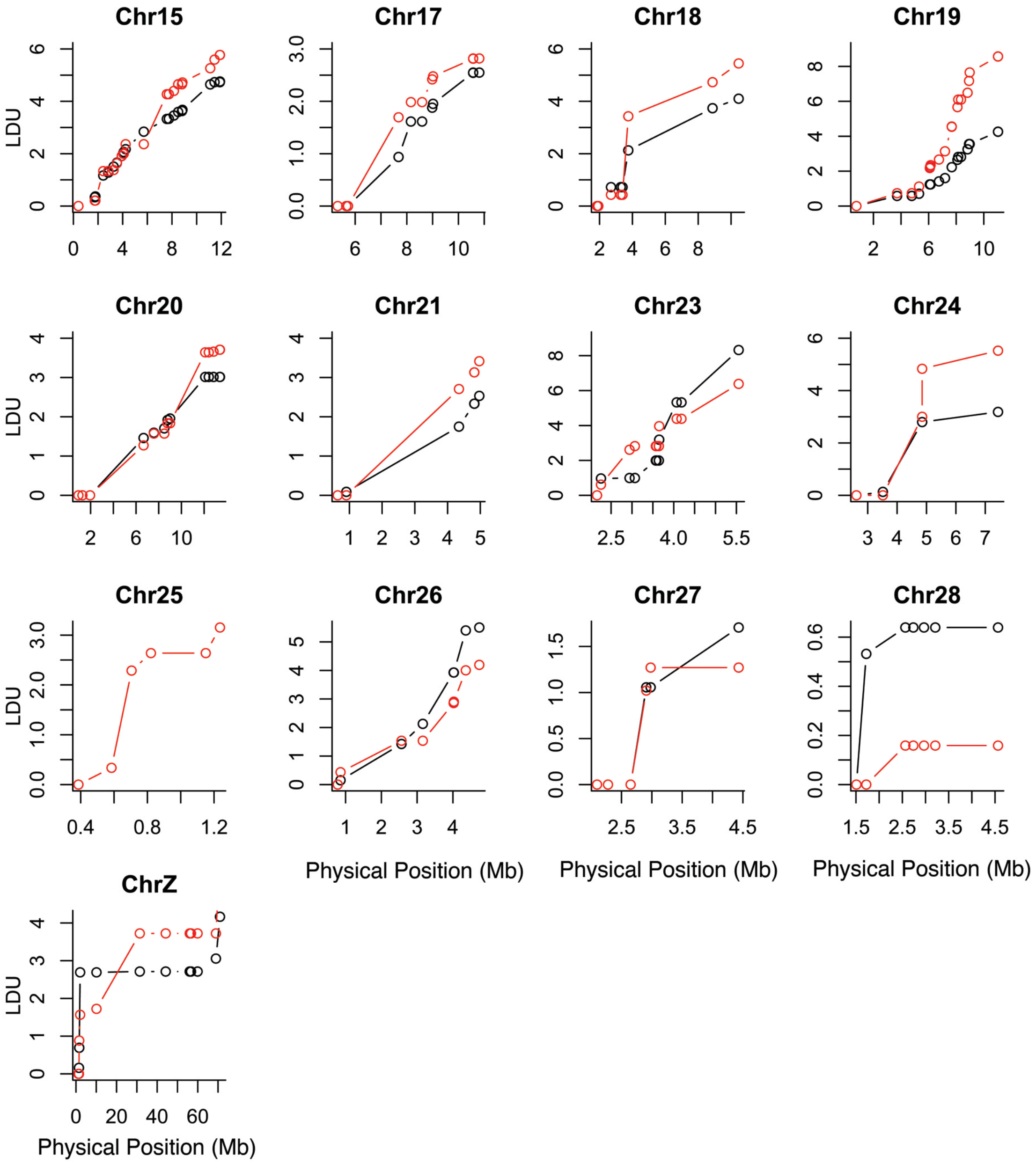


Figure S3. Correlation matrix of GC content and GC sequence motifs (CCTCCT , CTCTCCC, CCCCCCC, CTCF Con (CCNCCNGGNGG). Upper triangle of the matrix gives correlation coefficient and significance level (0 ***, <0.001 **, <0.05 *), on the diagonal is histograms of data and scatter plots on the lower triangle. All data are log transformed.

