ddrad phase 2 project

Caccone PostDoc

Gus Dunn

April, 2015

Contents

1 Tasks

1.1 BEAST

1.1.1 –To DO–

-

1.1.2 –Completed–

- `[wont do]` Convert BAMs to NEXSUS
  - waiting to hear back from admins about getting permissions to AndreaG's BAMs
- `[wont do]` BEAST configuration
- `[wont do]` attempt BEAST run
- `[2015-03-13]` meeting with GisellaC and Aris 2015-03-13 at 11
- `[2015-03-12]` conversation with Aris
- `[wont do]` write up conversation with Aris for GisellaC and get clearance to proceed.

1.2 Linkage disequilibrium thresholds for SNP-pairs

1.2.1 –To Do–

- `[ ]`

1.2.2 –Completed–

- `[2015-03-12]` set up and yield models
- `[2015-03-12]` take model and return parameters
- `[2015-03-12]` take parameters and df and set value for each SNP-pair's probability ($1 - \text{CDF}$)
- `[2015-03-12]` take df and set value for each SNP-pair's BH corrected probability

---

2 Contig proximity graph

2.1 2015-03-10 (Tuesday)

- calculate LD only between **INTER-** contig SNPS **[Conversation with JoshM]**

2.1.1 Calculate interchromosomal LD with `vcftools`

2.1.1.1 Attempt 1 [FAILED: bug in v0.1.12b]

- -INPUT- -

```
SNP_DIR="/home2/wd238/data/genomes/glossina_fuscipes/annotations/SNPs"

VCF="${SNP_DIR}/tsetseFINAL_14Oct2014_f2_53.recode.renamed_scaffolds.maf0_05.vcf"

OUT_PREFIX="${SNP_DIR}/vcftools_out/tsetseFINAL_14Oct2014_f2_53.recode.renamed_scaffolds

mkdir -p ${SNP_DIR}/vcftools_out/

vcftools --vcf $VCF  --out $OUT_PREFIX --interchrom-geno-r2
```

- -OUTPUT- -

```
VCFtools - v0.1.12b
(C) Adam Auton and Anthony Marcketta 2009

Parameters as interpreted:
        --vcf /home2/wd238/data/genomes/glossina_fuscipes/annotations/SNPs/tsetseFINAL_1
        --max-alleles 2
        --min-alleles 2
        --interchrom-geno-r2
        --out /home2/wd238/data/genomes/glossina_fuscipes/annotations/SNPs/vcftools_out/

After filtering, kept 53 out of 53 Individuals
Outputting Interchromosomal Pairwise Genotype LD (bi-allelic only)
Error: Require phased haplotypes for r^2 calculation (use --phased)
```

2.1.1.1.1 Email to vcftools-help

I have recently tried to run the following command

```
$ vcftools --vcf $VCF  --out $OUT_PREFIX --interchrom-geno-r2
```

and was answered with the following error/output

```
VCFtools - v0.1.12b
(C) Adam Auton and Anthony Marcketta 2009

Parameters as interpreted:
        --vcf /long/path/to/snps.vcf
        --max-alleles 2
        --min-alleles 2
        --interchrom-geno-r2
        --out /long/path/to/out/snps.vcf

After filtering, kept 53 out of 53 Individuals
Outputting Interchromosomal Pairwise Genotype LD (bi-allelic only)
Error: Require phased haplotypes for r^2 calculation (use --phased)
```

I was under the impression from the docs that these options (`--geno-r2` and `--interchrom-geno-r2`) only require phased data for `D` and `D'` metrics:

> `--geno-r2`
>
> Calculates the squared correlation coefficient between genotypes encoded as 0, 1 and 2 to represent the number of non-reference alleles in each individual. This is the same as the LD measure reported by PLINK. The D and D' statistics are only available for phased genotypes. The output file has the suffix ".geno.ld".

Can anyone spot what is going wrong for me or am I confused?

Thanks,

Gus

2.1.1.1.2 [RESPONSE] Email to vcftools-help

- said its a bug and they will fix

2.1.1.2 Attempt 2 [FAILED: ran out of space]

I installed vcftools_0.1.12a and it began without complaint.

- -INPUT- -

```
SNP_DIR="/home2/wd238/data/genomes/glossina_fuscipes/annotations/SNPs"
VCF="${SNP_DIR}/tsetseFINAL_14Oct2014_f2_53.recode.renamed_scaffolds.maf0_05.vcf"
OUT_PREFIX="${SNP_DIR}/vcftools_out/tsetseFINAL_14Oct2014_f2_53.recode.renamed_scaffolds

mkdir -p ${SNP_DIR}/vcftools_out/
```

```
module load vcftools/0.1.12a
vcftools --vcf $VCF  --out $OUT_PREFIX --interchrom-geno-r2
```

- -OUTPUT- -

- Ran out of disk space.

