

Daily Records

Caccone PostDoc

Gus Dunn

January, 2015

Contents

1	2015-01-02 (Friday)	5
1.1	Writing Methods	5
1.2	Linkage Analysis	5
1.2.1	PLINK	5
1.2.2	Plot PLINK results	6
1.3	TODO for Gisella	6
2	2015-01-03 (Saturday)	6
2.1	Linkage Analysis	6
2.1.1	Plot PLINK results	6
3	2015-01-05 (Monday)	7
3.1	Linkage Analysis	7
3.1.1	Plot PLINK results	7
3.1.2	Re-Filter original VCF	7
3.1.3	PLINK - rerun	8
3.2	Recover dead positives	8
3.2.1	Dissections	8
4	2015-01-06 (Tuesday)	9
4.1	Linkage Analysis	9
4.1.1	PLINK --make-bed	9
4.1.2	PLINK --blocks	10
4.1.3	Plot PLINK results	11

4.2	Recover dead positives	11
4.2.1	Dissections	11
4.3	Bonizzoni <i>et al</i> : Insecticide Resistance	11
5	2015-01-07 (Wednesday)	11
5.1	Bonizzoni <i>et al</i> : Insecticide Resistance	11
5.2	Meeting with Serap Aksoy	12
5.2.1	Discussed	12
5.2.2	Action items	12
5.3	Meeting with Andrea	13
5.3.1	Discussed	13
5.4	Recover dead positives	13
6	2015-01-08 (Thursday)	13
6.1	Admin stuff	13
6.1.1	Serap's postdoc advertisement	13
6.1.2	Lab meetings	13
6.1.3	Current pc_labs list	14
6.2	Recover dead positives	15
6.2.1	Meeting with Kirstin	15
7	2015-01-09 (Friday)	15
7.1	Sarah Licensing Exam	15
7.2	New project brainstorming	15
7.3	PLINK: Fst	15
8	2015-01-10 (Saturday)	16
8.1	PLINK: Fst	16
8.2	Install EggLib	16
8.3	Install Bio++ (bpp)	16
9	2015-01-11 (Sunday)	16
9.1	Install EggLib	16
9.2	Install Bio++ (bpp)	17
10	2015-01-12 (Monday)	17
10.1	Install EggLib	17
10.2	Install Bio++ (bpp)	17
11	2015-01-13 (Tuesday)	17
11.1	Recover dead positives	17

11.2	Maps stuff	18
11.3	TsetseSampleDB	18
11.4	Helping Aris	18
12	2015-01-15 (Thursday)	18
12.1	Science Fair	18
12.2	Meeting about Kenya Tsetse	18
13	2015-01-16 (Friday)	19
13.1	Collection Spreadsheet review	19
13.1.1	Email explanation	19
13.2	Updating maps: current trap locations	19
13.2.1	spartan dev: GPS stuff	19
13.3	Phone for Dan	20
14	2015-01-18 (Sunday)	20
14.1	Sarah is sick	20
14.2	Updating maps: current trap locations	20
14.2.1	spartan dev: GPS stuff	20
15	2015-01-19 (Monday)	20
15.1	Andrea: quick chat	20
15.2	Updating maps: current trap locations	21
15.2.1	spartan dev: GPS stuff	21
15.3	Install Stacks	21
15.4	Collection Spreadsheet review	21
16	2015-01-20 (Tuesday)	21
16.1	Sarah still sick	21
16.2	Meeting with Gisella and Andrea: ddRAD paper	22
16.3	ddRAD todos	22
16.4	Install Stacks	22
16.4.1	Prerequisites	22
16.4.2	Stacks	23
16.5	Install SparseHash	23
17	2015-01-21 (Wednesday)	24
17.1	Family still VERY sick	24
17.2	Manual install of Samtools/htslib	24
17.2.1	htslib (built with Arch ABS on jupiter)	24

17.3	EasyBuild installs	25
17.3.1	EasyBuild 1.16.1	25
17.3.2	Install samtools-1.1	25
18	2015-01-22 (Thursday)	25
18.1	Collection documentation files	25
18.2	EasyBuild installs	26
18.2.1	Install GSL-1.16	26
18.2.2	Install sparsehash-2.0.2	26
18.2.3	Install Stacks-1.03	26
18.2.4	Install Stacks-1.24	27
18.2.5	Install zlib-1.2.8	28
19	2015-01-23 (Friday)	29
19.1	Meeting with Alexis	29
19.2	EasyBuild installs	29
19.3	Doc appointment	29
20	2015-01-24 (Saturday)	29
20.1	Sarah sprained/broke? her ankle	29
21	2015-01-25 (Sunday)	30
21.1	EasyBuild installs	30
22	2015-01-26 (Monday)	31
22.1	Carl Zimmer Writing Workshop	31
23	2015-01-27 (Tuesday)	31
23.1	SNOW-pocalypse	31
24	2015-01-28 (Wednesday)	31
24.1	EasyBuild installs	31
24.1.1	Stacks	31
24.2	tmux	32
24.3	ddRAD stuff	32
25	2015-01-29 (Thursday)	32
25.1	ddRAD stuff	32
25.2	Robert's stuff	33
25.3	Rob H Jobs	34
25.3.1	Iowa samples	34

25.3.2	<i>G. f. fuscipes</i> samples 2014	34
25.4	Updating maps: current trap locations	34
25.4.1	spartan dev: GPS stuff	34

1 2015-01-02 (Friday)

1.1 Writing Methods

- [x] Functional Annotations
 - done-ish at 2015-01-02 08:58
- [] Linkage
 - **STILL NEED TO DO THIS ANALYSIS...**

1.2 Linkage Analysis

- Still not understanding how you can calculate LD without phased data but it **seems** like many programs claim to...