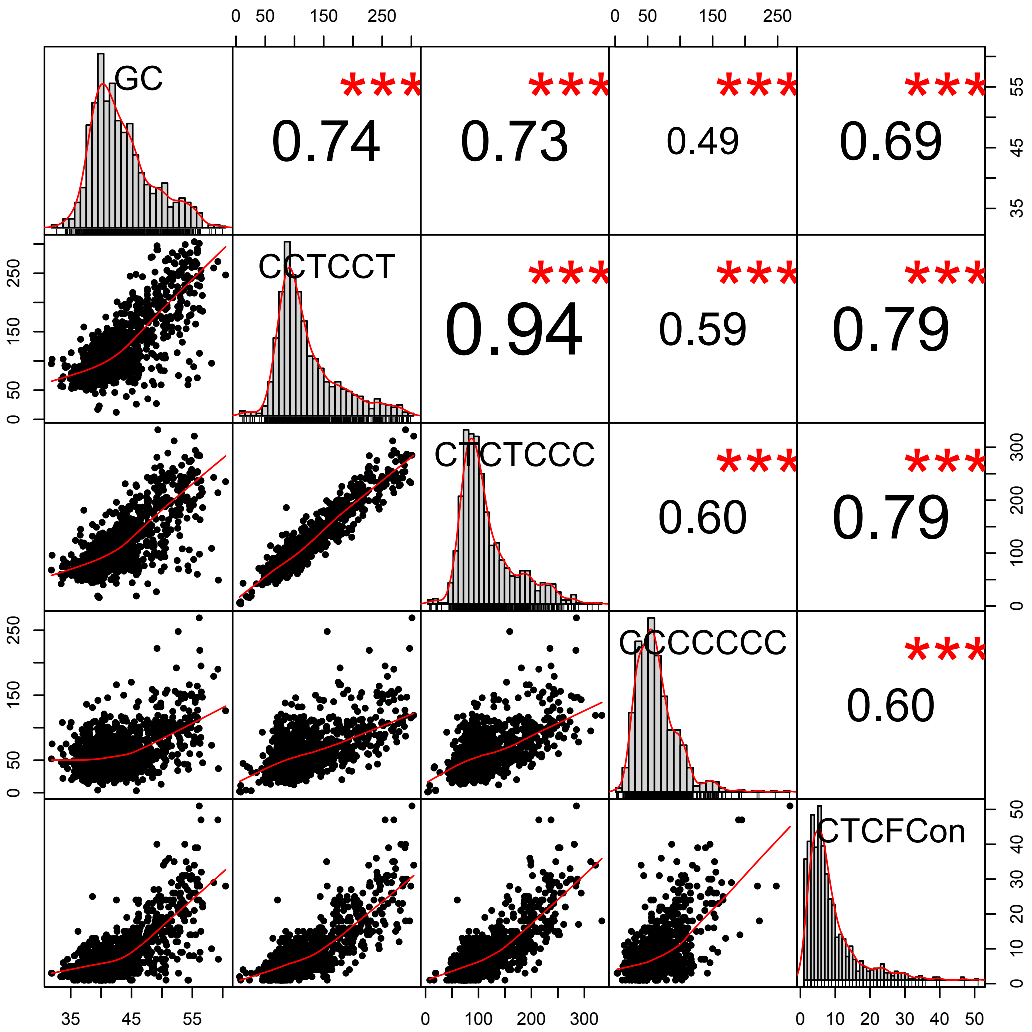


Figure S4. Linkage disequilibrium (r^2) between syntenic pairs of SNPs plotted against: a) physical distance (Mb), solid line represents mean r^2 for 1Mb bins, dashed line is the Sved's equation, for all the macrochromosome (left) and microchromosomes (right); b) genetic distance (cM), solid line represents mean r^2 for 1cM bins, dashed line is the Sved's equation, for all the macrochromosome (left) and microchromosomes (right).

