Genome Center Rebels:
Next-Gen Analysis Outside the Machine
Cofactor Genomics

Experimental design
Molecular biology
DNA sequencing
Data analysis
Visualization

Feedback-driven optimization of data generation

Advanced visualization tools enable complex design
The Transition

- Unlimited Sequencing/Reagent Resources
- 1000+ CPU cluster, 100s of TBs of Disk
- Platform LSF job scheduler
- IT staff
- Flexible deadlines
- 99% Human Re-sequencing

- Tight budget with no wiggle room
- Modest compute and storage resources
- Manual job execution
- …No way.
- Customers with grant deadlines!
- 1% human resequencing
Top Requested Applications

**Fragment**  Whole genome characterization by single-pass shotgun sequencing of fragments from total DNA, PCR products, etc.

**Paired-end**  Whole genome characterization by shotgun sequencing from both ends of DNA fragments with ~200bp inserts.

**RNA-Seq**  Quantitative transcriptome profiling by sequencing cDNAs constructed from messenger RNA isolated from total RNA.

**miRNA**  Discovery & quantitation of novel microRNAs and isoforms by sequencing cDNAs of microRNAs isolated from total RNA.

**Bisulfite**  Genome methylation profiling by sequencing DNA fragments bisulfite treated to convert non-methylated C’s into U’s.

**ChIP-Seq**  Discovery & quantitation of protein-DNA interactions by sequencing DNA from immunoprecipitations.
First 10 Months of Libraries

50  RNA-seq
39  Genomic
24  small RNA
20  Genomic Reduced Representation Sequencing (RRS)
19  ChIP-seqs of Transcription Factors and Histone Mods
10  Pooled Patient/Crop PCR
  6  Pooled Loci PCR
  3  ChIP-seq of RNA-binding proteins
  3  Bisulphite converted *Not currently recommended!
  3  DNAse 1 Hypersensitivity (DHS)
  2  “PCR-free” genomic for 2 high AT genomes
  1  Pooled Bacterial Genomes

180  Total Libraries, most Paired-End if relevant
### Recurring Analysis Issues & Idioms

<table>
<thead>
<tr>
<th>Category</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Too Many Tools</strong></td>
<td>A plethora of substitutable tools, few of which are worth using, such as: MAQ, Mosaik, SOAP, SHRiMP, BowTie, NovoAlign or Velvet, Euler-SR, Edena, All-Paths, AMOScmp-shortReads, AbySS</td>
</tr>
<tr>
<td><strong>Poor Algorithms</strong></td>
<td>Single-threaded, compromise-accuracy-for-performance, memory-hog applications like MAQ, Velvet, Euler-SR, ELAND...</td>
</tr>
<tr>
<td><strong>Poor Data Formats</strong></td>
<td>Giant uncompressed TXT files from Illumina &amp; AB, useless design-by-committee formats like SRF, non-standardized formats like GFF</td>
</tr>
<tr>
<td><strong>Recurring Intermediates</strong></td>
<td>Base-by-base coverage, lists of intervals (like annotations or clusters), base-by-base nucleotide count and quality, FASTA, alignments</td>
</tr>
<tr>
<td><strong>Recurring Idioms</strong></td>
<td>Base-by-base whole genome iteration, annotation directed base-by-base iteration, assemblies of non-mapping reads and mapping of original reads to assemblies...</td>
</tr>
</tbody>
</table>
Analysis on a Shoestring - Illumina Barcoded ChIP

- Re-write Illumina barcode binning (saves 20MB -- an old 454 run’s worth of data -- per lane)
- Compute base-by-base coverage from alignments
- Iterate through bases to find regions of contiguous coverage
**Analysis on a Shoestring - SNPs in Pooled Samples**

- Tally base-by-base nucleotide frequencies and qualities
- Pick thresholds from simple error model
- 83% validated at predicted population level by RT-PCR!
- MAQ - frequent false positives and false negatives