1.2.1 PLINK

- [v1.90 user manual: LD section](#)
- [x] create the files needed from the master VCF file (tsetseF1-NAL_14Oct2014_f2_53.recode.renamed_scaffolds.vcf)
 - looks like `plink` now reads VCF (v1.90): will try this first.
- [x] split data into smaller pieces to parallelize the `plink` analysis.
 - looks like the `--parallel` flag will allow `plink` to take care of this.
- [x] start `run(s)` on louise.
- [] *try running a `--blocks` PLINK analysis for haplotype blocks to see if its useful?*

plink commands run and kept:

```

plink --vcf tsetseFINAL_14Oct2014_f2_53.recode.renamed_scaffolds.vcf --allow-extra-chr \
  --r gz with-freqs \
  --out plink_out/tsetseFINAL_14Oct2014_f2_53.recode.renamed_scaffolds.vcf\
  /ld/r_none_freqs

plink --vcf tsetseFINAL_14Oct2014_f2_53.recode.renamed_scaffolds.vcf --allow-extra-chr \
  --r gz in-phase with-freqs \
  --out plink_out/tsetseFINAL_14Oct2014_f2_53.recode.renamed_scaffolds.vcf\
  /ld/r_none_phase_freqs

plink --vcf tsetseFINAL_14Oct2014_f2_53.recode.renamed_scaffolds.vcf --allow-extra-chr \
  --r triangle gz \
  --out plink_out/tsetseFINAL_14Oct2014_f2_53.recode.renamed_scaffolds.vcf\
  /ld/r_tri

```

1.2.2 Plot PLINK results

- [x] create ipython notebook file
 - [YALE/ddrad58/2015-01-02_Plot_PLINK_results.ipynb](#)
 - [2015-01-13]: [YALE/ddrad58/2015-01-05_Plot_PLINK_results.ipynb](#)
- [] write code to plot

1.3 TODO for Gisella

- [] re-read grant bit about bioinformatics and think about how to use Hongyu Zhao.
-

2 2015-01-03 (Saturday)

2.1 Linkage Analysis

2.1.1 Plot PLINK results

- [x] **WHAT** should be plotted?
 - [x] what *exactly* is the r value telling us?

- * [x] does it already take into account the distance?
 - according to [Wikipedia](#), r is simply the correlation coefficient between pairs of loci:

$$r = \frac{D}{\sqrt{p_1 p_2 q_1 q_2}}$$

– It seems like plotting $\frac{r}{l_a - l_b}$ (r divided by distance) **is** warranted where:

- * l_a is location of SNP_a
- * l_b is location of SNP_b

- [] write code to plot

2.1.1.1 Questions for Andrea

- Some MAFs are zero which causes the LD (r) to fail. [link](#)
-

3 2015-01-05 (Monday)

3.1 Linkage Analysis

3.1.1 Plot PLINK results

- met with Andrea after showing her what I had and specifically the $\text{MAF} = 0$ for about half the scaffold0 comparisons.
 - (see meeting notes for more details)

3.1.2 Re-Filter original VCF

- the incorrect (`--min-allele/--max-allele`) filter was used to generate: `tsetseFINAL_14Oct2014_f2_53.recode.vcf`
- the correct filter is `--maf`.
- I am doing it myself with $\text{MAF} = 0.05$ (see below).
- **retained only 47.7% sites**
- **I will be re-running my PLINK LD analysis just in case.**

wd238 at compute-1-4 in ~GENOMES/glossina_fuscipes/annotations/SNPs (py278)

```
$ vcftools \  
  --vcf tsetseFINAL_140ct2014_f2_53.recode.renamed_scaffolds.vcf \  
  --maf 0.05 \  
  --out tsetseFINAL_140ct2014_f2_53.recode.renamed_scaffolds.maf0_05 \  
  --recode
```

VCFTools - v0.1.12b

(C) Adam Auton and Anthony Marcketta 2009

Parameters as interpreted:

```
--vcf tsetseFINAL_140ct2014_f2_53.recode.renamed_scaffolds.vcf  
--maf 0.05  
--out tsetseFINAL_140ct2014_f2_53.recode.renamed_scaffolds.maf0_05  
--recode
```

After filtering, kept 53 out of 53 Individuals

Outputting VCF file...

After filtering, kept 73297 out of a possible 153650 Sites

Run Time = 21.00 seconds

3.1.3 PLINK - rerun

```
plink --vcf tsetseFINAL_140ct2014_f2_53.recode.renamed_scaffolds.maf0_05.vcf \  
--allow-extra-chr \  
--r gz with-freqs dprime \  
--out plink_out/tsetseFINAL_140ct2014_f2_53.recode.renamed_scaffolds.maf0_05.vcf/ld/r_none_
```

3.2 Recover dead positives

3.2.1 Dissections

- Prepped for dissections and pre-filled the worksheets
- but we are out of the 1.5 ml tubes that I bought for this and I will have to go get some more tomorrow morning.

4 2015-01-06 (Tuesday)

4.1 Linkage Analysis

- emailed Dan about looking over the results.
- Should probably run them by Jeff if he has time too.

