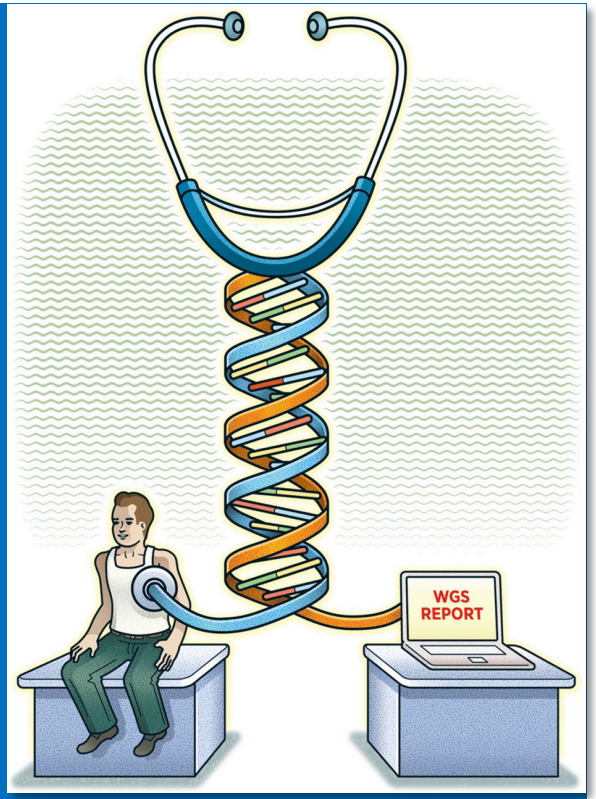
Customizing Galaxy for a Hospital Environment



from- Drmanac Science 6/1/12

GCC2012 Lightening Talks

Larry Helseth, PhD
Bioinformatician, Center for Molecular Medicine
NorthShore University HealthSystem
Clinical Researcher, University of Chicago
Evanston, Illinois
27 July 2012

NorthShore
University HealthSystem

Installation challenges:

http://gorancson.wordpress.com/

- Everything is locked down for HIPAA

NorthShore University HealthSystem

Other Challenges

- Corporate browser standard = IE8
  - Galaxy status won't auto refresh and can't download result files
  - Can't visualize alignment using Trackster
    - Visualize using IGV/IGB
- Intranet access only
- Built Galaxy without root access

NorthShore University HealthSystem

Advantages

- Human only (+/- mm10)
- No storage or backup concerns (co-located with EMR system which has existing **petabyte** storage & off-site backup)
- Corporate IT already provides 7 X 24 support
- Fast intranet (no need for sneaker-net)
- One of the best EMR installs in US w/ several dozen EPIC workflow developers