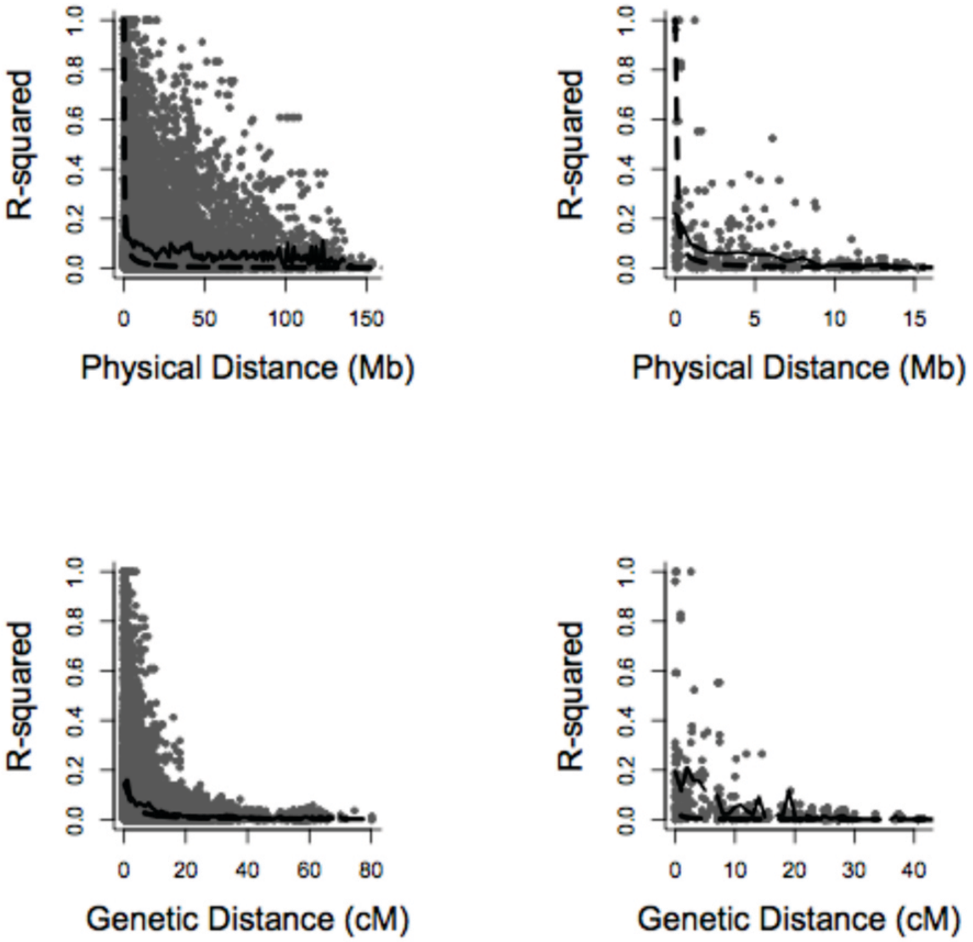


Figure S5.1. LD maps (LDU) and genetic maps (cM) plotted against physical distance along each chromosome. Solid circles and black line indicate LD map and open red squares and red line indicate genetic map.

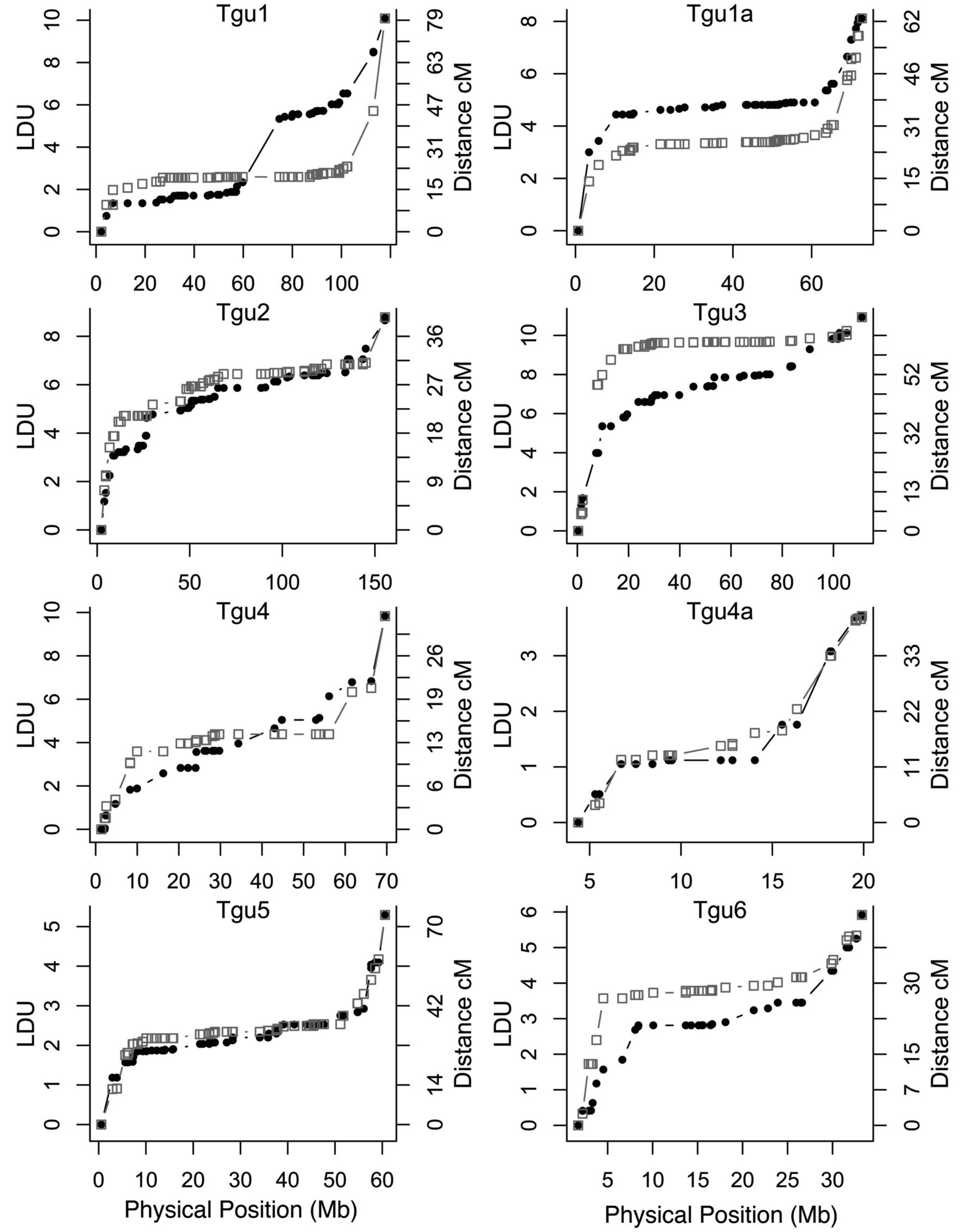


Figure S5.2. LD maps (LDU) and genetic maps (cM) plotted against physical distance along each chromosome. Solid circles and black line indicate LD map and open red squares and red line indicate genetic map.