<table>
<thead>
<tr>
<th>CONTIG POSITION</th>
<th>DEPTH</th>
<th>VARIANT</th>
<th>V.DEPTH</th>
<th>%CALLS</th>
<th>%QUALS</th>
<th>V.QUAL</th>
<th>QUALS</th>
</tr>
</thead>
<tbody>
<tr>
<td>scaffold_185</td>
<td>11915</td>
<td>G</td>
<td>59</td>
<td>0.9516</td>
<td>0.9891</td>
<td>1266</td>
<td>1280</td>
</tr>
<tr>
<td>scaffold_152</td>
<td>238099</td>
<td>T</td>
<td>85</td>
<td>0.9444</td>
<td>0.9221</td>
<td>1266</td>
<td>1373</td>
</tr>
<tr>
<td>scaffold_15</td>
<td>393751</td>
<td>A</td>
<td>75</td>
<td>0.9740</td>
<td>0.9945</td>
<td>1266</td>
<td>1273</td>
</tr>
<tr>
<td>scaffold_147</td>
<td>12004</td>
<td>G</td>
<td>50</td>
<td>0.8333</td>
<td>0.9806</td>
<td>1266</td>
<td>1291</td>
</tr>
<tr>
<td>scaffold_1133</td>
<td>6860</td>
<td>C</td>
<td>70</td>
<td>0.9859</td>
<td>0.9969</td>
<td>1266</td>
<td>1270</td>
</tr>
<tr>
<td>scaffold_52</td>
<td>103072</td>
<td>G</td>
<td>70</td>
<td>1.0000</td>
<td>1.0000</td>
<td>1265</td>
<td>1265</td>
</tr>
<tr>
<td>scaffold_3</td>
<td>680248</td>
<td>C</td>
<td>59</td>
<td>0.9833</td>
<td>0.9976</td>
<td>1265</td>
<td>1268</td>
</tr>
<tr>
<td>scaffold_229</td>
<td>75856</td>
<td>C</td>
<td>92</td>
<td>0.8679</td>
<td>0.9426</td>
<td>1265</td>
<td>1342</td>
</tr>
<tr>
<td>scaffold_213</td>
<td>71821</td>
<td>T</td>
<td>83</td>
<td>0.8737</td>
<td>0.9664</td>
<td>1265</td>
<td>1309</td>
</tr>
<tr>
<td>scaffold_15</td>
<td>391030</td>
<td>C</td>
<td>61</td>
<td>0.8841</td>
<td>0.9664</td>
<td>1265</td>
<td>1309</td>
</tr>
<tr>
<td>scaffold_140</td>
<td>135979</td>
<td>T</td>
<td>51</td>
<td>1.0000</td>
<td>1.0000</td>
<td>1265</td>
<td>1265</td>
</tr>
<tr>
<td>scaffold_14</td>
<td>953190</td>
<td>A</td>
<td>61</td>
<td>0.8971</td>
<td>0.9708</td>
<td>1265</td>
<td>1303</td>
</tr>
<tr>
<td>scaffold_58</td>
<td>76069</td>
<td>C</td>
<td>72</td>
<td>0.9730</td>
<td>0.9961</td>
<td>1264</td>
<td>1269</td>
</tr>
<tr>
<td>scaffold_49</td>
<td>374940</td>
<td>G</td>
<td>60</td>
<td>0.9623</td>
<td>0.9929</td>
<td>1264</td>
<td>1273</td>
</tr>
<tr>
<td>scaffold_35</td>
<td>1027666</td>
<td>C</td>
<td>61</td>
<td>0.9474</td>