NorthShore University HealthSystem

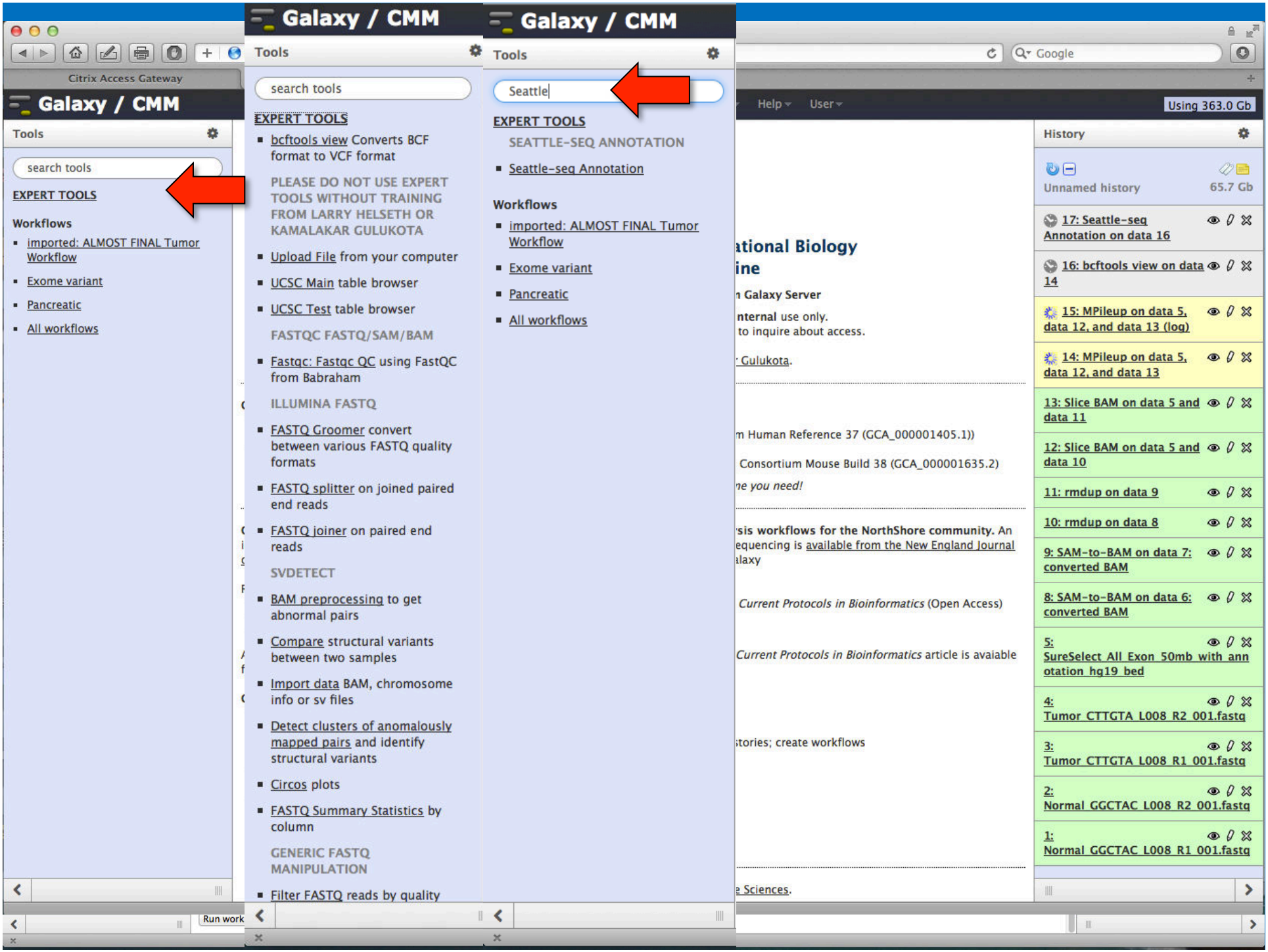
Current Environment

- Current environment (cloned from Galaxy-Dist on 5/30/12):
  - **24 core (48 virtual)**
  - 32 GB RAM
  - Currently 5 TB storage (expandable on demand)
  - Hacked BWA wrapper to use 20 cores
  - Apache load balancing with web0 ,web1 ,web2 ,web3, handler0, handler1 & manager per Galaxy Admin docs
  - Modified Linux tmp environment. Use galaxy_env
- Analyzed three pairs of exome sequences simultaneously (6 FASTQ files -> VCF -> SeattleSeq annotation) in ~8 hours
  - Time sensitive information. Don't just want genomic autopsy

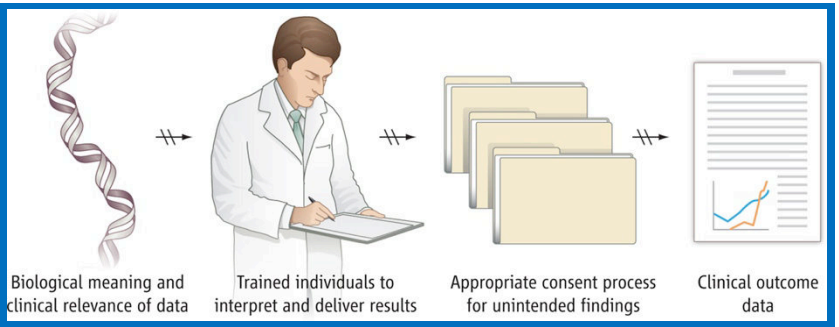
NorthShore University HealthSystem

Customize Interface using Workflows

- Simplify Galaxy UI for clinicians by stripping out menu options
  - Consolidate tools under “Expert Tools”
- Clinicians rely on workflows curated by bioinformatician
  - Tumor-Normal exome pairs
  - RNAseq, etc.
- Manage data using Galaxy Data Libraries so users don’t have to upload or FTP


