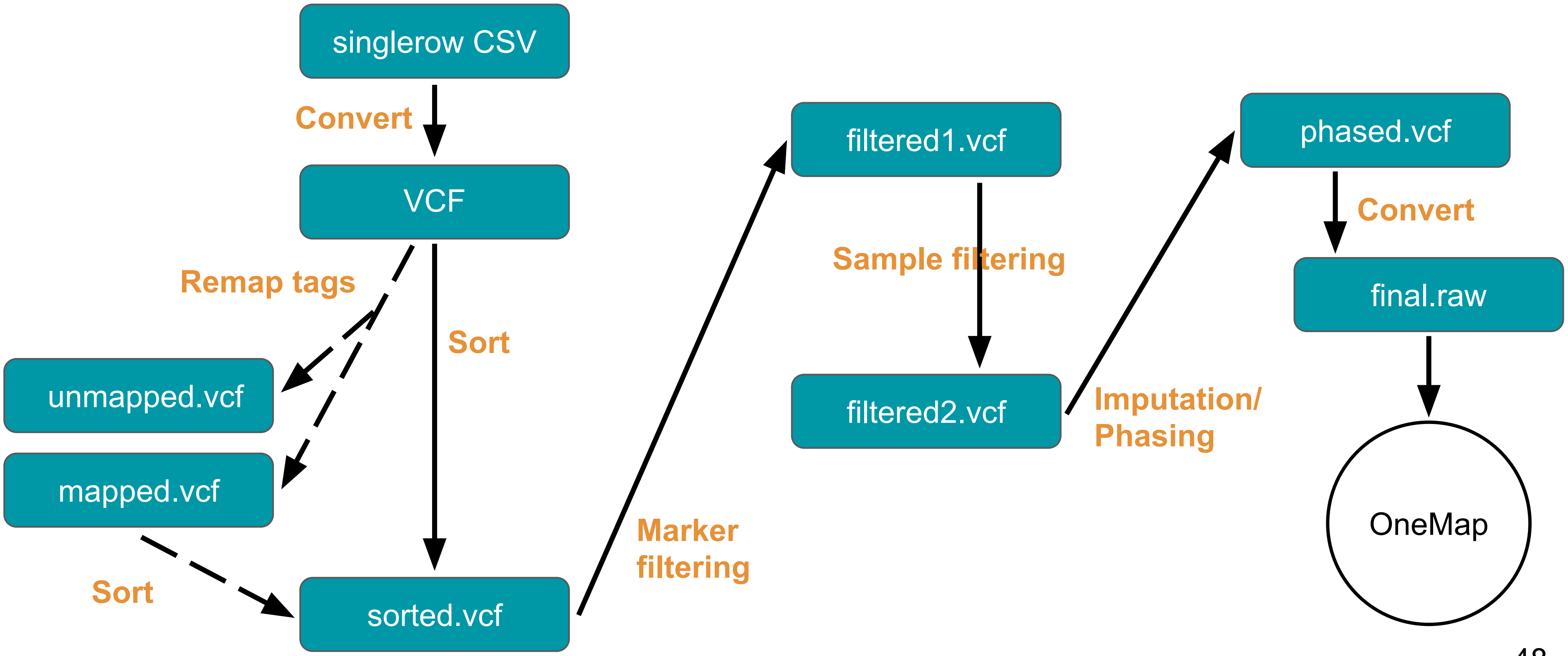
Wednesday, 25 July 2018

- Marker filtering
- Sample Filtering
- Imputation

Windows users will need WinSCP today

+ Daily reminder to copy your commands for later reference :-)

