NGS Data Analysis Using Partek® Software

Xiaowen Wang
Field Application Specialist
Who is Partek?

• Founded in 1993
• Building tools for statistics & visualization
• Focused on genomics
• Thousands of customers worldwide
• Worldwide, world-class customer support
Comprehensive Solution for NGS Analysis

Start to Finish Analysis for NGS and Microarray Experiments
Partek® Flow®

- Web based application
- Intuitive interface
- Schedule tasks
- Share resources/data/pipelines
Partek® Genomics Suite™

- NGS
- Microarray
- Workflows
- Integration
RNA-Seq Analysis
RNA-seq Objectives

- **Quantification**
  - How much expression each gene has
- **Differential expression detection**
  - Which genes are differently expressed between groups
- **Alternative splicing detection**
  - Treatment group expressed differently among splicing variants
- **Allele specific expression**
  - Nucleotide sequence variation affect gene expression
- **Novel transcripts detection**
- **Fusion gene detection**
- **Biological Interpretation: GO, Pathway**
RNA-seq Data Analysis

Flow:
- Sequencer
  - fastq
- Alignment
- Pre/Post QA/QC
- Trim/Filter
- Quantification
- DGE

Genomics Suite:
- Import
- QA/QC
- Quantification
- Exploratory Analysis
- Statistical Analysis
- Visualization
- Biological Interpretation

(QA/QC)

.zip

.bam
Pipeline

- Data smart interface
- Save/reuse pipelines
- Different layers display
- Details recorded
### Resource Management

**Home > Workers dashboard**

#### Queue stats
- **Running tasks**: 5
- **Pending tasks**: 12
- **All tasks end time**: 23 Oct 2013, 11:42 AM CDT

#### Licensing
- **Available cores licenses**: 274 out of 400
- **Available worker licenses**: 87 out of 100

#### Active workers

<table>
<thead>
<tr>
<th>Name</th>
<th>Worker CPU</th>
<th>Worker Memory</th>
<th>Server CPU</th>
<th>Server memory</th>
<th>Uptime</th>
<th>Time to completion</th>
<th>Type</th>
<th>Stop</th>
</tr>
</thead>
<tbody>
<tr>
<td>eucalyptus.partek.com</td>
<td>33%</td>
<td>31%</td>
<td>27%</td>
<td>57%</td>
<td>8 days</td>
<td>33:20:00</td>
<td>internal</td>
<td></td>
</tr>
<tr>
<td>compute-1.partek.com</td>
<td>24%</td>
<td>17%</td>
<td>24%</td>
<td>17%</td>
<td>02:00:06</td>
<td>82:53:55</td>
<td>unmanaged</td>
<td></td>
</tr>
<tr>
<td>compute-2.partek.com</td>
<td>95%</td>
<td>88%</td>
<td>95%</td>
<td>83%</td>
<td>01:59:23</td>
<td>24:50:37</td>
<td>unmanaged</td>
<td></td>
</tr>
<tr>
<td>compute-3.partek.com</td>
<td>28%</td>
<td>36%</td>
<td>28%</td>
<td>36%</td>
<td>02:00:15</td>
<td>60:33:29</td>
<td>unmanaged</td>
<td></td>
</tr>
<tr>
<td>compute-4.partek.com</td>
<td>3%</td>
<td>25%</td>
<td>3%</td>
<td>25%</td>
<td>00:01:13</td>
<td>23:50:37</td>
<td>unmanaged</td>
<td></td>
</tr>
</tbody>
</table>
FASTQ File Formats

Line 1) begins with a '@' character and is followed by a sequence identifier and an optional description (like a FASTA title line).

Line 2) is the raw sequence letters (ACGT).

Line 3) begins with a '+' character and is optionally followed by the same sequence identifier (and any description) again.

Line 4) encodes the quality values for the sequence in Line 2, and must contain the same number of symbols as letters in the sequence.