Galaxy / CMM
Tools
search tools
EXPERT TOOLS
Workflows
imported: ALMOST FINAL Tumor Workflow
Exome variant
Pancreatic
All workflows
Galaxy / CMM
Tools
search tools
EXPERT TOOLS
bcftools view Converts BCF format to VCF format
PLEASE DO NOT USE EXPERT TOOLS WITHOUT TRAINING FROM LARRY HELSETH OR KAMALAKAR GULUKOTA
Upload File from your computer
UCSC Main table browser
UCSC Test table browser
FASTQC FASTQ/SAM/BAM
Fastqc: Fastqc QC using FastQC from Babraham
ILLUMINA FASTQ
FASTQ Groomer convert between various FASTQ quality formats
FASTQ splitter on joined paired end reads
FASTQ joiner on paired end reads
SVDETECT
BAM preprocessing to get abnormal pairs
Compare structural variants between two samples
Import data BAM, chromosome info or sv files
Detect clusters of anomalously mapped pairs and identify structural variants
Circos plots
FASTQ Summary Statistics by column
GENERIC FASTQ MANIPULATION
Filter FASTQ reads by quality
Galaxy / CMM
Tools
Seattle
EXPERT TOOLS
SEATTLE-SEQ ANNOTATION
Seattle-seq Annotation
Workflows
imported: ALMOST FINAL Tumor Workflow
Exome variant
Pancreatic
All workflows
Google
Help
User
Using 363.0 Gb
History
Unnamed history
65.7 Gb
17: Seattle-seq Annotation on data 16
16: bcftools view on data 14
15: MPileup on data 5, data 12, and data 13 (log)
14: MPileup on data 5, data 12, and data 13
13: Slice BAM on data 5 and data 11
12: Slice BAM on data 5 and data 10
11: rmdup on data 9
10: rmdup on data 8
9: SAM-to-BAM on data 7: converted BAM
8: SAM-to-BAM on data 6: converted BAM
5: SureSelect All Exon 50mb with annotation hg19 bed
4: Tumor CTTGTA L008 R2 001.fastq
3: Tumor CTTGTA L008 R1 001.fastq
2: Normal GGCTAC L008 R2 001.fastq
1: Normal GGCTAC L008 R1 001.fastq
Run work

Next Steps:



Brunham & Hayden, Science 6/1/12

- Automate local annotation using Annovar → Excel
  http://www.openbioinformatics.org/annovar/annovar_accessary.html#excel
- Annotation, annotation and annotation.
  - » Evaluating Seattleseq code provided by Martijn Vermaat, Leiden University Medical Center, Belgium and Annovar code provided by Dr. Yasukazu Nakamura, DDBJ Center, National Institute of Genetics, Japan
  - » SomaticSnipper →SeattleSeq or Annovar (still manual)
  - » Visualization with Circos (Bruno Zeitouni, Inst. Curie)
  - » Integration with VAT (Gerstein lab)?
- Implement Galaxy sample tracking and automated workflow

NorthShore University HealthSystem

Next Steps (cont):

- Currently ”**Research Use Only**". Planning to lock down version of server and VALIDATE before full clinical use.
- Interface annotation results with EMR system possibly using Illumina-like visualization
- Scale current environment as needed
- User training

➢ **SUGGESTIONS/FEEDBACK TO lhelseth@gmail.com**

NorthShore University HealthSystem

NorthShore
University HealthSystem

How'd You Do That?

