Hands-on analysis of DArTseq data for linkage mapping

Jessen Bredeson
and
Jess Lyons

Research to Nourish Africa

Points of order

- Please be on time and minimize disruptions

- GitHub site is home base for everything pertaining to the course
  - https://github.com/bredeson/HandsOnDArT
  - Current agenda, resources, exercises, *etc*.
  - Check there first!

- Course goals
  - Understand DArTseq data, and use it for mapping
  - Strengthen UNIX skills in an applied context
  - Teach teachers

On your index card:

Name, and preferred name

Research focus

What will you be using DArTseq data for?

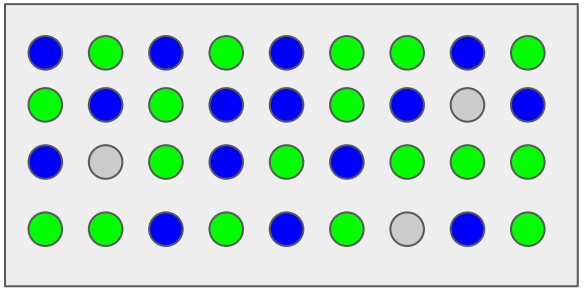
Experience working with genotyping data/mapping

GBS/DArTseq overview

DArT genotyping platforms

1. DArT (traditional): Oligo hybridization array chip

- Restriction digestion + bacterial cloning
- Relative fluorescent color signal
- Presence/Absence
- 100s–1,000s of markers



2. DArTseq: Reduced representation sequencing

- Restriction enzyme double-digest
- Methylation sensitivity
- Markers biased to genic portions of genome
- 1,000s–10,000s of markers

GBS/DArTseq library construction

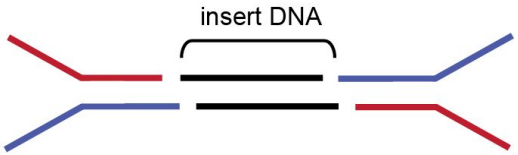
Restriction digest

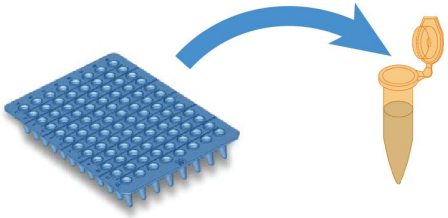
```
ApeKI: G¦C W G C
       C G W C¦G

W= A or T
```

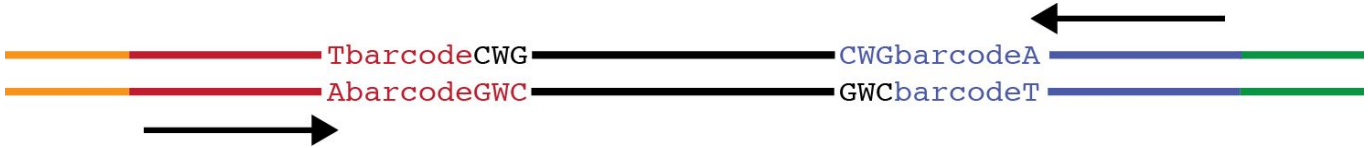
Ligate barcoded adapters

```
5′ ACACTCTTTCCCTACACGACGCTCTTCCGATCTbarcode   3′
                                              5′
3′ GAGCCGTAAGGACGACTTGGCGAGAAGGCTAGAbarcodeGWC
```



