Daily Records

Caccone PostDoc

Gus Dunn

February, 2015

Contents

---

1 2015-02-01 (Sunday)

1.1 Updating maps: current trap locations

1.1.1 `spartan` dev: GPS stuff

- writing autovivification version of `GPSCoordTree._grow_branch`.

---

2 2015-02-02 (Monday)

2.1 Updating maps: current trap locations

2.1.1 `spartan` dev: GPS stuff

- testing (test_utils_maps_gps.py):
  - [x] `GPSCoordTree._grow_branch`
  - [x] `GPSCoordTree._get_subtree`
  - [x] `GPSCoordTree.mean`

2.2 Creating Uganda Data Repo

- **local location:**
  /home/gus/Dropbox/uganda_data/data_repos/field_data
- **github address:** https://github.com/CacconeLabYale/field_data.git

---

3 2015-02-03 (Tuesday)

3.1 Updating maps: current trap locations

- established comprehensive lists of village-ID-map and trap GPS locations for Uganda:
  - **village-ID-map:**
    field_data/locations/names/uganda_village_id_map.csv
  - **trap GPS coords:**
    field_data/locations/gps/traps/uganda_traps_gps.csv

---

4 2015-02-04 (Wednesday)

4.1 General ToDo

- [x] email to confirm HR got my letter
- [x] meet with Gisella and Andrea [1130]
  - [X] write up notes from meeting: gisella_andrea_2015-02-04.pdf
- [x] Talk to Ben E about the MAD idea.
- [x] create git repo for this paper
- [ ] begin development of the MAD idea
- [X] install LDna and R-studio
- [X] Located space to move the EPH *G. pallidipes* samples here to ESC with Rob

4.2 ddRAD stuff

4.2.1 LD: detect 'outlier' SNP-pairs

- **I propose this method:**
  1. for each distance group: collect r^2 from $\pm \sim 5$ bp distance window
     a. across genome
     b. across scaffold
  2. calculate modified z-score (based on *median absolute deviation* rather than standard deviation: **MAD is more robust than SD for HTS-type data**)
  3. flag any SNP-pair with $z \geq 3.5$
  4. *possibly randomize data and calculate FDR to evaluate performance.*
     a. perhaps vary the window from step 1 to use FDR to chose window that minimizes FDR.
- **Ben E's thoughts:**
  - basically: this is probably a waste of time and energy
    * other more sophisticated methods have already been applied to this data with not much significance detected
    * why do we expect this work to yield better/more results?
- **Gisella's thoughts:**
  - still should do it bc we will need it when we have more data

4.2.2 Install LDna

- github page: github.com/petrikemppainen/LDna
- paper reference: http://onlinelibrary.wiley.com/doi/10.1111/1755-0998.12369/abstract
- installed `devtools` with `RStudio` gui: **[successful]**
- installed `LDna` with `devtools`: **[successful]**

```
devtools::install_github("petrikemppainen/LDna")
```

  - documentation: LDna/html/00Index.html

4.2.3 LDna notes

- operates on:

> Lower diagonal matrix of pairwise LD values, r^2 is strongly recommended

- the code below should generate what I want (**I think**):

```
plink --vcf tsetseFINAL_14Oct2014_f2_53.recode.renamed_scaffolds.maf0_05.vcf \
--allow-extra-chr \
--r2 bin  \
--out plink_out/tsetseFINAL_14Oct2014_f2_53.recode.renamed_scaffolds.maf0_05.vcf/ld/r2_bin
```

4.2.4 PLINK run for LDna

- ran the command below:

```
wd238 at compute-23-2 in ~GENOMES/glossina_fuscipes/annotations/SNPs (py278)
$ plink --vcf tsetseFINAL_14Oct2014_f2_53.recode.renamed_scaffolds.maf0_05.vcf \
> --allow-extra-chr \
> --r2 bin  \
> --out plink_out/tsetseFINAL_14Oct2014_f2_53.recode.renamed_scaffolds.maf0_05.vcf/ld/r2_bi
```