DArTseq to mapping: analysis flow chart



singlerow CSV

Convert

VCF

Remap tags

Sort

unmapped.vcf

mapped.vcf

Sort

sorted.vcf

Marker filtering

filtered1.vcf

Sample filtering

filtered2.vcf

Imputation/ Phasing

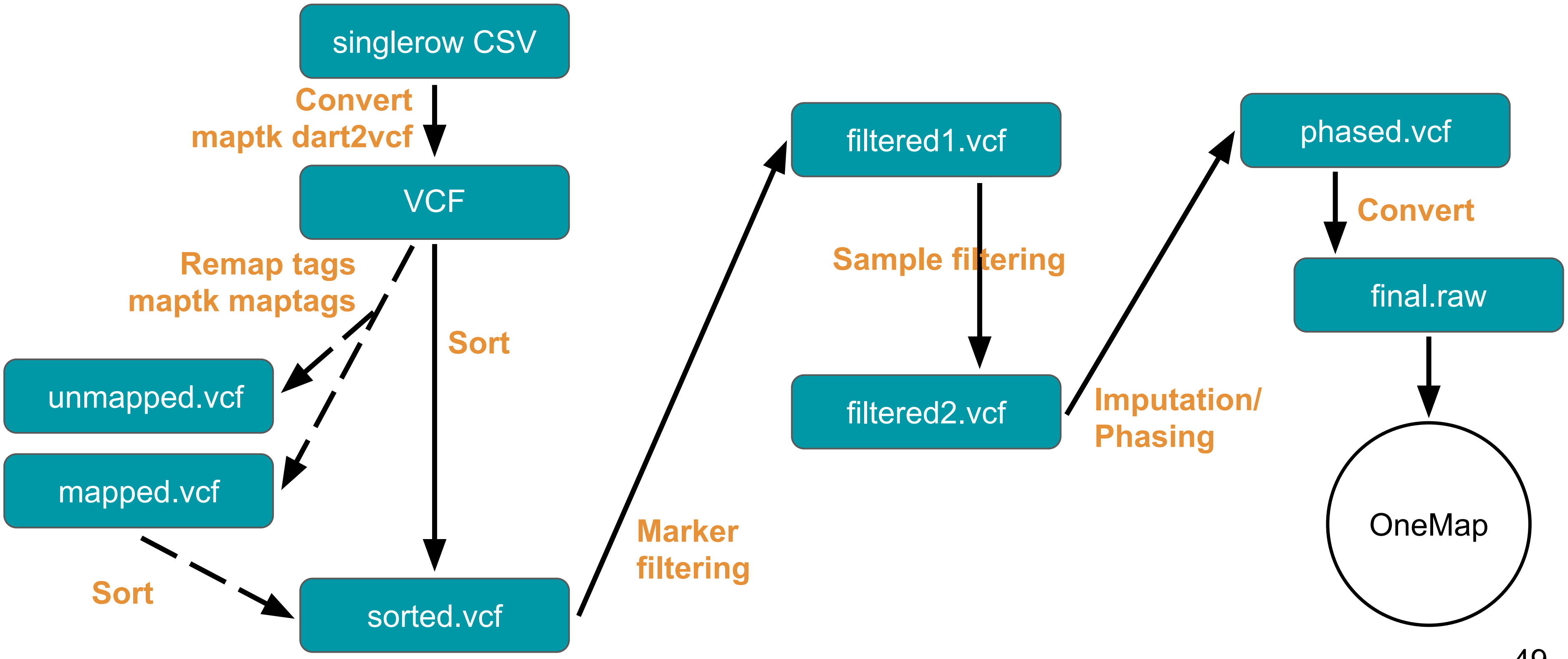
phased.vcf

Convert

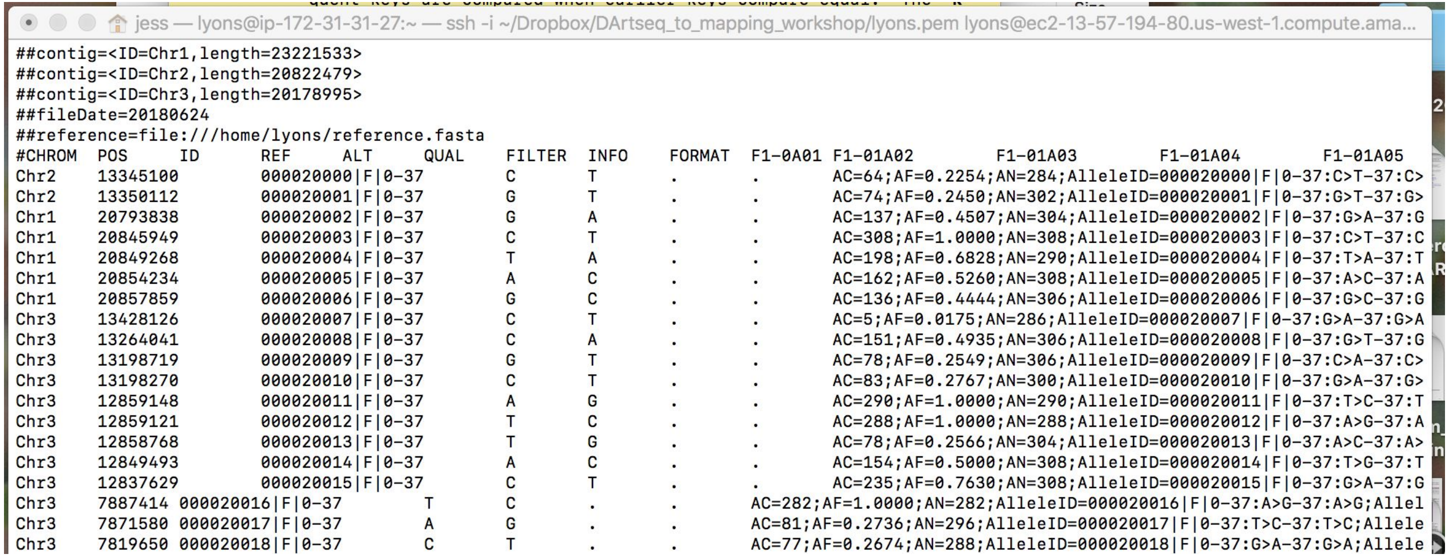
final.raw

OneMap

DArTseq to mapping: analysis flow chart



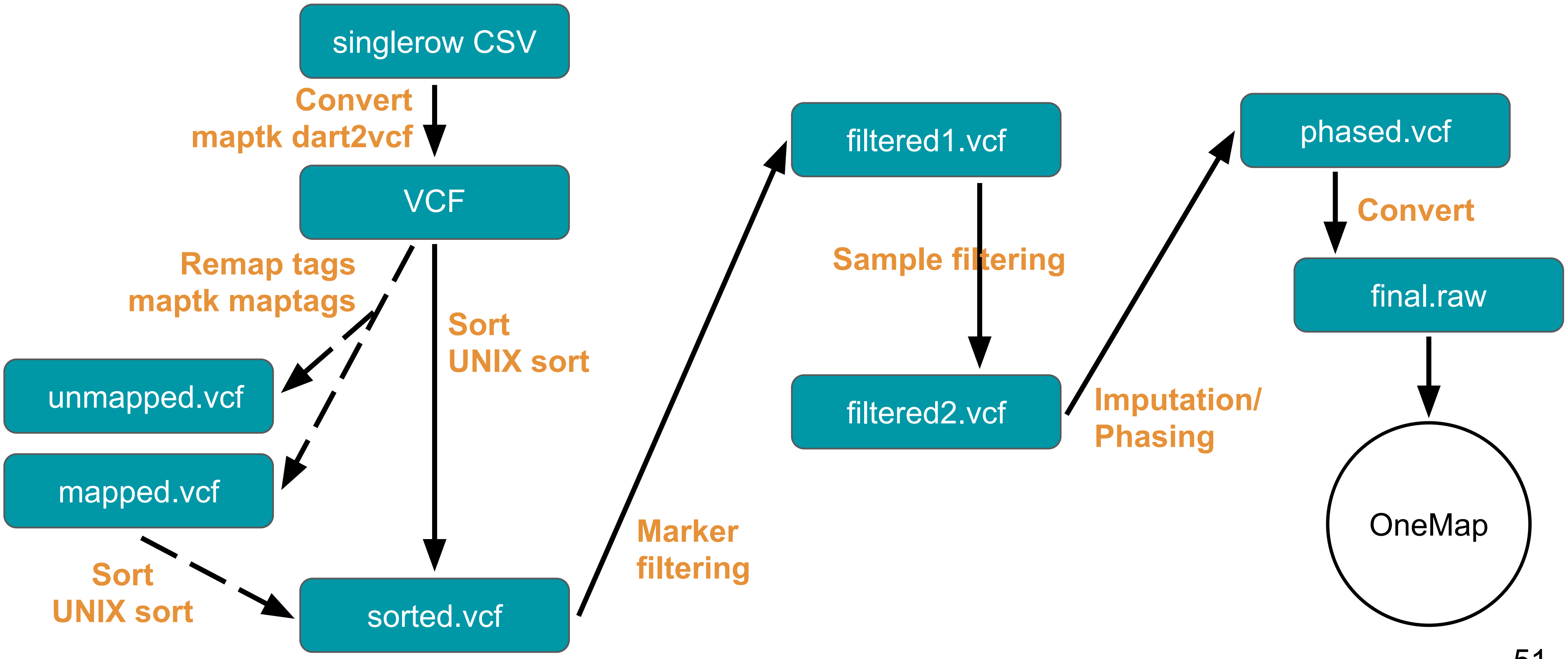
Preparing the (mapped) VCF file for filtering


