Supplementary material

ngsLD: evaluating linkage disequilibrium using genotype likelihoods

Emma A. Fox [1], Alison E. Wright [2], Matteo Fumagalli [1], Filipe G. Vieira [3]

[1] Department of Life Sciences, Silwood Park Campus, Imperial College London, Ascot, United Kingdom
[2] Department of Animal and Plant Sciences, University of Sheffield, United Kingdom
[3] Center for GeoGenetics, Natural History Museum of Denmark, University of Copenhagen, Denmark

1 Supplementary Methods

1.1 Likelihood functions

We implemented two different algorithms to estimate LD levels from short-read sequencing data, both based on genotype likelihoods. The first one is a maximum likelihood (ML) approach to estimate haplotype frequencies between pairs of sites using an expectation maximization (EM) algorithm (**?**). From such estimated haplotype frequencies, we then calculate D, D' and r^2 for each pair of sites. This popular method has been previously used, for instance, to estimate ancestral LD from admixed populations from genotype data (**?**). Here, we have adapted the latter method to a single unadmixed population and extended it to deal with genotype likelihoods rather than known genotypes.

Following the notation in the original study (**?**), let's denote $G = (G_1, G_2, ..., G_n)$ as the genotypes of n samples with $G_i = (G_i^1, G_i^2)$ being the genotypes of sample i for the pair of SNPs of interest. Likewise, we denote $D_i = (D_i^1, D_i^2)$ as the sequencing reads data for sample i at the pair of SNPs. Lastly, let's denote the frequency of haplotype j as p_j, the unobserved pair of haplotypes for a sample as $h = (h_1, h_2)$, and H as the set of all possible haplotypes.

The likelihood for haplotype frequencies $p = (p_j)$ given the sequencing data D is given by:

$$
\begin{aligned}
L(p) &= P(D|G) && (1)\\
&= \prod_{i=1}^{n} P(D_i|G_i)P(G_i|p) && (2)\\
&= \prod_{i=1}^{n} \sum_{h\in H} P(D_i|G_i)P(G_i, h_1, h_2|p) && (3)\\
&= \prod_{i=1}^{n} \sum_{h\in H} P(D_i|G_i)P(G_i|h_1, h_2)P(h_1, h_2|p) && (4)\\
&= \prod_{i=1}^{n} \sum_{h\in h(G_i)} P(D_i|G_i)p_{h_1}p_{h_2} && (5)\\
&= \prod_{i=1}^{n} \sum_{g_i\in G_i} \sum_{h\in h(g_i)} P(D_i|g_i)p_{h_1}p_{h_2} && (6)
\end{aligned}
$$

Note that $P(G_i|h) = 1$ when G_i is consistent with haplotypes h and $P(G_i|h) = 0$ otherwise. Thus, $h(G_i)$ is the set of all pairs of haplotypes that are consistent with the genotype of individual i. We take into account the data uncertainty by iterating over all possible unknown genotypes, weighting them by their likelihood. In fact, $P(D_i|g_i)$ is the genotype likelihood for sample i which can be easily calculated using any method implemented in ANGSD (**?**). We then calculated ML estimates of haplotype frequencies by maximizing the above likelihood using an EM algorithm previously proposed (**?**).

The second implementation is based on the squared Pearson correlation (r^2) between expected genotypes $E[G]$ which is calculated as

$$E[G] = \sum_{g \in \{0,1,2\}} g \cdot P(g|D) \tag{7}$$

where g is the genotype (as the number of minor alleles) and $P(g|D)$ is the genotype posterior probability using a HWE-based prior probability.

1.2 Simulations

Simulations were performed using `msprime v0.4.0` (**?**) to generate 100 chromosomes sampled from a putative population of European descent under a previously proposed demographic model (**?**). We simulated sequences comparable to the human chromosome 22 using the HapMap2 genetic map and a mutation rate of $1e-8$ per base per generation. The genotypes were then converted into genotype likelihoods with `msToGlf` (part of `ANGSD` package) (**?**) under seven hypothetical mean read depths per site per sample (0.5, 1, 2, 5, 10, 20, and 50) and assuming an error rate of 0.01. Since singleton and doubleton SNPs provide very little information, they were removed from the dataset when calculating LD. `ngsLD` was run on these datasets under two flavors: genotype likelihoods (GL), and calling genotypes (CG) by calling the one with the highest likelihood. The data set with called genotypes at a read depth of 50 was used as the reference (ground truth) set. To remove possible SNP discovery biases, we only used SNPs present in the reference dataset for the calculation of root mean square deviation (RMSD; Equation **??**) and mean standard bias (MSB; Equation **??**):

$$RMSD = \sqrt{\frac{\sum_n (\hat{x} - x)^2}{n}} \tag{8}$$

$$MSB = \frac{\sum_n \frac{\hat{x}-x}{x}}{n} \tag{9}$$

where, $\hat{x}$ and x stand for the estimated and true values, respectively, and n to the total number of data points.