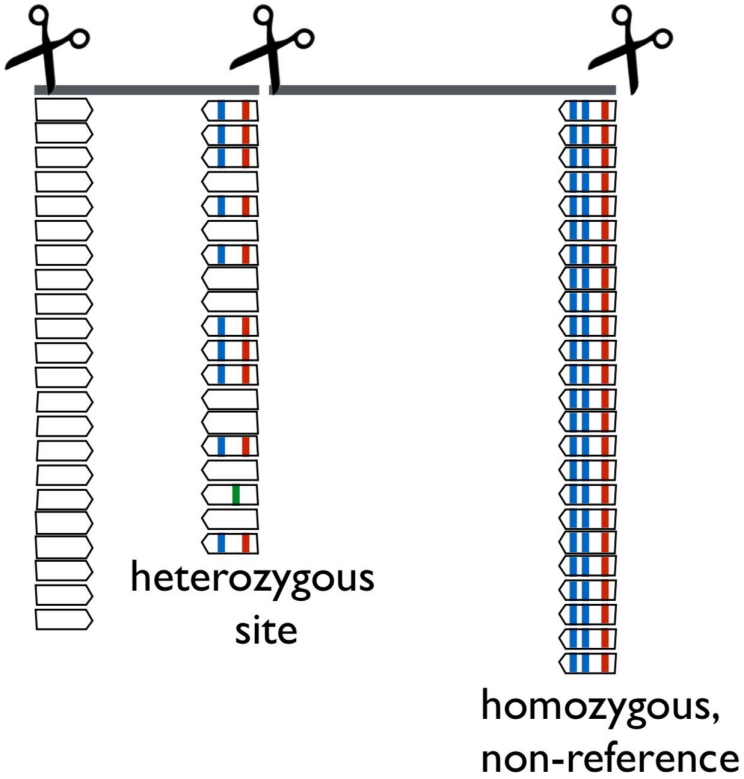


Pool samples
PCR



Sequencing

GBS/DArTseq variant sampling



Diversity Arrays Technology

IGSS Africa

DArTseq data types

1. SNP: Nucleotide differences observed in tag sequences

- “Co-dominant” genotypes:

```
Aa x aa => 1 Aa : 1 aa
Aa x Aa => 1 AA : 2 Aa : 1 aa
AA x Aa => 1 AA : 1 Aa
```

0	aa
1	Aa
2	AA
-	No data

2. SilicoDArTs:

- Presence/absence variation
- “Dominant” genotypes:

G^1 x G^0 : Aa x aa => 1 Aa : 1 aa => 1 G^1 : 1 G^0
: AA x aa => 2 Aa : 0 aa => 2 G^1 : 0 G^0

0	A allele not present
1	A allele present
-	No data

DArT bioinformatic analysis

- Data generation:
  - Libraries sequenced using Illumina HiSeq 2000/2500
  - Sequencing reads filtered for 90% confidence over half of read
- Genome-independent variants and genotyping calling
  - Reads clustered to form ~77 bp marker sequence tags
  - SNPs called as differences between tags
- Allows markers to be mapped/re-mapped to a genomic sequence
  - No need to re-map all reads and re-call SNPs

A brief UNIX review

Useful UNIX commands for working with files

Tools specifically useful for looking at files/directories:

1. Navigation: ls, cd, pwd
   relative vs. absolute paths
2. File viewing: less, more, head, tail, od, column
3. File manipulation: mv, cp, rm, mkdir, grep, cat, cut, tr, sed
4. File compression and archiving: gzip, bzip2, zip, tar
5. File transfer: scp, wget, curl
6. Getting help: man, apropos

```
/
├── files/
└── home/
    ├── bredeson/
    └── lyons/
        └── cuffdiff.txt
```

Absolute: `ls /home/lyons/cuffdiff.txt`
Relative: `ls ./cuffdiff.txt`

Common cross-platform file issues

- Invisible characters:
  - CR/LF/CRLF
  - Tabs vs spaces
- Special characters:
  - File encoding (ASCII vs. Unicode)

Carriage Return (CR)	Old Mac OS versions	\r
Line Feed (LF)	MacOS X, UNIX, LINUX	\n
CRLF	Windows	\r\n

!!!

Invisible characters

od -c is your friend!

