“Choose your own adventure” clinical scenario

A patient has been diagnosed with a glioblastoma multiforme brain tumor, a disease with very poor survival. You are the treating physician and would like to know if the patient harbors any genetic alterations associated with prognosis or treatment.

You would like to look at multiple exons of 50 cancer associated genes to assess the presence of a genetic alteration.

What test should you run?

- You could use single or multiplex PCR followed by Sanger sequencing. However, this would be time consuming and costly.
- You could use a targeted sequencing Next Generation Sequencing (NGS) panel to sequence all 50 genes at once.
Next Generation Sequencing (NGS)

• NGS is a quantitative, high-throughput method of digitally sequencing nucleic acid

• Different types of NGS
  • Second generation sequencers
    • Illumina (Sequencing by synthesis)
    • Ion Torrent (Semi-conductor sequencing)
  • Third generations sequencers
    • Pac-Bio (Single Molecule, sequencing by synthesis)

• Take advantage of natural DNA replication mechanisms

• Quantitative nature allows the determination of the variant allele fraction within a sample
DNA sequencing timeline

1953 - DNA Structure Discovery
1977 - Sanger sequencing developed
1990-2003 Human Genome Project
2005-2006 Roche 454 developed
2006-2007 Illumina sequencer on market
2010-2011 Ion Torrent on market
2011-2014 PacBio on market
2014-2014 MinION on market
**Sanger Sequencing**

- Use of incorporating fluorescent nucleotides (ddNTPs) during DNA replication to sequence DNA
- Also called “sequencing by synthesis”
- Developed in 1977 by Fred Sanger
  - Won the Nobel Prize in 1980 for methodology
- Allows for sequencing up to ~700bp
  - Single reactions w/ no multiplexing capacity
  - Limited in scope
  - Qualitative, not quantitative
Sanger Sequencing

2 Pyrophosphates and a hydrogen released after nucleotide incorporation

Lack of hydroxyl group terminates growing nucleotide chain
Human Genome Project

- Many of the advances in genome sequencing were propelled by the Human Genome Project (HGP)

- HGP started in 1990, was an international, collaborative effort to complete map all genes in the human genome.
  - Baylor College of Medicine was one of the original sites that conducted sequencing for HGP

- HGP characterized the location and sequence of ~20,500 genes in the 3 billion bp human genome.

- The International Human Genome Sequencing Consortium published the first draft of the human genome in the journal Nature in February 2001

- The full sequence was completed and published in April 2003.
  - One of the only government projects completed on time and under budget

- "It's [Human Genome] a history book - a narrative of the journey of our species through time. It's a shop manual, with an incredibly detailed blueprint for building every human cell. And it's a transformative textbook of medicine, with insights that will give health care providers immense new powers to treat, prevent and cure disease." - Francis Collins

- The HGP provided the human reference sequence allowing for a standard reference when mapping sequences to the human genome
Shotgun Sequencing

- Used to sequence whole genomes
- Steps:
  - DNA is broken up randomly into smaller fragments
  - Dideoxy method produces reads
  - Look for overlap of reads

<table>
<thead>
<tr>
<th>Strand</th>
<th>Sequence</th>
</tr>
</thead>
<tbody>
<tr>
<td>First Shotgun Sequence</td>
<td>AGCATGCTGCA GTCATGCT-----------------------TAGGCTA</td>
</tr>
<tr>
<td>Second Shotgun Sequence</td>
<td>AGCATG--------------------CTGCA GT CATGCT TAGGCTA</td>
</tr>
<tr>
<td>Reconstruction</td>
<td>AGCATGCTGCA GT CATGCT TAGGCTA</td>
</tr>
</tbody>
</table>
The Human Genome Project was one of only a few government funded projects to finish under-budget and on-time.
Two main NGS platforms that are widely used today...