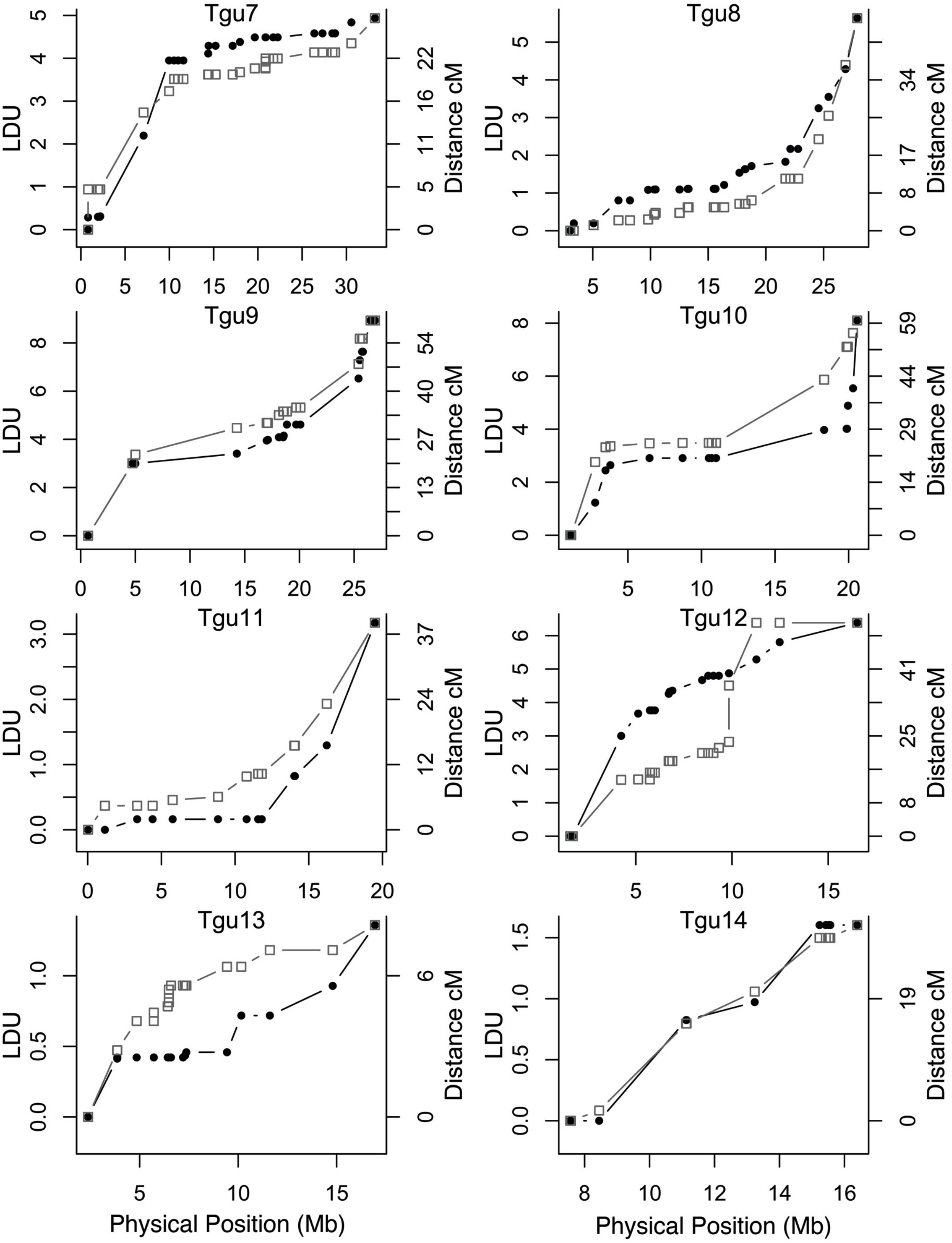


Figure S5.3. LD maps (LDU) and genetic maps (cM) plotted against physical distance along each chromosome. Solid circles and black line indicate LD map and open red squares and red line indicate genetic map.

