

A Practical Guide to Choosing the Right SNP Detection Tool

--Part 2: Practical Usage of the SNP Detection Tools

Ming Yi, Ph.D.

Advanced Biomedical Computing Center (ABCC)

Bioinformatics Training and Education Program (BTEP)

Center for Cancer Research

February 26th, 2013

Introduction to SNP discovery tools used for Next Generation Sequencing data

- **Comparison Study of NGS SNP Detection Tools**
 - ❖ Brief background and introduction for the current status of SNP detection field and each of the selected tools to be compared
 - ❖ Description of our benchmark exome-seq data with pedigree info and SNP array data from matched-samples and why they are useful for comparison of these tools for SNP call quality
 - ❖ Comparison and validation results of these tools using the benchmark data
 - ❖ Conclusion and take-home message
 - ❖ Q & A session

- **Detailed Illustration of the Practical Usage of Each SNP Detection Tool**
 - ❖ Brief introduction of practical aspects of the tools (e.g., download, installation, interface, running environment, basic system requirement etc)
 - ❖ Practical command lines for command-driven tool(s), parameter options, wrapper script examples for the command-driven tools, interface for commercial tools
 - ❖ Brief discussion of result files and some diagnosis plots, etc.
 - ❖ Q & A session

NGS-based SNP Discovery Tools

- Atlas-SNP2 (Baylor). *Genome Res.* 2010,20(2):273-80
- SOAPsnp (BGI). *Bioinformatics* 2008, 24(5):713-4
- Crossbow (UM). *Nature Biotech* 2010, 28:691-693
- Bambino (NCI, Beutow). *Bioinformatics* 2011,5;27(6):865-6
- GigaBayes→FreeBayes (Boston College). *Nature Method* 2008, 5(2):183-8
- CLCbio Genomics Workbench (Commercial)
- Genomatix Mining Station (GMS) (Commercial)
- Partek SNP tool in Genomics Suite (Commercial)
- Avadis NGS (Commercial)
- Illumina Casava (Commercial)
- SAMtools (Sanger Institute). *Bioinformatics* 2009, 25:2078-9
- VarScan (Washington Univ). *Bioinformatics* 2009; *Genome Res* 2012
- GATK (Broad Institute). *Genome Res* 2010; *Nature Genet* 2011
-

ABCC-Hosted SNP Discovery Tools

Browse Applications - Mozilla Firefox

tools-abcc.ncicrf.gov/apps/resources/browseCategories

AppDB

Online documentation and access details to application

Application Categories

- Alignment Tools
- Gene prediction and Primer design
- In-house applications
- Licensed Softwares
- Linkage and Phylogenetic Analysis
- Mathematics and Statistics
- Molecular modeling
- Next Generation Sequencing
 - 454
 - ABySS
 - ALLPATHS-LG
 - AMOS
 - ANNOVAR
 - APT
 - BAMtools
 - BEAGLE
 - BEDtools
 - BFAST
 - biotoolbox
 - Blat Run Online
 - Bowtie
 - Bowtie2
 - BSMAP
 - BWA
 - CEAS
 - cgatools
 - CisGenome
 - cnv-seq
 - CNVnator
 - Consed
 - Cufflinks

View Record - Mozilla Firefox

tools-abcc.ncicrf.gov/apps/resources/record/1282

GATK At ABCC

Name	GATK
Current Version	1.6-7-g2be5704
Old Version(s)	1.2-26-g43b0c98 (09/28/2011 - 05/28/2012)
	1.1-33-gcf24303
	1.1-23-g8072bd9
	1.0-6148-g7688bda
	1.0.5974
	1.0.5777
	1.0.5336
	1.0.4418
	1.0.4168
Category	NGS, Next generation sequencing
Author(s)/Vendor(s)	Broad Institute.
Online Documentation	http://www.broadinstitute.org/gsa/wiki/index.php/The_Genome_Analysis_Toolkit
Source Website	http://www.broadinstitute.org/gsa/wiki/index.php/The_Genome_Analysis_Toolkit
Computer platform(s)	Linux
Location	/opt/nasapps/stow/GenomeAnalysisTK-1.6-7-g2be5704/bin
ABCC Contact Person	Jigui Shan
Access	This package was written in JAVA, see user guide to run programs: http://www.broadinstitute.org/gsa/wiki/index.php/The_Genome_Analysis_Toolkit
Program Description	The Genome Analysis Toolkit (GATK) is a structured software library that makes writing efficient analysis tools using next-generation sequencing data very easy, and second it's a suite of tools for working with human medical resequencing projects such as 1000 Genomes and The Cancer Genome Atlas. These tools include things like a depth of coverage analyzers, a quality score recalibrator, a SNP/indel caller and a local realigner.
Created By	Jigui Shan

Tool By Tool Highlighting Major Aspects of Practical Usage

- **GATK**
- Samtools
- VarScan
- CLCBio
- CASAVA
- Partek Genomic Suite

Each tool is keeping evolving on its own schedule. So the session only give snapshot of “current” status of the tools.

GATK: A Variant Discovery Tool from Broad Institute

TECHNICAL REPORTS

nature
genetics

A framework for variation discovery and genotyping using next-generation DNA sequencing data

Mark A DePristo¹, Eric Banks¹, Ryan Poplin¹, Kiran V Garimella¹, Jared R Maguire¹, Christopher Hartl¹, Anthony A Philippakis¹⁻³, Guillermo del Angel¹, Manuel A Rivas^{1,4}, Matt Hanna¹, Aaron McKenna¹, Tim J Fennell¹, Andrew M Kernytsky¹, Andrey Y Sivachenko¹, Kristian Cibulskis¹, Stacey B Gabriel¹, David Altshuler^{1,3,4} & Mark J Daly^{1,3,4}

¹Program in Medical and Population Genetics, Broad Institute of Harvard and MIT, Cambridge, Massachusetts, USA. ²Brigham and Women's Hospital, Boston, Massachusetts, USA. ³Harvard Medical School, Boston, Massachusetts, USA. ⁴Center for Human Genetic Research, Massachusetts General Hospital, Richard B. Simches Research Center, Boston, Massachusetts, USA. Correspondence should be addressed to M.A.D. (depristo@broadinstitute.org).

Received 27 August 2010; accepted 17 March 2011; published online 10 April 2011; doi:10.1038/ng.806



The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data

Aaron McKenna, Matthew Hanna, Eric Banks, et al.

Genome Res. 2010 20: 1297-1303 originally published online July 19, 2010
Access the most recent version at doi:[10.1101/gr.107524.110](https://doi.org/10.1101/gr.107524.110)

Old Webpage-Best Practice v2

Best Practice Variant Detection with the GATK v2

From GSA

Contents

- 1 Introduction
 - 1.1 Lanes, Samples, Cohort
 - 1.2 Testing data: 64x HiSeq on chr20 for NA12878
- 2 Phase I: Raw data processing
 - 2.1 Initial read mapping
 - 2.2 Raw BAM to realigned, recalibrated BAM
 - 2.2.1 Previous recommendation: lane-level recalibration, sample-level realignment
 - 2.2.2 Fast: lane-level realignment at known sites only and lane-level recalibration
 - 2.2.3 Fast + sample-level realignment
 - 2.2.4 Better: sample-level realignment with known indels and recalibration
 - 2.2.5 Best: multi-sample realignment with known sites and recalibration
 - 2.2.6 Misc. notes on the process
- 3 Initial variant discovery and genotyping
 - 3.1 Input BAMs for variant discovery and genotyping
 - 3.2 Multi-sample SNP and indel calling
 - 3.2.1 Selecting an appropriate quality score threshold
 - 3.3 Protocol
- 4 Integrating analyses: getting the best call set possible
 - 4.1 Analysis read VCF protocol
 - 4.2 Basic indel filtering
 - 4.3 Basic SNP filtering
 - 4.4 Filtering around indels
 - 4.5 Making analysis ready calls SNP calls with hard filtering
 - 4.6 Making analysis ready calls with variant quality score recalibration
- 5 Expected SNP call quality
 - 5.1 Summary results for deep whole genome, multi-sample low-pass, and whole exome
 - 5.2 Expected Ti/Tv ratios

Introduction

Old Webpage-Best Practice v3 (Up to GATK v1.6)

Best Practice Variant Detection with the GATK v3

From GSA

Contents

- 1 Data Processing Pipeline Script
- 2 Introduction
 - 2.1 Lanes, Samples, Cohort
 - 2.2 Testing data: 64x HiSeq on chr20 for NA12878
- 3 Phase I: Raw data processing
 - 3.1 Initial read mapping
 - 3.2 Raw BAM to realigned, recalibrated BAM
 - 3.2.1 Previous recommendation: lane-level recalibration, sample-level realignment
 - 3.2.2 Fast: lane-level realignment at known sites only and lane-level recalibration
 - 3.2.3 Fast + sample-level realignment
 - 3.2.4 Better: sample-level realignment with known indels and recalibration
 - 3.2.5 Best: multi-sample realignment with known sites and recalibration
 - 3.2.6 Misc. notes on the process
- 4 Initial variant discovery and genotyping
 - 4.1 Input BAMs for variant discovery and genotyping
 - 4.2 Multi-sample SNP and indel calling
 - 4.2.1 Selecting an appropriate quality score threshold
 - 4.3 Protocol
- 5 Integrating analyses: getting the best call set possible
 - 5.1 Whole Genome Shotgun experiments
 - 5.1.1 Analysis ready VCF protocol
 - 5.1.2 SNP specific recommendations
 - 5.1.3 Indel specific recommendations
 - 5.2 Whole Exome Experiments
 - 5.2.1 Analysis ready VCF protocol
 - 5.2.2 SNP specific recommendations
 - 5.2.3 Indel specific recommendations
 - 5.3 Making analysis ready SNP and indel calls with hand filtering when VQSR is not possible
- 6 Expected SNP call quality
 - 6.1 Summary results for deep whole genome, multi-sample low-pass, and whole exome
 - 6.2 Expected Ti/Tv ratios
- 7 Previous versions of Best Practices (now outdated)
 - 7.1 Version 2 -- hybrid VQSR and hard filters
 - 7.2 Version 1 -- hard filters

New Website-Best Practice v4-GATK v2.0

The screenshot shows a Mozilla Firefox browser window displaying the GATK Best Practices website. The browser's address bar shows the URL www.broadinstitute.org/gatk/guide/topic?name=best-practices. The website's navigation bar includes links for Home, About, Guide, Community, and Downloads. The main content area is titled "Best Practices" and features a sidebar with a "Guide" menu containing links to Guide Index, Introductory Materials, Technical Documentation, Methods and Workflows, Best Practices (highlighted), FAQs, Tutorials, and Videos. The main text area contains the following sections:

Best Practices

Official guidelines on how to best use our tools for data processing and analysis

Best Practice Variant Detection with the GATK v4, for release 2.0

Introduction

Our current best practice for making SNP and indel calls is divided into four sequential steps: initial mapping, refinement of the initial reads, multi-sample indel and SNP calling, and finally variant quality score recalibration. These steps are the same for targeted resequencing, whole exomes, deep whole genomes, and low-pass whole genomes. Example commands for each tool are available on the individual tool's wiki entry. [There is also a list of which resource files to use with which tool](#)

Note that due to the specific attributes of a project the specific values used in each of the commands may need to be selected/modified by the analyst. Care should be taken by the analyst running our tools to understand what each parameter does and to evaluate which value best fits the data and project design.

Lane, Library, Sample, Cohort

There are four major organizational units for next-generation DNA sequencing processes:

- **Lane:** The basic machine unit for sequencing. The lane reflects the basic independent run of an NGS machine. For Illumina machines, this is the physical sequencing lane.
- **Library:** A unit of DNA preparation that at some point is physically pooled together. Multiple lanes can be run from aliquots from the same library. The DNA library and its preparation is the natural unit that is being sequenced. For example, if the library has limited complexity, then many sequences are duplicated and will result in a high duplication rate across lanes.
- **Sample:** A single individual, such as human CEPH NA12878. Multiple libraries with different properties can be constructed from the original sample DNA source. Here we treat samples as independent individuals whose genome sequence we are attempting to determine. From this perspective, tumor / normal samples are different despite coming from the same individual.
- **Cohort:** A collection of samples being analyzed together. This organizational unit is the most subjective and depends intimately on the design goals of the sequencing project. For population discovery projects like the 1000 Genomes, the analysis cohort is the ~100 individual in each population. For exome projects with many samples (e.g., ESP with 800 EOMI samples) deeply sequenced we divide up the complete set of samples into cohorts of ~50 individuals for multi-sample analyses.

This document describes how to call variation within a single analysis cohort, comprised for one or many samples, each of one or many libraries that were sequenced on at least one lane of an NGS machine.

Note that many GATK commands can be run at the lane level, but will give better results seeing all of the data for a single sample, or even all of the data for all samples. Unfortunately, there's a trade-off in computational cost by

New GATK (v2.0) Website-Download and Guide

GATK Main Page - Mozilla Firefox

File Edit View History Bookmarks Tools Help


View Record x GATK Main Page x +

www.broadinstitute.org/gatk/

COG Pages - aroma.affymet... R Linear Models for Mi... WPS Biobase Authorization Metabolon.com - Vie...

gatk

The Genome Analysis Toolkit or GATK is a software package developed at the Broad Institute to analyse next-generation resequencing data. The toolkit offers a wide variety of tools, with a primary focus on variant discovery and genotyping as well as strong emphasis on data quality assurance. Its robust architecture, powerful processing engine and high-performance computing features make it capable of taking on projects of any size.



[Learn more »](#)

About
The GATK and the people behind it

Guide
Detailed documentation, guidelines and tutorials

Community
Forum for questions and announcements

Videos
Tutorial videos, Q&A and event recordings

Latest stable release
2.0-35
[Release Notes \(2012-07-23\)](#)
[Download now](#)

Download the GATK - Mozilla Firefox

File Edit View History Bookmarks Tools Help

View Record x Download the GATK x +

www.broadinstitute.org/gatk/download

COG Pages - aroma.affymet... R Linear Models for Mi... WPS Biobase Authorization Metabolon.com - Vie...

gatk

The current version is 2.0-35
Please see [this page](#) for complete details on available packages and limitations.

Download GATK 2.0 (beta)
GATK 2.0 includes all of the original GATK 1.x tools as well as many newer and more advanced tools for error modeling, data compression, and variant calling. The version of Queue provided below is built for GATK 2.0.
Please be aware that the GATK 2.0 beta tool chain may be unstable, slow, not scalable, poorly documented, or not interact seamlessly among each other or with other tools in the suite, so could require more effort from users. With these caveats, these tools provide radically improved calling sensitivity, specificity, and performance so are worth the exposure as beta software.

[Download GATK 2.0](#) [Download Queue](#)

Download GATK-lite
GATK-lite is a subset of the full GATK 2.0 release that is free-to-use for all entities, including commercial ones. It includes all of the capabilities (if not the exact tools) from GATK 1.6 but none of the exclusive 2.0 tools. The version of Queue provided below is built for GATK-lite.
For the tech-savvy, GATK-lite is the binary distribution corresponding to the public GATK source released in the Github repository. Everything in GATK-lite is licensed under the MIT license.

New GATK (v2.0) Website-Introduction

The screenshot shows a Mozilla Firefox browser window titled "Intro to the GATK - Mozilla Firefox". The address bar shows the URL "www.broadinstitute.org/gatk/about#high-performance". The browser's menu bar includes "File", "Edit", "View", "History", "Bookmarks", "Tools", and "Help". The browser's toolbar shows "View Record", "Intro to the GATK", and a search icon. The browser's status bar shows several open tabs, including "COG", "Pages - aroma.affymet...", "R Linear Models for Mi...", "WPS", "Biobase Authorization", "Metabolon.com", "Genome.gov", and "The Ca...".

The website content includes a navigation menu with "Home", "About", "Guide", "Community", and "Downloads". The main heading is "Introduction to the GATK". Below this, there are several navigation links: "What is the GATK?", "Using the GATK", "Typical Workflows", "High Performance" (highlighted with a blue circle), "Getting Help", and "Licensing".

The "High Performance" section is titled "High Performance" and has the subtitle "Built for scalability and parallelism". It contains the following text:

The GATK was built from the ground up with performance in mind.

Map/Reduce: it's not just for Google anymore
Every GATK walker is built using the Map/Reduce framework, which is basically a strategy to speed up performance by breaking down large iterative tasks into shorter segments then merging overall results.

Multi-threading
The GATK takes advantage of the latest processors using multi-threading, i. e. run using multiple cores on the same machine, sharing the RAM. To enable multi-threading in the GATK, simply add the `-nt x` argument to your command line, where `x` is the number of threads, or cores, you want to use.

The diagram illustrates multi-threading. It shows a blue box labeled "GATK" with an arrow pointing to a blue box labeled "OS". From the "OS" box, four arrows labeled "threads" point to four orange boxes labeled "CPU core". Below the diagram is the text: "The GATK does multi-threading."

Out on the farm with Queue
Queue is a companion program that allows the GATK to take parallelization to the next level: running jobs on a high-performance computing cluster, or server farm. Queue manages the entire process of breaking down big jobs into many smaller ones (scatter) then collecting and merging results when they are done (gather).

At the Broad, we use a Queue pipeline to run GATK analyses on hundreds, even thousands of exomes, on our cluster of hundreds of nodes.

The diagram illustrates the scatter-gather process. It shows a blue box labeled "scatter" at the top, with many arrows pointing down to a large number of small orange boxes representing individual jobs. From these small boxes, many arrows point up to a blue box labeled "gather" at the bottom.

Queue uses a scatter-gather process to parallelize operations.

The Flagship Features of GATK WorkFlow

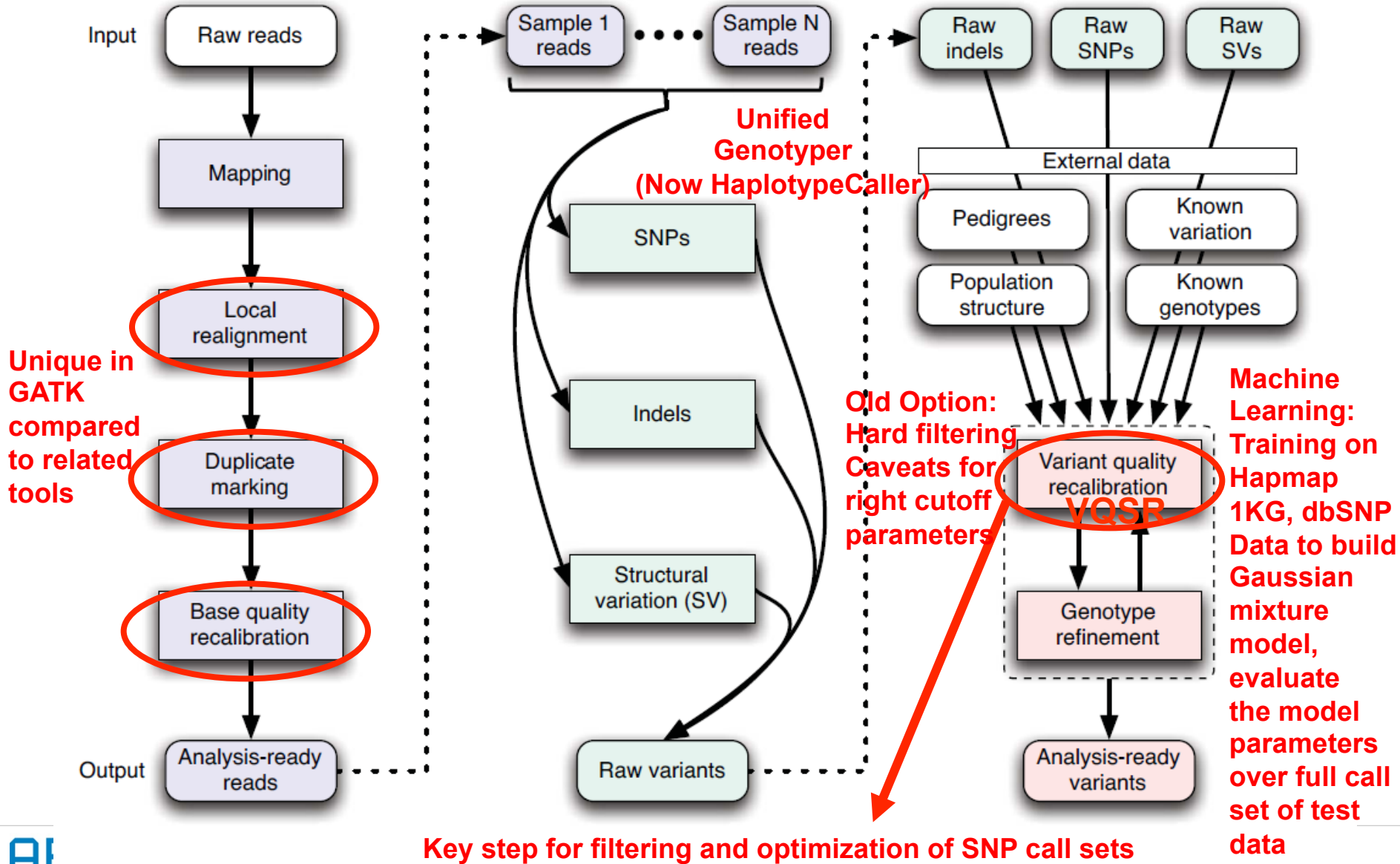
Phase 1: NGS data processing

Phase 2: Variant discovery and genotyping

Phase 3: Integrative analysis

— Typically by lane —

— Typically multiple samples simultaneously but can be single sample alone —



What you have to know about GATK

- Command-line driven and modular-wise framework: need write wrapper program(s) and include GATK commands as system calls
- Web portal-based customizable pipeline may be constructed possibly through pipeline platforms such as Pipeline Pilot, Galaxy etc to publish the wrappers on web for easy usage.
- Allow flexible scenario-based variant detection schemes based on users' need and concern on computational cost
- Dynamic evolving of the toolkit and documentation issues .
- Experimental and work-in-progress types of features in some steps or function tools in the toolkit: e.g. Variant quality score recalibration

Tips or Pre-steps For Preparation to Run GATK Best Practice Procedure

- **first make the “properly” (chromosomes) ordered reference file**
- **Use samtools faidx to create index file for reference file**
- **Use picard CreateSequenceDictionary and the reference file to create the fasta sequence dictionary file**
- **Use picard AddOrReplaceReadGroups to add read group tags and info**
- **Use picard CreateSequenceDictionary to use the reference file to create the fasta sequence dictionary file**
- **use picard ReorderSam to re-order your input bam file(s) for their chromosomes order based on that in the “properly ordered” reference fasta file**
- **use picard SortSam to sorts the alignments of reads in the bam file(s) for coordinate-sorted. (samtools sorted bam files still with issue)**
- **Use picard ValidateSamFile to validate the input bam file(s), relatively stringent**
- **Always index newly created bam file(s) during the GATK steps by using samtools index**
- **Target interval (region) list file(s)**

How Reference File Looks like?

--In Fasta Format

```
tork.ncifcrf.gov - PuTTY
torkv:/banas/nextgen2/illumina/PROC/RefGenomes/hg19/ordered> more hg19_chrM_1st.fa
>chrM
GATCACAGGTCTATCACCCCTATTAACCACTCACGGGAGCTCTCCATGCAT
TTGGTATTTTCGTCTGGGGGGTGTGCACGCGATAGCATTGCGAGACGCTG
GAGCCGGAGCACCCCTATGTGCGAGTATCTGTCTTTGATTCTGCCTCATT
CTATTATTTATCGCACCTACGTTCAATATTACAGGCGAACATACTACTA
AAGTGTGTTAATTAATTAATGCTTGTAGGACATAATAATAACAATTGAAT
GTCTGCACAGCCGCTTTCCACACAGACATCATAACAAAAAATTTCCACCA
AACCCCCCTCCCCCGCTTCTGGCCACAGCACTTAAACACATCTCTGC
CAAACCCCAAAAACAAGAACCCTAACACCAGCCTAACCCAGATTTCAAAT
TTTATCTTTAGGCGGTATGCACTTTTAACAGTCACCCCCCACTAACACA
TTATTTTCCCCTCCCCTCCCATACTACTAATCTCATCAATACAACCCCC
GCCATCCTACCCAGCACACACACACCGCTGCTAACCCCATACCCCGAAC
CAACCAACCCCAAGACACCCCCACAGTTTATGTAGCTTACCTCCTCA
AAGCAATACACTGAAAATGTTTAGACGGGCTCACATCACCCCATAAACAA
ATAGGTTTGGTCCTAGCCTTCTATTAGCTCTTAGTAAGATTACACATGC
AAGCATCCCCGTTCCAGTGAGTTCACCCCTCTAAATCACCCACGATCAAAAG
GGACAAGCATCAAGCACGCAGCAATGCAGCTCAAAACGCTTAGCCTAGCC
ACACCCCCACGGGAAACAGCAGTGATTAACCTTTAGCAATAAACGAAAGT
TTAACTAAGCTATACTAACCCAGGGTTGGTCAATTTTCGTGCCAGCCACC
GCGGTACACGATTAACCCAAGTCAATAGAAGCCGGCGTAAAGAGTGT'TT
TAGATCACCCCTCCCAATAAAGCTAAAACCTCACCTGAGTTGTAAAAAA
CTCCAGTTGACACAAAATAGACTACGAAAGTGGCTTTAACATATCTGAAC
ACACAATAGCTAAGACCCAAACTGGGATTAGATACCCACTATGCTTAGC
CCTAAACCTCAACAGTTAAATCAACAAAACCTGCTCGCCAGAACACTACGA
GCCACAGCTTAAAACCTCAAAGGACCTGGCGGTGCTTCATATCCCTCTAGA
GGAGCCTGTTCTGTAATCGATAAACCCCGATCAACCTCACCCACTCTTGC
TCAGCCTATATACCGCCATCTTCAGCAAACCCCTGATGAAGGCTACAAAGT
AAGCGCAAGTACCCACGTAAAGACGTTAGGTCAAGGTGTAGCCCATGAGG
TGGCAAGAAATGGGCTACATTTTCTACCCAGAAAACCTACGATAGCCCTT
```

How Reference File Looks like?

