

Tuesday, 24 July 2018

- Quick review of Problem Set solutions
- DArTseq file formats
- Anchoring genotypes to a genome assembly
- Marker filtering

DArT file formats

DArTseq files

Report_DY17-1234/

Report_DY17-1234_SNP_singlerow.csv

Report_DY17-1234_SNP.csv

Report_DY17-1234_SilicoDArTs.csv

metadata.json

SNP (singlerow)

[illegible]

SNP (standard/two-row)

[illegible]

SilicoDArTs

* , * , * , * , * , * , * , * , * , * , * , DY17-2486,DY17-2486,DY17-2486,DY17-2486,DY17-2
*, *, *, *, *, *, *, *, *, *, *, *, *, 910917261001,910917261001,910917261001,9109172
*, *, *, *, *, *, *, *, *, *, *, *, *, 1,
*, *, *, *, *, *, *, *, *, *, *, *, *, A,A,A,A,A,A,A,A,A,A,B,B,B,B,B,B,B,B,B,B,
*, *, *, *, *, *, *, *, *, *, *, *, 1,10,11,12,2,3,4,5,6,7,8,9,1,10,11,12,2,3,4,5,
*, *, *, *, *, *, *, *, *, *, *, *, Progeny,Progeny,Progeny,Progeny,Progeny,Progen
CloneID,AlleleSequence,TrimmedSequence,Chrom_Assembly,ChromPos_Assembly,Al
100018673,TGCAGCCAAATACTAAGCTTTAGTTCCTACGATCAACCAAGATGCTAACCTTGAATCTTGATGA
100062117,TGCAGAACTTGAATTTATCAAGAAATCATTTATCAGGACAACCTTCAGAAAAATCGGCCAGCT
100070803,TGCAGCAATTGAGAACACACTGTCAACTAAGCTAGAGGATGCTCGGCAGCTATTGCAGTTACAG
100000052,TGCAGCTGGTGAATTTGTTATATTTATCTTTGCTGTCATCTACTGAAGGATTTGAAAGTAGTTT
100069972,TGCAGAACTAATGCTGTTCAAACCACAGAGAATGTAATACAGATCTCATTGGACCATTGTAAG
100022056,TGCAGGCTAGCTTCTTGATGATAGAAAAGATGAGCCATTGGATCTTCATATTTAGTAAGCTGC

JSON: JavaScript Object Notation

metadata.json

```
"orders" : {  
  "DY17-1234" : {  
    "clientemail" : "your@address.org",  
    "sequencing" : {  
      "1" : {  
        "HiSeq 2500" : "2017-06-02 16:14:10"  
      },  
      "2" : {  
        "HiSeq 2500" : "2017-06-23 14:47:38"  
      }  
    },  
    "orderdatetime" : "2017-05-05 03:19:36",  
    "clientname" : "Your Name",  
    "productname" : "DARtseq (1.0)"  
  }  
},
```

Mapping SNP tags to a genome

MapTK: Introduction to the UI

MapTK is a tool kit for linkage mapping

- Many useful tools/commands
- Every tool has its own usage page, options, and arguments

Not compatible with Windows machines--yet!

Note: Non-required and required arguments: [arg] vs. <arg>

MapTK UI: list of commands

```
jess — lyons@ip-172-31-31-27:~ — ssh -i ~/Dropbox/DArtseq_to_mapping_workshop/lyons.pe
[lyons@ip-172-31-31-27 ~]$ maptk
Prototype mismatch: sub main::assert: none vs (&@) at /usr/local/lib/Exporter.pm line 66.
  at /usr/local/bin/maptk line 40.

Program: maptk (Utilities for working with mapping data)
Version: 0.1.0
Contact: Jessen Bredeson <jessenbredeson@berkeley.edu>

Usage:  maptk <command> [options]

Command: dart2vcf  Convert DArT singlerow CSV format to VCF
maptags  Extract and re-map marker tags to an assembly
chif1    Calc. Mendelian segregation stats for F1 crosses
mtvr     Calc. Mendelian transmission violation rate
mclust   Model-based clustering of relatedness2 output
vcf2loc  Convert VCF to JoinMap .loc format for mapping
vcf2raw  Convert VCF to OneMap .raw format for mapping

dotplot  Plot the genetic distance of one map against another
h2k      Convert from Haldane to Kosambi centiMorgan scale
nn       Est marker genetic position by nearest neighbors
predict  Est marker genetic position by physical position
paint    Color a genetic map by component contig
rescale  Rescale a map by regressing onto a second map
rev      Reverse some or all LG orientations

anchor   Anchor and orient contigs onto a genetic map
break    Break scaffolds at user-defined locations
join     Join scaffolds with user-defined join information
```