Last Saved: 7/22/12 9:18:32 PM
File Path ▾ : ~/Desktop/tool_conf.sample.xml

tool_conf.sample.xml | (no symbol selected)

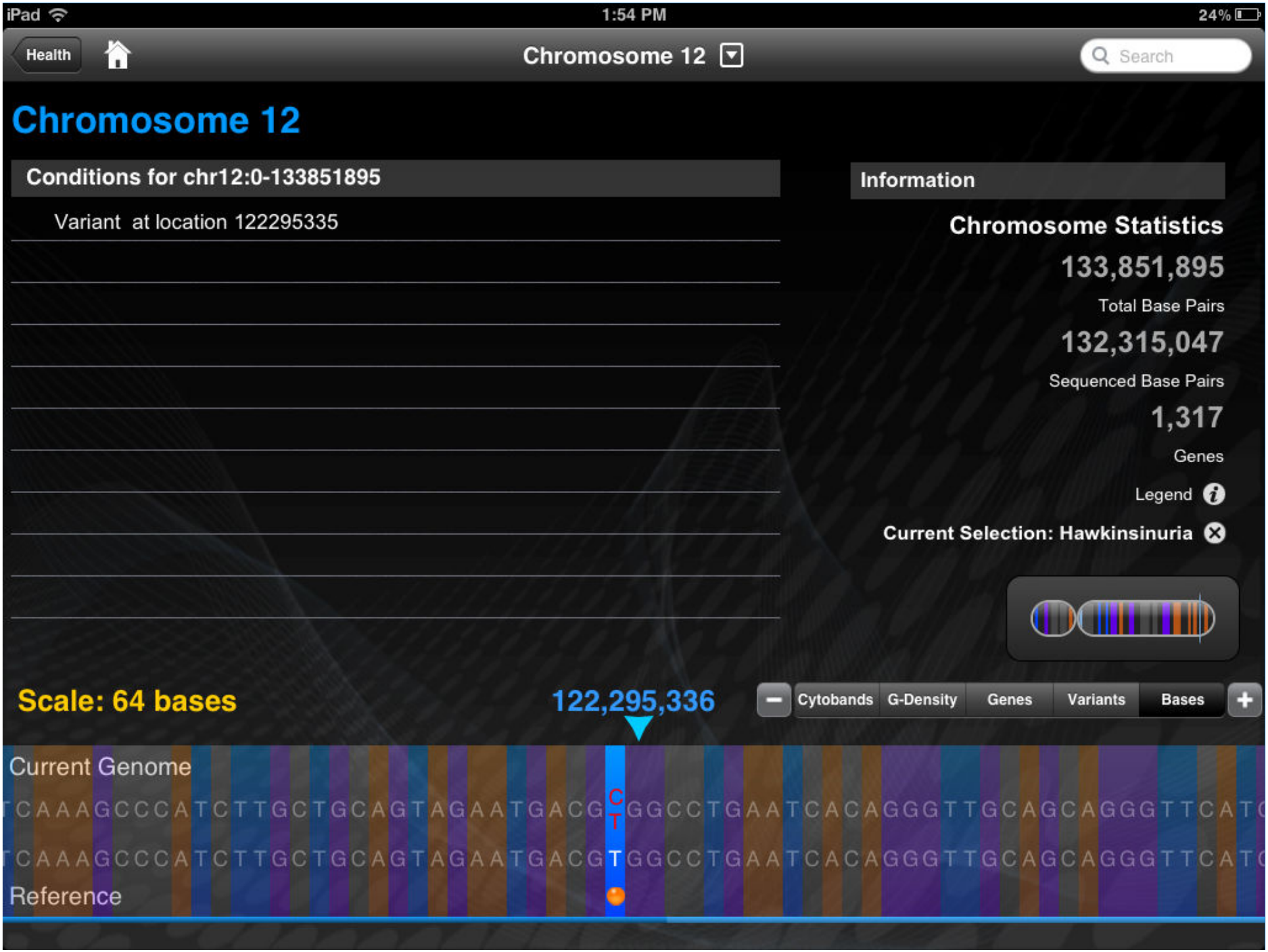
```
<toolbox>
  <section name="Get Data" id="getext">
    <tool file="data_source/upload.xml"/>
    <tool file="data_source/ucsc_tablebrowser.xml" />
    <tool file="data_source/ucsc_tablebrowser_test.xml" />
    <tool file="data_source/ucsc_tablebrowser_archaea.xml" />
    <tool file="data_source/bx_browser.xml" />
    <tool file="data_source/ebi_sra.xml"/>
    <tool file="data_source/microbial_import.xml" />
    <tool file="data_source/biomart.xml" />
    <tool file="data_source/biomart_test.xml" />
    <tool file="data_source/cbi_rice_mart.xml" />
    <tool file="data_source/gramene_mart.xml" />
    <tool file="data_source/fly_modencode.xml" />
    <tool file="data_source/flymine.xml" />
    <tool file="data_source/flymine_test.xml" />
    <tool file="data_source/modmine.xml" />
    <tool file="data_source/ratmine.xml" />
    <tool file="data_source/yeastmine.xml" />
    <tool file="data_source/metabolicmine.xml" />
    <tool file="data_source/worm_modencode.xml" />
    <tool file="data_source/wormbase.xml" />
    <tool file="data_source/wormbase_test.xml" />
    <tool file="data_source/eupathdb.xml" />
