

## SNP Discovery

### Background:

1. Single Nucleotide Polymorphisms (SNPs) are discovered by comparing aligned DNA sequences. These sequences must be allelic (have variation!)
  - a. Physical Mapping
  - b. Functional Studies
  - c. Association Studies
  - d. Evolutionary Studies

The ability to locate a SNP will depend on a couple of factors:

- a. Frequency of the less-common allele (minor allele frequency)
  - i. If T = .65 and C = .35, then MAF = .35
  - ii. Harder to find rare alleles
- b. Number of individuals sampled

### 2. Sequencing Approaches

- a. Which DNA to sequence?
  - i. Genomic

BAC (Bacterial Artificial Chromosome)  
RRG (Reduced Representative Genome)  
Shotgun Sequencing

- ii. mRNA -> cDNA libraries

- b. Sequencing Methods

- i. First Generation Method - Sanger sequencing

During sample preparation, different sized fragments of DNA are made - each starting from the same location. Each fragment ends with a particular base that is labeled with one of four fluorescent dyes corresponding to that particular base. Then all of the fragments are distributed in order of their length by driving them through a gel. Information regarding the last base is used to determine the original sequence.

Under standard conditions, this method results in a read length that is approximately 700 bases on average, but may be extended to 1,000 bases. These are relatively long read lengths compared with other sequencing methods. However, first generation sequencing is limited by the small amounts of data that can be processed per unit of time, referred to as throughput.

- ii. Next-Generation Sequencing

Commercial second generation DNA sequencing tools emerged in 2004 in response to the low throughput of first generation methods. To address this problem, second generation sequencing tools achieve much higher throughput by sequencing a large number of DNA molecules in parallel. Roche 454 Sequencing

1. Illumina Solexa
2. ABI SOLiD
3. Roche 454
4. Heliscope/Pacific Biosciences (3rd Generation)

## SNP Discovery: Using BWA and SamTools

We will start processing our Next Generation sequence data first since this process takes the longest to complete. In this lab, we will be working with Roche 454 data from e.coli downloaded from the DnaNexus server (the SRA archive).

**//Make sure Xming is running prior to connecting via Putty**

All of the files that you will use today are in the lab8.tar.gz file in your home directory on plantgenome.plantsciences.ucdavis.edu  
First uncompress and untar this file using a combined command: `tar xvfz filename.tar.gz`

Go back into the lab8 directory and go into the bowtie\_samtools directory.

You will find:

`NC_000913.fna` E.Coli reference genome file  
`SRR001355.fastq` 454 data in Fastq format (reads and quality scores have already been merged)

Make a **454 directory** and move the two files into it (`mkdir`).

Index the reference genome for bwa:

`bwa index NC_000913.fna`

Map the run to the reference to generate a sam file:

`bwa bwasw NC_000913.fna SRR001355.fastq | gzip > 454-1.sam.gz`

**//Start Sanger Exercise Here!**

Using samtools to generate a bam file from the sam file:

`samtools calmd -uS 454-1.sam.gz NC_000913.fna | samtools sort - 454-1`

Index the reference genome for samtools:

`samtools faidx NC_000913.fna`

Prepare bam file for TVIEW and IGV:

`samtools index 454-1.bam`

**To view in TVIEW**, make sure X-ming is running and type:

`samtools tview 454-1.bam NC_000913.fna`

You can scroll across the sequences using the arrow keys.

Hit escape twice to quit tview.

**To view in IGV (Integrated Genome Viewer)**

Download your 454-1.bam, 454-1.bai, NC\_000913.fna onto your desktop.

Run IGV from <http://www.broadinstitute.org/igv/>

Import Genome with:

ID: `ecoli_<initials>`

Name: `ecol_NC00913`

Choose the NC\_00913.fna for FASTA file

ID \*  (unique id, e.g. hg18)

Name \*

Fasta file is a directory

Fasta file \*  ...

Cytoband file  ...