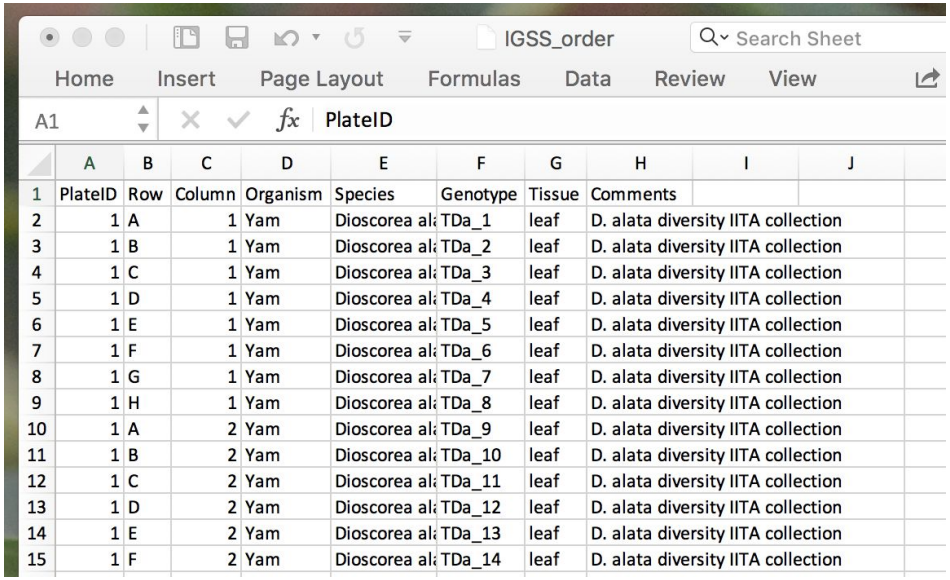
jess — lyons@ip-172-31-31-27:~ — ssh -i ~/Dropbox/DArtseq_to_mapping_wo...

```
0007740  A  T  A  C  A  T  C  T  T  T  T  A  A  T  T  C
0007760  A  T  C  T  A  G  C  G  G  A  A  T  A  T  T  G
0010000  T  C  A  A  A  A  A  G  T  T  C  A  G  A  A  T
0010020  A  T  C  G  A  \n C  G  G  A  G  G  G  C  A  A
0010040  A  A  T  T  G  T  T  C  A  G  C  G  C  C  A  C
0010060  A  C  A  T  A  G  G  A  T  T  C  A  C  A  G  A
0010100  T  C  A  G  A  T  G  A  A  A  C  A  A  A  G  T
0010120  T  G  T  C  T  G  C  T  G  T  T  A  A  A  G  C
0010140  T  A  A  G  A  A  \n C  C  A  C  A  A  A  A  T
0010160  T  T  T  C  T  G  A  G  T  A  T  A  G  T  G  T
0010200  T  C  C  T  G  C  A  A  G  G  C  T  T  T  T  T
0010220  T  T  A  T  T  T  T  T  T  T  T  A  T  T  T  T
0010240  T  T  G  C  T  T  T  C  T  T  G  T  G  A  A  T
0010260  G  T  T  T  G  C  A  \n G  C  T  T  T  G  G  T
0010300  T  G  T  G  T  C  C  G  T  T  G  T  T  G  T  T
0010320  C  T  C  T  T  G  G  C  A  A  C  T  T  C  T  A
0010340  G  T  G  T  T  T  G  C  A  T  G  T  A  T  T  G
0010360  C  T  A  T  A  G  G  C  T  G  T  C  T  G  G  T
0010400  T  A  C  T  G  C  A  T  \n T  G  T  T  T  C  C
```

