Starling-May18
Projects/Katarina Stuart/KStuart.Starling-Aug18/At1_Genome/Analysis/2023-02-07.Recombination

PDF Version generated by

Katarina Stuart (z5188231@ad.unsw.edu.au)

on

Aug 23, 2023 @09:49 AM NZST

Table of Contents

2023-02-07.Recombination

Recombination

Regression based inference | LDjump:

https://onlinelibrary.wiley.com/doi/full/10.1111/1755-0998.12994?af=R

https://rdrr.io/github/PhHermann/LDJump/f/README.md

LD based inference| LDhat:

https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1933511/

Or reMIX:
https://github.com/adreau/ReMIX

LDhelmet:

https://github.com/popgenmethods/LDhelmet

Methods reference:

https://www.frontiersin.org/articles/10.3389/fgene.2022.738105/full (LDjump)

https://academic.oup.com/hmg/article/30/R1/R11/6096959 (has methods comparison table)

Regression based inference: FastEPRR

https://academic.oup.com/g3journal/article/6/6/1563/6029955

https://www.picb.ac.cn/evolgen/softwares/download/FastEPRR/FastEPRR2.0/FastEPRR2.0_manual.pdf

Good paper to follow for methods:

https://doi.org/10.1111/mec.16824

Kessler/CH_01/CH01_recombination_rate.Rmd · master · WiDGeT_TrentU / Graduate_theses · GitLab

Download

```
ls -lh /nesi/nobackup/uoa02613/kstuart_projects/programs/FastEPRR
FastEPRR_2.0.tar.gz
```

Filter the VCF

```
#!/bin/bash -e

#SBATCH --job-name=2023_02_07.recombination_filtering.sl
#SBATCH --account=uoa02613
```

```
#SBATCH --time=00-12:00:00
#SBATCH --mem=2GB
#SBATCH --output=%x_%j.errout
#SBATCH --mail-user=katarina.stuart@auckland.ac.nz
#SBATCH --mail-type=ALL
#SBATCH --nodes=1
#SBATCH --ntasks=1
#SBATCH --cpus-per-task=2
#SBATCH --profile task

# load modules
module load picard/2.26.10-Java-11.0.4
module load SAMtools/1.15.1-GCC-11.3.0

module load psmc/0.6.5-gimkl-2018b
module load VCFtools/0.1.15-GCC-9.2.0-Perl-5.30.1
module load BCFtools/1.15.1-GCC-11.3.0

cd /nesi/nobackup/uoa02613/kstuart_projects/At1_MynaGenome/analysis/recombination/FastEPRR

VCF=/nesi/nobackup/uoa02613/kstuart_projects/At1_MynaGenome/analysis/variant_calling/SNP_bcftools/myna_42inds.vcf.gz

vcftools --gzvcf ${VCF} --keep NZ_LEI_individuals.txt --min-meanDP 5 --max-meanDP 50 --mac 1 --recode --out recombination_NZ_LEI
vcftools --gzvcf ${VCF} --keep IND_nonoutlier_individuals.txt --min-meanDP 5 --max-meanDP 50 --mac 1 --recode --out recombination_IND
```

Prep the data and file paths (focusing in just IND for dry run)

```
cd /nesi/nobackup/uoa02613/kstuart_projects/At1_MynaGenome/analysis/recombination/FastEPRR
VCF=/nesi/nobackup/uoa02613/kstuart_projects/At1_MynaGenome/analysis/recombination/FastEPRR/recombination_IND.recode.vcf
GENOME=/nesi/nobackup/uoa02613/kstuart_projects/At1_MynaGenome/analysis/curation/step4_scaffolding/ragtag_atris_synteny/renamed/AcTris_vAus2.0.fasta
grep ">" $GENOME | cut -f1 | head -n 35 | sed 's/>//g' > primary_autosomes.txt
```

split the vcf

```
cd /nesi/nobackup/uoa02613/kstuart_projects/At1_MynaGenome/analysis/recombination/FastEPRR/vcf_scaffolds

grep "^#" $VCF > vcf_header.txt

for i in $(cat ../primary_autosomes.txt)
do
cat vcf_header.txt <(grep "^${i}\b" $VCF) > recombination_IND_${i}.vcf
done
```

convert in beagle and shapeit

```
#!/bin/bash -e

#SBATCH --job-name=2023_02_07.recombination_phasing.sl
#SBATCH --account=uoa02613
#SBATCH --time=00-12:00:00
#SBATCH --mem=1GB
#SBATCH --output=%x_%j.errout
#SBATCH --mail-user=katarina.stuart@auckland.ac.nz
#SBATCH --mail-type=ALL
#SBATCH --nodes=1
#SBATCH --ntasks=1
#SBATCH --cpus-per-task=2
#SBATCH --profile task
#SBATCH --array=1-35

CHROM=$(sed "${SLURM_ARRAY_TASK_ID}q;d" /nesi/nobackup/uoa02613/kstuart_projects/At1_MynaGenome/analysis/recombination/FastEPRR/primary_autosomes.txt)
echo "working with chrom:" $CHROM

PROGRAMS=/nesi/nobackup/uoa02613/kstuart_projects/programs

cd /nesi/nobackup/uoa02613/kstuart_projects/At1_MynaGenome/analysis/recombination/FastEPRR/vcf_scaffolds

java -Xmx1g -jar ${PROGRAMS}/beagle.27Jan18.7e1.jar gt=recombination_IND_${CHROM}.vcf out=recombination_IND_${CHROM}.beagle

# Main run
${PROGRAMS}/shapeit.v2.904.3.10.0-693.11.6.el7.x86_64/bin/shapeit -V recombination_IND_${CHROM}.beagle.vcf.gz -O shapeit_${CHROM} --output-log phase_logs/${CHROM}.main

# Convert back to phased vcf, containing only the phased sites that shapeit used
```

```
${PROGRAMS}/shapeit.v2.904.3.10.0-693.11.6.el7.x86_64/bin/shapeit -convert --input-haps shapeit_${CHROM} --output-vcf phase_shapeit_${CHROM} --output-log phase_logs/${CHROM
```