4.1.1 PLINK --make-bed

```
wd238 at compute-1-4 in ~GENOMES/glossina_fuscipes/annotations/SNPs (py278)
$ plink --vcf tsetseFINAL_14Oct2014_f2_53.recode.renamed_scaffolds.maf0_05.vcf \
> --allow-extra-chr \
> --maf 0.05 \
> --make-bed \
> --out tsetseFINAL_14Oct2014_f2_53.recode.renamed_scaffolds.maf0_05.plink

PLINK v1.90b2o 64-bit (25 Nov 2014)      https://www.cog-genomics.org/plink2
(C) 2005-2014 Shaun Purcell, Christopher Chang  GNU General Public License v3
Logging to tsetseFINAL_14Oct2014_f2_53.recode.renamed_scaffolds.maf0_05.plink.log.
48251 MB RAM detected; reserving 24125 MB for main workspace.
--vcf: 73k variants complete.
tsetseFINAL_14Oct2014_f2_53.recode.renamed_scaffolds.maf0_05.plink-temporary.bed
+
tsetseFINAL_14Oct2014_f2_53.recode.renamed_scaffolds.maf0_05.plink-temporary.bim
+
tsetseFINAL_14Oct2014_f2_53.recode.renamed_scaffolds.maf0_05.plink-temporary.fam
written.
73297 variants loaded from .bim file.
53 people (0 males, 0 females, 53 ambiguous) loaded from .fam.
Ambiguous sex IDs written to
tsetseFINAL_14Oct2014_f2_53.recode.renamed_scaffolds.maf0_05.plink.nosex .
Using 1 thread (no multithreaded calculations invoked).
Before main variant filters, 53 founders and 0 nonfounders present.
Calculating allele frequencies... done.
Total genotyping rate is 0.965098.
0 variants removed due to MAF threshold(s) (--maf/--max-maf).
73297 variants and 53 people pass filters and QC.
Note: No phenotypes present.
--make-bed to
tsetseFINAL_14Oct2014_f2_53.recode.renamed_scaffolds.maf0_05.plink.bed +
```

```
tsetseFINAL_14Oct2014_f2_53.recode.renamed_scaffolds.maf0_05.plink.bim +
tsetseFINAL_14Oct2014_f2_53.recode.renamed_scaffolds.maf0_05.plink.fam ...
done.
```

4.1.2 PLINK --blocks

- running with --blocks option to look at estimated haplotype blocks

```
wd238 at compute-1-4 in ~GENOMES/glossina_fuscipes/annotations/SNPs (py278)
$ plink --vcf tsetseFINAL_14Oct2014_f2_53.recode.renamed_scaffolds.maf0_05.vcf \
> --allow-extra-chr \
> --blocks no-pheno-req no-small-max-span \
> --out plink_out/tsetseFINAL_14Oct2014_f2_53.recode.renamed_scaffolds.maf0_05.vcf\
  /ld/blocks_nophenoreq_nosmallmaxspan

PLINK v1.90b2o 64-bit (25 Nov 2014)      https://www.cog-genomics.org/plink2
(C) 2005-2014 Shaun Purcell, Christopher Chang  GNU General Public License v3
Logging to plink_out/tsetseFINAL_14Oct2014_f2_53.recode.renamed_scaffolds.maf0_05.vcf\
  /ld/blocks_nophenoreq_nosmallmaxspan.log.
48251 MB RAM detected; reserving 24125 MB for main workspace.
--vcf: 73k variants complete.

...

73297 variants loaded from .bim file.
53 people (0 males, 0 females, 53 ambiguous) loaded from .fam.
Ambiguous sex IDs written to
plink_out/tsetseFINAL_14Oct2014_f2_53.recode.renamed_scaffolds.maf0_05.vcf\
  /ld/blocks_nophenoreq_nosmallmaxspan.nosex
.
Using 1 thread (no multithreaded calculations invoked).
Before main variant filters, 53 founders and 0 nonfounders present.
Calculating allele frequencies... done.
Total genotyping rate is 0.965098.
73297 variants and 53 people pass filters and QC.
Note: No phenotypes present.
--blocks: 8040 haploblocks written to
plink_out/tsetseFINAL_14Oct2014_f2_53.recode.renamed_scaffolds.maf0_05.vcf\
  /ld/blocks_nophenoreq_nosmallmaxspan.blocks
.
Extra block details written to
```

```
plink_out/tsetseFINAL_14Oct2014_f2_53.recode.renamed_scaffolds.maf0_05.vcf\  
  /ld/blocks_nophenoreq_nosmallmaxspan.blocks.det
```

.
Longest span: 199.985kb.

4.1.3 Plot PLINK results

- cleaned up a few things
- added residual plots following the regplots

4.2 Recover dead positives

- need to meet with Kirsten (emailed her to schedule a time)
 - subject: *"Short meeting to talk about the dead positives screen?"*

4.2.1 Dissections

- getting more:
 - tubes
 - PBS
 - Pens
- need to get receipt(s) from Kirsten regarding the dissection dish order
 - emailed with subject: *"did you ever send me the receipt(s) for the stuff you ordered for the dissections over the internet?"*

4.3 Bonizzoni *et al*: Insecticide Resistance

- I am SUPER late on this!
-

5 2015-01-07 (Wednesday)

5.1 Bonizzoni *et al*: Insecticide Resistance

Status: COMPLETE

- finished reviewing the main text
- emailed it to her
- will not be going over the legends or figs

