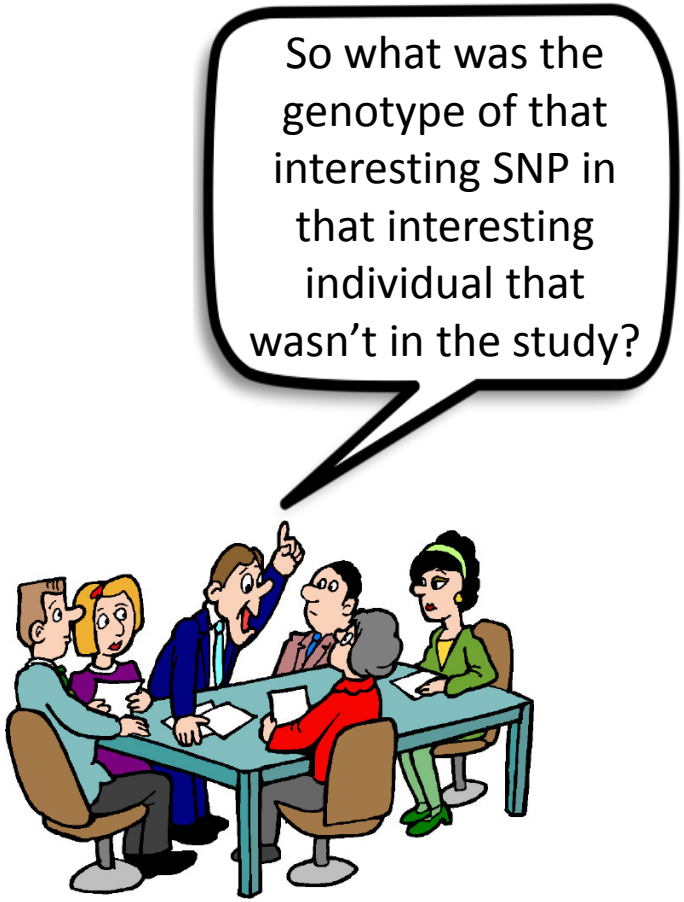
TagDigger software

Lindsay Clark, Oct. 29 2015

Background

- We identify SNPs of interest in one RAD-seq study, and then want to be able to rapidly genotype those same SNPs in other samples.
- Read count data is important for evaluating genotype quality, and for future things I want to do with imputation and polyploids.
- TASSEL and other software doesn't let you specify a subset of SNPs to search for.



Some notes on UNEAK

- Exports read counts in a nice tabular format, but does not report counts higher than 127
- Exported tag sequences include a number that is the length of the tag. Sequence after that is polyA that replaces adapter sequence, and should be removed.

rs	10ES-006-	2011-13_n	2011-18_n	2011-23-1	2011-23-2	2011-26_n	2011-30_n	2011-35_n	2011-38_n	DK96-044_	EBI-2008-5	E
TP28	2|0	96|0	37|0	5|0	31|0	0|0	67|0	0|0	7|0	13|0	1|0	0
TP37	0|0	29|0	24|1	26|0	0|6	31|0	0|0	2|0	4|0	6|0	0|0	0
TP41	11|0	26|4	30|1	75|0	16|0	34|21	38|0	34|14	50|21	5|5	113|0	6
TP49	0|7	18|10	0|22	13|10	10|19	0|68	0|25	9|53	8|30	0|0	1|17	3
TP87	0|0	0|20	0|19	0|19	0|0	0|79	0|29	0|40	0|47	0|3	0|69	0
TP88	0|15	0|74	70|55	13|36	0|58	0|107	0|85	0|127	0|104	26|19	0|127	0
TP95	3|0	26|0	10|0	22|0	6|0	1|0	11|0	3|0	6|0	0|0	3|0	1
TP98	0|0	0|0	0|0	3|0	0|0	7|0	0|0	7|0	8|0	0|0	0|0	0
TP116	0|1	0|0	0|0	0|0	0|0	0|0	0|0	0|0	0|0	0|0	0|14	0
TP119	8|0	3|0	0|0	0|0	7|0	7|1	0|0	0|0	0|7	0|0	0|20	0
TP122	4|0	0|0	0|0	0|0	0|0	0|0	0|0	0|0	0|0	0|0	20|0	1
TP129	13|0	21|0	54|0	5|0	33|0	44|0	52|0	26|0	2|0	9|0	59|0	1
TP132	6|0	14|0	0|0	0|0	0|0	1|0	0|0	19|0	7|0	6|0	0|0	0
TP134	0|2	0|16	0|5	0|0	0|2	0|2	0|0	0|8	0|0	0|0	0|2	0
TP162	0|5	0|80	60|6	0|21	0|50	0|52	0|47	0|48	0|122	0|45	0|127	0
TP176	2|10	4|26	14|15	12|17	0|6	19|25	0|7	0|35	16|16	0|0	6|4	2
TP192	0|0	20|0	12|0	3|0	5|0	23|26	9|0	1|0	1|0	3|0	4|0	1
TP193	0|1	8|15	29|0	0|0	0|0	0|16	15|29	58|0	45|0	5|0	112|0	3

```
>TP4874_query_64
TGCATAAACACCTTCCCGCTTTACTTCCCCTCCCAGTTGGCAATCCTGGTTTTCTACTGCTGGA
>TP4874_hit_64
TGCATAAACACCTTCCCGCTTTGCTTCCCCTCCCAGTTGGCAATCCTGGTTTTCTACTGCTGGA
>TP4921_query_46
TGCATAAACACTAAGTCCCAAGGGTGCCAAGCATATGCCAAGGCCGAAAAAAAAAAAAAAAAAA
>TP4921_hit_46
TGCATAGACACTAAGTCCCAAGGGTGCCAAGCATATGCCAAGGCCGAAAAAAAAAAAAAAAAAA
>TP4934_query_64
TGCATAAACACTCGTAGTTGGATCAGATCCGAATATAAATAGTAGGCTAGGGTTTGGGTGGTCG
>TP4934_hit_64
TGCATAAACACTCGTAGTTGGATCAGATCCGAATATAAATAGTAGGCTAGGGTTTGTGTGGTCG
```

