Working with the Data
Module 4: Overview
Sequencing Workflow

Sample Preparation

Cluster Generation

Sequencing

Data Analysis
The goal of this step is to take raw data (images, intensities or base calls) and turn them into raw sequence reads or alignments for use in GenomeStudio or other downstream applications.

- Real Time Analysis (RTA) produces files that contain raw sequence read of each cluster
- Pipeline takes these raw reads and aligns to a reference genome
- CASAVA aggregates sequencing runs, counts features and calls SNPs

GenomeStudio allows visualization of results
Data Analysis Workflows

<table>
<thead>
<tr>
<th>Outputs</th>
<th>Pipeline Names</th>
<th>Operating System</th>
</tr>
</thead>
<tbody>
<tr>
<td>GA and PEM</td>
<td>Images/TIFF files</td>
<td>Not Applicable</td>
</tr>
<tr>
<td>SCS/RTA</td>
<td>Intensities</td>
<td>Windows</td>
</tr>
<tr>
<td>Base Calling</td>
<td>Firecrest</td>
<td>Windows</td>
</tr>
<tr>
<td></td>
<td>Bustard</td>
<td>Linux</td>
</tr>
<tr>
<td>Pipeline Server</td>
<td>Alignments</td>
<td></td>
</tr>
<tr>
<td></td>
<td>GOAT</td>
<td></td>
</tr>
<tr>
<td></td>
<td>GERALD</td>
<td></td>
</tr>
</tbody>
</table>
Real-Time Analysis (RTA) Module

- The RTA module analyzes data as it leaves the Genome Analyzer
  - Performs image analysis, generation of cluster intensities
  - Produces base calls, including Phred-like quality scores
  - Generates reports, to assess run and library quality

- RTA simplifies the data management process
  - Eliminates the need to transfer images from computer to computer
  - Includes optional mechanisms for complete or selective archiving of images
  - Includes optional mechanisms for archiving of intensities

- RTA improves the system performance
  - Minimizes time to results – base calls and qualities generated within hours of the end of the run
  - Removes dependencies on network availability
  - Minimizes the time spent analyzing data after the run
RunFolders

- Initiated on the Genome Analyzer computer
- Transferred to the Pipeline server where additional data is computed
- Contains all the data for a particular run
- Each of the processes, image analysis, base calling and sequence analysis write their out data to the run folder
Firecrest

Module identifies clusters in image files and assigns intensities to them.
Bustard

Algorithm transforms intensities into base calls

Intensity Files

Sequence Files

Bustard
Bustard

- Cross-talk correction
  - Overlap of X & Y signal frequencies

- Normalization
  - Correcting for maximum intensity differences between X & Y
Phasing

Prephasing

Phase matrix corrects for chemistry errors
Base with highest corrected intensity is called C.
### Data Analysis Workflows

<table>
<thead>
<tr>
<th>Outputs</th>
<th>Pipeline Names</th>
<th>Operating System</th>
</tr>
</thead>
<tbody>
<tr>
<td>GA and PEM</td>
<td>Images/TIFF files</td>
<td>Not Applicable</td>
</tr>
<tr>
<td>SCS/RTA</td>
<td>Intensities</td>
<td>Windows</td>
</tr>
<tr>
<td></td>
<td>Base Calling</td>
<td>Windows</td>
</tr>
<tr>
<td></td>
<td>Firecrest</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Bustard</td>
<td></td>
</tr>
<tr>
<td></td>
<td>GOAT</td>
<td></td>
</tr>
<tr>
<td>Pipeline Server</td>
<td>Alignments</td>
<td>LINUX</td>
</tr>
<tr>
<td></td>
<td>GERALD</td>
<td></td>
</tr>
</tbody>
</table>

**Data Analysis Workflows**

- **GA and PEM**
  - Images/TIFF files
- **SCS/RTA**
  - Intensities
  - Base Calling
  - Firecrest
  - Bustard
- **Pipeline Server**
  - Alignments
  - GERALD
- **Operating System**
  - LINUX
GERALD

► GERALD module performs alignments and filtering of the output

Sequence Files

Output Files

Intensity Plots
Filtering removes low quality base calls

Chastity is the ratio of the highest intensity to the sum of the 1st and 2nd highest.