2.2 2015-03-11 (Wednesday)

2.2.1 Calculate interchromosomal LD with `vcftools`

2.2.1.1 Attempt 3 [?]

- attempting to use `fastscratch` to allow for extra space.

- -INPUT- -

```
FAST_SCRATCH=/fastscratch/wd238
SNP_DIR="/home2/wd238/data/genomes/glossina_fuscipes/annotations/SNPs"
VCF="${SNP_DIR}/tsetseFINAL_14Oct2014_f2_53.recode.renamed_scaffolds.maf0_05.vcf"
OUT_PREFIX="${FAST_SCRATCH}/vcftools_out/tsetseFINAL_14Oct2014_f2_53.recode.renamed_scaf

mkdir -p ${FAST_SCRATCH}/vcftools_out/

module load vcftools/0.1.12a
vcftools --vcf $VCF  --out $OUT_PREFIX --interchrom-geno-r2
```

3 Linkage disequilibrium thresholds for SNP-pairs

3.1 General

3.1.1 2015-03-10 (Tuesday) [Status]

- Decided its best to use the Beta distribution on data binned by distance and scaled thusly:

$$((x_i - 0.5) \cdot 0.999) + 0.5)$$

- So far the MAP estimation is coming out VERY close to the MCMC results, so I think I will simply use that since it is **MUCH** faster.
- [ ] does multiple testing correction need to be done?
  - I am pretty sure it does
- p-values will be obtained for each r^2 as: $1 - \mathrm{CDF}(x_i)$
- see 2015-02-27_overview_of_LD_work_in_Gff.ipynb for extra info.

3.2 Thresholds by binning: notebook

- notebook file: 2015-03-12_LD_thresholds_via_binning.ipynb
- script version: 2015-03-12_LD_thresholds_via_binning.py

3.2.1 2015-03-13 (Friday)

- got the whole data set to run
  - those bins which fail MAP go on to run MCMC
  - had to re-write a bit to get the model object to save the MCMC runner so that we can look at the traces to asses convergence
- running as script in IPython to view.
- SUCCESS. Finally.
- saved resulting table in pickle: ddrad58/ld_thresholds/post_MAP_calc.plk
- use above to avoid re calculating the MAPs that take HOURS.
- started new ipython notebook file for results analysis: 2015-03-13_LD_thresholds_via_binning_RESULTS.ipy