<td>0.9937</td>
<td>1264</td>
<td>1272</td>
</tr>
<tr>
<td>scaffold_153</td>
<td>150</td>
<td>G</td>
<td>36</td>
<td>0.9855</td>
<td>0.9922</td>
<td>1264</td>
<td>1274</td>
</tr>
<tr>
<td>scaffold_152</td>
<td>131015</td>
<td>G</td>
<td>68</td>
<td>0.9800</td>
<td>0.9875</td>
<td>1263</td>
<td>1279</td>
</tr>
<tr>
<td>scaffold_15</td>
<td>392560</td>
<td>G</td>
<td>70</td>
<td>0.9722</td>
<td>0.9961</td>
<td>1262</td>
<td>1267</td>
</tr>
<tr>
<td>scaffold_4739</td>
<td>644</td>
<td>C</td>
<td>73</td>
<td>0.9865</td>
<td>0.9992</td>
<td>1262</td>
<td>1263</td>
</tr>
<tr>
<td>scaffold_22</td>
<td>460616</td>
<td>T</td>
<td>53</td>
<td>1.0000</td>
<td>1.0000</td>
<td>1262</td>
<td>1263</td>
</tr>
<tr>
<td>scaffold_170</td>
<td>42834</td>
<td>T</td>
<td>53</td>
<td>1.0000</td>
<td>1.0000</td>
<td>1262</td>
<td>1263</td>
</tr>
<tr>
<td>scaffold_125</td>
<td>31030</td>
<td>G</td>
<td>51</td>
<td>0.8644</td>
<td>0.9776</td>
<td>1262</td>
<td>1263</td>
</tr>
<tr>
<td>scaffold_105</td>
<td>314020</td>
<td>C</td>
<td>64</td>
<td>0.9103</td>
<td>0.9813</td>
<td>1262</td>
<td>1286</td>
</tr>
<tr>
<td>scaffold_45</td>
<td>5453</td>
<td>T</td>
<td>66</td>
<td>0.9103</td>
<td>0.9813</td>
<td>1262</td>
<td>1286</td>
</tr>
<tr>
<td>scaffold_36</td>
<td>232766</td>
<td>T</td>
<td>76</td>
<td>0.9268</td>
<td>0.9785</td>
<td>1262</td>
<td>1278</td>
</tr>
<tr>
<td>scaffold_21</td>
<td>455896</td>
<td>G</td>
<td>71</td>
<td>0.9103</td>
<td>0.9813</td>
<td>1262</td>
<td>1286</td>
</tr>
<tr>
<td>scaffold_195</td>
<td>125468</td>
<td>T</td>
<td>57</td>
<td>0.7231</td>
<td>0.9575</td>
<td>1262</td>
<td>1318</td>
</tr>
<tr>
<td>scaffold_18</td>
<td>100541</td>
<td>A</td>
<td>65</td>
<td>0.9559</td>
<td>0.9760</td>
<td>1262</td>
<td>1293</td>
</tr>
<tr>
<td>scaffold_167</td>
<td>59962</td>
<td>G</td>
<td>64</td>
<td>0.9846</td>
<td>0.9992</td>
<td>1262</td>
<td>1263</td>
</tr>
<tr>
<td>scaffold_167</td>
<td>112530</td>
<td>A</td>
<td>48</td>
<td>0.9796</td>
<td>0.9968</td>
<td>1262</td>
<td>1266</td>
</tr>
</tbody>
</table>
Analysis on a Shoestring - De Novo Transcriptome Assembly