Illumina
MiSeq/NextSeq/HiSeq
“Sequencing by Synthesis”

Ion
PGM/Proton
Semi-conductor sequencing
Illumina Sequencing Technology

Illumina video detailing sequencing technology: http://www.youtube.com/embed/HMyCqWhwB8E?iframe&rel=0&autoplay=1
Ion Torrent Sequencing Technology

Ion Torrent Quality Report

Ion Torrent Quality Report

298 M
Total Reads

98
Base Quality

ISP Density

2,600,563
Total Reads

47 %
Undersampled

90 %
Loading

5,709,140

10 %
Empty Wells

100 %
Enrichment

5,705,817

0 %
No Template

51 %
Clonal

2,917,116

49 %
Polyclonal

92 %
Final Library

2,690,563
1 %
Test Fragments
0 %
Adapter Dimer
7 %
Low Quality

111 bp
Mean

109 bp
Median

89 bp
Mode

Road Length

100
200
300
Read Length

275 M
Total Aligned Bases

0.1X
Reference Coverage

2,576,120

97 %
Aligned Bases

3 %
Unaligned

99.4 %
Mean Read Alignment %

261 M
All Alignment Reads

Alignment Quality

AG1 F
AG2 F
Perfect

Total Number of Bases [bp]
251 M
242 M
217 M

Mean Length [bp]
100
102
94

Longest Alignment [bp]
322
310
293

Mean Coverage Depth [
0.1
0.1
0.1

<table>
<thead>
<tr>
<th>Count</th>
<th>%</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total Reads</td>
<td>2,576,120</td>
</tr>
<tr>
<td>Aligned Reads</td>
<td>2,553,799</td>
</tr>
<tr>
<td>Unaligned Reads</td>
<td>22,321</td>
</tr>
</tbody>
</table>

Aligned to Homo sapiens
Ion Torrent Community and Plugin Store

https://ioncommunity.thermofisher.com/community/

Recent Content

- Plugin: AmpliCompare
  7 days ago
  by Thermo Fisher Administrator

- Plugin: AmpliconCoveragePlots
  7 days ago
  by Thermo Fisher Administrator

- Plugin: Assembler
  7 days ago
  by Thermo Fisher Administrator

- Plugin: AssemblerSPAdes
  7 days ago
  by Thermo Fisher Administrator

- Plugin: coverageAnalysis
  7 days ago
  by Thermo Fisher Administrator

- Plugin: DataXfer
  7 days ago
  by Thermo Fisher Administrator

- Plugin: snpEff
  7 days ago
  by Thermo Fisher Administrator

- Plugin: macsPeakFinder
  1 week ago
  by Thermo Fisher Administrator

- Plugin: variantCallerForMIDNA
  1 week ago
  by Thermo Fisher Administrator

- Plugin: wholeTranscriptomeAnalysisPreview
  1 week ago
  by Thermo Fisher Administrator

- Plugin: SNVQ
  1 week ago
  by Thermo Fisher Administrator

- Plugin: sampleIdentifier
  1 week ago
  by Thermo Fisher Administrator
Ion Torrent Output

coverageAnalysis

<table>
<thead>
<tr>
<th>Barcode Name</th>
<th>Sample</th>
<th>Mapped Reads</th>
<th>On Target</th>
<th>Mean Depth</th>
<th>Uniformity</th>
</tr>
</thead>
<tbody>
<tr>
<td>T335-0_002</td>
<td>T335-0_new kit</td>
<td>558,644</td>
<td>68.77%</td>
<td>5.11x</td>
<td>46.13%</td>
</tr>
<tr>
<td>T333-0_003</td>
<td>T333-0_new kit</td>
<td>1,154,053</td>
<td>76.24%</td>
<td>4.23x</td>
<td>97.03%</td>
</tr>
</tbody>
</table>

variantCaller

<table>
<thead>
<tr>
<th>Barcode Name</th>
<th>Sample</th>
<th>Mapped Reads</th>
<th>Variants</th>
<th>Hotspot Variants</th>
<th>Overdispersed</th>
</tr>
</thead>
<tbody>
<tr>
<td>T335-0_001</td>
<td>T335-0_new kit</td>
<td>558,644</td>
<td>33</td>
<td>VCF0.2</td>
<td>X8.3</td>
</tr>
<tr>
<td>T335-0_002</td>
<td>T335-0_new kit</td>
<td>1,154,053</td>
<td>29</td>
<td>VCF0.2</td>
<td>X8.3</td>
</tr>
</tbody>
</table>

Output Files

<table>
<thead>
<tr>
<th>Barcode Name</th>
<th>Sample</th>
<th>Mapped Reads</th>
<th>Mean Read Length</th>
<th>Read Length Histogram</th>
<th>Files</th>
</tr>
</thead>
<tbody>
<tr>
<td>T335-0_001</td>
<td>T335-0_new kit</td>
<td>558,644</td>
<td>759,863</td>
<td>0-200</td>
<td>USAM, BAM, BAI</td>
</tr>
<tr>
<td>T335-0_002</td>
<td>T335-0_new kit</td>
<td>1,154,053</td>
<td>1,108,148</td>
<td>0-200</td>
<td>USAM, BAM, BAI</td>
</tr>
</tbody>
</table>
## Platform Comparisons