Sanger format can encode a Phred quality score from 0 to 93 using ASCII 33 to 126.
Each base is assigned a quality score that indicates the probability that the base was accurately identified.

Range is 0-93 (Pred 33) or 0-62 (Pred 64).

Phred quality scores are logarithmically linked to error probabilities.

<table>
<thead>
<tr>
<th>Phred Quality Score</th>
<th>Probability of incorrect base call</th>
<th>Base call accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>10</td>
<td>1 in 10</td>
<td>90%</td>
</tr>
<tr>
<td>20</td>
<td>1 in 100</td>
<td>99%</td>
</tr>
<tr>
<td>30</td>
<td>1 in 1000</td>
<td>99.9%</td>
</tr>
<tr>
<td>40</td>
<td>1 in 10000</td>
<td>99.99%</td>
</tr>
<tr>
<td>50</td>
<td>1 in 100000</td>
<td>99.999%</td>
</tr>
</tbody>
</table>

Phred Quality Score = -10*\log_{10}(\text{Probability of Error})
Sample Manager

- Sample files can be on the server or local
- File type accepted: fastq, fna+qual, sff, sra, bam etc
- Admin/project owner manages vocabulary

Controlled vocabulary
Pre-alignment QA/QC

- Pre-alignment QA/QC to help set alignment parameters
- Output: Table and graphs (per project or per sample)
QA/QC on ERCC Spike-In Controls

Total number of alignments versus actual concentration
Trim Bases and Adapter

Trim: fastq

- Trim based on 3’ or 5’ end
- Trim both end
- Trim based on quality score
- Trim adapter--cutadapt
What is Alignment?

- Read comes off a short read sequencing machine
  
  \[
  \begin{array}{ccccccc}
  A & T & G & G & T & C & A \\
  \end{array}
  \]

- Goal: Determine where on the genome that read belongs

- Method: Match sequence of read to sequence from a reference genome
  
  \[
  \begin{array}{cccccccc}
  G & G & C & A & T & G & G & T & C & A & T & T & C \\
  \end{array}
  \]

  (reference genome)

  \[
  \begin{array}{ccccccc}
  A & T & G & G & T & C & A \\
  \end{array}
  \]

  (read)

  **Result:** Genomic Location of read
Choosing a Reference

• Align to a genomic reference
  • hg19, mm10, rn4, etc.
    • Standard reference sequence for model organisms
    • Identification of novel transcripts

• Align to transcriptome
  • RefSeq, ENSEMBL, Aceview, etc.
    • There is no standard reference sequence for the organism of interest, but the transcriptome sequences are available.
    • Only interested in known mRNAs
    • Speed of alignment
SAM/BAM Format

- Sequence Alignment/MAP (SAM) is TAB-delimited
- Header section (optional)
  - Begins with @
- Alignment section

```
@HD VN:1.3 SQ:coordinate
@SQ SN:ref LN:45
r001 163 ref 7 30 8M2I4M1D3M = 37 39 TTAGATAAAGGATACCTG *
r002 0 ref 9 30 3S6M1P1I4M * 0 0 AAAAGATAAGGATA *
r003 0 ref 9 30 5H6M * 0 0 AGCTAA * NM:i:1
r004 0 ref 16 30 6M14N5M * 0 0 ATAGCTTCAGC *
r003 16 ref 29 30 6H5M * 0 0 TACGC * NM:i:0
r001 83 ref 37 30 9M = 7 -39 CAGCGCCAT *
```

BAM = binary form of SAM format
Alignment -- Methods

Alignment input: fastq, fasta, sra, sff etc.