- Fix bugs in AMOScmp-shortRead ... makes no joins after 3 days
- Try to install Euler-SR...failed & one we know can get it installed either ;-) 
- Add 2 new strategies to Velvet (default De Bruijn graphs are too naïve)
  - Each increases mean/median contig length by up to 2X (4X total, constant total size)
- Simple alignments + graph theory to aggregate results of Multi-stage assemblies
- Align reads back to assembly for expression (& to avoid generating/parsing Velvet TXT files)
- HMMER3 alpha identifies all of the Core Eukaryotic Genes (CEGMA set of 437)

<table>
<thead>
<tr>
<th>gene</th>
<th>167</th>
<th>0 length</th>
<th>1602 frame</th>
<th>5 Ribonucleotide reductase, alpha subunit</th>
</tr>
</thead>
<tbody>
<tr>
<td>gene</td>
<td>149</td>
<td>0 length</td>
<td>2221 frame</td>
<td>3 N-arginine dibasic convertase NRD1 and related Zn2+ ATPase</td>
</tr>
<tr>
<td>gene</td>
<td>115</td>
<td>5.20E-277 length</td>
<td>1550 frame</td>
<td>2 DNA replication licensing factor, MCM5 component</td>
</tr>
<tr>
<td>gene</td>
<td>89</td>
<td>3.50E-276 length</td>
<td>1348 frame</td>
<td>1 DNA replication licensing factor, MCM5 component</td>
</tr>
<tr>
<td>gene</td>
<td>278</td>
<td>1.20E-261 length</td>
<td>1451 frame</td>
<td>4 US snRNP splicing subunit</td>
</tr>
<tr>
<td>gene</td>
<td>56</td>
<td>6.50E-256 length</td>
<td>2198 frame</td>
<td>6 RNA helicase</td>
</tr>
<tr>
<td>gene</td>
<td>102</td>
<td>3.90E-231 length</td>
<td>1263 frame</td>
<td>5 Phosphoglucomutase</td>
</tr>
<tr>
<td>gene</td>
<td>104</td>
<td>4.00E-228 length</td>
<td>1379 frame</td>
<td>4 WD40 repeat nuclear protein Bop1, involved in ribosome biogenesis</td>
</tr>
<tr>
<td>gene</td>
<td>169</td>
<td>1.10E-204 length</td>
<td>1192 frame</td>
<td>6 RNA polymerase II transcription initiation/nucleotide</td>
</tr>
<tr>
<td>gene</td>
<td>295</td>
<td>1.30E-190 length</td>
<td>924 frame</td>
<td>2 Predicted P-loop ATPase fused to an acetyltransferase</td>
</tr>
<tr>
<td>gene</td>
<td>88</td>
<td>4.90E-183 length</td>
<td>819 frame</td>
<td>4 DNA replication licensing factor, MCM2 component</td>
</tr>
<tr>
<td>gene</td>
<td>38</td>
<td>1.20E-179 length</td>
<td>980 frame</td>
<td>1 RNA polymerase III, large subunit</td>
</tr>
<tr>
<td>gene</td>
<td>188</td>
<td>2.30E-174 length</td>
<td>1378 frame</td>
<td>3 GDP-mannose pyrophosphorylase/mannose-1-phosphate uridyltransferase</td>
</tr>
<tr>
<td>gene</td>
<td>31</td>
<td>1.10E-170 length</td>
<td>750 frame</td>
<td>3 Alanyl-tRNA synthetase</td>
</tr>
<tr>
<td>gene</td>
<td>159</td>
<td>4.00E-170 length</td>
<td>979 frame</td>
<td>1 Vesicle coat complex COP1, beta subunit</td>
</tr>
<tr>
<td>gene</td>
<td>182</td>
<td>7.10E-170 length</td>
<td>949 frame</td>
<td>5 Glucosamine 6-phosphate synthetases, contain amido transferase</td>
</tr>
<tr>
<td>gene</td>
<td>351</td>
<td>2.20E-169 length</td>
<td>1020 frame</td>
<td>6 Phenylalanyl-tRNA synthetase</td>
</tr>
<tr>
<td>gene</td>
<td>83</td>
<td>3.80E-168 length</td>
<td>1416 frame</td>
<td>4 ATPase component of ABC transporters with duplicated fungal protein</td>
</tr>
<tr>
<td>gene</td>
<td>86</td>
<td>2.80E-159 length</td>
<td>883 frame</td>
<td>5 Mitochondrial translation elongation factor Tu</td>
</tr>
<tr>
<td>gene</td>
<td>147</td>
<td>3.40E-156 length</td>
<td>688 frame</td>
<td>3 US snRNP-specific protein</td>
</tr>
<tr>
<td>gene</td>
<td>47</td>
<td>1.50E-154 length</td>
<td>1155 frame</td>
<td>2 Nuclear exosomal RNA helicase MTR4, DEAD-box superfamily</td>
</tr>
<tr>
<td>gene</td>
<td>90</td>
<td>1.60E-153 length</td>
<td>857 frame</td>
<td>2 Ribosome Assembly protein</td>
</tr>
<tr>
<td>gene</td>
<td>176</td>
<td>1.40E-150 length</td>
<td>810 frame</td>
<td>1 HAT repeat protein</td>
</tr>
<tr>
<td>gene</td>
<td>180</td>