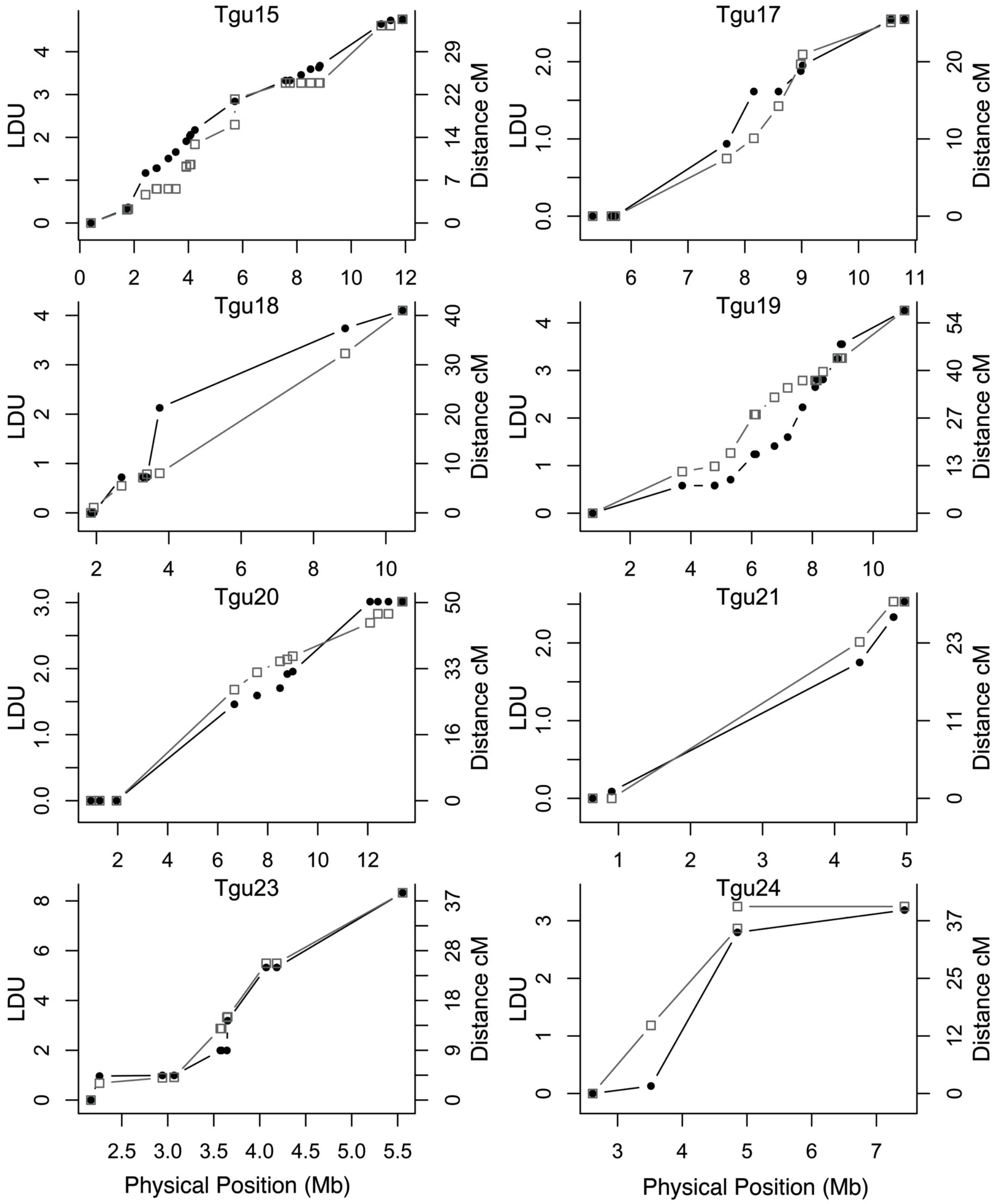


Figure S5.4. LD maps (LDU) and genetic maps (cM) plotted against physical distance along each chromosome. Solid circles and black line indicate LD map and open red squares and red line indicate genetic map.