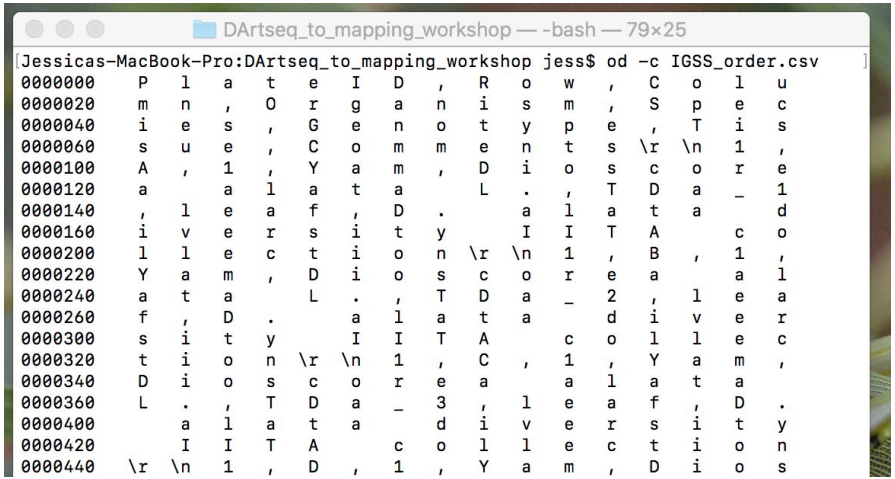
jess — lyons@ip-172-31-31-27:~/problemsets — ssh -i ~/Dropbox/DArtseq

```
2605320  1  \t 1  \t n  o  \n G  R  M  Z  M  2  G  1  5
2605340  8  5  6  8  \t G  R  M  Z  M  2  G  1  5  8  5
2605360  6  8  \t G  R  M  Z  M  2  G  1  5  8  5  6  8
2605400  \t C  h  r  5  :  2  0  3  3  1  3  4  9  3  -
2605420  2  0  3  3  1  5  9  6  9  \t B  7  3  .  s  \t
2605440  M  o  1  7  .  s  \t O  K  \t 5  9  9  .  4  7
2605460  8  \t 1  7  0  .  1  6  3  \t -  1  .  2  5  9
2605500  3  \t 0  .  6  4  8  6  4  4  \t 0  .  5  1  6
2605520  5  6  9  \t 0  .  7  9  8  9  9  2  \t n  o  \n
2605540  G  R  M  Z  M  2  G  1  5  8  5  7  5  \t G  R
2605560  M  Z  M  2  G  1  5  8  5  7  5  \t G  R  M  Z
2605600  M  2  G  1  5  8  5  7  5  \t C  h  r  2  :  1
2605620  6  1  5  9  7  3  2  8  -  1  6  1  6  0  9  2
2605640  7  2  \t B  7  3  .  s  \t M  o  1  7  .  s  \t
2605660  O  K  \t 1  1  .  4  3  1  3  \t 1  6  .  1  8
2605700  5  8  \t 0  .  3  4  7  7  8  \t -  0  .  4  3
2605720  1  3  5  9  \t 0  .  6  6  6  2  0  8  \t 0  .
2605740  9  0  1  2  7  5  \t n  o  \n G  R  M  Z  M  2
2605760  G  1  5  8  6  2  7  \t G  R  M  Z  M  2  G  1
2606000  5  8  6  2  7  \t G  R  M  Z  M  2  G  1  5  8
2606020  6  2  7  \t C  h  r  6  :  9  6  6  7  8  6  5
2606040  0  -  9  6  6  8  4  4  6  3  \t B  7  3  .  s
2606060  \t M  o  1  7  .  s  \t O  K  \t 2  0  .  3  4
2606100  0  6  \t 4  3  .  5  9  0  4  \t 0  .  7  6  2
```

