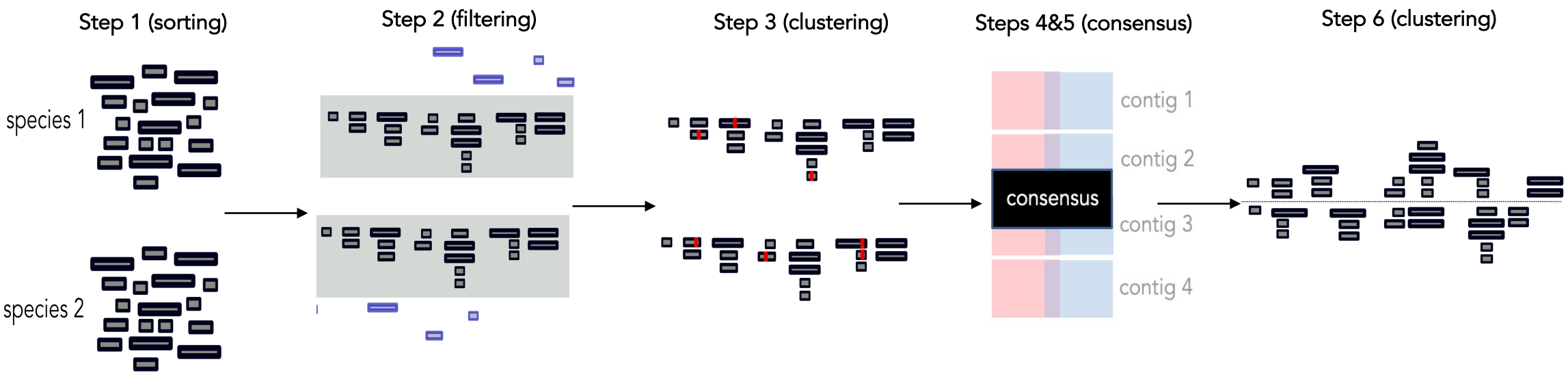
iPyrad – stats and output files

Let’s go through the stats files for each step together



What happens after we've run iPyrad?

Step	Role of step	Stats file
1	Demultiplexing	s1_demultiplex_stats.txt
2	Filtering	s2_rawedit_stats.txt
3	Clustering within inds	s3_cluster_stats.txt
4	Error rate & heterozygosity	s4_joint_estimate.txt
5	Consensus reads	s5_consens_stats.txt
6	Clustering among inds	s6_cluster_stats.txt
7	Cleaning up	[project]_stats.txt

Step 1: demultiplexing & sorting

```
annes-mbp:s123456_stats_files eac$ cat ranaddrad_total_4_s1_demultiplex_stats.txt
raw_file                              total_reads    cut_found  bar_matched
64T64_P29_S1_L005_R1_001.fastq           71486579     71486579     69923714

sample_name                           total_reads
Rber_T1113a                              8159746
Rber_T1113b                              6091513
Rber_T1114                               3232832
Rbla_D2864                               6239552
Rbla_D2865                               4801774
Rchi_T2034a                              4598334
Rchi_T2034b                              5507763
Rchi_T2049                               4418342
Rneo_T480                                6657867
Rneo_T527                                6653066
Rsph_T25870                              6761860
Rsph_T26064                              6801065

sample_name                             true_bar       obs_bar     N_records
Rber_T1113a                                ATTAC         ATTAC       8159746
Rber_T1113b                                CATAT         CATAT       6091513
Rber_T1114                                 CGAAT         CGAAT       3232832
Rbla_D2864                                 ACTGG         ACTGG       6239552
Rbla_D2865                                 ACTTC         ACTTC       4801774
Rchi_T2034a                                CGGCT         CGGCT       4598334
Rchi_T2034b                                CGGTA         CGGTA       5507763
Rchi_T2049                                 CGTAC         CGTAC       4418342
Rneo_T480                                  ATACG         ATACG       6657867
Rneo_T527                                  ATGAG         ATGAG       6653066
Rsph_T25870                                CGTCG         CGTCG       6761860
Rsph_T26064                                CTGAT         CTGAT       6801065
no_match                                                             1562865
```

Step 2: trimming and filtering

```
annes-mbp:s123456_stats_files eac$ cat ranaddrad_total_4_s2_rawedit_stats.txt
             reads_raw  trim_adapter_bp_read1  trim_quality_bp_read1  reads_filtered_by_Ns  reads_filtered_by_minlen  reads_passed_filter
Rber_T1113a    8159746                 197667                2292799                     3                      3120              8156623
Rber_T1113b    6091513                 143901                1672565                     3                      2199              6089311
Rber_T1114     3232832                  74364                 894147                     2                      1115              3231715
Rbla_D2864     6239552                 129260                1756460                     1                      2390              6237161
Rbla_D2865     4801774                  98276                1365355                     0                      1835              4799939
Rchi_T2034a    4598334                 106437                1210241                     0                      1974              4596360
Rchi_T2034b    5507763                 127490                1448659                     3                      2484              5505276
Rchi_T2049     4418342                 101244                1188569                     6                      1928              4416408
Rneo_T480      6657867                 148247                1855150                     4                      2567              6655296
Rneo_T527      6653066                 151370                1834309                     4                      2641              6650421
Rsph_T25870    6761860                 198727                2052813                     2                      2799              6759059
Rsph_T26064    6801065                 186264                1983466                     0                      2944              6798121
```

Step 3: clustering within samples

```
annes-mbp:s123456_stats_files eac$ cat ranaddrad_total_4_s3_cluster_stats.txt
           clusters_total  hidepth_min clusters_hidepth avg_depth_total avg_depth_mj avg_depth_stat sd_depth_total sd_depth_mj sd_depth_stat filtered_bad_align
Rber_T1113a         802220          5.0           186114            8.22        29.33          29.33         304.57      631.86       631.86               931
Rber_T1113b         686970          5.0           153957            7.21        25.88          25.88         244.56      516.16       516.16               743
Rber_T1114          360145          5.0            98887            7.40        22.21          22.21         194.27      370.33       370.33                 0
Rbla_D2864          502812          5.0           127110            9.87        33.71          33.71         449.35      893.29       893.29                 0
Rbla_D2865          432516          5.0           115912            8.70        27.53          27.53         364.51      703.78       703.78                 0
Rchi_T2034a         536519          5.0           115983            6.54        23.71          23.71         444.86      956.60       956.60               686
Rchi_T2034b         591567          5.0           130989            7.04        25.42          25.42         501.90     1066.41      1066.41               779
Rchi_T2049          520647          5.0           113175            6.49        23.37          23.37         420.26      901.19       901.19                 0
Rneo_T480           722831          5.0           157618            7.42        27.48          27.48         235.57      503.96       503.96               843
Rneo_T527           605126          5.0           143238            8.87        31.60          31.60         276.25      567.20       567.20                 0
Rsph_T25870         820907          5.0           182201            6.76        24.04          24.04         212.45      450.52       450.52                 0
Rsph_T26064         843798          5.0           186490            6.71        23.87          23.87         212.95      452.55       452.55               841
```