3.3 Investigate LD bin-data pattern

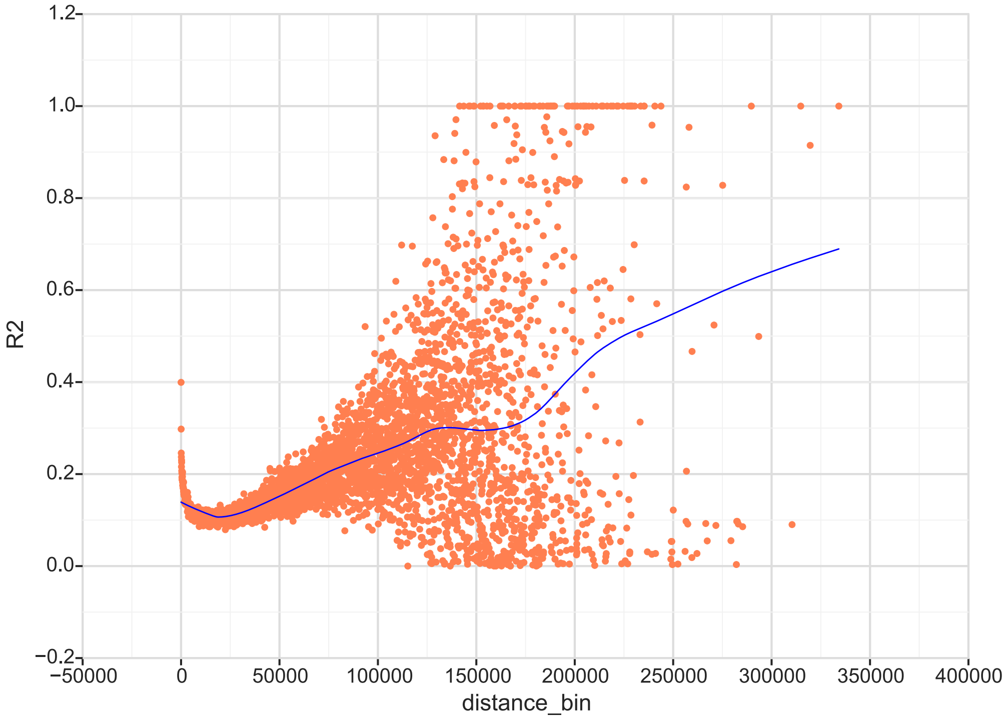

Figure 1: Distance vs r^2 overall

3.3.1 Bin-data membership quantity

Is the reason for the bizarre data shape due to loss of signal to noise as shorter contigs are eliminated from data pool?