Invisible characters



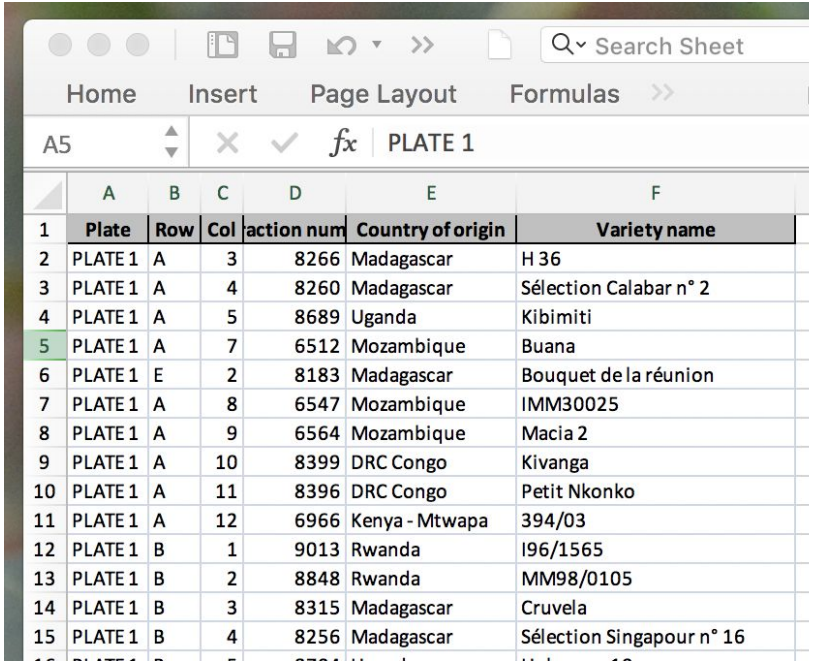
	A	B	C	D	E	F	G	H	I	J
1	PlateID	Row	Column	Organism	Species	Genotype	Tissue	Comments		
2	1	A	1	Yam	Dioscorea al	TDa_1	leaf	D. alata diversity IITA collection		
3	1	B	1	Yam	Dioscorea al	TDa_2	leaf	D. alata diversity IITA collection		
4	1	C	1	Yam	Dioscorea al	TDa_3	leaf	D. alata diversity IITA collection		
5	1	D	1	Yam	Dioscorea al	TDa_4	leaf	D. alata diversity IITA collection		
6	1	E	1	Yam	Dioscorea al	TDa_5	leaf	D. alata diversity IITA collection		
7	1	F	1	Yam	Dioscorea al	TDa_6	leaf	D. alata diversity IITA collection		
8	1	G	1	Yam	Dioscorea al	TDa_7	leaf	D. alata diversity IITA collection		
9	1	H	1	Yam	Dioscorea al	TDa_8	leaf	D. alata diversity IITA collection		
10	1	A	2	Yam	Dioscorea al	TDa_9	leaf	D. alata diversity IITA collection		
11	1	B	2	Yam	Dioscorea al	TDa_10	leaf	D. alata diversity IITA collection		
12	1	C	2	Yam	Dioscorea al	TDa_11	leaf	D. alata diversity IITA collection		
13	1	D	2	Yam	Dioscorea al	TDa_12	leaf	D. alata diversity IITA collection		
14	1	E	2	Yam	Dioscorea al	TDa_13	leaf	D. alata diversity IITA collection		
15	1	F	2	Yam	Dioscorea al	TDa_14	leaf	D. alata diversity IITA collection		




```
Jessicas-MacBook-Pro:DArtseq_to_mapping_workshop jess$ od -c IGSS_order.csv
0000000   P   l   a   t   e   I   D   ,   R   o   w   ,   C   o   l   u
0000020   m   n   ,   O   r   g   a   n   i   s   m   ,   S   p   e   c
0000040   i   e   s   ,   G   e   n   o   t   y   p   e   ,   T   i   s
0000060   s   u   e   ,   C   o   m   m   e   n   t   s  \r  \n   1   ,
0000100   A   ,   1   ,   Y   a   m   ,   D   i   o   s   c   o   r   e
0000120   a       a   l   a   t   a       L   .   ,   T   D   a   _   1
0000140   ,   l   e   a   f   ,   D   .       a   l   a   t   a       d
0000160   i   v   e   r   s   i   t   y       I   I   T   A       c   o
0000200   l   l   e   c   t   i   o   n  \r  \n   1   ,   B   ,   1   ,
0000220   Y   a   m   ,   D   i   o   s   c   o   r   e   a       a   l
0000240   a   t   a       L   .   ,   T   D   a   _   2   ,   l   e   a
0000260   f   ,   D   .       a   l   a   t   a       d   i   v   e   r
0000300   s   i   t   y       I   I   T   A       c   o   l   l   e   c
0000320   t   i   o   n  \r  \n   1   ,   C   ,   1   ,   Y   a   m   ,
0000340   D   i   o   s   c   o   r   e   a       a   l   a   t   a
0000360   L   .   ,   T   D   a   _   3   ,   l   e   a   f   ,   D   .
0000400       a   l   a   t   a       d   i   v   e   r   s   i   t   y
0000420       I   I   T   A       c   o   l   l   e   c   t   i   o   n
0000440  \r  \n   1   ,   D   ,   1   ,   Y   a   m   ,   D   i   o   s
```