**Table 1: Price comparison of benchtop instruments and sequencing runs**  
*Loman et al 2012*

<table>
<thead>
<tr>
<th>Platform</th>
<th>List price</th>
<th>Approximate cost per run</th>
<th>Minimum throughput (read length)</th>
<th>Run time</th>
<th>Cost/Mb</th>
<th>Mb/h</th>
</tr>
</thead>
<tbody>
<tr>
<td>454 GS Junior</td>
<td>$108,000</td>
<td>$1,100</td>
<td>35 Mb (400 bases)</td>
<td>8 h</td>
<td>$31</td>
<td>4.4</td>
</tr>
<tr>
<td>Ion Torrent PGM (314 chip)</td>
<td>$80,490a,b</td>
<td>$225c</td>
<td>10 Mb (100 bases)</td>
<td>3 h</td>
<td>$22.5</td>
<td>3.3</td>
</tr>
<tr>
<td>(316 chip)</td>
<td>$425</td>
<td></td>
<td>100 Mb (100 bases)</td>
<td>3 h</td>
<td>$4.25</td>
<td>33.3</td>
</tr>
<tr>
<td>(318 chip)</td>
<td>$625</td>
<td></td>
<td>1,000 Mb (100 bases)</td>
<td>3 h</td>
<td>$0.63</td>
<td>333.3</td>
</tr>
<tr>
<td>MiSeq</td>
<td>$125,000</td>
<td>$750</td>
<td>1,500 Mb (2 × 150 bases)</td>
<td>27 h</td>
<td>$0.5</td>
<td>55.5</td>
</tr>
</tbody>
</table>

**Table 1: Insertion/deletion and substitution errors on read level for benchtop NGS platforms**  
*Junemann et al 2012*

<table>
<thead>
<tr>
<th>Platform</th>
<th>Sequencing kit</th>
<th>Library</th>
<th>Strain</th>
<th>Date of sequencing</th>
<th>Indels per 100 bp</th>
<th>Indels per read</th>
<th>Substitutions per 100 bp</th>
<th>Substitutions per read</th>
</tr>
</thead>
<tbody>
<tr>
<td>GSJ</td>
<td>GSJ Titanium</td>
<td>Nebulization / AMPure XP</td>
<td>Sakai</td>
<td>June 2012</td>
<td>0.4011</td>
<td>1.8351</td>
<td>0.0543</td>
<td>0.2484</td>
</tr>
<tr>
<td>MiSeq</td>
<td>2 × 150-bp PE</td>
<td>Nextera</td>
<td>Sakai</td>
<td>June 2012</td>
<td>0.0009</td>
<td>0.0013</td>
<td>0.0921</td>
<td>0.1318</td>
</tr>
<tr>
<td>MiSeq</td>
<td>2 × 250-bp PE</td>
<td>Nextera</td>
<td>Sakai</td>
<td>September 2012</td>
<td>0.0009</td>
<td>0.0018</td>
<td>0.0940</td>
<td>0.2033</td>
</tr>
<tr>
<td>PGM</td>
<td>100 bp</td>
<td>Bioruptor / Ion Fragment Library</td>
<td>Sakai</td>
<td>July 2011</td>
<td>0.3520</td>
<td>0.3878</td>
<td>0.0929</td>
<td>0.1024</td>
</tr>
<tr>
<td>PGM</td>
<td>200 bp</td>
<td>Ion Xpress Plus Fragment</td>
<td>Sakai</td>
<td>July 2012</td>
<td>0.3955</td>
<td>0.6811</td>
<td>0.0303</td>
<td>0.0521</td>
</tr>
<tr>
<td>PGM</td>
<td>300 bp</td>
<td>Ion Xpress Plus Fragment</td>
<td>Sakai</td>
<td>August 2012</td>
<td>0.7054</td>
<td>1.4457</td>
<td>0.0861</td>
<td>0.1765</td>
</tr>
<tr>
<td>PGM</td>
<td>400 bp</td>
<td>Ion Xpress Plus Fragment</td>
<td>Sakai</td>
<td>November 2012</td>
<td>0.6722</td>
<td>1.8726</td>
<td>0.0790</td>
<td>0.2202</td>
</tr>
</tbody>
</table>
Next Generation Sequencing Analysis