- waiting for it to finish: **[failed]**

4.2.5 Louise Scratch Request Email

> **netid:** wd238 **group:** caccone **anticipated usage:**
>
> - ~ 100G
> - $<$ 100 files **purpose of usage:**
> - running `plink` *all_v_all* linkage disequilibrium calculations on ~40K SNPs
> - current attempt (documented below) gave a write failure which I think may be bc of some rather large tmp files generated during the process?
> - Does bumping up against our space quota have hard/immediate consequences like that?
>
> **error log:**

```
wd238 at compute-23-2 in ~GENOMES/glossina_fuscipes/annotations/SNPs (py278)
$ plink --vcf tsetseFINAL_14Oct2014_f2_53.recode.renamed_scaffolds.maf0_05.vcf \
> --allow-extra-chr \
> --r2 bin  \
> --out plink_out/tsetseFINAL_14Oct2014_f2_53.recode.renamed_scaffolds.maf0_05.vcf/ld/r2_
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/
516842 MB RAM detected; reserving 258421 MB for main workspace.
--vcf: 73k variants complete.
plink_out/tsetseFINAL_14Oct2014_f2_53.recode.renamed_scaffolds.maf0_05.vcf/ld/r2_bin-temp
+
plink_out/tsetseFINAL_14Oct2014_f2_53.recode.renamed_scaffolds.maf0_05.vcf/ld/r2_bin-temp
+
plink_out/tsetseFINAL_14Oct2014_f2_53.recode.renamed_scaffolds.maf0_05.vcf/ld/r2_bin-temp
written.
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/r2_bin.nose
.
Using up to 63 threads (change this with --threads).
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.
--r2 square bin to
plink_out/tsetseFINAL_14Oct2014_f2_53.recode.renamed_scaffolds.maf0_05.vcf/ld/r2_bin.ld.b
... done.

Error: File write failure.
```

4.2.6 Github repo for this paper

- github page:
  https://github.com/CacconeLabYale/gloria_soria_ddRAD_2015.git

---

5 2015-02-05 (Thursday)

5.1 Mariangela `blacktie` install

- turns out i did NOT send Mariangela install instructions for the development branch
- wrote quick install script for her to use and sent it

5.2 MAD idea

1. for each group of SNPs x bp apart: collect r^2 from $\pm \sim 5$ bp distance window around x:
    a. across genome
    b. across scaffold
2. calculate modified z-score (based on *median absolute deviation* rather than standard deviation: **MAD is more robust than SD for HTS-type data**)
3. flag any SNP-pair with $z \geq 3.5$
4. possibly randomize data and calculate FDR to evaluate performance.
    a. perhaps vary the window-size from step 1 to use FDR to chose window-size that minimizes FDR.

5.2.1 Development

- ipython notebook: ddrad58/2015-02-05_MAD_idea.ipynb

5.3 *G. pallidipes*

- Rob brought most to ESC this morning
- doesn't expect to need my truck for the rest

---

6 2015-02-06 (Friday)

6.1 MAD idea

6.1.1 Development

- LOTS of progress at ipython notebook: ddrad58/2015-02-05_MAD_idea.ipynb
- See notes about plotting median and MAD with bootstrapped CIs near the bottom of above (commit dd7fe5da5733406edeaab6ce3c25b523b94552f2)

---

7 2015-02-09 (Monday)

Goals:

- [x] Zimmer Workshop
- [x] Start Professional Development notebook
- [x] Find out how to process health reimbursement
  - [ ] Get them ready for mailing
    - [x] form
    - [ ] receipts
  - [X] Assemble list of information I need from Sarah and send it to her
- [ ] Progress on MAD idea
- [ ] Generate strategy for the week
- [ ] Sketch out abstract for Keystone? meeting
- [ ] find out if there is data available on tsetse control by area in Uganda
  - chemicals sold
  - etc

7.1 Health reimbursement

- http://yalehealth.yale.edu/claims
- Supplemental Claim form: http://yalehealth.yale.edu/sites/default/files/supplemental_claims_form.pdf
- pharmacy claim form: http://yalehealth.yale.edu/sites/default/files/pharmacy_claim_form_restat_catamaran.pdf

7.1.1 Instructions for pharmacy process

- from website above

> Include copies of prescription receipts showing the following information:
>
> - Pharmacy Name, Address & Phone Number
> - Patient Name
> - Prescription Number
> - Prescription Fill Date
> - Drug Name, Strength and NDC Code
> - Drug Quantity & Days supply
> - Drug Cost
> - Amount Paid
>
> Please mail the Prescription Drug Claim Form and receipts to:
>
> Restat
> Patient Reimbursement
> 11900 W. Lake Park Drive
> Milwaukee, WI 53224
>
> Claims are honored for one year from the date of service. If you haven't received a response to a claim within 60 days of filing, contact the Claims Department. You may call sooner to inquire if the claim has been received and is in process.