--chromosomes in certain order, refer to the VQSR training files for needed chromosome order

```
tork.ncifcrf.gov - PuTTY
torkv:/banas/nextgen2/illumina/PROC/RefGenomes/hg19/ordered> grep "chr" hg19_chrM_1st.fa
>chrM
>chr1
>chr2
>chr3
>chr4
>chr5
>chr6
>chr7
>chr8
>chr9
>chr10
>chr11
>chr12
>chr13
>chr14
>chr15
>chr16
>chr17
>chr18
>chr19
>chr20
>chr21
>chr22
>chrX
>chrY
torkv:/banas/nextgen2/illumina/PROC/RefGenomes/hg19/ordered> █
```

GATK order: chrM, chr1, chr2,,chr22, chrX, chrY

Create Needed GATK Input Files of Reference Genome

Resource URL from BROAD:

http://www.broadinstitute.org/gsa/wiki/index.php/Preparing_the_essential_GATK_input_files:_the_reference_genome

```
tork.ncifcrf.gov - PuTTY
torkv:/banas/nextgen2/illumina/PROC/RefGenomes/hg19/ordered> ls -l
total 6167360
-rw-r--r-- 1 yiming abcc 3524 2011-08-11 21:52 hg19_chrM_1st.dict
-rw-r--r-- 1 yiming abcc 3157608038 2011-08-11 21:35 hg19_chrM_1st.fa
-rw-r--r-- 1 yiming abcc 783 2011-08-11 22:05 hg19_chrM_1st.fa.fai
-rw-r--r-- 1 yiming abcc 3299 2011-08-02 10:51 hg19.dict
-rwxr-xr-x 1 zhaoyong abcc 3157608038 2011-07-14 00:49 hg19.fa
-rwxrwxrwx 1 yiming abcc 788 2011-07-19 16:47 hg19.fa.fai
torkv:/banas/nextgen2/illumina/PROC/RefGenomes/hg19/ordered> samtools faidx hg19_chrM_1st.fa
```

- Use samtools faidx to create corresponding .fai index file hg19_chrM_1st.fa.fai for the reference hg19_chrM_1st.fa file
- Use picard CreateSequenceDictionary to create .dict dictionary file for the reference hg19_chrM_1st.fa file

```
tork.ncifcrf.gov - PuTTY
torkv:/banas/nextgen2/illumina/PROC/RefGenomes/hg19/ordered> ls -l
total 6167360
-rw-r--r-- 1 yiming abcc 3524 2011-08-11 21:52 hg19_chrM_1st.dict
-rw-r--r-- 1 yiming abcc 3157608038 2011-08-11 21:35 hg19_chrM_1st.fa
-rw-r--r-- 1 yiming abcc 783 2011-08-11 22:05 hg19_chrM_1st.fa.fai
-rw-r--r-- 1 yiming abcc 3299 2011-08-02 10:51 hg19.dict
-rwxr-xr-x 1 zhaoyong abcc 3157608038 2011-07-14 00:49 hg19.fa
-rwxrwxrwx 1 yiming abcc 788 2011-07-19 16:47 hg19.fa.fai
torkv:/banas/nextgen2/illumina/PROC/RefGenomes/hg19/ordered> java -Xms4g -Xmx4g -jar /opt/nasapps/stow/picard-tool
s-1.70/CreateSequenceDictionary.jar R=hg19_chrM_1st.fa O=hg19_chrM_1st.dict
```

Add read group tags to the bam files

Resource URL from BROAD:

<http://www.broadinstitute.org/gsa/wiki/index.php/ReplaceReadGroups>

<http://picard.sourceforge.net/command-line-overview.shtml#AddOrReplaceReadGroups>

https://getsatisfaction.com/gsa/topics/the_unified_genotyper_complains_about_a_missing_read_group

Typical Command Used:

```
java -Xms4g -Xmx4g -jar
/opt/nasapps/stow/picard-tools-1.70/AddOrReplaceReadGroups.jar
INPUT=/PathToBamFile/before.bam
OUTPUT=/PathToProcessedBamFile/samples_w_@RG/after.bam
RGID=708BRAAXX_Sample_F18
RGLB=F18_illumina
RGPL=Illumina
RGPU=708BRAAXX.lane_7
RGSM=F18
RGCN=NCI-CCR_SF
VALIDATION_STRINGENCY=SILENT
```

- Use picard AddOrReplaceReadGroups to add read group tags to the bam file
- Using a description file and a wrapper program (system call for picard command) would be easier (loop for all bam files)
- It will add read group header (@RG) to the header of bam files (see next slide)
- It will add read group tag (RG:Z:) to each read (see next slide)
- Makes sure using samtools index for the newly created bam file

Header and RG Tags in the bam file after adding read group

reads declared as unsorted

chromosomes unordered

```
tork.ncifcrf.gov - PuTTY
torkkv:/banas/kebebew/AllUsers/GATK_better_s6_7_8/samples_w_@RG> samtools view -h F18_w_RG.bam | more
@HD
VN:1.0 SO:unsorted
@SQ
SN:chr10 LN:135534747
@SQ
SN:chr11 LN:135006516
@SQ
SN:chr12 LN:133851895
@SQ
SN:chr13 LN:115169878
@SQ
SN:chr14 LN:107349540
@SQ
SN:chr15 LN:102531392
@SQ
SN:chr16 LN:90354753
@SQ
SN:chr17 LN:81185210
@SQ
SN:chr18 LN:78077248
@SQ
SN:chr19 LN:59128983
@SQ
SN:chr1 LN:249250621
@SQ
SN:chr20 LN:63025520
@SQ
SN:chr21 LN:48129895
@SQ
SN:chr22 LN:51304566
@SQ
SN:chr2 LN:243199373
@SQ
SN:chr3 LN:198022430
@SQ
SN:chr4 LN:191154276
@SQ
SN:chr5 LN:180915260
@SQ
SN:chr6 LN:171115067
@SQ
SN:chr7 LN:159138663
@SQ
SN:chr8 LN:146364022
@SQ
SN:chr9 LN:141213431
@SQ
SN:chrM LN:16571
@SQ
SN:chrX LN:155270560
@SQ
SN:chrY LN:59373566
@RG
ID:708BRAAXX_Sample_F18 PL:illumina PU:708BRAAXX.lane_7 LB:F18_illumina SM:F18 CN:NCI-CCR
@PG
ID:illumina_export2sam.pl VN:2.0.0 CL:/banas/nextgen2/illumina/PROC/bin/illumina_export2sam.pl --read1=/banas/nextgen23/illumina/data/110427_NCI-GA3_00039_FC_708BRAAXX_Kebebew/Gerald_PE_110512/s_7_1_export.txt --read2=/banas/nextgen23/illumina/data/110427_NCI-GA3_00039_FC_708BRAAXX_Kebebew/Gerald_PE_110512/s_7_2_export.txt --nofilter
NCI-GA3 39:7:47:10371:19281 153 chr10 68144 44 107M * 0 0 GTCAGCAGAG
TAAACAGACAACCCACAGAGTGGGAGAAAATCTTCATAATCTATACATCTGACAGAGGACTAATATCCAGAATCCACAACAAACTCGAACAAATCAG CB
BB>BBBBC@GCACAABDDGGDRGGIGIHHIHHIHHGEE@GIEIIDGDDGBECAGCDGDGAG>GGGGAGBGGD@GGGEGGBGG@GGGGGGGGGGIIGIHHIHDIX
D:Z:107 RG:Z:708BRAAXX_Sample_F18 SM:i:44 AS:i:0
NCI-GA3 39:7:8:7523:6808 163 chr10 70001 44 107M = 70033 161 TTACCAAGG
CTGGGAAGGATAGTGGGGAGCTAGGGTGGAGTGGGCATTGCTCATGGGTACAAAAATAATTAGAATGAATGAGAGTCACTATTTGATAGCACAAATA LI
HHIIIIIGIGII@GGGBFFFCFGGG7GDGDGGG8DGF?FF@ACCEEGB@DEG8DDDGGDIIDDIIIBCHID<PCEEBDRHFIHHIHH8A<ACEEBHBB
XD:Z:107 RG:Z:708BRAAXX_Sample_F18 SM:i:3 AS:i:44
NCI-GA3 39:7:8:8055:20117 163 chr10 70017 44 107M = 70033 123 AGGATAGTGG
GGAGCTAGGGGGAGGGGGGCATTGCTCATGGGTACAAAAATAATTAGAATGAATGAGAGTCACTATTTGATAGCCCAATAGGGGGACAATGGTCAA GG
```

@RG header line added

RG tags added for each read

Reorder the chromosomes in the bam files

Typical Command Used:

```
java -Xms4g -Xmx4g -jar
/opt/nasapps/stow/picard-tools-1.70/ReorderSam.jar INPUT=/PathToBamFile/
samples_w_@RG/F18_w_RG.bam OUTPUT=/PathToBamFile/
samples_w_@RG_Reorder/F18_w_RG_reorder.bam
REFERENCE= /PathToReferenceFile/hg19_chrM_1st.fa
VALIDATION_STRINGENCY=SILENT
```

- Use picard ReorderSam to reorder the chromosome order in the bam file
- ReorderSam is to change the chromosomal order the reference sequences, which is different from sorting the alignment (using picard SortSam; e.g., in coordinate order)
- Makes sure using samtools index for the newly created bam file

Sort the reads in the bam files

Typical Command Used:

```
java -Xms4g -Xmx4g -jar /opt/nasapps/stow/picard-tools-1.70/SortSam.jar  
INPUT=/PathToBamFile/samples_w_@RG_Reorder/F18_w_RG_reorder.bam  
OUTPUT=/PathToBamFile/samples_w_@RG_Reorder_SamSort/  
F18_w_RG_reorder_sort.bam  
SORT_ORDER=coordinate
```

- Use picard SortSam to sort the alignment reads in the bam file
- Makes sure use SORT_ORDER=coordinate, which is required by many tools used in the GATK pipeline (e.g., MarkDuplicates)
- Use picard SortSam to make sure the header is declared as “coordinate” sorted, (samtools sort sorts the reads but won't change the header)
- Alternative option: use ASSUME_SORTED=true option (e.g. in MarkDuplicates)
- Makes sure using samtools index for the newly created bam file

Make Sure All the Training Data, Reference Genome, dbSNP Library files in the Same Chromosomal Order

Error caused by incompatibleness between the GATK training data and our reference

```
##### ERROR MESSAGE: Input files hapmap and reference have incompatible contigs: No overlapping contigs found.  
##### ERROR hapmap contigs = [1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, X]  
##### ERROR reference contigs = [chr1, chr2, chr3, chr4, chr5, chr6, chr7, chr8, chr9, chr10, chr11, chr12, chr13, chr14, chr15, chr16, chr17, chr18, chr19, chr20, chr21, chr22, chrX, chrY, chrM]
```

Error caused by incompatibleness between the dbsnp library file and our reference

```
##### ERROR MESSAGE: Input files dbsnp and reference have incompatible contigs: Order of contigs differences, which is unsafe.  
##### ERROR dbsnp contigs = [chrM, chr1, chr2, chr3, chr4, chr5, chr6, chr7, chr8, chr9, chr10, chr11, chr12, chr13, chr14, chr15, chr16, chr17, chr18, chr19, chr20, chr21, chr22, chrX, chrY]  
##### ERROR reference contigs = [chr1, chr2, chr3, chr4, chr5, chr6, chr7, chr8, chr9, chr10, chr11, chr12, chr13, chr14, chr15, chr16, chr17, chr18, chr19, chr20, chr21, chr22, chrX, chrY, chrM]
```

New version with `-B vcf` option to replace old `-D` option for dbsnp rod library file

Validate the bam files for GATK usage

Typical Command Used:

```
java -Xms4g -Xmx4g -jar /opt/nasapps/stow/picard-tools-1.70/ValidateSamFile.jar  
INPUT=/PathToBamFile/samples_w_@RG_Reorder_SamSort/  
F18_w_RG_reorder_sort.bam OUTPUT=/PathToBamFile/  
samples_w_@RG_Reorder_SamSort/F18_w_RG_reorder_sort.bam.ValidReport
```

- Use picard ValidateSamFile to validate bam file
- Report varied, checking the FAQ page of picard tool to gain help:
http://sourceforge.net/apps/mediawiki/picard/index.php?title=Main_Page
- Depend upon the bam file and platforms, one can use option as
IGNORE=MISSING_TAG_NM, VALIDATION_STRINGENCY=LENIENT etc.

Overview of GATK Phase I Steps (up to v1.6)

--Raw data processing

Resource URL from BROAD:

http://www.broadinstitute.org/gsa/wiki/index.php/Best_Practice_Variant_Detection_with_the_GATK_v3


Multiple Scenarios:

Fast+: sample-level realignment

Better: sample-level realignment with known indels and recalibration

Best: multi-sample realignment with known sites and recalibration

Better: For each sample, merged lane.bams for sample

- 
- **MarkDuplicate**
 - **RealignerTargetCreator** (Local realignment around indels)
 - **IndelRealigner** (Local realignment around indels) (Option: only Known site for large scale project e.g. 1kg)
 - **CountCovariates** (after realignment before recalibration)
 - **TableRecalibration** (Base Quality Recalibration)
 - **CountCovariates** (after recalibration)
 - **AnalyzeCovariates** for data before recalibration
 - **AnalyzeCovariates** for data after recalibration

for each
sample

- **Write a wrapper program to loop the samples (or parallel processing of each sample) and connect steps**

- **Samtools index the newly created bam file for each step**

GATK Phase I Steps: Action commands for sample F18

MarkDuplicates:

```
java -Xms4g -Xmx4g -jar /path/picard-tools-1.70/MarkDuplicates.jar INPUT=/Path/  
F18_w_RG_reorder_sort.bam OUTPUT=/Path/phase_I/F18_w_RG_reorder_sort_dedup.bam  
METRICS_FILE=/Path/phase_I/F18_w_RG_reorder_sort.metricFile VALIDATION_STRINGENCY=SILENT
```



RealignerTargetCreator :

```
java -Xmx4g -jar /Path/GenomeAnalysisTK-1.6-7-g2be5704/bin/GenomeAnalysisTK.jar -T  
RealignerTargetCreator -I /path/F18_w_RG_reorder_sort_dedup.bam -R /path/hg19_chrM_1st.fa -o /path/  
phase_I/F18_w_RG_reorder_sort.output.intervals  
-known /path/bundle/hg19/Mills_and_1000G_gold_standard.indels.hg19.sites.vcf  
-known /path/bundle/hg19/1000G_phase1.indels.hg19.vcf
```

Used by next step



IndelRealigner :

```
java -Xmx10g -jar /path/GenomeAnalysisTK-1.6-7-g2be5704/bin/GenomeAnalysisTK.jar -T IndelRealigner  
-I /Path/phase_I/F18_w_RG_reorder_sort_dedup.bam -R /Path/hg19_chrM_1st.fa  
-targetIntervals /Path/phase_I/F18_w_RG_reorder_sort.output.intervals  
-o /Path/phase_I/F18_w_RG_reorder_sort_realignedBam.bam  
-known /Path/bundle/hg19/Mills_and_1000G_gold_standard.indels.hg19.sites.vcf  
-known /Path/bundle/hg19/1000G_phase1.indels.hg19.vcf
```



GATK Phase I Steps: Action commands for sample F18 (Continued I)

CountCovariates (for data before recalibration):

```
java -Xmx10g -jar /Path/GenomeAnalysisTK-1.6-7-g2be5704/bin/GenomeAnalysisTK.jar  
-T CountCovariates -I INFO -R /Path/hg19_chrM_1st.fa  
-knownSites /Path/bundle/hg19/dbsnp_135.hg19.vcf  
-I /Path/phase_I/F18_w_RG_reorder_sort_realignedBam.bam  
-cov ReadGroupCovariate -cov QualityScoreCovariate -cov CycleCovariate -cov DinucCovariate  
-recalFile /Path/phase_I/F18_w_RG_reorder_sort_CovarTable_beforeRecal.csv
```

TableRecalibration:

```
java -Xmx10g -jar /Path/GenomeAnalysisTK-1.6-7-g2be5704/bin/GenomeAnalysisTK.jar  
-T TableRecalibration -I INFO -R /Path/hg19_chrM_1st.fa  
-I /Path/phase_I/F18_w_RG_reorder_sort_realignedBam.bam  
-o /Path/phase_I/F18_w_RG_reorder_sort_realign_recalBam.bam  
-recalFile /Path/phase_I/F18_w_RG_reorder_sort_CovarTable_beforeRecal.csv
```

Used by next step

CountCovariates (for data after recalibration):

```
java -Xmx10g -jar /Path/GenomeAnalysisTK-1.6-7-g2be5704/bin/GenomeAnalysisTK.jar  
-T CountCovariates -I INFO -R /Path/hg19_chrM_1st.fa  
-knownSites /Path/bundle/hg19/dbsnp_135.hg19.vcf  
-I /Path/phase_I/F18_w_RG_reorder_sort_realign_recalBam.bam  
-cov ReadGroupCovariate -cov QualityScoreCovariate -cov CycleCovariate -cov DinucCovariate  
-recalFile /Path/phase_I_2012/F18_w_RG_reorder_sort_CovarTable_afterRecal.csv
```

Used by next step

GATK Phase I Steps: Action commands for sample F18 (Continued II)

AnalyzeCovariates (for data before recalibration):

```
java -Xmx10g -jar /Path/GenomeAnalysisTK-1.6-7-g2be5704/bin/AnalyzeCovariates.jar  
-recalFile /Path/phase_I/F18_w_RG_reorder_sort_CovarTable_beforeRecal.csv  
-outputDir /Path/phase_I/BeforeRecalfileAnlysis -ignoreQ 5
```



Directory for diagnosis result
of data before recalibration

AnalyzeCovariates (for data after recalibration):

```
java -Xmx10g -jar /Path/GenomeAnalysisTK-1.6-7-g2be5704/bin/AnalyzeCovariates.jar  
-recalFile /Path/phase_I/F18_w_RG_reorder_sort_CovarTable_afterRecal.csv  
-outputDir /Path/phase_I/AfterRecalfileAnlysis -ignoreQ 5
```



Directory for diagnosis result
of data after recalibration

Make sure loop for each sample



Check diagnosis plots



Proceed to GATK phase II

Overview of GATK Phase I Steps (Version 2.0 or Above)

--Raw data processing (simplified commands)

Resource URL from BROAD:

<http://gatkforums.broadinstitute.org/categories/methods-and-workflows>


Multiple Scenarios:

Fast+: sample-level realignment

Better: sample-level realignment with known indels and recalibration

Best: multi-sample realignment with known sites and recalibration

Better: For each sample, merged lane.bams for sample

- 
- **MarkDuplicate**
 - **RealignerTargetCreator** (Local realignment around indels)
 - **IndelRealigner** (Local realignment around indels) (Option: only Known site for large scale project e.g. 1kg)
 - **BaseRecalibrator** (call command twice)
 - **PrintReads**

for each sample

- **Write a wrapper program to loop the samples (or parallel processing of each sample) and connect steps**
- **Samtools index the newly created bam file for each step**

GATK Phase I Steps: Action commands for sample F18

(Major changes for Version 2.0 or above)

BaseRecalibrator (create the initial grp file for the next step) :

```
java -Xmx10g -jar /Path/GenomeAnalysisTK-2.1-13-g1706365/bin/GenomeAnalysisTK.jar  
-T BaseRecalibrator -R /Path/hg19_chrM_1st.fa  
-knownSites /Path/bundle/hg19/dbsnp_135.hg19.vcf  
-I /Path/phase I/F18 w RG reorder sort realignedBam.bam  
-O /Path/phase I/F18 w RG reorder CovarTable Orig.grp
```

BaseRecalibrator (create another grp file for recal and plotting):

```
java -Xmx10g -jar /Path/GenomeAnalysisTK-2.1-13-g1706365/bin/GenomeAnalysisTK.jar  
-T BaseRecalibrator -R /Path/hg19_chrM_1st.fa  
-I /Path/phase I/F18 w RG reorder sort realignedBam.bam  
-BQSR /Path/phase I/F18 w RG reorder CovarTable Orig.grp  
-o /Path/phase I/F18 w RG reorder sort CovarTable Recal.grp
```

Used by next step

For plotting

PrintReads (create recalibrated bam file)

```
java -Xmx10g -jar /Path/GenomeAnalysisTK-2.1-13-g1706365/bin/GenomeAnalysisTK.jar  
-T PrintReads -R /Path/hg19_chrM_1st.fa  
-I /Path/phase I/F18 w RG reorder sort realign recalBam.bam  
-BQSR /Path/phase I/F18 w RG reorder CovarTable Orig.grp  
-O /Path/phase I/F18 w RG reorder sort realignedBam_recalBam.bam
```

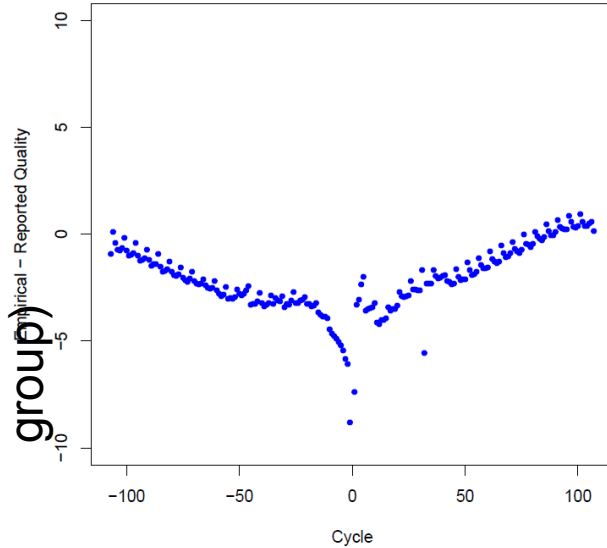
Used by next step

Benefit of Base Quality Score Recalibration

Residual Error by
Machine Cycle
(@PL tag of
read's read
group)