    <tool file="data_source/encode_db.xml" />
    <tool file="data_source/epigraph_import.xml" />
    <tool file="data_source/epigraph_import_test.xml" />
    <tool file="data_source/hbvar.xml" />
    <tool file="genomespace/genomespace_file_browser_prod.xml" />
    <tool file="genomespace/genomespace_importer.xml" />
    <tool file="validation/fix_errors.xml" />
  </section>
  <section name="Send Data" id="send">
    <tool file="data_destination/epigraph.xml" />
    <tool file="data_destination/epigraph_test.xml" />
    <tool file="genomespace/genomespace_exporter.xml" />
  </section>
  <section name="ENCODE Tools" id="EncodeTools">
    <tool file="encode/gencode_partition.xml" />
    <tool file="encode/random_intervals.xml" />
  </section>
  <section name="Lift-Over" id="liftOver">
    <tool file="extract/liftOver_wrapper.xml" />
  </section>
  <section name="Text Manipulation" id="textutil">
    <tool file="filters/fixedValueColumn.xml" />
    <tool file="stats/column_maker.xml" />
    <tool file="filters/catWrapper.xml" />
    <tool file="filters/cutWrapper.xml" />
    <tool file="filters/mergeCols.xml" />
    <tool file="filters/convert_characters.xml" />
    <tool file="filters/CreateInterval.xml" />
    <tool file="filters/cutWrapper.xml" />
    <tool file="filters/changeCase.xml" />
```