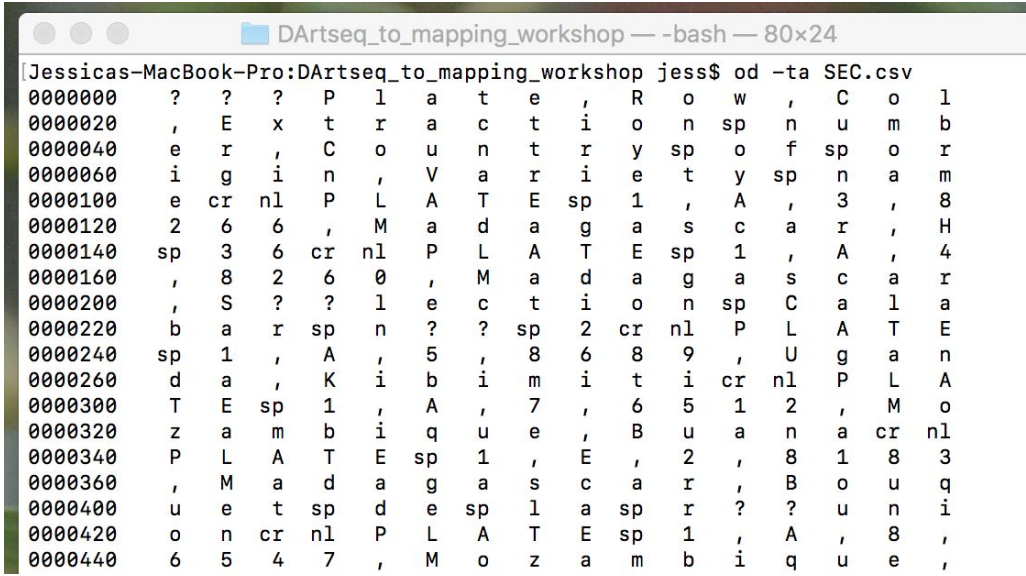
.csv file made on a mac

Special characters



A5 | PLATE 1

	A	B	C	D	E	F
1	Plate	Row	Col	action num	Country of origin	Variety name
2	PLATE 1	A	3	8266	Madagascar	H 36
3	PLATE 1	A	4	8260	Madagascar	Sélection Calabar n° 2
4	PLATE 1	A	5	8689	Uganda	Kibimiti
5	PLATE 1	A	7	6512	Mozambique	Buana
6	PLATE 1	E	2	8183	Madagascar	Bouquet de la réunion
7	PLATE 1	A	8	6547	Mozambique	IMM30025
8	PLATE 1	A	9	6564	Mozambique	Macia 2
9	PLATE 1	A	10	8399	DRC Congo	Kivanga
10	PLATE 1	A	11	8396	DRC Congo	Petit Nkonko
11	PLATE 1	A	12	6966	Kenya - Mtwapa	394/03
12	PLATE 1	B	1	9013	Rwanda	I96/1565
13	PLATE 1	B	2	8848	Rwanda	MM98/0105
14	PLATE 1	B	3	8315	Madagascar	Cruvela
15	PLATE 1	B	4	8256	Madagascar	Sélection Singapour n° 16



```
DArtseq_to_mapping_workshop — -bash — 80×24
[Jessicas-MacBook-Pro:DArtseq_to_mapping_workshop jess$ od -ta SEC.csv
0000000    ?   ?   ?   P   l   a   t   e   ,   R   o   w   ,   C   o   l
0000020    ,   E   x   t   r   a   c   t   i   o   n  sp   n   u   m   b
0000040    e   r   ,   C   o   u   n   t   r   y  sp   o   f  sp   o   r
0000060    i   g   i   n   ,   V   a   r   i   e   t   y  sp   n   a   m
0000100    e  cr  nl   P   L   A   T   E  sp   1   ,   A   ,   3   ,   8
0000120    2   6   6   ,   M   a   d   a   g   a   s   c   a   r   ,   H
0000140   sp   3   6  cr  nl   P   L   A   T   E  sp   1   ,   A   ,   4
0000160    ,   8   2   6   0   ,   M   a   d   a   g   a   s   c   a   r
0000200    ,   S   ?   ?   l   e   c   t   i   o   n  sp   C   a   l   a
0000220    b   a   r  sp   n   ?   ?  sp   2  cr  nl   P   L   A   T   E