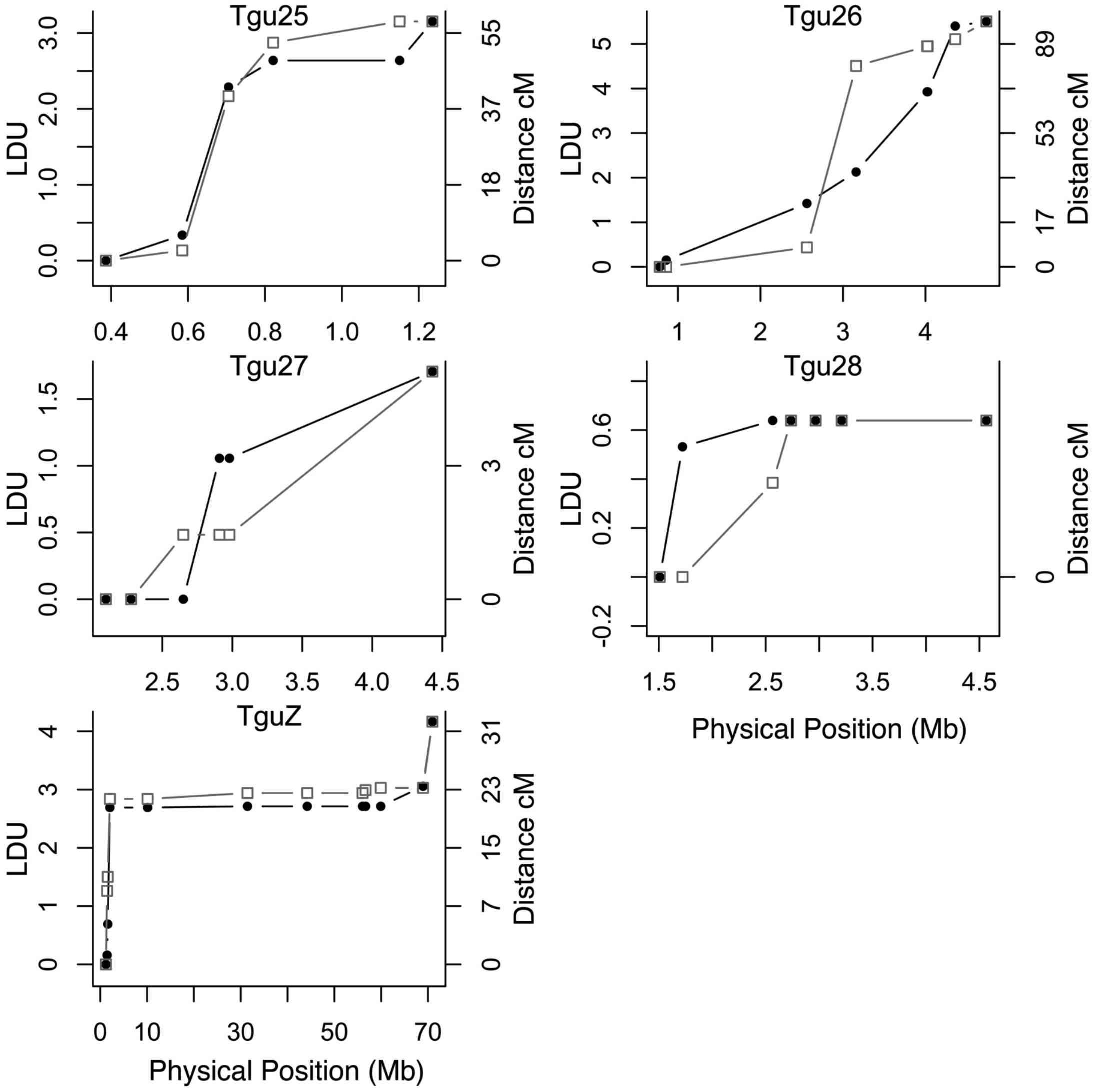


Figure S6. Relationship between sequence features per Megabase (Mb) (number of genes, GC content, heterozygosity) and log LDU per Mb. Correlation estimates based on Kendall's *tau*, *** denotes p-value<0.001). Red lines are smoothed splines.