Original

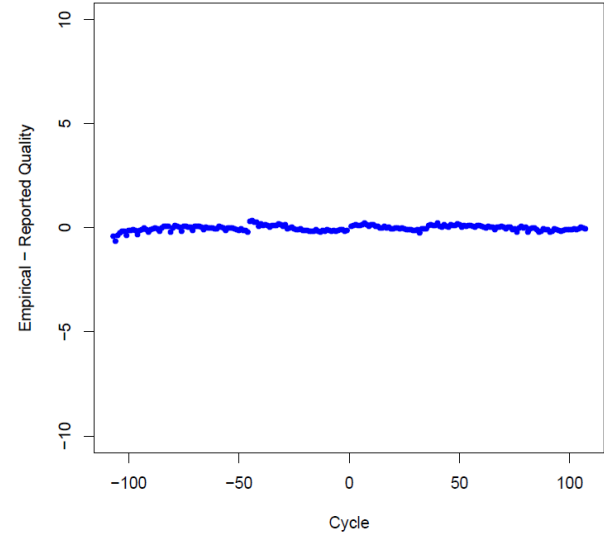
RMSE = 2.673



RMSE = 4.373

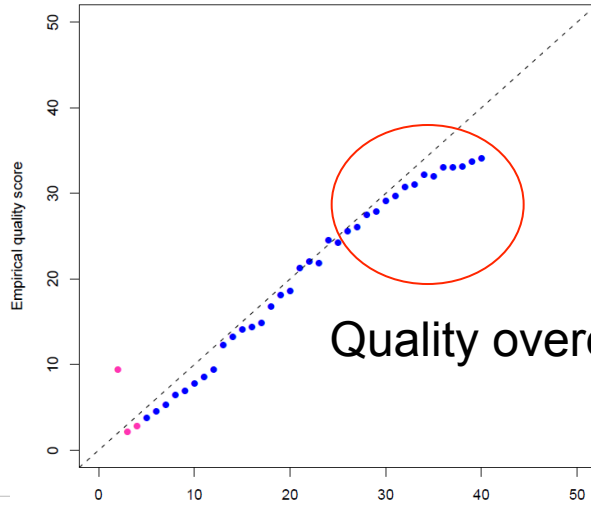
After

RMSE = 0.127

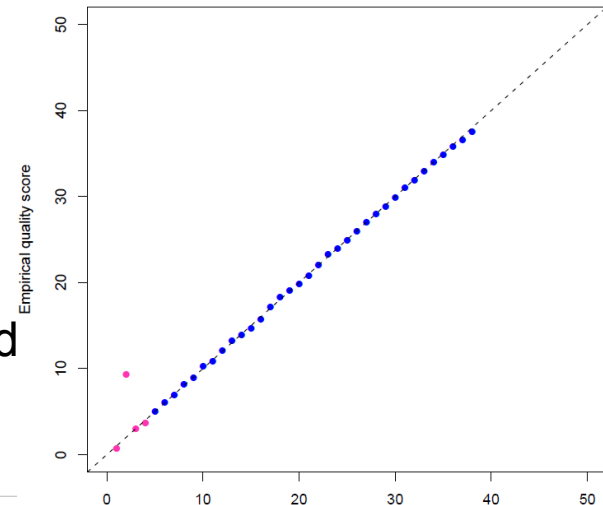


RMSE = 0.15

Reported vs Empirical
Quality



Quality overestimated

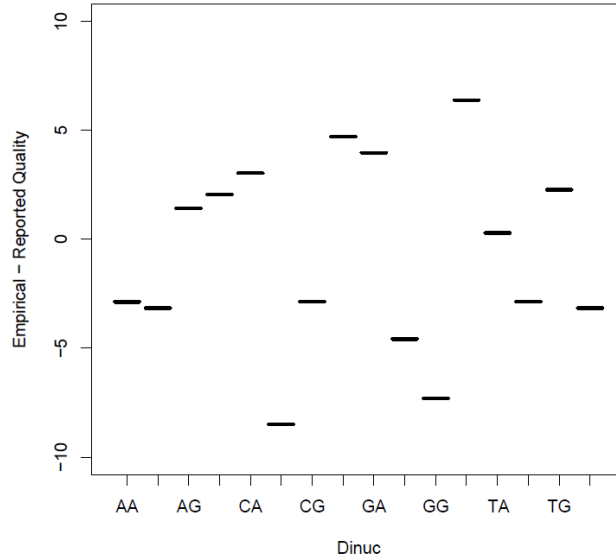


Benefit of Base Quality Score Recalibration

Residual Error by
Dinucleotide

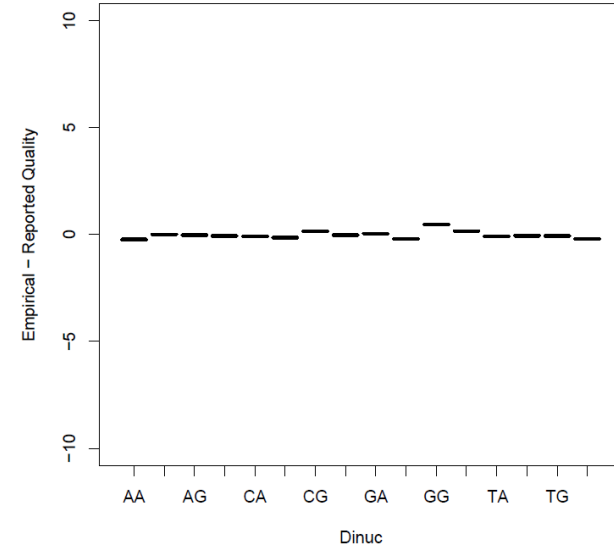
Original

RMSE = 4.122

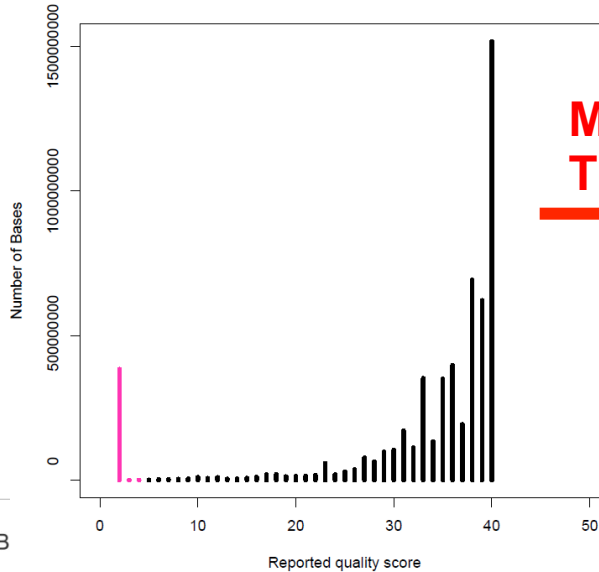


After

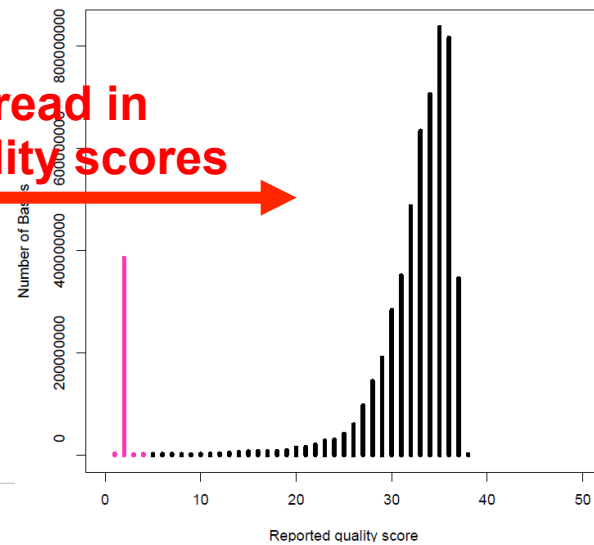
RMSE = 0.164



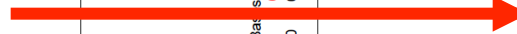
Reported quality score histogram, entropy = 3.528



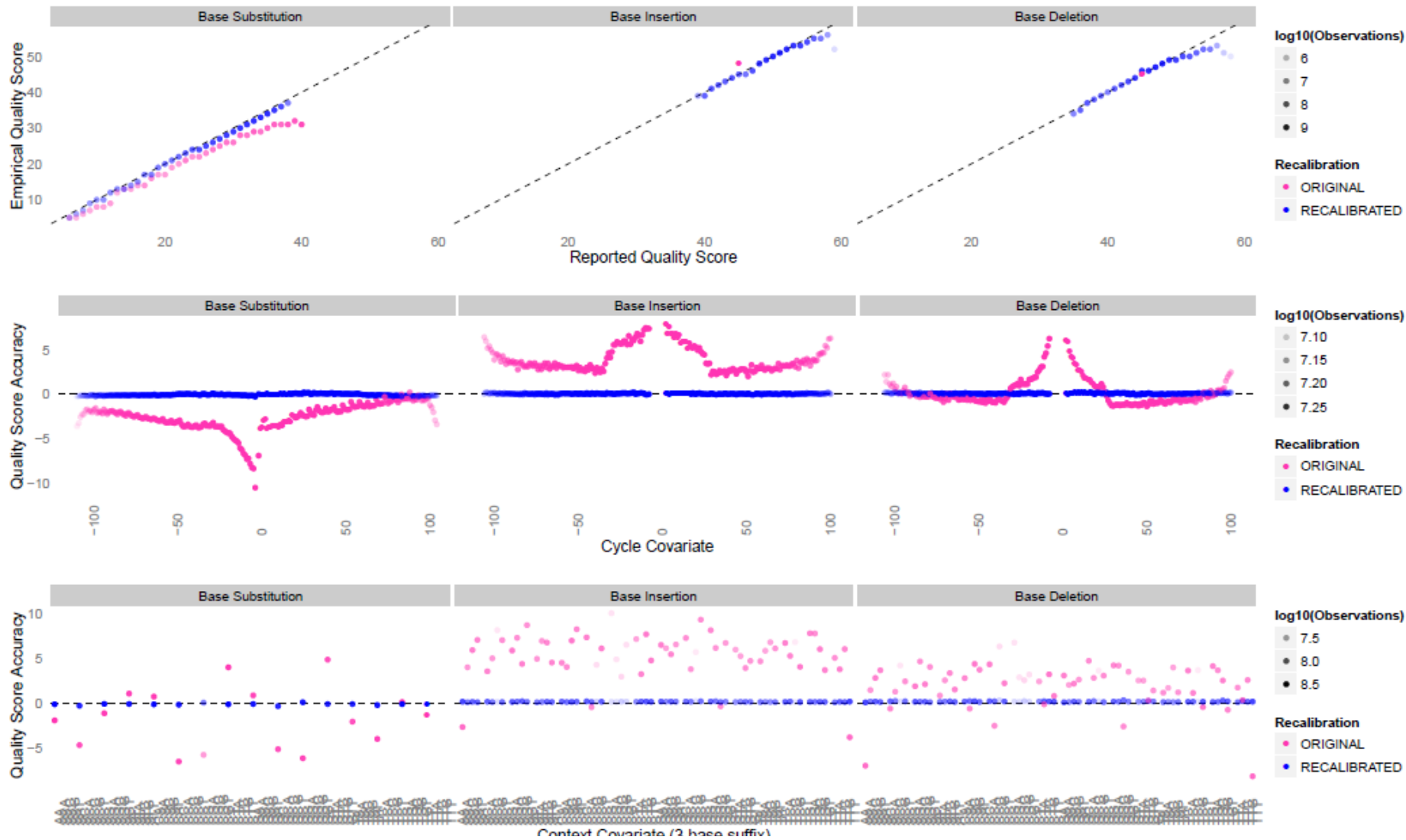
Reported quality score histogram, entropy = 3.54



More spread in
The quality scores



Example Plots of Base Quality Score Recalibration For GATK Version 2.0



Overview of GATK Phase II Steps

--Initial variant discovery and genotyping

Resource URL from BROAD:

http://www.broadinstitute.org/gsa/wiki/index.php/Best_Practice_Variant_Detection_with_the_GATK_v3

- **Call Unified Genotyper (single sample; multiple samples; SNP and Indel separately or simultaneously; many options (e.g., considering coverage, phred-scaled confidence threshold for calls); option use interval list file for target region -L)**
- **Now in V2.0 or above, HaplotypeCaller (performance issue, time cost)**
- **SelectVariants for SNP (option use interval list file for target region -L)**
- **SelectVariants for Indel (option use interval list file for target region -L)**

Used to be separated for SNP (Unified Genotyper) and Indel (Dindel): Now In V2 and V3 all in single command with options.

GATK Phase II Steps: Action commands for all samples

UnifiedGenotyper (call all samples altogether and only variants at target interval):

```
java -jar /Path/GenomeAnalysisTK-1.6-7-g2be5704/bin/GenomeAnalysisTK.jar
-T UnifiedGenotyper -R /Path/hg19_chrM_1st.fa
-I /Path/phase_I/S1_w_RG_reorder_sort_realign_recalBam.bam
-I /Path/phase_I/S2_w_RG_reorder_sort_realign_recalBam.bam
-I /Path/phase_I/S3_w_RG_reorder_sort_realign_recalBam.bam
.....
-I /Path/phase_I/S19_w_RG_reorder_sort_realign_recalBam.bam
--dbsnp /Path/bundle/hg19/dbsnp_135.hg19.vcf
-glm BOTH
-o /Path/phase_II_initialSNPCalls/GATK_UG_AllSamples_snps_indel.raw.afterRecal.vcf
-stand_call_conf 50
-stand_emit_conf 10 -dcov 50
-L /Path/Exome_Target_Region_SureSelect_AllExon_50mb_Bedfiles/029720_D_BED_20101013.bed
```

Select out only SNPs

Select out only Indels

Target interval
bed file
Downloaded from
Agilent

GATK Phase II Steps: Action commands for all samples

HypotypeCaller (call all samples altogether and only variants at target interval):

```
java -jar /Path/GenomeAnalysisTK-2.2-4-g4a174fb/bin/GenomeAnalysisTK.jar
-T HaplotypeCaller -R /Path/hg19_chrM_1st.fa
-I /Path/phase_I/S1_w_RG_reorder_sort_realign_recalBam.bam
-I /Path/phase_I/S2_w_RG_reorder_sort_realign_recalBam.bam
-I /Path/phase_I/S3_w_RG_reorder_sort_realign_recalBam.bam
.....
-I /Path/phase_I/S19_w_RG_reorder_sort_realign_recalBam.bam
--dbSNP /Path/bundle/hg19/dbSNP_135.hg19.vcf
-glm BOTH
-o /Path/phase_II_initialSNPCalls/GATK_HTC_AllSamples_snps_indel.raw.afterRecal.vcf
-stand_call_conf 50
-stand_emit_conf 10
-minPruning 5
-L /Path/Exome_Target_Region_SureSelect_AllExon_50mb_Bedfiles/029720_D_BED_20101013.bed
```

Select out only SNPs

Select out only Indels

Target interval
bed file
Downloaded from
Agilent

--minPruning option: The minimum allowed pruning in assembly graph.

--enable_experimental_downsampling -dcov 10: no more than 10 reads starting at the exact same position will be included in the analyzed data

GATK Phase II Steps: Select out SNPs and Indels

Select out only SNPs:

```
java -jar /Path/GenomeAnalysisTK-1.6-7-g2be5704/bin/GenomeAnalysisTK.jar
-T SelectVariants -R /Path/hg19_chrM_1st.fa
-L /Path/Exome_Target_Region_SureSelect_AllExon_50mb_Bedfiles/029720_D_BED_20101013.bed
--variant /Path/phase_II_initialSNPCalls/GATK_UG_AllSamples_snps_indel.raw.afterRecal.vcf
-selectType SNP
-o /Path/phase_II_initialSNPCalls/GATK_UG_AllSamples_SelSNP.vcf
```

Select out only Indels:

```
java -jar /Path/GenomeAnalysisTK-1.6-7-g2be5704/bin/GenomeAnalysisTK.jar
-T SelectVariants -R /Path/hg19_chrM_1st.fa
-L /Path/Exome_Target_Region_SureSelect_AllExon_50mb_Bedfiles/029720_D_BED_20101013.bed
--variant /Path/phase_II_initialSNPCalls/GATK_UG_AllSamples_snps_indel.raw.afterRecal.vcf
-selectType INDEL
-o /Path/phase_II_initialSNPCalls/GATK_UG_AllSamples_SelIndel.vcf
```

Overview of GATK Phase III Steps

Integrating analyses: getting the best call set possible

- **Making analysis ready SNP and indel calls with hand filtering when VQSR is not possible (e.g., targeted resequencing of a small region); Indel data lacks of data points for modeling**
 - Used to have Basic Indel Filtering, Basic SNP filtering, Filtering around Indels (in GATK v2, now eliminated and simplified in v3)**

- **Variant Quality Score Recalibration-VQSR (Whole Exome, Whole Genome Shotgun experiments etc, SNP vs Indel)**
 - 1. VariantRecalibrator for VQSR model (Ti/Tv-free approach).**
 - 2. ApplyRecalibration Select SNPs by Chosen Cutoffs set up by the VQSR model (Options for truth sensitivity level 0.90, 0.99 etc)**

New Version TiTv-Free VQSR Over TiTv-Targeted Approach

Requires an additional truth data set, and cuts the VQSLOD at given sensitivities to the truth set.

Advantages

- The truth sensitivity (TS) approach gives you back the novel Ti/Tv as a QC metric
- The truth sensitivity (TS) approach is conceptual cleaner than deciding on a novel Ti/Tv target for your dataset
- The TS approach is easier to explain and defend, as saying "I took called variants until I found 99% of my known variable sites" is easier than "I took variants until I dropped my novel Ti/Tv ratio to 2.07"

GATK Phase III Steps: Action commands for all samples

VariantRecalibrator:

```
java -Xmx4g -jar /Path/GenomeAnalysisTK-1.6-7-g2be5704/bin/GenomeAnalysisTK.jar
-T VariantRecalibrator -R /Path/hg19_chrM_1st.fa
-input /Path/phase_II_initialSNPCalls/GATK_UG_AllSamples_SeISNP.vcf
-resource:hapmap,known=false,training=true,truth=true,prior=15.0 /Path/bundle/hg19/
hapmap_3.3.hg19.sites.vcf
-resource:omni,known=false,training=true,truth=false,prior=12.0 /Path/bundle/
hg19/1000G_omni2.5.hg19.sites.vcf
-resource:dbsnp,known=true,training=false,truth=false,prior=8.0 /Path/bundle/hg19/dbsnp_135.hg19.vcf
-an QD -an HaplotypeScore -an MQRankSum -an ReadPosRankSum -an FS -an MQ -an InbreedingCoeff
--maxGaussians 6
-recalFile /Path/phase_II_initialSNPCalls/phase_III/GATK_UG_AllSamples_SeISNP.VarRecal
-tranchesFile /Path/phase_II_initialSNPCalls/phase_III/GATK_UG_AllSamples_SeISNP.tranches
-rscriptFile /Path/phase_II_initialSNPCalls/phase_III/GATK_UG_AllSamples_SeISNP.R
-mode SNP
```

VariantRecalibrator (at truth sensitivity 0.99 or 99%):

```
java -Xmx4g -jar /Path/GenomeAnalysisTK-1.6-7-g2be5704/bin/GenomeAnalysisTK.jar
-T ApplyRecalibration -R /Path/hg19_chrM_1st.fa
-input /Path/phase_II_initialSNPCalls/GATK_UG_AllSamples_SeISNP.vcf
-ts_filter_level 99.0
-tranchesFile /Path/phase_II_initialSNPCalls/phase_III/GATK_UG_AllSamples_SeISNP.tranches
-recalFile /Path/phase_II_initialSNPCalls/phase_III/GATK_UG_AllSamples_SeISNP.VarRecal
-o /Path/phase_II_initialSNPCalls/phase_III/GATK_UG_AllSamples_SeISNP.recalibrated.filtered.099.vcf
-mode SNP
```

Used by
next step

Option change to 90.0 for more stringent 90% truth sensitivity

How the resulting vcf file looks like?

GATK resource:

http://www.broadinstitute.org/gsa/wiki/index.php/Understanding_the_Unified_Genotyper%27s_VCF_files

Header part of the vcf file:

```
tork.ncifcrf.gov - PuTTY
##fileformat=VCFv4.1
##FILTER=<ID=LowQual,Description="Low quality">
##FORMAT=<ID=AD,Number=.,Type=Integer,Description="Allelic depths for the ref and alt alleles in the order listed">
##FORMAT=<ID=DP,Number=1,Type=Integer,Description="Read Depth (only filtered reads used for calling)">
##FORMAT=<ID=GQ,Number=1,Type=Float,Description="Genotype Quality">
##FORMAT=<ID=GT,Number=1,Type=String,Description="Genotype">
##FORMAT=<ID=PL,Number=.,Type=Float,Description="Normalized, Phred-scaled likelihoods for AA,AB,BB genotypes where A=ref and B=alt; if site is not biallelic, number of likelihoods if n*(n+1)/2">
##INFO=<ID=AC,Number=.,Type=Integer,Description="Allele count in genotypes, for each ALT allele, in the same order as listed">
##INFO=<ID=AF,Number=.,Type=Float,Description="Allele Frequency, for each ALT allele, in the same order as listed">
##INFO=<ID=AN,Number=1,Type=Integer,Description="Total number of alleles in called genotypes">
##INFO=<ID=BaseQRankSum,Number=1,Type=Float,Description="Z-score from Wilcoxon rank sum test of Alt Vs. Ref base qualities">
##INFO=<ID=DB,Number=0,Type=Flag,Description="dbSNP Membership">
##INFO=<ID=DP,Number=1,Type=Integer,Description="Filtered Depth">
--More-- (0%)
```

Data part of the vcf file:

The column names of the data matrix

```
tork.ncifcrf.gov - PuTTY
ndel.recal_data.csv indelDebug=false dovit=false GSA_PRODUCTION_ONLY=false exactCalculation=LINEAR_EXPERIMENTAL ignoreSNPAleles=false output_all_callable_bases=false genotype=false out=org.broadinstitute.sting.gatk.io.stubs.VCFWriterStub NO_HEADER=org.broadinstitute.sting.gatk.io.stubs.VCFWriterStub sites_only=org.broadinstitute.sting.gatk.io.stubs.VCFWriterStub debug_file=null metrics_file=null annotation=[]"
##source=SelectVariants
#CHROM POS ID REF ALT QUAL FILTER INFO FORMAT F17 F18 F19
chr1 14522 rs17149433 G A 106.19 PASS AC=2;AF=0.33;AN=6;BaseQRankSum=0.750;DB;DP=31;Dels=0.00;FS=2.017;HRun=0;HaplotypeScore=0.5548;MQ=35.31;MQ0=0;MQRankSum=-0.528;QD=5.90;ReadPosRankSum=-1.806;SB=-35.91 GT:AD:DP:GQ:PL 0/1:2,3:6:21.40:88,0,21 0/0:11,2:13:18.04:0,18,189 0/1:9,3:12:56.73:57,0,184
chr1 14542 rs17149429 A G 103.81 PASS AC=3;AF=0.50;AN=6;BaseQRankSum=-0.231;DB;DP=30;Dels=0.00;FS=2.098;HRun=1;HaplotypeScore=0.2629;MQ=36.96;MQ0=0;MQRankSum=-0.643;QD=3.46;ReadPosRankSum=0.437;SB=-62.45 GT:AD:DP:GQ:PL 0/1:2,2:4:30.29:63,0,30 0/1:9,5:14:48.12:48,0,161 0/1:8,4:12:31.20:31,0,201
chr1 14653 rs62635297 C T 327.09 PASS AC=3;AF=0.50;AN=6;BaseQRankSum=1.633;DB;DP=80;Dels=0.00;FS=1.490;HRun=0;HaplotypeScore=0.8720;MQ=37.96;MQ0=0;MQRankSum=0.897;QD=4.09;ReadPosRankSum=-2.968;SB=-52.06 GT:AD:DP:GQ:PL 0/1:6,6:12:81.98:186,0,82 0/1:27,7:35:53.93:54,0,549 0/1:23,8:33:99:126,0,475
--More-- (0%)
```

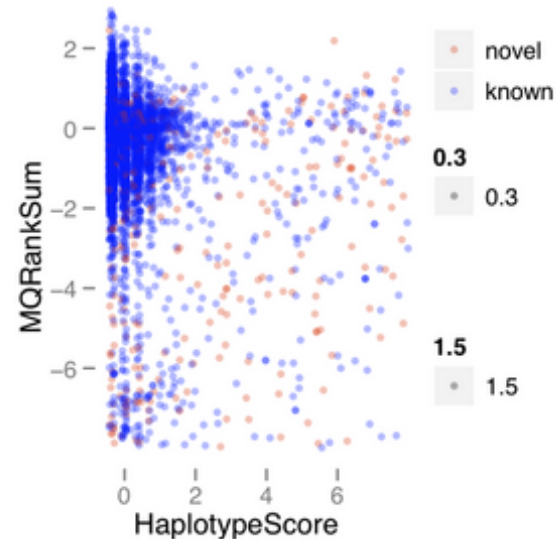
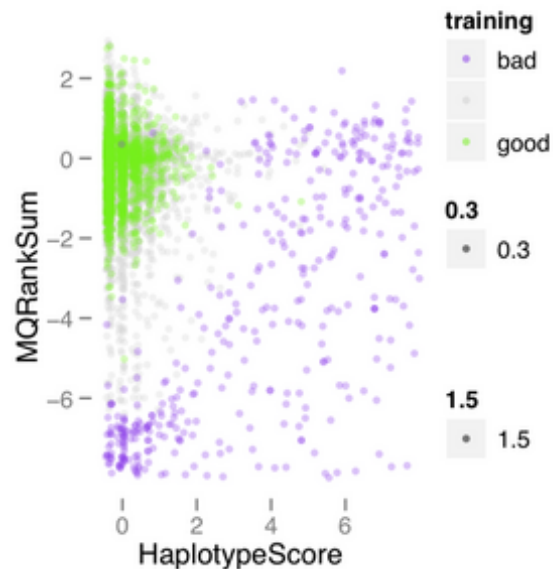
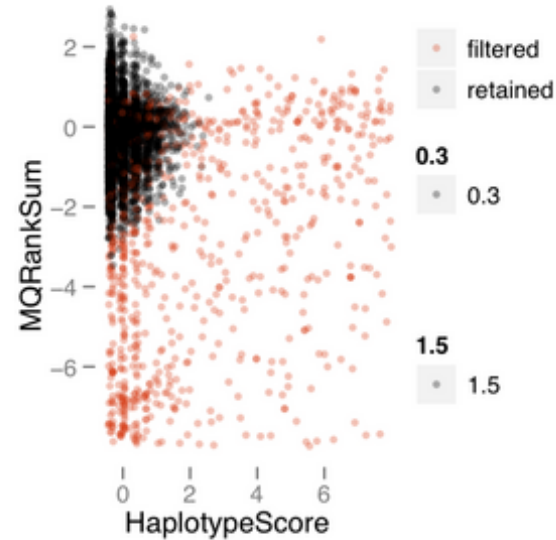
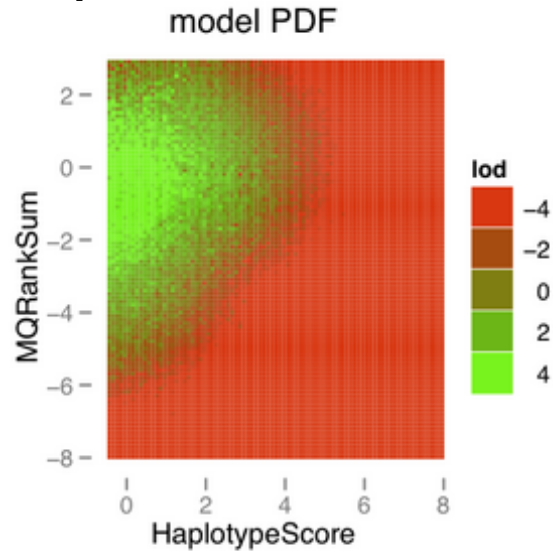
VCF File QC

- **AD: Allelic depths for the ref and alt alleles in the order listed**
- **DP: Read Depth (only filtered reads used for calling)**
- **GQ: Genotype Quality**
- **BaseQRankSum: Z-score from Wilcoxon rank sum test of Alt vs. Ref base qualities**
- **FS: Phred-scaled p-value using Fisher's exact test to detect strand bias**
- **MQ: RMS Mapping Quality**
- **MQ0: Total Mapping Quality Zero Reads**
- **MQRankSum: Z-score From Wilcoxon rank sum test of Alt vs Ref read mapping qualities**
- **QD: Varinat Confidence/Quality by Depth**
- **ReadPosRankSum: Z-score from Wilcoxon rank sum test of Alt vs Ref read position bias**
- **SB: Strand Bias**
- **HaplotypeScore: Consistency of the site with two (and only two) segregating haplotypes**

.....