Largest 3 chroms (1, 2, 3) did not run to completion. Rerun! 1-4

Not sure about the step? Change chrom names to numeric?

```
cd /nesi/nobackup/uoa02613/kstuart_projects/At1_MynaGenome/analysis/recombination/FastEPRR
VCF=/nesi/nobackup/uoa02613/kstuart_projects/At1_MynaGenome/analysis/recombination/FastEPRR/recombination_IND.recode.vcf
cat vcf_scaffolds/vcf_header.txt <(grep -f primary_autosomes.txt $VCF) | sed 's/Superscaffold_chr//g' > recombination_IND_autosomes_renamed.vcf
sed 's/Superscaffold_chr//g' primary_autosomes.txt  > primary_autosomes_renamed.txt
```

The step1-3, and rho2r scripts

```
#step1_FastEPRR.R
setwd("/nesi/nobackup/uoa02613/kstuart_projects/At1_MynaGenome/analysis/recombination/FastEPRR")
require("FastEPRR")
library("FastEPRR")

FastEPRR_VCF_step1(vcfFilePath = "/nesi/nobackup/uoa02613/kstuart_projects/At1_MynaGenome/analysis/recombination/FastEPRR/vcf_scaffolds/phase_shapeit_Superscaffold_chr12.v
"/nesi/nobackup/uoa02613/kstuart_projects/At1_MynaGenome/analysis/recombination/FastEPRR/step1_IND/chr_12/chr_12")

#step2_FastEPRR.R
setwd("/nesi/nobackup/uoa02613/kstuart_projects/At1_MynaGenome/analysis/recombination/FastEPRR")
#require("FastEPRR")
library("FastEPRR")

FastEPRR_VCF_step2(srcFolderPath = "/nesi/nobackup/uoa02613/kstuart_projects/At1_MynaGenome/analysis/recombination/FastEPRR/step1_IND/chr_12", jobNumber=1, currJob=1, DX
"/nesi/nobackup/uoa02613/kstuart_projects/At1_MynaGenome/analysis/recombination/FastEPRR/step2_IND/chr_12/")

#step3_FastEPRR.R
setwd("/nesi/nobackup/uoa02613/kstuart_projects/At1_MynaGenome/analysis/recombination/FastEPRR")
#require("FastEPRR")
library("FastEPRR")

FastEPRR_VCF_step3(srcFolderPath = "/nesi/nobackup/uoa02613/kstuart_projects/At1_MynaGenome/analysis/recombination/FastEPRR/step1_IND/chr_12", DXFolderPath =
"/nesi/nobackup/uoa02613/kstuart_projects/At1_MynaGenome/analysis/recombination/FastEPRR/step2_IND/chr_12/", finalOutputFolderPath = "/nesi/nobackup/uoa02613/kstuart_projects/A

#rho2r_FastEPRR.R
setwd("/nesi/nobackup/uoa02613/kstuart_projects/At1_MynaGenome/analysis/recombination/FastEPRR")
library("FastEPRR")

FastEPRR_rho2r(inputFilePath = "/nesi/nobackup/uoa02613/kstuart_projects/At1_MynaGenome/analysis/recombination/FastEPRR/step3_IND/chr12", outputFilePath =
"/nesi/nobackup/uoa02613/kstuart_projects/At1_MynaGenome/analysis/recombination/FastEPRR/step3_IND/transformed_chr12", Ne = 74745.06)
```

Editing the step1-3 + rho2r scripts

```
for CHROM in $(cat primary_autosomes_renamed.txt)
do
mkdir /nesi/nobackup/uoa02613/kstuart_projects/At1_MynaGenome/analysis/recombination/FastEPRR/step1_IND/chr_${CHROM}/
sed "s/12/${CHROM}/g" step1_FastEPRR.R > step1_IND_scripts/step1_FastEPRR_${CHROM}.R
done
```

```
#need to change from names
for CHROM in $(cat primary_autosomes_renamed.txt)
do
#sed 's/Superscaffold_chr//g'
/nesi/nobackup/uoa02613/kstuart_projects/At1_MynaGenome/analysis/recombination/FastEPRR/step1_IND/chr_${CHROM}/chr_${CHROM} > /nesi/nobackup/uoa02613/kstuart_projects/At
rm /nesi/nobackup/uoa02613/kstuart_projects/At1_MynaGenome/analysis/recombination/FastEPRR/step1_IND/chr_${CHROM}/chr_${CHROM}
mv
/nesi/nobackup/uoa02613/kstuart_projects/At1_MynaGenome/analysis/recombination/FastEPRR/step1_IND/chr_${CHROM}/chr_${CHROM}_edit /nesi/nobackup/uoa02613/kstuart_projects
done
```

```
for CHROM in $(cat primary_autosomes_renamed.txt)
do
#mkdir /nesi/nobackup/uoa02613/kstuart_projects/At1_MynaGenome/analysis/recombination/FastEPRR/step2_IND/chr_${CHROM}
sed "s/12/${CHROM}/g" step2_FastEPRR.R > step2_IND_scripts/step2_FastEPRR_${CHROM}.R
done
```

```
for CHROM in $(cat primary_autosomes_renamed.txt)
do
```

```
sed "s/12/${CHROM}/g" step3_FastEPRR.R > step3_IND_scripts/step3_FastEPRR_${CHROM}.R
done
```

```
for CHROM in $(cat primary_autosomes_renamed_rho2r.txt)
do
sed "s/12/${CHROM}/g" rho2r_FastEPRR.R > rho2r_IND_scripts/rho2r_FastEPRR_${CHROM}.R
done
```