0000240   sp   1   ,   A   ,   5   ,   8   6   8   9   ,   U   g   a   n
0000260    d   a   ,   K   i   b   i   m   i   t   i  cr  nl   P   L   A
0000300    T   E  sp   1   ,   A   ,   7   ,   6   5   1   2   ,   M   o
0000320    z   a   m   b   i   q   u   e   ,   B   u   a   n   a  cr  nl
0000340    P   L   A   T   E  sp   1   ,   E   ,   2   ,   8   1   8   3
0000360    ,   M   a   d   a   g   a   s   c   a   r   ,   B   o   u   q
0000400    u   e   t  sp   d   e  sp   l   a  sp   r   ?   ?   u   n   i
0000420    o   n  cr  nl   P   L   A   T   E  sp   1   ,   A   ,   8   ,
0000440    6   5   4   7   ,   M   o   z   a   m   b   i   q   u   e   ,
```

Non-ASCII characters are not read properly

Common HTS data formats

Common HTS formats you may encounter

- FASTA
- FASTQ
- VCF
- SAM

FASTA sequence file format

"Greater-than" symbol (>) signifies a new sequence record

Sequence ID
(required)

Sequence description*
(optional)

(both on one line)

```
>AF140613.1 Manihot esculenta N-hydroxylating cytochrome P450 (CYP79D1) mRNA, complete cds
GTTCAGGGCATATCAATATGGCCATGAACGTCTCCACCACCATCGGTTTACTTAACGCCACCTCCTTCGC
CTCCTCCTCCTCCATCAACACGGTCAAGATCTTGTTCGTCACCCTCTTTATTTCCATTGTTAGTACTATT
GTAAAACTTCAAAAGAGTGCTGCTAACAAGGAAGGTAGCAAGAAACTCCCACTCCCTCCTGGCCCTACTC
CATGGCCACTCATCGGAAACATCCCGGAAATGATCCGGTACAGACCCACGTTTCGGTGGATTCACCAACT
CATGAAGGACATGAACACTGATATTTGTCTCATTCGTTTTGGAAGAACTAACTTTGTTCCTATAAGCTGT
CCTGTTCTTGCTCGTGAAATACTAAAAAAGAATGACGCTATCTTCT
```

Sequence string
May be "wrapped flush" to an arbitrary width, or not at all

*Everything after the first whitespace (one or more spaces or tabs) is interpreted as description

