# 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:

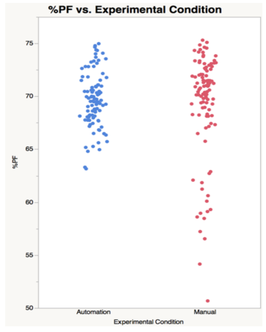
- Increasing percent of clusters passing filter, while limiting data loss
- Increasing throughput and incorporating automation
- Scaling up to a 7 day process

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:** Automate exclusion amplification striptube preparation to meet HiSeqX throughput and capacity as well as reduce variability between lanes and samples.




Automating sample striptube creation:

- Reduces potential for sample swaps
- Reduces failures related to pipetting errors
- Reduces process time
- Capable of preparing 96 individual libraries for cluster amplification

**Figure 1: %PF per lane of automation vs manual validation flowcells:** The automated striptube workflow was helped to increase the average % PF of samples and to decrease % PF variability between samples

Longterm flowcell storage:

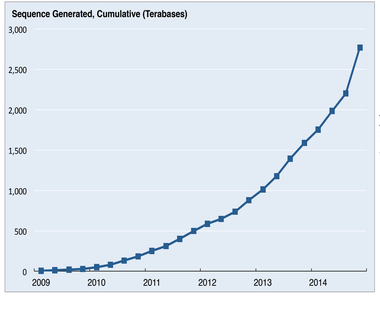
- Prepared flowcells may be stored up to 3 days at 4°C
- Inventory creation allows for sequencing runs to occur 7 days a week
- Maximizes instrument utilization by minimizing instrument downtime

**Observed Improvements:**

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

| | 2013 | 2014 | Total |
|---|---|---|---|
| 30X Equivalent Genomes Completed | 3,021 | 8,031 | ***11052*** |
| PF Bases (Tb) Generated-Genomes | 253.8 | 769.3 | ***1023.1*** |
| Run Time (Days) | 10.8 | 2.9 | |

(A)




(B)

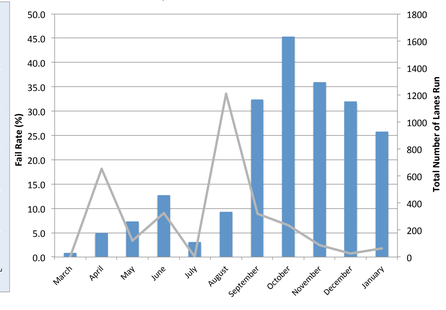



**Figure 2: Data generation at scale.** With the aforementioned process improvements in place, data generated via the HiSeqX platform were observed to increase significantly (A) while lane failure rate decreased (B) over time.

## 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.

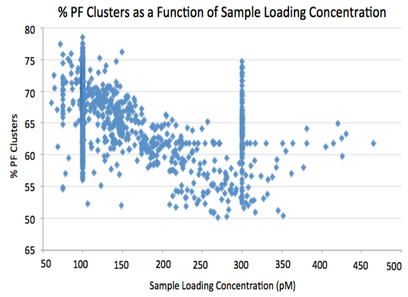



**Figure 3: Understanding loading concentration.** % PF clusters was found to increase as loading concentration decreased.

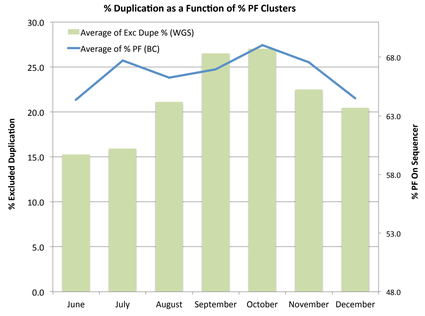



**Figure 4: Duplication and data loss.** Efforts to improve % PF Clusters resulted in an increase in overall % duplication and data loss.

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

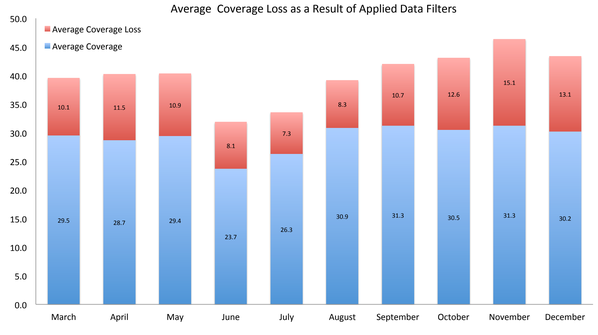



**Figure 5: True genome coverage**. Using these new data filters, we are able to truly assess how well we meet our coverage goals over time.

## 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.

| Protocol | Sequencing Technology | # Lanes of Sequencing | Library Size | Mean Coverage | % 15X | % 20X |
|---|---|---|---|---|---|---|
| PCR-free WGS | HiSeq 2500 | 2 | 9,635,612,469 | 25 | 89.3 | 75.4 |
| PCR-free WGS | HiSeqX | 1 | 6,159,578,978 | 32 | 96.7 | 94.2 |
| Standard WGS | HiSeqX | 1 | 2,803,046,955 | 28 | 92.1 | 81.7 |

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

- Improved coverage across the genome
- Reduction in base specific biases that are attributed with DNA polymerases

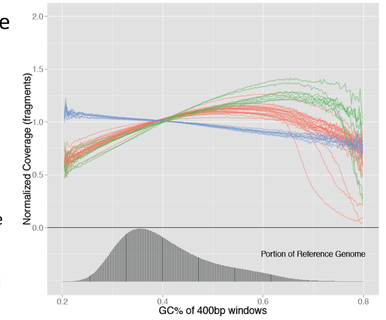


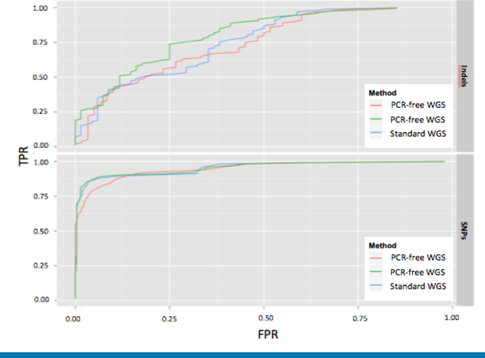

**Figure 6: GC bias comparison.** PCR-free WGS libraries (blue) show significantly more even coverage across the GC spectrum traditional WGS libraries (red and green).

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

| Condition | CNV's called (+5kb) | Estimated False Detection Rate |
|---|---|---|
| Standard WGS HISeqX v2.0 | 467 | 8.3% |
| PCR-free WGSHiSeqX v2.0 | 499 | 3.4% |

**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

National Human Genome Research Institute
illumina®