- Sequencing generates a FASTQ file for each sample
  - File contains the individual read sequence data for every read sequenced
- Sequences are then aligned to the reference genome of species
  - Produces a BAM file that contains all the alignment information
  - Can be visualized with programs such as IGV
- Variant calling
  - Produces VCF (Variant Calling Format) file that contains the chromosomal position of identified variants, reference allele(s), and the variant allele(s)
Next Generation Sequencing Analysis

- Multiple open-source programs developed for analysis
  - Genome Analysis ToolKit (GATK) from Broad and VarScan (SNV/InDels)
  - deFuse and SoapFuse (Gene Fusion)
  - PinDel (InDels)
  - MACS (ChIP-seq)
  - BS Seeker and Bismark (Methyl-Seq)
  - Galaxy
    (easy to use tool that integrates commonly used tools in a user friendly manner- does not require bioinformatics capabilities)
Galaxy:
A beginners tool for NGS analysis

• https://galaxyproject.org/
Visualization of NGS data by IGV: SNV Variant Example

MD-294, DDX3X, c.1583G>A, p.R528H, chrX:41205843
Visualization of NGS data by IGV: InDel Variant Examples

MD-208, CTDNEP1, c.635_636delCA, p.T212fs, chr17:7147908
Visualization of NGS data by IGV: Splice Site Variant Example
Next Generation Sequencing Analysis

- Multiple challenges for analysis
  - High computational requirement
  - Nomination of multiple variants (both SNPs and somatic/germline single nucleotide variants) often requires extensive filtering methodologies
  - Tumors often require a matched or pooled normal comparison to remove uninformative germline SNPs
  - Require expert contextual interpretation in accordance with clinical information
  - Current standard requires orthogonal validation by a secondary method
Multiple types of applications

- Whole genome (DNA) or whole transcriptome sequencing (RNA)
  - Most comprehensive
  - Identification of variants, translocations,
- Whole exome (DNA/RNA) sequencing
  - Coding regions only
- ChIP-Seq (DNA)
  - Identifies locations in genome for transcription factor binding
- Methyl-Seq (Bi-sulfite sequencing) (DNA)
  - Discover methylation patterns across genome
- Targeted NGS panels (DNA/RNA)
  - Sequencing limited to subset of genes

In 2015, 6844 publications mentioned “Next Generation Sequencing”
Applications of NGS: Tumor Evolution

Genomic architecture of renal cell carcinoma: sequencing

Marco Gerlinger, Stuart Horswell, Varela, Rosalie Fisher, Nicholas Martinez, Benjamin Phillimore, Sakshi Gulati, Paul A Bates, Gordon Steven Hazell, P Andrew Futreal,
Applications of NGS: De Novo Germline Mutations

De Novo Mutations in Essential for Brain Development and Intellectual-Disability

Jung-Hyun Kim, Deepali N. Shir, Michael E. Belmonte, Gregory R. Wilson, Dan T. Peterson, Joshua K. Stone, Eu Stumpel, Jos M. Draaisma, Joost Helger G. Yntema, Kristin Lindstro
Applications of NGS: Pathogen Detection

KlebSeq: A Diagnostic Detection, and Monitor

Jolene R. Bowersa#, Darrin Lemme
Elizabeth M. Driebea, Bette Wojac
and Paul Keima,b

Figure 4. Maximum parsimony tree with 100 bootstraps of the SNPs among 548 K. pneumoniae genomes. Major clonal groups are colored, and locations of canonical SNPs for strain identification assays are marked with stars. All branches labeled with canonical SNPs had >99% bootstrap support, except on the three branches indicated.
DNA sequencing of maternal plasma reliably identifies trisomy 18 and trisomy 13 as well as Down syndrome: an international collaborative study

Glenn E. Palomaki PhD, Cosmin Deciu MS, Edward M. Kloza MS, Geralyn M. Lambert-Messerlian PhD, James E. Haddow MD, Louis M. Neveux BA, Mathias Ehrich MD, Dirk van den Boom PhD, Allan T. Bombard MD, MBA, Wayne W. Grody MD, PhD, Stanley F. Nelson MD & Jacob A. Canick PhD
Applications of NGS: Identification of causal familial disease genes