TASSEL-GBS v. 2 output

- VCF format; read counts are there but not accessible without programming skill.

```
E:\TASSEL151023test>more 151026test.vcf
##fileformat=VCFv4.0
##Tassel=<ID=GenotypeTable,Version=5,Description="Reference allele is not known.
 The major allele was used as reference allele">
##FORMAT=<ID=GT,Number=1,Type=String,Description="Genotype">
##FORMAT=<ID=AD,Number=.,Type=Integer,Description="Allelic depths for the refere
nce and alternate alleles in the order listed">
##FORMAT=<ID=DP,Number=1,Type=Integer,Description="Read Depth (only filtered rea
ds used for calling)">
##FORMAT=<ID=GQ,Number=1,Type=Float,Description="Genotype Quality">
##FORMAT=<ID=PL,Number=3,Type=Float,Description="Normalized, Phred-scaled likeli
hoods for AA,AB,BB genotypes where A=ref and B=alt; not applicable if site is no
t biallelic">
##INFO=<ID=NS,Number=1,Type=Integer,Description="Number of Samples With Data">
##INFO=<ID=DP,Number=1,Type=Integer,Description="Total Depth">
##INFO=<ID=AF,Number=.,Type=Float,Description="Allele Frequency">
#CHROM  POS     ID      REF     ALT     QUAL    FILTER  INFO    FORMAT  2011-30
2011-35 2011-39 2011-38 JA26a   RU2012-005      RU2012-006      RU2012-007
RU2012-008      RU2012-009      RU2012-001      RU2012-002      Gifu-2010-009B1a
        X2011-2 Koike-13e       X2011-6 RU2012-010      RU2012-012      RU2012-0
13      Gifu-2010-009B1d        RU2012-221      JA41d   2011-13 2011-18 Blank
2011-23-1       2011-23-2       RU2012-217      RU2012-218      RU2012-219
Tohoku-2010-003a        X2011-10        JA54c   Koike-08e       JA44c   2011-15-
2       Koike-11f       2011-26-1       JA21-2c 2011-26-2       JA46b   EBI-2008
-29a    JA48e
01      4035    S01_4035        G       C       .       PASS    QualityScore=0.0
;DP=1473        GT:AD:DP:GQ:PL  0/1:6,5:11:99:147,0,183 1/0:4,18:22:99:255,0,78
0/0:92,0:92:100:0,255,255       0/0:21,2:23:65:3,0,255  0/0:75,0:75:100:0,225,25
5       0/0:48,0:48:99:0,144,255        0/0:73,0:73:100:0,219,255       0/0:84,0
:84:100:0,253,255       0/0:38,0:38:99:0,114,255        0/0:28,0:28:99:0,84,255
0/0:59,0:59:100:0,177,255       0/0:132,0:132:100:0,255,255     0/0:15,0:15:99:0
,45,255 0/0:13,2:15:99:27,0,255 0/0:15,0:15:99:0,45,255 0/1:17,6:23:99:147,0,255
        0/0:8,0:8:99:0,24,255   0/0:15,0:15:99:0,45,255 0/0:40,0:40:99:0,120,255
        ./.:0,0:0       ./.:0,0:0       0/0:5,0:5:96:0,15,180   0/0:56,9:65:99:1
29,0,255        1/1:1,19:20:99:255,24,0 ./.:0,0:0       1/1:0,22:22:99:255,66,0
1/1:0,9:9:99:255,27,0   0/0:13,0:13:99:0,39,255 0/0:49,0:49:99:0,147,255
0/0:50,0:50:99:0,150,255        0/0:21,0:21:99:0,63,255 0/0:54,6:60:99:36,0,255
0/0:27,0:27:99:0,81,255 0/0:21,0:21:99:0,63,255 0/0:77,0:77:100:0,231,255
0/0:50,0:50:99:0,150,255        0/0:23,0:23:99:0,69,255 1/0:5,15:20:99:255,0,120
        0/0:61,0:61:100:0,183,255       ./.:0,0:0       0/0:9,0:9:99:0,27,255
0/0:4,0:4:94:0,12,144   0/0:51,0:51:99:0,153,255
01      19747   S01_19747       C       T       .       PASS    QualityScore=0.0
;DP=26  GT:AD:DP:GQ:PL  ./.:0,0:0       ./.:0,0:0       ./.:0,0:0       1/1:0,4:
4:94:144,12,0   ./.:0,0:0       ./.:0,0:0       ./.:0,0:0       0/0:3,0:3:88:0,9
,108    ./.:0,0:0       ./.:0,0:0       ./.:0,0:0       ./.:0,0:0       ./.:0,0:
0       ./.:0,0:0       ./.:0,0:0       ./.:0,0:0       ./.:0,0:0       ./.:0,0:
0       ./.:0,0:0       ./.:0,0:0       ./.:0,0:0       ./.:0,0:0       ./.:0,0:
0       1/1:0,9:9:99:255,27,0   ./.:0,0:0       ./.:0,0:0       ./.:0,0:0
./.:0,0:0       ./.:0,0:0       ./.:0,0:0       ./.:0,0:0       ./.:0,0:0
-- More (0%) --
```

3 programs in TagDigger

- **Tagdigger_interactive.py:** Read a FASTQ file and extract read counts for any barcode and tag sequences you specify
- **Barcode_splitter.py:** Split a FASTQ file into multiple FASTQ files by barcode, and remove adapter sequence
- **Tag_manager.py:** Manage RAD-seq markers across multiple projects

Philosophy behind TagDigger

- Everything works on a laptop, on any operating system
- Programming experience not required
- No confusing error messages (hopefully)
- Input and output can be opened in Excel or other spreadsheet software

TagDigger Input: Barcodes

For tagdigger_interactive.py:

File	Barcode	Sample name
RU2012-lib1_C4BADACXX_s_7_fastq.txt.gz	CCGA	RU2012-014
RU2012-lib1_C4BADACXX_s_7_fastq.txt.gz	TTCTA	RU2012-016
RU2012-lib1_C4BADACXX_s_7_fastq.txt.gz	AATGCA	RU2012-017
RU2012-lib1_C4BADACXX_s_7_fastq.txt.gz	GGACGCATTT	RU2012-018
RU2012-lib1_C4BADACXX_s_7_fastq.txt.gz	GTGA	blank
RU2012-lib1_C4BADACXX_s_7_fastq.txt.gz	CACTG	RU2012-020
RU2012-lib1_C4BADACXX_s_7_fastq.txt.gz	TGTGCA	RU2012-023
RU2012-lib1_C4BADACXX_s_7_fastq.txt.gz	ACACACATTT	RU2012-027
RU2012-lib1_C4BADACXX_s_7_fastq.txt.gz	AACG	RU2012-028
RU2012-lib1_C4BADACXX_s_7_fastq.txt.gz	GGTCG	RU2012-029
RU2012-lib1_C4BADACXX_s_7_fastq.txt.gz	CCATCA	RU2012-030
RU2012-lib1_C4BADACXX_s_7_fastq.txt.gz	TTGAACATTT	RU2012-031
RU2012-lib1_C4BADACXX_s_7_fastq.txt.gz	CGAT	RU2012-034
RU2012-lib1_C4BADACXX_s_7_fastq.txt.gz	TCGAA	RU2012-035
RU2012-lib1_C4BADACXX_s_7_fastq.txt.gz	ATCGCA	RU2012-036
RU2012-lib1_C4BADACXX_s_7_fastq.txt.gz	GATCGCATTT	RU2012-037
RU2012-lib1_C4BADACXX_s_7_fastq.txt.gz	TGCA	RU2012-040
RU2012-lib1_C4BADACXX_s_7_fastq.txt.gz	ACTGG	RU2012-042
RU2012-lib1_C4BADACXX_s_7_fastq.txt.gz	GTACAA	RU2012-043
RU2012-lib1_C4BADACXX_s_7_fastq.txt.gz	CAGTCCATTT	RU2012-044
RU2012-lib1_C4BADACXX_s_7_fastq.txt.gz	CATG	RU2012-045
RU2012-lib1_C4BADACXX_s_7_fastq.txt.gz	TTACA	RU2012-046
RU2012-lib1_C4BADACXX_s_7_fastq.txt.gz	ACGTCA	RU2012-047
RU2012-lib1_C4BADACXX_s_7_fastq.txt.gz	GGCAGCATTT	RU2012-048
RU2012-lib1_C4BADACXX_s_7_fastq.txt.gz	GCGG	RU2012-049
RU2012-lib1_C4BADACXX_s_7_fastq.txt.gz	CTCAG	RU2012-050
RU2012-lib1_C4BADACXX_s_7_fastq.txt.gz	TATTAA	RU2012-052
RU2012-lib1_C4BADACXX_s_7_fastq.txt.gz	AGACCCATTT	RU2012-053
RU2012-lib1_C4BADACXX_s_7_fastq.txt.gz	TCAA	RU2012-055
RU2012-lib1_C4BADACXX_s_7_fastq.txt.gz	ATGGA	RU2012-056
RU2012-lib1_C4BADACXX_s_7_fastq.txt.gz	GACCCA	RU2012-058
RU2012-lib1_C4BADACXX_s_7_fastq.txt.gz	CGTTGCATTT	RU2012-059
RU2012-lib1_C4BADACXX_s_7_fastq.txt.gz	ATTA	RU2012-060
RU2012-lib1_C4BADACXX_s_7_fastq.txt.gz	GAAGA	RU2012-061
RU2012-lib1_C4BADACXX_s_7_fastq.txt.gz	CGGCCA	RU2012-062