<http://gatkforums.broadinstitute.org/discussion/1268/how-should-i-interpret-vcf-files-produced-by-the-gatk>

VQSR: Pair-wise combination of annotations used in modeling 2D projection of the Gaussian mixture model is shown (MQRankSum vs HaplotypeScore)



GATK example

Evaluating SNP call quality

Expected number of calls?

- The number of SNP calls should be close to the average human heterozygosity of 1 variant per 1000 bases
- Only detects gross under/over calling

Concordance with genotype chip calls?

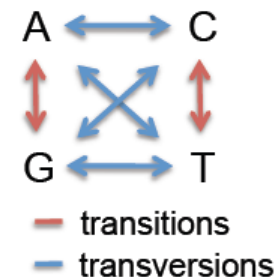
- Often we have genotype chip data that indicates the hom-ref, het, hom-var status at millions of sites
- Good SNP calls should be >99.5% consistent these chip results, and >99% of the variable sites should be found
- The chip sites are in the better parts of the genome, and so are not representative of the difficulties at novel sites

What fraction of my calls are already known?

- dbSNP catalogs most common variation, so most of the true variants found will be in dbSNP
- For single sample calls, ~90 of variants should be in dbSNP
- Need to adjust expectation when considering calls across samples

Transition to transversion ratio (Ti/Tv)?

- Transitions are twice as frequent as transversions (see *Ebersberger, 2002*)
 - Validated human SNP data suggests that the Ti/Tv should be ~2.1 genome-wide and ~2.8 in exons
- FP SNPs should have Ti/Tv around 0.5
- Ti/Tv is a good metric for assessing SNP call quality



Use GATK VariantEval to Evaluate the TiTv Ratio of SNPs

Resource URL from BROAD:

http://www.broadinstitute.org/gsa/gatkdocs/release/org_broadinstitute_sting_gatk_walkers_varianteval_VariantEvalWalker.html

Command example (old version):

```
java -Xmx4g -jar /Path/GenomeAnalysisTK-1.1-23-g8072bd9/bin/GenomeAnalysisTK.jar
-T VariantEval -R /Path/hg19_chrM_1st.fa
-B:dbsnp,VCF /Path/bundle/hg19/dbsnp_132.hg19.vcf
-B:eval,VCF /Path/phase_II_initialSNPCalls/GATK_UG_AllSamples_SelSNP.vcf
-o /Path/phase_II_initialSNPCalls/GATK_UG_AllSamples_SelSNP.vcf.eval.gatkreport.txt
```

Command example (new version):

```
java -Xmx4g -jar /Path/GenomeAnalysisTK-1.6-7-g2be5704/bin/GenomeAnalysisTK.jar
-T VariantEval -R /Path/hg19_chrM_1st.fa
--eval: /Path/phase_II_initialSNPCalls/GATK_UG_AllSamples_SelSNP.vcf
--dbsnp /Path/bundle/hg19/dbsnp_135.hg19.vcf
-o /Path/phase_II_initialSNPCalls/GATK_UG_AllSamples_SelSNP.vcf.eval.gatkreport.txt
```

Note: Although commands slightly different in syntax between new and old version of GATK, which does not cause any difference in results, the versions of dbSNP would have impact to cause difference in TiTv ratio obtained (e.g., dbsnp_132.hg19.vcf vs dbsnp_135.hg19.vcf). The old version dbsnp_132.hg19.vcf is only suggested to be used in VariantEval and would make the TiTv appear better than that if using new dbSNP (v135), since more “known” SNPs in new version, which are novel for old version.

Impact of Target Interval (Region) for Exome-seq

Target Region/Interval:

- Agilent Sure Select Human All Exon 50 Mb kit for the library
- Corresponding target region file (a bed format file, 0-based) downloadable from Agilent eArray website was used as target Interval list file for GATK

TiTv Ratio Indicates SNP Quality

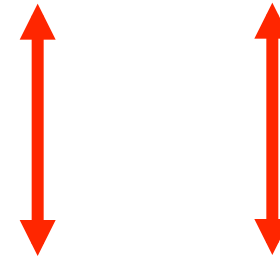
(GATK SNP Calls Within Target Interval Supposed to Have Better Quality and Indeed Have Higher TiTv Ratio)

3-sample dataset

SNP call set without selection by target interval

targetTruthSensitivity	numKnown	numNovel	knownTiTv	novelTiTv	minVQSLod
90	295920	12049	2.1756	1.2272	5.8503
99	330447	18801	2.1816	1.0939	2.1305
99.9	371310	42581	2.1085	0.9996	-3.181
100	389096	79834	2.065	1.032	-infinity

The TiTv ratios are greatly improved using target interval



targetTruthSensitivity	numKnown	numNovel	knownTiTv	novelTiTv	minVQSLod
90	36166	1473	2.8955	2.1609	6.8598
99	43532	2598	2.7907	1.8026	3.0361
99.9	50037	7660	2.6545	1.273	-4.3851
100	52844	15120	2.578	1.217	-infinity

SNP call set within target interval

Comparison of GATK SNP Calls Before and After VQSR (V3)

SNP call sets with 3 samples within target interval

After VQSR
Filter-level 0.90

Novelty	nTi	nTv	TiTvRatio
all	27890	9750	2.860513
known	26883	9284	2.895627
novel	1007	466	2.160944

After VQSR
Filter-level 0.99

Novelty	nTi	nTv	TiTvRatio
all	33719	12411	2.716864
known	32048	11484	2.790665
novel	1671	927	1.802589

The TiTv ratios are greatly improved
After VQSR

Novelty	nTi	nTv	TiTvRatio
all	46375	21589	2.14808467
known	38075	14769	2.57803507
novel	8300	6820	1.2170088

UG SNP call set Before VQSR

GATK SNP Calls with More Samples Have Better Quality —After VQSR V3

SNP call set with 3-samples dataset

targetTruthSensitivity	numKnown	numNovel	knownTiTv	novelTiTv	minVQSLod
90	36166	1473	2.8955	2.1609	6.8598
99	43532	2598	2.7907	1.8026	3.0361
99.9	50037	7660	2.6545	1.273	-4.3851
100	52844	15120	2.578	1.217	-infinity

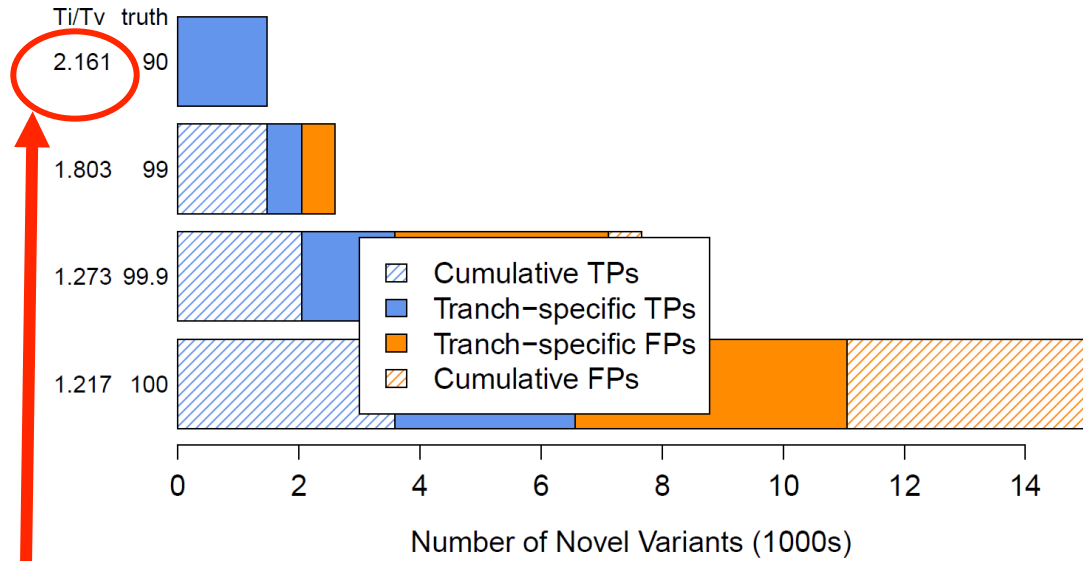
The TiTv ratios are greatly improved
Calling with more samples

targetTruthSensitivity	numKnown	numNovel	knownTiTv	novelTiTv	minVQSLod
90	52780	2873	2.9927	2.8154	6.4358
99	65078	6059	2.8113	2.0555	2.6208
99.9	74784	14900	2.6661	1.4251	-2.4309
100	80336	34062	2.5624	1.175	-infinity

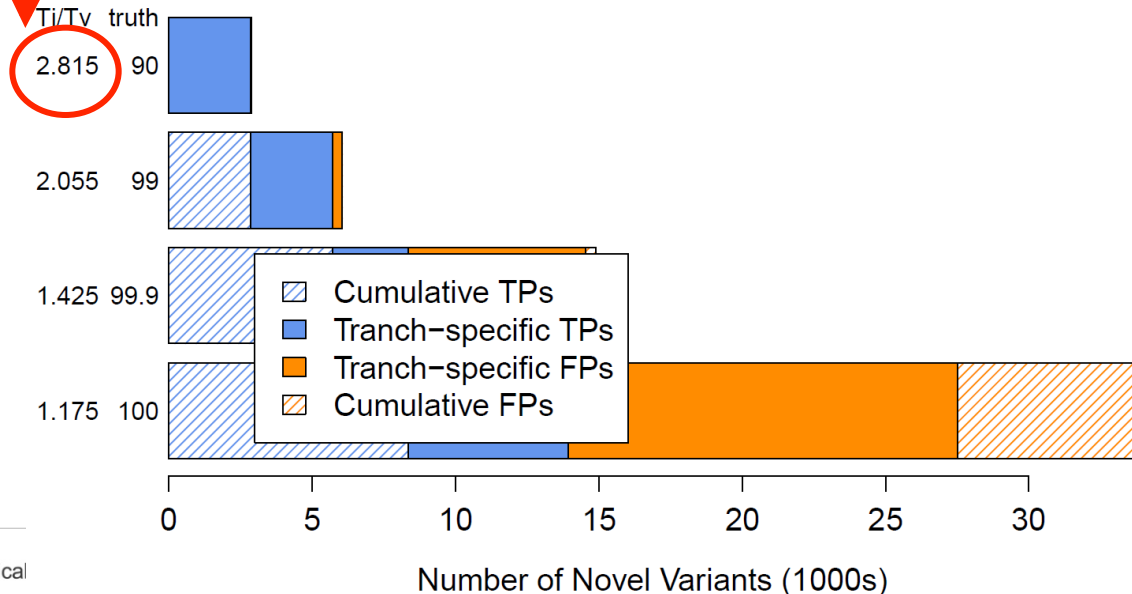
SNP call set with 19-sample dataset

Novel SNP Calls with More Samples Have Better Quality —After VQSR V3 (both call sets within target interval)

SNP callset
with only 3
samples



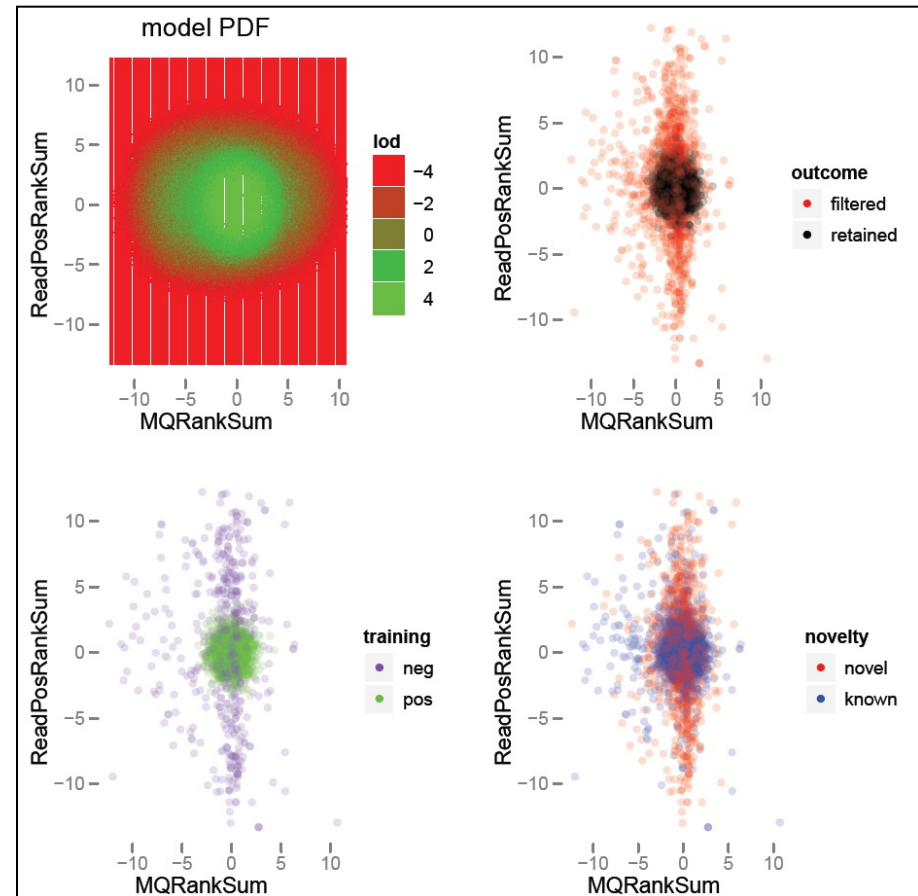
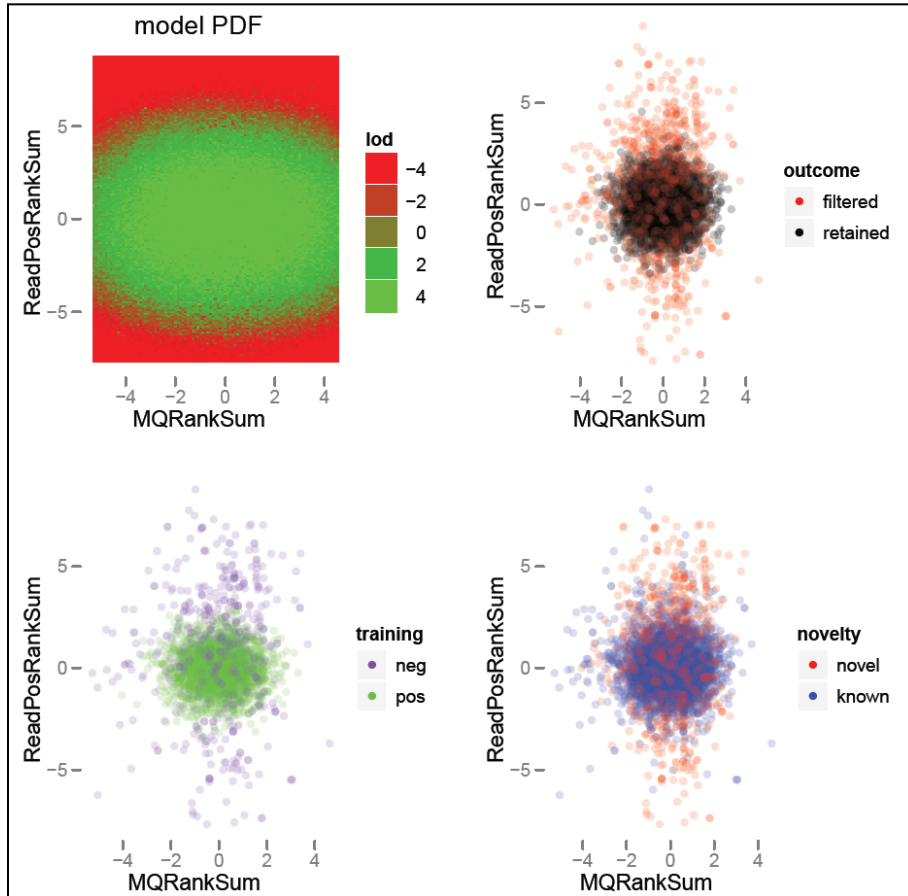
SNP callset
with all 19
samples



VQSR: Pair-wise combination of annotations used in modeling 2D projection of the Gaussian mixture model is shown (ReadPosRankSum vs MQRankSum)

3-sample dataset

19-sample dataset

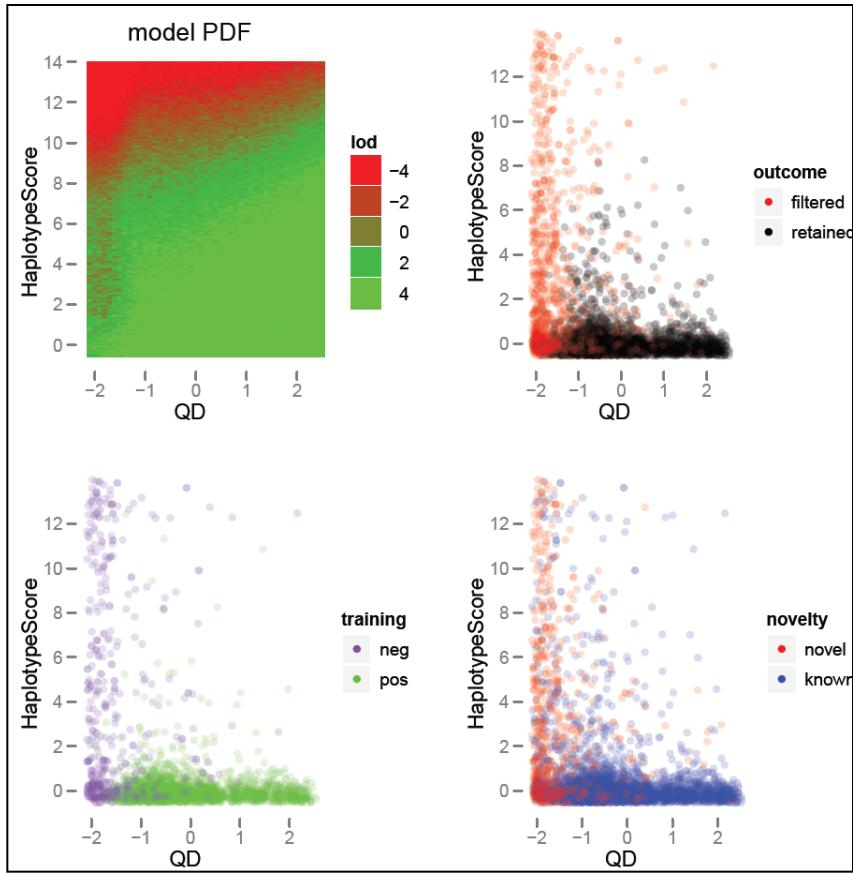


Clustered more tightly in 19-sample data indicates improved separation

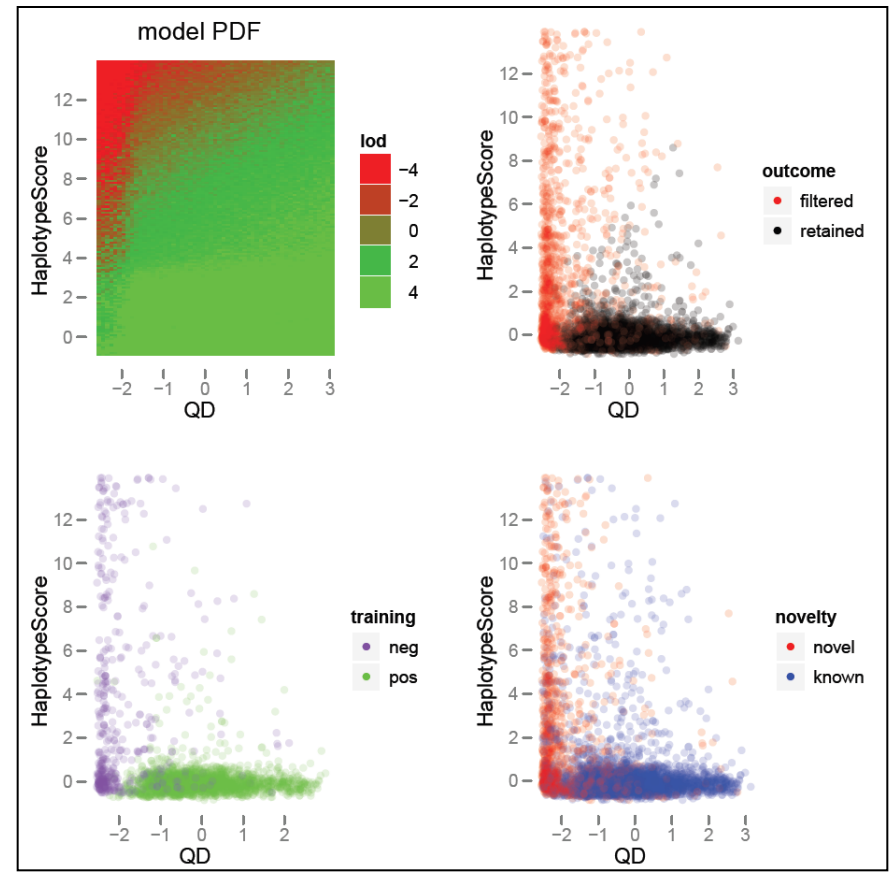
VQSR: Pair-wise combination of annotations used in modeling

2D projection of the Gaussian mixture model is shown (HaplotypeScore vs QD)

SNP call set with only 3 samples



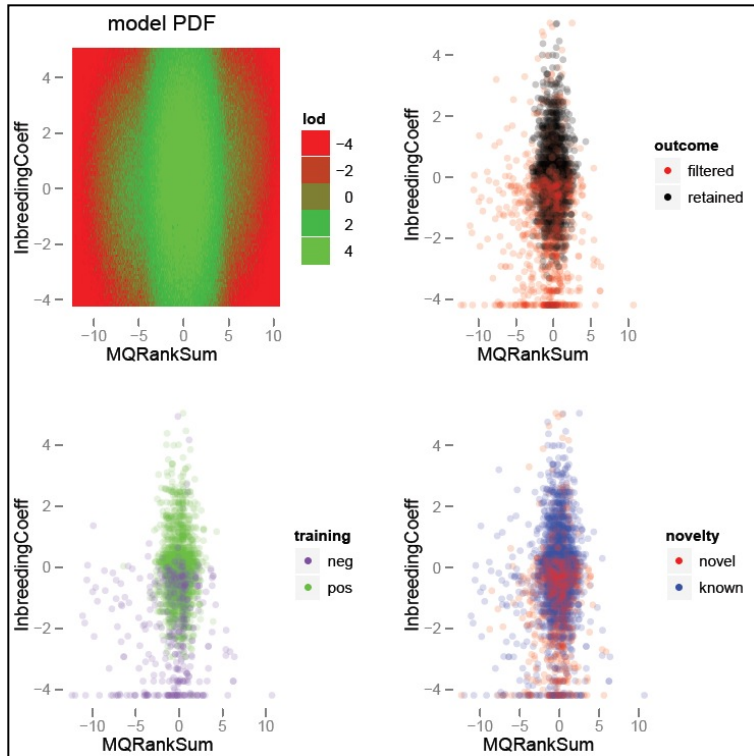
SNP call set with all 19 samples



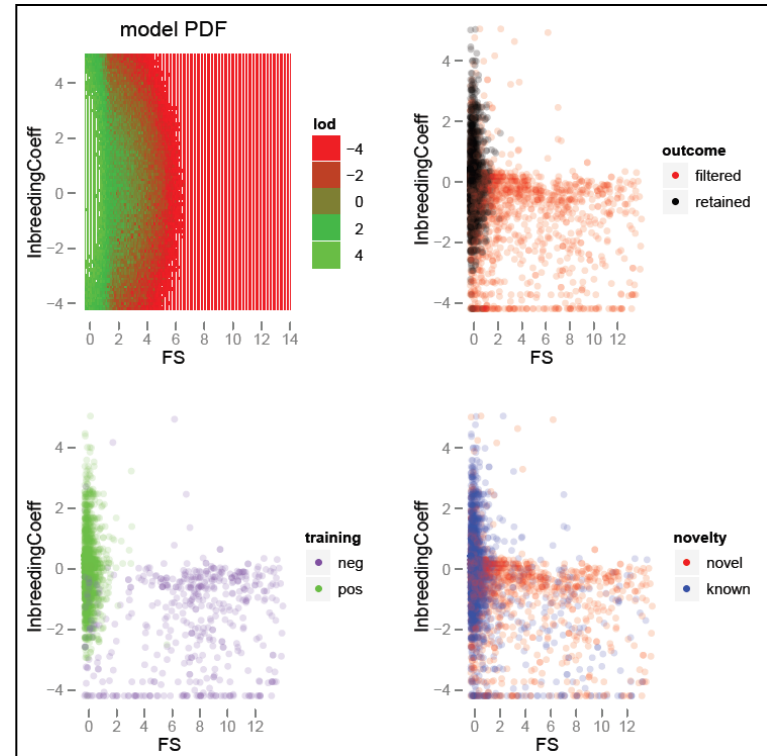
Many annotations used to help modeling

Some annotations require large sample size (e.g., InbreedingCoeff required at least 10 samples)