Chastity: (ideally, >0.6)

\[ C = \frac{I_A}{I_A + I_B} \]

Other Filters:
- Purity
- Similarity
- Neighbor
- Neighborhood
Downstream Applications

- Output from GERALD
  - Aggregates runs from multiple flow cells
  - Calls SNPs in areas of sufficient coverage
  - Counts incidence of sequence in features
  - Produces GenomeStudio-ready data build

- GenomeStudio allows visualization of sequencing results

- Other downstream applications from third parties
  - Assemblers
  - SNP callers
  - Data analysis
  - Visualization
Flexible Output for Myriad Applications

- Instrument Name
- Read
- Quality (symbolic)
- Aligned Locus
- Strand Designation
Genome Studio

- An integrated data analysis platform incorporating microarray and sequencing visualization and analysis tools
- A platform for analysis of genomics data from DNA, RNA and ChIP sequencing experiments, alongside microarray genotyping, gene expression and DNA methylation assay data
Genome Studio Modules

- Framework (FW)
- Genotyping (GT)
- Gene Expression (GX)
- Methylation (M)
- Protein Analysis (PT)
- ChIP Sequencing (CS)
- RNA Sequencing (RS)
- DNA Sequencing (DS)
Genome Studio Tables
(Squence, Samples, Lanes, Alleles)
Every flow cell should have a positive control with a balanced genome
  - PhiX acts as a control for both run performance and matrix/phasing

A subset of images can be saved to assist in troubleshooting.

Chastity filtering defaults to value of 0.6
  - Experimentally found to be a good balance when comparing the two highest base signals.
  - Increasing Chastity value is more stringent, producing smaller %PF, but having higher base quality values.

Run Pipeline with nohup command to retain log and error messages
  - Example: `nohup make recursive -j n &`
  - Unix based command to direct standard output to file even if ‘make’ process is interrupted or if the user logs out.

Maintain run folder structure and locations of files
Working with the Data
Best Practices from The Broad Institute
Looking at Run Data

Overview

► Using RTA 1.5

► Looking at Run Data -- Start at the top and dig deeper as necessary:
  – GERALD Summary Files
  – Intensity Plots
  – RTA Charts
  – Images
  – Internal Controls
  – At base-calling: Crosstalk Matrix
  – At base-calling: Phasing, Prephasing
  – Tracked information from throughout the process
Multiple scales are used throughout the process (especially for intensity). “Good” values depend on which scale is being used.

- First Base Report
- RTA
- GERALD Summary Files
- Intensity Plots

<table>
<thead>
<tr>
<th></th>
<th>First Base Report</th>
<th>RTA</th>
<th>GERALD Summary File (PF)</th>
<th>IVC Plots: All</th>
<th>IVC Plots: Called</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>1st cycle intensity</strong></td>
<td>~1150</td>
<td>~750</td>
<td>~100</td>
<td>~100</td>
<td>~400</td>
</tr>
<tr>
<td><strong># Clusters</strong></td>
<td>160,000 /tile</td>
<td>365,000 /mm²</td>
<td>185,000 /tile</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Real-Time Analysis (RTA) Software

- Using RTA 1.5
- Looking at Run Data -- Start at the top and dig deeper as necessary:
  - GERALD Summary Files
  - Intensity Plots
  - RTA Charts
  - Images
  - Internal Controls
  - At base-calling: Crosstalk Matrix
  - At base-calling: Phasing, Prephasing
  - Tracked information from throughout the process
Real-Time Analysis (RTA) Overview

► On-rig image processing / base calling / quality scoring

► Dashboard style real time monitoring of run
  – Opens automatically upon run start*  *We cover manually starting RTA in the lab.
  – Configuration can be modified at run start when SCS is opened.

► INPUT:
  – Image files

► OUTPUT:
  – QSEQ files one per tile
    ▪ base calls and associated quality scores
  – Cluster intensity files (.cif) one per tile per cycle
  – Cluster noise files (.cnf) one per tile per cycle

Act as Input to GERALD

Allow offline base calling in Bustard
Real-Time Analysis (RTA) Overview

- RTA also saves information to run folder* during run
  - Assists with post-run troubleshooting
  - Allows quick visualization of flow cell through all cycles

* Note that run folder is set up differently in RTA than in previous versions of the pipeline

<run folder>/Data/reports
Real-Time Analysis (RTA) Software

- RTA navigation bar remains at the top of the display.
- “Run Info” is default screen
Real-Time Analysis (RTA) Software

“Extracted/Called/Scored” indicates which cycle is undergoing each analysis step, respectively.

- **85/83/82:**
  - Cycle 85 is undergoing Extraction (Determining Cluster Intensities)
  - Cycle 83 is undergoing Base Calling
  - Cycle 82 is undergoing Quality Scoring
“Tile Status” displays the analysis status of each tile on the flow cell.

- Color corresponds to extraction status
- Numbers in tile boxes indicate which cycles are currently being called and scored, respectively.
Real-Time Analysis (RTA) Software

“Cluster Density” shows the number of clusters per mm²

Note: Clusters/mm² is ~2X Clusters/tile

<run folder>/Data/reports: “NumClusters By Lane.png”
Real-Time Analysis (RTA) Software

“Intensity & Focus Quality” plots show intensity or focus quality by cycle, for any base.

- Two plots allow comparison.
- Intensity plots can be used to monitor intensity decay.

<run folder>/Data/reports: 
“Intensity_By_Cycle_a.png”
“FWHM_By_Cycle_a.png”
Real-Time Analysis (RTA) Software

“Charts” allows comparative visualization of flow cell on 3 adjacent maps.