For barcode_splitter.py:

Input File	Barcode	Output File
RU2012-lib1_C4BADACXX_s_7_fastq.txt.gz	CCGA	RU1_CCGA_RU2012-014.fq
RU2012-lib1_C4BADACXX_s_7_fastq.txt.gz	TTCTA	RU1_TTCTA_RU2012-016.fq
RU2012-lib1_C4BADACXX_s_7_fastq.txt.gz	AATGCA	RU1_AATGCA_RU2012-017.fq
RU2012-lib1_C4BADACXX_s_7_fastq.txt.gz	GGACGCATTT	RU1_GGACGCATTT_RU2012-018.fq
RU2012-lib1_C4BADACXX_s_7_fastq.txt.gz	GTGA	RU1_GTGA_blank.fq
RU2012-lib1_C4BADACXX_s_7_fastq.txt.gz	CACTG	RU1_CACTG_RU2012-020.fq
RU2012-lib1_C4BADACXX_s_7_fastq.txt.gz	TGTGCA	RU1_TGTGCA_RU2012-023.fq
RU2012-lib1_C4BADACXX_s_7_fastq.txt.gz	ACACACATTT	RU1_ACACACATTT_RU2012-027.fq
RU2012-lib1_C4BADACXX_s_7_fastq.txt.gz	AACG	RU1_AACG_RU2012-028.fq
RU2012-lib1_C4BADACXX_s_7_fastq.txt.gz	GGTCG	RU1_GGTCG_RU2012-029.fq
RU2012-lib1_C4BADACXX_s_7_fastq.txt.gz	CCATCA	RU1_CCATCA_RU2012-030.fq
RU2012-lib1_C4BADACXX_s_7_fastq.txt.gz	TTGAACATTT	RU1_TTGAACATTT_RU2012-031.fq
RU2012-lib1_C4BADACXX_s_7_fastq.txt.gz	CGAT	RU1_CGAT_RU2012-034.fq
RU2012-lib1_C4BADACXX_s_7_fastq.txt.gz	TCGAA	RU1_TCGAA_RU2012-035.fq
RU2012-lib1_C4BADACXX_s_7_fastq.txt.gz	ATCGCA	RU1_ATCGCA_RU2012-036.fq
RU2012-lib1_C4BADACXX_s_7_fastq.txt.gz	GATCGCATTT	RU1_GATCGCATTT_RU2012-037.fq
RU2012-lib1_C4BADACXX_s_7_fastq.txt.gz	TGCA	RU1_TGCA_RU2012-040.fq
RU2012-lib1_C4BADACXX_s_7_fastq.txt.gz	ACTGG	RU1_ACTGG_RU2012-042.fq
RU2012-lib1_C4BADACXX_s_7_fastq.txt.gz	GTACAA	RU1_GTACAA_RU2012-043.fq
RU2012-lib1_C4BADACXX_s_7_fastq.txt.gz	CAGTCCATTT	RU1_CAGTCCATTT_RU2012-044.fq
RU2012-lib1_C4BADACXX_s_7_fastq.txt.gz	CATG	RU1_CATG_RU2012-045.fq
RU2012-lib1_C4BADACXX_s_7_fastq.txt.gz	TTACA	RU1_TTACA_RU2012-046.fq
RU2012-lib1_C4BADACXX_s_7_fastq.txt.gz	ACGTCA	RU1_ACGTCA_RU2012-047.fq
RU2012-lib1_C4BADACXX_s_7_fastq.txt.gz	GGCAGCATTT	RU1_GGCAGCATTT_RU2012-048.fq
RU2012-lib1_C4BADACXX_s_7_fastq.txt.gz	GCGG	RU1_GCGG_RU2012-049.fq
RU2012-lib1_C4BADACXX_s_7_fastq.txt.gz	CTCAG	RU1_CTCAG_RU2012-050.fq
RU2012-lib1_C4BADACXX_s_7_fastq.txt.gz	TATTAA	RU1_TATTAA_RU2012-052.fq
RU2012-lib1_C4BADACXX_s_7_fastq.txt.gz	AGACCCATTT	RU1_AGACCCATTT_RU2012-053.fq
RU2012-lib1_C4BADACXX_s_7_fastq.txt.gz	TCAA	RU1_TCAA_RU2012-055.fq
RU2012-lib1_C4BADACXX_s_7_fastq.txt.gz	ATGGA	RU1_ATGGA_RU2012-056.fq
RU2012-lib1_C4BADACXX_s_7_fastq.txt.gz	GACCCA	RU1_GACCCA_RU2012-058.fq
RU2012-lib1_C4BADACXX_s_7_fastq.txt.gz	CGTTGCATTT	RU1_CGTTGCATTT_RU2012-059.fq
RU2012-lib1_C4BADACXX_s_7_fastq.txt.gz	ATTA	RU1_ATTA_RU2012-060.fq
RU2012-lib1_C4BADACXX_s_7_fastq.txt.gz	GAAGA	RU1_GAAGA_RU2012-061.fq
RU2012-lib1_C4BADACXX_s_7_fastq.txt.gz	CGGCCA	RU1_CGGCCA_RU2012-062.fq

TagDigger Input: Tags (six formats)

- HapMap.fas.txt from TASSEL-UNEAK
- SAM from TASSEL-GBS
- Catalog files from Stacks
- CSV with tag sequences in rows
- CSV with tag sequences in two columns
- CSV with merged tag sequences e.g. AACTA[G/C]TTACCCG

For all formats, you can provide a separate file listing a subset of markers that you want to use.

Example use

```
C:\Users\lvclark\Documents\GitHub\tagdigger>python tagdigger_interactive.py

                    TagDigger v. 0.0
              Copyright Lindsay V. Clark
    Released under GNU General Public License v3

Known restriction enzymes are:
ApeKI EcoT22I NcoI NsiI PstI SbfI

What restriction cut site should be found immediately
after the barcode sequence?  Type the name of one of the
above enzymes, OR type the restriction cut site as it
should appear in the sequence data (i.e. not including
bases before the beginning of the overhang) using
characters ACGTRYSWKMBDHVN (IUPAC codes for ambiguous
nucleotides).

Restriction site: PstI
Cut site: TGCAG

Current directory is:
C:\Users\lvclark\Documents\GitHub\tagdigger
Use different directory for reading and writing files? (y/n) y
New directory: E:\tagdiggertest

Contents of current directory:
131108keyfile_E10.csv
test.catalog.alleles.tsv.txt
test.catalog.snps.tsv.txt
test.catalog.tags.tsv.txt

Do you wish to supply a list of marker names?  If provided, this list
will be used to subset the list of markers in the tag file.
Y/N: n


Available tag file formats are:
  1: UNEAK FASTA
  2: Merged tags
  3: Tags in columns
  4: Tags in rows
  5: Stacks catalog

Enter the number of the format of your tag file: 5
```

Yellow rectangles indicate input that the user types in.