InbreedingCoeff vs MQRankSum

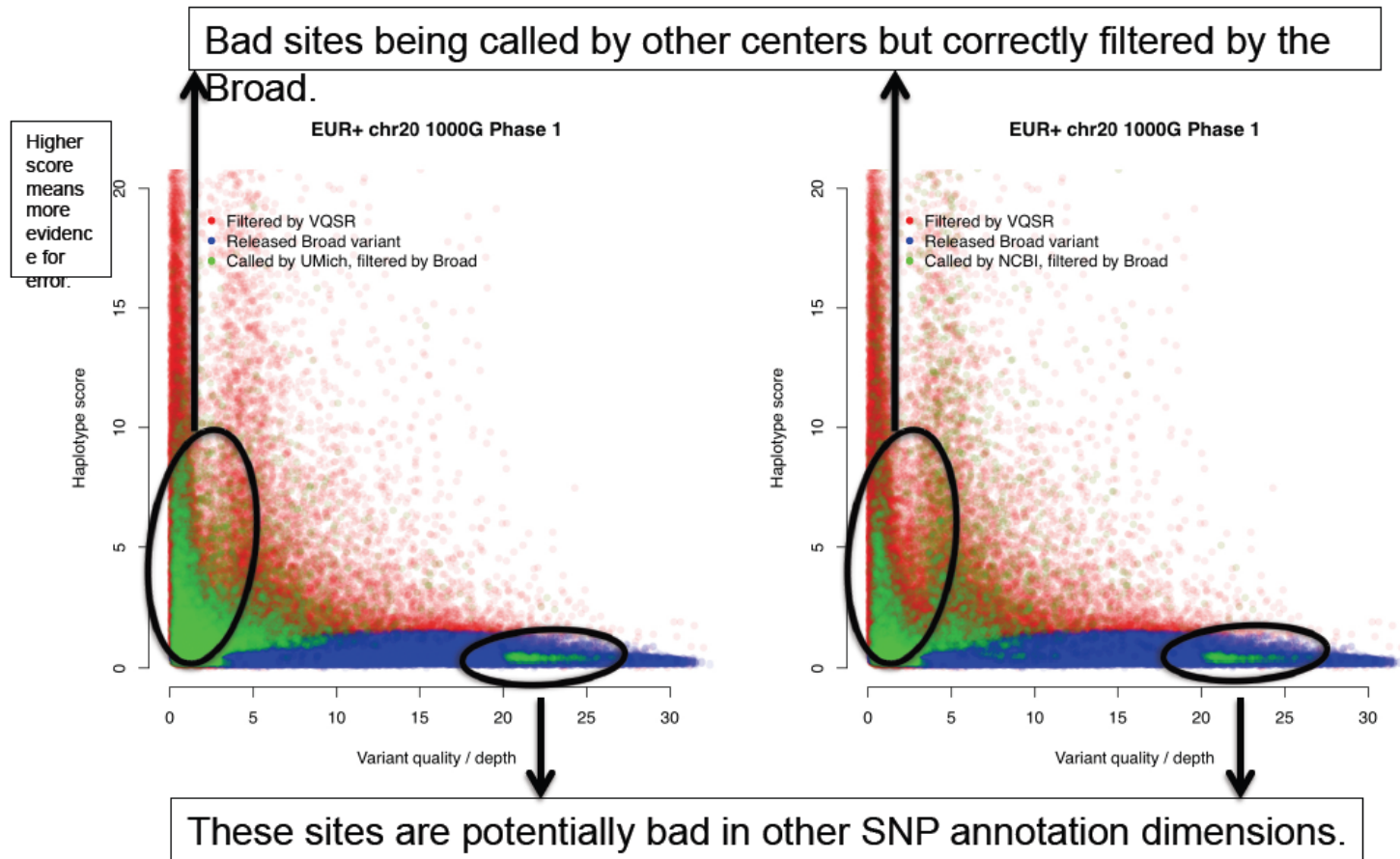


InbreedingCoeff vs FS



Adding the InbreedingCoefficient in 19-sample dataset is a huge bonus and its effect permeates every plot

Variants with bad Haplotype Scores often exhibit good Ti/Tv ratios and are included in other centers' callsets, but are likely FPs



Ryan Poplin (Broad GATK Team)

IT infrastructure and system stability is critical for NGS

Run GATK Phase I At the step CountCovariates (w/o recalibration)

Log file has 36089 lines and wrapper still runs to the end w/o interruption

```
tork.ncifcrf.gov - PuTTY
INFO 04:01:15,662 TraversalEngine - chr11:116080962 7.14e+08 78.2 m 6.6 s 62.4% 2.1 h 47.1 m
INFO 04:01:45,670 TraversalEngine - chr11:123910647 7.18e+08 78.7 m 6.6 s 62.7% 2.1 h 46.9 m
INFO 04:02:15,671 TraversalEngine - chr11:13449761 7.23e+08 79.2 m 6.6 s 63.0% 2.1 h 46.5 m
INFO 04:02:45,713 TraversalEngine - chr12:7476606 7.26e+08 79.7 m 6.6 s 63.3% 2.1 h 46.3 m
INFO 04:03:15,716 TraversalEngine - chr12:15483048 7.29e+08 80.2 m 6.6 s 63.5% 2.1 h 46.1 m
INFO 04:03:45,718 TraversalEngine - chr12:28463141 7.34e+08 80.7 m 6.6 s 63.9% 2.1 h 45.5 m
INFO 04:04:15,747 TraversalEngine - chr12:43896599 7.39e+08 81.2 m 6.6 s 64.4% 2.1 h 44.8 m
INFO 04:04:45,756 TraversalEngine - chr12:52501670 7.42e+08 81.7 m 6.6 s 64.7% 2.1 h 44.5 m
INFO 04:05:15,773 TraversalEngine - chr12:57554366 7.45e+08 82.2 m 6.6 s 64.9% 2.1 h 44.5 m
INFO 04:06:15,782 TraversalEngine - chr12:99488666 7.59e+08 83.7 m 6.6 s 66.2% 2.1 h 42.7 m
INFO 04:13:15,947 TraversalEngine - chr14:50844872 8.21e+08 90.2 m 6.6 s 72.7% 2.1 h 33.9 m
INFO 04:13:45,949 TraversalEngine - chr14:61181338 8.25e+08 90.7 m 6.6 s 73.0% 2.1 h 33.5 m
INFO 04:14:15,953 TraversalEngine - chr14:71801604 8.29e+08 91.2 m 6.6 s 73.4% 2.1 h 33.1 m
INFO 04:14:46,002 TraversalEngine - chr14:81744883 8.33e+08 91.7 m 6.6 s 73.7% 2.1 h 32.7 m
INFO 04:15:16,020 TraversalEngine - chr14:94753142 8.38e+08 92.2 m 6.6 s 74.1% 2.1 h 32.2 m
INFO 04:15:46,022 TraversalEngine - chr14:104645266 8.43e+08 92.7 m 6.6 s 74.4% 2.1 h 31.8 m
INFO 04:16:16,023 TraversalEngine - chr15:26841125 8.47e+08 93.2 m 6.6 s 75.4% 2.1 h 30.4 m
INFO 04:16:46,025 TraversalEngine - chr15:37980267 8.51e+08 93.7 m 6.6 s 75.8% 2.1 h 30.0 m
INFO 04:17:16,078 TraversalEngine - chr15:44089392 8.54e+08 94.2 m 6.6 s 76.0% 2.1 h 29.8 m
INFO 04:17:46,082 TraversalEngine - chr15:52899970 8.57e+08 94.7 m 6.6 s 76.2% 2.1 h 29.5 m
```


Broad discovered the most variants at very high quality levels in 1000G chr20 bake-off exercise

# samples	Center	Total # variants	dbSNP % (129)	# knowns	Known ti/tv	# novels	Novel ti/tv	Includes genotype refinement?
1004	Broad	765,365	24.82	190,000	2.36	575,365	2.37	No
1004	BC	733,155	25.34	185,787	2.37	547,368	2.32	No
1004	Sanger	728,374	25.31	184,341	2.36	544,033	2.36	No
1004	UMich	721,250	26.46	190,871	2.33	530,379	2.35	Yes
1004	Oxford	660,024	27.44	181,095	2.38	478,929	2.38	Yes
1004	BCM	605,274	29.98	181,444	2.33	423,830	2.29	Yes
1004	NCBI	601,907	29.26	176,150	2.39	425,757	2.57	No

Ryan Poplin (Broad GATK Team)

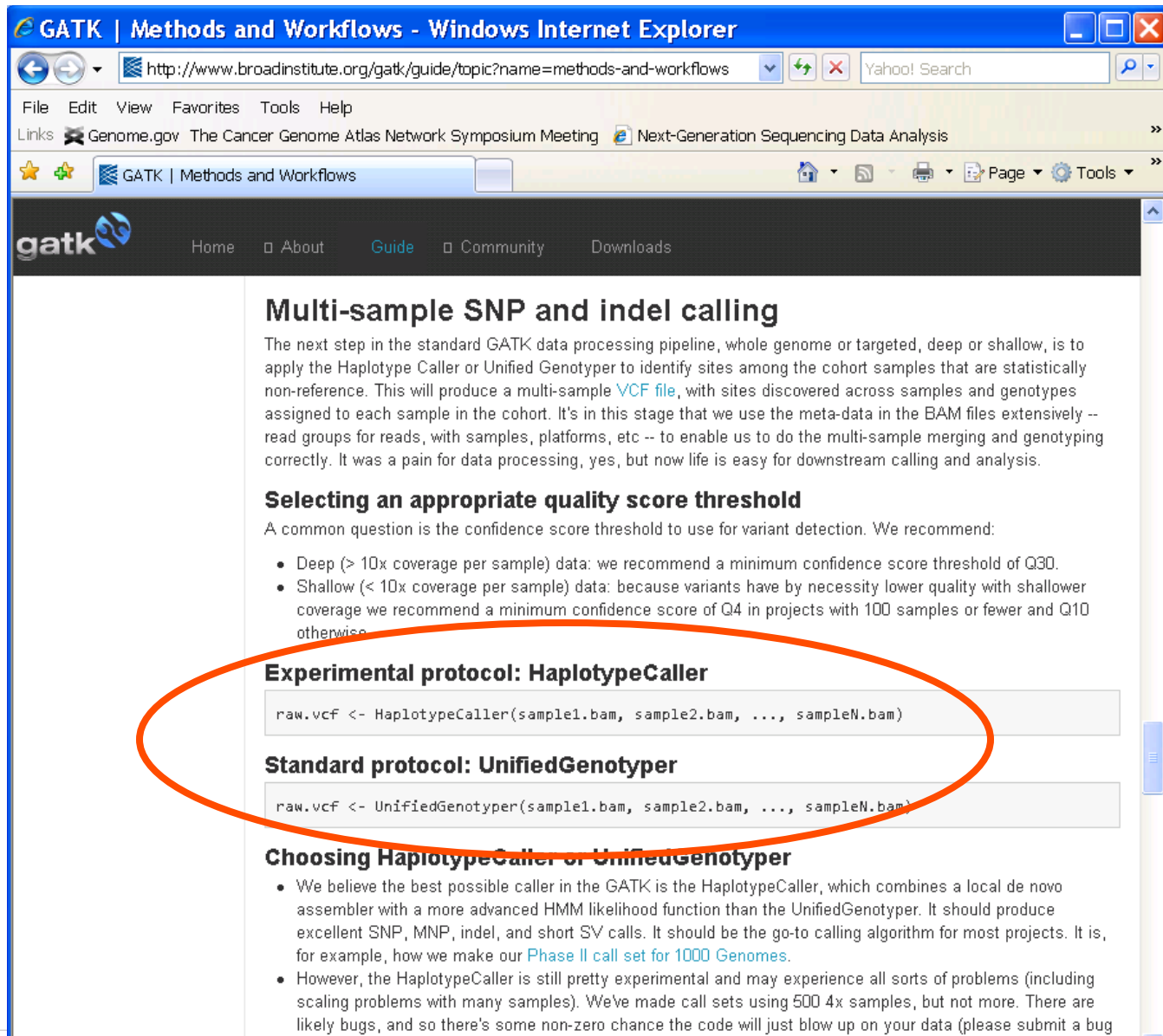
Summary of GATK Testing

- **GATK is a great tool with a lot of thoughts and strategies behind the scene. Not just do the job, but desire to do a great job for high quality SNP discovery.**
- **All the steps account for the quality of the final SNP call set, although VQSR looking most promising.**
- **More samples used for SNP calls in general help the quality**
- **Steps within its modularized procedure may be used to combine with or “help” other SNP discovery tools**
- **VQSR is only good for indel of whole Genome Shotgun Experiment and hand filtering is recommended for exome indels.**

New GATK (v2.0) Website-Documentation

The screenshot shows a Mozilla Firefox browser window displaying the GATK documentation index. The browser's address bar shows the URL www.broadinstitute.org/gatk/gatkdocs/. The website header includes the GATK logo and navigation links: Home, About, Guide, Community, and Downloads. The main content area is titled "GATK documentation index 2.0-35-g2d70733" and features a sidebar on the left with a "Guide" section containing links to "Guide Index", "Introductory Materials", "Technical Documentation" (highlighted), "Methods and Workflows", "Best Practices", "FAQs", "Tutorials", and "Videos". The main content area lists various tool categories in a vertical stack of boxes: "BAM Processing and Analysis Tools", "Cancer-specific Variant Discovery Tools", "Companion Utilities", "GATK Engine", "Quality Control and Simple Analysis Tools", "Read filters", "Reference ordered data (ROD) codecs", "User exceptions", "Validation Utilities", "VariantAnnotator annotations", "Variant Discovery Tools", and "Variant Evaluation and Manipulation Tools". At the bottom of the main content area, there is a footer with the text "See also [Documentation index](#) | [GATK Site](#) | [GATK support forum](#)" and "GATK version 2.0-35-g2d70733 built at 2012/08/03 15:13:40." The footer of the website includes the Broad Institute logo and the text "© Broad Institute 2012".

New GATK (v2.0) Website-New Raw SNV Callers



GATK | Methods and Workflows - Windows Internet Explorer

http://www.broadinstitute.org/gatk/guide/topic?name=methods-and-workflows

File Edit View Favorites Tools Help

Links Genome.gov The Cancer Genome Atlas Network Symposium Meeting Next-Generation Sequencing Data Analysis

GATK | Methods and Workflows

gatk Home About Guide Community Downloads

Multi-sample SNP and indel calling

The next step in the standard GATK data processing pipeline, whole genome or targeted, deep or shallow, is to apply the Haplotype Caller or Unified Genotyper to identify sites among the cohort samples that are statistically non-reference. This will produce a multi-sample VCF file, with sites discovered across samples and genotypes assigned to each sample in the cohort. It's in this stage that we use the meta-data in the BAM files extensively -- read groups for reads, with samples, platforms, etc -- to enable us to do the multi-sample merging and genotyping correctly. It was a pain for data processing, yes, but now life is easy for downstream calling and analysis.

Selecting an appropriate quality score threshold

A common question is the confidence score threshold to use for variant detection. We recommend:

- Deep (> 10x coverage per sample) data: we recommend a minimum confidence score threshold of Q30.
- Shallow (< 10x coverage per sample) data: because variants have by necessity lower quality with shallower coverage we recommend a minimum confidence score of Q4 in projects with 100 samples or fewer and Q10 otherwise.

Experimental protocol: HaplotypeCaller

```
raw.vcf <- HaplotypeCaller(sample1.bam, sample2.bam, ..., sampleN.bam)
```

Standard protocol: UnifiedGenotyper

```
raw.vcf <- UnifiedGenotyper(sample1.bam, sample2.bam, ..., sampleN.bam)
```

Choosing HaplotypeCaller or UnifiedGenotyper

- We believe the best possible caller in the GATK is the HaplotypeCaller, which combines a local de novo assembler with a more advanced HMM likelihood function than the UnifiedGenotyper. It should produce excellent SNP, MNP, indel, and short SV calls. It should be the go-to calling algorithm for most projects. It is, for example, how we make our [Phase II call set for 1000 Genomes](#).
- However, the HaplotypeCaller is still pretty experimental and may experience all sorts of problems (including scaling problems with many samples). We've made call sets using 500 4x samples, but not more. There are likely bugs, and so there's some non-zero chance the code will just blow up on your data (please submit a bug

Unified Genotyper works phenomenally well

- SNPs
 - > 98.5% confirmation rate for variation discovery in 1100 4x samples in 1000G
 - At least for “easy” sites in the genome
 - 98% of singletons in 2500 deep exomes
 - 78/79 *de novo* SNPs confirmed in Autism trios
- Indels
 - 1000G validation underway, unknown confirmation rate
 - Significant false negative rates for large events, especially large insertions
 - E.g., ~50% false negative rate for large (>15 bp) indels
 - Indel calling is the future challenge!

Ryan Poplin (Broad GATK Team)

Phase 2 Data Processing Overview

- Baseline production release (Khalid Shakir)
 - Whole genome and whole exome SNP / indel site list using current Broad best practices for baseline comparisons
- Methods development for improving indels
 - Calibrated indel model parameters (Mauricio Carneiro)
 - Updated BQSR to calculate empirically accurate base insertion and base deletion quality scores for use in indel models
 - Haplotype caller (Ryan Poplin)
 - Call SNPs and indels simultaneously via local de-novo assembly
 - Updated exact model (Eric Banks)
 - Generalized mathematical formulation for genotyping multi-allelic SNPs and indels
 - Updated VQSR (Chris Hartl)
 - GMM + Random forest greatly outperforms on indel callsets

Ryan Poplin (Broad GATK Team)

Contrasting indel calling workflows

Unified Genotyper

Propose
Haplotypes

Look for coincident events in the read data. Must be seen at least 5 times.



Evaluate
Haplotypes

Pair HMM evaluates the likelihood of the proposed event with the reference. Affine gap penalties based on homopolymer context.



Assign
Genotypes

“Exact model” from Heng Li chooses optimal configuration.

Haplotype Caller

Local de novo assembly via DeBruijn graphs. Paths through graph are weighted by number reads which support each kmer.

Same Pair HMM chooses the best two haplotypes which explain the read data. Gap penalties derived from data per read group via new Indel Quality Score Recalibrator.

No change from UG.

Ryan Poplin (Broad GATK Team)

Haplotype Caller greatly increases sensitivity to larger indel events over the Unified Genotyper

Caller	Mullikin		Mills	
	Variant Sensitivity (strict)	Genotype Concordance (strict)	Variant Sensitivity (strict)	Genotype Concordance (strict)
Unified Genotyper	51.9% (40 / 77)	51.9% (40 / 77)	49.0% (97 / 198)	49.0% (97 / 198)
Haplotype Caller	90.9% (70 / 77)	89.6% (69 / 77)	81.8% (162 / 198)	81.8% (162 / 198)

- Input data is NA12878 b37+decoy WGS HiSeq high coverage
- Sites chosen to be very difficult (het) but high confidence in being real (require family transmission)
- Evaluation sets
 - Mullikin Fosmids and Mills et al, GR, 2011 (2x hit, double center)
 - Large events (> 15 bp), largest is 106bp (which we don't yet call)

Ryan Poplin (Broad GATK Team)

Tool By Tool Highlighting Major Aspects of Practical Usage

- GATK
- SAMtools
- VarScan
- CLCBio
- CASAVA
- Partek Genomic Suite

Samtools: A Variant Discovery Tool from Sanger Institute

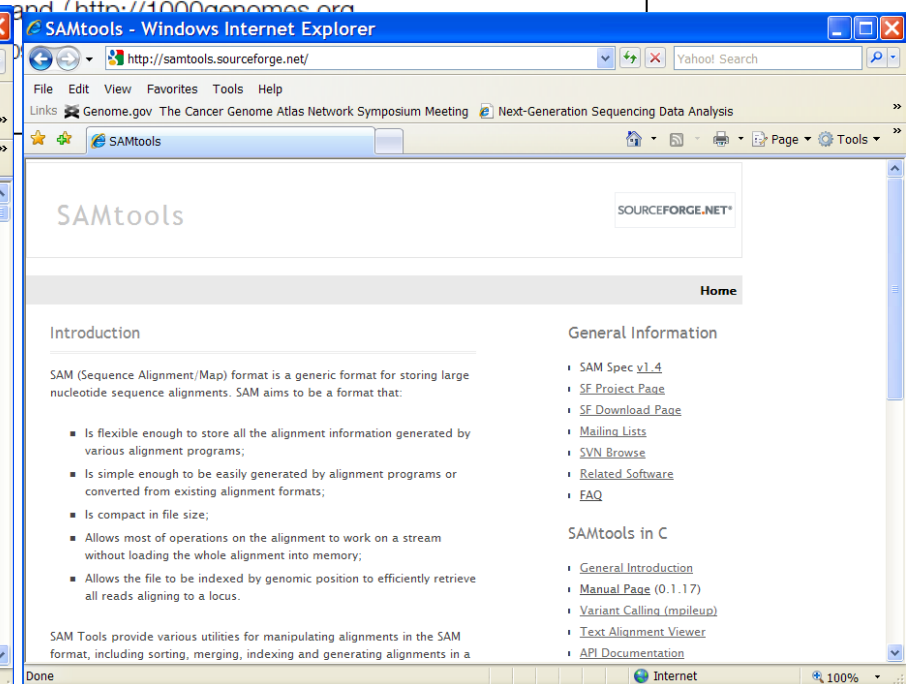
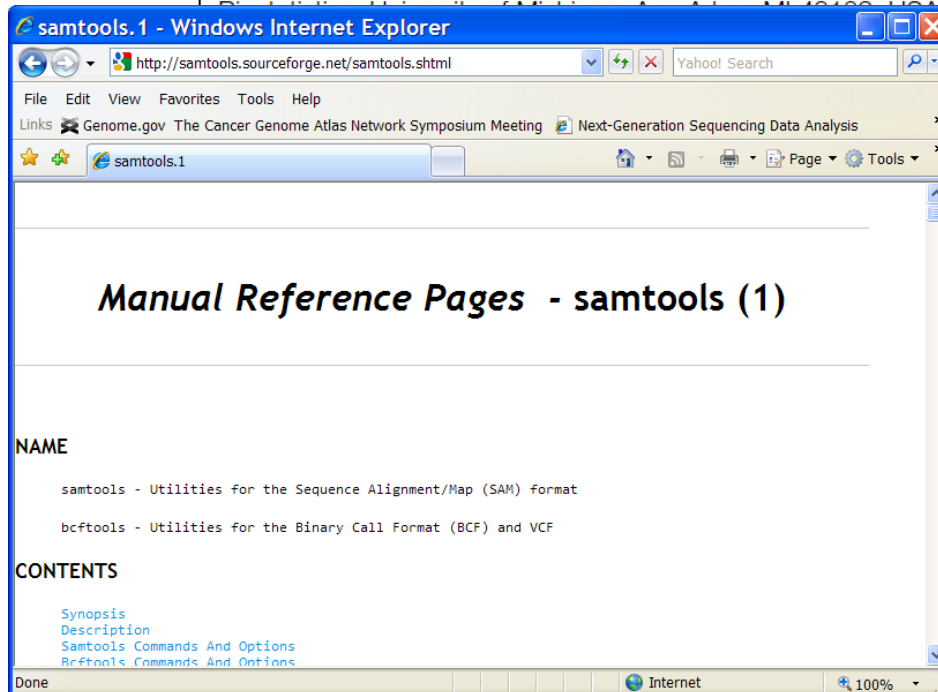
BIOINFORMATICS APPLICATIONS NOTE Vol. 25 no. 16 2009, pages 2078–2079
doi:10.1093/bioinformatics/btp352