Gene file  ...

Alias file  ...

Supply a sequence URL if defining a web-hosted genome (optional, not common). See user guide for more details.

Sequence URL

\* required.  
The sequence file (required) can be a FASTA file, a directory of FASTA files, or a zip of FASTA files. Optionally, specify a cytoband file to display the chromosome ideogram and an annotation file to display the gene track. See the documentation for descriptions of supported annotation formats.

Then 'Load with File' and choose the 454-1.bam file.

Zoom in more than halfway with the top right zoom tool.

Scroll across the program and what does the letters/lines represent?

## Sanger SNP Discovery

### Overview of software tools

This project will contain three sets of Sanger sequenced data for SNP discovery. The first is haploid (extracted from megagametophyte tissue) DNA from loblolly pine (*Pinus taeda*). The second set contains diploid (extracted from leaf tissue) DNA resequenced (forward and reverse) from black cottonwood (*Populus trichocarpa*). In both sets, we are looking at one amplified region (amplicon) across multiple individuals.

**Phred**, **Phrap**, **Consed**, and **PolyPhred** are UNIX programs from the University of Washington that work as a group for the analysis of DNA sequences. They perform the following functions:

**Phred**: Base calling and quality assignments.

**Phrap**: Sequence assemblies and new quality assignments.

**Consed**: Visual X-Windows graphic interface, to view and edit alignments and contigs, and to view the original traces.

**PolyPhred**: Finds polymorphisms in Phrap contigs, gives quality calls, adds data to Phrap files to permit Consed finding and visualizing polymorphisms.

### PHRED

Phred observes the base trace, makes base calls, and assigns quality values (qv) of bases in the sequence.

- a. Writes base calls and qv to output files for Phrap assembly.
- b. Quality values (phred scores) range from 0 to 60. 20 and above is considered a confident base call (1 in 100 chance that is has been called wrong ~99% accuracy)
- c. Useful for consensus sequence construction
- d. Formula:  $q = -10 \times \log_{10}(p)$   
where  
 $q$  - quality value  
 $p$  - estimated probability error for a base call
- e. Why use Phred?  
Output sequence might contain errors.
  - I. Vector contamination
  - II. Dye-terminator reaction might not occur.
  - III. Weak or variable signal strength of peak corresponding to a base.

### PHRAP

Phrap constructs the contig sequence as a mosaic of the highest quality parts of the reads rather than a statistically computed "consensus".

- a. Avoids the complex algorithm issues associated with multiple alignment methods
- b. Speed and accuracy
- c. Phrap is an assembler NOT an aligner
- d. The sequence produced by Phrap is quite accurate (less than 1 error per 10 kb in typical datasets)
- e. Phrap considers sequence quality at a given position (determined by Phred)

### CONSED

Consed is a program for viewing and editing assemblies produced by Phrap. Currently supports Solexa, 454, and Sanger Reads and allows mixtures of these read types

- a. Assembly viewer - allows for visualization of contigs, assembly (aligned reads), quality values of reads and final sequence.
- b. Trace file viewer - of a given sequence or several reads.
- c. Navigation – identify and list regions which are below a given quality threshold, contain high quality discrepancies, single-strand coverage, etc.
- d. Autofinish – automatic set of functions for: gap closure, improvement of sequence quality, determination of relative orientation of contigs, identification of regions covered by a single read or by reads of a single strand. The program automatically performs primer picking and chooses the templates.

### Preparing the Files/Directories

In the **lab8** directory, create a directory called **cad**.

In the **cad** directory, create two more directories: **edit\_dir** and **phd\_dir**.

In the **lab8** folder, go into the **snp** directory and you will find **tracefiles** which will contain two folders relating to two different candidate loci: **cad** and **sam2**. Each directory represents a locus of interest and contains tracefiles of forward and reverse sequences of several individuals.

Since we are working with the candidate locus **cad**, we need to move the **cad tracefiles** into the **cad** main subdirectory, and rename them as **chromat\_dir** subdirectory.