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.compute.ama...

```
##contig=<ID=Chr1,length=23221533>
##contig=<ID=Chr2,length=20822479>
##contig=<ID=Chr3,length=20178995>
##fileDate=20180624
##reference=file:///home/lyons/reference.fasta
#CHROM  POS     ID      REF     ALT     QUAL    FILTER  INFO    FORMAT  F1-0A01 F1-01A02        F1-01A03        F1-01A04        F1-01A05
Chr2    13345100        000020000|F|0-37        C       T       .       .       AC=64;AF=0.2254;AN=284;AlleleID=000020000|F|0-37:C>T-37:C>
Chr2    13350112        000020001|F|0-37        G       T       .       .       AC=74;AF=0.2450;AN=302;AlleleID=000020001|F|0-37:G>T-37:G>
Chr1    20793838        000020002|F|0-37        G       A       .       .       AC=137;AF=0.4507;AN=304;AlleleID=000020002|F|0-37:G>A-37:G
Chr1    20845949        000020003|F|0-37        C       T       .       .       AC=308;AF=1.0000;AN=308;AlleleID=000020003|F|0-37:C>T-37:C
Chr1    20849268        000020004|F|0-37        T       A       .       .       AC=198;AF=0.6828;AN=290;AlleleID=000020004|F|0-37:T>A-37:T
Chr1    20854234        000020005|F|0-37        A       C       .       .       AC=162;AF=0.5260;AN=308;AlleleID=000020005|F|0-37:A>C-37:A
Chr1    20857859        000020006|F|0-37        G       C       .       .       AC=136;AF=0.4444;AN=306;AlleleID=000020006|F|0-37:G>C-37:G
Chr3    13428126        000020007|F|0-37        C       T       .       .       AC=5;AF=0.0175;AN=286;AlleleID=000020007|F|0-37:G>A-37:G>A
Chr3    13264041        000020008|F|0-37        C       A       .       .       AC=151;AF=0.4935;AN=306;AlleleID=000020008|F|0-37:G>T-37:G
Chr3    13198719        000020009|F|0-37        G       T       .       .       AC=78;AF=0.2549;AN=306;AlleleID=000020009|F|0-37:C>A-37:C>
Chr3    13198270        000020010|F|0-37        C       T       .       .       AC=83;AF=0.2767;AN=300;AlleleID=000020010|F|0-37:G>A-37:G>
Chr3    12859148        000020011|F|0-37        A       G       .       .       AC=290;AF=1.0000;AN=290;AlleleID=000020011|F|0-37:T>C-37:T
Chr3    12859121        000020012|F|0-37        T       C       .       .       AC=288;AF=1.0000;AN=288;AlleleID=000020012|F|0-37:A>G-37:A
Chr3    12858768        000020013|F|0-37        T       G       .       .       AC=78;AF=0.2566;AN=304;AlleleID=000020013|F|0-37:A>C-37:A>
Chr3    12849493        000020014|F|0-37        A       C       .       .       AC=154;AF=0.5000;AN=308;AlleleID=000020014|F|0-37:T>G-37:T
Chr3    12837629        000020015|F|0-37        C       T       .       .       AC=235;AF=0.7630;AN=308;AlleleID=000020015|F|0-37:G>A-37:G
Chr3    7887414 000020016|F|0-37        T       C       .       .       AC=282;AF=1.0000;AN=282;AlleleID=000020016|F|0-37:A>G-37:A>G;Allel
Chr3    7871580 000020017|F|0-37        A       G       .       .       AC=81;AF=0.2736;AN=296;AlleleID=000020017|F|0-37:T>C-37:T>C;Allele
Chr3    7819650 000020018|F|0-37        C       T       .       .       AC=77;AF=0.2674;AN=288;AlleleID=000020018|F|0-37:G>A-37:G>A;Allele
```

Why sort?