- Chart 1:
  - Cluster Density
  - % PF Clusters

- Charts 2,3 (by cycle & base):
  - Intensity
  - Q30 scores
  - Focus quality

<run folder>/Data/reports:
  “NumClusters_Chart.png”
  “PassedFilter25_Chart.png”
  “Intensity_Chart_1_a.png”
  “NumGT30_Chart_1.png”
  “FWHM_Chart_1_a.png”
Real-Time Analysis (RTA) Reports

Plots and charts are saved: <run folder>/Data/reports
- FWHM by Color and Cycle.txt (1 file per flow cell)
- FWHM_By_Cycle_a.png (1 plot per base)
- FWHM_Chart_1_a.png (1 chart per cycle per base)
- Intensity by Color and Cycle.txt (1 file per flow cell)
- Intensity_By_Cycle_a.png (1 plot per base)
- Intensity_Chart_1_a.png (1 chart per cycle per base)
- NumClusters by lane.txt (1 file per flow cell)
- NumClusters By Lane.png (1 plot per flow cell)
- NumClusters_Chart.png (1 chart per flow cell)
- NumGT30_Chart_1.png (1 chart per cycle per base)
- PassedFilter25_Chart.png (1 chart per flow cell)
Real-Time Analysis (RTA) Reports

- Plots and charts are saved: <run folder>/Data/reports
  - FWHM by Color and Cycle.txt (1 file per flow cell)
  - FWHM_By_Cycle_a.png (1 plot per base)
  - FWHM_Chart_1_a.png (1 chart per cycle per base)
Real-Time Analysis (RTA) Reports

- Plots and charts are saved: <run folder>/Data/reports

- Intensity by Color and Cycle.txt  (1 file per flow cell)
- Intensity_By_Cycle_a.png  (1 plot per base)
- Intensity_Chart_1_a.png  (1 chart per cycle per base)
Real-Time Analysis (RTA) Reports

- Plots and charts are saved: <run folder>/Data/reports
  - NumClusters by lane.txt  (1 file per flow cell)
  - NumClusters By Lane.png  (1 plot per flow cell)
  - NumClusters_Chart.png  (1 chart per flow cell)
Real-Time Analysis (RTA) Reports

- Plots and charts are saved: `<run folder>/Data/reports`
  - `NumGT30_Chart_1.png` *(1 chart per cycle per base)*
  - `PassedFilter25_Chart.png` *(1 chart per flow cell)*
Real-Time Analysis (RTA) Reports

- If GERALD Summary files show Tile-to-Tile variability
- RTA charts can help diagnose problems:

Tip: Scrolling through saved charts allows view over time
Looking at Run Data after Run Completion

- We have begun to use a fairly standard method of troubleshooting, which is also useful for generally understanding a run.

![Troubleshooting a Run Diagram](image)
Using RTA 1.5

Looking at Run Data -- Start at the top and dig deeper as necessary:

- **GERALD Summary Files**
- Intensity Plots
- RTA Charts
- Images
- Internal Controls
- At base-calling: Crosstalk Matrix
- At base-calling: Phasing, Prephasing
- Tracked information from throughout the process
GERALD Summary Files

- Stored within the Run Folder
- Contains most of the critical metrics
  - Overall run statistics
  - Lane-by-lane summary statistics
- More info between the lines
GERALD Summary Files

- Chip Summary
- Chip Results Summary
- Lane Parameter Summary
- Lane Results Summary (by read)
- Expanded Lane Summary (by read)
- Per-Tile Statistics (by lane by read)
- Plots and Graphs
- Pair Summary Information (by lane)
  - Relative Orientation Statistics
  - Insert Size Statistics
  - Insert Statistics
GERALD Summary Files

- **Lane Results Summary (by read)**
  - Does intensity start high enough and decay slowly enough to avoid noise problems later?
    - 1st cycle intensity PF > 100
    - % Intensity PF after 20 cycles > 50
  - Is % Error Rate (PF) high?
    - % called bases in PF aligned reads that don’t match the reference.

**Note:** This intensity is NOT the same as that reported in RTA

---

**Lane Results Summary : Read 1**

<table>
<thead>
<tr>
<th>Lane</th>
<th>Lane Yield (kbases)</th>
<th>Clusters (raw)</th>
<th>Clusters (PF)</th>
<th>1st Cycle Int (PF)</th>
<th>% intensity after 20 cycles (PF)</th>
<th>% PF Clusters</th>
<th>% Align (PF)</th>
<th>Alignment Score (PF)</th>
<th>% Error Rate (PF)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1120134</td>
<td>170659 +/- 9170</td>
<td>110904 +/- 17931</td>
<td>194 +/- 48</td>
<td>173.32 +/- 469.73</td>
<td>64.66 +/- 9.64</td>
<td>71.64 +/- 6.17</td>
<td>43.79 +/- 6.46</td>
<td>1.40 +/- 0.46</td>
</tr>
<tr>
<td>2</td>
<td>1203974</td>
<td>177225 +/- 6908</td>
<td>119205 +/- 4878</td>
<td>217 +/- 13</td>
<td>82.63 +/- 2.46</td>
<td>67.28 +/- 1.58</td>
<td>62.81 +/- 4.70</td>
<td>35.68 +/- 4.48</td>
<td>1.30 +/- 0.11</td>
</tr>
</tbody>
</table>
GERALD Summary Files

- Lane Results Summary (by read)
- Tile Mean +/- SD for each lane
- High Standard Deviation? Check per-tile stats.