Now we need to move the **cad poly\_files** into the **cad** main subdirectory, and rename them as **poly\_dir** subdirectory.

### CONSENSUS SEQUENCE

When **Polyphred** is used here for SNP and polymorphism determination, it is most useful to have the “wildtype” sequence available for comparison and to provide a “backbone” for the multiple sequence alignment. This is often not available, especially for species without a complete genome. In this case, you will use the best possible consensus sequence derived from multiple alignments. This sequence has been provided for you and is in FASTA format. What is needed is a phred file **\*.phd** file for this sequence.

To convert the FASTA formatted sequence to a phred file, a script has been provided with the phredPhrap package. Incidentally, phred and phrap come with many scripts that can assist users with some basic file manipulation and sequence analysis.

Go into the **consensus** directory and use the Perl script **fasta2Phd.perl** to convert the **cad.fasta** file to a **cad.phd.1** file:

```
fasta2Phd.perl cad.fasta
```

This script will quickly complete and you should see the following:

```
[malenita@plantgenome consensus]$ fasta2Phd.perl cad.fasta
creating file cad.phd.1
[malenita@plantgenome consensus]$ █
```

Copy this phd-formatted file to the **phd\_dir** subdirectories for the candidate locus

This **cad.phd.1** file must be used in the **Phred, Phrap, Consed, Polyphred** analysis.

tran

```
cp cad.phd.1 ../cad/phd_dir/
```

#### EXECUTION OF PHRED AND PHRAP

The **Phred** and **Phrap** packages can be run independently or as one combined scripts. Most users use the combined option unless they are making specific modifications to one script of the other. The combined script, **phredPhrap** must always be run from inside the **edit\_dir**. Go into the **cad's edit\_dir** and run:

```
phredPhrap -forcelevel 0
```

This executes RepeatMasker, Phred, Phrap, in that order. Forcelevel is just one of the many parameters that can be selected. The higher the forcelevel, the more liberal the assembly (generally produces fewer contigs). If it is necessary to increase the forcelevel, it often means the sequence quality is low in specific regions or there are certain reads causing a problem. If an assembly is more forced, more manual editing will be necessary downstream and the polymorphisms should be carefully reviewed to ensure they reflect true sequence variation.

You will see a lot of messages come up to the screen as the programs executes... Once finished, examine your new files:

You should see the following:

```
[malenita@plantgenome edit_dir]$ ls
cad.fasta                cad.fasta.screen.contigs.qual  cad.fasta.screen.singlets
cad.fasta.log            cad.fasta.screen.log           cad.fasta.screen.view
cad.fasta.screen         cad.fasta.screen.problems      cadNewChromats.fof
cad.fasta.screen.ace.1   cad.fasta.screen.problems.qual cad.phrap.out
cad.fasta.screen.contigs cad.fasta.screen.qual          cad.screen.out
[malenita@plantgenome edit_dir]$
```

Some of the outputs of the **phredPhrap** run contain:

**.contigs** – FASTA file containing the contigs and singletons. The contigs contain multiple reads and the singletons are those that match one of the contigs but could not be merged consistently.

**.singlets** – FASTA file of the single unmatched reads

**.ace** – formatted for viewing the full assembly with tags in Consed

**.view** – formatted viewing the assembly with tags in Phrapview (another viewer)

#### EXECUTION OF CONSED

Make sure you are still in the **edit\_dir** subdirectory for the appropriate run

```
consed
```

Open the **\*.ace** file. Take some time to look at your alignment. Select one of the contigs and view the coverage and sequence quality.

Try scrolling back and forth. Try scrolling by dragging the thumb of the scrollbar. Also try scrolling by clicking on the 4 << > >> buttons for scrolling by small amounts. For scrolling by tiny amounts, click on the arrows at either end of the scrollbar. For scrolling by huge amounts, use the middle mouse button and just click on some location on the scrollbar.