A. Family 12946 - TFAP2β R285Q
- PDA surgically closed
- heart-related death
- heart problems
- normal ECHO toe syndactyly
- P2946 PDA surgically closed
- Closure via cath
- murmured detected normal ECHO
- toe syndactyly

B. Family 10861 - TBX5 Y407X
- ASD Pacemaker
- ASD/VSD Pacemaker
- ASD/VSD

C. Family 10006 - TBX5 D166S
- Died age 27 yrs
- heart-related
- ? holes in heart
- no contact
- Died during CHD surgery
- perimembranous VSD (small)
- ASD (surgery) perimembranous VSD (spontaneous closure)
- ASD muscular VSD (spontaneous closure)
- ASD + muscular VSD (surgically closed)
- ? ASD closure

D. Family 12637 - ELN (c.950-3C>G)
- SVAS surgery
- SVAS surgery age 12 yrs
- ? normal abnormality on cardiac review

E. Family 11756 - NOTCH1 G200R
- TOF multiple surgeries
- 01720 PTA / VSD
- PA / VSD / MAPCAS
- ALCAPA
NGS Core Services at BCM

• Genomic and RNA Profiling Core
  • Illumina MiSeq, HiSeq 2000, and HiSeq 2500 Systems
  • Service provided:
    • DNA Sequencing
    • DNA-Protein Interaction Analysis (ChIP-Seq)
    • FFPE Transcriptome Analysis
    • Gene Expression Analysis
    • Low-Quality RNA-Seq Analysis
    • Microbial Sequencing Analysis
    • Next-Generation Sequencing Data Services*
    • Sequencing-Based Methylation Analysis
    • Sequencing-Based Transcriptome Analysis
    • Small RNA Discovery
    • Targeted (Amplicon) Resequencing Analysis
    • Targeted RNA Expression Analysis

• For more information:
  https://www.bcm.edu/research/advanced-technology-core-labs/lab-listing/genomic-and-rna-profiling-core/services/next-generation-sequencing-services
“Choose your own adventure”
clinical scenario

NGS identified 3 variants in your brain tumor patient:

• TP53 mutation
  • Indicator of poor prognosis; may decrease chemosensitivity to temozolomide
  • Temozolomide is common front-line therapy for GBM

• H3F3A K27 mutation
  • Indicator of poor prognosis

• A novel NTRK3 fusion
  • Targetable fusion; may qualify for clinical trial- NCT02576431

Genetic findings may indicate a relatively poorer prognostic outcome and therefore, may indicate a more aggressive treatment regimen, especially since data indicates a reduced sensitivity to Temozolomide. The presence of the NTRK3 fusion may indicate the potential use of investigation NTRK3 inhibitors currently in clinical trials.
Lab Exercises
FASTQ generation

Illumina

bcl2fastq2 Conversion Software Guide

Version 2.17 for MiSeq®, HiSeq®, NextSeq®, and HiSeq® X Systems
For Research Use Only. Not for use in diagnostic procedures.

Ion Torrent

FileExporter (v4.4.0.0) [528]

Create SFF? False
Create FASTQ? True
Moved to compressed file? False
Include variant caller files? True
Move TVC files to compressed file: BAM/BAI: False
VCF: False
XLS: False

DELIMITER: "_"
SELECTIONS:
2016-08-16
SAMPLEID@
BARINFO@

Output Files:
2016-08-16_C1032D_IonXpress_011.bam
2016-08-16_C1032D_IonXpress_011.bam.bai
2016-08-16_C1032D_IonXpress_011.fastq
2016-08-16_C1032D_IonXpress_011.vcf
2016-08-16_C1032D_IonXpress_011_alleles.xls
2016-08-16_C1032D_IonXpress_011_variants.xls
### Flowcell Summary

<table>
<thead>
<tr>
<th>Clusters (Raw)</th>
<th>Clusters(PF)</th>
<th>Yield (MBases)</th>
</tr>
</thead>
<tbody>
<tr>
<td>10,902,614</td>
<td>9,851,872</td>
<td>2,956</td>
</tr>
</tbody>
</table>