Step 1-3 r script: basic stencil

```
#!/bin/bash -e

#SBATCH --job-name=2023_02_10.recombination_step1.sl
#SBATCH --account=uoa02613
#SBATCH --time=00-12:00:00
#SBATCH --mem=1GB
#SBATCH --output=%x_%j.errout
#SBATCH --mail-user=katarina.stuart@auckland.ac.nz
#SBATCH --mail-type=ALL
#SBATCH --nodes=1
#SBATCH --ntasks=1
#SBATCH --cpus-per-task=2
#SBATCH --profile task
#SBATCH --array=1-35

CHROM=$(sed "${SLURM_ARRAY_TASK_ID}q;d" /nesi/nobackup/uoa02613/kstuart_projects/At1_MynaGenome/analysis/recombination/FastEPRR/primary_autosomes_renamed.txt)
echo "working with kemr:" $CHROM

module load R/4.1.0-gimkl-2020a

cd /nesi/nobackup/uoa02613/kstuart_projects/At1_MynaGenome/analysis/recombination/FastEPRR
echo "Executing R ..."
srun Rscript step1_IND_scripts/step1_FastEPRR_${CHROM}.R
echo "R finished."

cd /nesi/nobackup/uoa02613/kstuart_projects/At1_MynaGenome/analysis/recombination/FastEPRR
echo "Executing R ..."
srun Rscript step2_IND_scripts/step2_FastEPRR_${CHROM}.R
echo "R finished."

cd /nesi/nobackup/uoa02613/kstuart_projects/At1_MynaGenome/analysis/recombination/FastEPRR
echo "Executing R ..."
srun Rscript step3_IND_scripts/step3_FastEPRR_${CHROM}.R
echo "R finished."
```

Compute Ne for Rho to r transformation. Start by calculating pi

```
#!/bin/bash -e

#SBATCH --job-name=2023_02_14.recombination_pi.sl
#SBATCH --account=uoa02613
#SBATCH --time=00-12:00:00
#SBATCH --mem=1GB
#SBATCH --output=%x_%j.errout
#SBATCH --mail-user=katarina.stuart@auckland.ac.nz
#SBATCH --mail-type=ALL
#SBATCH --nodes=1
#SBATCH --ntasks=1
#SBATCH --cpus-per-task=2
#SBATCH --profile task

module load VCFtools/0.1.15-GCC-9.2.0-Perl-5.30.1

cd /nesi/nobackup/uoa02613/kstuart_projects/At1_MynaGenome/analysis/recombination/FastEPRR/pi_IND

VCF=/nesi/nobackup/uoa02613/kstuart_projects/At1_MynaGenome/analysis/recombination/FastEPRR/recombination_IND.recode.vcf
vcftools --vcf ${VCF} --window-pi 50000 --window-pi-step 5000 --out recombination_IND

module load R/4.1.0-gimkl-2020a
R
setwd("/nesi/nobackup/uoa02613/kstuart_projects/At1_MynaGenome/analysis/recombination/FastEPRR/pi_IND")
```

```
pi_IND_table <- read.table("/nesi/nobackup/uoa02613/kstuart_projects/At1_MynaGenome/analysis/recombination/FastEPRR/pi_IND/recombination_IND.windowed.pi", sep="\t", header=T)
pi_IND <- mean(pi_IND_table$PI, na.rm = TRUE)

Ne_funct <- function(pi, mu){
  ne <- (1/4) * (pi/mu)
  return(ne)
}

# mu used from msmc analysis
mu <- 1.33e-8

# compute ne
ne_IND <- Ne_funct(pi_IND, mu) #74745.06
ne_IND
```

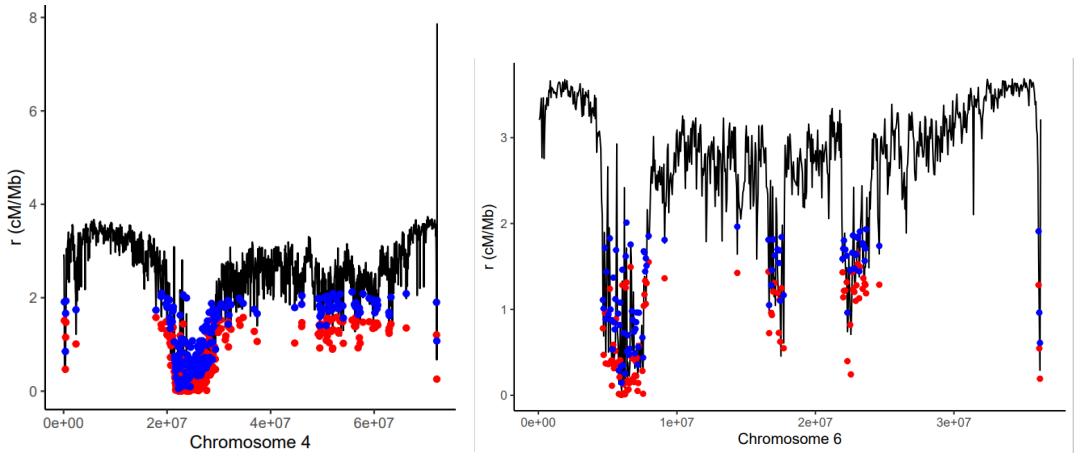
run the final step for conversion of rho to r
note, that different individual list file run to remove the characters from e.g. from 5c -> 5.3