Sequence analysis

The Sequence Alignment/Map format and SAMtools

Heng Li^{1,†}, Bob Handsaker^{2,†}, Alec Wysoker², Tim Fennell², Jue Ruan³, Nils Homer⁴, Gabor Marth⁵, Goncalo Abecasis⁶, Richard Durbin^{1,*} and 1000 Genome Project Data Processing Subgroup⁷

¹Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Cambridge, CB10 1SA, UK, ²Broad Institute of MIT and Harvard, Cambridge, MA 02141, USA, ³Beijing Institute of Genomics, Chinese Academy of Science, Beijing 100029, China, ⁴Department of Computer Science, University of California Los Angeles, Los Angeles, CA 90095, ⁵Department of Biology, Boston College, Chestnut Hill, MA 02467, ⁶Center for Statistical Genetics, Department of Biostatistics, Harvard School of Public Health, Boston, MA 02115, USA and ⁷<http://1000genomes.org>



Overview of samtools procedure

Resource For Samtools:

Samtools commands used are based on samtools documentations:

Version 1 at: <http://samtools.sourceforge.net/mpileup.shtml>

```
samtools mpileup -uf ref.fa aln1.bam aln2.bam | bcftools view -bvcg - > var.raw.bcf  
bcftools view var.raw.bcf | vcfutils.pl varFilter -D100 > var.flt.vcf
```

also info from URL:<http://samtools.sourceforge.net/samtools.shtml>

Version 2 at: <http://samtools.sourceforge.net/samtools.shtml>

on example: Call SNPs and short indels for multiple diploid individuals:

```
samtools mpileup -P ILLUMINA -ugf ref.fa *.bam | bcftools view -bvcg - > var.raw.bcf  
bcftools view var.raw.bcf | vcfutils.pl varFilter -D 2000 > var.flt.vcf
```

using `-d` option for filtering

```
bcftools view var.raw.bcf | vcfutils.pl varFilter -d 10 > var.flt.vcf
```

Call Samtools Sample Options:

- Call samtools with individual sample (bam file)
- Call samtools with all samples altogether (bam files)

Samtools SNP Filtering Options:

- Filtering with `-d 10` (`-d`: minimum read depth [2])
- Filtering `-D 2000` (`-D`: maximum read depth [10000000])

Write a wrapper program to loop the samples and/or connect steps

Samtools Steps: Action commands for samples altogether

Samtools mpile pipe into bcftools view for all samples altogether

```
samtools mpileup -ugf /Path/hg19_chrM_1st.fa  
/Path/S1.bam  
/Path/S2.bam  
/Path/S3.bam  
.....  
/Path/S19.bam | bcftools view -bcvg - > /Path/samtools_mpileup_AllSamples_snps.raw.bcf
```

bcftools/varFilter filtered by `-D 2000`:



```
bcftools view /Path/samtools_mpileup_AllSamples_snps.raw.bcf | vcfutils.pl varFilter -D 2000 > /Path/  
samtools_mpileup_AllSamples_snps.raw.fild2000.vcf
```

bcftools/varFilter filtered by `-d 10`:

Or

```
bcftools view /Path/samtools_mpileup_AllSamples_snps.raw.bcf | vcfutils.pl varFilter -d 10 > /Path/  
samtools_mpileup_AllSamples_snps.raw.fild10.vcf
```

bcftools/varFilter filtered by `-d` and `-D`:

Or

```
bcftools view /Path/samtools_mpileup_AllSamples_snps.raw.bcf | vcfutils.pl varFilter -d 10 -D 2000  
> /Path/samtools_mpileup_AllSamples_snps.raw.fild10D2000.vcf
```

No Filtering:

Or

```
bcftools view /Path/samtools_mpileup_AllSamples_snps.raw.bcf > /Path/  
samtools_mpileup_AllSamples_snps.raw.Nofil.vcf
```

Samtools Steps: Action commands for individual sample

Samtools mpile for individual sample

```
samtools mpileup -ugf /Path/hg19_chrM_1st.fa  
/Path/S1.bam | bcftools view -bcvg - > /Path/samtools_mpileup_S1_snps.raw.bcf
```



bcftools/varFilter filtered by -D 2000:

```
bcftools view /Path/samtools_mpileup_S1_snps.raw.bcf | vcfutils.pl varFilter -D 2000 > /Path/  
samtools_mpileup_S1_snps.raw.fild2000.vcf
```

bcftools/varFilter filtered by -d 10:

Or

```
bcftools view /Path/samtools_mpileup_S1_snps.raw.bcf | vcfutils.pl varFilter -d 10 > /Path/  
samtools_mpileup_S1_snps.raw.fild10.vcf
```

bcftools/varFilter filtered by -d and -D: Or

```
bcftools view /Path/samtools_mpileup_S1_snps.raw.bcf | vcfutils.pl varFilter -d 10 -D 2000 > /Path/  
samtools_mpileup_S1_snps.raw.fild10D2000.vcf
```

No Filtering: Or

```
bcftools view /Path/samtools_mpileup_S1_snps.raw.bcf > /Path/  
samtools_mpileup_S1_snps.raw.Nofil.vcf
```

Loop
to the
next
bam
file

Multiple-sample SNP calling procedure enhances the power for calling SNPs between samples but reduced the power for singleton SNPs -Heng Li

Use GATK SelectVariants to Select out SNPs and Indels within the target interval regions

Select out only SNPs:

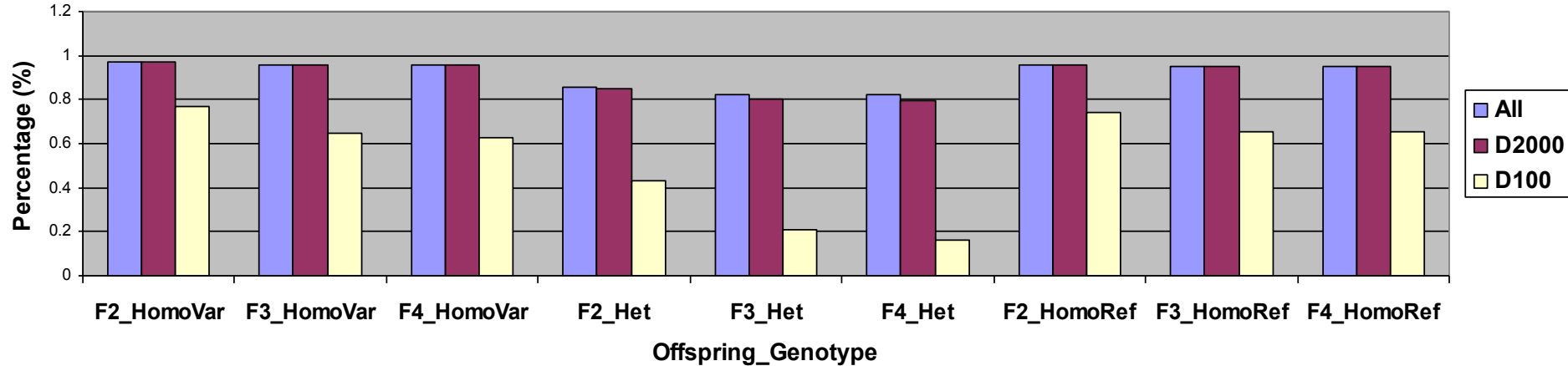
```
java -jar /Path/GenomeAnalysisTK-1.6-7-g2be5704/bin/GenomeAnalysisTK.jar
-T SelectVariants -R /Path/hg19_chrM_1st.fa
-L /Path/Exome_Target_Region_SureSelect_AllExon_50mb_Bedfiles/029720_D_BED_20101013.bed
--variant /Path/samtools_mpileup_AllSamples_snps.raw.fild10.vcf
-selectType SNP
-o /Path/samtools_mpileup_AllSamples_snps.raw.fild10_SelSNP.vcf
```

Select out only Indels:

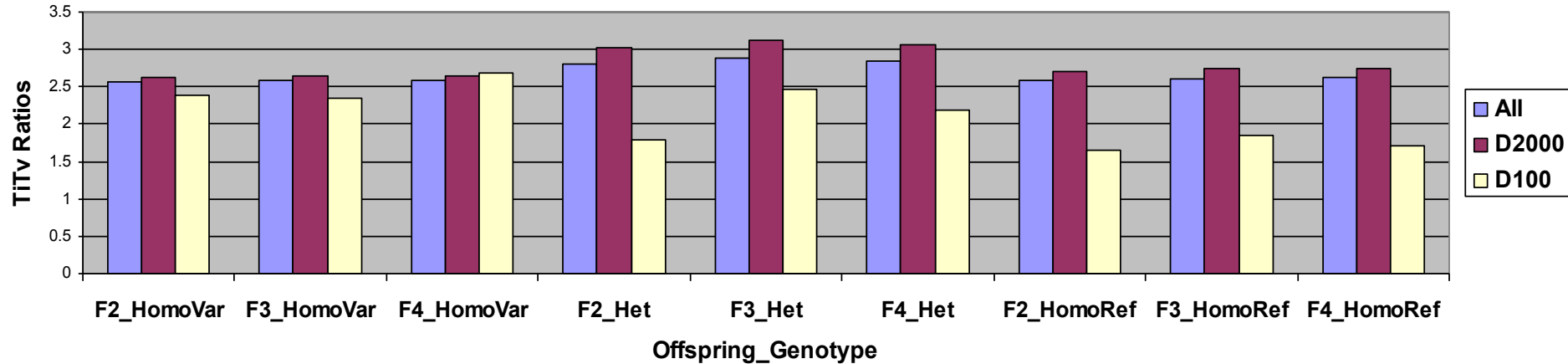
```
java -jar /Path/GenomeAnalysisTK-1.6-7-g2be5704/bin/GenomeAnalysisTK.jar
-T SelectVariants -R /Path/hg19_chrM_1st.fa
-L /Path/Exome_Target_Region_SureSelect_AllExon_50mb_Bedfiles/029720_D_BED_20101013.bed
--variant /Path/samtools_mpileup_AllSamples_snps.raw.fild10.vcf
-selectType INDEL
-o /Path/samtools_mpileup_AllSamples_snps.raw.fild10_SelIndel.vcf
```

Samtools Call Sets: Filtering not necessarily help!

Offspring_Hits_Percentage_in_Parent_Common_Variant



TiTv Ratios in the Hit Call Sets



All: All variant call; D2000: Filtered by -D 2000; D100: Filtered by -D 100

Tool By Tool Highlighting Major Aspects of Practical Usage

- GATK
- Samtools
- **VarScan**
- CLCBio
- CASAVA
- Partek Genomic Suite

VarScan: A Variant Discovery Tool from WashU

BIOINFORMATICS APPLICATIONS NOTE Vol. 25 no. 17 2009, pages 2283–2285
doi:10.1093/bioinformatics/btp373

Sequence analysis

VarScan: variant detection in massively parallel sequencing of individual and pooled samples

Daniel C. Koboldt*, Ken Chen, Todd Wylie, Elaine R. Mardis, George M. Weinstock, Richard

The Genome Center at Washington University School of
Received on April 16, 2009; revised on June 11, 2009; accepted on June 19, 2009
Advance Access publication June 19, 2009

Associate Editor: Dmitrij Frishman



VarScan 2: Somatic mutation and copy number alteration discovery in cancer by exome sequencing

Daniel C. Koboldt, Qunyuan Zhang, David E. Larson, et al.

Genome Res. 2012 22: 568-576 originally published online February 2, 2012
Access the most recent version at doi:10.1101/gr.129684.111

VarScan - Variant Detection in Massively Parallel Sequencing Data

http://varscan.sourceforge.net/

File Edit View Favorites Tools Help

Links Genome.gov The Cancer Genome Atlas Network Symposium meeting Next-Generation Sequencing Data Analysis

VarScan - Variant Detection in Massively Parallel Sequencing Data

VarScan
variant detection in massively parallel sequencing data

Home Project Download Manual FAQ Documentation

VarScan

- About/Wiki
- Project Page
- User's Manual
- JavaDoc
- Download Now
- Support/FAQ

VarScan Overview

The advent of massively parallel sequencing technologies has fundamentally changed the study of genetics. New platforms like the Illumina HiSeq2000 yield unprecedented levels of sequencing throughput. The analysis and interpretation of data from next-generation sequencing (NGS) platforms presents a substantial informatics challenge. VarScan is a platform-independent software tool developed at the Genome Institute at Washington University to detect variants in NGS data.

Sequencing Platforms and Variant Types
Why Use VarScan?
Installing VarScan
Citing VarScan

Overview of VarScan procedure

Resource For VarScan:

Now at Version 2.2.8 at: <http://varscan.sourceforge.net/>

Written in Java and run on any operating system (Linux, Unix, Mac OSX, Windows)

Command line driven, cutoff choice

Use samtools mpileup for variant calling

May need write wrapper

Write a wrapper program to loop the samples and/or connect steps

VarScan Steps: Action commands for samples altogether

Command to pipe the samtools mpileup result into VarScan mpileup2snp:

```
samtools mpileup -f /Path/hg19_chrM_1st.fa
-q 10
-E /Path/S1.bam /Path/S2.bam /Path/S3.bam ...../Path/S19.bam
| java -jar /Path/VarScan_v2.2.8/bin/VarScan.v2.2.8.jar mpileup2snp
--output-vcf 1
--min-coverage 4
--min-var-freq 0.20
--p-value 0.05 > /Path/SNPs_p0_05/
VarScan_mpileup_AllSamples_snps.raw_VarScanAuthorCutoff.vcf
```

- E: extended BAQ for higher sensitivity but lower specificity
- E parameter will become the default in future SAMtools releases

Parameter Options:

- Author suggested as above
- Others used: `mpileup2snp --output-vcf 1 --min-coverage 10 --min-avg-qual 20 --min-var-freq 0.25 --p-value 1e-06`
(BMC Bioinformatics 2011, 12:267)

Note: Problem found in v2.2.8 vcf format issue at column “ID” in vcf file and sample names not in vcf files (inconvenience)

Check samtools mpileup options

```
tork.ncifcrf.gov - PuTTY
torqv:~> samtools mpileup

Usage: samtools mpileup [options] in1.bam [in2.bam [...]]

Input options:

    -6          assume the quality is in the Illumina-1.3+ encoding
    -A          count anomalous read pairs
    -B          disable BAQ computation
    -b FILE     list of input BAM files [null]
    -C INT      parameter for adjusting mapQ; 0 to disable [0]
    -d INT      max per-BAM depth to avoid excessive memory usage [250]
    -E          extended BAQ for higher sensitivity but lower specificity
    -f FILE     faidx indexed reference sequence file [null]
    -G FILE     exclude read groups listed in FILE [null]
    -l FILE     list of positions (chr pos) or regions (BED) [null]
    -M INT      cap mapping quality at INT [60]
    -r STR      region in which pileup is generated [null]
    -R          ignore RG tags
    -q INT      skip alignments with mapQ smaller than INT [0]
    -Q INT      skip bases with baseQ/BAQ smaller than INT [13]

Output options:

    -D          output per-sample DP in BCF (require -g/-u)
    -g          generate BCF output (genotype likelihoods)
    -O          output base positions on reads (disabled by -g/-u)
    -s          output mapping quality (disabled by -g/-u)
    -S          output per-sample strand bias P-value in BCF (require -g/-u)
    -u          generate uncompress BCF output

SNP/INDEL genotype likelihoods options (effective with '-g' or '-u'):

    -e INT      Phred-scaled gap extension seq error probability [20]
    -F FLOAT    minimum fraction of gapped reads for candidates [0.002]
    -h INT      coefficient for homopolymer errors [100]
    -I          do not perform indel calling
    -L INT      max per-sample depth for INDEL calling [250]
    -m INT      minimum gapped reads for indel candidates [1]
    -o INT      Phred-scaled gap open sequencing error probability [40]
```

Tool By Tool Highlighting Major Aspects of Practical Usage

- GATK
- SAMtools
- VarScan
- **CLCBio**
- CASAVA
- Partek Genomic Suite

NGS-based SNP Discovery Tools

- Atlas-SNP2 (Baylor). *Genome Res.* 2010,20(2):273-80
- SOAPsnp (BGI). *Bioinformatics* 2008, 24(5):713-4
- Crossbow (UM). *Nature Biotech* 2010, 28:691-693
- Bambino (NCI, Beutow). *Bioinformatics* 2011,5;27(6):865-6
- GigaBayes→FreeBayes (Boston College). *Nature Method* 2008, 5(2):183-8
- CLCbio Genomics Workbench (Commercial) (used v4.8 for the comparison)
- Genomatix Mining Station (GMS) (Commercial)
- Partek SNP tool in Genomics Suite (Commercial)
- Avadis NGS (Commercial)
- Illumina Casava (Commercial)
- SAMtools (Sanger Institute). *Bioinformatics* 2009, 25:2078-9
- VarScan (Washington Univ). *Bioinformatics* 2009; *Genome Res* 2012
- GATK (Broad Institute). *Genome Res* 2010; *Nature Genet* 2011
-

CLCbio solution for NGS data analysis

The screenshot shows a Windows Internet Explorer browser window displaying the CLCbio website. The address bar shows the URL <http://www.clcbio.com/index.php?id=1240>. The browser's menu bar includes File, Edit, View, Favorites, Tools, and Help. The address bar also contains a search box with the text "Yahoo! Search".

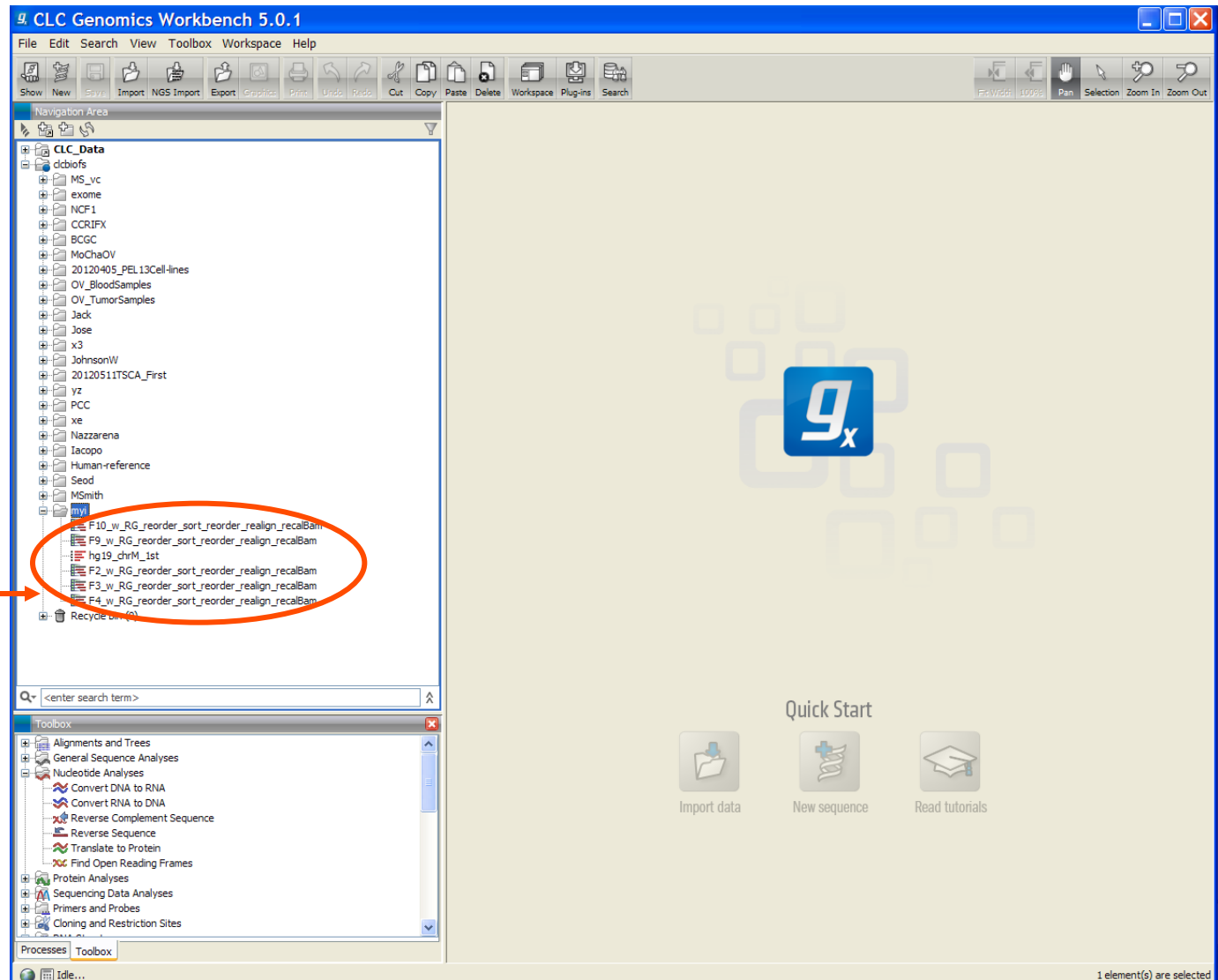
The website's header features the CLCbio logo and navigation links: Buy Online, Newsletter, and a Google Custom Search box. Below the header is a main navigation menu with the following items: Next Gen Sequencing, Enterprise Platform, Desktop Applications, Customized Solutions, and Contact.

The main content area is titled "CLC Genomics Workbench" and includes a "Request a quote" button, a "Buy" button, a "Download" button, and a "Presentation" button. The main heading is "Domingating the High-Throughput Sequencing data analysis challenge". Below this heading is a paragraph of text: "We have overcome the challenge to analyze High-Throughput Sequencing data faster than it is produced by implementing a SIMD-accelerated assembly algorithm in our Next Generation Sequencing solution, CLC Genomics Workbench - a cross-platform desktop application with a graphical user-interface." A "Share" button is located to the right of this paragraph.

Below the text is a circular logo with the text "GENOMICS" at the top, "TRANSCRIPTOMICS" at the bottom, and "CLC" in the center. To the right of the main content is a sidebar with a list of links: Front page, Solution Overview, All Downloads, Desktop software, Compare desktop applications, CLC Genomics Workbench, Genomics Gateway, Latest Genomics Workbench news, Product features, User manual, Latest improvements, Download a trial, CLC Main Workbench, CLC Sequence Viewer, CLC Developer Kit, Pricing and Licensing Options, System requirements, Enterprise solutions, High-Performance Computing, CLC Consulting Solutions, Science, Support, Corporate, and Contact.

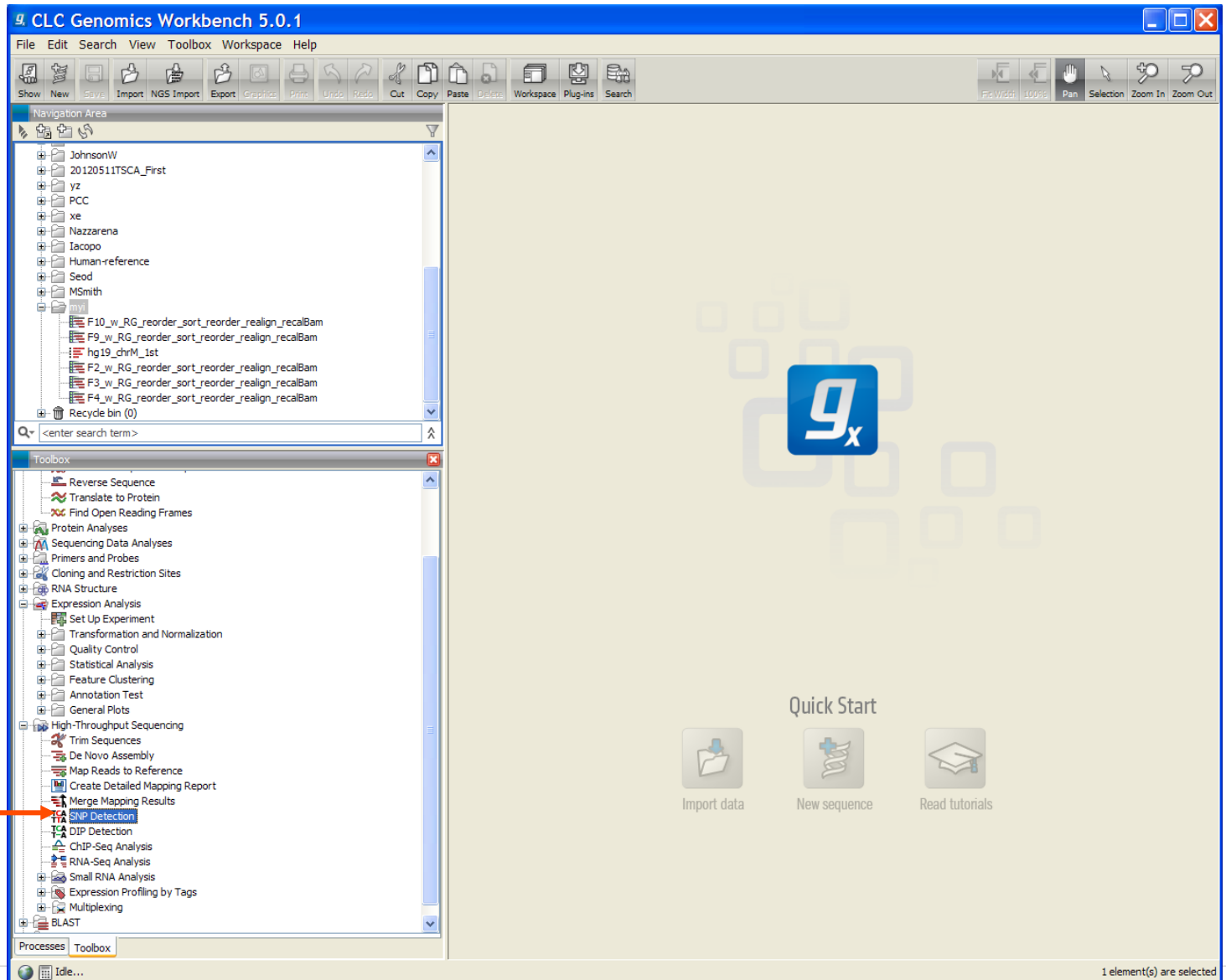
At the bottom of the page, there is a section titled "Some of the key Next Gen Sequencing applications of CLC Genomics Workbench are listed below" with a list of items, including "Genomics" and "Transcriptomics".