The simulations were run on a AMD Opteron(tm) 6380 using 10 threads and took, on average, 34 minutes (Sup. Table **??**).

1.3 Real data

The sequencing data for gonad and spleen from 10 mallard duck and 11 wild turkey was downloaded from SRA (PRJNA271731) and analyzed separately through the PALEOMIX pipeline (**?**). Briefly, we used `Trimmomatic v0.36` (**?**) to remove adapters and trim low quality regions from reads, also excluding those with less than 36bp. Afterwards, each sample was mapped against the duck (**?**) or the turkey (**?**) reference genomes using `bwa-mem v0.7.15` (**?**), and mapped reads were filtered for PCR and optical duplicates

using `picard v2.6.0` (https://broadinstitute.github.io/picard). Finally, `GATK v3.6` (**?**) was used to perform a local realignment step around indels. For each species, `ANGSD` was used to calculate genotype likelihoods, excluding sites with extremely high coverage (greater than the 95th percentile of the empirical distribution) or with a minor allele frequency lower than 0.05 (to remove singletons and doubletons). Finally, we used `ngsLD` to calculate pairwise LD between all sites at less than $1000kbp$, and fitted a 3-parameters decay curve to the relationship between LD strength and distance between SNP pairs using the `fit_LDdecay.R` script (see Supplementary Code for details).

We also compared the performances of `ngsLD` and `GUS-LD` (**?**) on a real dataset. Since we do not know the truth, we sub-sampled the dataset to 10% and 50% of the original number of reads, and compared the results of both `ngsLD` and `GUS-LD` assuming the results from `GUS-LD` on the original (full) dataset as reference. Since `GUS-LD` is very memory intensive, we restricted our analysis to scaffold `NC_015011.2` of the turkey genome ($191Mbps$). In fact, `GUS-LD` as-is is not well suited for large-scale analyses, mainly due to its intensive memory requirements. Furthermore, it also lacks several important features, like (i) processing data from multiple chromosomes, (ii) limiting the comparison to SNPs within a certain distance, (iii) random sub-sampling of pairs of sites, (iv) filtering by minor allele frequency, (v) printing directly to gzipped standard output. As such, we changed the source code of `GUS-LD` to make it feasible to be run on this dataset.

1.4 Auxiliary scripts

1.4.1 LD prunning

Apart from `ngsLD`, we also provide some auxiliary scripts to perform some common LD-related analyses. One of the most common is probably the pruning of linked SNPs (to obtain a set of independent SNPs). The script `prune_graph.pl` represents the pattern of linked sites in the dataset as a network, with nodes as SNPs and edges as linkage between two SNPs (with LD level as a weight). The script finds the most connected (or linked) node and excludes all directly connected nodes; it proceeds iteratively until no other connected nodes are available. The script takes various options to reduce computation time and memory usage, such as maximum distance or minimum LD level between SNPs.

1.4.2 LD decay

Another common population genomic analysis is to infer LD strength and decay over physical (or genetic) distance. The script `fit_LDdecay.R` fits a decay curve to a plot of LD strength versus physical (or genetic) distance between the involved SNPs. The expected value of r^2 under a drift-recombination equilibrium is (**?**):

$$E[r^2] = \frac{1}{1+C} \tag{10}$$

where $C = 4N_e\rho$, N_e is the effective population size, and ρ the recombination fraction between sites. Since equation **??** is a theoretical expectation rarely followed in natural populations, we chose no to implement it in its native form but rather two other formulations of it. The first formulation derived this expectation by adjusting for sample size and assuming a low level of mutation (**?**):

$$E[r^2] = \left[\frac{10+C}{(2+C)(11+C)}\right] \cdot \left[1 + \frac{(3+C)(12+12C+C^2)}{n(2+C)(11+C)}\right] \tag{11}$$

where n is the sample size. The second formulation, is an extension of equation **??** to account for the range of observed r^2 values (note that this formulation requires the estimation of three parameters):

$$E[r^2] = \frac{r^2_{high} - r^2_{low}}{1+C} + r^2_{low} \tag{12}$$

where r^2_{high} and r^2_{low} stand for the maximum and minimum (respectively) observed r^2 values.