Notice the colors. The bases that are in red are the ones that disagree with the consensus.

These quality values are shown in grey scales:

Quality 0 through 4 is given by dark grey

Quality 5 through 9 is given by a shade lighter

Quality 10 through 14 is given a a shade still lighter

...

Quality of 40 through 97 is given by white (the brightest shade)

The ends of reads shows bases that are grey and have a black background. These are the low quality ends of reads or the unaligned ends of reads, as determined by phrap.

Blue: agrees with consensus

Orange: disagrees with consensus

Yellow: this stretch of this read was used to form the consensus

Grey: Low quality or unaligned ends of reads

Consed also contains a large array of editing functions, personalized tags, and primer design.

Use the **Quit Consed** command from the **Consed** Main Menu to shut down **Consed**.

#### EXECUTION OF POLYPHRED

**Polyphred** is one of several open source options for SNP calling that integrate well with **Phred**, **Phrap**, **Consed** pipeline, adequate for Sanger or long-read resequencing data (could be used for 454).

*The detection of single-nucleotide polymorphisms (SNPs) and short insertion/deletions (INDELs) in DNA sequences is challenging because one must align and compare sequences from varied sources, and differentiate true polymorphisms from sequencing errors.*

**PolyPhred** - identifies potential heterozygous single-base substitutions by going the all bases in the Phrap-generated contigs, and examining the information about the sequence quality and peaks in each trace. For increased accuracy, PolyPhred ignores the lower quality sequences at the beginning and end of sequence traces.

Run **PolyPhred** from the **edit\_dir** subdirectory:

```
polyphred -ace *.ace.1 -tag p -snp hom -indel -f 50 > *.polyphred.out
```

where **\*.ace.1** is the .ace file present in the **edit\_dir** directory.

NOTE: This command will give polymorphisms of quality ranks 1 through 6, where 1 is the highest quality and 6 is the lowest quality.

The qualifier **-f 50** is used to list 50bp of flanking sequence on each side of the detected polymorphism.

The qualifier **-indel** is used to identify insertions or deletions.

The qualifier **-snp hom** is used to identify homozygous SNPs only (haploid DNA).

The qualifier **-tag p** is used to list the tagged polymorphisms in the text file **\*.polyphred.out**

To see ALL polymorphisms (ranks 1-6), add the **-rank 6** option in the above command. Polymorphisms of ranks 4, 5, or 6 are very seldom real, as is true also of many of rank 2 or 3.

Lots of messages will come up to the screen as **PolyPhred** executes.

The output is written to a text file with the extension: **polyphred.out**

**PolyPhred** also MODIFIES the **\*.fasta.screen.ace** file, the **\*.phd** files in the **phd\_dir** directory for sequences with polymorphisms, and the **\*.poly** files in the **poly\_dir** directory for sequences with polymorphisms.

Examine the contents of the **\*.polyphred.out** output file and look for the section beginning with:

**BEGIN\_POLY**

Examine these data to the end of this section:

**END\_POLY**

You should see the following:

```
[malenita@plantgenome edit_dir]$ more cad.fasta.screen.ace.1.polyphred.out
BEGIN_MESSAGE

BEGIN_HEADER
-----
POLYPHRED      Version 5.02          Build 2005.07.11
-----
POLYPHRED_THUMBPRINT  1192995210
TIME                21/10/07 12:33:30
CURRENT_DIRECTORY   /home/malenita/Lab9files/cad/edit_dir/
END_HEADER

BEGIN_COMMAND_LINE
-dir      /home/malenita/Lab9files/cad/
-ace      cad.fasta.screen.ace.1
-score    70
-quality  25
-window   20
-tag      polymorphism
END_COMMAND_LINE

BEGIN_CONTIG
Contig5

BEGIN_POLY
230      AAAGAAGAAGCCATGGAAGTCTCGGCGCCGATGCTTATCTTGTAGCAA  A  G  GATACTGAAAAGATGATGGTGCGCCAATTTAATT
CGATAACATTGTTTCT 75
END_POLY
```

These data show the position of a polymorphism in a contig, 5' and 3' sequences, the SNP, and quality of the polymorphism.