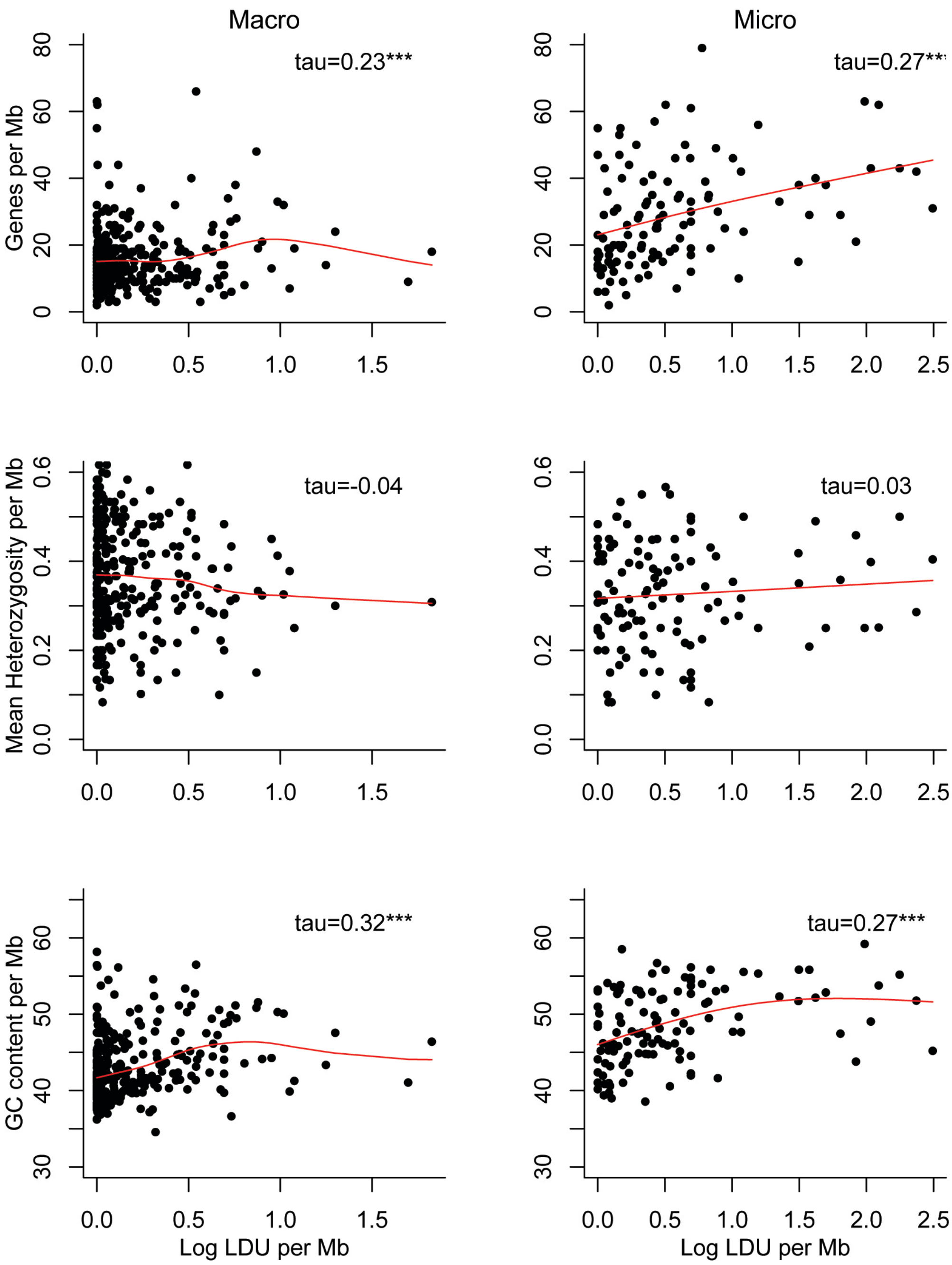


Figure S7.1. The total number of linkage disequilibrium units (LDU), GC content (GC), number of genes (Genes) and mean heterozygosity (Het) per megabase (Mb) along zebra finch chromosomes.