For D', we fit the expectation derived by **?**, assuming a recombination rate of $1cm = 1Mb$, fixing $D'_0 = 1$ (representing the initial value of D') and estimating the three other parameters t, D'_{high} and D'_{low}, representing the number of generations since $D' = D'_0$, and the maximum and minimum expected D' between markers, respectively:

$$E[D'] = D'_{low} + (D'_{high} - D'_{low}) * D'_0 * (1-\theta)^t \tag{13}$$

To fit the above mentioned equations, we used the *optim* package from R to minimize the residual variability measured as the sum of squares:

$$SS = \sum (LD - E[LD])^2 \tag{14}$$

where LD is the observed disequilibrium coefficient between two SNPs, and $E[LD]$ the expected linkage disequilibrium under the model. This script can also be used to bin data points into windows, perform bootstrap analyses, and plot 95% confidence intervals.

1.4.3 LD blocks

We also provide the `LD_blocks.sh` script to plot LD blocks for a specific region. Given a specific region and coordinates, it plots r^2 using the R package `LDheatmap` (**?**)

2 Supplementary Figures and Tables

	0.5x	1x	2x	5x	10x	20x	50x
GL	50	48	36	35	36	37	16
CG	42	41	36	30	24	23	16

Supplementary Table 1: Running times (in minutes) for the simulated dataset when using genotype likelihoods (GL) and called genotypes (CG).

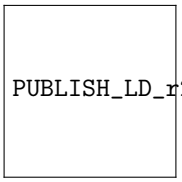
Coverage	Type	Distance	$r^2 - Pearson$	D	D'	r^2
1x	CG	$<= 10kb$	0.4802691	0.1030863	0.4331485	0.365254
		$> 10kb <= 100kb$	0.4217236	0.097468	0.4338424	0.3174724
		$> 100kb$	0.3959294	0.09383327	0.4416293	0.2971232
	GL	$<= 10kb$	0.4778938	0.1067788	0.3037001	0.215965
		$> 10kb <= 100kb$	0.4193353	0.1027622	0.3409216	0.1929541
		$> 100kb$	0.3935483	0.09991455	0.3771209	0.1830678
2x	CG	$<= 10kb$	0.3787519	0.07232258	0.412428	0.3298213
		$> 10kb <= 100kb$	0.3261336	0.06602788	0.4004184	0.2829039
		$> 100kb$	0.3023516	0.06249778	0.3916957	0.2619124
	GL	$<= 10kb$	0.3699717	0.06605041	0.2533213	0.1235468
		$> 10kb <= 100kb$	0.3175826	0.0610496	0.2799534	0.1084968
		$> 100kb$	0.2938332	0.05833835	0.3031421	0.09878242
5x	CG	$<= 10kb$	0.1622496	0.03200612	0.2651638	0.1684726
		$> 10kb <= 100kb$	0.1346846	0.02528141	0.2633643	0.1417432
		$> 100kb$	0.1213574	0.02196447	0.2588149	0.1286726
	GL	$<= 10kb$	0.1495284	0.02620193	0.1643936	0.05057308
		$> 10kb <= 100kb$	0.1231976	0.01794112	0.1804204	0.04482478
		$> 100kb$	0.1101826	0.01375448	0.19292	0.03949049
10x	CG	$<= 10kb$	0.05548174	0.0164721	0.1376595	0.05639049
		$> 10kb <= 100kb$	0.04409223	0.01052652	0.1396062	0.04517365
		$> 100kb$	0.0379471	0.007968536	0.1401131	0.03924509
	GL	$<= 10kb$	0.04397528	0.01560279	0.07938792	0.01963039
		$> 10kb <= 100kb$	0.03559623	0.009026443	0.08837916	0.01756353
		$> 100kb$	0.03090511	0.005908159	0.0962099	0.01535589
20x	CG	$<= 10kb$	0.01021865	0.01214253	0.02930025	0.01023027
		$> 10kb <= 100kb$	0.008050431	0.006628229	0.02996349	0.008032566
		$> 100kb$	0.00673809	0.003639248	0.03000047	0.006763703
	GL	$<= 10kb$	0.007953477	0.01210349	0.01927902	0.004274684
		$> 10kb <= 100kb$	0.006363818	0.006551845	0.0204608	0.003753527
		$> 100kb$	0.005448971	0.003507605	0.02220685	0.003317526
50x	CG	$<= 10kb$	0	0	0	0
		$> 10kb <= 100kb$	0	0	0	0
		$> 100kb$	0	0	0	0
	GL	$<= 10kb$	0.0001386686	2.17522e-05	0.0003077681	0.0001575763
		$> 10kb <= 100kb$	0.0001002384	2.1248e-05	0.0002597385	0.0001155164
		$> 100kb$	0.0001091353	2.29711e-05	0.0003527515	0.0001264611

