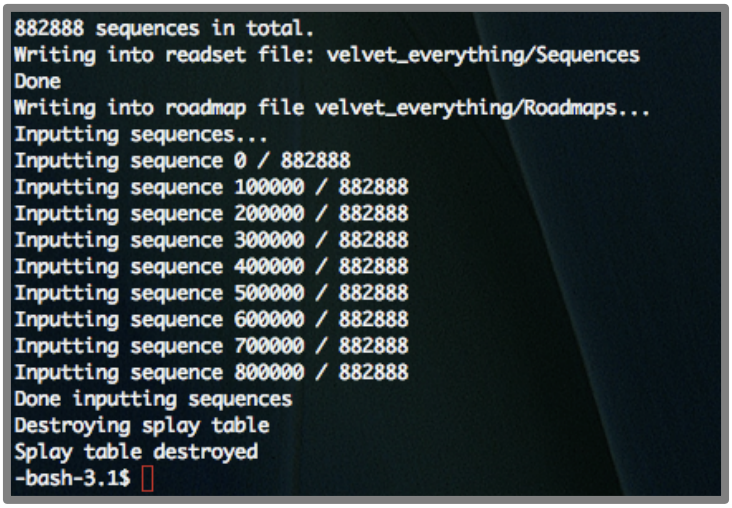
Bioinformatics 465 : Midterm

Duane Johnson

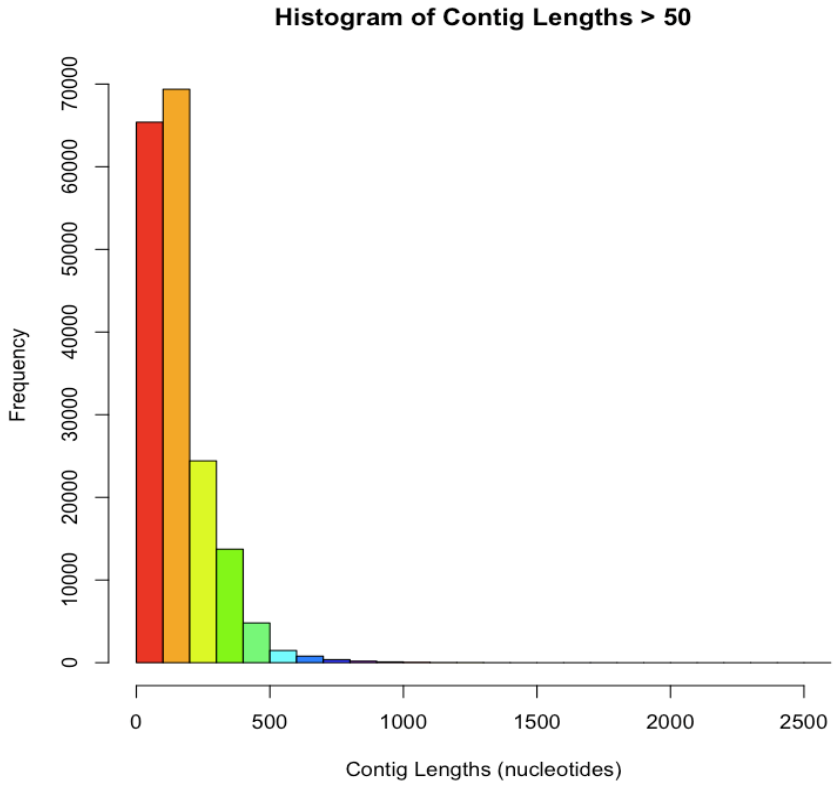
Feb. 26, 2009

1. Using Velvet on psoda64, I assembled the sequences from the raspberry reads into 370,301 contigs. The assembly is a good start, but it needs more depth in order it to be more complete--and for us to be more confident in the assembly of the raspberry genome. The 454 reads should have about 8x coverage in order to have reasonable surety of a consensus.



Running velvetg on psoda64 using the raspberry sequence data

Unfortunately, even with the "long" reads from the 454 machine (381 bp long, i.e. 337 Mbp / 882,000 sequences), we are still not be able to reassemble any long repeat regions in the genome. After assembly, velvet produced contigs averaging 110 bp, with a max of 2570.