Notes:

1. Input .map files are required to adhere to the PLINK MAP file spec (<http://pngu.mgh.harvard.edu/~purcell/plink/data.shtml#map>).
2. Many of the above tools include a '-M' option which, when enabled, requires marker IDs to be formatted as a concatenation of each marker's contig ID and physical position separated by any one of the characters in the set [:/], e.g. "Contig1:1000". This marker ID format is intended to allow these tools to reference a marker's coordinate in both sequenced contig space and chromosome space should they differ (as when anchoring scaffolds onto a map to form pseudomolecular representations of the chromosome).

MapTK dart2vcf command

```
jess — lyons@ip-172-31-31-27:~ — ssh -i ~/Dropbox/DArtseq_to_mapping_workshop/lyons.pem
```

```
[lyons@ip-172-31-31-27 ~]$ maptk dart2vcf
```

```
Prototype mismatch: sub main::assert: none vs (&@) at /usr/local/lib/Exporter.pm line 66.  
at /usr/local/bin/maptk line 40.
```

```
Usage: maptk dart2vcf [options] <singlerow.csv>
```

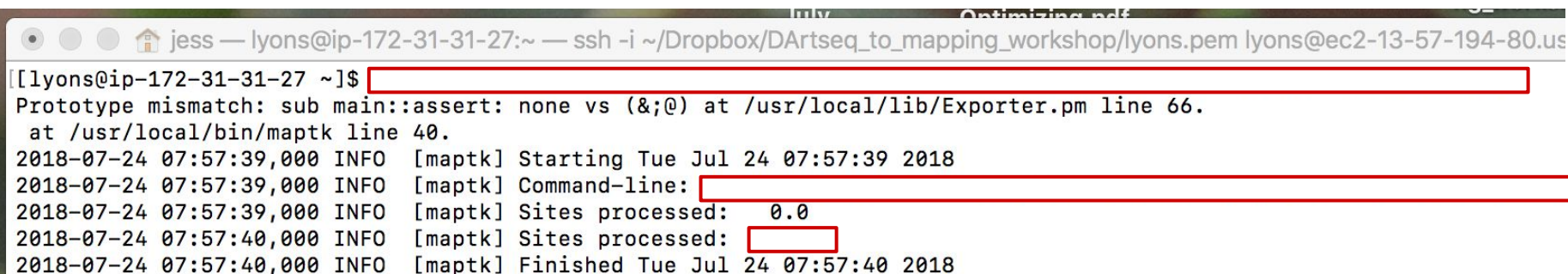
```
Options: -o <file>      Write output to FILE [stdout]  
         -A <str>       Assembly identifier (see Note 1)  
         -R <file>       Specify the reference fasta FILE (see Note 2)  
         -h              This help document
```

Notes:

1. The assembly identifier is that observed in the "Chrom_<str>" and "ChromPos_<str>" fields. In some cases, the AlleleSequence will have been reported to be mapped to multiple assemblies in the singlerow CSV. The {-A} option specifies which assembly to report in the CHROM and POS fields in the output VCF file. If no assembly is specified, the loci are reported in the VCF as unmapped.
2. If a reference fasta file is given (via {-R}) an assembly identifier (via {-A}) must also be given. Additionally, the sequence names in the reference must agree with those reported in the input singlerow CSV file.

MapTK dart2vcf command

How do you know if dart2vcf ran correctly?



```
jess — lyons@ip-172-31-31-27:~ — ssh -i ~/Dropbox/Dartseq_to_mapping_workshop/lyons.pem lyons@ec2-13-57-194-80.us
[[lyons@ip-172-31-31-27 ~]$ 
Prototype mismatch: sub main::assert: none vs (&@) at /usr/local/lib/Exporter.pm line 66.
  at /usr/local/bin/maptk line 40.
2018-07-24 07:57:39,000 INFO [maptk] Starting Tue Jul 24 07:57:39 2018
2018-07-24 07:57:39,000 INFO [maptk] Command-line: 
2018-07-24 07:57:39,000 INFO [maptk] Sites processed: 0.0
2018-07-24 07:57:40,000 INFO [maptk] Sites processed: 
2018-07-24 07:57:40,000 INFO [maptk] Finished Tue Jul 24 07:57:40 2018
```