Step 4: error and heterozygosity estimates

```
annes-mbp:s123456_stats_files eac$ cat ranaddrad_total_4_s4_joint_estimate.txt
             hetero_est  error_est
Rber_T1113a    0.017304   0.003614
Rber_T1113b    0.016984   0.003947
Rber_T1114     0.013945   0.003735
Rbla_D2864     0.012272   0.003026
Rbla_D2865     0.011767   0.003267
Rchi_T2034a    0.012805   0.004117
Rchi_T2034b    0.012760   0.003980
Rchi_T2049     0.012621   0.004130
Rneo_T480      0.013120   0.003240
Rneo_T527      0.013061   0.003019
Rsph_T25870    0.015218   0.003336
```

Step 5: consensus reads

```
annes-mbp:s123456_stats_files eac$ cat ranaddrad_total_4_s5_consens_stats.txt
           clusters_total filtered_by_depth filtered_by_maxH filtered_by_maxN reads_consens   nsites nhetero heterozygosity
Rber_T1113a         802220            616117             9508            51440        125155 14441503   94306        0.00653
Rber_T1113b         686970            533023             7927            41265        104755 12086721   77660        0.00643
Rber_T1114          360145            261266             4359            22034         72486  8354958   49676        0.00595
Rbla_D2864          502812            375714             5299            25988         95811 11041382   43048        0.00390
Rbla_D2865          432516            316609             4783            23661         87463 10080198   39347        0.00390
Rchi_T2034a         536519            420546             5312            28675         81986  9443051   37534        0.00397
Rchi_T2034b         591567            460588             5957            31686         93336 10751754   42516        0.00395
Rchi_T2049          520647            407482             5150            27666         80349  9257660   36004        0.00389
Rneo_T480           722831            565227             7209            37100        113295 13050647   46843        0.00359
Rneo_T527           605126            461902             6419            30792        106013 12213584   43050        0.00352
Rsph_T25870         820907            638714             8217            47983        125993 14527911   88732        0.00611
Rsph_T26064         843798            657318             8364            48896        129220 14895268   89563        0.00601
```

Step 6: clustering among samples

```
annes-mbp:s123456_stats_files eac$ cat ranaddrad_total_4_s6_cluster_stats.txt
vsearch v2.0.3_linux_x86_64, 441.6GB RAM, 64 cores
/home1/02576/rdtarvin/miniconda2/lib/python2.7/site-packages/bin/vsearch-linux-x86_64 -c
shuf.tmp -strand plus -query_cov 0.75 -minsl 0.5 -id 0.91 -userout /scratch/02576/rdtarv
n/ddrad_rana-R1/clust_91_across/clust_91.htemp -userfields query+target+qstrand -maxacce
76/rdtarvin/ddrad_rana-R1/clust_91_across/s6_cluster_stats.txt
Started  Sat Apr 14 17:40:19 2018140144637 nt in 1215862 seqs, min 35, max 130, avg 115


      Alphabet  nt
    Word width  8
     Word ones  8
        Spaced  No
        Hashed  No
         Coded  No
       Stepped  No
         Slots  65536 (65.5k)
       DBAccel  100%

Clusters: 544377 Size min 1, max 216, avg 2.2
Singletons: 302459, 24.9% of seqs, 55.6% of clusters


Finished Sat Apr 14 18:21:11 2018
Elapsed time 40:52
Max memory 1.7GB
```

Step 7: final stats file

This file has four sections:

```
annes-mbp:ranaddrad_total_4_outfiles eac$ cat ranaddrad_clust_91_stats.txt


## The number of loci caught by each filter.
## ipyrad API location: [assembly].stats_dfs.s7_filters

                          total_filters  applied_order  retained_loci
total_prefiltered_loci           241918              0         241918
filtered_by_rm_duplicates          3698           3698         238220
filtered_by_max_indels              598            598         237622
filtered_by_max_snps               2758            405         237217
filtered_by_max_shared_het         4605           4106         233111
filtered_by_min_sample           155829         154606          78505
filtered_by_max_alleles           10303           3112          75393
total_filtered_loci               75393              0          75393
```

Step 7: final stats file

This file has four sections:

```
## The number of loci recovered for each Sample.
## ipyrad API location: [assembly].stats_dfs.s7_samples

             sample_coverage
Rber_T1113a            53940
Rber_T1113b            50492
Rber_T1114             42060
Rbla_D2864             38768
Rbla_D2865             37157
Rchi_T2034a            23740
Rchi_T2034b            25406
Rchi_T2049             23533
Rneo_T480              50410
Rneo_T527              51033
Rsph_T25870            38178
Rsph_T26064            38113


## The number of loci for which N taxa have data.
## ipyrad API location: [assembly].stats_dfs.s7_loci

    locus_coverage  sum_coverage
1                0             0
2                0             0
3                0             0
4            20462         20462
5            16781         37243
6             8956         46199
7             9561         55760
8             5745         61505
9             6073         67578
10            3031         70609
11            1921         72530
12            2863         75393
```

Step 7: final stats file

This file has four sections:

```
## The distribution of SNPs (var and pis) per locus.
## var = Number of loci with n variable sites (pis + autapomorphies)
## pis = Number of loci with n parsimony informative site (minor allele in >1 sample)
## ipyrad API location: [assembly].stats_dfs.s7_snps

     var  sum_var    pis  sum_pis
0   5966        0  16402        0
1   7425     7425  11992    11992
2   8015    23455  10407    32806
3   8283    48304   8717    58957
4   8051    80508   7341    88321
5   7412   117568   5896   117801
6   6714   157852   4605   145431
7   5618   197178   3441   169518
8   4621   234146   2459   189190
9   3892   269174   1714   204616
10  2857   297744   1059   215206
11  2122   321086    620   222026
12  1511   339218    335   226046
13   996   352166    185   228451
14   697   361924    118   230103
15   456   368764     51   230868
16   280   373244     27   231300
17   211   376831     17   231589
18   124   379063      3   231643
19    86   380697      4   231719
20    56   381817      0   231719
```

Step 7: final stats file

This file has four sections:

Final Sample stats summary

	state	reads_raw	reads_passed_filter	clusters_total	clusters_hidepth	hetero_est	error_est	reads_consens	loci_in_assembly
Rber_T1113a	7	8159746	8156623	802220	186114	0.017304	0.003614	125155	53940
Rber_T1113b	7	6091513	6089311	686970	153957	0.016984	0.003947	104755	50492
Rber_T1114	7	3232832	3231715	360145	98887	0.013945	0.003735	72486	42060
Rbla_D2864	7	6239552	6237161	502812	127110	0.012272	0.003026	95811	38768
Rbla_D2865	7	4801774	4799939	432516	115912	0.011767	0.003267	87463	37157
Rchi_T2034a	7	4598334	4596360	536519	115983	0.012805	0.004117	81986	23740
Rchi_T2034b	7	5507763	5505276	591567	130989	0.012760	0.003980	93336	25406
Rchi_T2049	7	4418342	4416408	520647	113175	0.012621	0.004130	80349	23533
Rneo_T480	7	6657867	6655296	722831	157618	0.013120	0.003240	113295	50410
Rneo_T527	7	6653066	6650421	605126	143238	0.013061	0.003019	106013	51033
Rsph_T25870	7	6761860	6759059	820907	182201	0.015218	0.003336	125993	38178
Rsph_T26064 _	7	6801065	6798121	843798	186490	0.014984	0.003384	129220	38113

iPyrad –output files

Two main files I want to go over:

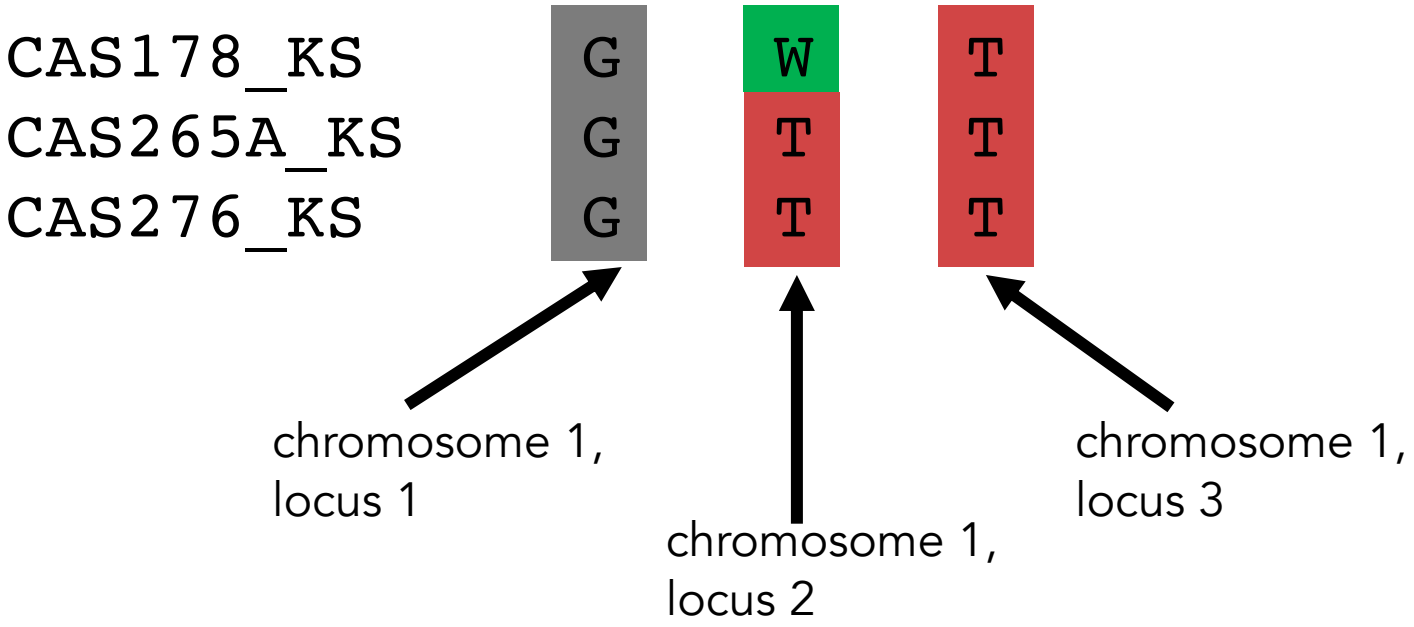
- **.vcf** file: this is a file commonly used as input into many other (mostly population genetics-based) programs
- **.loci** file: this is iPyrad-specific and gives us some important information

Variant call format file (.vcf, .vcf.gz)

- Contains information about genotypes, quality, and read depth, along with sequence data *(sometimes even more!)*
- *vcftools* is a useful software for analyzing vcf files

Variant call format file (.vcf, .vcf.gz)

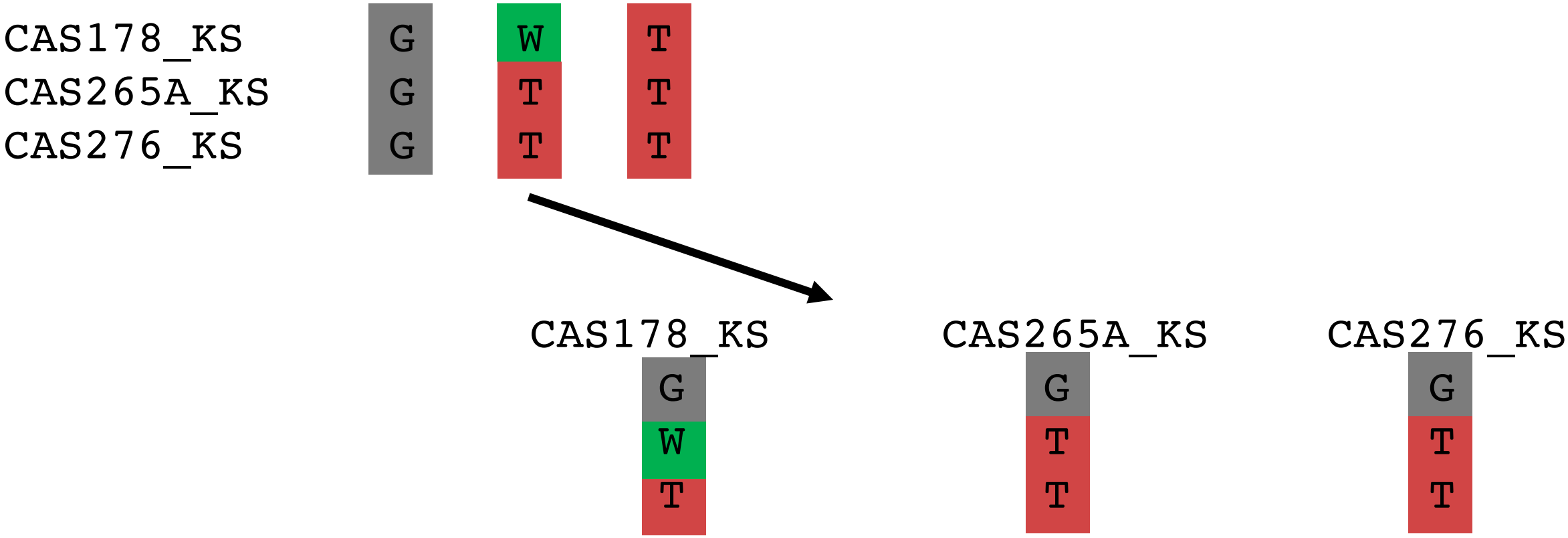
Let's take a look at some **SNP** sequence data for three samples:



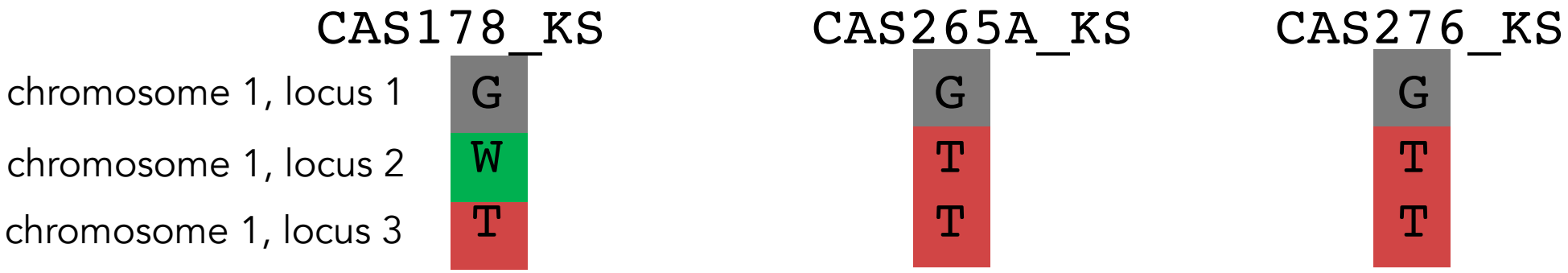
(I realize that if this was your entire dataset, sites 1 and 3 wouldn't be considered SNPs because they're invariant, but just assume that you have other samples with variable sites at sites 1 and 3)

Variant call format file (.vcf, .vcf.gz)

Transpose the data:

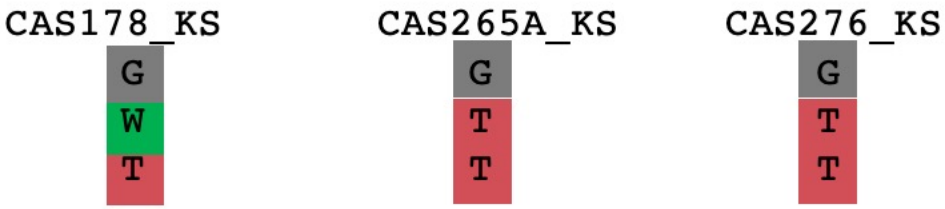


Variant call format file (.vcf, .vcf.gz)



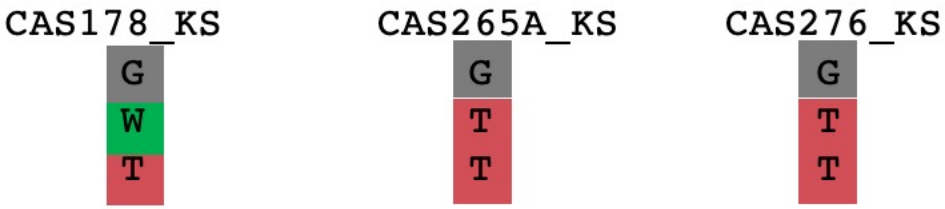
This is exactly the format that the vcf file uses!

Variant call format file (.vcf, .vcf.gz)



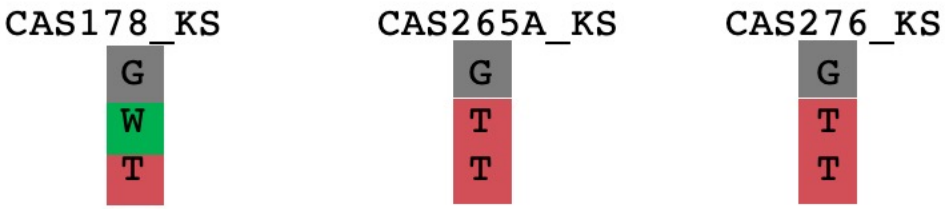
#CHROM	POS	ID	REF	ALT	QUALFILTER	INFO	FORMAT	CAS178_KS	CAS265A_KS	CAS276_KS

Variant call format file (.vcf, .vcf.gz)



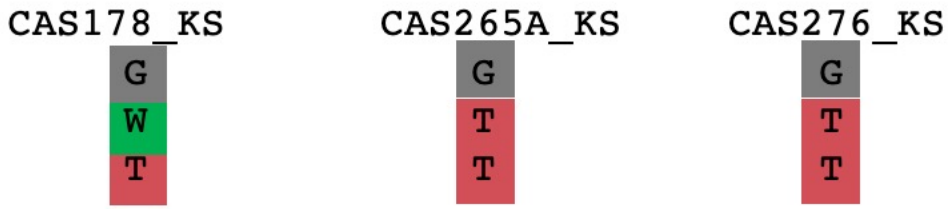
#CHROM	POS	ID	REF	ALT	QUAL	FILTER	INFO	FORMAT	CAS178_KS	CAS265A_KS	CAS276_KS
RAD_0	12	loc0_pos11									
RAD_0	15	loc0_pos14									
RAD_0	23	loc0_pos22									

Variant call format file (.vcf, .vcf.gz)



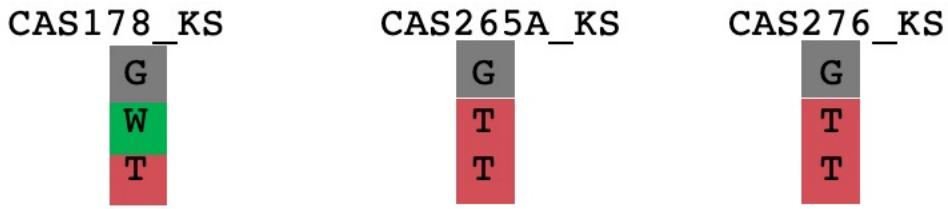

```
##fileformat=VCFv4.0
##fileDate=2021/09/28
##source=ipyrad_v.0.9.81
##reference=pseudo-reference (most common base at site)
##phasing=unphased
##INFO=<ID=NS,Number=1,Type=Integer,Description="Number of Samples With Data">
##INFO=<ID=DP,Number=1,Type=Integer,Description="Total Depth">
##FORMAT=<ID=GT,Number=1,Type=String,Description="Genotype">
##FORMAT=<ID=DP,Number=1,Type=Integer,Description="Read Depth">
##FORMAT=<ID=CATG,Number=1,Type=String,Description="Base Counts (CATG)">
#CHROM  POS ID          REF ALT QUALFILTER  INFO            FORMAT      CAS178_KS          CAS265A_KS          CAS276_KS
RAD_0   12  loc0_pos11
RAD_0   15  loc0_pos14
RAD_0   23  loc0_pos22
```

Variant call format file (.vcf, .vcf.gz)




```
##fileformat=VCFv4.0
##fileDate=2021/09/28
##source=ipyrad v.0.9.81
##reference=pseudo-reference (most common base at site)
##phasing=unphased
##INFO=<ID=NS,Number=1,Type=Integer,Description="Number of Samples With Data">
##INFO=<ID=DP,Number=1,Type=Integer,Description="Total Depth">
##FORMAT=<ID=GT,Number=1,Type=String,Description="Genotype">
##FORMAT=<ID=DP,Number=1,Type=Integer,Description="Read Depth">
##FORMAT=<ID=CATG,Number=1,Type=String,Description="Base Counts (CATG)">
#CHROM  POS ID          REF ALT QUALFILTER  INFO            FORMAT      CAS178_KS       CAS265A_KS          CAS276_KS
RAD_0   12  loc0_pos11  G   A   13  PASS    NS=113;DP=4086  GT:DP:CATG
RAD_0   15  loc0_pos14  T   A   13  PASS    NS=112;DP=4086  GT:DP:CATG
RAD_0   23  loc0_pos22  T   A   13  PASS    NS=114;DP=4086  GT:DP:CATG
```