Supplementary Table 2: Root Mean Square Deviation (RMSD) values for all four LD statistics, assuming genotypes called (CG) at 50× as the ground truth.

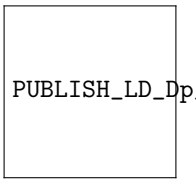
Supplementary Figure 1: Boxplots of the Root Mean Square Deviation (RMSD) for all four LD statistics (rows) at three different distance ranges (columns), and using both called genotypes (red) and genotype likelihoods (blue); we assumed genotypes called at 50× as the ground truth.

Coverage	Type	Distance	$r^2 - Pearson$	D	D'	r^2
1x	CG	$<= 10kb$	6.7883788	-0.5861226	0.3025183	10.9656254
		$> 10kb <= 100kb$	14.6509973	-0.5882485	0.8173968	30.0483104
		$> 100kb$	25.263428	-0.5770944	1.50188263	57.9692215
	GL	$<= 10kb$	6.7232779	-0.4425493	0.541295542	19.4748357
		$> 10kb <= 100kb$	14.560605	-0.4405534	1.201174	56.23428165
		$> 100kb$	25.232488	-0.4040141	2.094521	115.384347
2x	CG	$<= 10kb$	5.5455226	-0.4135590	0.06441694	3.0581046
		$> 10kb <= 100kb$	12.3154713	-0.4159735	0.3506162	9.726297
		$> 100kb$	21.149223	-0.4119061	0.7421672	19.4726552
	GL	$<= 10kb$	5.521689	-0.20016074	0.3653961131	8.13551182
		$> 10kb <= 100kb$	12.1869499	-0.18677621	0.7754511335	23.2872497
		$> 100kb$	20.85082	-0.16392394	1.3317896247	46.066568
5x	CG	$<= 10kb$	3.035657	-0.1775426	0.06367574	1.4430677
		$> 10kb <= 100kb$	5.780543	-0.1798473	0.1893561	3.804511
		$> 100kb$	9.2135537	-0.1853287	0.3627173	6.7409611
	GL	$<= 10kb$	2.8938465	-0.06297014	1.943278e-01	1.899796
		$> 10kb <= 100kb$	5.4874868	-0.05491939	3.455406e-01	5.119789
		$> 100kb$	8.641384	-0.050130268	5.549339e-01	9.206552
10x	CG	$<= 10kb$	0.90019823	-0.04025752	0.05486018	0.6515784
		$> 10kb <= 100kb$	1.551388	-0.04003615	0.1027084	1.284561
		$> 100kb$	2.471098	-0.03901899	0.1690646	2.181826
	GL	$<= 10kb$	0.73146353	-1.774423e-02	0.06607049	0.4186113
		$> 10kb <= 100kb$	1.271201	-1.545134e-02	0.1137282	0.9589022
		$> 100kb$	1.978590	-1.279278e-02	0.1755965	1.759859
20x	CG	$<= 10kb$	0.03453458	-0.006300113	0.003706502	0.0308322
		$> 10kb <= 100kb$	0.07100035	-0.005701256	0.005896093	0.06044355
		$> 100kb$	0.1137127	-0.005143937	0.0102679	0.1028088
	GL	$<= 10kb$	3.332211e-02	-4.836275e-03	0.003660898	1.902307e-02
		$> 10kb <= 100kb$	0.06258531	-4.300109e-03	0.007512918	5.249849e-02
		$> 100kb$	0.09664223	-3.777175e-03	0.01038638	7.914358e-02
50x	CG	$<= 10kb$	0	0	0	0
		$> 10kb <= 100kb$	0	0	0	0
		$> 100kb$	0	0	0	0
	GL	$<= 10kb$	0.0000121828	-1.602524e-05	0.0000387346	4.093477e-05
		$> 10kb <= 100kb$	9.899304e-06	-0.0000124297	2.976227e-05	0.0000311882
		$> 100kb$	2.586288e-05	-1.518837e-05	4.413672e-05	7.298129e-05

Supplementary Table 3: Mean Standard Bias (MSB) values for all four LD statistics, assuming genotypes called (CG) at 50× as the ground truth.