Example use

```
Enter the number of the format of your tag file: 5
Enter the name of the *.catalog.tags.tsv file: test.catalog.tags.tsv
Enter the name of the *.catalog.snps.tsv file: test.catalog.snps.tsv
Enter the name of the *.catalog.alleles.tsv file: test.catalog.alleles.tsv
Only retain binary markers? y/n: y
Files not readable.

Enter the number of the format of your tag file: 5
Enter the name of the *.catalog.tags.tsv file: test.catalog.tags.tsv.txt
Enter the name of the *.catalog.snps.tsv file: test.catalog.snps.tsv.txt
Enter the name of the *.catalog.alleles.tsv file: test.catalog.alleles.tsv.txt
Only retain binary markers? y/n: y
6_GA skipped for having non-ACGT nucleotides.
6_GG skipped for having non-ACGT nucleotides.
6_TA skipped for having non-ACGT nucleotides.
20_A skipped for having non-ACGT nucleotides.
20_G skipped for having non-ACGT nucleotides.
32_ATA skipped for having non-ACGT nucleotides.
32_GCG skipped for having non-ACGT nucleotides.
34_A skipped for having non-ACGT nucleotides.
34_G skipped for having non-ACGT nucleotides.
43_CCTGT skipped for having non-ACGT nucleotides.
43_TCGGT skipped for having non-ACGT nucleotides.
43_TCTGG skipped for having non-ACGT nucleotides.
43_TCTGT skipped for having non-ACGT nucleotides.
43_TGTAT skipped for having non-ACGT nucleotides.
43_TGTGT skipped for having non-ACGT nucleotides.
46_CGGCAGCCAAGAANN skipped for having non-ACGT nucleotides.
46_CGGCATCACTNNNNN skipped for having non-ACGT nucleotides.
46_CGGCCCACCGNNNNN skipped for having non-ACGT nucleotides.
46_CGGNTGGTGGTTGTT skipped for having non-ACGT nucleotides.
48_C skipped for having non-ACGT nucleotides.
48_T skipped for having non-ACGT nucleotides.

22 tag sequences read.


Sanitizing tags...
22 tag sequences remain.

Name of key file with barcodes:
```

Yellow rectangles indicate input that the user types in.

Example use

```
Name of key file with barcodes: 131108keyfile_E10.csv
File D2HD4ACXX_2_fastq.txt.gz: 96 barcodes


File name for output of read counts: mycounts.csv

Output CSV of diploid numeric genotypes? Y/N y
File name for output of genotypes: mygenotypes.csv

Press enter to begin processing FASTQ files.
Reads: 50000 With barcode and cut site: 48694 With tag: 6
Reads: 100000 With barcode and cut site: 97890 With tag: 8
Reads: 150000 With barcode and cut site: 147075 With tag: 11
```

Yellow rectangles indicate input that the user types in.

Progress printed as it runs.

- Takes ~2 hours to go through a FASTQ file, depending on how many barcodes and tags you are looking for.
- Before it starts going through the FASTQ file, it builds indexing trees for the barcodes and tags to that it can search as quickly as possible.

```
Reads: 202900000 With barcode and cut site: 198130070 With tag: 20324
Reads: 202950000 With barcode and cut site: 198174370 With tag: 20328
D2HD4ACXX_2_fastq.txt.gz
Reads: 203000000 With barcode and cut site: 198218533 With tag: 20332
Reads: 203050000 With barcode and cut site: 198262475 With tag: 20337
Reads: 203100000 With barcode and cut site: 198306107 With tag: 20340
Reads: 203150000 With barcode and cut site: 198349508 With tag: 20344
Reads: 203200000 With barcode and cut site: 198393012 With tag: 20349
Reads: 203250000 With barcode and cut site: 198436109 With tag: 20352
Reads: 203300000 With barcode and cut site: 198479096 With tag: 20356
Reads: 203350000 With barcode and cut site: 198521818 With tag: 20361
Reads: 203400000 With barcode and cut site: 198564809 With tag: 20366
Reads: 203450000 With barcode and cut site: 198607628 With tag: 20369
Reads: 203500000 With barcode and cut site: 198650430 With tag: 20372

Press enter to quit.
```

TagDigger output: read counts

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W
1		4_A_0	4_C_1	7_C_0	7_G_1	18_CC_0	18_TT_1	19_C_0	19_G_1	21_A_0	21_G_1	25_A_0	25_C_1	35_G_0	35_T_1	36_A_0	36_G_1	39_C_0	39_T_1	41_C_0	41_T_1	47_A_0	47_G_1
2	X2011-10	58	8	111	241	0	0	0	1	274	293	0	0	68	4	0	73	0	228	40	0	23	26
3	X2011-2	0	49	61	85	0	0	0	0	189	197	0	0	58	0	42	65	0	145	45	0	1	11
4	X2011-6	0	0	151	160	0	0	0	0	196	168	0	0	113	0	0	46	0	247	26	0	3	12
5	2011-13	10	25	126	160	0	0	0	0	178	172	0	0	77	0	1	25	0	159	6	0	0	12
6	2011-15-2	0	11	0	49	0	0	0	0	41	31	0	0	24	0	0	0	0	75	0	0	0	0
7	2011-18	0	0	107	79	0	0	0	0	108	120	0	0	64	0	0	48	0	103	55	0	0	26
8	2011-23-1	0	0	25	38	0	0	0	0	53	38	0	0	48	0	0	42	0	25	9	0	0	0
9	2011-23-2	0	0	48	99	0	0	0	0	45	62	0	0	37	0	0	43	0	53	9	0	0	12
10	2011-26-1	0	0	71	43	0	0	0	0	58	69	0	0	32	0	0	38	0	73	0	0	0	0
11	2011-26-2	0	0	59	79	0	0	0	0	67	64	0	0	22	0	0	66	0	57	13	0	0	0
12	2011-30	0	6	71	124	0	0	0	0	61	84	0	0	56	0	0	30	0	119	53	0	0	0
13	2011-35	0	0	70	123	0	0	0	0	68	192	0	0	89	1	12	58	0	132	29	0	14	17
14	2011-38	0	0	98	85	0	0	0	0	66	44	0	0	31	0	0	33	0	108	27	0	0	10
15	2011-39	0	41	0	119	0	0	0	0	119	75	0	0	6	4	0	41	0	77	0	0	0	0
16	RU2012-001	0	39	0	62	0	0	0	0	77	62	0	0	21	28	0	51	0	119	0	0	0	11
17	RU2012-002	0	121	0	128	0	0	0	0	116	166	0	9	13	0	0	107	0	154	0	5	0	10
18	RU2012-005	0	21	0	48	0	0	0	0	23	32	0	0	3	0	0	20	0	27	0	0	0	6
19	RU2012-006	0	31	0	71	0	0	0	0	78	52	0	0	4	5	0	83	0	96	0	0	0	21
20	RU2012-007	0	63	0	138	0	0	0	0	131	139	0	0	7	8	0	121	0	178	0	0	0	36
21	RU2012-008	0	20	0	50	0	0	0	0	48	45	0	0	0	9	0	46	0	79	0	0	0	18
22	RU2012-009	0	27	0	23	0	0	0	0	27	46	0	0	0	7	0	34	0	38	0	0	0	8
23	RU2012-010	0	53	0	56	0	0	0	0	98	52	0	0	0	0	0	15	0	47	0	0	0	0
24	RU2012-012	0	103	0	53	0	0	0	0	30	61	0	0	11	7	0	6	0	33	0	45	0	7
25	RU2012-013	0	2	0	21	0	0	0	0	47	33	0	0	0	3	0	13	0	15	0	0	0	8
26	RU2012-217	0	47	0	75	0	0	0	0	100	58	0	0	3	1	0	25	0	46	0	6	0	11
27	RU2012-218	0	41	0	59	0	0	0	0	72	63	0	7	13	0	0	24	0	16	0	8	0	0
28	RU2012-219	0	0	0	87	0	0	0	0	70	93	0	0	14	0	0	27	0	81	0	9	0	0
29	RU2012-221	0	5	0	106	0	0	0	0	85	55	0	0	2	21	0	96	0	113	0	39	0	10
30	Blank	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
31	Koike-13e	0	46	0	85	0	0	0	0	144	125	0	0	26	13	0	55	0	94	0	8	0	0
32	Tohoku-2010-003a	0	46	0	47	0	0	0	0	62	58	0	0	4	0	0	32	0	58	0	10	0	0
33	EBI-2008-29a	1	61	0	73	0	0	0	0	38	7	0	0	0	3	0	27	0	13	0	23	0	0
34	Gifu-2010-009B1a	0	17	0	58	0	0	0	0	63	46	0	0	10	0	0	8	0	25	0	2	0	0
35	Gifu-2010-009B1d	0	14	0	55	0	0	0	0	107	39	0	0	6	0	0	8	0	90	0	6	0	0
36	JA21-2c	0	31	0	97	0	0	0	0	109	178	0	0	10	1	0	52	0	52	0	37	0	15
37	JA26a	0	18	0	85	0	0	0	0	65	96	0	0	0	3	0	46	0	62	0	0	0	0
38	JA41d	0	23	0	78	0	0	0	0	30	87	0	0	10	1	0	59	0	38	0	9	0	3
39	JA44c	0	68	0	59	0	0	0	0	64	22	0	0	0	1	0	26	0	75	0	0	0	9
40	JA46b	0	38	0	71	0	0	0	0	45	68	0	7	0	2	0	7	0	41	0	3	0	0