Alignment output:
- aligned data node – downstream analysis
- unaligned data node – allow multiple stages alignment

Aligners:
- Bowtie
- Bowtie2
- Tophat
- Tophat2
- BWA
- TMAP
- SHRiMP 2
- GSNAP
- STAR
- More coming
Combine Alignments

- Single alignment doesn’t always give the perfect solution
- Multiple stages alignment can increase the alignment rate
- **Combine Alignments** is a very useful tool in this case.
Post-alignment QA/QC in Flow

Input -- .bam files
Check alignment quality and genomic coverage

<table>
<thead>
<tr>
<th>Sample name</th>
<th>Total reads</th>
<th>Aligned</th>
<th>Unique</th>
<th>Coverage</th>
<th>Avg. coverage depth</th>
<th>Avg. length</th>
<th>Avg. quality</th>
<th>Avg. mapping quality</th>
<th>%GC</th>
</tr>
</thead>
<tbody>
<tr>
<td>SRR015306_chr22</td>
<td>45545</td>
<td>76.10%</td>
<td>76.10%</td>
<td>0.60%</td>
<td>3.61 (SD 29.69)</td>
<td>32</td>
<td>23.34</td>
<td>256 (SD 0)</td>
<td>54.07%</td>
</tr>
<tr>
<td>SRR015305_chr22</td>
<td>49890</td>
<td>78.08%</td>
<td>78.08%</td>
<td>0.64%</td>
<td>3.80 (SD 32.47)</td>
<td>32</td>
<td>23.44</td>
<td>256 (SD 0)</td>
<td>54.26%</td>
</tr>
<tr>
<td>SRR015257_chr22</td>
<td>44125</td>
<td>75.12%</td>
<td>75.12%</td>
<td>0.88%</td>
<td>2.36 (SD 4.67)</td>
<td>32</td>
<td>22.60</td>
<td>256 (SD 0)</td>
<td>52.67%</td>
</tr>
<tr>
<td>SRR015302_chr22</td>
<td>46172</td>
<td>76.56%</td>
<td>76.56%</td>
<td>0.60%</td>
<td>3.67 (SD 30.69)</td>
<td>32</td>
<td>23.57</td>
<td>256 (SD 0)</td>
<td>53.83%</td>
</tr>
<tr>
<td>SRR015261_chr22</td>
<td>43714</td>
<td>75.12%</td>
<td>75.12%</td>
<td>0.87%</td>
<td>2.36 (SD 4.82)</td>
<td>32</td>
<td>22.78</td>
<td>256 (SD 0)</td>
<td>51.88%</td>
</tr>
</tbody>
</table>
• On average, each base has been sequenced a certain number of times on the genome

• Number of reads * read length / genome size
Mapping reads to transcriptome – Measure gene/transcript abundances

Algorithm:
- E/M algorithm (Xing, Y. et al. Nucl. Acids Res. 2006 34:3150-3160)
- Cufflinks
How to Assign Reads

- Read vs alignment
- Read type:
  - Single end (one alignment)
  - Paired end (two alignments)
  - Junction (one alignment)
- Region type:
  - Exonic
  - Partial exonic
  - Intronic
  - Intergenic
Differential Expression Detection

- **Cuffdiff**
- **Gene specific analysis:**
  - Identify a statistical model that is the best for a specific gene
  - Use the best mode to test for differential expression for each gene independently

![Best model type](image1)