---

8 2015-02-10 (Tuesday)

Goals:

- [ ] Get pharm claim ready for mailing
  - [x] form
  - [ ] receipts
- [ ] Progress on MAD idea
- [ ] Generate strategy for the week
- [ ] Sketch out abstract for Keystone? meeting
- [ ] find out if there is data available on tsetse control by area in Uganda
  - chemicals sold

  - etc
- [ ] figure out how to download zimmer files

8.1 Health reimbursement

- printed form

8.2 Met with Postdoc applicant (Christina)

- had lunch

---

9 2015-02-12 (Thursday)

9.1 Health reimbursement

- Need Catherine's member ID

9.2 MAD idea

9.2.1 Development

- *yesterday:*
  - bootstrap confidence intervals are functional
  - modified z-score is functional
  - used ggplot to provide nice figure showing rough progression of z-scored r^2 through distance between snps

---

10 2015-02-13 (Friday)

10.1 *G. pallidipes* Sample catalog

10.1.1 Summary table

- data types:
  - location
  - symbols when present (*I assume you mean location symbol?*)
  - number of individuals
  - date range
  - is tissue?
  - is extraction?
  - analysis status
- will be done in `python` for increased flexibility by **[Gus]**
- notebook file: 2015-02-12_sample_catalog_summary.ipynb
- Showed output to Gisella and she signed off on it after asking whether I could accommodate GEO COORDS when we get them.
- **STATUS: [completed]**

10.1.2 Primers etc

- RobH reports that he and KirstinD found many primers etc that were either designed for *G. pallidipes* or shown to work with it in the past.
- testing on the primers will begin next week.

10.1.3 Leg extractions

- Rob did Xymogen extractions on 5 legs
- NanoDrop indicates absorption at 260 but peaks look weird
  - probably bc the kit leaves EVERYTHING still in solution
  - [ ] RobH will check with KirstinD about her extraction traces on *G. f. fuscipes* legs

10.2 MAD idea

10.2.1 Development

- **[completed]:** functions to
  - update df with `distance_bin` and `mad_z`
  - plot mad_z by bins
- **[to do]:**
  - implement printing/saving snp-pairs that pass the z-filter

---

11 2015-02-14 (Saturday)

11.1 MAD idea

11.1.1 Development

- implement printing/saving snp-pairs that pass the z-filter

---

12 2015-02-16 (Monday)

12.1 *G. f. fuscipes*: infection summaries

- ipython to get pivot table for infected flies
  - file: 2015-02-16_g_f_fuscipes_pandas_import.ipynb
    - file of dumped pandas table of collection records for 2014 in hdf5 format:
- add PCR detected fly statuses to main DB

12.2 *G. pallidipes*: MicroSat extraction pilot

- RobH spoke with KirstinD about strange NanoDrop traces:
  - KirstinD: hers looked the same, just used 260/280 values as presented
  - likely explanation is that the extraction kit is EXTREMELY dirty by design so the spec peaks are shifted around
- RobH is beginning PCRs with ITS primers (same that KirstinD is using on the *G. f. fuscipes*) today.
- RobH is researching location names on the SerapA tubes (n ~ 6) bc GisellaC is not convinced the sheet SerapA included makes since.
  - RobH will google first
  - GusD will get GIS admin layers to search if google fails
- v0.2.1.2-1.tar.gz

---

13 2015-02-17 (Tuesday)

13.1 meeting

- escarpment Nguruman:
  - GisellaC try to get samples from extremes and in the middle
- [ ] GusD send most recent version of protocol to BrianW

14 2015-02-18 (Wednesday)

14.1 *G. pallidipes* status update meeting

- GusD
- RobH
- KirstinD
- extractions not working for a while with KirstinD
- trouble shooting

- KirstinD moving forward with extractions now

---

15 2015-02-21 (Saturday)

GOALS:

- `[worked on]` *G. f. fuscipes* infection summaries/maps for GisellaC meeting
- `[no work]` script for MariangelaB
- `[small work]` r^2 per bin model

15.1 *G. f. fuscipes*: infection summaries

15.1.1 Converting dates to YYYY-MM-DD

- 2014_spring_summer_from_rob.xlsx
  - added new function to `TsetseCheckout`: TsetseCheckout/data/utils.py:convert_brit_dates_to_yyyy_mm_dd(string)
  - added new cell magic to ipython to send variable to clipboard: clip_magic.py
  - used new function and the cell magic to copy, change, then paste back into spreadsheet.
- 2014_fall_for_pandas.xlsx
  - dates already fine

15.1.2 Adding Village names to the spring/summer excel file

- **[COMPLETED]:** 2015-02-22
- created python hack to use the summary sheet info to generate the Village rows YalePostDoc/project_stuff/g_f_fucipes_uganda/collection_data/traps_to_villages.py
  - summary sheets: 2014_full_surveyreport_20140820/summary survey data.xlsx

15.1.3 ALERT: errors detected in fly name code combinations

- during this process i detected instances where the fly number code combinations (example: `OLW-14 038`) were **NOT** correct!
- the following IDs illustrate this:
  - `OLO-14 033` is Olobo
  - `OLO-14 034` is Olobo
  - `OLW-14 035` is Olwi
  - `OLW-14 036` is Olwi
  - `OLW-14 037` is Olobo
  - `OLW-14 038` is Olobo
- additionally, the `Dissection Data-Kole` sheet has **ALL** fly IDs starting `KO` regardless of the source village.
- **RECOMEND *NOT* DEPENDING ON FLY ID FOR VILLAGE SOURCE!**

---

16 2015-02-22 (Sunday)

GOALS:

- `[worked on]` *G. f. fuscipes* infection summaries/maps for GisellaC meeting
- `[none]` script for MariangelaB
- `[none]` r^2 per bin model

16.1 *G. f. fuscipes*: infection summaries

16.1.1 HDF5 import and data cleaning

- standardized the spreadsheet column titles by hand to allow import and correct dataframe referencing
- file: 2015-02-16_g_f_fuscipes_pandas_import.ipynb
- `recode_villages(df, map_func=map_func)`:
  - renaming villages to letter codes

  - **[degenerate names discovered]** and accommodated in uganda_village_id_map.csv by mapping the letter code to more than one long form:
    * AKAYODEBE vs AKAYO-DEBE
  - corrected misspellings of
    * "Orubakulemi" from "Orubakulem"
    * "JIAKO" from "JAIKO"

- `recode_positives(df)`:
  - recode `prob`, `midgut`, `sal.gland` as `0` or `1`.
  - **[NOTE]** this will change to a trivalent state (class `Tristate`) soon

- `recode_tenerals(df)`
  - implemented but needs conversion to `Tristate`

- `recode_dead(df)`
  - implemented but needs conversion to `Tristate`

- `add_infection_state_col(df)`
  - implemented but failing to actually alter the dataframe

- `spartan.utils.misc.Tristate`
  - implements three state logic that *mostly* supports normal boolean arithmetic (just ignoring the `None` state)

17 2015-02-23 (Monday)

GOALS:

- [ ] *G. f. fuscipes* infection summaries/maps for GisellaC meeting
- [ ] script for MariangelaB
- [ ] r^2 per bin model

17.1 *G. f. fuscipes*: infection summaries

17.1.1 HDF5 import and data cleaning

- `spartan.utils.misc.Tristate`

  - I found an existing "Tribool" class on github and forked it: https://github.com/xguse/python_tribool
  - it did not support boolean arithmetic but was much more sophisticated in all other ways.
  - I added support for boolean addition but will also add *, -, and / before writing the tests and submitting a pull request to upstream.
  - I am now using `Tribool` instead of `Tristate`

- running into serious hashable issues `df.midgut.unique()` throws `__nonezero__`'s `ValueError`.
  - possible solutions:
    * override `__new__` might allow me to mimic the "always the same mem address" behavior of `True` etc?
    * Look into implementation of Factories in Python
    * perhaps a hint in behavior/class code for `np.NaN`?
    * **[best bet]** use `enum` class

- looking for more fertile ground to cover while I think