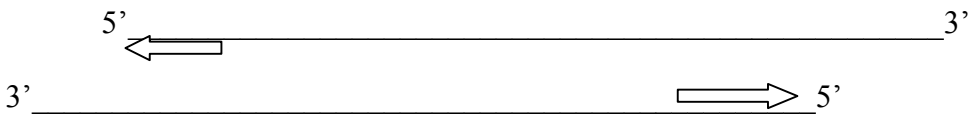
END DETERMINATION

Your sequence is in one contig. Primers have been run and coverage is sufficient. There are no weak areas (gray colored bases in the consensus).

The Consed screen shows two possible choices for ends: defined ends or circular (these represent more than two possible physical realities, but for the purposes of annotation, these two are all we consider). Read up on circularly permuted, cohesive, and terminally redundant ends to find out what is happening in the phage and bacterial infection.

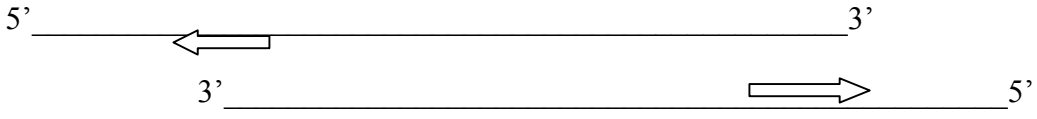
Defined ends are characterized by a pile-up of clones that begin at each end (they are overrepresented). On the left end on rightward reads, one can sometimes find the "GAT" half of the EcoRV cloning site, and on the leftward reads at the right end one can sometimes see the "ATC" half of the EcoRV cloning site.

For defined ends, in the actual phage particle, there is usually a "sticky" overhang. Each end is single stranded and its sequence is complementary to the other end. There can be 5' or 3' extensions. Because the polymerase that does the sequences only adds bases 5' to 3' direction, this means that the sequence "falls off" the 3' extensions and all of the bases are not read. A false "A" is added to rightward reads and a "T" to leftward



Remember the polymerase is using the one strand as a template, but producing something equivalent to the other strand. In the diagram above, the leftward facing arrow is copying the top strand, but generating a sequence that is found on the bottom strand. Except it reaches an end of the top strand and "falls off" without making the few bases on the lower left. Only by ligating the DNA (joining the sticky ends) and then sequencing can you learn what those bases are (10 is a common number of bases present in the overhang.)

The situation is different with 5' overhangs. The polymerase reads to the ends and then falls off. Ligating and sequencing yields no new bases, the read just "goes around the block" and picks up reading the other end, which will show up, right under the false A or T.



The second possibility of Consed is that you see the same sequence at both ends. This can mean your genome is circularly permuted (each genome has somewhat more than one unit length genome, and where these begin is not the same for each phage particle) or terminally redundant (actually has the same sequence at each end, and difficultly enough, this may or may not be the sequence that is shown as the ends on Consed). In practice, we usually treat these cases the same, unless we know it is terminally redundant and then we do biochemistry (DNA Digests, etc) to figure out what the repeated sequence actually is. So what do we do with this type of genome? First Jen or Alexis cuts it to "unit length" essentially throwing away the repeats at each end leaving one phage genome. Now…where do we define base one? It depends! Try to find a terminase. Take the DNA sequence as a text Fasta file and blastX it against the terminase database. Also, see if it is similar to other phage genomes. (BlastX chunks of it against genbank and the pbi database, then if you find a similar phage, try aligning two sequences in Blast or dottering those two sequences.) For highly similar phages, define your base 1 according to similarity with their base 1. (Cut off the preceding bases and paste them to the end)
If there are no highly similar phages, use terminase as gene 1. If you can find a large subunit, search around the surrounding open reading frames and blastp their products to see if one is a small terminase subunit. If so, use it as base 1.

Some other things to consider: try to not end the genome with a "wrap-around" gene. This becomes a pain later. It's a lot easier to annotate rightward facing orfs, so look at the overall pattern before finalizing base 1. You might want to flip the whole genome around before getting going! To address these, it is good to run glimmer before printing your 6 frame to be sure you like where you have called base 1.

First choice for end: physical end if one is known. Three prime sticky ends are by convention placed at the right end. Five prime sticky ends can go at the left end. Only one copy of the extensions should be in the fasta text file of unit length genome.

Second choice for end: near identity to a highly similar phage

Third choice for end: terminase small subunit

Fourth choice for end: terminase large subunit if you cannot identify a small subunit

Finally: orientation or architecture (leftward arms typically go on the right of the genome and rightward arms go on the left end of the genome.) Not all phages have two opposing transcriptional units.