![Summary of factors influencing expression](image2)
### Generate Gene list

<table>
<thead>
<tr>
<th>Row</th>
<th>Gene symbol</th>
<th>Transcript</th>
<th>Total reads</th>
<th>Q-value</th>
<th>P-value</th>
<th>Ratio</th>
<th>Fold change</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>SF3A1</td>
<td>NM_001005409</td>
<td>1070</td>
<td>0.0265</td>
<td>0.00695</td>
<td>2.01</td>
<td>2.01</td>
</tr>
<tr>
<td>2</td>
<td>MED15</td>
<td>NM_015889</td>
<td>551</td>
<td>0.00363</td>
<td>0.00986</td>
<td>2.02</td>
<td>2.02</td>
</tr>
<tr>
<td>3</td>
<td>POLDIP3</td>
<td>NM_179136</td>
<td>612</td>
<td>5.05E-4</td>
<td>0.0011</td>
<td>2.06</td>
<td>2.06</td>
</tr>
<tr>
<td>4</td>
<td>OSBP2</td>
<td>NM_030758</td>
<td>426</td>
<td>1.7E-5</td>
<td>2.52E-5</td>
<td>2.08</td>
<td>2.08</td>
</tr>
<tr>
<td>5</td>
<td>SLC25A1</td>
<td>NR_046298</td>
<td>220</td>
<td>0.0131</td>
<td>0.0433</td>
<td>2.11</td>
<td>2.11</td>
</tr>
<tr>
<td>6</td>
<td>PPIL2</td>
<td>NM_014337</td>
<td>250</td>
<td>0.108</td>
<td>0.503</td>
<td>2.11</td>
<td>2.11</td>
</tr>
<tr>
<td>7</td>
<td>PES1</td>
<td>NM_001243225</td>
<td>319</td>
<td>5.26E-5</td>
<td>8.88E-5</td>
<td>2.13</td>
<td>2.13</td>
</tr>
<tr>
<td>8</td>
<td>SMCR7L</td>
<td>NM_019008</td>
<td>728</td>
<td>2.4E-7</td>
<td>2.4E-7</td>
<td>2.15</td>
<td>2.15</td>
</tr>
<tr>
<td>9</td>
<td>DESI1</td>
<td>NM_015704</td>
<td>1196</td>
<td>3.5E-8</td>
<td>2.93E-8</td>
<td>2.23</td>
<td>2.23</td>
</tr>
<tr>
<td>10</td>
<td>RANBP1</td>
<td>NM_002882</td>
<td>382</td>
<td>3.22E-5</td>
<td>5.19E-5</td>
<td>2.25</td>
<td>2.25</td>
</tr>
<tr>
<td>11</td>
<td>TNRC6B</td>
<td>NM_015088</td>
<td>713</td>
<td>1.46E-5</td>
<td>2.11E-5</td>
<td>2.26</td>
<td>2.26</td>
</tr>
<tr>
<td>12</td>
<td>AP1B1</td>
<td>NM_001166019</td>
<td>442</td>
<td>0.00208</td>
<td>0.00534</td>
<td>2.28</td>
<td>2.28</td>
</tr>
<tr>
<td>13</td>
<td>TAB1</td>
<td>NM_005116</td>
<td>621</td>
<td>1.04E-5</td>
<td>1.19E-5</td>
<td>2.29</td>
<td>2.29</td>
</tr>
<tr>
<td>14</td>
<td>CASB1</td>
<td>NM_001201429</td>
<td>404</td>
<td>0.00428</td>
<td>0.0118</td>
<td>2.29</td>
<td>2.29</td>
</tr>
<tr>
<td>15</td>
<td>SUN2</td>
<td>NM_015374</td>
<td>1349</td>
<td>9.44E-8</td>
<td>8.75E-8</td>
<td>2.3</td>
<td>2.3</td>
</tr>
<tr>
<td>16</td>
<td>T T L 1 2</td>
<td>NM_015140</td>
<td>602</td>
<td>1.43E-5</td>
<td>1.68E-5</td>
<td>2.32</td>
<td>2.32</td>
</tr>
<tr>
<td>17</td>
<td>ATXN10</td>
<td>NM_001167621</td>
<td>887</td>
<td>1.93E-5</td>
<td>2.87E-5</td>
<td>2.33</td>
<td>2.33</td>
</tr>
</tbody>
</table>
Dot Plot on Differential Expressed Gene