### Lane Summary

<table>
<thead>
<tr>
<th>Lane</th>
<th>Project</th>
<th>Sample</th>
<th>Barcode sequence</th>
<th>PF Clusters</th>
<th>% of the lane</th>
<th>% Perfect barcode</th>
<th>% One mismatch barcode</th>
<th>Yield (Mbases)</th>
<th>% PF Clusters</th>
<th>% &gt;= Q30 bases</th>
<th>Mean Quality Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>T1005R-ASF</td>
<td>T1005R-ASF</td>
<td>TAAGCGCA+GCGTAAC</td>
<td>898,940</td>
<td>9.12</td>
<td>100.00</td>
<td>NaN</td>
<td>270</td>
<td>95.12</td>
<td>84.31</td>
<td>34.62</td>
</tr>
<tr>
<td>1</td>
<td>T1006R-ASF</td>
<td>T1006R-ASF</td>
<td>CGTACTAG+GTGATCCA</td>
<td>947,472</td>
<td>9.62</td>
<td>100.00</td>
<td>NaN</td>
<td>284</td>
<td>96.64</td>
<td>86.04</td>
<td>35.05</td>
</tr>
<tr>
<td>1</td>
<td>T1007R-ASF</td>
<td>T1007R-ASF</td>
<td>AGGCGAGA+GTGCAGTT</td>
<td>774,376</td>
<td>7.86</td>
<td>100.00</td>
<td>NaN</td>
<td>232</td>
<td>95.25</td>
<td>83.78</td>
<td>34.51</td>
</tr>
<tr>
<td>1</td>
<td>T1008R-ASF</td>
<td>T1008R-ASF</td>
<td>TCCTGAGC+GACCGTAA</td>
<td>701,214</td>
<td>7.12</td>
<td>100.00</td>
<td>NaN</td>
<td>210</td>
<td>96.72</td>
<td>88.38</td>
<td>35.56</td>
</tr>
<tr>
<td>1</td>
<td>T1009R-ASF</td>
<td>T1009R-ASF</td>
<td>GGACTCTC+TGATCGTC</td>
<td>818,491</td>
<td>8.31</td>
<td>100.00</td>
<td>NaN</td>
<td>246</td>
<td>95.37</td>
<td>86.92</td>
<td>35.25</td>
</tr>
<tr>
<td>1</td>
<td>T1010R-ASF</td>
<td>T1010R-ASF</td>
<td>TAGGCATG+GTCCAAGG</td>
<td>896,040</td>
<td>9.10</td>
<td>100.00</td>
<td>NaN</td>
<td>269</td>
<td>95.65</td>
<td>85.92</td>
<td>35.02</td>
</tr>
<tr>
<td>1</td>
<td>T1011R-ASF</td>
<td>T1011R-ASF</td>
<td>CTCTCTACTGGATTCGTT</td>
<td>1,091,985</td>
<td>11.08</td>
<td>100.00</td>
<td>NaN</td>
<td>328</td>
<td>95.30</td>
<td>88.13</td>
<td>35.51</td>
</tr>
<tr>
<td>1</td>
<td>T1012R-ASF</td>
<td>T1012R-ASF</td>
<td>CAGAGGAG+AGTATGCC</td>
<td>683,722</td>
<td>6.94</td>
<td>100.00</td>
<td>NaN</td>
<td>205</td>
<td>95.64</td>
<td>85.85</td>
<td>34.99</td>
</tr>
<tr>
<td>1</td>
<td>default</td>
<td>Undetermined</td>
<td>unknown</td>
<td>3,039,632</td>
<td>30.85</td>
<td>100.00</td>
<td>NaN</td>
<td>912</td>
<td>80.34</td>
<td>89.95</td>
<td>35.85</td>
</tr>
</tbody>
</table>
Galaxy

https://usegalaxy.org/

Data intensive biology for everyone.

Galaxy is an open, web-based platform for data intensive biomedical research. Whether on the free public server or your own instance, you can perform, reproduce, and share complete analyses.

Use Galaxy

Use project's free server or other public servers

Get Galaxy

Install locally or in the cloud or get Galaxy on SlipStream

Learn Galaxy

Screencasts, Galaxy 101, ...

Get Involved

Mailing lists, Tool Shed, wiki

Search all resources

The Galaxy Team is a part of the Center for Comparative Genomics and Bioinformatics at Penn State University, and the Department of Biology at Johns Hopkins University. The Galaxy Project is supported in part by NHGRI, NSF, The Huck Institutes of the Life Sciences, The Institute for CyberScience at Penn State, and Johns Hopkins University.
Galaxy: Upload data

Galaxy is an open source, web-based platform for data intensive biomedical research. If you are new to Galaxy start here or consult our help resources. You can install your own Galaxy by following the tutorial and choose from thousands of tools from the Tool Shed.