(Marker names are simply numbers in this example, since tags came from Stacks.)

TagDigger output: genotypes

	A	B	C	D	E	F	G	H	I	J	K	L
1		4	7	18	19	21	25	35	36	39	41	47
2	X2011-10	1	1		2	1		1	2	2	0	1
3	X2011-2	2	1			1		0	1	2	0	1
4	X2011-6		1			1		0	2	2	0	1
5	2011-13	1	1			1		0	1	2	0	2
6	2011-15-2	2	2			1		0		2		
7	2011-18		1			1		0	2	2	0	2
8	2011-23-1		1			1		0	2	2	0	
9	2011-23-2		1			1		0	2	2	0	2
10	2011-26-1		1			1		0	2	2		
11	2011-26-2		1			1		0	2	2	0	
12	2011-30	2	1			1		0	2	2	0	
13	2011-35		1			1		1	1	2	0	1
14	2011-38		1			1		0	2	2	0	2
15	2011-39	2	2			1		1	2	2		
16	RU2012-0(	2	2			1		1	2	2		2
17	RU2012-0(	2	2			1	2	0	2	2	2	2
18	RU2012-0(	2	2			1		0	2	2		2
19	RU2012-0(	2	2			1		1	2	2		2
20	RU2012-0(	2	2			1		1	2	2		2
21	RU2012-0(	2	2			1		2	2	2		2
22	RU2012-0(	2	2			1		2	2	2		2
23	RU2012-01	2	2			1			2	2		
24	RU2012-01	2	2			1		1	2	2	2	2
25	RU2012-01	2	2			1		2	2	2		2
26	RU2012-21	2	2			1		1	2	2	2	2
27	RU2012-21	2	2			1	2	0	2	2	2	
28	RU2012-219		2			1		0	2	2	2	
29	RU2012-22	2	2			1		1	2	2	2	2
30	Blank											
31	Koike-13e	2	2			1		1	2	2	2	
32	Tohoku-2(	2	2			1		0	2	2	2	
33	EBI-2008-2	1	2			1		2	2	2	2	
34	Gifu-2010-	2	2			1		0	2	2	2	
35	Gifu-2010-	2	2			1		0	2	2	2	
36	JA21-2c	2	2			1		1	2	2	2	2
37	JA26a	2	2			1		2	2	2		
38	JA41d	2	2			1		1	2	2	2	2
39	JA44c	2	2			1		2	2	2		2

(Marker names are simply numbers in this example, since tags came from Stacks.)

Barcode splitter program

- We can't deposit our raw sequencing data to NCBI
- Must split the FASTQ file by barcode and remove barcode and adapter sequence
- -> Each file corresponds to one individual, and all sequence is genomic

NCBI Resources How To

SRA SRA Advanced

Display Settings: Full

SRX316140: PMS-601
1 ILLUMINA (Illumina HiSeq 2000) run: 1.3M spots, 107.8M bases, 99Mb downloads

Design: "To detect variation and population structure in a broad sample of Miscanthus
DNA was digested with PstI and MspI and ligated to barcoded PstI adapters and a univ
selection of 200-500 bp was performed on a 2% agarose gel. After sequencing, the soft
barcode sequence, trim the remaining sequence to 80 nt, and filter out low quality read

Submitted by: UNIVERSITY OF ILLINOIS, URBANA-CHAMPAIGN

Study: Genetic diversity of Miscanthus sinensis
PRJNA207721 • SRP026347 • All experiments • All runs
show Abstract

Sample: Generic sample from Miscanthus sinensis
SAMN02213367 • SRS453423 • All experiments • All runs
Organism: Miscanthus sinensis

Library:
Name: PMS-601
Instrument: Illumina HiSeq 2000
Strategy: WGS
Source: GENOMIC
Selection: Reduced Representation
Layout: SINGLE

Spot descriptor:
1 forward

Runs: 1 run, 1.3M spots, 107.8M bases, 99Mb

Run	# of Spots	# of Bases	Size	Published
SRR922822	1,347,240	107.8M	99Mb	2014-06-26

ID: 441274

Barcode splitter program

- Although other programs do this, it is nice not to have to use Biocluster
- We use Megan Hall's adapter sequences, which are non-standard
- Get MD5 checksums for each file without using Biocluster

- Uses TagDigger algorithm for rapid searching of barcodes

Tag Manager program

- Using UNEAK across multiple projects, the same SNPs do not end up getting the same names
- We want to have a consolidated list of all markers for the lab, and a way to quickly match new and old markers