Legend:
- Hispanic
- African
- Caucasian
- European
- Egyptian

Dataset: Raw reads
Group by: Phenotype
Color by: Ethnicity

PABPN1
Genome View on Differential Expressed Gene
Detect Fusion Genes

<table>
<thead>
<tr>
<th>Chromosome1</th>
<th>Start1</th>
<th>Stop1</th>
<th>Chromosome2</th>
<th>Start2</th>
<th>Stop2</th>
<th>Sample ID</th>
<th>Counts</th>
<th>p-value</th>
<th>Genes1</th>
<th>Genes2</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>17</td>
<td>39680649</td>
<td>39680870</td>
<td>17</td>
<td>39767624</td>
<td>ZR751_1_trimmed</td>
<td>12</td>
<td>6.40E-4</td>
<td>KRT19</td>
<td>KRT16</td>
</tr>
<tr>
<td>2</td>
<td>17</td>
<td>39680676</td>
<td>39680883</td>
<td>17</td>
<td>39778659</td>
<td>ZR751_1_trimmed</td>
<td>6</td>
<td>2.56E-3</td>
<td>KRT19</td>
<td>KRT17</td>
</tr>
<tr>
<td>3</td>
<td>20</td>
<td>47580395</td>
<td>47580595</td>
<td>17</td>
<td>37884775</td>
<td>ZR751_1_trimmed</td>
<td>2</td>
<td>0.07</td>
<td>ARFGEF2</td>
<td>ERBB2</td>
</tr>
<tr>
<td>4</td>
<td>17</td>
<td>37884006</td>
<td>37884206</td>
<td>17</td>
<td>39684298</td>
<td>ZR751_1_trimmed</td>
<td>2</td>
<td>0.07</td>
<td>ERBB2</td>
<td>KRT19</td>
</tr>
<tr>
<td>5</td>
<td>17</td>
<td>39021081</td>
<td>39021282</td>
<td>17</td>
<td>39680649</td>
<td>ZR751_1_trimmed</td>
<td>2</td>
<td>0.07</td>
<td>KRT12</td>
<td>KRT19</td>
</tr>
<tr>
<td>6</td>
<td>17</td>
<td>39681121</td>
<td>39681321</td>
<td>11</td>
<td>118917122</td>
<td>ZR751_1_trimmed</td>
<td>2</td>
<td>0.07</td>
<td>KRT19</td>
<td>HYOU1</td>
</tr>
<tr>
<td>7</td>
<td>17</td>
<td>15673265</td>
<td>1679565</td>
<td>7</td>
<td>17171285</td>
<td>ZR751_1_trimmed</td>
<td>2</td>
<td>0.07</td>
<td>ERBB2</td>
<td>KRT19</td>
</tr>
</tbody>
</table>
RNA-seq Workflows in PGS

- **RNA-Seq**
  - Quantification
  - Differential expression detection
  - Alt-splicing detection
  - Allele specific expression

- **MicroRNA-Seq**
  - Quantification
  - Differential expression detection
  - Integration with mRNA
DNA-seq Analysis
DNA-seq Objectives

• Variant calling – find variants that are statistically significant
• Variant annotation:
  • Known SNP? (dbSNP)
  • Population specific (1000 Genome)
• Gene section annotation
  • Exonic, intronic, promoter
• Functional annotation
  • Amino acid change
• Comparing SNVs across samples
  • How man samples share this change? Only in disease samples?
• Structural variation: copy number, chromosomal translocation
• De novo assembly
DNA-seq Data Analysis

**Sequencer** → **.fastq**
- **Alignment**
- **Pre/Post QA/QC**
- **Trimming**
- **SNV Detection**
- **SNV Annotation**

**Flow**

**.bam**
- **Import**
- **QA/QC**
- **SNV Detection**
- **SNV Annotation**
- **SeqDuo/Trio**
- **Visualization**
- **Biological Interpretation**

**Genomics Suite**

**.zip**
Variant Detection in Flow

- **Input**: .bam
- **Output**: .vcf
- **Algorithms**:
  - Genotype likelihood test
    - Efficient for genome wide SNV detection
  - Samtools mpileup
    - Detect SNV and small indels
- **Annotate SNV**
  - dbSNP
  - Gene body
  - Function effect
  - Custom database e.g. COSMIC
Variant Detection in PGS