3.3.2 Bin-data pattern of individual populations

4 Dating the North/South population split

4.1 Converting the BAMS to NEXSUS for BEAST

- using PGDSpider2 to convert to NEXUS

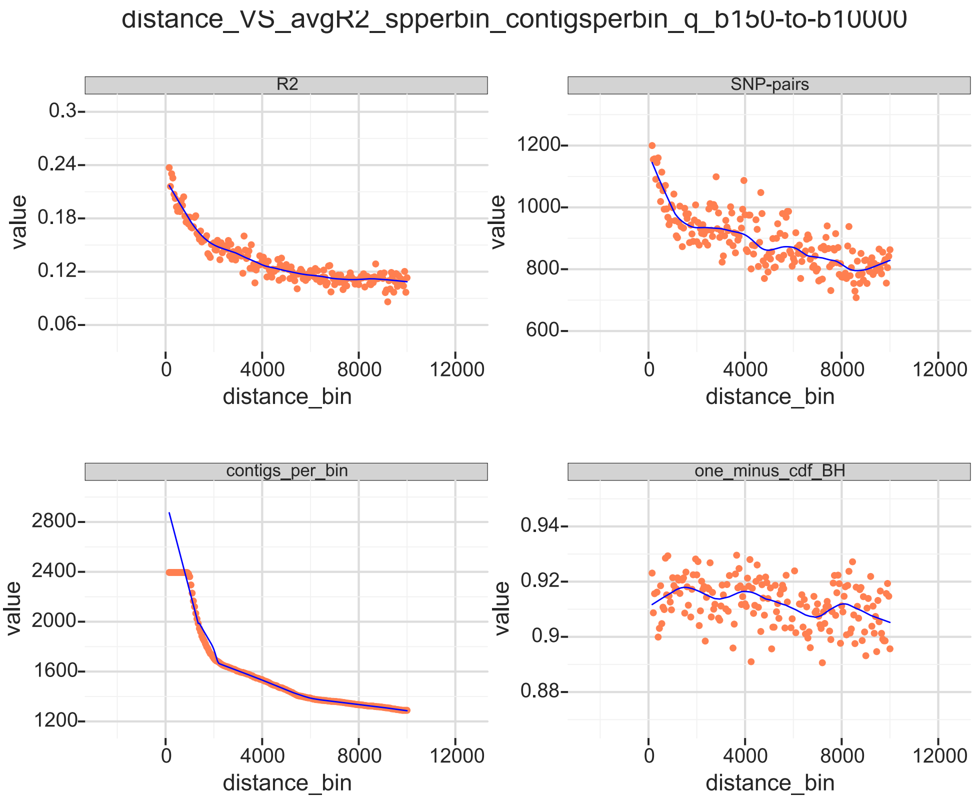

Figure 2: Distance vs avg r^2, contigs and q for bins 150-10000

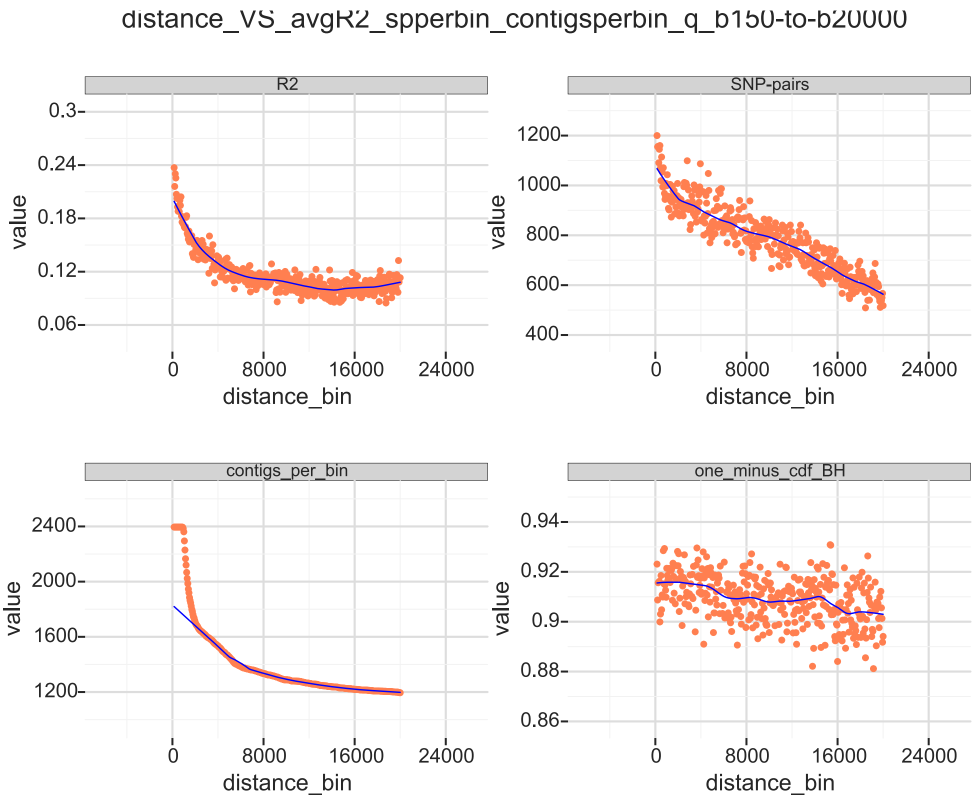

Figure 3: Distance vs avg r^2, contigs and q for bins 150-20000

- BAM location: /scratch/ag674/sample_mappedSC
- SPID file: bam_to_nex_for_BEAST.spid
- BAMS to use:
  - find /scratch/ag674/sample_mappedSC -name * | grep -P "\d\.sorted" > $HOME/data/projects/ddrad58/PGDSpider_files/bam_to_nex_for_BEAST/bam_to_nex_fo
  - bam_to_nex_for_BEAST.bam_list.txt
- ref for bam: Glossina-fuscipes-IAEA_SCAFFOLDS_GfusI1.fa

4.1.1 2015-03-11 (Wednesday)

- stymied by permissions issues with the bams.
- see tomorrow

4.1.2 2015-03-12 (Thursday)

4.1.2.1 Attempt 1 [FAILED: write permissions]

```
module load PGDSpider/2.0.8.0 samtools-bcftools-htslib/1.0

java -Xmx2048m -Xms512m -jar /home2/wd238/.local/easybuild/software/PGDSpider/2.0.8.0/PG
```

NOTES:

- PGDSpider seems to write a bunch of temporary files in the same dir as the inputfile.
- this breaks because I only have READ access to the data dir
- proceeding with copying the BAMs to a place I have write access to and trying again

4.1.2.2 Attempt 2 [FAILED: memory limit]

```
$ java -Xmx2048m -Xms512m -jar /home2/wd238/.local/easybuild/software/PGDSpider/2.0.8.0/

-[  output  ]-
INFO  16:27:47 - load PGDSpider configuration from: /home2/wd238/.local/easybuild/softwa
initialize convert process...
read input file...
INFO  16:28:04 - Run samtools/bcftools...
INFO  16:28:33 - [bam_sort_core] merging from 3 files...
ERROR 16:30:24 - not enough memory. To increase the allowed memory see help.
read input file done.
```

```
write output file...
write output file done.
```

NOTES:

- PGDSpider ran out of mem.
- I am going to bump up the mem and try again.

4.1.2.3 Attempt 3 [FAILED: reference file issue]

```
$ java -Xmx16384m -Xms16000m -jar /home2/wd238/.local/easybuild/software/PGDSpider/2.0.8
```

```
-[  output   ]-
INFO  17:23:52 - load PGDSpider configuration from: /home2/wd238/.local/easybuild/softwa
initialize convert process...
read input file...
INFO  17:24:16 - Run samtools/bcftools...
INFO  17:24:51 - [bam_sort_core] merging from 3 files...
INFO  17:26:38 - ...done
ERROR 17:29:37 - reference file does not contain *!
read input file done.
write output file...
write output file done.
```

NOTES:

- PGDSpider ran out of mem.
- I am going to bump up the mem and try again.

4.1.3 2015-03-13 (Friday)

- ABANDONING THIS AND LETTING ARIS TRY TO START FROM SCRATCH via PYRAD.
- thank GAWD.

5 Meeting

- Introduce Joshua and suggest a meeting