Tag Manager program

```
C:\Users\lvclark\Documents\GitHub\tagdigger>python tag_manager.py

        TagDigger v. 0.0 Tag Manager
         Copyright Lindsay V. Clark
 Released under GNU General Public License v3


Current directory is:
C:\Users\lvclark\Documents\GitHub\tagdigger
Use different directory for reading and writing files? (y/n) y
New directory: ../../tagdigger/151016Miscanthus

Contents of current directory:
AOBfreqs.csv
AOB_markerlist.txt
ClareM1LGs.csv
ClareM1_markerlist.txt
database151016preA.csv
database151016preB.csv
database151016preC.csv
database151016preD.csv
HapMapAOB.fas.txt
HapMapClareM1.fas.txt
HapMapJXB.fas.txt
HapMapRU2012.fas.txt
JapanSNPsRetained.csv
MiscanthusTagDatabase151016.csv
MiscanthusTagDatabase151016.xlsx
non-paralog RU marker names.csv
old
preDalign.sam
preDtags.fa

        Options are:
1. Look up markers by sequence in existing database
2. Add markers to existing database
3. Add alignment data to database
4. Start new database

Select option:
```

Tag Manager program

- Reads tags in any of the six formats mentioned above
- Assigns new names with prefix + number, e.g. UIMiscanthus000102.
- Can include column with original marker names
- Can output FASTA and import SAM to add alignment information
- Output in CSV to read with Excel or use with TagDigger