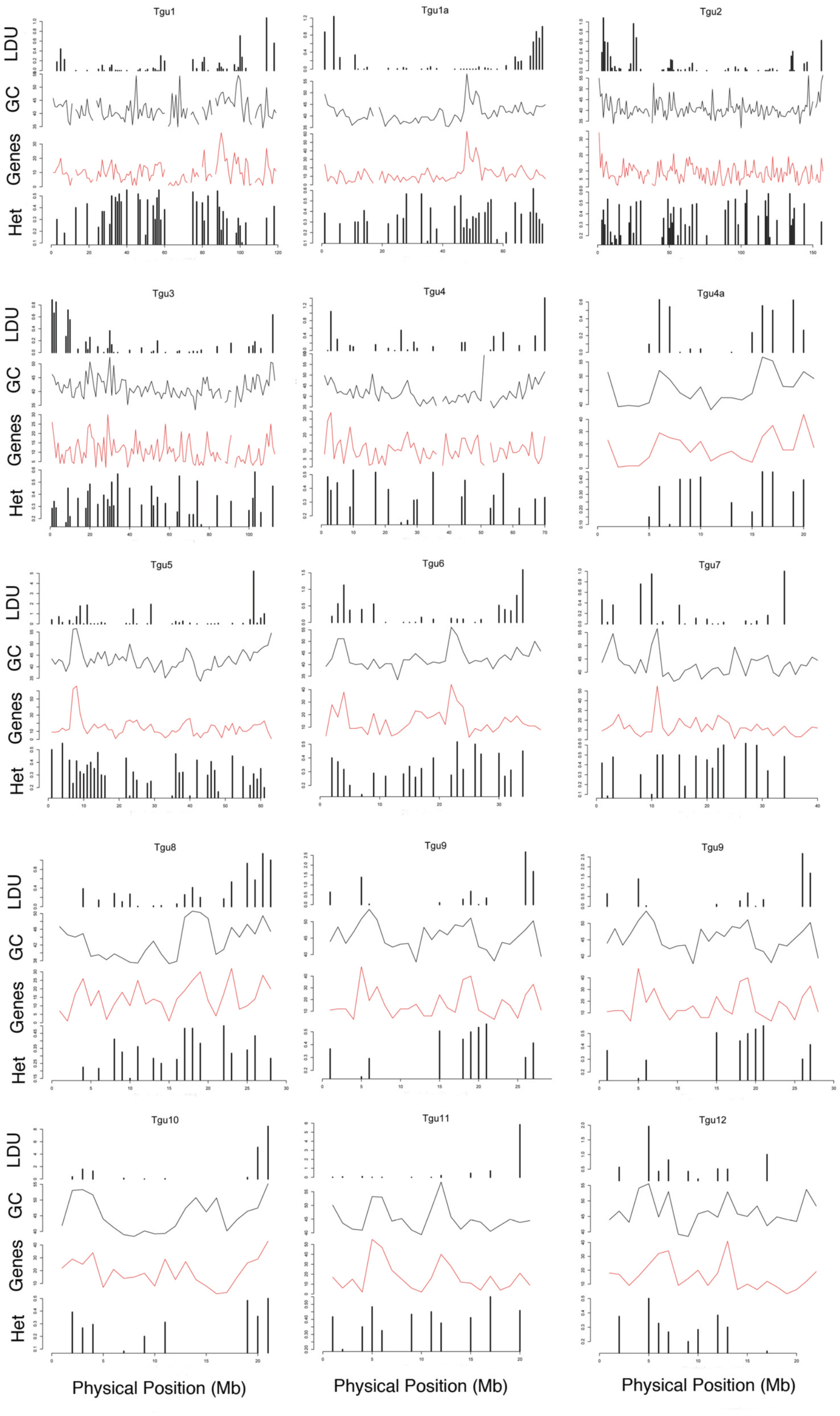
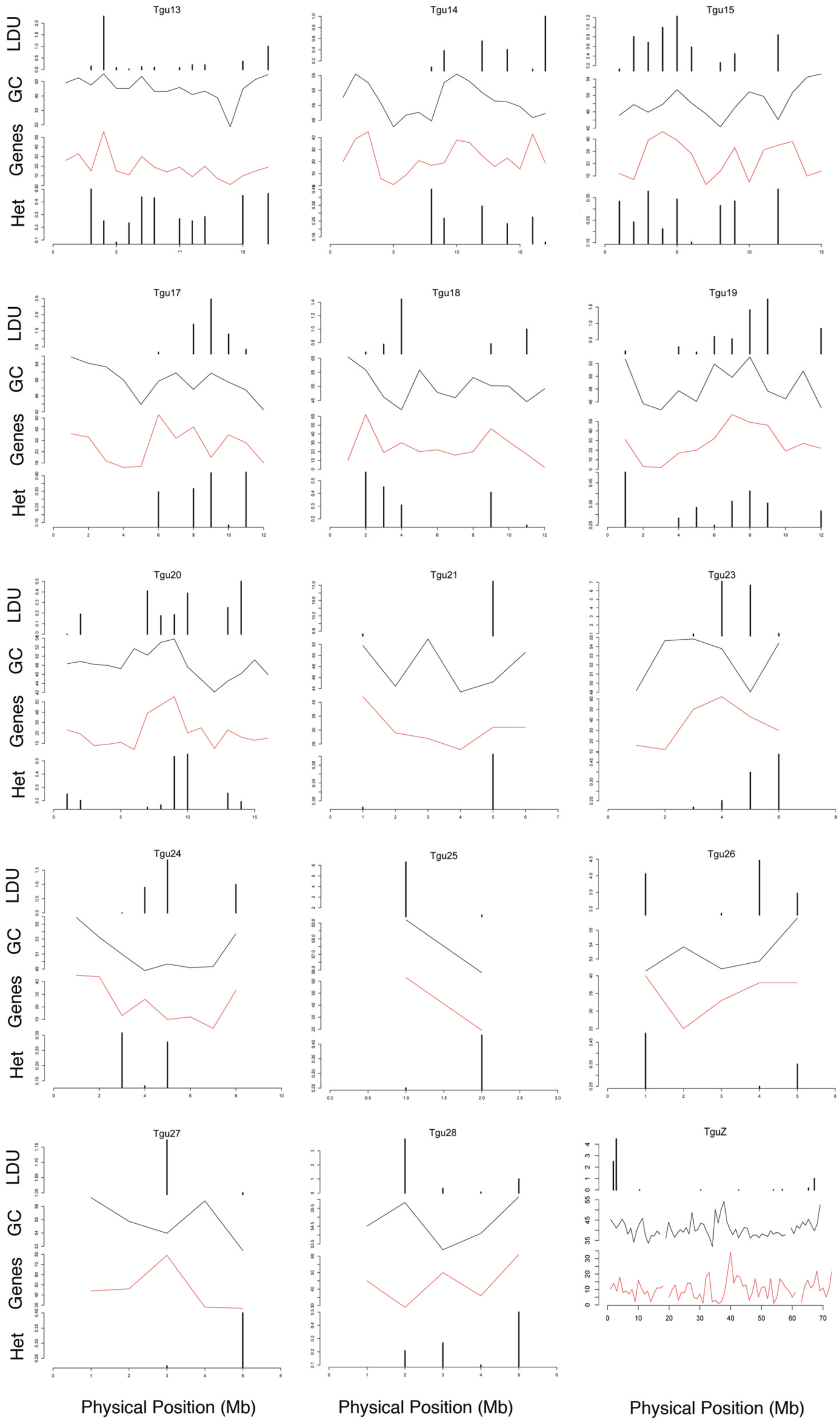


Figure S7.2. The total number of linkage disequilibrium units (LDU), GC content (GC), number of genes (Genes) and mean heterozygosity (Het) per megabase (Mb) along zebra finch chromosomes.


Tgu13
Tgu14
Tgu15
Tgu17
Tgu18
Tgu19
Tgu20
Tgu21
Tgu23
Tgu24
Tgu25
Tgu26
Tgu27
Tgu28
TguZ
LDU
GC
Genes
Het
Physical Position (Mb)
Physical Position (Mb)
Physical Position (Mb)