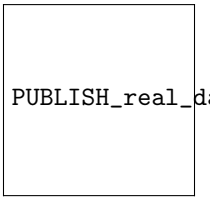


Supplementary Figure 2: Fitting of r^2 decay between called genotypes (CG) and genotype likelihoods (GL), across different simulated coverages (rows). The best-fitted curve (solid-line) and confidence interval (shaded area) was based on 250*bps* bins but, for sake of clarity, data is represented as bins of 500*bps* (points) and Y-axis us truncated at 0.5. Confidence interval was based on 100 bootstraps.

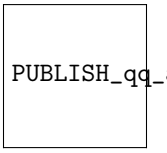


Supplementary Figure 3: Fitting of D' decay between called genotypes (CG) and genotype likelihoods (GL), across different simulated coverages (rows). The best-fitted curve (solid-line) and confidence interval (shaded area) was based on 10*bps* bins but, for sake of clarity, data is represented as bins of 100*bps* (points). Confidence intervals were based on 100 bootstraps.

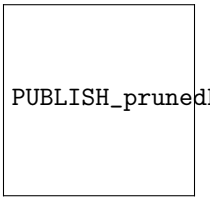
Supplementary Figure 4: Fitting of r^2 decay between called genotypes (CG) and genotype likelihoods (GL) from a sample size of 500 individuals, across different simulated coverages (rows). The best-fitted curve (solid-line) and confidence interval (shaded area) was based on 250*bps* bins but, for sake of clarity, data is represented as bins of 500*bps* (points). Confidence intervals were based on 100 bootstraps.



Supplementary Figure 9: Results of LD (r^2) decay fitting to two datasets of duck (dotted line) and turkey (solid line). The best fitted curves were based on 1000 bp bins (points).



Supplementary Figure 10: QQ plot of LD estimates on a subset of the real dataset, comparing GUS-LD (X-axis) against inferences from both 'ngsLD' (blue) and 'gusLD' (red) on datasets sub-sampled at 50% (triangles) and 10% (circles).



Supplementary Figure 5: Pair-wise r^2 after pruning, calculated from called genotypes (CG) simulated at 50× (columns). For sake of clarity, data is represented as bins of 1000*bps* (points).

Supplementary Figure 6: Pairwise LD levels inferred by *ngsLD*, across a 15*kb* region (28 SNPs) of the simulated chromosomes. Colors depict the LD (r^2) strength between SNPs, together with their physical location (diagonal line).

Supplementary Figure 7: Per-sample depth distribution for each of the duck samples.

Supplementary Figure 8: Per-sample depth distribution for each of the turkey samples.