5.2 Meeting with Serap Aksoy

Time: 10:00 AM to 11:30AM

5.2.1 Discussed

- how to log Iowa tsetse samples
- student to do much of the logging after a spreadsheet is devised
- location of the other RNA midguts
 - she said she thought they didn't get any but then thought she remembered that Brian tried extracting RNA from at least a few infected midguts with no success.
 - ****She said she needs the carcasses of the infected flies too which I did not remember (I need to bring this up with Gisella bc this is a major reduction of our expected infected material. . .).**
 - * *this doesn't really make sense anyway since i don't think we preserved the bodies for RNA.*
- having me send an ad or two for a postdoc position for her lab to my friends

5.2.2 Action items

Status: IN PROGRESS

- [x] create simple excel sheet to track Iowa samples
 - [x] email sheet to Brian and Serap
 - * subject: *Spreadsheet to record Iowa sample materials*
 - **NOTE:** sheet is a google sheet named [Iowa_tsetse_material_inventory](#)
- [x] locate extra RNA midguts in our freezers
 - [x] email Aksoy, Brian, Michelle to schedule pickup
 - * subject: *Many more midguts and heads for RNA*
- [] send feelers and ads to friends about postdoc position in her lab
- [] contact Gisella about Serap wanting the carcasses. . .

5.3 Meeting with Andrea

5.3.1 Discussed

- problem re-running the figure generation R script
- couldn't open the PNG writer bc no X11 on the cluster and `ssh -Y` wasn't working even though it did last time. . .
- I added some code to the R script to specify `'png(type="cairo")`
- waiting to hear the outcome
 - *program ran but output was not what was expected: many more graphs than last time*
 - I expect user error

5.4 Recover dead positives

- Meeting with Kirstin tomorrow at 1 or 2 PM
-

6 2015-01-08 (Thursday)

6.1 Admin stuff

6.1.1 Serap's postdoc advertisement

- posted to facebook

6.1.2 Lab meetings

- email from Jeff:
 - subject: *lab meetings*
 - body:

Gus,

Any lab meetings set up? We should grab Aris and Yiota who might give a joint one. Then also the new Anthropology guy, Eduardo Fernandez-Duque.

Jeff

- doodle poll:
 - [link](#) sent to pc_labs
 - not certain if Maggie is on that list yet
 - emailed Carol for her email in case not
 - * She **WAS** on the list when I sent the link
 - * Carol replied with current list: recorded below

6.1.3 Current pc_labs List

#separator-CACCONE#

adalgisa.caccone@yale.edu
carol.mariani@yale.edu
danielle.edwards@yale.edu
nphavill@fs.fed.us
jrichardson@providence.edu
giovanna.carpi@yale.edu
katharine.walter@yale.edu
gus.dunn@yale.edu

#separator-POWELL#

jeffrey.powell@yale.edu
kirstin.dion@yale.edu
andrea.gloria-soria@yale.edu
b.evans@yale.edu
joshua.richardson@yale.edu

#separator-TEMP-ROTATION-UNDERGRAD#

christian.hernandez@yale.edu
elaine.guevara@yale.edu
andres.valdivieso@yale.edu
mkcorley@gmail.com
alexis.halyard@yale.edu
pkotsakiozi@hotmail.com
aristeidis.parmakelis@yale.edu

6.2 Recover dead positives

6.2.1 Meeting with Kirstin

- see: [Kirsten-2015-01-08.md](#)
 - sent above for confirmation or amendment to Kirstin
 - She approves
-

7 2015-01-09 (Friday)

7.1 Sarah Licensing Exam

- I was taking care of the kids all morning

7.2 New project brainstorming

7.3 PLINK: Fst

- defining population ID file for:
 - tsetseFINAL_14Oct2014_f2_53.recode.renamed_scaffolds.maf0_05.vcf
 - tsetseFINAL_14Oct2014_f2_53.recode.renamed_scaffolds.maf0_05.vcf.popdef

```
plink --bfile tsetseFINAL_14Oct2014_f2_53.recode.renamed_scaffolds.maf0_05.plink \
--allow-extra-chr \
--within tsetseFINAL_14Oct2014_f2_53.recode.renamed_scaffolds.maf0_05.vcf.popdef \
--fst \
--out plink_out/tsetseFINAL_14Oct2014_f2_53.recode.renamed_scaffolds.maf0_05.vcf/fst/out
```

- This keeps giving me errors:

```
Warning: No samples named in --within file remain in the current analysis.
Using 1 thread (no multithreaded calculations invoked).
Before main variant filters, 53 founders and 0 nonfounders present.
Calculating allele frequencies... done.
Total genotyping rate is 0.965098.
73297 variants and 53 people pass filters and QC.
Note: No phenotypes present.
Error: --fst requires at least two nonempty clusters.
```

8 2015-01-10 (Saturday)

8.1 PLINK: Fst

- Still getting errors
- looks like its bc there is no sex info attached to the samples
- looking for other was to do this stuff: found EggLib-py

8.2 Install EggLib

- installation needs bio++
- see tomorrow

8.3 Install Bio++ (bpp)

- install script ([bpp-setup.sh](#)) obtained from [bio++ website](#).
 - altered install script to fit system (louise) and renamed [install_bpp_2.2.0.sh](#).
- see tomorrow

9 2015-01-11 (Sunday)

9.1 Install EggLib

- finished bio++install
- installing other things (actually may just module-ize versions already installed on louise:
 - [x] gsl (already installed)
 - [x] clustalw (linked to louise/~MAIN_APPS/clustalw/clustalw-2.0.12-linux-i686-libc
 - [x] muscle
 - [x] paml