<td>4.90E-150 length</td>
<td>785 frame</td>
<td>1 Acyl-CoA synthetase</td>
</tr>
<tr>
<td>gene</td>
<td>68</td>
<td>2.20E-146 length</td>
<td>673 frame</td>
<td>5 Karyopherin (importin) beta 1</td>
</tr>
<tr>
<td>gene</td>
<td>153</td>
<td>3.30E-144 length</td>
<td>688 frame</td>
<td>3 Serine/threonine specific protein phosphatase involved in biological regulation</td>
</tr>
<tr>
<td>gene</td>
<td>80</td>
<td>3.30E-144 length</td>
<td>615 frame</td>
<td>5 Vestec protein clathrin, heavy chain</td>
</tr>
<tr>
<td>gene</td>
<td>233</td>
<td>4.50E-144 length</td>
<td>642 frame</td>
<td>3 Isocitryl-tRNA synthetase</td>
</tr>
<tr>
<td>gene</td>
<td>148</td>
<td>7.70E-144 length</td>
<td>757 frame</td>
<td>4 DNA/RNA helicase MER3/SLH1, DEAD-box superfamili</td>
</tr>
<tr>
<td>gene</td>
<td>229</td>
<td>1.20E-141 length</td>
<td>611 frame</td>
<td>256S proteasome regulatory complex, subunit RPNI</td>
</tr>
<tr>
<td>gene</td>
<td>46</td>
<td>5.20E-141 length</td>
<td>666 frame</td>
<td>3 Vesicle coat complex COP1, alpha subunit</td>
</tr>
<tr>
<td>gene</td>
<td>170</td>
<td>1.90E-136 length</td>
<td>578 frame</td>
<td>1 mRNA cleavage and polyadenylation factor II complex</td>
</tr>
<tr>
<td>gene</td>
<td>219</td>
<td>4.90E-135 length</td>
<td>837 frame</td>
<td>6 NADP-dependent isocitrate dehydrogenase</td>
</tr>
<tr>
<td>gene</td>
<td>69</td>
<td>1.10E-133 length</td>
<td>673 frame</td>
<td>6 Serine/threonine specific protein phosphatase involved in biological regulation</td>
</tr>
<tr>
<td>gene</td>
<td>168</td>
<td>1.10E-133 length</td>
<td>758 frame</td>
<td>1 RNA polymerase II transcription initiation/nucleotide</td>
</tr>
<tr>
<td>gene</td>
<td>140</td>
<td>4.50E-132 length</td>
<td>1077 frame</td>
<td>2 DEAH-box RNA helicase</td>
</tr>
<tr>
<td>gene</td>
<td>4</td>
<td>4.70E-131 length</td>
<td>736 frame</td>
<td>1 Structural maintenance of chromosomes protein 1 (sMC1)</td>
</tr>
<tr>
<td>gene</td>
<td>320</td>
<td>1.70E-130 length</td>
<td>1023 frame</td>
<td>2 Phosphoglucomutase/phosphomannomutase</td>
</tr>
<tr>
<td>gene</td>
<td>232</td>
<td>1.10E-126 length</td>
<td>678 frame</td>
<td>4 Conserved protein Mo25</td>
</tr>
<tr>
<td>gene</td>
<td>145</td>
<td>3.10E-126 length</td>
<td>615 frame</td>
<td>1 Adaptor complexes medium subunit family</td>
</tr>
<tr>
<td>gene</td>
<td>33</td>
<td>2.20E-124 length</td>
<td>571 frame</td>
<td>4 P-type ATPase</td>
</tr>
<tr>
<td>gene</td>
<td>172</td>
<td>3.10E-124 length</td>
<td>555 frame</td>
<td>4 Glutaminyl-tRNA synthetase</td>
</tr>
<tr>
<td>gene</td>
<td>319</td>
<td>5.70E-122 length</td>
<td>512 frame</td>
<td>6 RNA polymerase II elongator complex, subunit ELPS13</td>
</tr>
<tr>
<td>gene</td>
<td>280</td>
<td>2.70E-121 length</td>
<td>685 frame</td>
<td>1 Ubiquitin fusion-degradation protein</td>
</tr>
<tr>
<td>gene</td>
<td>152</td>
<td>2.10E-119 length</td>
<td>555 frame</td>
<td>2 DNA polymerase delta, catalytic subunit</td>
</tr>
<tr>
<td>gene</td>
<td>114</td>
<td>3.10E-118 length</td>
<td>589 frame</td>
<td>6 AAA+ type ATPase containing the peptidase M41 domain</td>
</tr>
<tr>
<td>gene</td>
<td>210</td>
<td>4.00E-118 length</td>
<td>587 frame</td>
<td>5 26S proteasome regulatory complex, subunit RPNI</td>
</tr>
<tr>
<td>gene</td>
<td>406</td>
<td>9.30E-118 length</td>
<td>673 frame</td>
<td>4 Predicted RNA-binding protein Pno1p interacting with Nsp1 and Nsp2</td>
</tr>
<tr>
<td>gene</td>
<td>363</td>
<td>4.20E-117 length</td>
<td>1352 frame</td>
<td>5 sn-1,2-diacylglycerol ethanolamine, and cholenoepinephrine 2-deacyltransferase</td>
</tr>
<tr>
<td>gene</td>
<td>122</td>
<td>1.00E-116 length</td>
<td>777 frame</td>
<td>2 Dehydrogenase kinase</td>
</tr>
</tbody>
</table>
Analysis on a Shoestring - Allele Specific Expression