3 Supplementary Code

```
#########################################################
# Pt. 1 - Simulated data                                #
#########################################################
# Set run variables
READ_NUM=[one of 1, 2, 5, 10, 20, 50]
SET_NAME=OUTofAFRICA$READ_NUM

# Run simulation with MSPRIME (msprime script available from authors upon request)
msprime.out_of_africa.py 0,100,0 1 chr22 $SET_NAME
N_HAP=`head -n 1 $SET_NAME.pos | cut -f 1`
N_SNPS=`head -n 1 $SET_NAME.pos | cut -f 2`
N_SITES=`head -n 1 $SET_NAME.pos | cut -f 3`
N_IND=$((N_HAP / 2))
MINMAF=$(echo "scale=3; 2/$N_HAP" | bc)

# Convert MS output to genotype likelihoods (10 GLs format), assuming a given sequencing depth
    and error rate
msToGlf -in $SET_NAME.ms -out $SET_NAME.ms -regLen $N_SITES -singleOut 1 -depth $READ_NUM -err
    0.01 -pileup 0 -Nsites 0 -printSNP 1
hexdump -s 4 -v -e '1/4 "%d" "\n"' $SET_NAME.ms.vPos | awk '{print "chrSim\t"$1}' >
    $SET_NAME.ms_pos

# Convert genotype likelihoods to BEAGLE format
angsd -isSim 1 -glf $SET_NAME.ms.glf.gz -fai reference.fa.fai -nInd $N_IND -doMajorMinor 1
    -doMaf 1 -doGlf 2 -minMaf $MINMAF -out $SET_NAME

# Create POSITION file
zcat $SET_NAME.mafs.gz | tail -n +2 | awk 'NR==FNR{x[FNR]=$0} NR!=FNR{print x[$2]"\t"$6}'
    $SET_NAME.ms_pos - > $SET_NAME.pos
NS=`cat $SET_NAME.pos | wc -l`

# Run NGSLD from GL
ngsLD --verbose 1 --n_threads $N_THREADS --n_ind $N_IND --n_sites $NS --geno
    $SET_NAME.beagle.gz --probs --pos $SET_NAME.pos --max_kb_dist 500 | gzip >
    ${SET_NAME}_GL.ld.gz

# Run NGSLD from Called Genotypes
$NGSLD --verbose 1 --n_threads $N_THREADS --n_ind $N_IND --n_sites $NS --geno
    $SET_NAME.beagle.gz --probs --pos $SET_NAME.pos --max_kb_dist 500 --call_geno | gzip >
    ${SET_NAME}_CG.ld.gz

### LD decay
# Subsample dataset
parallel "zcat {} | awk 'rand()<0.001' | gzip --best > {.}_sampled.gz" ::: *.ld.gz

# Plot r^2 LD decay
ls $NAME*.ld_sampled.gz | sort -V | awk 'BEGIN{OFS="\t"; print "File\tCoverage\tType"}
    {split($1,a,"_|.ld"); sub("OUTofAFRICA","",a[1]); $1=$1"\t"a[1]"x\t"a[2]; print $0}' |
    Rscript --vanilla --slave fit_LDdecay.R --header --col 5 --max_kb_dist 200 --fit_level 100
    --fit_boot 100 --plot_group Coverage --plot_wrap_formula 'Coverage~Type' --fit_bin_size
    250 --plot_bin_size 500 --plot_data -o PUBLISH.LD_r2_data.pdf

# Plot D' LD decay
ls $NAME*.ld_sampled.gz sort -V | awk 'BEGIN{OFS="\t"; print "File\tCoverage\tType"}
    {split($1,a,"_|.ld"); sub("OUTofAFRICA","",a[1]); $1=$1"\t"a[1]"x\t"a[2]; print $0}' |
    Rscript --vanilla --slave fit_LDdecay.R --header --col 5 --ld Dp --max_kb_dist 500
    --fit_level 100 --fit_boot 100 --plot_group Coverage --plot_wrap_formula 'Coverage~Type'
    --fit_bin_size 10 --plot_bin_size 100 --plot_data -o PUBLISH.LD_Dp_data.pdf

### LD blocks
```

```
zcat ${NAME}50_CG.ld.gz | cut -f 1,3,5- | scripts/LD_blocks.sh chrSim 35000 50000
mv LD_blocks.pdf PUBLISH.LD_blocks.pdf

### LD prunning
# Prune SNPs based on r^2 estimates from 50x coverage data
zcat ${NAME}50_CG.ld.gz | cut -f 1,3,5- | perl scripts/prune_graph.pl --max_kb_dist 200
    --min_weight 0.1 --weight_type a | sort -V > pruned_${NAME}50_CG.a.id

# Extract 0.1 fraction to plot
zcat ${NAME}50_CG.ld.gz | awk 'NR==FNR{x[$1]++} NR!=FNR && x[$1] && x[$3]{print}'