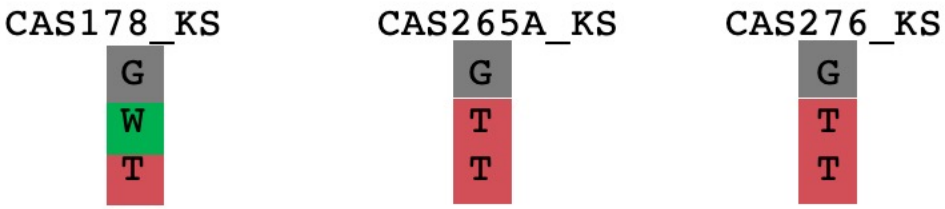
Variant call format file (.vcf, .vcf.gz)




```
##fileformat=VCFv4.0
##fileDate=2021/09/28
##source=ipyrad_v.0.9.81
##reference=pseudo-reference (most common base at site)
##phasing=unphased
##INFO=<ID=NS,Number=1,Type=Integer,Description="Number of Samples With Data">
##INFO=<ID=DP,Number=1,Type=Integer,Description="Total Depth">
##FORMAT=<ID=GT,Number=1,Type=String,Description="Genotype">
##FORMAT=<ID=DP,Number=1,Type=Integer,Description="Read Depth">
##FORMAT=<ID=CATG,Number=1,Type=String,Description="Base Counts (CATG)">
#CHROM  POS ID          REF ALT QUALFILTER  INFO            FORMAT      CAS178_KS          CAS265A_KS          CAS276_KS
RAD_0   12  loc0_pos11  G   A   13   PASS   NS=113;DP=4086  GT:DP:CATG
RAD_0   15  loc0_pos14  T   A   13   PASS   NS=112;DP=4086  GT:DP:CATG
RAD_0   23  loc0_pos22  T   A   13   PASS   NS=114;DP=4086  GT:DP:CATG
```

quality information

Variant call format file (.vcf, .vcf.gz)



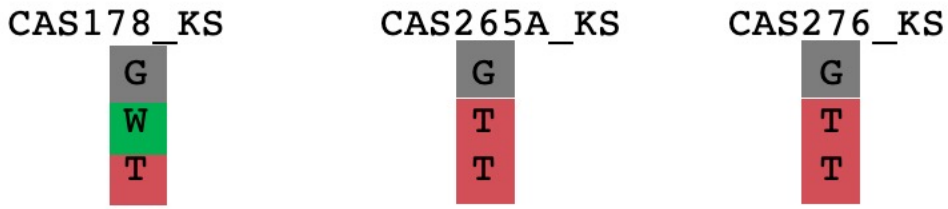

```
##fileformat=VCFv4.0
##fileDate=2021/09/28
##source=ipyrad_v.0.9.81
##reference=pseudo-reference (most common base at site)
##phasing=unphased
##INFO=<ID=NS,Number=1,Type=Integer,Description="Number of Samples With Data">
##INFO=<ID=DP,Number=1,Type=Integer,Description="Total Depth">
##FORMAT=<ID=GT,Number=1,Type=String,Description="Genotype">
##FORMAT=<ID=DP,Number=1,Type=Integer,Description="Read Depth">
##FORMAT=<ID=CATG,Number=1,Type=String,Description="Base Counts (CATG)">
#CHROM  POS ID          REF ALT QUALFILTER  INFO            FORMAT      CAS178_KS          CAS265A_KS          CAS276_KS
RAD_0   12  loc0_pos11  G   A   13  PASS    NS=113;DP=4086  GT:DP:CATG
RAD_0   15  loc0_pos14  T   A   13  PASS    NS=112;DP=4086  GT:DP:CATG
RAD_0   23  loc0_pos22  T   A   13  PASS    NS=114;DP=4086  GT:DP:CATG
```

Variant call format file (.vcf, .vcf.gz)

CAS178_KS	CAS265A_KS	CAS276_KS
G	G	G
W	T	T
T	T	T

```
##fileformat=VCFv4.0
##fileDate=2021/09/28
##source=ipyrad_v.0.9.81
##reference=pseudo-reference (most common base at site)
##phasing=unphased
##INFO=<ID=NS,Number=1,Type=Integer,Description="Number of Samples With Data">
##INFO=<ID=DP,Number=1,Type=Integer,Description="Total Depth">
##FORMAT=<ID=GT,Number=1,Type=String,Description="Genotype">
##FORMAT=<ID=DP,Number=1,Type=Integer,Description="Read Depth">
##FORMAT=<ID=CATG,Number=1,Type=String,Description="Base Counts (CATG)">
#CHROM  POS ID          REF ALT QUALFILTER  INFO            FORMAT      CAS178_KS       CAS265A_KS          CAS276_KS
RAD_0   12  loc0_pos11  G   A   13  PASS    NS=113;DP=4086  GT:DP:CATG
RAD_0   15  loc0_pos14  T   A   13  PASS    NS=112;DP=4086  GT:DP:CATG
RAD_0   23  loc0_pos22  T   A   13  PASS    NS=114;DP=4086  GT:DP:CATG
```

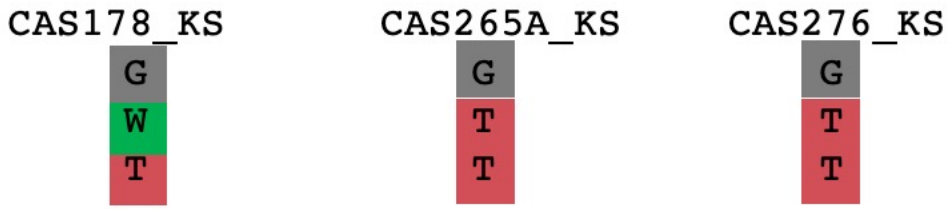
Variant call format file (.vcf, .vcf.gz)




```
##fileformat=VCFv4.0
##fileDate=2021/09/28
##source=ipyrad_v.0.9.81
##reference=pseudo-reference (most common base at site)
##phasing=unphased
##INFO=<ID=NS,Number=1,Type=Integer,Description="Number of Samples With Data">
##INFO=<ID=DP,Number=1,Type=Integer,Description="Total Depth">
##FORMAT=<ID=GT,Number=1,Type=String,Description="Genotype">
##FORMAT=<ID=DP,Number=1,Type=Integer,Description="Read Depth">
##FORMAT=<ID=CATG,Number=1,Type=String,Description="Base Counts (CATG)">
#CHROM  POS ID          REF ALT QUALFILTER  INFO            FORMAT      CAS178_KS       CAS265A_KS          CAS276_KS
RAD_0   12  loc0_pos11  G   A   13  PASS    NS=113;DP=4086  GT:DP:CATG
RAD_0   15  loc0_pos14  T   A   13  PASS    NS=112;DP=4086  GT:DP:CATG
RAD_0   23  loc0_pos22  T   A   13  PASS    NS=114;DP=4086  GT:DP:CATG
```

First item: **GT (genotype)**; this is commonly coded with 0s and 1s, separated by a /