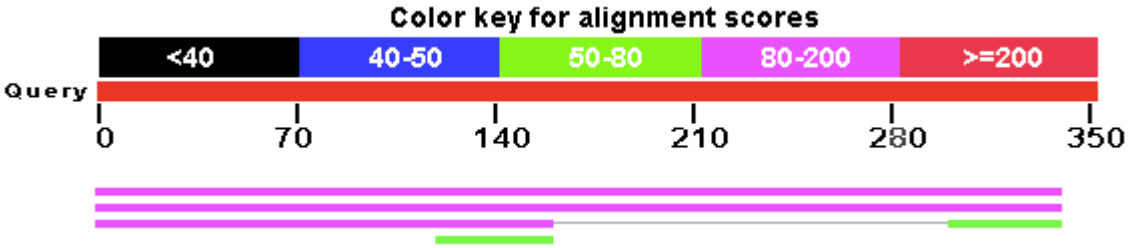
By removing contigs of trivial length (< 50 nucleotides), I generated the above histogram confirming that the distribution is right skewed--most of the contigs are in the range 50 to 450 nucleotides long.

2. Using NCBI's ORF Finder, I tried about a dozen contigs and then mapped them through NCBI's Blast program. Only one of the contigs yielded a positive Blast search result:

```
 37 atgccggcgccgagcagctgcggaatgacgctgacgatcacctcg
    M  P  A  P  S  S  C  G  M  T  L  T  I  T  S
 82 tcgagcagcccggccgccaggcaactaccggcgaggctgccgccg
    S  S  S  P  A  A  R  Q  L  P  A  R  L  P  P
127 ccggccagccagacccgccggcagccctgttcgccgaggcgtgcg
    P  A  S  Q  T  R  R  Q  P  C  S  P  R  R  A
172 aggccttcctggggcgtgtcatggcgcaactcgacgccttccacc
    R  P  S  W  G  V  S  W  R  N  S  T  P  S  T
217 gcgctctcccggggattgcgggtgagcacctggcagggcttgccc
    A  L  S  R  G  L  R  V  S  T  W  Q  G  L  P
262 ggatacggccagtcgccgaagccgcgcacgatatcgtag
    G  Y  G  Q  S  P  K  P  R  T  I  S  *
```

Thus, this is likely to be a gene because it has postive selection in another very unrelated organism. In addition, the start (atg) and stop (tag) codons are clearly present in this sequence, so at least it matches our nicely laid out central dogma of genetics (mRNA / ribosomes / transcription etc.).

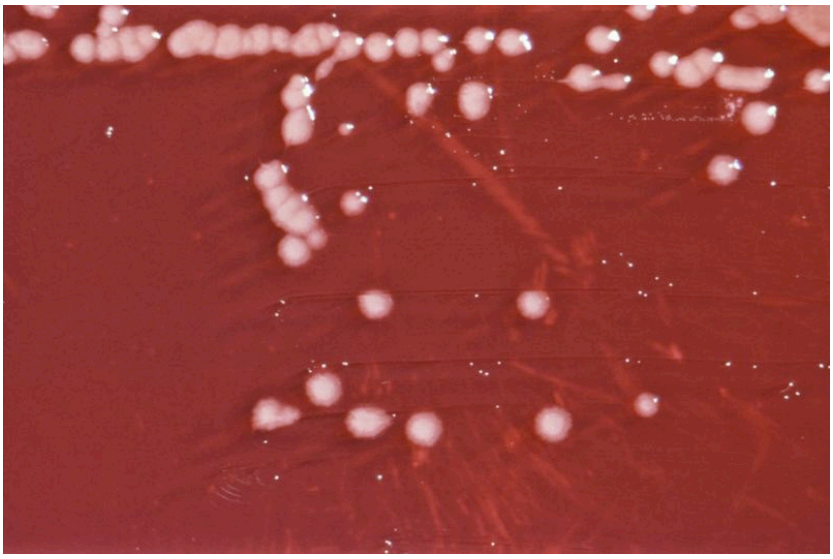
3. The match was made with a gene from Pseudomonas aeruginosa, called LESB58:



Wikipedia describes the bacteria:

> ***Pseudomonas aeruginosa*** is a common bacterium which can cause disease in animals and humans. It is found in soil, water, and most man-made environments throughout the world. It thrives not only in normal atmospheres, but also with little oxygen, and has thus colonised many natural and artificial environments.

The matched gene's functions have not yet been identified. Nevertheless, it is quite likely that this is an accurate match, as the sequence of 352 nucleotides has an E value of 2e-28. It is extremely unlikely that such a match could occur by chance alone.



Pseudomonas Aeruginosa