    pruned_${NAME}50_CG.a.id - | awk 'rand()<0.1' | gzip --best >
    pruned_${NAME}50_CG.a.ld_sampled.gz

# Plot pruned r^2 LD decay
ls pruned_$NAME*.a.ld_sampled.gz | awk 'BEGIN{OFS="\t"; print "File\tCoverage\tType"}
    {split($1,a,"_|\\."); sub("OUTofAFRICA","",a[2]); $1=$1"\t"a[2]"x\t"a[3]; print $0}' |
    Rscript --vanilla --slave fit_LDdecay.R --header --col 5 --fit_level 100 --plot_data
    --plot_no_legend -o PUBLISH.prunedLD_r2_data.pdf



#########################################################
# Pt. 2 - Real Data                                     #
#########################################################
# Duck:
#  SP=duck
#  REF=GCF_000355885.1_BGI_duck_1.0.fasta
#  BED=GCF_000355885.1_BGI_duck_1.0_genomic.exon.bed
# Turkey:
#  SP=turkey
#  REF=GCF_000146605.2_Turkey_5.0.fasta
#  BED=GCF_000146605.2_Turkey_5.0_genomic.exon.bed

# Trim low quality sections using trimmomatic
java -jar trimmomatic-0.36.jar PE -phred64 ${ID}_1.fastq ${ID}_2.fastq
${ID}_1P.fq ${ID}_1U.fq ${ID}_2P.fq ${ID}_2U.fq
ILLUMINACLIP:$ADAPTER_FILE:2:30:10 LEADING:3 TRAILING:3 SLIDINGWINDOW:4:15 MINLEN:36

# Use PALEOMIX to map the reads (PALEOMIX YAML files available from authors upon request)
bam_pipeline run --jar-root $PROGS_DIR --temp-root $TMP_DIR/paleomix/ --max-threads 10
    --bwa-max-threads 2 --destination _01.PALEOMIX $ID.PALEOMIX.yaml

# Set filters
N_IND=`cat _02.ANGSD/$SP.bamfiles | wc -l`
MINMAF=$(echo "scale=3; 2/(2*$N_IND)" | bc)

# Get EXON positions into ANGSD format
awk '{print $1"\t"($2+1)"\t"$3}' $BED > _02.ANGSD/$SP.sites
angsd sites index _02.ANGSD/$SP.sites

# Analise the quality of the sequences in the bam files
angsd -nThreads 10 -bam _02.ANGSD/$SP.bamfiles -sites _02.ANGSD/$SP.sites -ref $REF
    -uniqueOnly 1 -remove_bads 1 -only_proper_pairs 1 -C 50 -baq 1 -minMapQ 20 -minQ 20
    -doCounts 1 -doDepth 1 -maxDepth 1000 -doQsDist 1 -doPLots 3 -out _02.ANGSD/$SP.QC

# Calculate genotype likelihoods for each locus using ANGSD
angsd -nThreads 10 -bam _02.ANGSD/$SP.bamfiles -sites _02.ANGSD/$SP.sites -ref $REF
    -uniqueOnly 1 -remove_bads 1 -only_proper_pairs 1 -C 50 -baq 1 -minMapQ 20 -minQ 20
    -doCounts 1 -setMaxDepth 300 -minInd 7 -GL 1 -doMajorMinor 1 -doMaf 1 -minMaf $MINMAF
    -doGlf 2 -out _02.ANGSD/$SP

# Create position files and determine number of sites
zcat _02.ANGSD/$SP.mafs.gz | cut -f 1,2 | tail -n +2 > _02.ANGSD/$SP.pos
```

```
NS=‘cat _02.ANGSD/$SP.pos | wc -l‘

# Run ngsLD
ngsLD --n_threads 10 --geno _02.ANGSD/$SP.beagle.gz --pos _02.ANGSD/$SP.pos --probs --n_ind
    $N_IND --n_sites $NS --max_kb_dist 1000 | gzip --best > _03.LD/$SP.ld.gz

# Subsample LD results
zcat _03.LD/duck.ld.gz | awk 'rand()<0.02' | gzip --best > _03.LD/duck.ld_sampled.gz
zcat _03.LD/turkey.ld.gz | awk 'rand()<0.05' | gzip --best > _03.LD/turkey.ld_sampled.gz

# Run exponential curve fitting script
ls _03.LD/*.ld_sampled.gz | awk 'BEGIN{print "File\tSpecies"} {sp=$1; sub(".*/","",sp);
    sub("[.].*","",sp); sub(/\w/, substr(toupper(sp),1,1), sp); print $1"\t"sp}' | Rscript
    --vanilla --slave fit_LDdecay.R --header --ld r2 --max_kb_dist 200 --fit_bin_size 1000
    --fit_level 100 --fit_boot 100 --plot_group Species --plot_data --plot_no_legend
    --plot_scale 3 -o PUBLISH.real_data.LD_r2_data.pdf
```