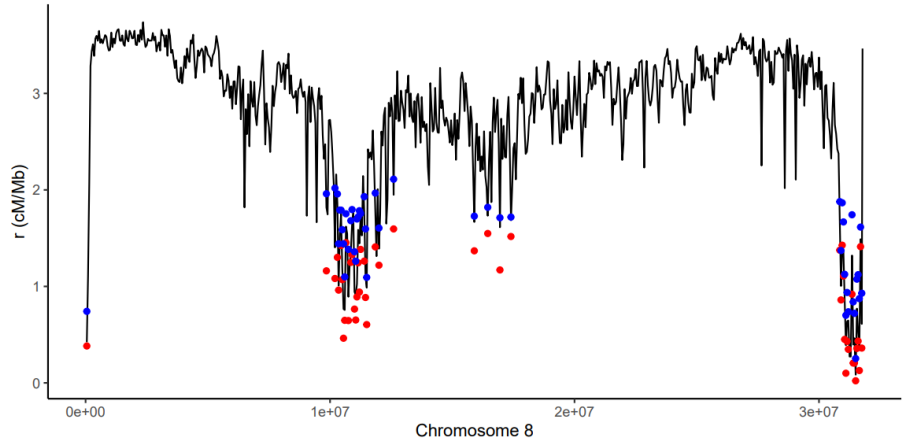
```
#!/bin/bash -e

#SBATCH --job-name=2023_02_20.recombination_rho2r.sl
#SBATCH --account=uoa02613
#SBATCH --time=00-12:00:00
#SBATCH --mem=1GB
#SBATCH --output=%x_%j.errout
#SBATCH --mail-user=katarina.stuart@auckland.ac.nz
#SBATCH --mail-type=ALL
#SBATCH --nodes=1
#SBATCH --ntasks=1
#SBATCH --cpus-per-task=2
#SBATCH --profile task
#SBATCH --array=1-35
#SBATCH --partition=milan

CHROM=$(sed "${SLURM_ARRAY_TASK_ID}q;d" /nesi/nobackup/uoa02613/kstuart_projects/At1_MynaGenome/analysis/recombination/FastEPRR/primary_autosomes_renamed_rho2r.txt
echo "working with kemr:" $CHROM

#FILENAME: rho2r_FastEPRR.R

module load R/4.1.0-gimkl-2020a

cd /nesi/nobackup/uoa02613/kstuart_projects/At1_MynaGenome/analysis/recombination/FastEPRR
echo "Executing R ..."
srun Rscript rho2r_IND_scripts/rho2r_FastEPRR_${CHROM}.R
echo "R finished."
```

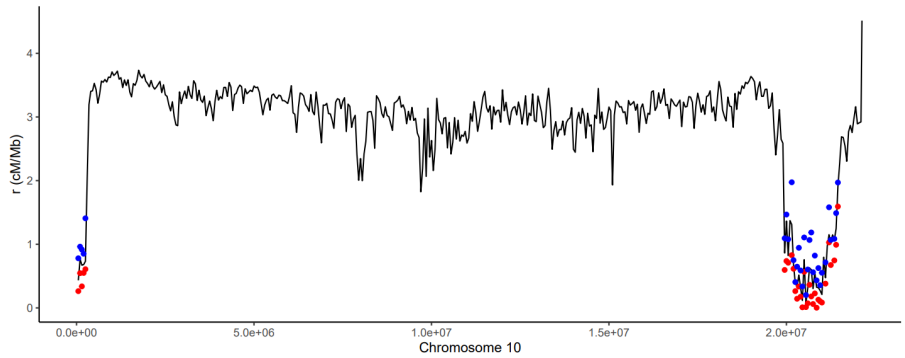
quick plot

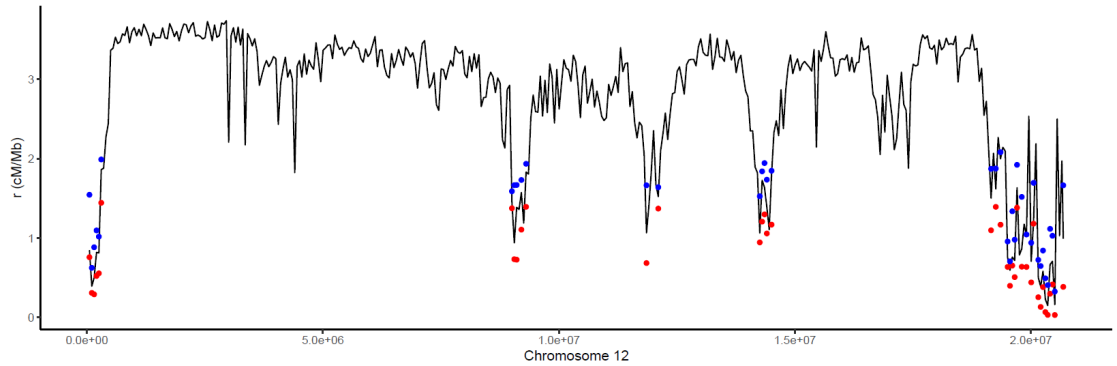
```
#!/bin/bash -e

#SBATCH --job-name=2023_02_20.recombination_plot.sl
#SBATCH --account=uoa02613
#SBATCH --time=00-12:00:00
#SBATCH --mem=1GB
#SBATCH --output=%x_%j.errout
#SBATCH --mail-user=katarina.stuart@auckland.ac.nz
#SBATCH --mail-type=ALL
#SBATCH --nodes=1
#SBATCH --ntasks=1
#SBATCH --cpus-per-task=2
#SBATCH --profile task
#SBATCH --array=1-35
#SBATCH --partition=milan

CHROM=$(sed "${SLURM_ARRAY_TASK_ID}q;d" /nesi/nobackup/uoa02613/kstuart_projects/At1_MynaGenome/analysis/recombination/FastEPRR/primary_autosomes_renamed_rho2r.txt
echo "working with kemr:" $CHROM


#FILENAME: plot_FastEPRR.R

module load R/4.1.0-gimkl-2020a

cd /nesi/nobackup/uoa02613/kstuart_projects/At1_MynaGenome/analysis/recombination/FastEPRR
```