Lane Results Summary : Read 1

<table>
<thead>
<tr>
<th>Lane</th>
<th>Lane Yield (pbases)</th>
<th>Tile Mean +/- SD for Lane</th>
<th>Clusters (raw)</th>
<th>Clusters (PF)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1120134</td>
<td>170659 +/- 9170</td>
<td>110904 +/- 17931</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>1203974</td>
<td>177225 +/- 6908</td>
<td>119205 +/- 4878</td>
<td></td>
</tr>
</tbody>
</table>

- Clusters (PF)
- 1st Cycle Int (PF)
- % intensity after 20 cycles (PF)

110904 +/- 17931
194 +/- 48
173.32 +/- 469.73

119205 +/- 4878
217 +/- 13
82.63 +/- 2.46
GERALD Summary Files

► Expanded Lane Summary (by read)
► Some of the same metrics as Lane Results
► Applied Phasing & Prephasing
  – Average across all lanes
  – Ideally: Phasing < 0.4%
  – Ideally: Prephasing < 0.3%
  – If values are higher, it’s worth finding lane-specific values (in Run Folder)

NOTE: Illumina says <1 but we want <2% error

---

**Expanded Lane Summary : Read 1**

<table>
<thead>
<tr>
<th>Lane</th>
<th>Clusters (tile mean) (raw)</th>
<th>% Phasing</th>
<th>% Prephasing</th>
<th>% Error Rate (raw)</th>
<th>% retained</th>
<th>Cycle 2-4 Av Int (PF)</th>
<th>Cycle 2-10 Av % Loss (PF)</th>
<th>Cycle 10-20 Av % Loss (PF)</th>
<th>% Align (PF)</th>
<th>% Error Rate (PF)</th>
<th>Equiv Perfect Clusters (PF)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>170660</td>
<td>0.3300</td>
<td>0.2400</td>
<td>1.71</td>
<td>64.66</td>
<td>204.02 +/- 16.28</td>
<td>-0.18 +/- 4.59</td>
<td>0.24 +/- 2.13</td>
<td>71.64</td>
<td>1.40</td>
<td>75715</td>
</tr>
<tr>
<td>2</td>
<td>177226</td>
<td>0.3300</td>
<td>0.2400</td>
<td>1.58</td>
<td>67.28</td>
<td>211.13 +/- 11.57</td>
<td>0.76 +/- 0.23</td>
<td>0.84 +/- 0.14</td>
<td>62.81</td>
<td>1.30</td>
<td>70818</td>
</tr>
</tbody>
</table>
GERALD Summary Files

» Expanded Lane Summary (by read)

» Intensity Decay Rates
  - (Cycle 2-10, 10-20 Av % Loss (PF))
  - Average across all tiles
  - Look for possible problems
    - Variation between lanes
    - High Standard Deviation
    - Fast decay, signaling probable noise in later cycles

Expanded Lane Summary : Read 1

<table>
<thead>
<tr>
<th>Lane Info</th>
<th>Phasing Info</th>
<th>Raw Data (tile mean)</th>
<th>Filtered Data (tile mean)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Clusters (tile mean) (raw)</td>
<td>% Phasing</td>
<td>% Error Rate (raw)</td>
<td>Equiv Perfect Clusters (raw)</td>
</tr>
<tr>
<td>Lane</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>170660</td>
<td>0.3300</td>
<td>0.2400</td>
</tr>
<tr>
<td>2</td>
<td>177226</td>
<td>0.3300</td>
<td>0.2400</td>
</tr>
</tbody>
</table>
Intensity: GERALD Summary File vs. RTA

**GERALD Summary File:**
~88% PF

~104 1st cycle int (PF)
~98 cyc 2-4 avg Int (PF)
~0.9 % Loss cyc 2-10 (PF)
~0.5% Loss cyc 10-20 (PF)
~82% int after 20 cyc (PF)

**RTA**
~800-1200

---

<table>
<thead>
<tr>
<th>First Base Report</th>
<th>RTA</th>
<th>GERALD Summary File (PF)</th>
<th>IVC Plots: All</th>
<th>IVC Plots: Called</th>
</tr>
</thead>
<tbody>
<tr>
<td>1st cycle Intensity</td>
<td>~1150</td>
<td>~750</td>
<td>~100</td>
<td>~100</td>
</tr>
</tbody>
</table>
## GERALD Summary Files