Phage Galaxy
An instance dedicated to sequencing, analysis, annotation, and comparative genomics of bacteriophage genomes.

Penn State
Johns Hopkins University
TACC
CyVerse

The Galaxy Team is a part of the Center for Comparative Genomics and Bioinformatics at Penn State, and the Department of Biology at Johns Hopkins University.

This instance of Galaxy is utilizing infrastructure generously provided by the CyVerse at the Texas Advanced Computing Center, with support from the National Science Foundation.
Galaxy: Sequencing QC

Galaxy is an open source, web-based platform for data intensive biomedical research. If you are new to Galaxy start here or consult our help resources. You can install your own Galaxy by following the tutorial and choose from thousands of tools from the Tool Shed.

Looking to learn?
New comprehensive tutorials on:
- Diploid variant calling Reference based RNAseq
- Processing multiple samples Introduction to NGS technologies
- Galaxy 181, parts 1 & 2

PennState
Johns Hopkins University
TACC
CyVerse

The Galaxy Team is a part of the Center for Comparative Genomics and Bioinformatics at Penn State, and the Department of Biology and at Johns Hopkins University.

This instance of Galaxy is utilizing infrastructure generously provided by the CyVerse at the Texas Advanced Computing Center, with support from the National Science Foundation.

The Galaxy Project is supported in part by NSF, NHGRI, The Huck Institutes of the Life Sciences, The Institute for CyberScience at Penn State, and Johns Hopkins University.

This is a fork of the main Galaxy repository. Data providers and datasets are not imported; if these are present on the main repository, data are stored...
Galaxy Sequencing QC: FastQC

FQQC aims to provide a simple way to do some quality control checks on raw sequence data coming from high throughput sequencing pipelines. It provides a modular set of analyses which you can use to give a quick impression of whether your data has any problems of which you should be aware before doing any further analysis.

The main functions of FastQC are:

- Import of data from BAM, SAM or Fastq files (any variant)
- Providing a quick overview to tell you in which areas there may be problems
- Summary graphs and tables to quickly assess your data
- Export of results to an HTML based permanent report
- Offline operation to allow automated generation of reports without running the interactive application

FastQC
This is a Galaxy wrapper. It merely exposes the external package FastQC which is documented at FastQC. Kindly acknowledge it as well as this tool if you use it. FastQC incorporates the Bioconda tools libraries for sam/bam processing.

The contaminants file parameter was borrowed from the independently developed fastqcwrapper contributed to the Galaxy Community Tool Shed by J. Johnson. Adaptation to version 0.11.2 by T. McGowan.

Inputs and outputs
FastQC is the best place to look for documentation - it's very good. A summary follows below for those in a tearing hurry.

This wrapper will accept a Galaxy fastq, sam or bam as the input read file to check. It will also take an optional file containing a list of contaminants information, in the form of a tab-delimited file with 2 columns, name and sequence. As another option the tool takes a custom limits.txt file that allows setting the warning thresholds for the different modules and also specifies which modules to include in the output.

The tool produces a basic text and a HTML output file that contain all of the results, including the following:

- Basic Statistics
- Per base sequence quality
- Per sequence quality scores
- Per base sequence content
- Per base GC content
Galaxy: Mapping/Alignment

Galaxy is an open source, web-based platform for data intensive biomedical research. If you are new to Galaxy, start here or consult our help resources. You can install your own Galaxy by following the tutorial and choose from thousands of tools from the Tool Shed.

Galaxy project

Running Your Own
Understanding how Galaxy works
An in-depth tutorial

PENNSTATE

JOHNS HOPKINS UNIVERSITY

TACC

CYVERSE

The Galaxy Team is a part of the Center for Comparative Genomics and Bioinformatics at Penn State, and the Department of Bioloay at Johns Hopkins University.

This instance of Galaxy is utilizing infrastructure generously provided by the CyVerse at the Texas Advanced Computing Center, with support from the National Science Foundation.

The Galaxy Project is supported in part by NSF, DOE, The Huck Institutes of the Life Sciences, The Institute for CyberScience at Penn State, and Johns Hopkins University.