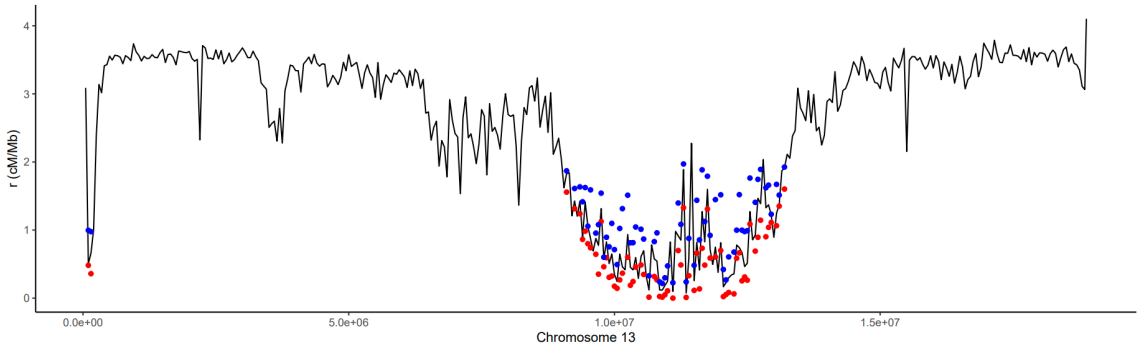
```
echo "Executing R ..."
srun Rscript plots_IND_scripts/plot_FastEPRR_${CHROM}.R
echo "R finished."


#edit script
for CHROM in $(cat primary_autosomes_renamed_rho2r.txt)
do
sed "s/12/${CHROM}/g" plot_FastEPRR.R > plots_IND_scripts/plot_FastEPRR_${CHROM}.R
done

#plot_FastEPRR.R
setwd("/nesi/nobackup/uoa02613/kstuart_projects/At1_MynaGenome/analysis/recombination/FastEPRR/plots_IND")
library(ggplot2)

r_IND <- read.table("/nesi/nobackup/uoa02613/kstuart_projects/At1_MynaGenome/analysis/recombination/FastEPRR/step3_IND/transformed_chr12", sep=" ", header=T)

pdf("At1_recombination_chr12_IND.pdf", width=12, height=4)
ggplot(data=r_IND, aes(x=End, y=r.cM.Mb.)) + geom_path()+
geom_point(data=r_IND, aes(x=End, y=CIL.cM.Mb.), color="red" ) +
geom_point(data=r_IND, aes(x=End, y=CIR.cM.Mb.), color="blue" ) +
theme_classic() + xlab("Chromosome 12") + ylab("r (cM/Mb)")
dev.off()
```

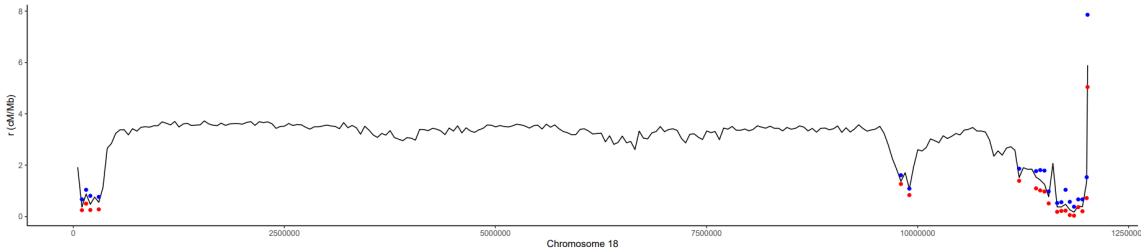












LD based inference: LDhat

https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1933511/
Manual: https://ldhat.sourceforge.net/manual.pdf
some methods:
https://github.com/QuentinRougemont/LDhat_workflow/blob/master/02-scripts/02.interval_iteration.sh

remove invariant sites: https://sourceforge.net/p/ldhat/mailman/ldhat-help/?page=3
Removing singletons **(REFERENCE - mention "use of LDhat with HQ genome)** does not seem to be important: Fine-Scale Recombination Maps of Func Hotspots - PMC (nih.gov)

we believe the utility of adding extra individuals to resolve species recombination patterns outweights the confounding factors. Early iterations ran on each r

Install

```
cd /nesi/nobackup/uoa02613/kstuart_projects/programs
git clone https://github.com/auton1/LDhat.git
cd LDhat
make
```

do once with these files to compare to below ethods.
then again with more restricted IND samples (but still small sample group)

Step 1: file conversion using vcftools
check filtering used in example

LDHAT cannot deal with too many snps:
https://sourceforge.net/p/ldhat/mailman/message/30599446/

Filter for just MP and TN individuals

```
#!/bin/bash -e

#SBATCH --job-name=2023_03_24.ldhat_filter.sl
#SBATCH --account=uoa02613
#SBATCH --time=00-12:00:00
#SBATCH --mem=1GB
#SBATCH --output=%x_%j.errout
#SBATCH --mail-user=katarina.stuart@auckland.ac.nz
#SBATCH --mail-type=ALL
#SBATCH --nodes=1
#SBATCH --ntasks=1
#SBATCH --cpus-per-task=2
#SBATCH --profile task

module load VCFtools/0.1.15-GCC-9.2.0-Perl-5.30.1

cd /nesi/nobackup/uoa02613/kstuart_projects/At1_MynaGenome/analysis/recombination/LDhat/vcf_split

#cp /nesi/nobackup/uoa02613/kstuart_projects/At1_MynaGenome/analysis/recombination/FastEPRR/IND_nonoutlier_individuals.txt IND_MP_TN_individuals.txt #(removed top 5 inds)

VCF=/nesi/nobackup/uoa02613/kstuart_projects/At1_MynaGenome/analysis/variant_calling/SNP_bcftools/myna_42inds.vcf.gz
vcftools --gzvcf ${VCF} --keep IND_MP_TN_individuals.txt  --min-meanDP 5 --max-meanDP 50 --mac 1 --recode --out recombination_IND_MP_TN
```

convert

```
#!/bin/bash -e

#SBATCH --job-name=2023_03_24.ldhat_convert.sl
#SBATCH --account=uoa02613
#SBATCH --time=00-12:00:00
#SBATCH --mem=1GB
#SBATCH --output=%x_%j.errout
#SBATCH --mail-user=katarina.stuart@auckland.ac.nz
#SBATCH --mail-type=ALL
#SBATCH --nodes=1
#SBATCH --ntasks=1
#SBATCH --cpus-per-task=2
#SBATCH --profile task
#SBATCH --array=1-35

#just doing subset for now
CHROM=$(sed "${SLURM_ARRAY_TASK_ID}q;d" /nesi/nobackup/uoa02613/kstuart_projects/At1_MynaGenome/analysis/recombination/FastEPRR/primary_autosomes.txt)
echo "working with chrom:" $CHROM

module load VCFtools/0.1.15-GCC-9.2.0-Perl-5.30.1

cd /nesi/nobackup/uoa02613/kstuart_projects/At1_MynaGenome/analysis/recombination/LDhat/vcf_split

VCF=recombination_IND_MP_TN.recode.vcf
vcftools --vcf ${VCF} --chr ${CHROM} --thin 1000 --ldhat-geno --out recombination_IND_MP_TN.${CHROM}  #to make ldhat output
vcftools --vcf ${VCF} --chr ${CHROM} --thin 1000 --recode --out recombination_IND_MP_TN.${CHROM}  #to mkae SNP output to link back to the LD stats
```

Step 2: LD calculations with LDhat complete

even if your data is unphased, it seems you need 2 number of samples in your data set. Only a single 1k file needed for all separate chromosomes.