- [x] phymL
- [x] primer3
- [x] phylip

9.2 Install Bio++ (bpp)

- had to ammend the install script to include the bpp install location in \$PATH so it can use itself to install/build parts of itself
 - as far as I can tell the only things that fail to install now are the GUI-based stuff that needs Qt. As I am using this on th cluster I dont need/want these so I am going to proceed as if this succeeded.
-

10 2015-01-12 (Monday)

10.1 Install EggLib

- finished installing external helper programs:
 - clustalw
 - phylip
- made a modules file for the whole group: egglib_helpers

10.2 Install Bio++ (bpp)

- forgot to write a modules file for this.
 - doing it now
-

11 2015-01-13 (Tuesday)

11.1 Recover dead positives

- meeting
- see meeting notes [gisella_kirsten_2015-01-14.md](#)

11.2 Maps stuff

- Need to update our Northern Uganda map with the latest Village location data
- `Spartan.utils.maps.gps`
 - coding the functions to take all trap GPS coords for a village and return one GPS coord set for each that represents the central tendency for simplified plotting

11.3 TsetseSampleDB

- adding names/village codes to the list of village-to-code map ([village_id_map.csv](#))

11.4 Helping Aris

- getting `mpirun mrbayes` to run on grace
 - took ~ 1 hour.
-

12 2015-01-15 (Thursday)

12.1 Science Fair

- 8 to 12:30

12.2 Meeting about Kenya Tsetse

Members:

- Gisella
- Serap
- Michelle
- Brian
- Gus

Notes (bad bc they were taken with my phone):

- Gpd samples
 - DNA samples
 - What column data will be needed for this Gpd Kenya collection
 - Re-circulate collection template excel and protocol etc
-

13 2015-01-16 (Friday)

13.1 Collection Spreadsheet review

- point is to make sure we can use this for the Kenya “simpler”
- after speaking with Gisella, I am going to add a few of the “simpler” column heading to the normal collection spreadsheet and write a bunch of notes explaining that not everything needs to be filled in for everything.
- collection spreadsheet: [Example_collection_template_kenya.xls](#)
 - **status:** finished
- summary spreadsheet: [Example_summary_template_kenya.xls](#)
 - **status:** in progress

13.1.1 Email explanation

Subject:

Body:

13.2 Updating maps: current trap locations

- created new spreadsheet: [collection_meta_data/meta_data.ods](#) to store current state of stuff like the village-to-symbol map, etc.
- [] collecting trap GPS data to file: [TsetseCheckout/data/village_id_map.csv](#)
- [] collecting all village-to-symbol maps that I have to [meta_data.ods](#)

13.2.1 spartan dev: GPS stuff

- pycharm and ipython

13.3 Phone for Dan

I emailed Dan the following:

Subject: Phone call for you

Body: Just fielded a call for you from Karan(Karen?) Peart from the Yale Public affairs and communications office.

She would like you to call her back at your earliest convenience (432-1326).

Gus

14 2015-01-18 (Sunday)

14.1 Sarah is sick

- short day: 10am to 1:45pm

14.2 Updating maps: current trap locations

- ipython: [2015-01-16_convert_fall2014_trap_gps_village_names.ipynb](#)

14.2.1 spartan dev: GPS stuff

- working on teaching GPSCoordTree how to get mean coordinates
-

15 2015-01-19 (Monday)

15.1 Andrea: quick chat

- wants to re-run the ddRAD pipeline since I(we) found some issues with at least one of the command line runs' options.
- I agree

- [] #todo: I am installing [Stacks 1.24](#) for her on louise and trying to set up the web-based analysis part
- should not change TOO much about the results and will end up being MUCH more replicatable due to the use of publicly accessible data from vectorbase.
- we can continue to work with the current data until the new set is done and just adjust the work to accommodate the new stuff at the end.

15.2 Updating maps: current trap locations

15.2.1 spartan dev: GPS stuff

- working on teaching GPSCoordTree how to get mean coordinates

15.3 Install Stacks

- [installation guide](#)

15.4 Collection Spreadsheet review

- adding explanation text to the [fly_collection_basic.md](#) document.
 - **STATUS:**
 - having issues getting validation and drop-down lists to carry over into “empty” rows
 - plan to fix it by copying a template row into like 1000 rows
 - still need to execute the above tomorrow bc computer is acting a fool and I have to go home to sick Sarah and Liam.
-

16 2015-01-20 (Tuesday)

16.1 Sarah still sick

- stayed home to help with particularly crazy morning

16.2 Meeting with Gisella and Andrea: ddRAD paper

- Met at 10:30 AM
- summarized in [gisella_andrea_2015-01-20.md](#).

16.3 ddRAD todos

- [in progress] read Mark's trypan paper for the LD stuff
- [] email Mark to have a short sit-down to go over my results and ask about his work

16.4 Install Stacks

- [installation guide](#)

16.4.1 Prerequisites

Visualization:

- [] DB2 Pear Module: <http://pear.php.net/package/MDB2/>
- [] MDB2 MySQL driver: http://pear.php.net/package/MDB2_Driver_mysql/
- [] PHP
- [] MySQL
- [] Perl DBI module installed with the MySQL driver [CPAN/dist/DBD-mysql/](http://cpan.org/dist/DBD-mysql/)

Spreadsheet export:

- [] Perl module: [Spreadsheet::WriteExcel](#)

Performance improvement:

- [X] samtools for reading BAM files (already installed)
- [] Google's SparseHash class to lower memory usage <http://code.google.com/p/sparsehash/>

16.4.2 Stacks

INSTALLATION PROBLEMS NOTES:

- Having issues getting the samtools includes and libs configured for the `./configure` command.
- plan to build on jupiter using the ARCH ABS and copy the include/lib directories to louise

download location:

- status: *in progress*
- downloaded [stacks-1.24.tar.gz](#) to gus@louise/src.

install script:

- status: *in progress*
- [gus@louise/scripts/installs/install_XXXX](#)

module file:

- status: *in progress*
- [gus@louise/.local/environment-modules/Modules/3.2.10/my_modulefiles/XXXXX/XXXXX](#)

software root:

- status: *in progress*
- [gus@louise/home/gus/remote_mounts/louise/.local/easybuild/software/XXXXX/XXXXX](#)

16.5 Install SparseHash

download location:

- status: *complete*
- downloaded [sparsehash-2.0.2.tar.gz](#) to gus@louise/src.

install script:

- status: *written and run*
- [gus@louise/scripts/installs/install_sparsehash_2.0.2.sh](#)

module file:

- status: *complete but not tested*
- [gus@louisie/.local/environment-modules/Modules/3.2.10/my_modulefiles/sparsehash/2.0.2](#)

software root:

- status: *installed*
 - [gus@louisie/home/gus/remote_mounts/louisie/.local/easybuild/software/sparsehash/2.0.2](#)
-

17 2015-01-21 (Wednesday)

17.1 Family still VERY sick

- stayed home with Liam while Sarah took Clementine and herself to the doctor
- got to work at 12:30
- had to go home so Sarah could sleep bc Clem was not letting her
- went home at 1:30
- came back at 3:30
- home again at 6:00
- work again at 8:30

17.2 Manual install of Samtools/htslib

17.2.1 htslib (built with Arch ABS on jupiter)

ABANDONING THIS METHOD FOR NOW. TOO MANY PROBLEMS WITH INTEGRATING CERTAIN INCLUDE AND LIB DIRECTRIES WHEN COMPILING ACCROSS DEPENDENCIES. TRYING EASYBUILD AND ITS *TOOLCHAIN* PARADIGM FOR NOW.

ABS build:

```
cd /home/gus/remote_mounts/louisie/src/ABS/
tar -xf htslib.tar.gz
cd htslib
makepkg
```

Install script:

- [gus@louisie/scripts/installs/install_htslib_1.1.sh](#)

17.3 EasyBuild installs

17.3.1 EasyBuild 1.16.1

Install script:

- [gus@louisie/scripts/installs/install_easybuild_1.16.1.sh](#)

17.3.2 Install samtools-1.1

```
wd238 at compute-21-15 in ~ (py278)
```

```
$ md load EasyBuild/1.16.1
```

```
wd238 at compute-21-15 in ~ (py278)
```

```
$ eb SAMtools-1.1-goolf-1.4.10.eb --try-toolchain=goolf,1.4.10-no-OFED --robot
```

```
...
```

- this will be installing the whole toolchain and all samtools dependencies so it was executed in a screen.
- I am going home to sleep while this works [2015-01-21 21:33].
- emailing Andrea first to let her know its not gonna be ready when I told her.

STATUS: [2015-01-22 07:52]

- Build seems to have **SUCCEEDED**
-

18 2015-01-22 (Thursday)

18.1 Collection documentation files

- fixed/kludged the collection template to keep cell-notes and verification by including “dummy” entries up to 500.
 - collection template: [Example_collection_template_kenya.xlsx](#)
 - summary template: [Example_summary_template_kenya.xlsx](#)
- amended and committed v0.2.1 of [fly_collection_basic.md](#) to the repo with custom message:

“protocols/fly_collection_basic/fly_collection_basic.md committed at version: v0.2.1”

18.2 EasyBuild installs

18.2.1 Install GSL-1.16

```
wd238 at compute-21-15 in ~ (py278)
$ eb GSL-1.16-goolf-1.4.10.eb --try-toolchain=goolf,1.4.10-no-OFED --robot
```

STATUS:

```
== COMPLETED: Installation ended successfully
== Results of the build can be found in the log file \
    /home2/wd238/.local/easybuild/software/GSL/1.16-goolf-1.4.10-no-OFED/\
    easybuild/easybuild-GSL-1.16-20150122.075231.log
== Build succeeded for 1 out of 1
```

18.2.2 Install sparsehash-2.0.2

```
wd238 at compute-21-15 in ~ (py278)
eb google-sparsehash-2.0.2-goolf-1.4.10.eb --try-toolchain=goolf,1.4.10-no-OFED --robot
```

STATUS:

```
== COMPLETED: Installation ended successfully
== Results of the build can be found in the log file \
    /home2/wd238/.local/easybuild/software/google-sparsehash/ \
    2.0.2-goolf-1.4.10-no-OFED/easybuild/ \
    easybuild-google-sparsehash-2.0.2-20150122.080547.log
== Build succeeded for 1 out of 1
```

18.2.3 Install Stacks-1.03

NOTE: this is not for use per se but to generate the config and module files so that I can modify them for the latest version of Stacks and install *THAT* version.

```
wd238 at compute-21-15 in ~ (py278)
$ eb Stacks-1.03-goolf-1.4.10.eb --try-toolchain=goolf,1.4.10-no-OFED --robot
```

STATUS:

```
== COMPLETED: Installation ended successfully
== Results of the build can be found in the log file \
  /home2/wd238/.local/easybuild/software/Stacks/ \
  1.03-goolf-1.4.10-no-OFED/easybuild/ \
  easybuild-Stacks-1.03-20150122.081337.log
== Build succeeded for 1 out of 1
```

18.2.4 Install Stacks-1.24

easyconfig file:

- altered the one generated when building Stacks-1.03
 - [gus@louise/.local/easybuild/ebfiles_repo/Stacks/Stacks-1.03-goolf-1.4.10-no-OFED.eb](#)
- [gus@louise/scripts/installs/easybuild/Stacks-1.24-goolf-1.4.10-no-OFED.eb](#)

18.2.4.1 Attempt 01

```
wd238 at compute-21-15 in ~ (py278)
eb Stacks-1.24-goolf-1.4.10-no-OFED.eb --try-toolchain=goolf,1.4.10-no-OFED --robot
```

STATUS: FAILED

- couldn't find Stacks-1.24-goolf-1.4.10-no-OFED.eb
- basically expected.