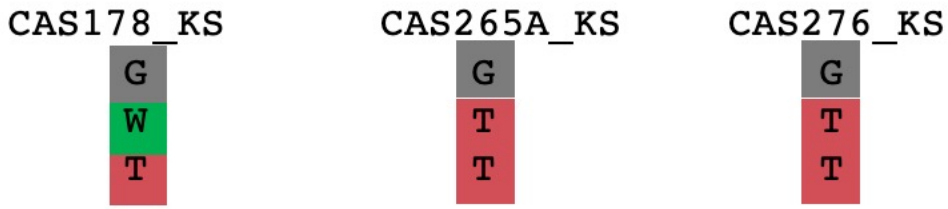
Variant call format file (.vcf, .vcf.gz)




```
##fileformat=VCFv4.0
##fileDate=2021/09/28
##source=ipyrad_v.0.9.81
##reference=pseudo-reference (most common base at site)
##phasing=unphased
##INFO=<ID=NS,Number=1,Type=Integer,Description="Number of Samples With Data">
##INFO=<ID=DP,Number=1,Type=Integer,Description="Total Depth">
##FORMAT=<ID=GT,Number=1,Type=String,Description="Genotype">
##FORMAT=<ID=DP,Number=1,Type=Integer,Description="Read Depth">
##FORMAT=<ID=CATG,Number=1,Type=String,Description="Base Counts (CATG)">
#CHROM  POS ID          REF ALT QUALFILTER  INFO            FORMAT      CAS178_KS       CAS265A_KS       CAS276_KS
RAD_0   12  loc0_pos11  G   A   13  PASS    NS=113;DP=4086  GT:DP:CATG  0/0             0/0              0/0
RAD_0   15  loc0_pos14  T   A   13  PASS    NS=112;DP=4086  GT:DP:CATG  0/1             0/0              0/0
RAD_0   23  loc0_pos22  T   A   13  PASS    NS=114;DP=4086  GT:DP:CATG  0/0             0/0              0/0
```

First item: **GT (genotype)**; this is commonly coded with 0s and 1s, separated by a /

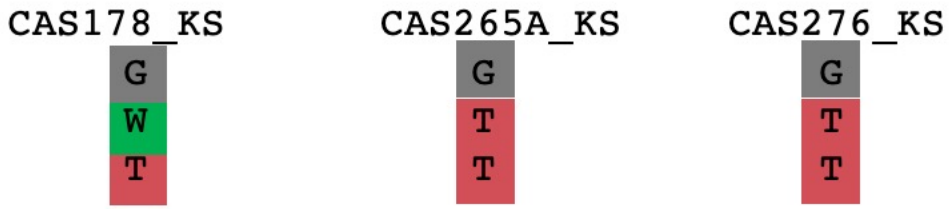
Variant call format file (.vcf, .vcf.gz)




```
##fileformat=VCFv4.0
##fileDate=2021/09/28
##source=ipyrad_v.0.9.81
##reference=pseudo-reference (most common base at site)
##phasing=unphased
##INFO=<ID=NS,Number=1,Type=Integer,Description="Number of Samples With Data">
##INFO=<ID=DP,Number=1,Type=Integer,Description="Total Depth">
##FORMAT=<ID=GT,Number=1,Type=String,Description="Genotype">
##FORMAT=<ID=DP,Number=1,Type=Integer,Description="Read Depth">
##FORMAT=<ID=CATG,Number=1,Type=String,Description="Base Counts (CATG)">
#CHROM  POS ID          REF ALT QUALFILTER  INFO            FORMAT      CAS178_KS      CAS265A_KS      CAS276_KS
RAD_0   12  loc0_pos11  G   A   13   PASS    NS=113;DP=4086  GT:DP:CATG  0/0:19         0/0:43          0/0:39
RAD_0   15  loc0_pos14  T   A   13   PASS    NS=112;DP=4086  GT:DP:CATG  0/1:19         0/0:43          0/0:39
RAD_0   23  loc0_pos22  T   A   13   PASS    NS=114;DP=4086  GT:DP:CATG  0/0:19         0/0:43          0/0:39
```

Second item: **DP (read depth)**

Variant call format file (.vcf, .vcf.gz)




```
##fileformat=VCFv4.0
##fileDate=2021/09/28
##source=ipyrad_v.0.9.81
##reference=pseudo-reference (most common base at site)
##phasing=unphased
##INFO=<ID=NS,Number=1,Type=Integer,Description="Number of Samples With Data">
##INFO=<ID=DP,Number=1,Type=Integer,Description="Total Depth">
##FORMAT=<ID=GT,Number=1,Type=String,Description="Genotype">
##FORMAT=<ID=DP,Number=1,Type=Integer,Description="Read Depth">
##FORMAT=<ID=CATG,Number=1,Type=String,Description="Base Counts (CATG)">
#CHROM  POS  ID          REF  ALT  QUAL  FILTER  INFO             FORMAT      CAS178_KS          CAS265A_KS          CAS276_KS
RAD_0   12   loc0_pos11  G    A    13    PASS    NS=113;DP=4086   GT:DP:CATG  0/0:19:0,0,0,19    0/0:43:0,0,0,43     0/0:39:0,0,0,39
RAD_0   15   loc0_pos14  T    A    13    PASS    NS=112;DP=4086   GT:DP:CATG  0/1:19:0,10,9,0    0/0:43:0,0,43,0     0/0:39:0,0,39,0
RAD_0   23   loc0_pos22  T    A    13    PASS    NS=114;DP=4086   GT:DP:CATG  0/0:19:0,0,19,0    0/0:43:0,0,43,0     0/0:39:0,0,39,0
```

Third item: **CATG (# reads per base)**

Variant call format file (.vcf, .vcf.gz)

Bioinformatics pipelines differ in the type (and amount) of information contained within the vcf file

```
##fileformat=VCFv4.2(angsd version)
##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=RA,Number=1,Type=String,Description="Reference Allele (included since ANGSD places the MAJOR allele under REF)">
##INFO=<ID=AA,Number=1,Type=String,Description="Ancestral Allele">
##INFO=<ID=AF,Number=A,Type=Float,Description="Minor Allele Frequency">
##FORMAT=<ID=GT,Number=1,Type=String,Description="Genotype">
##FORMAT=<ID=DP,Number=1,Type=Integer,Description="Read Depth">
##FORMAT=<ID=AD,Number=.,Type=Integer,Description="Allelic depths for the major and minor alleles in the order listed">
##FORMAT=<ID=GP,Number=G,Type=Float,Description="Genotype Probabilities">
##FORMAT=<ID=PL,Number=G,Type=Float,Description="Phred-scaled Genotype Likelihoods">
##FORMAT=<ID=GL,Number=G,Type=Float,Description="scaled Genotype Likelihoods (loglikeratios to the most likely (in log10))">
#CHROM  POS ID  REF ALT QUALFILTER  INFO                              FORMAT          ind0
chr1    373 .   G   T   .   PASS    NS=73;DP=1342;RA=G;AF=0.079071  GT:DP:AD:GP:GL  0/0:10:10,0:0.999832,0.000168,0.000000:0.000000,-3.010338,-15.012766
chr1    825 .   C   T   .   PASS    NS=76;DP=456;RA=C;AF=0.050129   GT:DP:AD:GP:GL  0/0:7:7,0:0.999176,0.000824,0.000000:0.000000,-2.107237,-14.923611
chr1    1019.   A   C   .   PASS    NS=73;DP=474;RA=A;AF=0.074571   GT:DP:AD:GP:GL  0/0:4:4,0:0.990028,0.009972,0.000000:0.000000,-1.204135,-10.716633
```