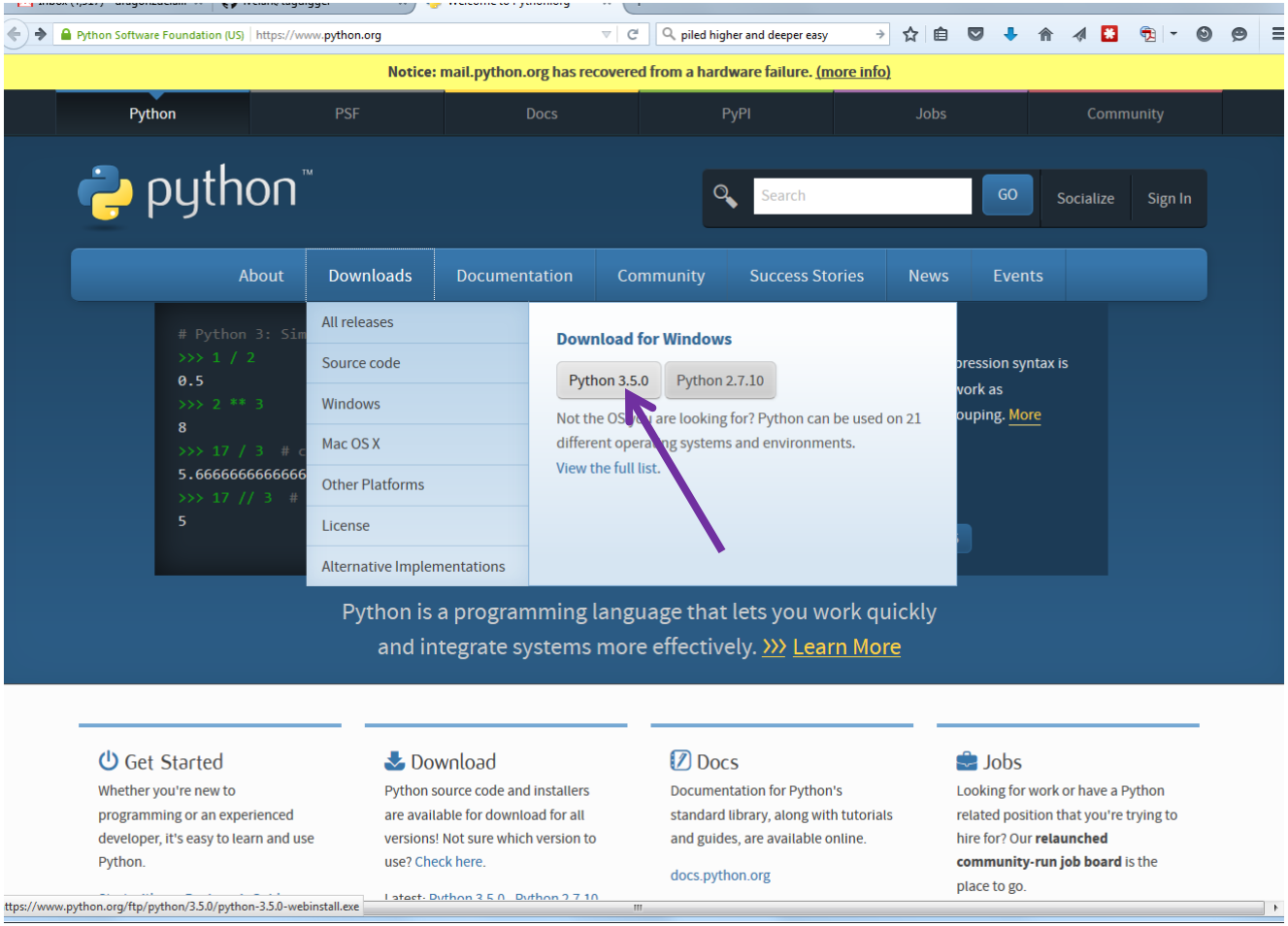
Example tag database

	A	B	C	D	L	M	N	O	P	Q
1	Marker name	Tag sequence	Miscanthus LG	ClareMap1 name	AOB Msi names	JXB Japan names	RU2012 names	Sbicolor 2.0 chromosome	Sbicolor 2.0 position	Sbicolor 2.0 quality
2	UIMiscanthus000001	TGCAGAAAAAA[A/G]TAGTGAGAAATGGGCCCGATAGCCATGGAACTTGGTCACTTGGAAGAATTGC	12	TP58	TP666	TP280				
3	UIMiscanthus000002	TGCAGAAAACACATGGGAAGGAAGCGGAAGAACATGGTGGCTACCC[A/G]CTCGAGCCTACGCCTCA	9	TP270				Chr04	64822460	22
4	UIMiscanthus000003	TGCAGAAAACATGGAGA[G/T]GGAGATGGCACGGCAGCACCACCGCTGGTCCGCTGCCCGTTTGCGG	8	TP316						
5	UIMiscanthus000004	TGCAGAAAACCACCAAATCATCAGTACTCTCTCCCTCTTTCGCTTCACCAGCCTAGTTCG[A/G]GGA	18	TP329				Chr10	7463679	12
6	UIMiscanthus000005	TGCAGAAAAGGTTCTTCCTTA[A/T]TGGCTTATTGCAGAAAAGGTTCTGACCAAGAAACTTAGCAGG	5	TP505		TP2566		Chr03	59281098	28
7	UIMiscanthus000006	TGCAGAAAATAGCCCAGGTGAAGAGACAGCCCACCAGCACGGCAACGAGGCC[A/G]ACGGCGGCCGC	12	TP566				Chr06	55707967	22
8	UIMiscanthus000007	TGCAGAAAATCAGAGTCTTTGATACTAACATAATCGATTCCCACACATT[G/T]ATTTAATAATTTAT	6	TP593		TP2910		Chr03	72696545	28
9	UIMiscanthus000008	TGCAGAAAATTGAAGAGCTAA[A/T]GCTTCGTGATGGCGGAGGGTGCAATGAGCCCAACTGGGATGG	16	TP672						
10	UIMiscanthus000009	TGCAGAAACA[A/G]TTTCGACCGCTGTCATCAGAACACGCCAAGCGTGATTGGAGAAGCAAACAGAA	19	TP742				Chr10	1130941	28
11	UIMiscanthus000010	TGCAGAAACACGGCAATGAAGGGCGTCCCAAACCTTCCTTG[C/T]GCGAGGACAGGCAGTGAAGATG	3	TP768						
12	UIMiscanthus000011	TGCAGA[A/G]ACAGAGTTGCAGCAGCTGCTGCCTTGGAACTTGGAAGCGTTCCAAGAAACCCCTCCT	3	TP794	TP9218					
13	UIMiscanthus000012	TGCAGAAACAGTAGTAAGA[A/G]ATGTTTACGAAACTAAGTAAAAAAAAAGTATATGAATTTGACCA	11	TP823						
14	UIMiscanthus000013	TGCAGAAACATCTAGGTAAGCATGTACAA[G/T]AAAATAGATTCAAGAAAATATTATTCATATGCCC	8	TP848	TP9697	TP3985				
15	UIMiscanthus000014	TGCAGAAACCAAACCAAG[A/C]ATGAACTTGTCAGCCTATACGCTGCTGGCCAGTCATCAGCTGGCA	8	TP877				Chr04	58972377	22
16	UIMiscanthus000015	TGCAGAAACCCAGCCCGATCGACATTCAGTTTCTTTTGTGCCTGTG[C/T]ACGCACGTACCTCTTCT	6	TP928						
17	UIMiscanthus000016	TGCAGAAACCTCAAAACCCCAACGA[C/T]GCCGCAACCGCAAGGGGACAAGGCGGGAGGGGTTTCTG	19	TP966				Chr10	8778807	24
18	UIMiscanthus000017	TGCAGAAACCTT[C/T]TGCCAAATCCTCCATGCCATGCCAATTGCCAGAGCCCAGAGAGGCGGTCAC	19	TP981	TP10852	TP4547		Chr10	57820630	36
19	UIMiscanthus000018	TGCAGAAACGATACTCAGCGACATGCCA[A/G]TCGCATTGGGGTGATTGGTGTCCGCCTCGTGTTGT	6	TP999				Chr03	54754855	42
20	UIMiscanthus000019	TGCAGAA[A/G]CTAGAGCCGAAGGTTTACACTGACACCATGATCACGGCTTTTTGCTCGGTGAGGGA	1	TP1086						
21	UIMiscanthus000020	TGCAGAAACTTCTCTCGCGCTTCACATGGGCTGTTGCATCTACCAC[C/T]ACAGTTTAGAGGCAGCT	13	TP1167						
22	UIMiscanthus000021	TGCAGAAAGAAGCCAGAGGCCGCGGCAGGCACGGACACTTCACGCAGGCCGCGCACTAG[A/C]GCCA	3	TP1197				Chr02	65321410	41
23	UIMiscanthus000022	TGCAGAAAGAGGGCGGCTTGCCCAT[C/T]GAGCTCACCGACGACATCGTCATGGCGCTCATGTTTGT	1	TP1243				Chr01	23701888	24
24	UIMiscanthus000023	TGCAGAAAGATAAACAGCAGAAGAGAAGATAAAGCTAAATGAGAGCTGGTTTGA[A/C]TATGATCAT	3	TP1248				Chr02	61769551	22
25	UIMiscanthus000024	TGCAGAAAGCAGC[A/C]CGCGCGGGGCACAATGAAACTGTTGAGTTTGAACTGCTGGGGCTTGGGGC	6	TP1300						
26	UIMiscanthus000025	TGCAGAA[A/C]GCCGCGACGAGCGTGTTGTAAGTCACGGCATTGGGGTCGCAGCCCGCGCCGCGCAT	19	TP1336			TP2167	Chr10	3253914	44
27	UIMiscanthus000026	TGCAGAAAGGAAGATA[C/T]GTGCGCGGGTAGATAGGAGTTGAGGAGAGTAGATCGAGAAAGACGCA	7	TP1380		TP6080				
28	UIMiscanthus000027	TGCAGAA[A/C]GGCGGCGGCACTTCTGAGAAAGGCGACGGACCCGAGGGCGGCGTCCACTGAGCCCG	1	TP1435						
29	UIMiscanthus000028	TGCAGAAAGGTT[A/G]GCAATCGCTACGTGCCCACGACGCGGCAGATACGGTTCGACCTGGGCCGTG	9	TP1473				Chr05	2177855	36
30	UIMiscanthus000029	TGCAGAAAGTAGTAAATTTTGGATGGAAATACTCTTTTAATTTTGTTTTCGGACGTTACGC[C/T]AA	9	TP1489						
31	UIMiscanthus000030	TGCAGAAATA[C/T]AGGCTGCACACGCGGCGCACGTCGTCGGACGTCGGCGGCGTCGACCAGCACGT	14	TP1578				Chr08	50640445	36
32	UIMiscanthus000031	TGCAGAAATCAAACGCAAGGTAAGCTTTTCAATTAGGCCATGATTTC[C/T]AGCCGAGAAACAAGTA	1	TP1640		TP7139		Chr01	60153827	36
33	UIMiscanthus000032	TGCAGAAATGCAGAATGCCAGCGTTG[C/G]CCGTTTCTTGCCCACGCAGACGGCAGACGCAGGCAGC	7	TP1752				Chr04	62658081	28
34	UIMiscanthus000033	TGCAGAAATG[C/T]TCTCTCTGTTCACTAGTGGCCAGAATTATGTAGATGCTGTTTATTGAAAGAAG	7	TP1785				Chr07	62509621	36
35	UIMiscanthus000034	TGCAGAAATGGAGAGTATAATGCATGTACAAATCTGTCAAATATATCGTGGAGCATTTAGGTA[C/T]	6	TP1794	TP19139	TP7765		Chr03	3372777	36
36	UIMiscanthus000035	TGCAGAAATGGTTCACC[A/T]GCTACTGGAATGGAGGCCAGCTCTGGCCGACCAAGTCGACAGCAGC	10	TP1819				Chr05	7084542	9
37	UIMiscanthus000036	TGCAGAAATGTACTGCACCACAAGTAGTAATAAACACATATAGGAAGCGTCCGCCGTTACA[C/T]GA	2	TP1828						
38	UIMiscanthus000037	TGCAGAAATTCAGAG[A/G]GAAAGGATCCAAGATAAAAAGAAATCAAGCCATGTTTCCAACCAATGA	16	TP1904				Chr10	42282565	22
39	UIMiscanthus000038	TGCAGAAATTCTTTTTAGGCCCACGTATAGCAGG[C/T]TGATTGCCCAACCAGGATTCACATGGGTC	5	TP1912						
40	UIMiscanthus000039	TGCAGAACAACCCCTCGTAGACGCCGAGCATGGCGTCGAAATCGACCACGGCG[C/G]CCGCCACGCC	16	TP2053	TP21349	TP8731	TP3354			
41	UIMiscanthus000040	TGCAGAACACGCACGCACCGA[C/T]CTCAGGAGTCAGGTTCTTCAGAAAATTACAACGTGATATTGA	18	TP2198						
42	UIMiscanthus000041	TGCAGAACAGAAGTGGGAAAGAGCGCGAGCGAGCCGACATCAGGTTTTGGGCCTT[A/G]GTGTAATA	5	TP2243	TP22949	TP9447				
43	UIMiscanthus000042	TGCAGAACAGAATGAATACCCTCCTTTCACTGGCTCCAGCACGTCGCTCCTCCGA[C/T]CCCGACGA	5	TP2246				Chr03	2878403	22