[likelihood file for different no. seqs than data · Issue #13 · auton1/LDhat (github.com)](https://github.com/auton1/LDhat/issues/13)

```
#!/bin/bash -e

#SBATCH --job-name=2023_02_24.recombination_complete.sl
#SBATCH --account=uoa02613
#SBATCH --time=00-12:00:00
#SBATCH --mem=1GB
#SBATCH --output=%x_%j.errout
#SBATCH --mail-user=katarina.stuart@auckland.ac.nz
#SBATCH --mail-type=ALL
#SBATCH --nodes=1
#SBATCH --ntasks=1
#SBATCH --cpus-per-task=2
#SBATCH --profile task

cd /nesi/nobackup/uoa02613/kstuart_projects/At1_MynaGenome/analysis/recombination/LDhat/

LDHAT=/nesi/nobackup/uoa02613/kstuart_projects/programs/LDhat/

#grabbing and inzipping prebaked 1k file. You can do this so long as you are reducing the '-n' value
cp ${LDHAT}/lk_files/lk_n100_t0.001.gz .

gunzip lk_n100_t0.001.gz

#Reduce -n from 100 to 30 (15 individuals * 2)
${LDHAT}/lkgen -lk lk_n100_t0.001 -nseq 30

mv new_lk.txt new_lk_30.txt

#if you have more individuals than the pre-prepared 1k files you will need to run complete yourself
```

Step 3: LD calculations with LDhat internval

```
#!/bin/bash -e

#SBATCH --job-name=2023_03_24.ldhat_interval.sl
#SBATCH --account=uoa02613
#SBATCH --time=00-12:00:00
#SBATCH --mem=5GB
#SBATCH --output=%x_%j.errout
#SBATCH --mail-user=katarina.stuart@auckland.ac.nz
#SBATCH --mail-type=ALL
#SBATCH --nodes=1
#SBATCH --ntasks=1
#SBATCH --cpus-per-task=2
#SBATCH --profile task
#SBATCH --array=1-35

#just doing subset for now
CHROM=$(sed "${SLURM_ARRAY_TASK_ID}q;d" /nesi/nobackup/uoa02613/kstuart_projects/At1_MynaGenome/analysis/recombination/FastEPRR/primary_autosomes.txt)
echo "working with chrom:" $CHROM

DIR=/nesi/nobackup/uoa02613/kstuart_projects/At1_MynaGenome/analysis/recombination/LDhat/
LDHAT=/nesi/nobackup/uoa02613/kstuart_projects/programs/LDhat/

cd ${DIR}/IND_chroms

mkdir ${CHROM} && cd $_

${LDHAT}/interval -seq ${DIR}/vcf_split/recombination_IND_MP_TN.${CHROM}.ldhat.sites -loc ${DIR}/vcf_split/recombination_IND_MP_TN.${CHROM}.ldhat.locs -lk ${DIR}/new_lk_30.txt -i


##############

#!/bin/bash -e

#SBATCH --job-name=2023_03_27.ldhat_interval_chr1.sl
#SBATCH --account=uoa02613
#SBATCH --time=00-12:00:00
#SBATCH --mem=15GB
#SBATCH --output=%x_%j.errout
```

```
#SBATCH --mail-user=katarina.stuart@auckland.ac.nz
#SBATCH --mail-type=ALL
#SBATCH --nodes=1
#SBATCH --ntasks=1
#SBATCH --cpus-per-task=2
#SBATCH --profile task

SLURM_ARRAY_TASK_ID=1

#just doing subset for now
CHROM=$(sed "${SLURM_ARRAY_TASK_ID}q;d" /nesi/nobackup/uoa02613/kstuart_projects/At1_MynaGenome/analysis/recombination/FastEPRR/primary_autosomes.txt)
echo "working with chrom:" $CHROM

DIR=/nesi/nobackup/uoa02613/kstuart_projects/At1_MynaGenome/analysis/recombination/LDhat/
LDHAT=/nesi/nobackup/uoa02613/kstuart_projects/programs/LDhat/

cd ${DIR}/IND_chroms

mkdir ${CHROM} && cd $_

${LDHAT}/interval -seq ${DIR}/vcf_split/recombination_IND_MP_TN.${CHROM}.ldhat.sites -loc ${DIR}/vcf_split/recombination_IND_MP_TN.${CHROM}.ldhat.locs -lk ${DIR}/new_lk_30.txt -i

echo "DONE"
```

Step 4: Using LDhat stat to summarise the output of LDhat interval. Produced file called 'res' which you can plot from.