More data, including genotype likelihoods

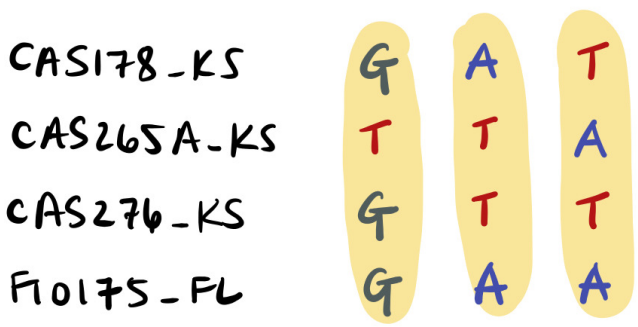
The iPyrad .loci file

Let's go back to the original data but change it so that it's actually three SNPs and add another sample:

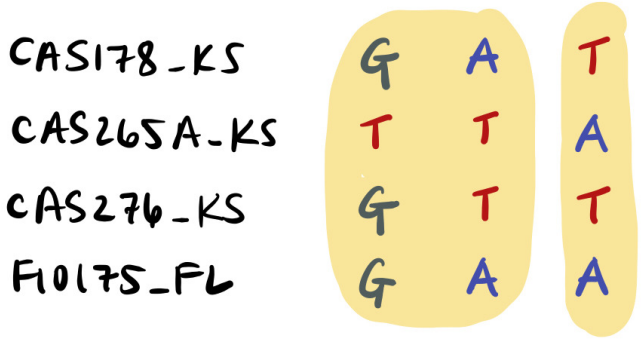
CAS178_KS	G	A	T
CAS265A-KS	T	T	A
CAS276_KS	G	T	T
F10175_FL	G	A	A

When we get our sequence data though, we *do not know* which SNPs occur on which loci

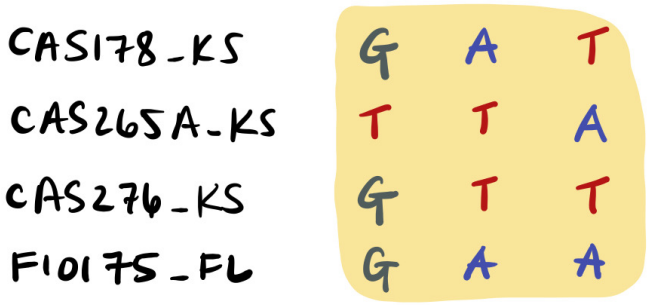
This information is what the .loci file tells us (using the entire locus, not just the SNPs)!



3 SNPs, 3 loci

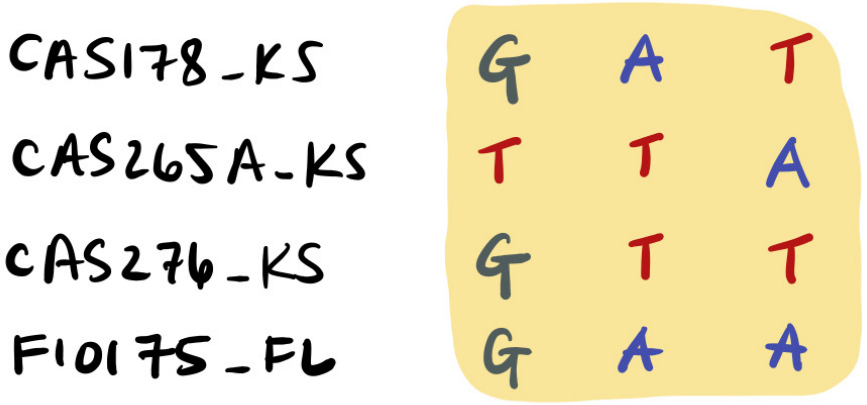


3 SNPs, 2 loci



3 SNPs, 1 locus

The iPyrad .loci file



3 SNPs, 1 locus

Before we look at the .loci file for these samples, let's think about what these SNPs are telling us (from a phylogenetic perspective)

The iPyrad .loci file

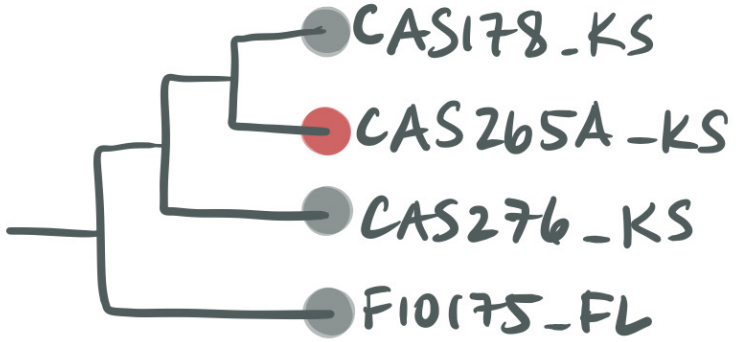
CAS178_KS	G	A	T
CAS265A_KS	T	T	A
CAS276_KS	G	T	T
F10175_FL	G	A	A

3 SNPs, 1 locus

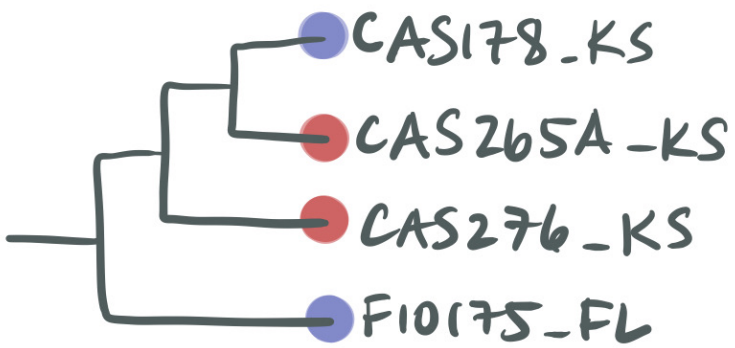
```
## The distribution of SNPs (var and pis) per locus.
## var = Number of loci with n variable sites (pis + autapomorphies)
## pis = Number of loci with n parsimony informative site (minor allele in >1 sample)
## ipyrad API location: [assembly].stats_dfs.s7_snps

     var  sum_var    pis  sum_pis
0   5966        0  16402        0
1   7425     7425  11992    11992
2   8015    23455  10407    32806
3   8283    48304   8717    58957
4   8051    80508   7341    88321
5   7412   117568   5896   117801
6   6714   157852   4605   145431
7   5618   197178   3441   169518
8   4621   234146   2459   189190
9   3892   269174   1714   204616
10  2857   297744   1059   215206
11  2122   321086    620   222026
12  1511   339218    335   226046
13   996   352166    185   228451
14   697   361924    118   230103
15   456   368764     51   230868
16   280   373244     27   231300
17   211   376831     17   231589
18   124   379063      3   231643
19    86   380697      4   231719
20    56   381817      0   231719
```