DArTseq to mapping: analysis flow chart



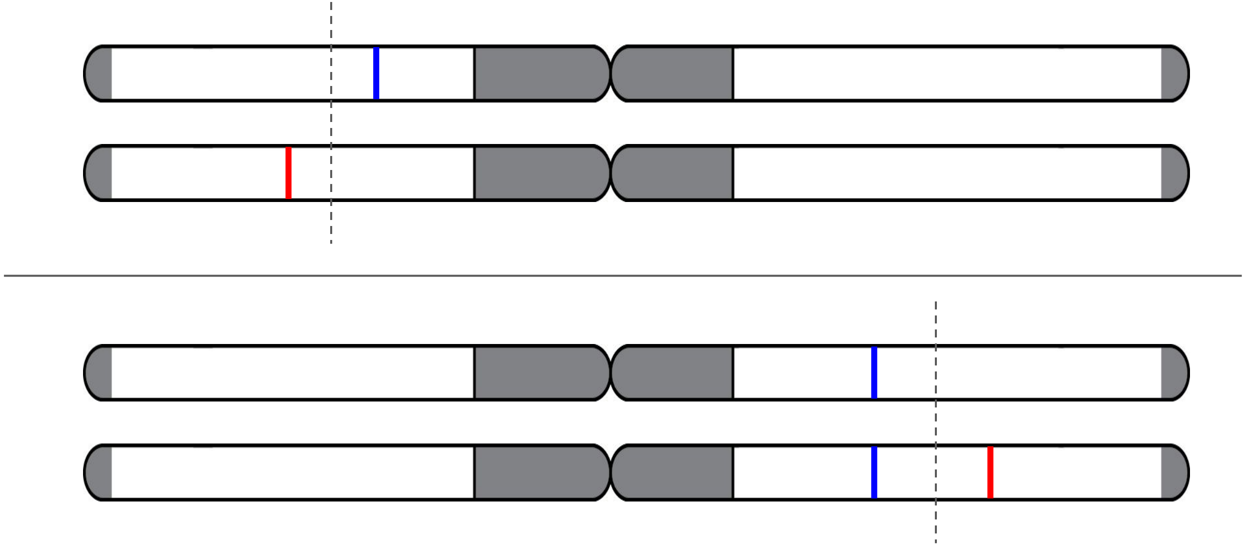
Filtering loci for segregating markers

What markers are useful for F1 linkage mapping?

- Useful markers have two or more alleles segregating in a Mendelian manner
- At least one of the parents must be heterozygous

	A	A
a	Aa	Aa
a	Aa	Aa

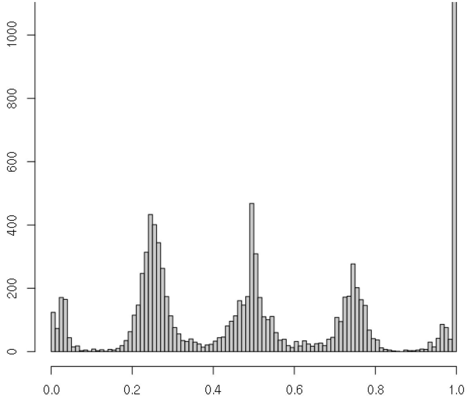
1 Aa



X^2 test for Mendelian segregation

Goodness-of-fit test:

- Expected counts vs. observed counts
- “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

