Delivering Population-Scale Genomics:
A Deep Dive into HiSeqX Performance and Capabilities

Justin Abreu, Kylee Bergin, Tim De Smet, Ryan Hegarty, Matthew Coole, Laurie Holmes, Katie Sullivan, Tim Fennell, DSDE, Niall Lennon, Danielle Perrin, Sheila Dodge and Stacey Gabriel
Genomics Platform, Broad Institute, 320 Charles Street Cambridge, MA

The introduction of the Illumina HiSeqX sequencers has enabled the Genomics Platform at the Broad Institute to generate an unparalleled amount of data. To fully maximize the benefits of the HiSeq X platform, a significant effort was undertaken to scale-up the output of high quality sequencing data. In order to accomplish this task, we focused on:

- Increasing percent of clusters passing filter, while limiting data loss
- Increasing throughput and incorporating automation
- Scaling up to a 7 day process

These efforts have resulted in:

- Increase in machine utilization to full capacity
- An unprecedented amount of data output
- Improved sequencing data yield and quality

**Increased Scale Through Process Optimization**

**Goal:** Automatic exclusion amplification strip tube preparation to meet HiSeqX throughput and capacity as well as reduce variability between lanes and samples.

Automating sample strip tube creation:

- Reduces potential for sample swaps
- Reduces failures related to pipetting errors
- Reduces process time
- Capable of preparing 56 individual libraries for cluster amplification

Long term flowcell storage:

- Prepared flowcells may be stored up to 3 days at 4°C
- Inventory creation allows for sequencing runs to occur 7 days a week
- Maximizes instrument utilization by minimizing instrument downtime

**Observed Improvements:**

- Improved throughput by 384%
- Increased overall data output
- Decreased flow cell failures

**Understanding Sequencing Yield and Data Quality**

**Goal:** Increase the amount of usable bases generated per lane of HiSeqX sequencing in order to maximize the PF Gb.

- Understanding the relationship between loading concentration and % PF has aided in maximizing the output of usable data.

- Initial testing revealed % PF Clusters increased as loading concentration decreased.

- Overall duplication was observed to increase as % PF Clusters increased resulting in a decrease in usable data. This revelation sparked an effort to find a balance between % PF Clusters and % Duplication.

- New quality filters were established with the goal of determining the true performance of a WGS library on the HiSeqX.

**PCT EXC DUPE:** Percentage of bases excluded from coverage calculations because reads are marked as duplicates.

**PCT EXC OVERLAP:** Percentage of bases excluded from coverage calculations because two observations of a single base from a single insert due to overlapping reads 1 and 2.

**PCT EXC TOTAL:** The sum of the above exclusions.

**PCR-free WGS on HiSeqX**

Combining the sequencing power of the HiSeqX along with the Broad’s PCR-free WGS protocol led to the generation of data of unprecedented quality and quantity.

**Advantages**

- Proven ability to generate ~30X coverage within a single lane of sequencing.
- Improved coverage across the genome
- Reduction in base specific biases that are attributed with DNA polymerases

- Increased sensitivity to detect and reduce in false-positive observations when calling indels and copy number variants.

**Table 2:** Comparison of WGS protocol performance on the HiSeqX and HiSeq 2500.

**Table 3:** CNV and false detection call rate comparison.

**Figure 7:** SNP/Indel analysis. Analysis has shown SNP and indel analysis to be equivalent between the two protocols.

**Acknowledgements**

[Image and text related to acknowledgements]