- Algorithms:
  - Genotype likelihood test
  - Allele Percentage test
    - Targeted region sequencing or small genome (<100Kbp)
    - Mutation occurs across a wider range of proportion

- Type:
  - Detect SNV against reference genome
  - Detect SNV among samples
  - Filter SNV
  - Annotate SNV
  - Find SNVs in multiple samples

- Find Regions in Multiple Samples: 714 in Heart (1043), 329 in Brain (1863), 1334 total.
SeqDuo (identity by state) in PGS

<table>
<thead>
<tr>
<th>Chromosome</th>
<th>Start</th>
<th>Stop</th>
<th>IBS Call</th>
<th>Log Odds</th>
<th>Known SNPs</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>554618</td>
<td>554619</td>
<td>1</td>
<td>2.64921</td>
<td>None</td>
</tr>
<tr>
<td>1</td>
<td>554619</td>
<td>554620</td>
<td>1</td>
<td>2.48908</td>
<td>None</td>
</tr>
<tr>
<td>1</td>
<td>555806</td>
<td>555807</td>
<td>1</td>
<td>1.72146</td>
<td>None</td>
</tr>
<tr>
<td>1</td>
<td>556679</td>
<td>556680</td>
<td>1</td>
<td>2.67347</td>
<td>rs2185537</td>
</tr>
<tr>
<td>1</td>
<td>556691</td>
<td>556692</td>
<td>1</td>
<td>2.69261</td>
<td>None</td>
</tr>
<tr>
<td>1</td>
<td>556955</td>
<td>556956</td>
<td>2</td>
<td>11.3844</td>
<td>rs9326622</td>
</tr>
<tr>
<td>1</td>
<td>556961</td>
<td>556962</td>
<td>2</td>
<td>14.6758</td>
<td>None</td>
</tr>
<tr>
<td>1</td>
<td>557172</td>
<td>557173</td>
<td>1</td>
<td>1.79664</td>
<td>None</td>
</tr>
<tr>
<td>1</td>
<td>557184</td>
<td>557185</td>
<td>1</td>
<td>2.39531</td>
<td>None</td>
</tr>
<tr>
<td>1</td>
<td>557186</td>
<td>557187</td>
<td>1</td>
<td>2.65104</td>
<td>None</td>
</tr>
<tr>
<td>1</td>
<td>557301</td>
<td>557302</td>
<td>1</td>
<td>2.39848</td>
<td>None</td>
</tr>
<tr>
<td>1</td>
<td>557311</td>
<td>557312</td>
<td>1</td>
<td>2.69919</td>
<td>None</td>
</tr>
</tbody>
</table>
SeqTrio in PGS

UPI-P

MI-S

BPI

<table>
<thead>
<tr>
<th>MI-D</th>
<th>MI-S</th>
<th>UPI-M</th>
<th>UPI-P</th>
<th>BPI</th>
<th>NI</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mother</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Father</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Child</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>MI-D</th>
<th>MI-S</th>
<th>UPI-M</th>
<th>UPI-P</th>
<th>BPI</th>
<th>NI</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mother</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Father</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Child</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>MI-D</th>
<th>MI-S</th>
<th>UPI-M</th>
<th>UPI-P</th>
<th>BPI</th>
<th>NI</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mother</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Father</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Child</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

UPI-M

MI-D

BPI

Mendelian Inconsistency – Double Alleles
Mendelian Inconsistency – Single Allele
Uniparental Inheritance – Maternal
Uniparental Inheritance – Paternal
Biparental Inheritance
Not Informative