FASTQ sequence file format

"At" symbol (@) signifies a new sequence record (but may also be observed in the quality string

Sequence ID
(required)
(both on one line)

Sequence description*
(optional; Paired-end Illumina sequences will have 1 or 2 in the leading colon-separated fields)

```
@SN638:1322:HHFYVBCXY:2:2213:17643:60632 1:N:0:2
GTTGGTTTGTTGCTTGGTTTCATTCTATGTGTGCTTGATTGTTG
+
DDDDDIIIIIIIIIIIHIIIIIIIIIIIIIIIIIIIHIIGHHI
```

Sequence string
Not wrapped

Quality ID
Must start with a '+';
Typically left blank

Quality string
Not wrapped; Characters represent ASCII+*offset* transformed numeric values; see https://en.wikipedia.org/wiki/FASTQ_format for more info

*Everything after the first whitespace (one or more spaces or tabs) is interpreted as description

VCF: Variant Call Format

Tab-separated, columnar file

Declarative Meta information header line(s)
Starts with two pound symbols (##)

File body header line
Starts with one pound symbol (#)

```
##fileformat=VCFv4.2
##FILTER=<ID=PASS,Description="PASSes all quality filters">
##FORMAT=<ID=GT,Number=1,Type=String,Description="Genotype">
##contig=<ID=Chr1,length=26803966>
##contig=<ID=Chr2,length=24424175>
#CHROM	POS	ID	REF	ALT	QUAL	FILTER	INFO	FORMAT	Sample
Chr1	187607	.	A	G	34149.20	PASS	.	GT	1/1
Chr1	1597049	.	ACT	A,AC	21739.90	PASS	.	GT	0/2
```

SNP locus

Indel locus

Homozygous Alternate allele

Alternate alleles indices = 1,2,...

Reference allele index = 0

Heterozygous

Please visit https://samtools.github.io/hts-specs/VCFv4.2.pdf for more detailed information.

SAM: Sequence Alignment/Map format

Tab-separated, columnar file

SAM header section

Format version information

Sequence definition(s)

Read Group(s) information

```
@HD     VN:1.5  SO:coordinate
@SQ     SN:dh-d00009    LN:10034360
@SQ     SN:dh-h04487            LN:15235
@RG     ID:L10KB        LB:DEC10KB1    SM:Sample   PL:ILLUMINA
HWI-EAS19X:0:15:16:199   177  dh-d00009  942   0  18M      dh-h04487  1259     0
        TGGATGTATTTCATATTTA     DIHE@BGFIEDHEDIEEE#   AS:i:18    RG:Z:L10KB
```

SAM alignment section

Please visit https://samtools.github.io/hts-specs/SAMv1.pdf for more detailed information.

UNIX & File Format Exercises

1. If you have not already completed the exercises on the PFB2017 website, please do so first. If you have, proceed to the next item.
2. Log into the AWS server
3. Make a directory called “exercises” and navigate into it
4. Copy all files starting with “data” from the /files directory into your exercises directory.
5. Examine the contents of data1 (hint: use less).
   a. What format of file is it? Rename the file with the appropriate file extension.
   b. Is the contents of the file single-end or paired-end?
6. Examine the contents of data2.
   a. What format of file is it? Rename the file with the appropriate file extension.
   b. What is incorrect about its formatting?

UNIX & File Format Exercises

7. Examine the contents of data3 (hint: use man to look at the -S option).
    a. What format of file is it? Rename the file with the appropriate file extension.
    b. How many loci are there? How many SNPs? How many Indels?
8. Examine the contents of data4.
    a. What format of file is it?
    b. What is incorrect about this file (hint: see od)?
    c. From which operating system might it have come from?
    d. Correct the file using UNIX commands (hint: see tr)
9. Examine the contents of data5.
    a. The file is tab-delimited, but what is incorrect about this file?
    b. Correct the file using UNIX commands.