* don’t include markers that are paralogous or that were not included in your analysis

Obtaining TagDigger

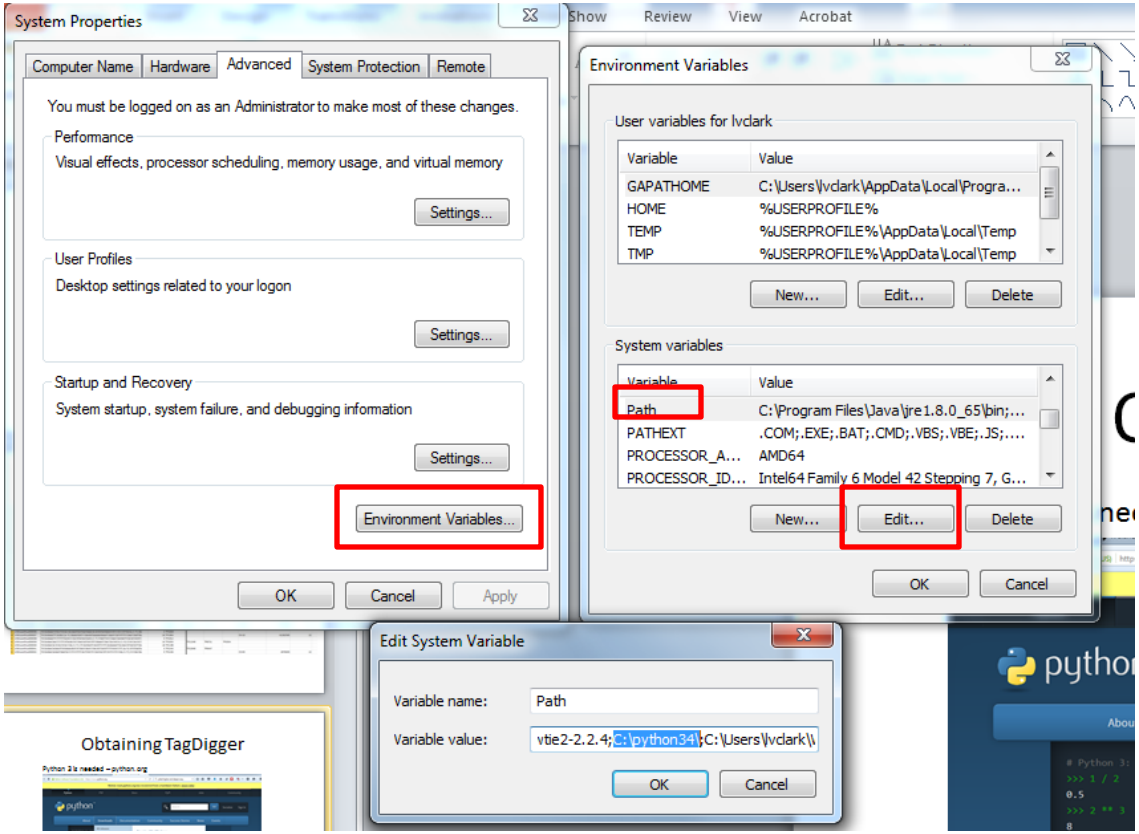
Python 3 is needed – python.org



(Follow installation instructions on the website.)

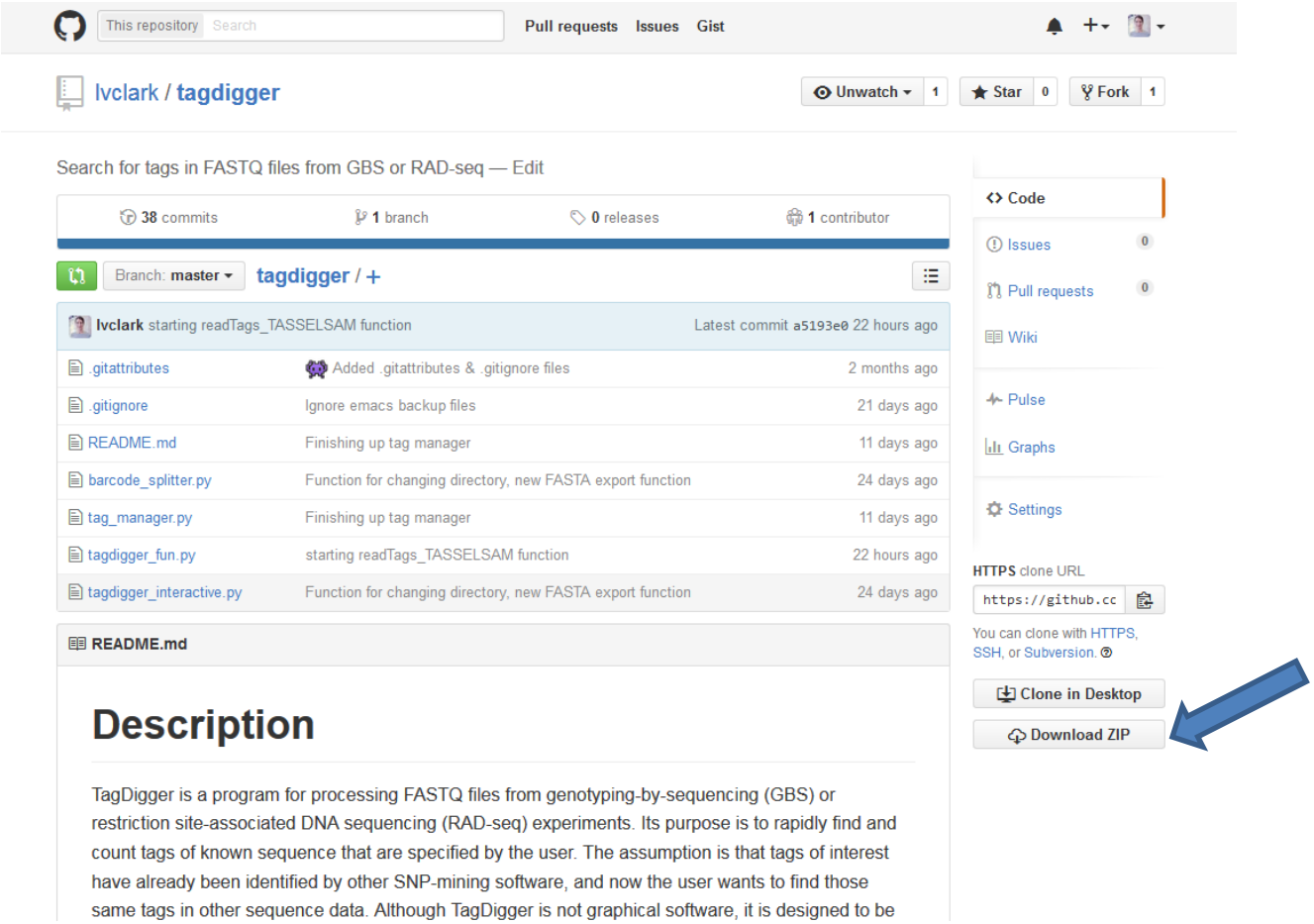
Obtaining TagDigger

Add Python 3 to your PATH variable (in Advanced System Settings):



Obtaining TagDigger

https://github.com/lvclark/tagdigger



Obtaining and running TagDigger

- Unzip the tagdigger directory
- Use ‘cd’ to navigate into that directory
- Type ‘python’ and the name of the program you want to run.