- Sequencing from F1 Generation of In-bred P's
- Competitive Mapping to P Genomes
- Compute fractional, parallel base-by-base coverage
- Look for genes with high coverage in only 1 genome

>My FASTA-Like Coverage File

```
0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00
0.23 0.23 0.35 0.57 0.57 0.78 0.78 1.10 1.10 1.23 1.23 1.23 1.23 1.33 1.33 1.33 1.33 1.49 1.49 1.58 1.58
1.18 1.18 2.31 2.31 2.31 2.31 2.31 2.31 2.31 2.31 2.31 2.31 2.31 2.31 2.31 2.31 2.31 2.31 2.31 2.31 2.31
4.72 4.72 4.72 4.72 4.72 4.72 4.72 4.72 4.72 4.72 4.72 4.72 4.72 4.72 4.72 4.72 4.72 4.72 4.72 4.72 4.72
5.21 5.21 5.21 5.21 5.21 5.21 5.21 5.21 5.21 5.21 5.21 5.21 5.21 5.21 5.21 5.21 5.21 5.21 5.21 5.21 5.21
```

-5 0 500 1000 1500 2000 2500

Reference Position

Mapped Read Depth
Cofactor Browser

nhr-80
Multi-Faceted Sequencing & Analysis

- Assembly
- Expression Analysis
- Gene Annotation
- Nucleosome Profiling
- Mass Spec & 2D PAGE
Even Higher Order: Temporal/Spatial

Time Point 1

Tissue 1

... 

Tissue Y

Time Point X

...
Highest Order Data & Computation

Public DB

Project 1

Project N

Investigator 1

...  

Investigator N
How Can We Get There Analysis-wise?

• No More Text files
• No More Perl Engines
• No More Single-threaded Apps
• No More Clusters, Configuration, or Installation
• Standard Indexed Binary Formats
• Free, Universal Computational Engine
• Open & Extensible System for Everyone
Our Solution - The Genome Operating System

GUI

PDF

expression

applications

Gentuition UI

‘CPU’

Programming API

SNPs

Network

‘File System’

PostgreSQL

Genomes

Reads

Annotations

Alignments

NovoAlign

PRB

Meta Data
**Genome OS - Design Principles**

- **No Cluster Required**  
*Efficient Algos, Fancy structs, 100% Multithreaded*

- **No Petabyte Disks Required**  
*Custom Binary Storage & Intermediates*

- **Zero-Configuration**

- **Platform Independent**  
*All Java ‘Binaries’*

- **Scales to Size of Project**  
*Small footprint when sparse, low overhead when dense*

- **Open & Extensible**  
*Open specification & stellar API*

- **Smart & Preemptive Computation**  
*Meta-Data you actually care about*
**Genome OS - Progress/Current Example**

genuition new named CancerProject
genuition CancerProject set ProjectType RNAseq
genuition CancerProject add sample named Tumor
genuition CancerProject add sample named Normal
genuition CancerProject add references from Human37.fasta
genuition CancerProject/Tumor add alignments from tumor_novos/
genuition CancerProject/Normal add alignments from normal_novos/
genuition CancerProject Compare