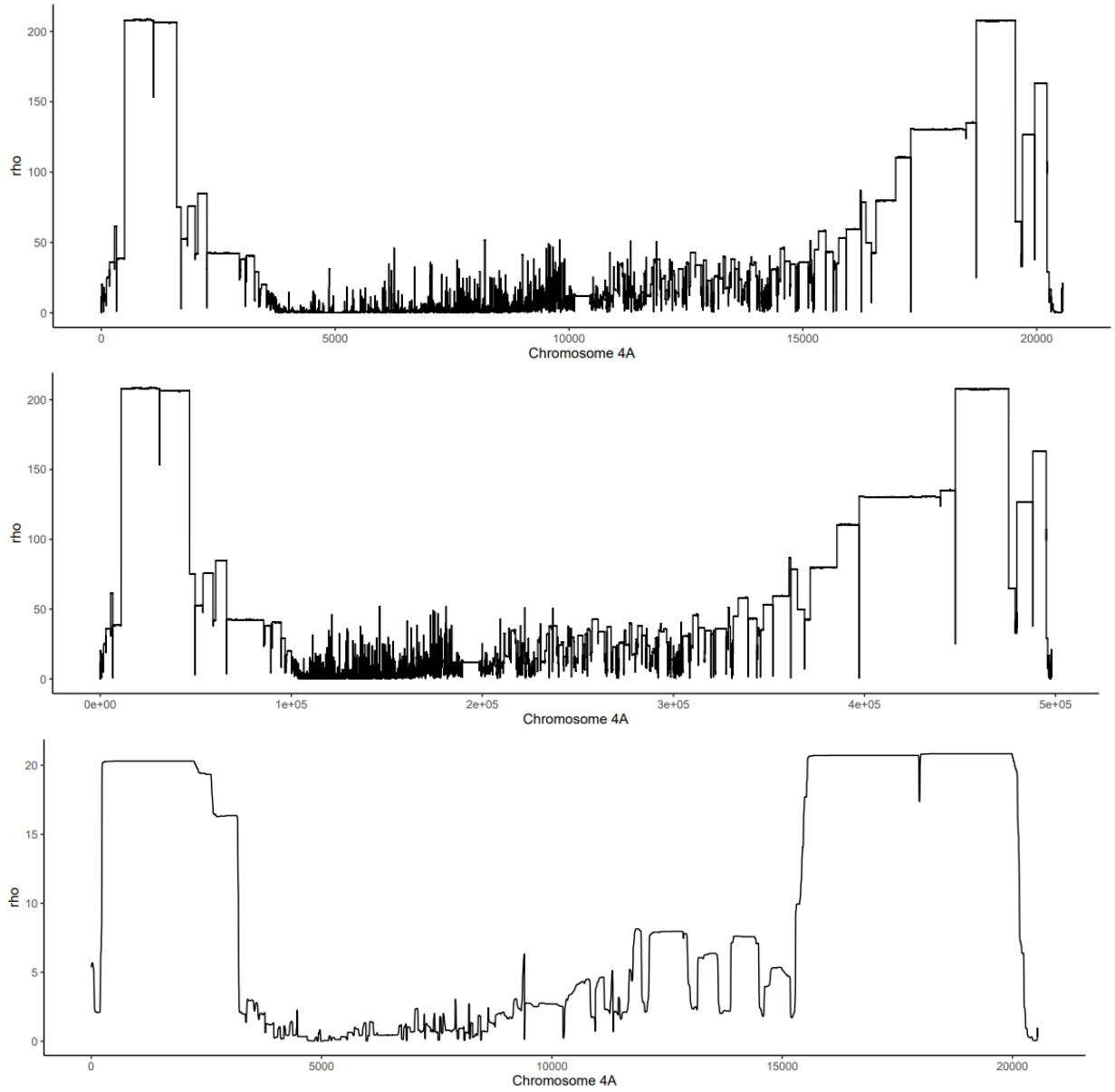
```
#!/bin/bash -e

#SBATCH --job-name=2023_03_27.ldhat_stat.sl
#SBATCH --account=uoa02613
#SBATCH --time=00-12:00:00
#SBATCH --mem=10GB
#SBATCH --output=%x_%j.errout
#SBATCH --mail-user=katarina.stuart@auckland.ac.nz
#SBATCH --mail-type=ALL
#SBATCH --nodes=1
#SBATCH --ntasks=1
#SBATCH --cpus-per-task=2
#SBATCH --profile task
#SBATCH --array=1-35

#just doing subset for now
CHROM=$(sed "${SLURM_ARRAY_TASK_ID}q;d" /nesi/nobackup/uoa02613/kstuart_projects/At1_MynaGenome/analysis/recombination/FastEPRR/primary_autosomes.txt)
echo "working with chrom:" $CHROM

DIR=/nesi/nobackup/uoa02613/kstuart_projects/At1_MynaGenome/analysis/recombination/LDhat/
LDHAT=/nesi/nobackup/uoa02613/kstuart_projects/programs/LDhat/

cd ${DIR}/IND_chroms/${CHROM}

${LDHAT}/stat -input rates.txt -burn 250 -loc ${DIR}/vcf_split/recombination_IND_MP_TN.${CHROM}.ldhat.locs

echo "DONE"
```

Step 5: plot

```
#!/bin/bash -e

#SBATCH --job-name=2023_03_27.ldhat_map.sl
#SBATCH --account=uoa02613
#SBATCH --time=00-12:00:00
#SBATCH --mem=10GB
#SBATCH --output=%x_%j.errout
#SBATCH --mail-user=katarina.stuart@auckland.ac.nz
#SBATCH --mail-type=ALL
#SBATCH --nodes=1
#SBATCH --ntasks=1
#SBATCH --cpus-per-task=2
#SBATCH --profile task
#SBATCH --array=1-35

#just doing subset for now
CHROM=$(sed "${SLURM_ARRAY_TASK_ID}q;d" /nesi/nobackup/uoa02613/kstuart_projects/At1_MynaGenome/analysis/recombination/FastEPRR/primary_autosomes.txt)
```

```
echo "working with chrom:" $CHROM

#prep res file for plotting
cd /nesi/nobackup/uoa02613/kstuart_projects/At1_MynaGenome/analysis/recombination/LDhat/IND_chroms/${CHROM}/

tail -n +3 res.txt | sed 's/ //g' > res_edit.txt

grep -v "^#" /nesi/nobackup/uoa02613/kstuart_projects/At1_MynaGenome/analysis/recombination/LDhat/vcf_split/recombination_IND_MP_TN.${CHROM}.recode.vcf | cut -f1,2 >
/nesi/nobackup/uoa02613/kstuart_projects/At1_MynaGenome/analysis/recombination/LDhat/vcf_split/recombination_IND_MP_TN.${CHROM}.snps

paste <(head -n -1 recombination_IND_MP_TN.${CHROM}.snps) res_edit.txt | awk -v OFS="," '{print $1,$2,$2+1,$4}' > recombination_IND_MP_TN.${CHROM}.recombination

##########all together now

cd /nesi/nobackup/uoa02613/kstuart_projects/At1_MynaGenome/analysis/recombination/LDhat/IND_chroms

#generate file list in the order I want them added together in cat
ls -lv */*recombination | sed -e "s/[[:space:]]\+/\t/g" | cut -f9 > recombination_file_list.txt

xargs -i cat '{}' < recombination_file_list.txt > recombination_rho_allchroms.txt

#try thinning the above?? delete every 10 SNPs?
awk 'NR % 10 == 0'  recombination_rho_allchroms.txt >  recombination_rho_allchroms_thin10.txt
awk 'NR % 100 == 0'  recombination_rho_allchroms.txt >  recombination_rho_allchroms_thin100.txt
```

```
###############

#plot_LDhat.R
module load R/4.1.0-gimkl-2020a
R
setwd("/nesi/nobackup/uoa02613/kstuart_projects/At1_MynaGenome/analysis/recombination/LDhat/IND_chroms/Superscaffold_chr3")
library(ggplot2)

r_IND <- read.table("res_edit.txt", sep="\t", header=F)

pdf("At1_recombination_ldhat_Superscaffold_chr3_IND.pdf", width=12, height=4)
ggplot(data=r_IND, aes(x=V1, y=V2)) + geom_path()+
#geom_point(data=r_IND, aes(x=V1, y=V4), color="red" ) +
#geom_point(data=r_IND, aes(x=V1, y=V5), color="blue" ) +
theme_classic() + xlab("Superscaffold_chr3") + ylab("rho")
dev.off()
```