**NOTE:** There are several other types of polymorphisms that can be identified and tagged using **Polyphred**. All of these polymorphisms can also be viewed with **Consed**.

**Consed** is used to examine visually the ab1 chromatogram characteristics for each potential polymorphism to decide if a given polymorphism is true variation. Annotations or comments can be made to the above file to designate decisions made concerning potential polymorphisms.

#### 9. MODIFICATIONS TO FILES DURING POLYPHRED RUN

To view modifications in the other files made by **PolyPhred**, examine the end of the contents of the **\*.fasta.screen.ace.1** file:

You should see the following:

```

RT{
Ncad_F8yR2_8-33 matchElsewhereLowQual phrap 138 206 071021:122632
}

RT{
Ncad_F8yR2_8-33 matchElsewhereLowQual phrap 9 195 071021:122632
}

WA{
phrap_params phrap 071021:122632
/apps/bin/phrap cad.fasta.screen -new_ace -view -forcelevel 0
phrap version 0.990329
}

CT{
Contig5 polymorphism polyPhred 250 250 071021:123330
COMMENT{
75
C)
}

CT{
Contig5 indelSite polyPhred 62 62 071021:123330
}
[malenita@plantgenome edit_dir]$ █

```

The .ace file will provide information on the time, version, and parameters of **Phrap**. You will also be able to find information on any SNPs and indels identified by **Polyphred**. The ace file is a good resource for determining the parameters applied. Multiple .ace files can be generated and compared depending on the application.

The \*.poly files found in the **poly\_dir** were also modified. The first line of the poly files contain the name of the corresponding trace file, followed by five numbers. The first number is the minimum of the following four numbers. These four numbers are scaling factors for the A trace, C trace, G trace and T trace, respectively. The remaining lines have information for each base in the sequence.

Go into the **poly\_dir** and examine the **CAD-F2y+R2\_8-76.poly** file.