- **Per-Tile Statistics (by lane by read)**
  - Critical when Lane Results Summary shows high Standard Deviations

### Lane Results Summary: Read 1

<table>
<thead>
<tr>
<th>Lane Info</th>
<th>Tile Mean +/- SD for Lane</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lane</td>
<td>Lane Yield (kbases)</td>
</tr>
<tr>
<td>-----------</td>
<td>---------------------</td>
</tr>
<tr>
<td>1</td>
<td>1120134</td>
</tr>
<tr>
<td>2</td>
<td>1203974</td>
</tr>
</tbody>
</table>
GERALD Summary Files

**Per-Tile Statistics (by lane by read)**

**Lane 1 : Read 1**

<table>
<thead>
<tr>
<th>Lane</th>
<th>Tile</th>
<th>Clusters (raw)</th>
<th>Av 1st Cycle Int (PP)</th>
<th>Av % intensity after 20 cycles (PF)</th>
<th>% PF Clusters</th>
<th>% Align (PP)</th>
<th>Av Alignment Score (PF)</th>
<th>% Error Rate (PF)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0001</td>
<td>184152</td>
<td>176.45</td>
<td>88.11</td>
<td>67.04</td>
<td>76.48</td>
<td>49.68</td>
<td>1.21</td>
</tr>
<tr>
<td>1</td>
<td>0002</td>
<td>186088</td>
<td>188.43</td>
<td>86.40</td>
<td>67.42</td>
<td>76.06</td>
<td>49.18</td>
<td>1.21</td>
</tr>
<tr>
<td>1</td>
<td>0003</td>
<td>183807</td>
<td>210.55</td>
<td>84.14</td>
<td>67.75</td>
<td>75.43</td>
<td>48.13</td>
<td>1.24</td>
</tr>
<tr>
<td>1</td>
<td>0004</td>
<td>183974</td>
<td>211.10</td>
<td>82.60</td>
<td>67.35</td>
<td>75.59</td>
<td>48.58</td>
<td>1.20</td>
</tr>
<tr>
<td>1</td>
<td>0005</td>
<td>182335</td>
<td>208.62</td>
<td>88.21</td>
<td>66.39</td>
<td>75.77</td>
<td>48.49</td>
<td>1.25</td>
</tr>
<tr>
<td>1</td>
<td>0006</td>
<td>177046</td>
<td>190.20</td>
<td>92.25</td>
<td>66.56</td>
<td>75.36</td>
<td>48.67</td>
<td>1.17</td>
</tr>
<tr>
<td>1</td>
<td>0007</td>
<td>177570</td>
<td>202.57</td>
<td>87.97</td>
<td>66.99</td>
<td>75.46</td>
<td>48.59</td>
<td>1.24</td>
</tr>
<tr>
<td>1</td>
<td>0008</td>
<td>171897</td>
<td>190.69</td>
<td>84.74</td>
<td>66.83</td>
<td>75.49</td>
<td>48.32</td>
<td>1.26</td>
</tr>
<tr>
<td>1</td>
<td>0009</td>
<td>166125</td>
<td>173.72</td>
<td>105.02</td>
<td>63.83</td>
<td>69.12</td>
<td>40.32</td>
<td>1.49</td>
</tr>
<tr>
<td>1</td>
<td>0010</td>
<td>132422</td>
<td>3.92</td>
<td>3136.71</td>
<td>10.78</td>
<td>67.52</td>
<td>30.72</td>
<td>2.13</td>
</tr>
<tr>
<td>1</td>
<td>0011</td>
<td>138515</td>
<td>5.38</td>
<td>3566.51</td>
<td>5.52</td>
<td>24.00</td>
<td>3.85</td>
<td>4.81</td>
</tr>
<tr>
<td>1</td>
<td>0012</td>
<td>171385</td>
<td>226.62</td>
<td>81.01</td>
<td>65.13</td>
<td>67.84</td>
<td>39.97</td>
<td>1.37</td>
</tr>
<tr>
<td>1</td>
<td>0013</td>
<td>170539</td>
<td>224.70</td>
<td>81.73</td>
<td>64.85</td>
<td>68.89</td>
<td>41.09</td>
<td>1.32</td>
</tr>
<tr>
<td>1</td>
<td>0014</td>
<td>182131</td>
<td>227.00</td>
<td>85.18</td>
<td>68.04</td>
<td>73.63</td>
<td>46.38</td>
<td>1.15</td>
</tr>
<tr>
<td>1</td>
<td>0015</td>
<td>183746</td>
<td>205.22</td>
<td>86.88</td>
<td>68.52</td>
<td>73.88</td>
<td>46.92</td>
<td>1.17</td>
</tr>
<tr>
<td>1</td>
<td>0100</td>
<td>183836</td>
<td>203.20</td>
<td>93.60</td>
<td>68.35</td>
<td>74.33</td>
<td>47.04</td>
<td>1.24</td>
</tr>
</tbody>
</table>
GERALD Summary Files