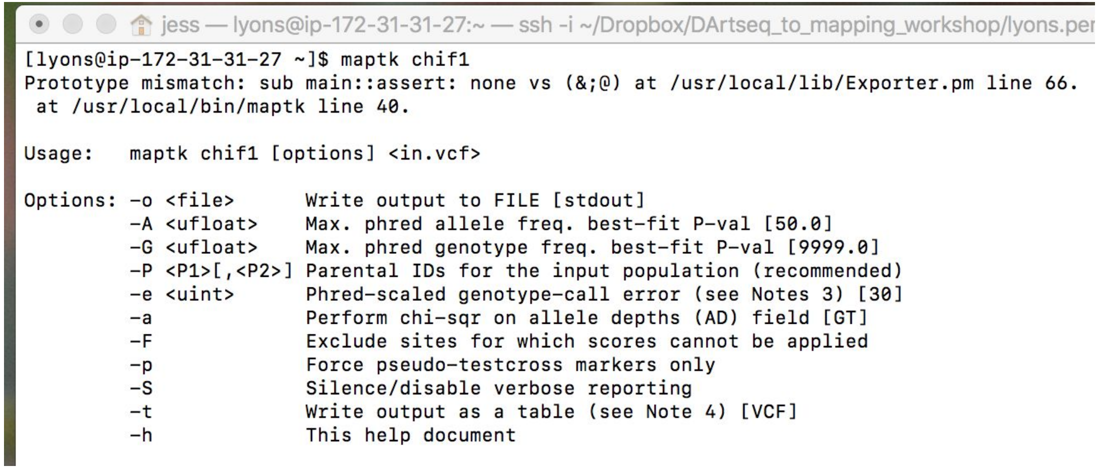
	A	a
A	AA	Aa
A	AA	Aa

1 AA : 1 Aa

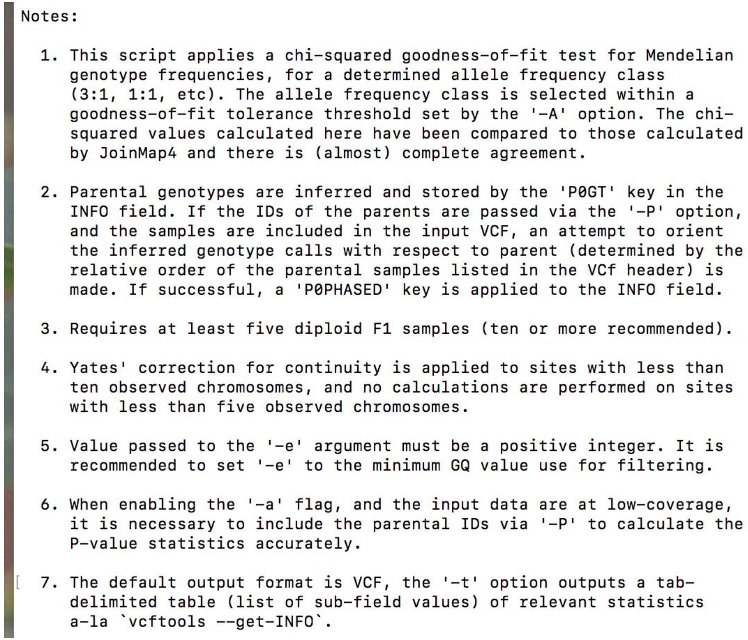
	a	a
A	Aa	Aa
a	aa	aa

1 Aa : 1 aa

MapTK chif1 command


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
```

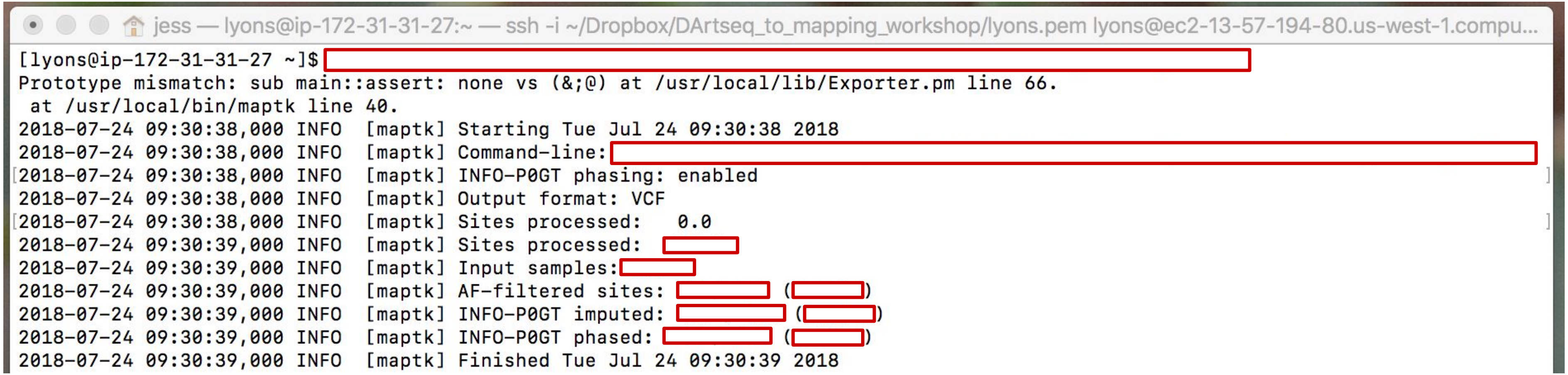


```
Notes:

  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`.
```

Note: this command requires all sites to have less than 50% missing data.
VCFtools can help you remove sites with >50% missing data.

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...

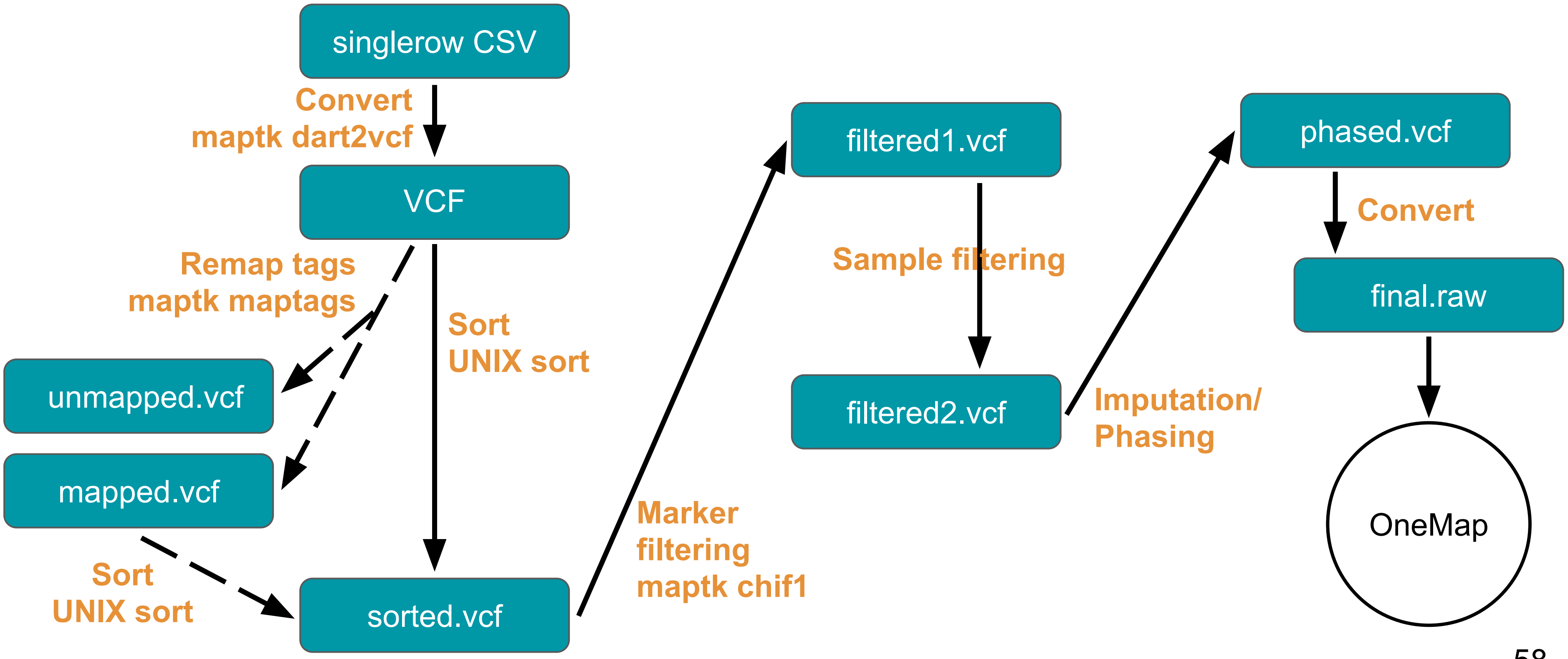
```
[lyons@ip-172-31-31-27 ~]$
Prototype mismatch: sub main::assert: none vs (&;@) at /usr/local/lib/Exporter.pm line 66.
 at /usr/local/bin/maptk line 40.
2018-07-24 09:30:38,000 INFO  [maptk] Starting Tue Jul 24 09:30:38 2018
2018-07-24 09:30:38,000 INFO  [maptk] Command-line:
2018-07-24 09:30:38,000 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:
2018-07-24 09:30:39,000 INFO  [maptk] Input samples:
2018-07-24 09:30:39,000 INFO  [maptk] AF-filtered sites:          (        )
2018-07-24 09:30:39,000 INFO  [maptk] INFO-P0GT imputed:           (        )
2018-07-24 09:30:39,000 INFO  [maptk] INFO-P0GT phased:           (        )
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 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 (and, if the option was enabled, the parental genotypes were able to be inferred)