MapTK maptags command

```
jess — lyons@ip-172-31-31-27:~ — ssh -i ~/Dropbox/DArtseq_to_mapping_workshop/lyons.pen  
[lyons@ip-172-31-31-27 ~]$ maptk maptags  
Prototype mismatch: sub main::assert: none vs (&@) at /usr/local/lib/Exporter.pm line 66.  
at /usr/local/bin/maptk line 40.  
  
Usage:  maptk maptags [options] <in.vcf> <genome.fasta> <out-prefix>  
  
Options: -p <str>    INFO tag representing variant position in tag [SnpPosition]  
         -s <str>    INFO tag representing variant tag sequence [AlleleSequence]  
         -q <uint>   Minimum mapping quality for confidently mapped reads [60]  
         -h          This help document  
[  
Notes:  Arguments to {-p} and {-s} are case sensitive
```

Our reference genome: /files/reference.fasta

Sample DArT data: /files/Report_DY17-1234

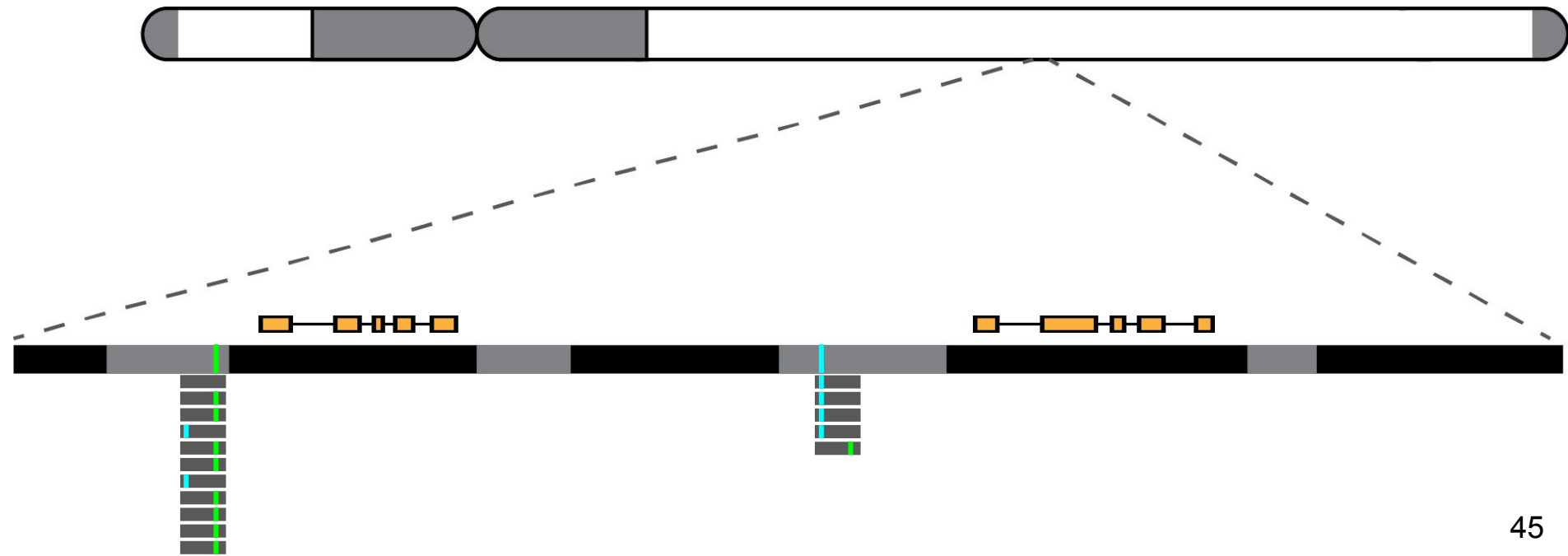
Copy to your home directory

MapTK maptags command

How do you know if maptags ran correctly?

```
jess — lyons@ip-172-31-31-27:~ — ssh -i ~/Dropbox/DArtseq_to_mapping_workshop/lyons.pem lyons@ec2-13-57-194-80.us-west-1.compu...  
[[lyons@ip-172-31-31-27 ~]$  
[lyons@ip-172-31-31-27 ~]$   
[Prototype mismatch: sub main::assert: none vs (&@) at /usr/local/lib/Exporter.pm line 66.  
at /usr/local/bin/maptk line 40.  