This is a free, open source software tool designed for data analysis, but not yet supported. If you have questions or feedback, please feel free to contact us.
Galaxy Mapping: BWA-MEM
Galaxy Variant Analysis

Galaxy is an open source, web-based platform for data-intensive biomedical research. If you are new to Galaxy, start here or consult our help resources. You can install your own Galaxy by following the tutorial and choose from thousands of tools from the Tool Shed.

Try Galaxy on the Cloud

Now you can have a personal Galaxy within the infinite Universe.

Tweets by @galaxyproject

Galaxy Project Retweeted
Nate Coraor @natecoro
Thanks to all the awesome people who made the @jetstream_cloud launch today possible! #usegalaxy.org runs some jobs here via @XSEDEscience

Galaxy Project Retweeted
Nate Coraor @natecoro
#usegalaxy.org now stages @jetstream_cloud jobs through dedicated nginx/uWSGI/VMs,

Penn State

Johns Hopkins University

TACC

CyVerse

The Galaxy Team is a part of the Center for Comparative Genomics and Bioinformatics at Penn State, and the Department of Biology and at Johns Hopkins University.

This instance of Galaxy is utilizing infrastructure generously provided by the CyVerse at the Texas Advanced Computing Center, with support from the National Science Foundation.
Galaxy Variant Analysis: Varscan

**VarScan** for variant detection (Galaxy Version 0.1)

- **Pileup dataset**: No pileup dataset available.
- **Analysis type**: single nucleotide variation
- **Minimum read depth**: 8
- **Minimum supporting reads**: 2
- **Minimum base quality at a position to count a read**: 15
- **Minimum variant allele frequency threshold**: 0.01
- **Minimum frequency to call homozygote**: 0.75
- **p-value threshold for calling variants**: 0.99
- **Ignore variants with >90% support on one strand**: No
- **sample_names**: Separate sample names by comma; leave blank to use default sample names.

**VarScan Overview**

VarScan performs variant detection for massively parallel sequencing data, such as exome, WGS, and transcriptome data. It calls variants from a pileup dataset and produces a VCF 4.1. Full documentation is available [online](#).

**Input**

- pileup file: The SAMtools pileup file
Galaxy Variant Analysis: Variant Annotation (ANNOVAR/SnpEff)
Visualizing Data: IGV

http://software.broadinstitute.org/software/igv/
Visualizing Data: IGV

Click hyperlink to download BAM file:

Click hyperlink to download BAI file:

Download IGV:
http://software.broadinstitute.org/software/igv/download

Navigate to chr20:18,469,759-18,475,788
View data
Getting Acquainted with VCFs

Download a VCF:

ftp://gsapubftp-anonymous@ftp.broadinstitute.org/bundle/hg38/hg38bundle/1000G_omni2.5.hg38.vcf.gz

Unzip the file
### Ion Torrent/NextGENe VCF Example

<table>
<thead>
<tr>
<th>CHROM</th>
<th>POS</th>
<th>ID</th>
<th>REF</th>
<th>ALT</th>
<th>QUAL</th>
<th>FILTER</th>
<th>INFO</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>100</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>200</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>300</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**FIELD Translator**:
- **#CHROM**: Chromosome number
- **POS**: Position
- **ID**: Variant ID
- **REF**: Reference allele
- **ALT**: Alternate allele
- **QUAL**: Quality score
- **FILTER**: Filter flag
- **INFO**: Additional information fields

**Example Records**:
- **CHROM**: 1
  - **POS**: 100
  - **ID**: None
  - **REF**: A
  - **ALT**: T
  - **QUAL**: 20
  - **FILTER**: PASS
  - **INFO**: None
- **CHROM**: 2
  - **POS**: 200
  - **ID**: None
  - **REF**: T
  - **ALT**: C
  - **QUAL**: 30
  - **FILTER**: PASS
  - **INFO**: None

---

**FORMAT**:
- **ID**: 5
  - **Number**: 1
  - **Type**: Integer
  - **Description**: NextGENe Read Count on Forward Strand
- **ID**: 6
  - **Number**: 1
  - **Type**: Integer
  - **Description**: NextGENe Read Count on Reverse Strand
- **ID**: 7
  - **Number**: 1
  - **Type**: Integer
  - **Description**: NextGENe Read Count on Positive Strand
- **ID**: 8
  - **Number**: 1
  - **Type**: Integer
  - **Description**: NextGENe Read Count on Negative Strand
- **ID**: 9
  - **Number**: 1
  - **Type**: Integer
  - **Description**: NextGENe Read Count on Forward and Reverse Strand