**You should see the following:**

```

[malenita@plantgenome edit_dir]$ cd ..
[malenita@plantgenome cad]$ cd poly_dir
[malenita@plantgenome poly_dir]$ more CAD-F2y+R2_8-76.poly
CAD-F2y+R2_8-76 17644.195080 18553.737029 19416.362538 17644.195080 20831.419940
C 7 19803.794376 1.000000 A 2 16504.610441 165046.104409 1406.642935 3270.762420 0.000000 1
628.694930
G 15 265261.631420 13.394485 C 20 2643.492914 0.092733 0.000000 224.024823 45015.589124 116
9.319437
C 30 2867.517738 0.020118 T 26 1085.796620 0.038089 0.000000 806.489364 1627.069486 41.7614
08
N 40 -1.000000 -1.000000 N -1 -1.000000 -1.000000 0.000000 0.000000 986.102719 0.000000
N 51 -1.000000 -1.000000 N -1 -1.000000 -1.000000 0.000000 0.000000 0.000000 0.000000
C 66 761.684399 0.007936 N -1 -1.000000 -1.000000 0.000000 179.219859 0.000000 0.000000
C 80 1881.808515 0.026073 N -1 -1.000000 -1.000000 0.000000 448.049647 0.000000 0.000000
N 86 -1.000000 -1.000000 N -1 -1.000000 -1.000000 0.000000 0.000000 0.000000 0.000000
N 98 -1.000000 -1.000000 N -1 -1.000000 -1.000000 0.000000 0.000000 0.000000 0.000000
A 111 12097.129244 0.208157 C 114 6272.695051 0.211086 2297.516794 627.269505 0.000000 0.00

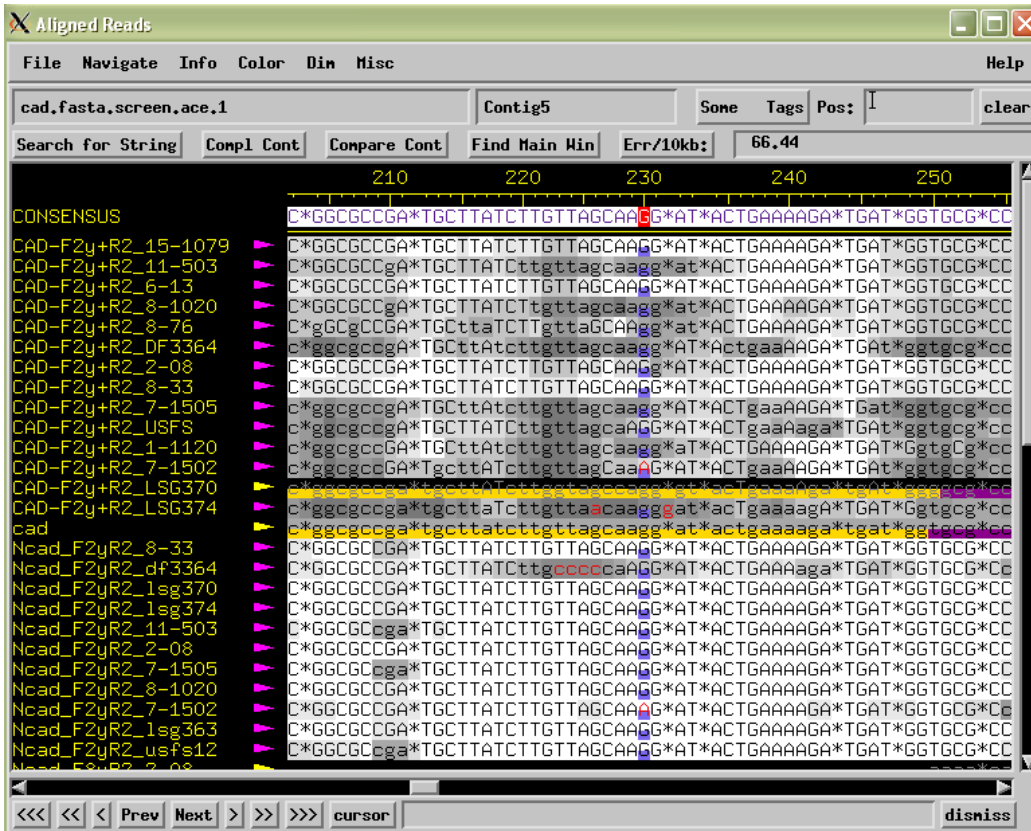
```

## 10. EXECUTION OF CONSED TO VIEW POLYMORPHISMS

Return to the edit\_dir and type in:

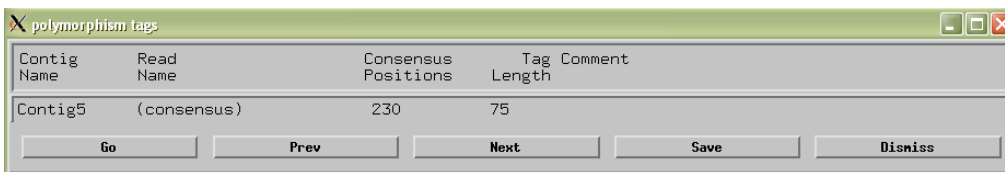
`consed`

In **Consed**, call up the appropriate \*.ace file and open **contig5**.