2018-07-24 08:35:11,000 INFO [maptk] Starting Tue Jul 24 08:35:11 2018  
2018-07-24 08:35:11,000 INFO [maptk] Command-line:   
2018-07-24 08:35:11,000 INFO [maptk] Extracting variants tags from VCF  
2018-07-24 08:35:11,000 INFO [maptk] Sites processed: 0.0  
2018-07-24 08:35:12,000 INFO [maptk] Mapping variants tags  
2018-07-24 08:35:12,000 INFO [maptk] Combining variants tags into VCF  
2018-07-24 08:35:12,000 INFO [maptk] Sites processed:   
2018-07-24 08:35:12,000 INFO [maptk] Sites processed:   
2018-07-24 08:35:12,000 INFO [maptk] Finished Tue Jul 24 08:35:12 2018
```

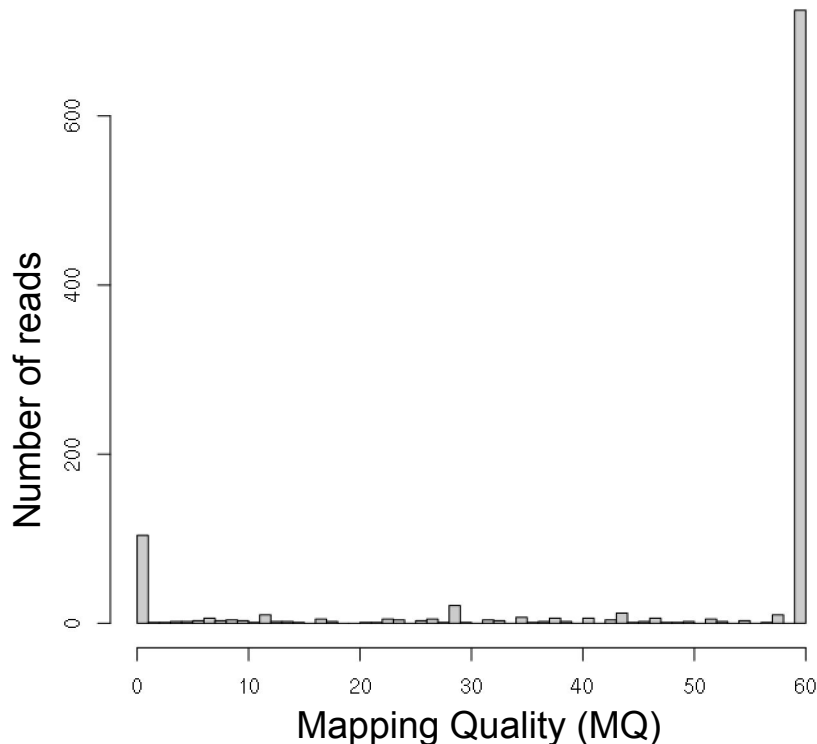
Genomic repeats frustrate proper sequence mapping



Mapping Quality: Scoring mapped sequences

- Range: 0–60
- MQ = 60: Read uniquely mapped
- MQ < 60: Has a sub-optimal mapping
- MQ = 0: read has two equivalently-scored mapped positions
- MQ score affected by fraction of mismatched bases in alignment

Phred = $-10 \log_{10}(\text{Prob})$

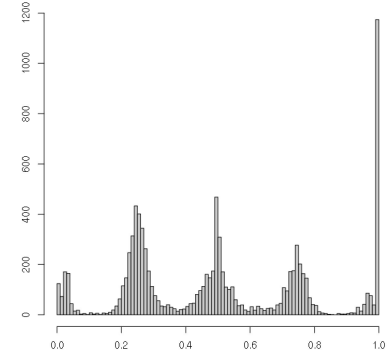


Filtering loci for segregating markers

χ^2 test for Mendelian segregation

Goodness-of-fit test:

- Expected Ratios vs. Observed Ratios
- “Test statistics that follow a chi-squared distribution arise from an assumption of independent normally distributed data” – Wikipedia