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

DArTseq to mapping: analysis flow chart



Filtering samples for full-sib progeny

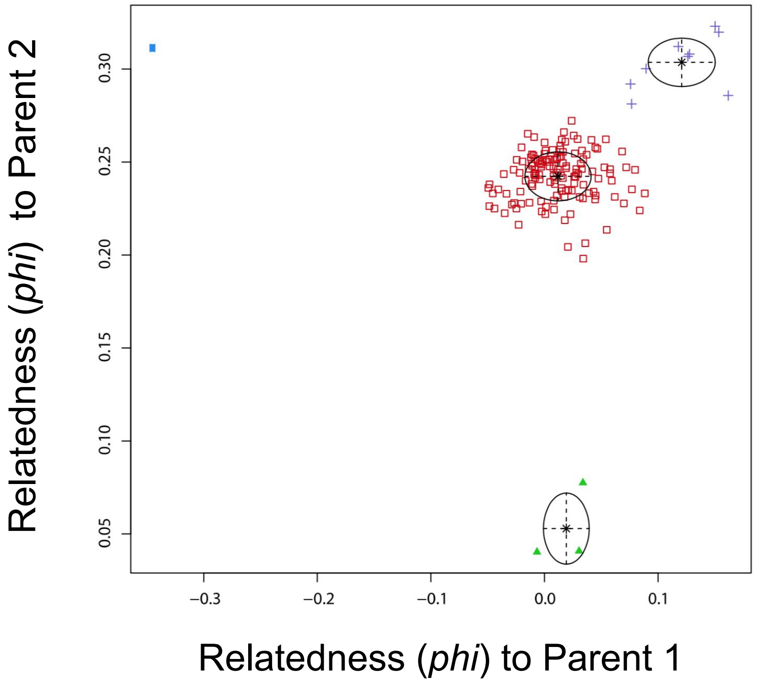
Sources of error in crossing experiments

1. Half-sibs
2. Selfs
3. Volunteer seedlings
4. Human error

Filtering samples for full-sib progeny

- Calculate relatedness* of putative offspring w.r.t. the parents.
- Relatedness *phi* about half of relatedness expected:

relationship	Expected	Reported ***phi***
clonal	1.0	0.5
parent–child	0.5	0.25
off-type	0.0	0.0



*Manichaikul. 2010. Bioinfo. doi: 10.1093/bioinformatics/btq559
ICGMC. 2015. G3 Journal. doi: 10.1534/g3.114.015008
https://bitbucket.org/rokhsar-lab/gbs-analysis/src/master/

Filtering samples for full-sib progeny

vcftools relatedness2

➔ Remember to check where it wants a file, and with a prefix

maptk mclust

➔ Check for files ending in .dat and .dat.pdf

➔ **Note**: mclust runs R to produce the plots, if the Mclust R package is not installed, R will not print the plots

Filtering samples for full-sib progeny

Download the plots to your laptop:

Mac/Linux:

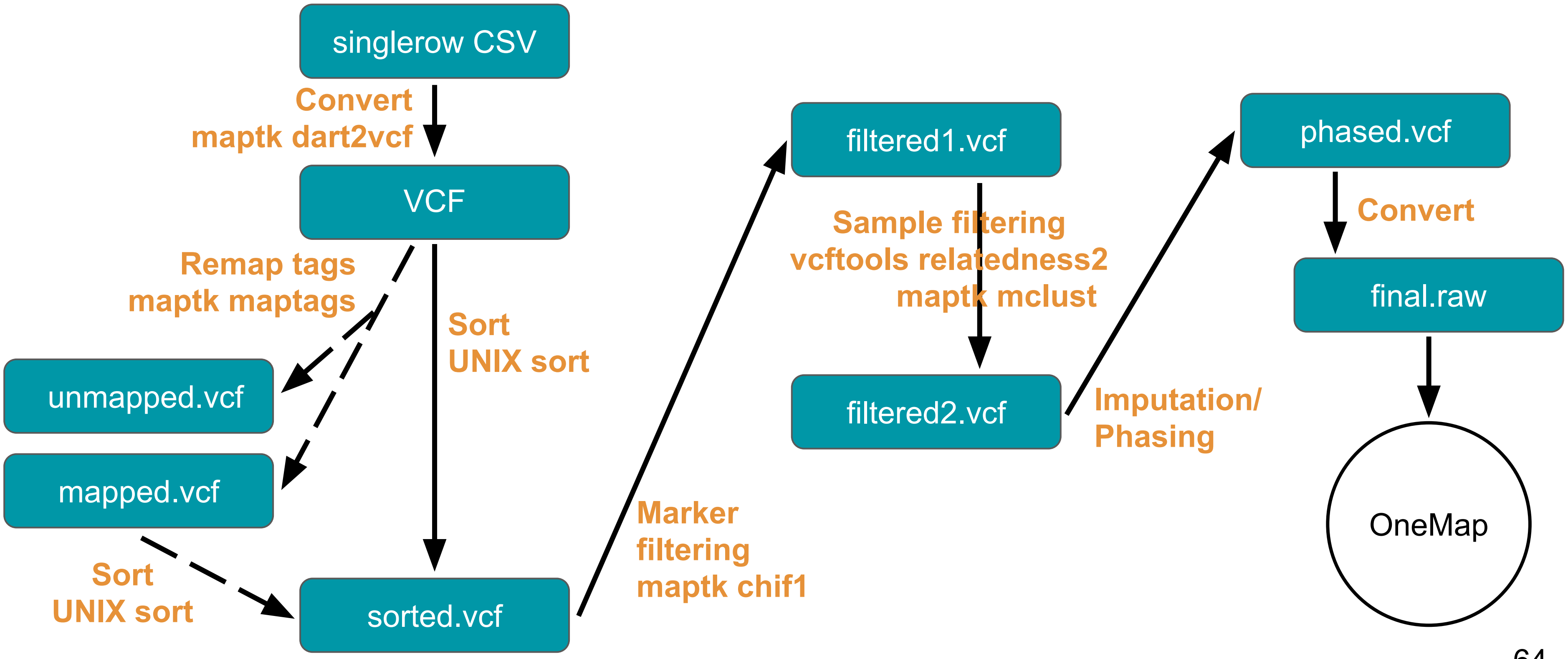
```
scp -i ~/fn.pem \
fn@ec2-13-57-194-80.us-west-1.compute.amazonaws.com:/home/fn/mclust.dat.pdf .
```