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>557403</td>
<td>557404</td>
<td>3</td>
<td>1e-006</td>
<td>rs1978001 rs55826372</td>
</tr>
<tr>
<td>2</td>
<td>149355827</td>
<td>149355828</td>
<td>3</td>
<td>1e-0006</td>
<td>None</td>
</tr>
<tr>
<td>9</td>
<td>82370452</td>
<td>82370453</td>
<td>4</td>
<td>10.783</td>
<td>None</td>
</tr>
<tr>
<td>11</td>
<td>10488300</td>
<td>10488301</td>
<td>3</td>
<td>9.7542</td>
<td>None</td>
</tr>
<tr>
<td>14</td>
<td>32024003</td>
<td>32024004</td>
<td>3</td>
<td>9.15236</td>
<td>None</td>
</tr>
<tr>
<td>M</td>
<td>14179</td>
<td>14180</td>
<td>3</td>
<td>9.15236</td>
<td>None</td>
</tr>
<tr>
<td>16</td>
<td>32295475</td>
<td>32295476</td>
<td>4</td>
<td>8.69637</td>
<td>None</td>
</tr>
</tbody>
</table>
Copy Number Analysis in PGS

- Need: genomic DNA of Samples & Reference Sample
- Genomic Binning to count reads in window
- Adjusted by total number of reads / sample
- Create LogRatio (Sample to Normal/reference sample)
- Convert to Copy Number space
- GC Adjustment
- Detect aberration regions
ChIP-seq Analysis
ChIP-seq Objectives

- Detect enriched regions
  - Transcription factor binding site
  - RNA binding proteins
  - Histone modifications
  - Methylated regions
- Motif detection
- Gene section annotation
ChIP-seq Data Analysis

Sequencer → .fastq → Alignment → Pre/Post QA/QC → Trimming Filtering

. bam → .zip → Genomics Suite

Import → QA/QC → Peak Detection → Peak Annotation → Motif Detection → Visualization → Biological Interpretation
ChIP-Seq Workflow in PGS

- Peak Detection
- Motif discovery
- Peak region annotation
- Measure distance of peaks to TSS
Powerful Statistics in PGS

- 60+ methods on data transformation
- Exploratory analysis
  - Principal components analysis
  - Clustering analysis
- Inferential analysis
  - Parametric test (ANOVA, t-test etc.)
  - Non-Parametric test (Chi-square etc.)
- Power analysis
- Survival analysis
- Predictive modeling
Visualization

- Genome Browser
- Heat map
- Dot plot
- Scatter plots
- Profile
- Venn Diagram
- Histogram
- Bar charts, Pie Charts
- Volcano Plots
- MA plots
- ...much more
Integration Approaches

- Combine data from different experiments
- Correlate data from different experiments/assays/platforms
- Visualize data from different experiments/assays/platforms
- Venn Diagram—5-way
- Various “Tools” commands
Biological Interpretation

- Biological relevance is not usually found in only a single gene

- Database:
  - Gene Ontology,
  - KEGG Pathways
  - Custom annotation
    - GMT, GAF, text file

- Method:
  - Enrichment: test if lead genes are overrepresented in any pathway
  - Pathway ANOVA: detect differentially expression pathway
Pathway Analysis

- Find enriched pathway
- Detect differentially expressed pathway
- Visualize gene relationships
- Search for specific pathway and gene
- Support 2000+ species in KEGG database

Partek Pathway
Extends Biological Interpretation
Pathway ANOVA
Partek Provide Solutions for Any Technology

Partek® Flow-GS-Pathway

RT-PCR
Partek Provide Solutions for Any Assay

Gene Expression

Partek® Flow-GS-Pathway

RNA-seq
sRNA-seq

Exon

miRNA

ChIP-chip

CN/LOH /ASCN

Methylation

ChIP-Seq

DNA-seq

SNP Proportion

A C G T N

DNA sequence

C A G C C A
Partek is Your Partner

Free Software Trials
• partek.com/trial
• Email: licensing@partek.com

Self-learning
• Help > On-line tutorials
• Recorded webinars

Regional Technical Support
• Email: support@partek.com
• Phone: +1-314-878-2329