**In only 36 MINUTES:**
On one 8-core Apple Xserve, 32 GB RAM, 1TB Disk
Parses 8 giga-bases of Paired-End alignments (160 GB)
Organizes isoforms of all genes
Computes gene-by-gene expression profile both samples
Computes unannotated expression patches and depths
Creates PDF graph of expression
Stores it all back in 340 MB (0.2% of input)
**Genome OS - Time/Space Complexity**

O(1) To select genome “partition” that easily fits into RAM

O(\log M) To seek to desired start location, \( M = \text{Min}(\text{partition length, data entries}) \)

O(1) To seek to successor of any location

O(1) Access to read, alignment, gene/transcript/exon, cluster, reference from any position

O(1) Access to custom defined fields

Amortized O(1) To keep search structures balanced

**No Locking** Absolutely no locks or spinning, not even to rebalance search structures

**All Single-Pass** Consensus, expression levels, novel read clusters, all computed as data is read from disk

**Load Balanced** Parallelized *within* chromosomes and *within* files to be free from balance assumptions

**No Re-Parsing** Data structures are stored as serialized Java objects

**All Binary** All data is stored in the smallest possible primitive type and then compressed

Amortized O(1) Space for search structure overhead, scales with lesser of genome or data size
public class ExonCoverageTask extends AWGAAAlignmentTask {

    ExonRegion exon = null;
    int exonTotalCoverage = 0;

    public ExonCoverageTask() {
        super(true);
    }

    public void setup(ExonRegion exon) {
        this.exon = exon;
        this.exonTotalCoverage = 0;
    }

    public void executeTask(Alignment aln) {
        int alnEnd = aln.alignmentLeftPosition + gentuition.readLength - 1;
        exonTotalCoverage += Math.max(Math.min(alnEnd, exon.getEnd()) - Math.max(aln.alignmentLeftPosition, exon.getStart()) + 1, 0);
    }

    public void cleanup() {
        this.exon.define("total-coverage", exonTotalCoverage);
    }
}
public class ExonCoverageTask extends AWGAAAlignmentTask {

    ExonRegion exon = null;
    int exonTotalCoverage = 0;

    public ExonCoverageTask(){
        super(true);
    }

    public void setup(ExonRegion exon){
        this.exon = exon;
        this.exonTotalCoverage = 0;
    }

    public void executeTask(Alignment aln){
        int alnEnd = aln.alignmentLeftPosition + gentuition.readLength - aln.getInsertions().size() + aln.getDeletions().size() - 1;
        exonTotalCoverage += Math.max(Math.min(alnEnd, exon.getEnd()) - Math.max(aln.alignmentLeftPosition, exon.getStart()) + 1, 0);
        for (Byte next : aln.getDeletions()) {
            int deleteLocation = aln.alignmentLeftPosition + next - 1;
            if (deleteLocation >= exon.getStart() && deleteLocation <= exon.getEnd()){
                exonTotalCoverage--;
            }
        }
    }

    public void cleanup(){
        this.exon.define("total-coverage", exonTotalCoverage);
    }
}
Genome OS - Current Development

- **GUI**
- **PDF**
- **Expression**
- **Applications**
- **SNPs**
- **Programming API**
- **Meta Data**
- **Network**

**Applications**
- **MapReads**
- **NovoAlign**
- **CSfasta**
- **PRB**

**Database**
- **'File System'**
  - Genomes
  - Reads
  - Annotations
  - Alignments

- **PostgreSQL**
- **Gentuition UI**
- **Gentuition API**
- **Expression**
- **Genes**
- **SNPs**
- **MapReads**
- **NovoAlign**
- **CSfasta**
- **Network**
- **GUI**
Genome OS - Release Timeline

September 1

- Programming API
  - ‘CPU’
  - Free & Open Spec

- Community Web Site

December 1

- Command Line
  - Open Source
  - PostgreSQL
  - PRB
  - S/BAM
  - CSfasta
  - MapReads

- GUI Display Protocol
  - Network
  - FS

- NovoAlign
- Ensembl GTF

Open Source

Copyright Cofactor Genomics
Special Thank You
to my Computer Science Interns:

Michael Fahey
Jonathan Wald