What does coverage refer to in a sequencing project?

- **Assume we have a 1 Gigabase genome. That is one billion nucleotides.**
  - **1x coverage** = you collect 1 billion nucleotide worth of sequence data

Illumina

- **150 bp paired end reads**; each end is a read
  - **1x coverage** of genome = 6,666,667 reads
    - =1,000,000,000 nt/150 nt/read
  - **10x coverage** of genome = 66,666,666 reads
    - Good for SNP calling
  - **30x coverage** of genome = 200,000,000 reads
    - Good for SNP calling
    - Small structural variant discovery

- **One Novaseq 6000 S1 flow cell**; one lane: ~1.4 billion reads, theoretically
  - **Experience says**: You will typically achieve 70% of that value = 980 million reads
    - Therefore:
      - You can sequence ~5 genotypes (1 Gb in size) at 30x coverage on one lane of the S1 flow cell (980 million/200 million)
        - This requires barcoding each genotype.

PacBio Revio CCS Sequencing

- **Advertise**: 90 Gb of data (15kb-18kb insert size) from a single run
  - **Typical**: ~75 Gb of data
    - 10x coverage of 1 Gb genome = 10 Gb
    - 30x coverage of 1 Gb genome = 30 Gb
    - 50x coverage of 1 Gb genome = 50 Gb
- **For a homozygous species of 1 Gb**
  - Typical to collect 30-50X coverage
  - One flow cell of the PacBio Revio would be recommended
- **For a heterozygous species of 1 Gb size**
  - Typical to collect 60x-100x coverage
  - Two flow cells of the PacBio Revio would be recommended
  - Why??
    - **Phasing**

- You need to **phase** the output to assemble each parental donor