Illumina Sequencing QC and Troubleshooting Workshop

Part 3: SAV Overview

April 17, 2017
Overview

1. Illumina sequencing overview/review
2. Common types of libraries
   – Common problems during library construction
3. SAV overview
   – Run stats, how to tell a good run from a bad one, PhiX, etc.
4. How to QC – best practices
Sequence Analysis Viewer (SAV)

• Sequencing stats/metrics during run
  – On sequencing platforms (MiSeq, HiSeq, etc)
  – Downloadable (PC only)
  – BaseSpace

• Requires several files from run folder (small)
  – InterOpt folder
  – runinfo.xml
  – runparameters.xml
Illumina SAV output

• SAV sequencing stats can be useful!
  – Overview of how well the run performed
  – Must have PhiX spike-in for most stats (calculated based on PhiX reads compared to reference)
  – Cluster density, Phasing/Prephasing, % PF, %Q30, % aligned

• Assessment of run quality before looking at actual sequencing data!
Use SAV locally
Use BaseSpace

BaseSpace Sequence Hub is a cloud-based genomics analysis and storage platform that directly integrates with all Illumina sequencers. Learn More

- Set up runs for Illumina sequencers and NeoPrep
- Monitor sequencing runs from the web or with the iOS app
- Stream data to the cloud directly from sequencers
- Analyze genomics data with pre-configured pipelines
- Share data instantaneously with anyone in the world
Terminology

• Clusters (raw): number of clusters detected through imaging
• % passed-filter (%PF): % of clusters or reads that pass a chastity filter (useable clusters)
• Intensity: Signal intensity
• FWHM: Average full-width-half-max measurement for clusters (focus)
• %>=Q30: % of bases that have a quality score greater than 30 (e.g. high-quality reads
Terminology

Based on PhiX reads alignment

• % aligned: percent of PF reads uniquely aligned to PhiX genome (should be close to the %PhiX spiked in)

• Error rate: calculated error rate based on alignment to PhiX

• Phasing/Prephasing: percentage of molecules in a cluster that fall behind (phasing) or ahead (prephasing) of the current cycle during sequencing
The importance of cluster density

- Illumina reports “optimal” cluster density for each platform
- pM amounts of libraries are used for sequencing
- Accurate QC and quantification are essential!
# Cluster Density

<table>
<thead>
<tr>
<th>Platform</th>
<th>Optimal Loading Concentration</th>
<th>Optimal Raw Cluster Density</th>
</tr>
</thead>
<tbody>
<tr>
<td>GAIlx (v5 cluster kit with SCS 2.8 or later)</td>
<td>12.0 pM</td>
<td>700-800K clusters/mm²</td>
</tr>
<tr>
<td>HiSeq 2000/2500 High Output v3</td>
<td>12.0 pM</td>
<td>750-850K clusters/mm²</td>
</tr>
<tr>
<td>HiSeq 2500 High Output v4</td>
<td>18.0 pM</td>
<td>950-1050K clusters/mm²</td>
</tr>
<tr>
<td>HiSeq 2500 Rapid Run v2</td>
<td>12.0 pM</td>
<td>850-1000K clusters/mm²</td>
</tr>
<tr>
<td>MiniSeq</td>
<td>1.8 pM</td>
<td>170-220K clusters/mm²</td>
</tr>
<tr>
<td>MiSeq v2 reagents</td>
<td>12.5 pM</td>
<td>1000-1200K clusters/mm²</td>
</tr>
<tr>
<td>MiSeq v3 reagents</td>
<td>20.0 pM</td>
<td>1200-1400K clusters/mm²</td>
</tr>
<tr>
<td>NextSeq 500 NextSeq Control Software v1.3 or later</td>
<td>1.8 pM</td>
<td>170-220K clusters/mm²</td>
</tr>
</tbody>
</table>

Nucleotide diversity

- Data processing on Illumina instruments
  1. First 4-7 cycles used to build template (cluster map)
  2. After 12 cycles the signal to noise and corrected intensity is calculated
  3. The first 25 cycles are used to calculate clusters passing filter
  4. Other stats only calculated for PF clusters
Nucleotide diversity

Initial low sequence diversity cycle

Wrong cluster boundaries inferred

Template generated based on initial low diversity cycle

Inferred cluster boundary

Later normal sequence cycle

Uncorrected cluster boundaries re-used

Wrong templates from initial cycle applied to new cycle

Actual clusters corresponding to A, C, T or G nucleotides (color-coded) at given sequence position

Strategies for Achieving High Sequencing Accuracy for Low Diversity Samples and Avoiding Sample Bleeding Using Illumina Platform
Mitra A et al. PLOS ONE. 2015 vol: 10 (4) pp: e0120520
Nucleotide diversity

- Solutions
  - Lower cluster density
  - Spike in high-diversity library
  - Back to library design
General guidelines

• Raw cluster density near optimal (or below for low-diversity libraries)
• Clusters PF >80%
• % aligned close to %PhiX spiked in
• Low phasing/prephasing (<0.5%)
• Good comparison to previous runs (on the same platform) for intensity, FWHM, etc.
Now time to play...

• Log into BaseSpace, download the NextSeq 500 V2 – RNA-Seq data set
Different types of libraries

• Amplicon, PCR-based
• Amplicon, Ligation-based
• Genomic, Ligation based (PCR free or not)
• Genomic, Nextera/Nextera XT
• RNA-Seq
• Lots of others...

They all run differently!
Why Runs Fail

• Clustering – low efficiency/can’t cluster
• Bad Quantification
• Low diversity at beginning
• Low-to-high or high-to-low diversity
• Over/under clustered
• <25 cycles
Clustering issues

• Missing/incomplete P5/P7 sequences
• Secondary structure
• Overclustering
  – Low PF, low Q30, high error rate
  – R2 failure
• Underclustering
  – Focus issues
Base diversity issues

• Low diversity at beginning
  – Problems assigning clusters

• Low-to-high or high-to-low diversity
  – Confuses instrument

• Low diversity/imbalanced index
  – Can’t distinguish indexes
Quantification issues

• How many methods were used, concordance between methods?
• How sensitive are the various methods to your sample type (2° structure, concatamers, size, etc)?
• What is in your sample and what does your method measure?
• Much more on this later...
Back to BaseSpace

• Low diversity, overclustered:
  – Run 2016_10Spear
  – 16S amplicon library (~400bp fragments, 2x251 with 8bp dual index reads), spike in 20% PhiX

• Same library at lower density:
  – Run 2016_Spear_Sharp-208_rerun
Back to BaseSpace

• Inline barcodes of same length
  – Run 170214_Erickson_Chatterjee-239P2
  – Custom small RNA library with 6bp inline barcodes followed by “T”
Back to BaseSpace

- Unable/inefficient clustering
  - Run 161205_Beeler_Tripp-225-Pippin
  - Custom ddRAD library with staggered-length inline barcodes, 15% PhiX spiked
Can (should) a bad library/run be fixed?

- Depends!
- Does Read1 work? Is the issue with the library or the loading?
- Do you know WHY it failed?
- Cost of re-sequencing (possibly multiple times) versus library prep
Questions?