Where “fn” is your family name

Windows:

Download and install WinSCP

DArTseq to mapping: analysis flow chart



singlerow CSV

Convert
maptk dart2vcf

VCF

Remap tags
maptk maptags

unmapped.vcf

mapped.vcf

Sort
UNIX sort

Sort
UNIX sort

sorted.vcf

Marker
filtering
maptk chif1

filtered1.vcf

Sample filtering
vcftools relatedness2
maptk mclust

filtered2.vcf

Imputation/
Phasing

phased.vcf

Convert

final.raw

OneMap

Imputing genotypes

Imputation and phasing

What is imputation:

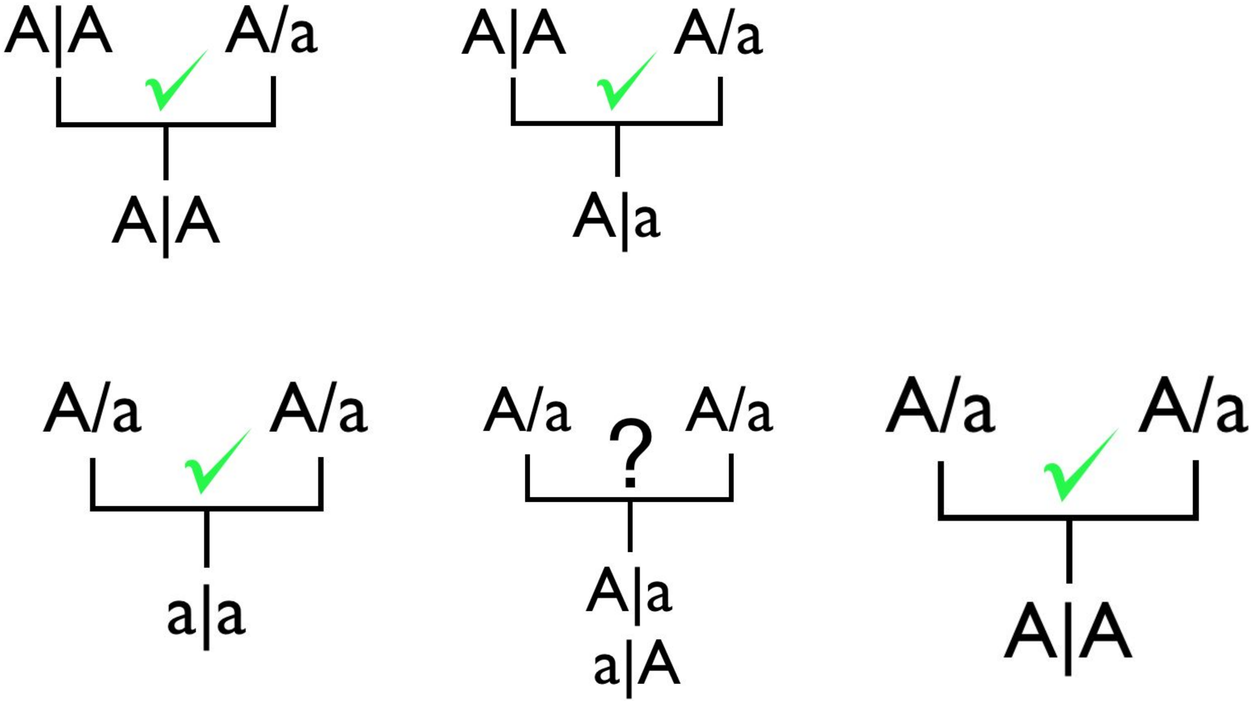
➔ Filling in missing values (*i.e.*, genotypes) with data inferred from patterns/correlations in the dataset.

What is phasing?

➔ Assigning alleles to their maternal or paternal chromosome.

phaseF1 script

Imputation and phasing



Imputation and phasing

Mapping algorithms are sensitive to missing data, but *more* sensitive to incorrect data.