	A	a
A	AA	Aa
a	Aa	aa

1 AA : 2 Aa : 1 aa

	A	a
A	AA	Aa
A	AA	Aa

1 AA : 1 Aa

	a	a
A	Aa	Aa
a	aa	aa

1 Aa : 1 aa

MapTK chif1: UI

```
jess — lyons@ip-172-31-31-27:~ — ssh -i ~/Dropbox/DARTseq_to_mapping_workshop/lyons.per
```

```
[lyons@ip-172-31-31-27 ~]$ maptk chif1
Prototype mismatch: sub main::assert: none vs (&@) at /usr/local/lib/Exporter.pm line 66.
at /usr/local/bin/maptk line 40.
```

```
Usage: maptk chif1 [options] <in.vcf>
```

```
Options: -o <file>      Write output to FILE [stdout]
        -A <ufloat>    Max. phred allele freq. best-fit P-val [50.0]
        -G <ufloat>    Max. phred genotype freq. best-fit P-val [9999.0]
        -P <P1>[,<P2>] Parental IDs for the input population (recommended)
        -e <uint>      Phred-scaled genotype-call error (see Notes 3) [30]
        -a             Perform chi-sqr on allele depths (AD) field [GT]
        -F             Exclude sites for which scores cannot be applied
        -p             Force pseudo-testcross markers only
        -S             Silence/disable verbose reporting
        -t             Write output as a table (see Note 4) [VCF]
        -h             This help document
```

Notes:

1. This script applies a chi-squared goodness-of-fit test for Mendelian genotype frequencies, for a determined allele frequency class (3:1, 1:1, etc). The allele frequency class is selected within a goodness-of-fit tolerance threshold set by the '-A' option. The chi-squared values calculated here have been compared to those calculated by JoinMap4 and there is (almost) complete agreement.
2. Parental genotypes are inferred and stored by the 'P0GT' key in the INFO field. If the IDs of the parents are passed via the '-P' option, and the samples are included in the input VCF, an attempt to orient the inferred genotype calls with respect to parent (determined by the relative order of the parental samples listed in the VCF header) is made. If successful, a 'P0PHASED' key is applied to the INFO field.
3. Requires at least five diploid F1 samples (ten or more recommended).
4. Yates' correction for continuity is applied to sites with less than ten observed chromosomes, and no calculations are performed on sites with less than five observed chromosomes.
5. Value passed to the '-e' argument must be a positive integer. It is recommended to set '-e' to the minimum GQ value use for filtering.
6. When enabling the '-a' flag, and the input data are at low-coverage, it is necessary to include the parental IDs via '-P' to calculate the P-value statistics accurately.
7. The default output format is VCF, the '-t' option outputs a tab-delimited table (list of sub-field values) of relevant statistics a-la 'vcftools --get-INFO'.

MapTK chif1

How do you know if chif1 ran correctly?

```
jess — lyons@ip-172-31-31-27:~ — ssh -i ~/Dropbox/DArtseq_to_mapping_workshop/lyons.pem lyons@ec2-13-57-194-80.us-west-1.compu...

[lyons@ip-172-31-31-27 ~]$ [REDACTED]
Prototype mismatch: sub main::assert: none vs (&@) at /usr/local/lib/Exporter.pm line 66.
  at /usr/local/bin/maptk line 40.
2018-07-24 09:30:38,000 INFO [maptk] Starting Tue Jul 24 09:30:38 2018
2018-07-24 09:30:38,000 INFO [maptk] Command-line: [REDACTED]
2018-07-24 09:30:38,000 INFO [maptk] INFO-P0GT phasing: enabled
2018-07-24 09:30:38,000 INFO [maptk] Output format: VCF
2018-07-24 09:30:38,000 INFO [maptk] Sites processed: 0.0
2018-07-24 09:30:39,000 INFO [maptk] Sites processed: [REDACTED]
2018-07-24 09:30:39,000 INFO [maptk] Input samples: [REDACTED]
2018-07-24 09:30:39,000 INFO [maptk] AF-filtered sites: [REDACTED] ([REDACTED])
2018-07-24 09:30:39,000 INFO [maptk] INFO-P0GT imputed: [REDACTED] ([REDACTED])
2018-07-24 09:30:39,000 INFO [maptk] INFO-P0GT phased: [REDACTED] ([REDACTED])
2018-07-24 09:30:39,000 INFO [maptk] Finished Tue Jul 24 09:30:39 2018
```

MapTK chif1: test statistics

- **F1AFP**: Phred-scaled P-value for X^2 goodness-of-fit test on locus allele frequencies
- **F1GTP**: Phred-scaled P-value for X^2 goodness-of-fit test on genotype frequencies
- **F1X2**: X^2 value for goodness-of-fit test on locus allele frequencies
- **P0GT**: The Inferred parental genotypes, ordered alphabetically by Sample ID
- **P0PHASED**: Boolean tag indicating the locus passed the specified X^2 test

Phred = $-10 \log_{10}(\text{Prob})$