18.2.4.2 Attempt 02

```
wd238 at compute-21-15 in ~ (py278)
eb $HOME/.local/easybuild/ebfiles_repo/Stacks/Stacks-1.24-goolf-1.4.10-no-OFED.eb \
  --try-toolchain=goolf,1.4.10-no-OFED --robot
```

STATUS: FAILED

- error log: [gus@louise/scripts/installs/easybuild/failure_logs/easybuild-Stacks-1.24-20150122.090021.pPmjV.log](#)
- looks like it cant find sparsehash for the linking
- will add samtools and sparsehash to the easyconfig file as dependencies and or build dependencies.

18.2.4.3 Attempt 03

- added the below to the easyconfig file:

```
– builddependencies = [('SAMtools', '1.1'), ('google-sparsehash',  
    '2.0.2')]
```

wd238 at compute-21-15 in ~ (py278)

```
eb $HOME/.local/easybuild/ebfiles_repo/Stacks/Stacks-1.24-goolf-1.4.10-no-OFED.eb \  
    --try-toolchain=goolf,1.4.10-no-OFED --robot
```

== temporary log file in case of crash /tmp/easybuild-OT7khn/easybuild-08eCl9.log

ERROR: EasyBuild crashed with an error \

```
(at easybuild/software/EasyBuild/1.16.1/lib/python2.7/site-packages/  
easybuild_framework-1.16.1-py2.7.egg/easybuild/tools/robot.py:232 in \  
resolve_dependencies): Irresolvable dependencies encountered: \  
SAMtools/1.1-goolf-1.4.10-no-OFED, google-sparsehash/2.0.2-goolf-1.4.10-no-OFED
```

STATUS: FAILED

- error log: [gus@louise/scripts/installs/easybuild/failure_logs/easybuild-08eCl9.log](#)

18.2.4.4 IRC session with author/devs

- one problem was that I dont need to keep using --try-toolchain=goolf,1.4.10-no-OFED since the local easyconfig (Stacks-1.24-goolf-1.4.10-no-OFED.eb) being passed to eb already defines the toolchain.
- the build still fails however

18.2.5 Install zlib-1.2.8

NOTE:

- this is because things seem to need it when building stacks-1.24?
- doesn't seem like this was the case?

wd238 at compute-21-15 in ~ (py278)

```
$ eb zlib-1.2.8-goolf-1.4.10.eb --try-toolchain=goolf,1.4.10-no-OFED --robot
```

STATUS: SUCCESSFUL

== COMPLETED: Installation ended successfully

== Results of the build can be found in the log file \

```
/home2/wd238/.local/easybuild/software/zlib/\
1.2.8-goolf-1.4.10-no-OFED/easybuild/\
easybuild-zlib-1.2.8-20150122.115448.log
== Build succeeded for 1 out of 1
```

19 2015-01-23 (Friday)

19.1 Meeting with Alexis

- 10:00 to 12:20
- talked about overall project and helped her come up with stuff to talk about for 2 minutes next Tuesday.
- Gisella was supposed to be here but double booked the time so will meet with Alexis individually.
- Gave Alexis my email and asked her to email me so that i would get hers
- so far [17:00] has not emailed me.

19.2 EasyBuild installs

- forked and cloned [easybuild-easyconfigs](#) git repo

19.3 Doc appointment

- left desk around 14:00 and got back around 15:20
-

20 2015-01-24 (Saturday)

20.1 Sarah sprained/broke? her ankle

- Sarah slipped while trying to shovel snow (?! WHY ?!)
 - had to go back home soon after arrival
-

21 2015-01-25 (Sunday)

21.1 EasyBuild installs

- adding new forked git-repo of easyconfigs to easybuild through environment variables in my [gus@louise/.zshrc](#).
- make check failing was caused by the easyconfig file setting runtest='check'.

– I removed this

- install still fails but seems to try to repeat itself and fails the SECOND TIME?

```
– ...
== building and installing Stacks/1.24-goolf-1.4.10-no-OFED...
== fetching files...
== creating build dir, resetting environment...
== unpacking...
== patching...
== preparing...
== configuring...
== building...
== testing...
== installing...
== creating build dir, resetting environment...
== unpacking...
== patching...
== preparing...
== configuring...
== building...
== FAILED: Installation ended unsuccessfully...
...
```

- checking the logs seems to suggest that the BAM include files are still not working

– testing by removing the reqs from the configure script

- * NO sparsehash and NO bam: **SUCCEEDS**
- * YES sparsehash and NO bam: **SUCCEEDS**
- * NO sparsehash and YES bam: **FAILS**

- **TO TRY TOMORROW:**

- ‘clone’ environment from one of the “test_reports” in [failure_logs](#) and try to run `./configure; make; make install`; manually.
-

22 2015-01-26 (Monday)