Tip: Copy from Summary File and paste into Excel to plot values.
GERALD Summary Files

- **Per-Tile Statistics (by lane by read)**
  - Gradient along the lane?
    - Peltier-related?
    - Density-related?
  - Values differ by column?
    - Footprint-related?
    - Edge effect?
  - Outliers?
    - Often occurs at the end of a lane
  - Scattering of tiles reporting “0”? 
    - Often signifies oil problem
  - Values reported in brackets?
    - No PF clusters (often signified bubble)

- Chip Summary
- Chip Results Summary
- Lane Parameter Summary
- **Lane Results Summary (by read)**
- Expanded Lane Summary (by read)
- **Per-Tile Statistics (by lane by read)**
- Plots and Graphs
- Pair Summary Information (by lane)
  - Relative Orientation Statistics
  - Insert Size Statistics
  - Insert Statistics
IntENSITY Plots

► Using RTA 1.5

► Looking at Run Data -- Start at the top and dig deeper as necessary:
  – GERALD Summary Files
  – Intensity Plots
  – RTA Charts
  – Images
  – Internal Controls
  – At base-calling: Crosstalk Matrix
  – At base-calling: Phasing, Prephasing
  – Tracked information from throughout the process
## Intensity Plots

<table>
<thead>
<tr>
<th>ALL</th>
<th>CALLED</th>
<th>% ALL</th>
<th>% CALLED</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total measured intensity in each channel.</td>
<td>After base-calling corrections, intensity of clusters determined to be in each channel.</td>
<td>Percent of clusters seen in each channel.</td>
<td>Percent of total intensities</td>
</tr>
<tr>
<td>All channels similar. Smooth decay.</td>
<td>All channels similar. Roughly 4X All. Smooth decay.</td>
<td>Shows changes in relative intensities between channels. No decay.</td>
<td>25% each in non-base-biased samples. Changes over time signify decreasing purity (phasing, etc)</td>
</tr>
</tbody>
</table>

Plots: Data/Intensities/Bustard/Plots
Data/Intensities/Basecalls/Plots
Intensity Plots

- Signal means file can also be plotted
- What to look for:
  - High starting intensity
  - Gradual, slow decay
  - Smooth lines
  - Good Read2 Regeneration

Tip: Scrolling through tile-to-tile plots allows view across a lane

Oil Problem:
Intensity Plots: % Basecalls

- Usually, base calls should be ~25% for each base at each cycle
- If single base shows higher proportion than other 3

![](image)
Run Images

- Using RTA 1.5

- Looking at Run Data -- Start at the top and dig deeper as necessary:
  - GERALD Summary Files
  - Intensity Plots
  - RTA Charts
  - Images
  - Internal Controls
  - At base-calling: Crosstalk Matrix
  - At base-calling: Phasing, Prephasing
  - Tracked information from throughout the process
Run Images

- GERALD Summary file showed Tile-to-Tile variability
- Charts from RTA v1.5 showed possible oil spread
- Images confirmed oil
Run Images

- RTA can be set to retain a subset of images
- Future versions of RTA will convert raw images to JPG
- 3\textsuperscript{rd} party software can be used to convert images to JPG using current versions of RTA
- You must determine which images you want to retain
Internal Controls (ICs)

- Using RTA 1.5

- Looking at Run Data -- Start at the top and dig deeper as necessary:
  - GERALD Summary Files
  - Intensity Plots
  - RTA Charts
  - Images
  - Internal Controls
  - At base-calling: Crosstalk Matrix
  - At base-calling: Phasing, Prephasing
  - Tracked information from throughout the process
Internal Controls (ICs)

- We currently use two types of control samples, each with benefits

<table>
<thead>
<tr>
<th>Phi X Lane</th>
<th>Broad Internal Controls (ICs)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Create a good crosstalk matrix for use in base-biased samples</td>
<td>Regain “real estate” by adding a small amount to each experimental lane (1/100 reads instead of 1/8 reads)</td>
</tr>
<tr>
<td>Well characterized small genome</td>
<td>Synthetic sequences require no error-inducing PCR enrichment</td>
</tr>
<tr>
<td>~ 50% GC</td>
<td>~ 50% GC</td>
</tr>
<tr>
<td>Simple data analysis</td>
<td>Easily identified monotemplates*</td>
</tr>
<tr>
<td>Simple workflow (treat as library)</td>
<td>Control in every lane</td>
</tr>
</tbody>
</table>

*Using Broad-designed pipeline software*
Internal Controls (ICs)

- The Internal Controls pool contains 4 synthetic monotemplate sequences, which don’t align to known genomes.
- Pool of 4 sequences added to all lanes (~1/100 clusters).
  - Pool contains equal base composition.
  - ICs with adaptors are digested/purified from plasmid.
  - IC fragment with final adapter sequence are flanked by unique restriction sites in plasmid.
  - ICs with adaptors are digested/purified from plasmid.
  - After final purification, all 4 ICs are quantified (KAPA qPCR) and pooled equimolarly to give even representation of each base at each cycle.