tool_conf.xml

Last Saved: 7/22/12 9:04:03 PM
File Path ▾ : ~/Desktop/tool_conf.xml

tool_conf.xml | (no symbol selected)

```
toolbox>
 <section name="EXPERT TOOLS" id="getext">
   <label text="Please do not use expert tools without training from Larry Helseth or Kamalakar Gulukota" id="larry" />
     <tool file="data_source/upload.xml"/>
     <tool file="data_source/ucsc_tablebrowser.xml" />
     <tool file="data_source/ucsc_tablebrowser_test.xml" />
   <label text="FastQC fastq/sam/bam" id="fastqcsambam" />
    <tool file="rgenetics/rgFastQC.xml" />
   <label text="Illumina fastq" id="illumina" />
      <tool file="fastq/fastq_groomer.xml" />
      <tool file="fastq/fastq_paired_end_splitter.xml" />
      <tool file="fastq/fastq_paired_end_joiner.xml" />
      <tool file="fastq/fastq_stats.xml" />
   <label text="SVDetect" id="svdetect" />
      <tool file="svdetect/BAM_preprocessingPairs.xml" />
      <tool file="svdetect/SVDetect_compare.xml" />
      <tool file="svdetect/SVDetect_import.xml" />
      <tool file="svdetect/SVDetect_run_parallel.xml" />
      <tool file="svdetect/circos_graph.xml" />
   <label text="Generic FASTQ manipulation" id="generic_fastq" />
      <tool file="fastq/fastq_filter.xml" />
      <tool file="fastq/fastq_trimmer.xml" />
      <tool file="fastq/fastq_trimmer_by_quality.xml" />
      <tool file="fastq/fastq_masker_by_quality.xml" />
      <tool file="fastq/fastq_paired_end_interlacer.xml" />
      <tool file="fastq/fastq_paired_end_deinterlacer.xml" />
      <tool file="fastq/fastq_manipulation.xml" />
      <tool file="fastq/fastq_to_fasta.xml" />
      <tool file="fastq/fastq_to_tabular.xml" />
      <tool file="fastq/tabular_to_fastq.xml" />
   <label text="FASTX-Toolkit for FASTQ data" id="fastx_toolkit" />
      <tool file="fastx_toolkit/fastq_quality_converter.xml" />
      <tool file="fastx_toolkit/fastx_quality_statistics.xml" />
      <tool file="fastx_toolkit/fastq_quality_boxplot.xml" />
      <tool file="fastx_toolkit/fastx_nucleotides_distribution.xml" />
      <tool file="fastx_toolkit/fastq_to_fasta.xml" />
      <tool file="fastx_toolkit/fastq_quality_filter.xml" />
      <tool file="fastx_toolkit/fastq_to_fasta.xml" />
      <tool file="fastx_toolkit/fastx_artifacts_filter.xml" />
      <tool file="fastx_toolkit/fastx_barcode_splitter.xml" />
      <tool file="fastx_toolkit/fastx_clipper.xml" />
      <tool file="fastx_toolkit/fastx_collapser.xml" />
      <tool file="fastx_toolkit/fastx_renamer.xml" />
      <tool file="fastx_toolkit/fastx_reverse_complement.xml" />
      <tool file="fastx_toolkit/fastx_trimmer.xml" />
   <label text="NGS Picard-Conversion" id="picard_beta" />
      <tool file="picard/picard_FastqToSam.xml" />
      <tool file="picard/picard_SamToFastq.xml" />
   <label text="NGS Picard-QC/Metrics for sam/bam" id="qcsambam"/>
      <tool file="picard/picard_BamIndexStats.xml" />
      <tool file="picard/rgPicardASMetrics.xml" />
      <tool file="picard/rgPicardGCBiasMetrics.xml" />
      <tool file="picard/rgPicardLibComplexity.xml" />
      <tool file="picard/rgPicardInsertSize.xml" />
      <tool file="picard/rgPicardHsMetrics.xml" />
   <label text="NGS Picard-bam/sam Cleaning" id="picard-clean" />
      <tool file="picard/picard_AddOrReplaceReadGroups.xml" />
      <tool file="picard/picard_ReorderSam.xml" />
      <tool file="picard/picard_ReplaceSamHeader.xml" />
      <tool file="picard/rgPicardFixMate.xml" />
```


iPad
1:54 PM
24%
Health
Chromosome 12
Search
Chromosome 12
Conditions for chr12:0-133851895
Variant at location 122295335
Information
Chromosome Statistics
133,851,895
Total Base Pairs
132,315,047
Sequenced Base Pairs
1,317
Genes
Legend
Current Selection: Hawkinsinuria
Scale: 64 bases
122,295,336
Cytobands
G-Density
Genes
Variants
Bases
Current Genome
TCAAAGCCCATCTTGCTGCAGTAGAATGACGCTGGCCTGAATCACAGGGTTGCAGCAGGGTTCATC
TCAAAGCCCATCTTGCTGCAGTAGAATGACGTGGCCTGAATCACAGGGTTGCAGCAGGGTTCATC
Reference

BWA Hack

- **~/galaxy-dist/tools/sr_mapping/bwa_wrapper.xml**

```
<tool id="bwa_wrapper" name="Map with BWA for Illumina"
    version="1.2.3">
    <description></description>
    <parallelism method="basic"></parallelism>
    <command interpreter="python">    bwa_wrapper.py      --
    threads="20" ## DLH hacked from 4 to 20 on 12 June to
    improve throughput
    #if $input1.ext == "fastqillumina":          --illumina1.3     #end if
    .
    .
    .
```

NorthShore University HealthSystem

How Bad Can It Be?!?

- It took 2 hours to download hg19.2bit from UCSF (~778 MB) through our firewall to my server using HTTP
- It took 20 minutes to download the same file to my laptop on my kitchen WiFi using HTTP
- It took 70 seconds to transfer hg19.2bit from my server to my desktop via HTTP