WVU GENOMIC DATA ANALYSIS

How to Choose a Good NGS Data Set

Access to a good set of sequence data is a crucial aspect of applied bioinformatics, but what makes a good data set anyway? While part of the answer depends upon your ultimate goal for the data, a few qualities are generally important:

1. Read length. Are the reads long enough for your purpose? In general, longer reads are almost always better. For Illumina reads, 100-150nt is a decent trade-off between length and quality loss: reads over 200nt are often degraded at their 3' ends which counteracts the gain from increased length. In addition, paired end (PE) reads are desired for most downstream analyses, although single-end (SE) data can be used for some re-sequencing or variant mapping protocols.

2. Read quality. Read quality is measured at each position as a probability that the base call is incorrect. A nominal quality threshold is Q>25; a superb threshold is Q>30. If you have to trim or throw away a significant number of reads because of poor quality, that decreases the amount of data you have to analyze. Eventually this could become a problem, though it is rare these days thanks to the immense output of most NGS platforms.

3. Read depth. Using the total number reads and the estimated (or known) genome length, you can calculate the mean coverage of your genome (also sometime called "read depth"). Increased read depth can contribute to a better assembly, though this is not always the case (it doesn't solve sequencing bias, for example). In general, read depth should be at least 50X, though the coverage necessary for a quality assembly varies widely from genome to genome.

4. NGS platform. Illumina is the most popular method for generating sequence data, and most algorithms are optimized for Illumina data. Long reads (PacBio, MinION) are growing in popularity but often require specialized tools and configurations.

Suggested Illumina Data Sets for Different Applications

De novo genome assembly	PE100-150	100X coverage	
Variant calling	PE50	200X coverage	
De novo transcriptome assembly	PE150	200X coverage	
Differential RNA-seq	PE50-PE100	50X coverage	5-20 reps/condition
Metagenome assembly	PE150-200	300X coverage (*est.*)	