- Cellular reproduction of plasmid (more accurate than PCR amplification).
- Distinguish sequencing quality from library quality (compare lanes).
- Assess lane quality regardless of reference sequence.
Internal Controls (ICs)

- Plotting error rates by cycle for ICs assists in overall run assessment
- Identification of run problems:
  - This run went bad after cycle 20 (intensity)
  - Bases called as “N” show up as errors in ICs
Analysis of Internal Controls (ICs)

- Use whatever data files you typically use:
  - qseq files
  - BAM files (this is easier for us, since we already have BAMs)

- Find IC clusters
  - For each read: Perform a base-by-base comparison to each of the ICs
  - If the read matches “well enough” (where “well enough” is user-defined), then it’s an IC
    - Base 1 of read is always Base 1 of IC
    - Define minimum match length (~20b)
    - Define allowable error rate (<~10%)

- Create metrics
  - Counts, proportions in pool, error rates by cycle (by IC &/or by base)
Internal Controls (ICs)

**Pairing Identification**: If a cluster is identified as an Internal Control in EITHER read, then the same cluster is included in the IC analysis in BOTH reads (even if no cluster is actually found there)

- Known cluster position (is it present in both reads?)
- Expected sequence vs Determined sequence
- More thorough calculation of error rates
  - Often slightly higher than genomic sample error rates
  - Poor cluster regeneration leads to high Read 2 IC error rates
Questions?

RUN MONITORING
Using RTA 1.5

BASIC TROUBLESHOOTING
GERALD Summary Files
Intensity Plots
RTA Charts
Images
Internal Controls

Next: What if the problem is still unclear?
Digging Deeper

- Using RTA 1.5

- Looking at Run Data -- Start at the top and dig deeper as necessary:
  - GERALD Summary Files
  - Intensity Plots
  - RTA Charts
  - Images
  - Internal Controls
  - At base-calling: Crosstalk Matrix
  - At base-calling: Phasing, Prephasing
  - Tracked information from throughout the process
Crosstalk Matrix (Color Matrix)

Applying matrix reduces effect of a cluster showing up in both X & Y channels.

Clusters in the “G” Channel are also present in the “T” Channel but not all clusters in the “T” Channel are bright in the “G” Channel.

# Auto-generated frequency response matrix

\[
\begin{array}{cccc}
> A & > C & > G & > T \\
1.09 & 0.17 & 0.07 & 0.08 \\
0.69 & 0.86 & 0.04 & 0.05 \\
0.05 & 0.04 & 1.17 & 0.02 \\
0.06 & 0.04 & 0.61 & 1.02 \\
\end{array}
\]
Crosstalk Matrix (Color Matrix)

- Crosstalk between channels causes some clusters to show up in two images, but their intensity is different in the two images.
  - Overlap of X & Y signal frequencies

- Normalization
  - Corrects for maximum intensity differences between X & Y
  - Reduces effect of a cluster showing up in both X & Y channels

![Diagram showing crosstalk and normalization](image-url)
Crosstalk Matrix (Color Matrix)

- Base-calling matrix is CALCULATED for each lane
- The APPLIED matrix is specified in analysis parameters
  - Average of all lanes’ calculated matrices
  - Control lane calculated matrix
  - Instrument default matrix

- Identifying APPLIED matrix:
  Runfolder/Data/Intensities/Bustard > Bustard.log

- Identifying CALCULATED matrices:
  Runfolder/Data/Intensities/Bustard/Matrix > s_[1]_[78]_matrix.txt
Crosstalk Matrix (Color Matrix)

- Identification of Matrix Problems using ICs
  - High sum error rate
  - Each IC’s error jumps by cycle
  - Each IC’s error range is huge
Crosstalk Matrix (Color Matrix)

- After application of new matrix, ICs show lower error rate
  - Total: Down from 10% to 1%
  - Individual: Smoother by cycle
Digging Deeper

- Using RTA 1.5

- Looking at Run Data -- Start at the top and dig deeper as necessary:
  - GERALD Summary Files
  - Intensity Plots
  - RTA Charts
  - Images
  - Internal Controls
  - At base-calling: Crosstalk Matrix
  - At base-calling: Phasing, Prephasing
  - Tracked information from throughout the process
Phasing & Prephasing

- **Values APPLIED to all lanes (as shown in Summary File):**
  - Data/Intensities/Bustard/Phasing > phasing.xml

- **CALCULATED for each lane:**
  - Data/Intensities/Bustard/Phasing > s_01_phasing.xml