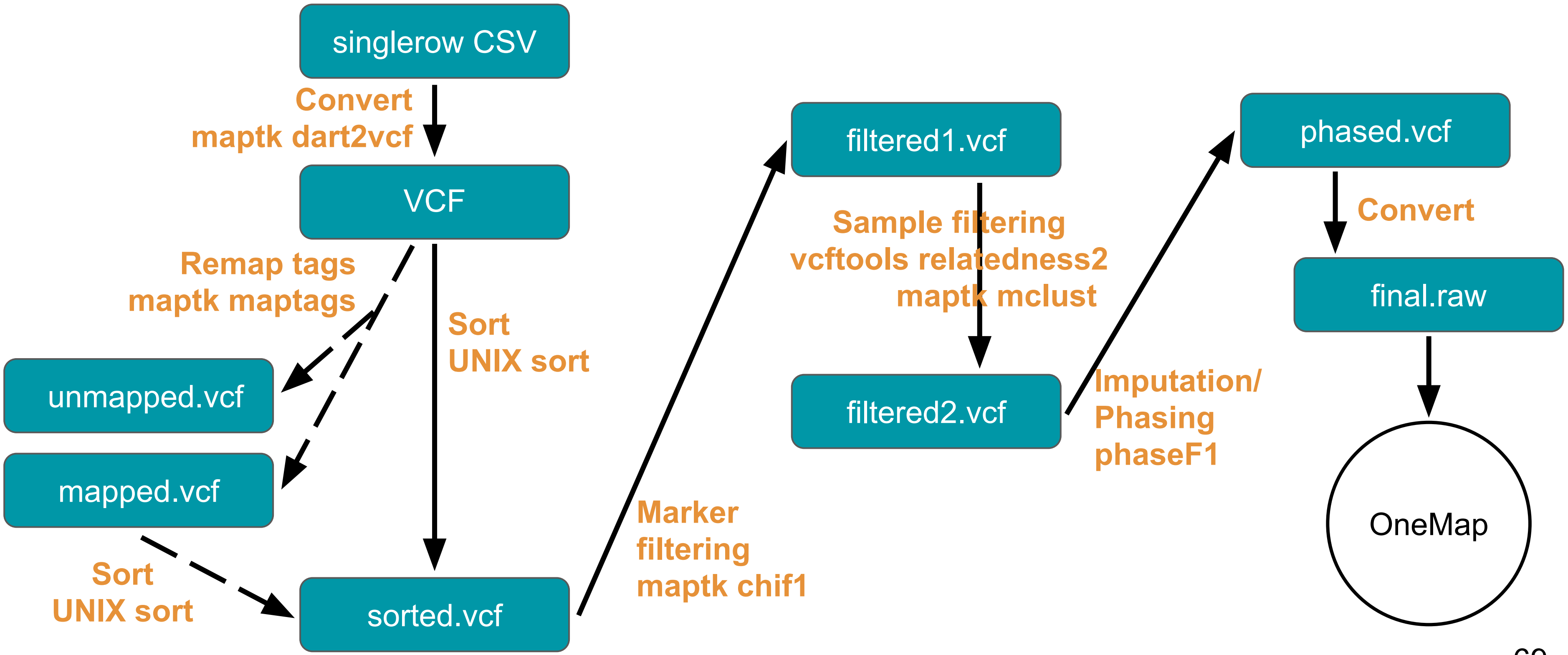
When to impute? When not?

- With OneMap we’ll be able to judge correctness of imputation.
- Iteration/parameter sweeps.

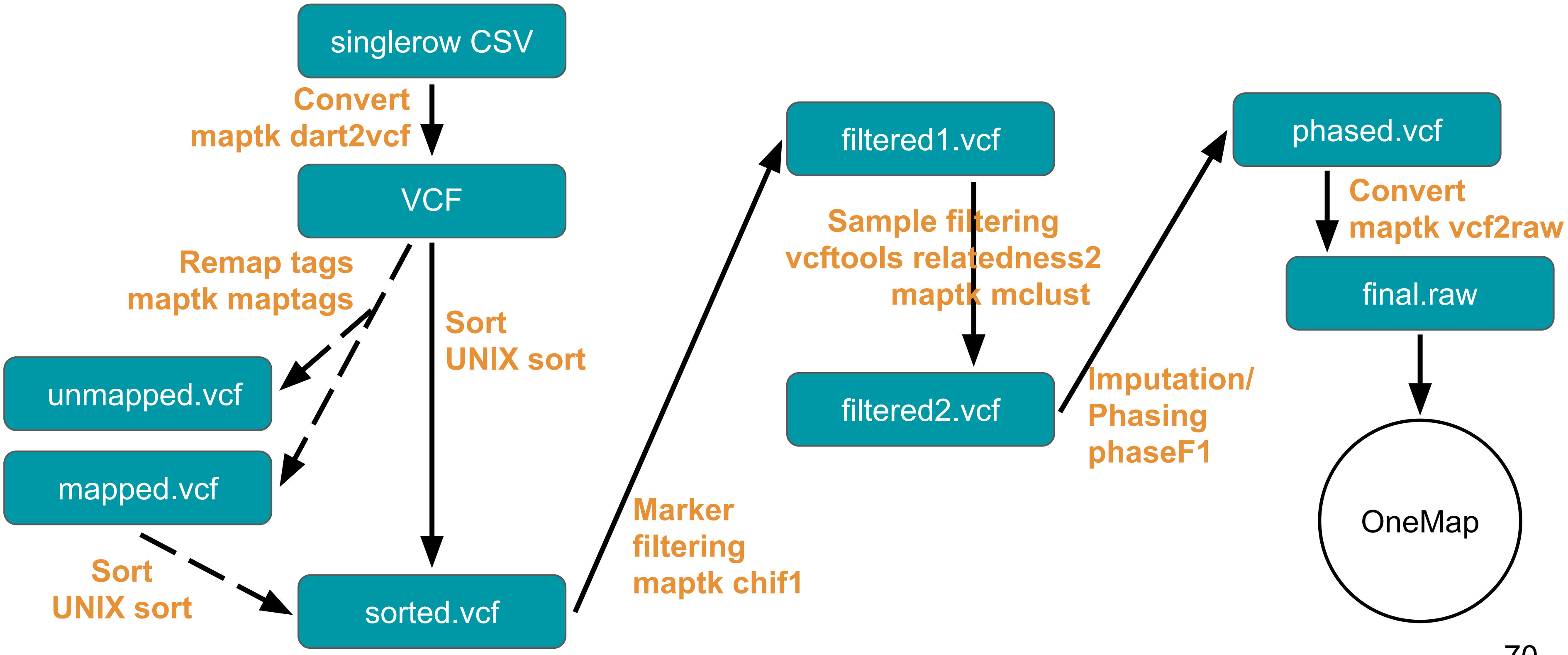
Imputation/phasing software:

1. Beagle
2. MACH
3. IMPUTE2

DArTseq to mapping: analysis flow chart



DArTseq to mapping: analysis flow chart



singlerow CSV

Convert
maptk dart2vcf

VCF

Remap tags
maptk maptags

unmapped.vcf

mapped.vcf

Sort
UNIX sort

Sort
UNIX sort

sorted.vcf

Marker
filtering
maptk chif1

filtered1.vcf

Sample filtering
vcftools relatedness2
maptk mclust

filtered2.vcf

Imputation/
Phasing
phaseF1

phased.vcf

Convert
maptk vcf2raw

final.raw

OneMap