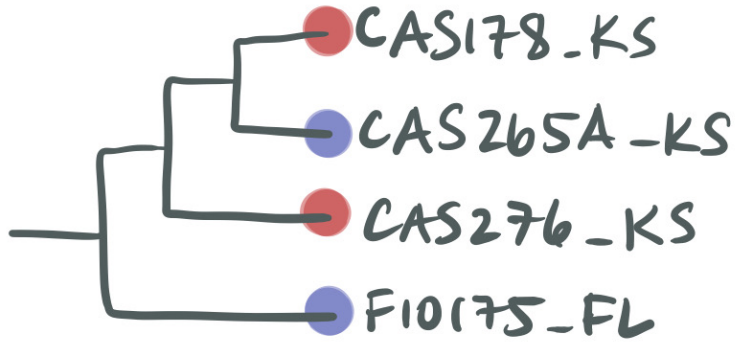
SNP 1 is an **autapomorphy**



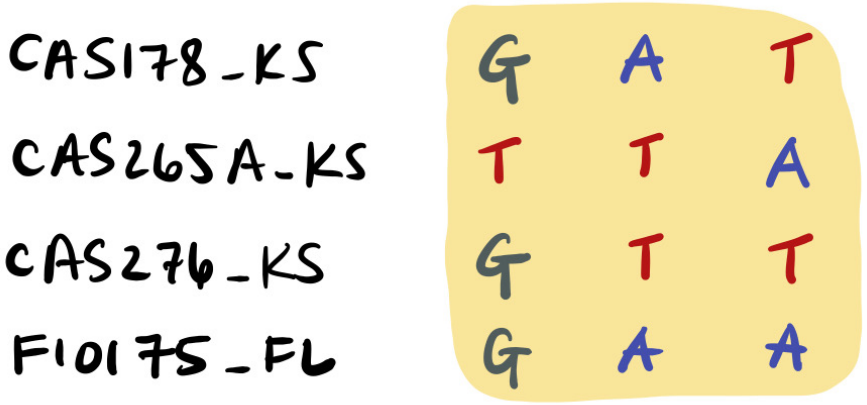
SNP 2 is a **synapomorphy (PI)**



SNP 3 is a **synapomorphy (PI)**



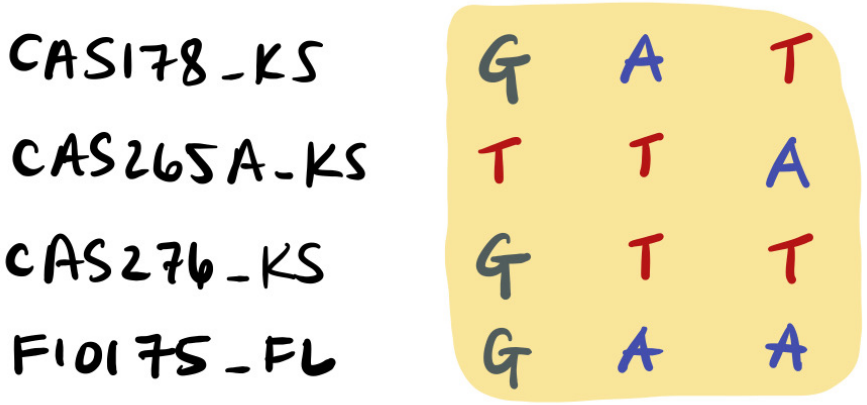
The iPyrad .loci file



3 SNPs, 1 locus

iPyrad's .loci file not only gives us information about which SNPs are contained within each locus, but it also tells us which SNPs are autapomorphies and which are synapomorphies (PIs)

The iPyrad .loci file



3 SNPs, 1 locus

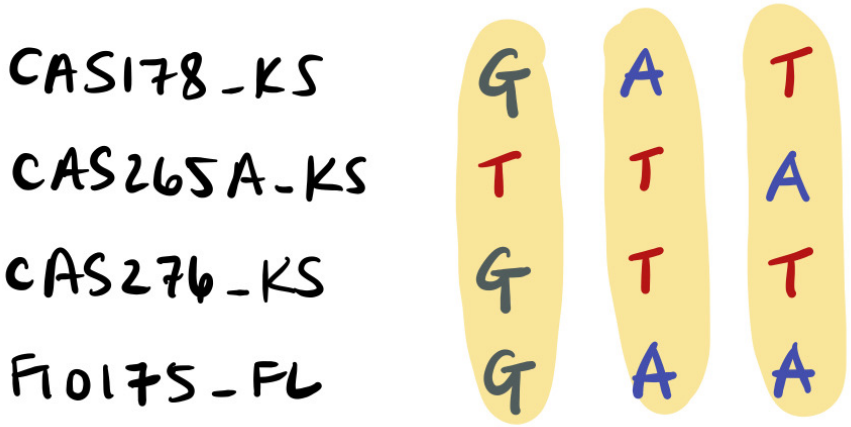
iPyrad's .loci file not only gives us information about which SNPs are contained within each locus, but it also tells us which SNPs are autapomorphies and which are synapomorphies (PIs)

This is what the .loci file looks like:

```
CAS178_KS      AATTCTCAAATGATGTGTAAATATATTGATTTCTGACCT
CAS265A_KS     AATTCTCAAATGATTTGTAAATATATTGTTTACTGACCT
CAS276_KS      AATTCTCAAATGATGTGTAAATATATTGTTTTCTGACCT
F10175_FL      AATTCTCAAATGATGTGTAAATATATTGATTACTGACCT
//                             –          *   *        | 0 | ←—— locus number
```

Site 1 is an autapomorphy ("variable") (–) and sites 2 and 3 are synapomorphies ("phylogenetically informative") (*)

The iPyrad .loci file



3 SNPs, 3 loci

Site 1 is variable (–) and
Sites 2 and 3 are phylogenetically informative (*)

```
CAS178_KS       AATTCTCAAATGATGTGTAAATATATTGATTTCTGACCT
CAS265A_KS      AATTCTCAAATGATTTGTAAATATATTGATTTCTGACCT
CAS276_KS       AATTCTCAAATGATGTGTAAATATATTGATTTCTGACCT
F10175_FL       AATTCTCAAATGATGTGTAAATATATTGATTTCTGACCT
//                             –                      |0|
CAS178_KS       GCTGCTCTCGACCCCGTTCTCATTGAGGACAAGGATAAG
CAS265A_KS      GCTGCTCTCGACCCCGTTCTCTTTGAGGACAAGGATAAG
CAS276_KS       GCTGCTCTCGACCCCGTTCTCTTTGAGGACAAGGATAAG
F10175_FL       GCTGCTCTCGACCCCGTTCTCATTGAGGACAAGGATAAG
//                                 *                  |1|
CAS178_KS       TAGACAGTGTGCAACGAAGAAGACTGGAAGGTAAATTGT
CAS265A_KS      TAGACAGAGTGCAACGAAGAAGACTGGAAGGTAAATTGT
CAS276_KS       TAGACAGTGTGCAACGAAGAAGACTGGAAGGTAAATTGT
F10175_FL       TAGACAGAGTGCAACGAAGAAGACTGGAAGGTAAATTGT
//                    *                               |2|
```