- **CALCULATED for each tile within all lanes:**
  - Data/Intensities/Bustard/Phasing > s_01_phasing.txt

<table>
<thead>
<tr>
<th>Lane</th>
<th>Tile</th>
<th>Phasing</th>
<th>Prephasing</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>0.0048</td>
<td>0.0035</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.0032</td>
<td>0.0054</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.0058</td>
<td>0.0046</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.0028</td>
<td>0.0026</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.0032</td>
<td>0.0052</td>
</tr>
<tr>
<td></td>
<td></td>
<td>-0.0012</td>
<td>-0.0048</td>
</tr>
<tr>
<td>1</td>
<td>2</td>
<td>0.0049</td>
<td>0.0035</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.0034</td>
<td>0.0052</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.0058</td>
<td>0.0052</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.0028</td>
<td>0.0024</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.0032</td>
<td>0.0055</td>
</tr>
<tr>
<td></td>
<td></td>
<td>-0.0024</td>
<td>-0.0054</td>
</tr>
<tr>
<td>1</td>
<td>3</td>
<td>0.0046</td>
<td>0.0033</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.0031</td>
<td>0.0048</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.0056</td>
<td>0.0049</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.0028</td>
<td>0.0021</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.0031</td>
<td>0.0052</td>
</tr>
<tr>
<td></td>
<td></td>
<td>-0.0003</td>
<td>-0.0039</td>
</tr>
</tbody>
</table>
Phasing & Prephasing

- Applied values are reported (average of all lanes)
- Applying an incorrect value can reduce call quality
- The s_xx_phasing.txt values can be used to plot phasing by tile for all lanes of a flow cell
- Can help diagnose problems with peltier block or reagent flow
Tracked information from throughout the process

- Using RTA 1.5

- Looking at Run Data -- Start at the top and dig deeper as necessary:
  - GERALD Summary Files
  - Intensity Plots
  - RTA Charts
  - Images
  - Internal Controls
  - At base-calling: Crosstalk Matrix
  - At base-calling: Phasing, Prephasing
  - Tracked information from throughout the process
Tracked information from throughout the process

Sample and Process details can hold critical information:
- Sample Source information
- Sample Preparation Details
- Cluster Generation Conditions
- QC Data (Shearing, Post-Enrichment, Gel images, SYBR QC)
- Instrument and Machine Identifications
- Reagent lots
- Event timing and location (cluster station, automation deck, etc.)
- Operator notes ("oil cleaned from flow cell after cycle 24")

These details are especially powerful when linked informatically
Identifying Problems in Resequencing Runs

We currently have an automated “red flag” JIRA system, which uses extremely conservative cut-offs to identify probable problems.

A run is manually investigated if any of the following flags occurs:

- Total reads in a lane < 2M
- %PF < 50%
- %PF aligned < 30%
- # Bad Cycles > 1 (cycles in which >80% calls were “N”)
- Mean Internal Control Error Rates > 5% for either read
- Concordance < 95% (based on known variations in tumor and/or normal tissue)

Additionally, collaborators may ask us to look into run details, if they find their run’s metrics to be questionable for any reason.
Next, we’ll lead you through some troubleshooting case studies

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>1a. Check Run &amp; Lane Summary Metrics</td>
<td>GERALD Summary File</td>
</tr>
<tr>
<td>1b. Check Tile Summary Metrics</td>
<td>GERALD Summary File</td>
</tr>
<tr>
<td>2. Check Intensity Plots</td>
<td>Intensity Versus Cycle (call.png, percent_base.png), Broad Generated plots</td>
</tr>
<tr>
<td>3. RTA Reports</td>
<td>(RTA 1.5) FWHM by Color and Cycle, Intensity by Color and Cycle, NumClusters (RTA 1.6) FWHM Chart, Intensity Charts, NumClusters, NumGT30</td>
</tr>
<tr>
<td>4. Check Images</td>
<td>SYBR, Selected saved tile jpegs</td>
</tr>
<tr>
<td>5. Check Internal Controls</td>
<td>Automated Error By Cycle plots, Summary metrics, By-Cycle Metrics</td>
</tr>
<tr>
<td>6. Check Basecalling Correction Parameters</td>
<td>Applied and Calculated: Matrix by Lane, Phasing by Lane, Prephasing by Lane</td>
</tr>
<tr>
<td>7. Dig Deeper Based on Findings in previous Troubleshooting Steps</td>
<td>Based on where previous steps looked problematic, other files and metrics can be assessed. GERALD Error plots, GERALD Rescore plots, Image Offsets, Sample identity information, lane waste volumes, hardware check, etc.</td>
</tr>
<tr>
<td>8. Provide Feedback</td>
<td>Reanalysis or Total Failure (actions to be determined by projects team)</td>
</tr>
</tbody>
</table>

Note: Broad Logo signifies some level of Broad-designed analysis (Intensity plots are recreated on-site, and Internal Controls are Broad-designed)