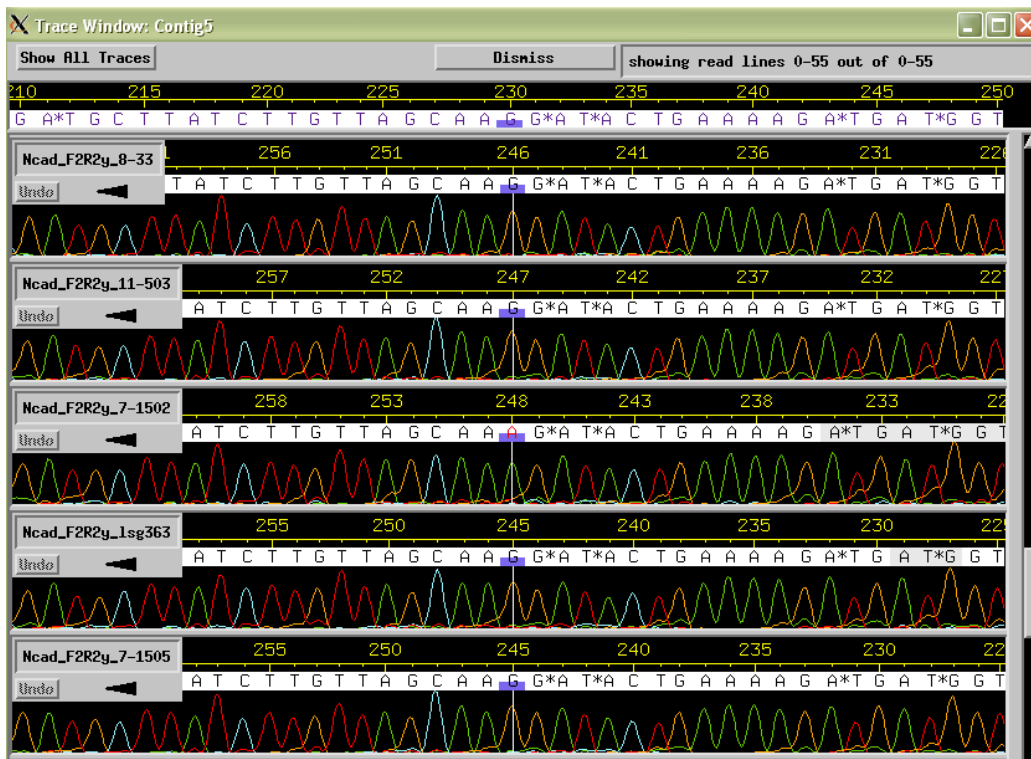


In the **Aligned Reads** window of **contig5**, under the **Navigate** pull-down menu, select **Toggle feature: when navigating to consensus location, pop up all traces (currently off)**. This will turn the feature on.

Again, under the **Navigate** pull-down menu, select **Tags**, and then in the **Select Tags** window select **polymorphism, OK**. This opens a new window called **polymorphism tags**. This window lists all individuals who have a putative SNP, and gives the consensus location for each SNP.



Double-click on the consensus location for a SNP in the **polymorphism tags** window. This will bring up all ab1 traces for each individual at this location.



Scroll through the traces and visualize each lane to determine if the SNP is real or not. The lanes thought to have a SNP by **Consed** are tagged in a bluish-purple color.

If a given SNP is determined to be “real”, then genotype information for the SNP could be copied from the **\*.polyphred.out** file into a Master file for the candidate locus. A SNP that is identified by **PolyPhred** but is not real is considered a false positive. A SNP that is true but is not called is considered a false negative.

### Diploid Sequences

In the **lab8** directory, look for the **hct6** directory. This directory has already been prepared with the required base directories, and tracefiles have been moved to the **chromat\_dir**. This set is ready for **phredPhrap**. Note: this set is from black cottonwood, leaf tissue and is thus, diploid sequence.

### Short Exercise in SNP Annotation

[gi|203308514](#)

This sequence has been identified with two SNPs:

Position 12 (C/T) and Position 32 is (G/A).

There is no reference genome

What is the species? What type of sequence? How long is the sequence?

Compare this sequence against a well-curated protein repository at NCBI.

Are these SNPs in coding regions?

Are they synonymous or non-synonymous?

Do the same for **hct6**.