22.1 Carl Zimmer Writing Workshop

- 10:00 to 12:00
 - Notes made in notebook to be transferred here when I have time (blizzard approaching)
-

23 2015-01-27 (Tuesday)

23.1 SNOW-pocalypse

- was told to stay home by Yale
 - came in by mistake but left soon after realizing
-

24 2015-01-28 (Wednesday)

24.1 EasyBuild installs

24.1.1 Stacks

24.1.1.1 Stacks no BAM

- testing to make sure it works
- abandoning this bc we decided that we dont need to run this step over right now

24.1.1.2 Stacks yes BAM

- still not building correctly

24.2 tmux

- starting point [easyconfig](#)
- abandoning this for now
- simply not crucial

24.3 ddRAD stuff

- [X] email Mark about Tryp LD analysis in his paper
 - [X] follow up with him on this “Re: Short meeting to chat about genomic scale LD analysis?”
 - [X] read Tryp paper for same
 - [X] install latest Stacks version on louise and make runnable by Andrea
 - [X] return hapFLK script to original code and copy altered version to new name
 - [X] let Andrea know (*acknowledged*)
 - [ip] Generate descriptive statistics and figures of the LD results as a whole rather than by contig where possible
-

25 2015-01-29 (Thursday)

25.1 ddRAD stuff

- [ip] Generate descriptive statistics and figures of the LD results as a whole rather than by contig where possible
- [2015-01-28_Plot_PLINK_results_cumulative.ipynb](#)

25.2 Robert's stuff

- [-ip-] Pick out 26 flies (13 M, 13 F) from each area we want to look at for Robert's work in March
 - sent first set to Kirstin and Alexis
 - * 13 M/13 F from Oyam/Kole trip in 2014-07 (see Table 1)
- [] make sure we have an updated map with all the villages from *Spring and Summer 2014*
- [X] Meet with Gisella to pick out which locations Robert's data will come from while looking at the updated map.
 - [] she told me to pick the areas and give her a table/report on which and why. The issues to consider are:
 - * wide representation of population areas in the North
 - * allows temporal comparisons as well

Table 1: Samples given to Alexis for DNA extraction from a single leg.

Collection Date	Species	Sex	Teneral	Village	Fly
2014-07-15	<i>G. f. fuscipes</i>	F	NT	OD	017
2014-07-15	<i>G. f. fuscipes</i>	F	NT	OCA	031
2014-07-15	<i>G. f. fuscipes</i>	F	NT	OCA	039
2014-07-16	<i>G. f. fuscipes</i>	F	NT	AKA	045
2014-07-16	<i>G. f. fuscipes</i>	F	NT	AKA	052
2014-07-16	<i>G. f. fuscipes</i>	F	NT	AKA	056
2014-07-16	<i>G. f. fuscipes</i>	F	NT	AKA	062
2014-07-16	<i>G. f. fuscipes</i>	F	NT	AKA	068
2014-07-16	<i>G. f. fuscipes</i>	F	NT	OCA	092
2014-07-16	<i>G. f. fuscipes</i>	F	NT	OCA	100
2014-07-16	<i>G. f. fuscipes</i>	F	NT	ACA	120
2014-07-16	<i>G. f. fuscipes</i>	F	NT	OD	137
2014-07-16	<i>G. f. fuscipes</i>	F	NT	OD	149
2014-07-15	<i>G. f. fuscipes</i>	M	NT	OD	020
2014-07-15	<i>G. f. fuscipes</i>	M	NT	OCA	022
2014-07-15	<i>G. f. fuscipes</i>	M	NT	OCA	025
2014-07-15	<i>G. f. fuscipes</i>	M	NT	OCA	026
2014-07-15	<i>G. f. fuscipes</i>	M	NT	OCA	035
2014-07-16	<i>G. f. fuscipes</i>	M	NT	AKA	049
2014-07-16	<i>G. f. fuscipes</i>	M	NT	AKA	063
2014-07-16	<i>G. f. fuscipes</i>	M	NT	OCA	095

Collection Date	Species	Sex	Teneral	Village	Fly
2014-07-16	<i>G. f. fuscipes</i>	M	NT	ACA	116
2014-07-16	<i>G. f. fuscipes</i>	M	NT	ACA	125
2014-07-16	<i>G. f. fuscipes</i>	M	NT	ACA	129
2014-07-16	<i>G. f. fuscipes</i>	M	NT	OCA	146
2014-07-16	<i>G. f. fuscipes</i>	M	NT	OD	155

25.3 Rob H Jobs

25.3.1 Iowa samples

- [X] catalog the boxes that we got here at **ESC**.
- [X] email Gisella about getting Rob over to **EPH** to start on their samples.

25.3.2 *G. f. fuscipes* samples 2014

- [X] organize boxes of *carcass* and *midgut* samples from 2014-03 to 2014-08 in Gisella freezer by month.
- [-ip-] standardize the collection spreadsheets to prepare for automated import to TsetseSampleDB.

25.4 Updating maps: current trap locations

25.4.1 spartan dev: GPS stuff

- [-ip-] teaching GPSCoordTree how to get mean coordinates
 - [X] make GPSCoord hashable
 - [-ip-] fix GPSCoordTree._add_levels(): getting an empty list somewhere or something that is causing a None to be returned.
- still not fixed but trying a new tactic
 - converting GPSCoordTree to use autovivification trick with an extra key at each node that holds links to all gps_objs found below it.
 - see [2014-12-26_functional_annotation_table_generator.ipynb](#) for example of autovivification method.