plot like:

https://genome.cshlp.org/content/20/4/496.full.pdf

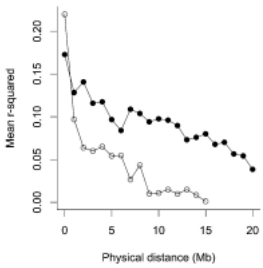


Figure 1. Mean r^2 for distance bins of 1 Mb for macrochromosomes and microchromosomes plotted against physical distance (Mb).

what about the smallest bin???

Trying to make the above plot

```
#!/bin/bash -e

#SBATCH --job-name=2023_03_29.ldhat_pairwise_marco.sl
#SBATCH --account=uoa02613
#SBATCH --time=00-48:00:00
#SBATCH --mem=5GB
#SBATCH --output=%x_%j.errout
#SBATCH --mail-user=katarina.stuart@auckland.ac.nz
#SBATCH --mail-type=ALL
#SBATCH --nodes=1
#SBATCH --ntasks=1
#SBATCH --cpus-per-task=2
#SBATCH --profile task

module load VCFtools/0.1.15-GCC-9.2.0-Perl-5.30.1

cd /nesi/nobackup/uoa02613/kstuart_projects/At1_MynaGenome/analysis/recombination/LDhat/pairwise

VAR=""

for CHROM in $(cat /nesi/nobackup/uoa02613/kstuart_projects/At1_MynaGenome/figures/circos/panels/macro_superscaffolds.txt)
do
VAR="${VAR} --chr $CHROM "
done

echo $VAR

VCF=/nesi/nobackup/uoa02613/kstuart_projects/At1_MynaGenome/analysis/variant_calling/SNP_bcftools/myna_42inds_filtered.recode.vcf
WIDTH=10000000
vcftools --vcf ${VCF} $VAR --thin 1000 --geno-r2 --ld-window-bp ${WIDTH} --out pairwise.MACRO.${WIDTH}
```

or just not split by chr, split my micro/macro?

but will be a very very big file

```
#!/bin/bash -e

#SBATCH --job-name=2023_03_30.pairwise_bin_macro.sl
#SBATCH --account=uoa02613
#SBATCH --time=00-12:00:00
#SBATCH --mem=1GB
#SBATCH --output=%x_%j.errout
#SBATCH --mail-user=katarina.stuart@auckland.ac.nz
#SBATCH --mail-type=ALL
#SBATCH --nodes=1
#SBATCH --ntasks=1
#SBATCH --cpus-per-task=2
#SBATCH --profile task

cd /nesi/nobackup/uoa02613/kstuart_projects/At1_MynaGenome/analysis/recombination/LDhat/pairwise

#want to obtain the distance in BP between sites. Minus one SNP pos off another. And remove Nans.
awk -v OFS="\t" '{print $1, $3-$2, $5}' pairwise.MACRO.10000000.geno.ld | grep -v "nan" > pairwise.MACRO.10000000.geno.ld_format.txt

touch MACRO_10Mb_outfile.txt

while read -r first second; do
    echo "$first" "$second"
awk -v start=${first} -v end=${second} '($2>start && $2<end)' pairwise.MACRO.10000000.geno.ld_format.txt | datamash mean 3 >> MACRO_10Mb_outfile.txt
echo "done"
done < interval_windows_10Mb.txt
```

```
#Move into R
module load R/4.1.0-gimkl-2020a
R
setwd("/nesi/nobackup/uoa02613/kstuart_projects/At1_MynaGenome/analysis/recombination/LDhat/pairwise")
library(ggplot2)

intervals <- read.table("interval_windows_10Mb.txt", sep="\t", header=F)

macro_r <- read.table("MACRO_10Mb_outfile.txt", sep="\t", header=F)
macro_r$order <- (intervals[,2] / 1000000)
```

```
macro_r$chrom <- c("Macro")

micro_r <- read.table("MICRO_10Mb_outfile.txt", sep="\t", header=F)
micro_r$order <- (intervals[,2] / 1000000)
micro_r$chrom <- c("Micro")

r2<-rbind(macro_r,micro_r)

pdf("At1_recombination.pdf")
ggplot(data=r2, mapping=aes(x=order,y=V1,group=chrom)) + geom_line(size=1.4,aes(color=chrom)) + geom_point(stroke=2,fill="white",size=3,shape=21,aes(color=chrom)) +
ylab("R^2") + xlab("1 Mb distance bins") + theme_classic(base_size = 18) + theme(axis.text=element_text(size=16))+scale_color_manual(values=c("#6d442d", "#3767a7"))+ scale_x_contin
dev.off()
```