CLCbio Genomic Workbench Interface



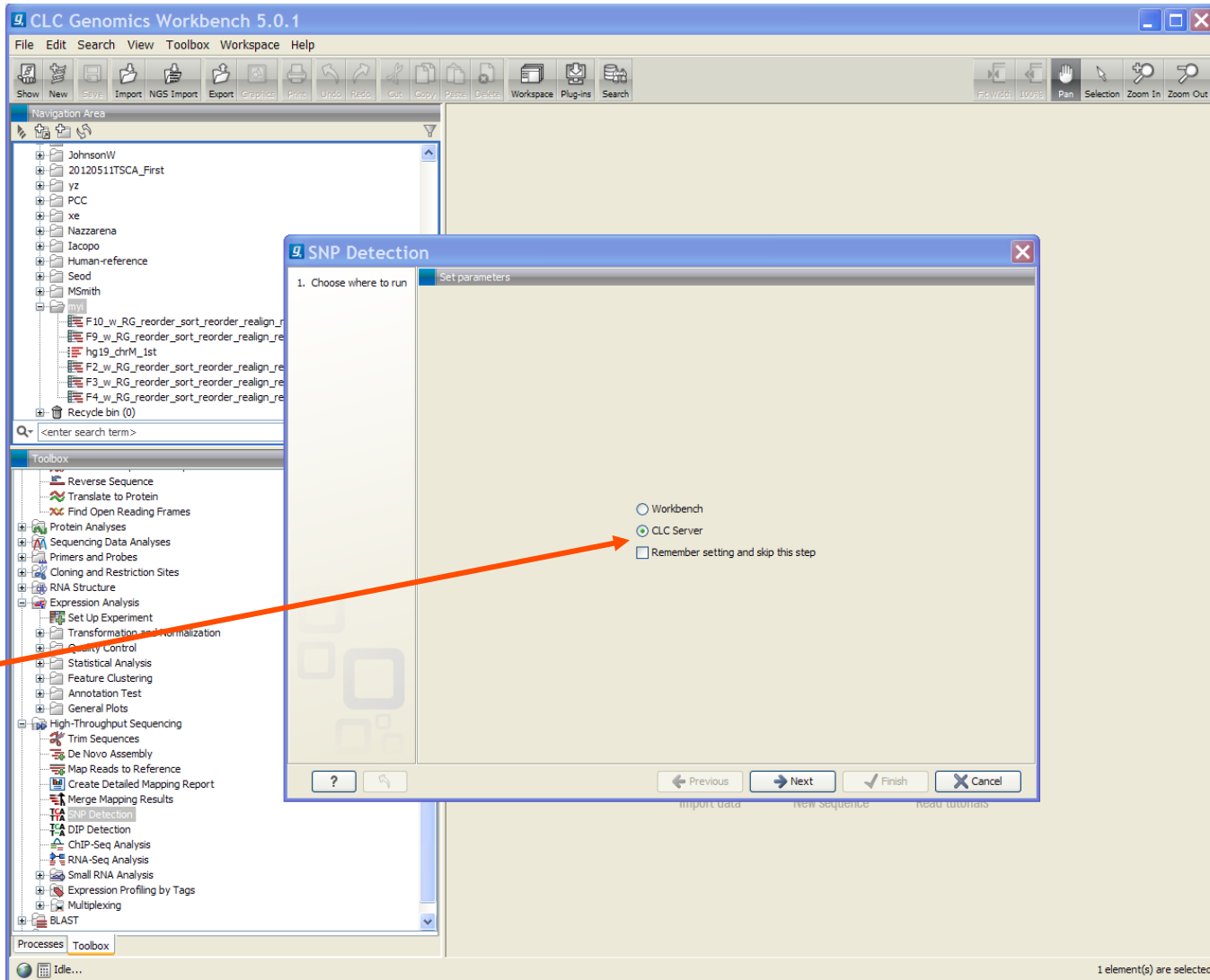
Need first to
import
Bam files
And reference
Into server

CLCbio SNP calling procedure

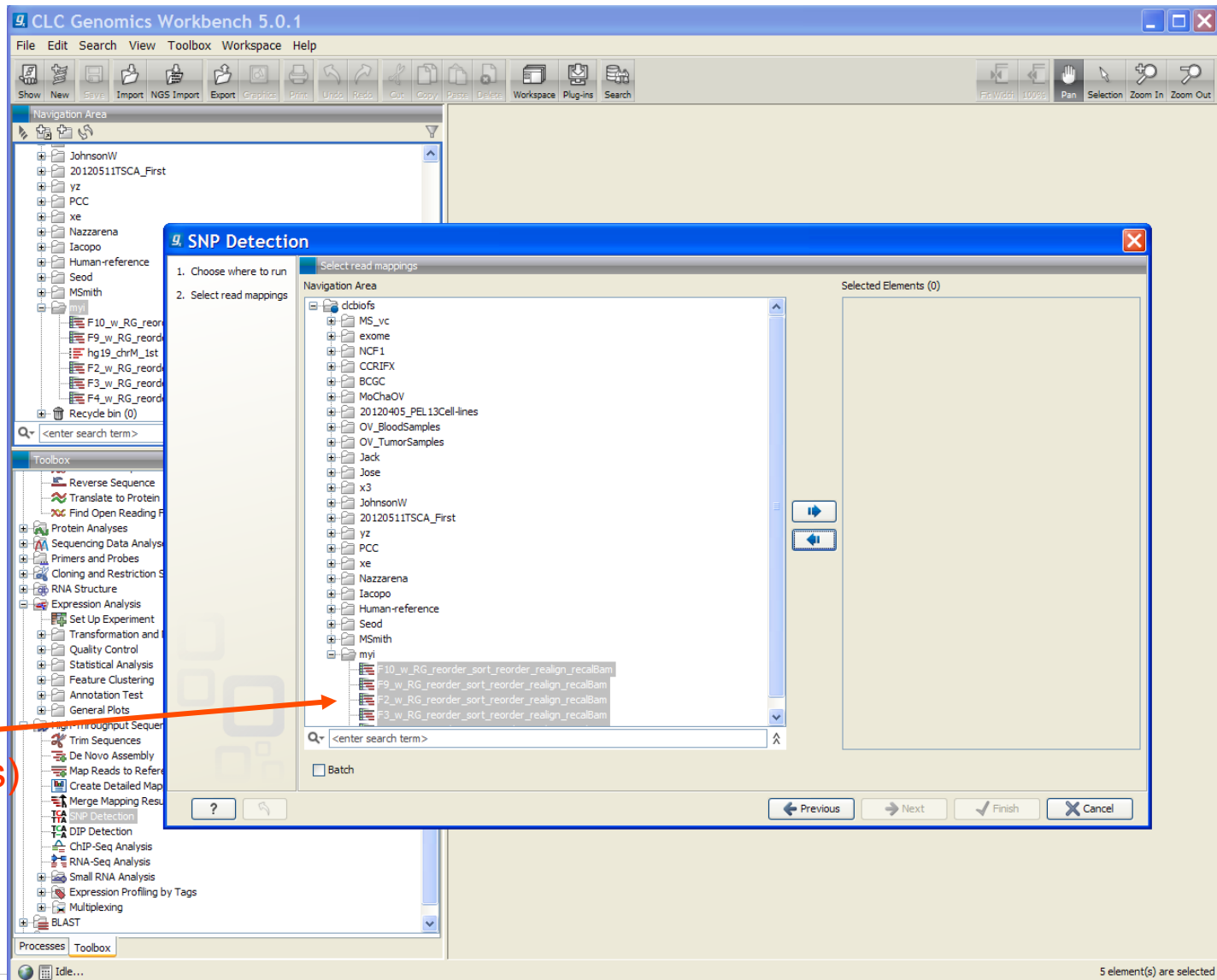


SNP detection
module

CLCbio SNP calling procedure

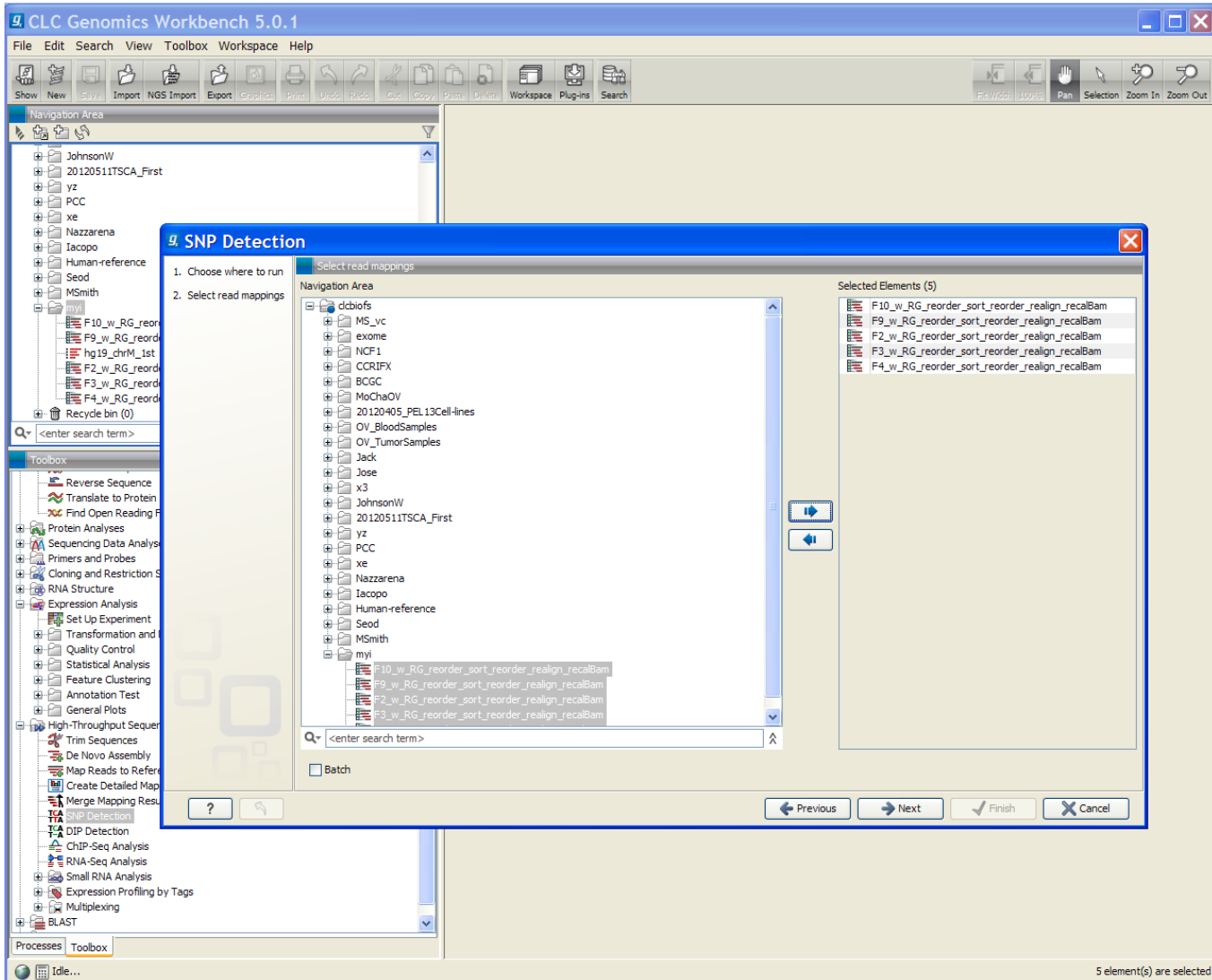


CLCbio SNP calling procedure



Select Samples' Bam file(s) For SNP calling

CLCbio SNP calling procedure



CLCbio SNP calling procedure

CLC Genomics Workbench 5.0.1

File Edit Search View Toolbox Workspace Help

Navigation Area

JohnsonW
20120511TSCA_First
yz
PCC
xe
Nazzarena
Iacopo
Human-reference
Seod
MSmith
myl
F10_w_RG_reord
F9_w_RG_reord
hg19_chrM_1st
F2_w_RG_reord
F3_w_RG_reord
F4_w_RG_reord
Recycle bin (0)

Toolbox

Reverse Sequence
Translate to Protein
Find Open Reading F
Protein Analyses
Cloning and Restriction S
RNA Structure
Expression Analysis
Set Up Experiment
Transformation and I
Quality Control
Statistical Analysis
Feature Clustering
Annotation Test
General Plots
High-Throughput Sequer
Trim Sequences
De Novo Assembly
Map Reads to Refer
Create Detailed Map
Merge Mapping Resu
SNP Detection
DIP Detection
ChIP-Seq Analysis
RNA-Seq Analysis
Small RNA Analysis
Expression Profiling by Tags
Multiplexing
BLAST

Processes Toolbox

Idle...

5 element(s) are selected

SNP Detection

1. Choose where to run
2. Select read mappings
3. Set SNP parameters

Set SNP parameters

Quality

Window length (must be odd) 11
Maximum number of gaps and mismatches 2
Minimum average quality of surrounding bases 15
Minimum quality of central base 20

Significance

Non-specific and low-quality matches are ignored during SNP detection.

Minimum coverage 4
Minimum variant frequency (%) 35.0

Advanced

Minimum paired coverage 0
Maximum coverage 25
Minimum variant count required 1 and sufficient 5

Ploidy

Maximum expected variations 2

? [Magnifying Glass] Previous Next Finish Cancel

Parameter
For
SNP
detection

CLCbio SNP calling procedure

CLC Genomics Workbench 5.0.1

File Edit Search View Toolbox Workspace Help

Navigation Area

- JohnsonW
- 20120511TSCA_First
- yz
- PCC
- ve
- Nazzarena
- Iacopo
- Human-reference
- Seod
- MSmith
- myl
 - F10_w_RG_reord
 - F9_w_RG_reord
 - hg19_chrM_1st
 - F2_w_RG_reord
 - F3_w_RG_reord
 - F4_w_RG_reord
- Recycle bin (0)

Search <center search term>

Toolbox

- Reverse Sequence
- Translate to Protein
- Find Open Reading F
- Protein Analyses
- Sequencing Data Analysis
- Primers and Probes
- Cloning and Restriction S
- RNA Structure
- Expression Analysis
- Set Up Experiment
- Transformation and
- Quality Control
- Statistical Analysis
- Feature Clustering
- Annotation Test
- General Plots
- High-Throughput Sequer
- Trim Sequences
- De Novo Assembly
- Map Reads to Refer
- Create Detailed Map
- Merge Mapping Resu
- SNP Detection
- DIP Detection
- ChIP-Seq Analysis
- RNA-Seq Analysis
- Small RNA Analysis
- Expression Profiling by Tags
- Multiplexing
- BLAST

Processes Toolbox

Idle...

5 element(s) are selected

SNP Detection

- Choose where to run
- Select read mappings
- Set SNP parameters
- Result handling

Result handling

Output options

- Annotate reference sequence(s)
- Annotate consensus sequence(s)
- Create table

Genetic code: 1 Standard

- Merge SNPs located within same codon

Result handling

- Open
- Save

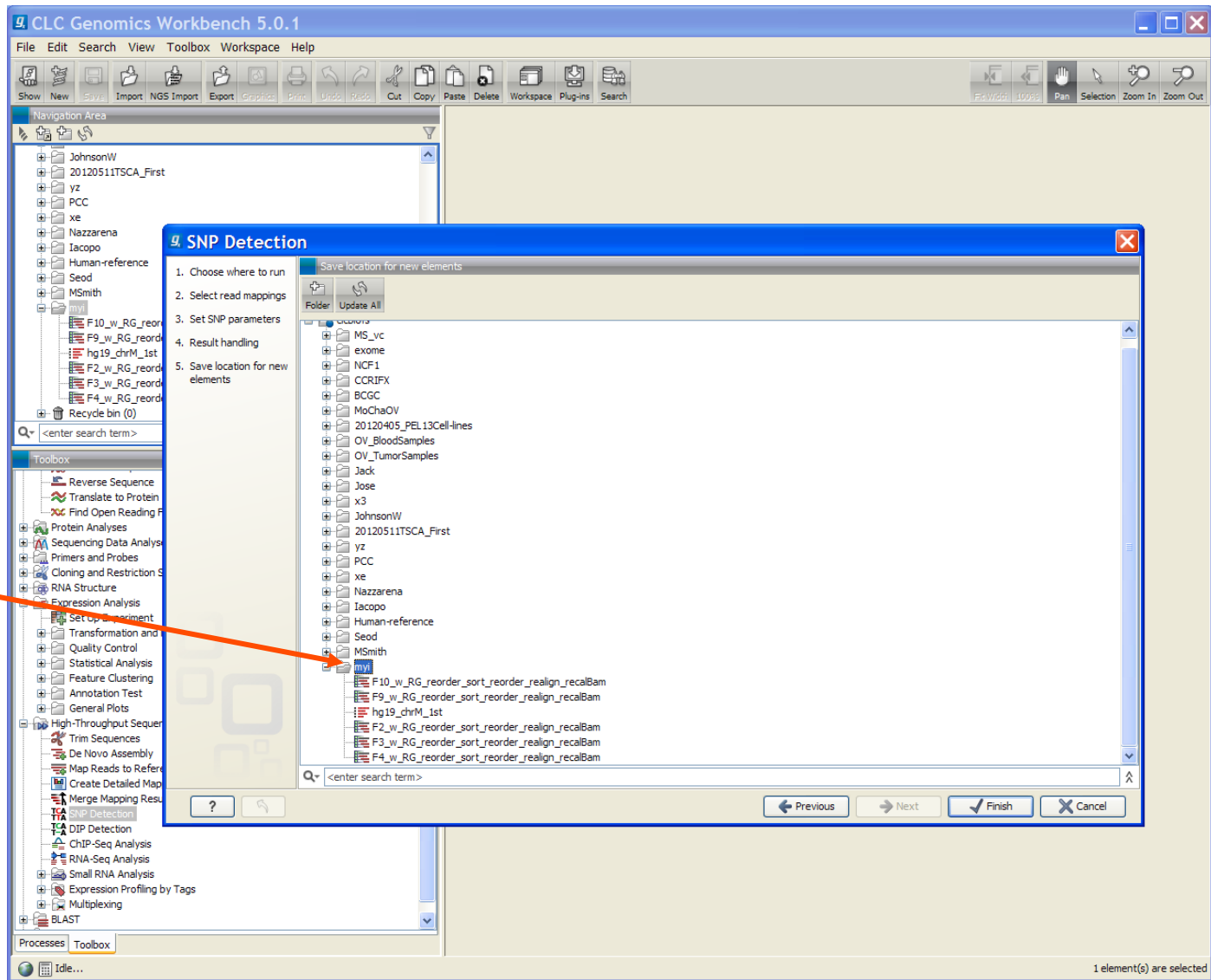
Log handling

- Make log

Previous Next Finish Cancel

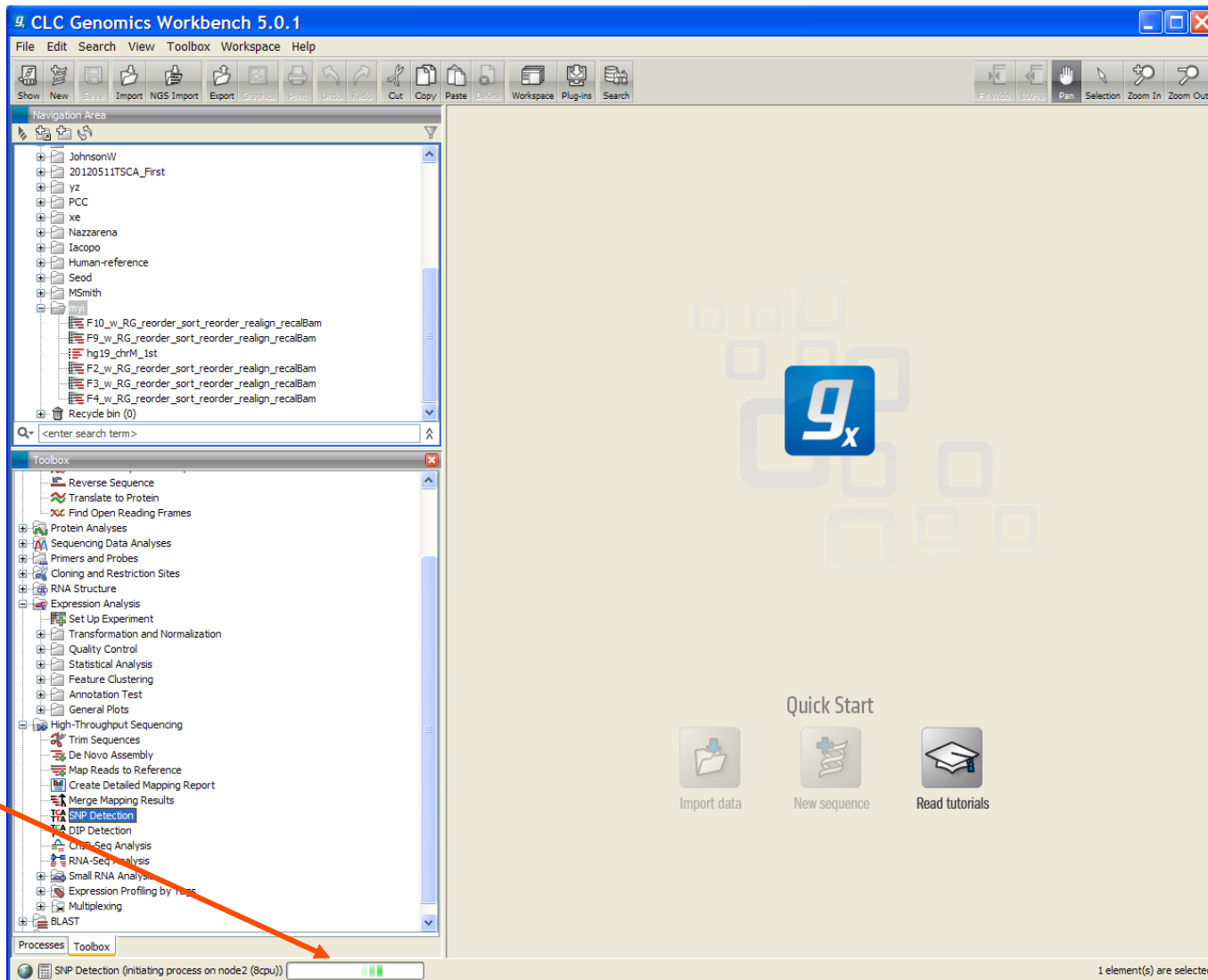
Still output
Not in
Vcf format

CLCbio SNP calling procedure

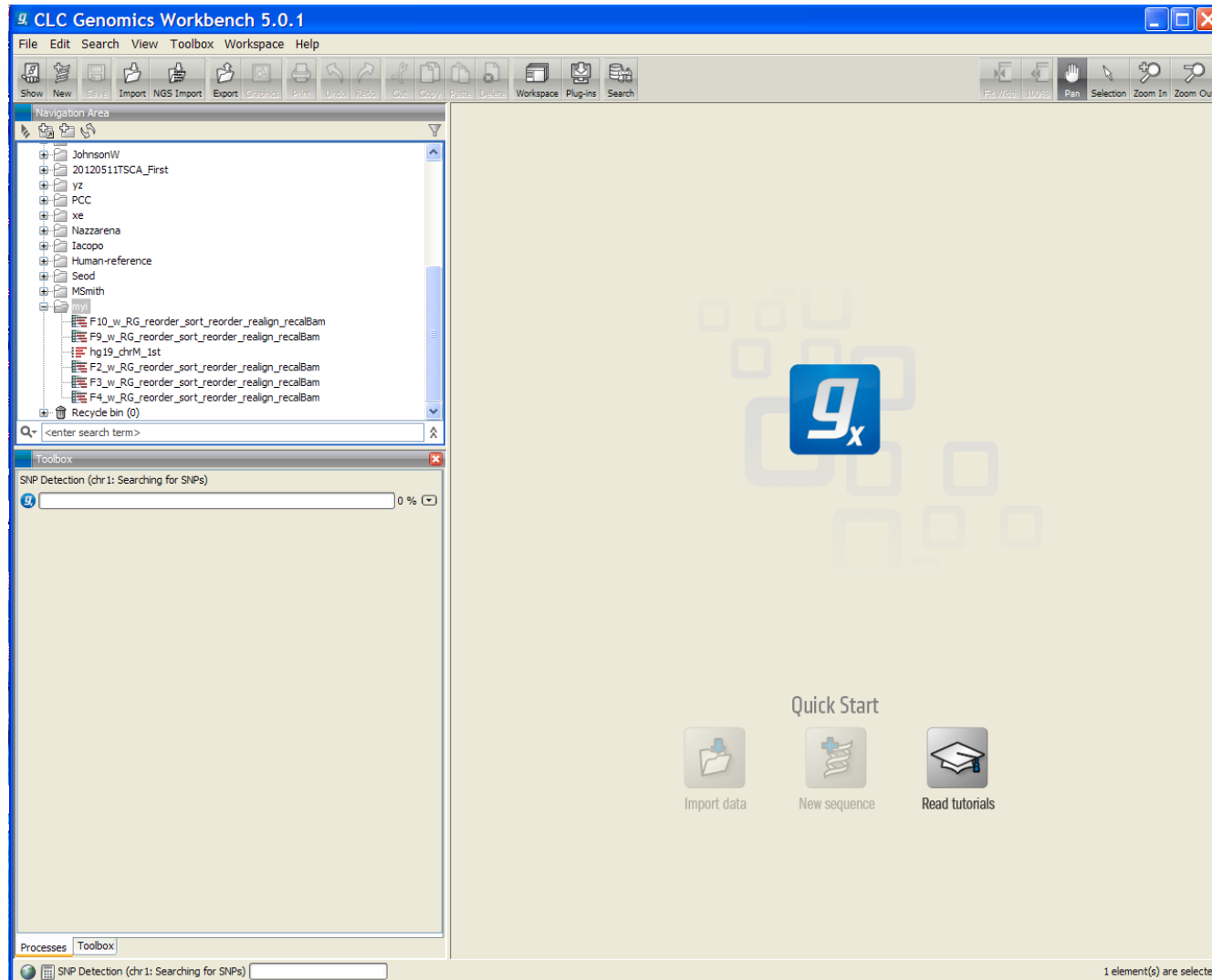


Choose location to save result

CLCbio SNP calling procedure



CLCbio SNP calling procedure



CLCbio SNP calling procedure

The screenshot displays the CLC Genomics Workbench 5.0.1 interface. The top menu bar includes File, Edit, Search, View, Toolbox, Workspace, and Help. The main window is divided into several panes:

- Navigation Area:** A tree view on the left showing a project structure for 'myi' with subfolders for 'F10_w_RG_reorder_sort_reorder_realign_recalBam', 'F9_w_RG_reorder_sort_reorder_realign_recalBam', 'F2_w_RG_reorder_sort_reorder_realign_recalBam', 'F3_w_RG_reorder_sort_reorder_realign_recalBam', and 'F4_w_RG_reorder_sort_reorder_realign_recalBam'. Each folder contains an 'SNP Detection log' and an 'SNP Detection Table' file. An orange arrow points to this area with the text 'SNP result files & log file'.
- Toolbox:** A panel at the bottom left showing the progress of SNP detection for each folder. Each entry is labeled 'SNP Detection (Done)' with a green progress bar at 100%. An orange arrow points to this area with the text 'Progress'.
- SNP Detection Table:** A large table on the right displaying the results for the selected 'F10_w_RG_reorder...' folder. The table has columns for Mapping, Reference, Consensus, Variation, Length, Reference, Variants, Allele Variants, Frequencies, Counts, and Cc. The first few rows are:

Mapping	Reference...	Consensu...	Variation ...	Length	Reference	Variants	Allele Vari...	Frequencies	Counts	Cc
chr1	14673	306	SNP	1	G	2	G/C	63.6/36.4	7/4	
chr1	14677	310	SNP	1	G	2	G/A	58.3/41.7	7/5	
chr1	14907	540	SNP	1	A	1	G	81.3	39	
chr1	14930	563	SNP	1	A	1	G	74.5	38	
chr1	15118	751	SNP	1	A	1	G	66.7	8	

An orange arrow points to the table with the text 'SNP result File content'. The table shows a list of SNPs across chromosome 1, including their positions, reference alleles, variant alleles, and associated frequencies and counts.

CLCbio SNP result file: not standard format (not in vcf format) One sample one SNP result file

```
tork.ncifcrf.gov - PuTTY
torkv:/banas/kebebew/AllUsers/CLCbio_SNPcalls> more "F10_w_RG_reorder_sort_reorder_realign_recalBam SNP Detection Table.txt"

"Mapping"      "Reference Position"  "Consensus Position"  "Variation Type"      "Length"      "Reference"      "Var
iants"  "Allele Variations"  "Frequencies"  "Counts"  "Coverage"  "Variant #1"  "Frequency of #1"  "Cou
nt of #1"  "Variant #2"  "Frequency of #2"  "Count of #2"  "Overlapping Annotations"  "Amino Acid Change"
chr1  14673  306  SNP  1  G  2  G/C  63.6/36.4  7/4  11  G  63.636  7  C  3
6.364  4
chr1  14677  310  SNP  1  G  2  G/A  58.3/41.7  7/5  12  G  58.333  7  A  4
1.667  5
chr1  14907  540  SNP  1  A  1  G  81.2  39  48  G  81.25  39
chr1  14930  563  SNP  1  A  1  G  74.5  38  51  G  74.51  38
chr1  15118  751  SNP  1  A  1  G  66.7  8  12  G  66.667  8
chr1  63516  3482  SNP  1  A  1  G  100  6  6  G  100  6
chr1  135982  4816  SNP  1  A  1  G  100  4  4  G  100  4
chr1  136048  4882  SNP  1  C  2  C/T  60.0/40.0  3/2  5  C  60  3  T  4
0  2
chr1  662029  5709  SNP  1  G  1  A  100  4  4  A  100  4
chr1  663097  6177  SNP  1  G  1  C  100  5  5  C  100  5
chr1  753269  8316  SNP  1  C  1  G  100  16  16  G  100  16
chr1  753405  8452  SNP  1  C  1  A  100  7  7  A  100  7
-More-- (0%)
```

CLCbio SNP result file annotation: (From user guide)

URL: http://www.clcbio.com/files/usermanuals/CLC_Genomics_Workbench_User_Manual.pdf

1. Reference position. The SNP's position on the reference sequence
2. Consensus position. The SNP's position on the consensus sequence.
3. Variation type. The SNP is described as complex, if it has more variations than specified in the ploidy setting in figure 19.99.
4. Length. The length of the SNP will always be one, as the name implies, unless two SNPs are found within the same codon.
5. Reference. The base found in the reference sequence. For results from de novo assembly, it will be the base found in the consensus sequence.
6. Variants. The number of variants among the reads.
7. Allele variations. Displays which bases are found at this position.
8. Frequencies. The frequency of a given variant.
9. Counts. This is similar to the frequency just reported in absolute numbers.
10. Coverage. The coverage at the SNP position. Note that only the reads that pass the quality filter will be reported here.
11. Variant numbers and frequencies. The information from the Allele variations, frequencies and counts are also split apart and reported for each variant individually
12. Overlapping annotations. This line shows if the SNP is covered by an annotation. The annotation's type and name will be displayed. For annotated reference sequences, this information can be used to tell if the SNP is found in e.g. a coding or non-coding region of the genome.
13. Amino acid change. If the reference sequence of the is annotated with ORF or CDS annotations, the SNP detection will also report whether the SNP is synonymous or nonsynonymous.

Tool By Tool Highlighting Major Aspects of Practical Usage

- GATK
- SAMtools
- VarScan
- CLCBio
- **CASAVA**
- Partek Genomic Suite

Illumina solution for NGS data analysis

--CASAVA SNP caller

Figure 1: SNP Caller in Illumina's DNA and RNA Sequencing Workflow

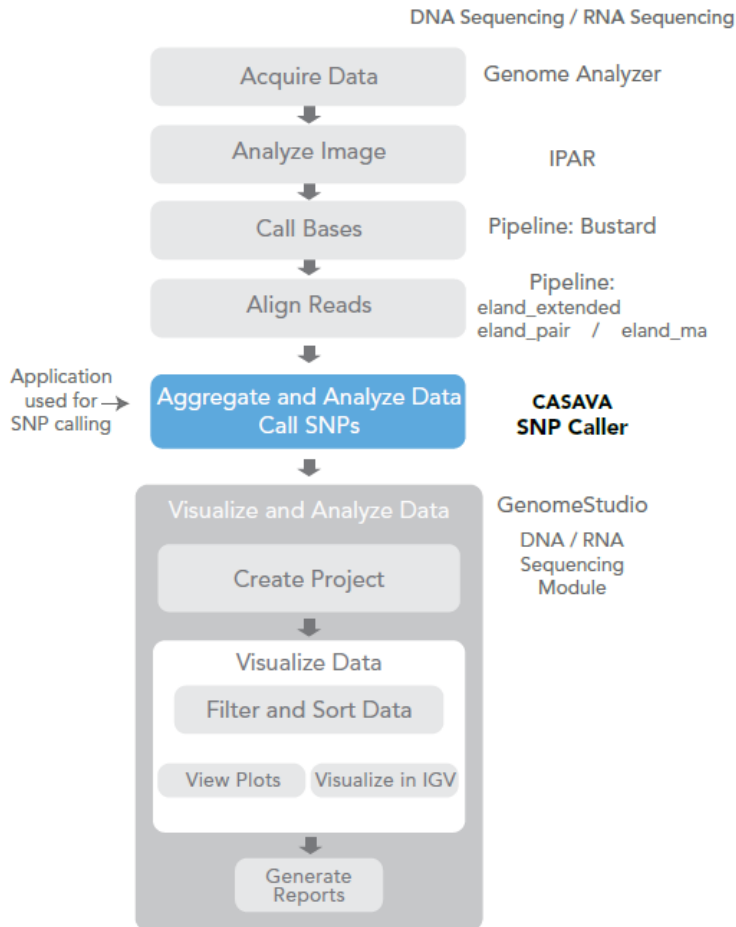
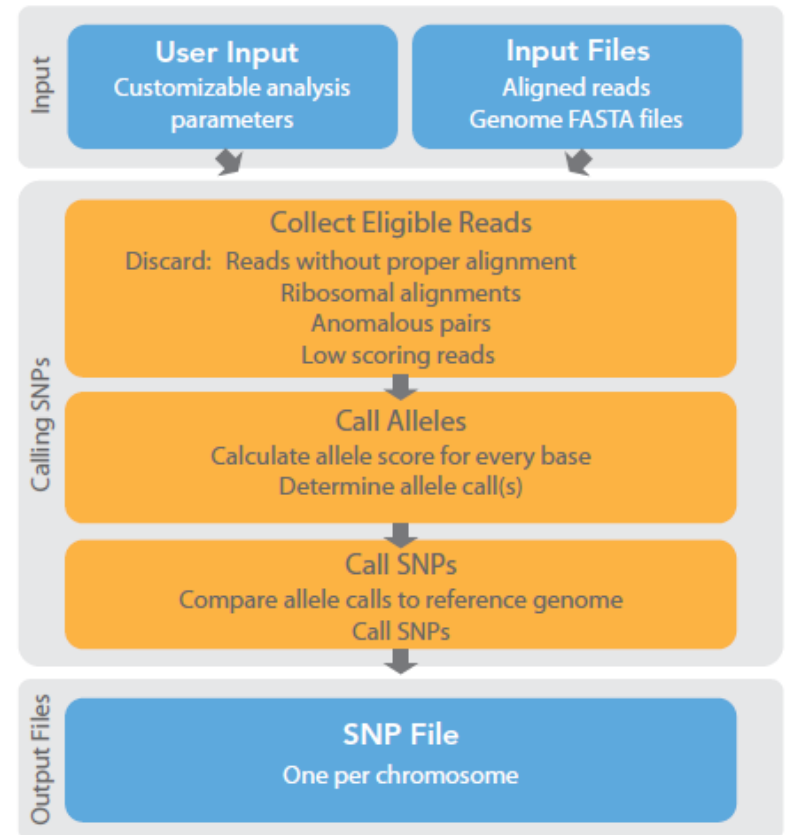


Figure 2: SNP Calling Workflow



Resources: http://www.illumina.com/documents/products/technotes/technote_snp_caller_sequencing.pdf

http://futo.cs.yale.edu/mw/images/7/77/CASAVA_UserGuide_15011196D.pdf

CASAVA1.8 SNP calling procedure

```
tork.ncifcrf.gov - PuTTY
torky:/bioinfoC/myi/Collaborators/R/NextGen> /opt/nasapps/bin/configureBuild.pl -h
[2012-07-24 11:19:05] [configureBuild.pl] INFO: Detected plugin sort
[2012-07-24 11:19:05] [configureBuild.pl] INFO: Detected plugin gsIndex
[2012-07-24 11:19:05] [configureBuild.pl] INFO: Detected plugin assembleIndels
[2012-07-24 11:19:05] [configureBuild.pl] INFO: Detected plugin refSeq
[2012-07-24 11:19:05] [configureBuild.pl] INFO: Detected plugin refSeqClean
[2012-07-24 11:19:05] [configureBuild.pl] INFO: Detected plugin rnaCounts
[2012-07-24 11:19:05] [configureBuild.pl] INFO: Detected plugin callSmallVariants
[2012-07-24 11:19:05] [configureBuild.pl] INFO: Detected plugin bam
Usage: configureBuild.pl [options]
  -id, --inSampleDir=PATH    - PATH to the aligned sample input directory
  -od, --outDir=PATH        - PATH to the build sample output directory
  -ref, --refSequences=PATH  - PATH of the reference genome sequences
  --samtoolsRefFile=FILE    - PATH to a single samtools-style reference file

OPTIONAL (BEHAVIOUR)
  -a, --applicationType=TYPE - type of analysis [DNA, RNA] default DNA
  --postRunCmd=CMDLINE      - executes CMDLINE after all tasks are finished
  -f, --force                - ignore errors from previous run
  --help[=TARGET]          - prints usage guide. If TARGET is specified, prints usage guide for the
target
  --tempDir                 - overrides default path for local temporary files
  --targets LIST            - space-separated LIST of targets to run (default: all)
  -w, --workflow            - instead of running the program generates the workflow definition file.
  -wa, --workflowAuto       - generates the workflow definition file and runs it. See --jobsLimit.
  -sa, --sgeAuto            - generates the workflow definition file and runs it on SGE (use with --
geQueue)
  --jobsLimit               - limit number of parallel jobs. Defaults: -1 (unlimited) for --sgeAuto.
1 for --workflowAuto.
  --sgeQueue                - SGE queue name - used with --sgeAuto or --workflow (e.g: all.q)
  --sgeQsubFlags            - Extra parameters to be passed to SGE qsub by the taskServer.pl
  --workflowFile=FILE       - overrides workflow file name. (default workflow.<date>.txt)
  --verbose=NUMBER          - sets the console log verbose level (default 0 - minimum)
  --version                 - prints version information

OPTIONAL (ANALYSIS)
  --refFlatFile=PATH        - PATH to UCSC refFlat.txt.gz file. The file must be gz-compressed.
  --seqGeneMdFile=PATH     - PATH to NCBI seq_gene.md.gz file. The file must be gz-compressed.
  --sortKeepAllReads       - Keep all purity filtered, duplicate and unmapped reads in the build.
                          These reads will be ignored during variant calling.
  --read                    - Limit input to the specified read only. Forces single-ended analysis
                          on one read of a double-ended dataset.
  --QVCutoff=NUMBER        - Sets the paired-end alignment score threshold to NUMBER (default 90)
  --QVCutoffSingle=NUMBER  - Sets the single-read alignment score threshold to NUMBER (default 10)
```

CASAVA1.8 SNP Call: Action commands for all samples

CASAVA (call sample individually):

```
/Path/illumina/casava_v1.8.2/bin/configureBuild.pl
--samtoolsRefFile /banas/nextgen2/illumina/PROC/RefGenomes/hg19/ordered/hg19.fa
--inSampleDir /Path/Sample_S1
--outDir Sample_S1_variants
--targets all
--wa
--variantsSnpCovCutoff=-1
--variantsIndelCovCutoff=-1
2>run_casava_build.err
1>run_casava_build.log&
```

Make sure `--variantsSnpCovCutoff=-1` to disable the filter for targeted resequencing, exome-seq etc; default as 3.0X mean chromosomal used-depth



Within output directory `Sample_S1_variants`, many files and subdirectories created



Within `Parsed_date` subdirectory, SNPs result file `snp.txt` in subdirectory of each chromosome (one SNP result file per chromo), need to combine

CASAVA1.8 SNP result file: not standard format (not in vcf format) and by chromosome (need to combine)

```
tork.ncifcrf.gov - PuTTY
torkv:/isl/projects/nextgen/scratch/illumina/PROJECT/kabebew_data/70BETAAXX/casava18_lane2_F4_variant/Parsed_08-02-12/chr11> more snps.txt
# ** CASAVA depth-filtered snp calls **
#$ CMDLINE /opt/nasapps/stow/illumina/casava_v1.8.2/libexec/CASAVA-1.8.2/filterSmallVariants.pl --projectDir=/isl/projects/nextgen/scratch/illumina/PROJECT/kabebew_data/70BETAAXX/casava18_lane2_F4_variant --chrom=chr11
#$ SEQ_MAX_DEPTH chr11 undefined
#
#$ COLUMNS seq name pos bcalls_used bcalls_filt ref Q(snp) max_gt Q(max_gt) max_gt|poly_site Q(max_gt|poly_site) A_
used C_used G_used T_used
chr11 128378 1 1 C 5 CT 2 CT 3 0 0 0 1
chr11 138790 1 0 A 1 AA 8 AG 3 0 0 1 0
chr11 139589 1 0 C 10 CG 3 CG 3 0 0 1 0
chr11 175373 1 0 A 5 AC 2 AC 3 0 1 0 0
chr11 175543 1 1 G 4 GG 2 CG 3 0 1 0 0
chr11 175566 2 0 T 6 CC 2 CC 4 0 2 0 0
chr11 178593 1 0 C 6 CG 2 CG 3 0 0 1 0
chr11 180116 1 0 C 10 CT 3 CT 3 0 0 0 1
chr11 180151 3 0 C 3 CC 4 CG 31 0 2 1 0
chr11 180153 3 0 A 3 AA 4 AG 31 2 0 1 0
chr11 180225 4 1 A 34 AG 33 AG 38 1 0 3 0
chr11 180623 1 0 A 10 AC 3 AC 3 0 1 0 0
chr11 184431 2 0 T 4 CT 4 CT 30 0 1 0 1
chr11 184475 3 0 A 3 AA 4 AG 31 2 0 1 0
chr11 184504 4 0 C 1 CC 5 AC 28 1 3 0 0
chr11 186232 1 0 T 5 TT 2 CT 3 0 1 0 0
chr11 186325 1 0 G 1 GG 7 AG 3 1 0 0 0
```

CASAVA1.8 SNP result file annotation: (From user guide)

- 1 seq_name Reference sequence label
- 2 Pos Sequence position of the site/snp
- 3 bcalls_used Basecalls used to make the genotype call for this site
- 4 bcalls_filt Basecalls mapped to the site but filtered out before genotype calling
- 5 Ref Reference Base
- 6 Q(snp) A Q-value expressing the probability of the homozygous reference genotype, subject to the expected rate of haplotype difference as expressed by the (Watterson) theta parameter
- 7 max_gt The most likely genotype (subject to theta, as above).
- 8 Q(max_gt) A Q-value expressing the probability that the genotype is not the most likely genotype above (subject to theta).
- 9 max_gt|poly_site The most likely genotype assuming this site is polymorphic with an expected allele frequency of 0.5 (theta is still used to calculate the probability of a third allele -- i.e. the chance of observing two non-reference alleles).
- 10 Q(max_gt|poly_site) A Q-value expressing the probability that the genotype is not the most likely genotype above assuming this site is polymorphic.
- 11 A_used 'A' basecalls used
- 12 C_used 'C' basecalls used
- 13 G_used 'G' basecalls used
- 14 T_used 'T' basecalls used

Tool By Tool Highlighting Major Aspects of Practical Usage

- GATK
- SAMtools
- VarScan
- CLCBio
- CASAVA
- **Partek Genomic Suite**

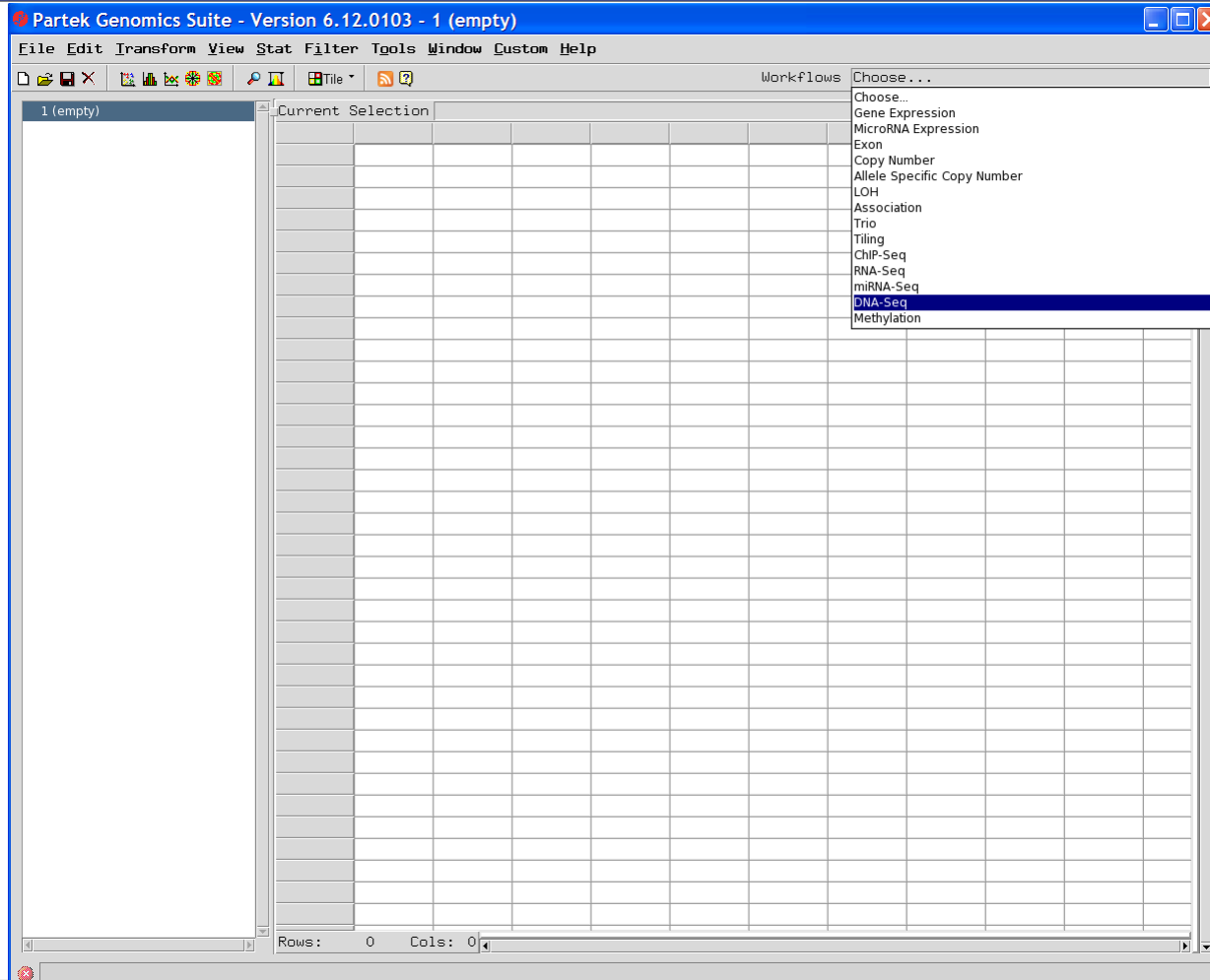
Partek solution for NGS data analysis

The screenshot shows the Partek website homepage. The browser title is "Partek Incorporated | Next Generation Sequencing & Microarray Software - Windows Internet Explorer". The address bar shows "http://www.partek.com/". The navigation menu includes "Home", "Free Trial", "Software", "Support", "Publications", "Company", and "日本語". The main banner reads "Meet the Partek® Family of Genomics Software" and "The Most Complete Start-to-Finish Next Generation Sequencing and Microarray Analysis Solution Available". It features three computer monitors with "LEARN MORE" buttons and logos for Partek Flow, Genomics Suite, and Pathway. Below the banner are logos for Affymetrix, Applied Biosystems, Illumina, Ingenuity, Ion Torrent, and NanoString. A central section highlights "Winner of the Illumina Data Excellence Award for Most Creative Algorithm" with a "Read more" link and a "Request Access to the Video" button. The bottom section is divided into three columns: "Next Generation Sequencing", "Microarray", and "Functional Genomics", each with a brief description of the technology.

Partek Flow uses external SNP detection methods (e.g, samtools),
Genomic suite has its own SNP detection method

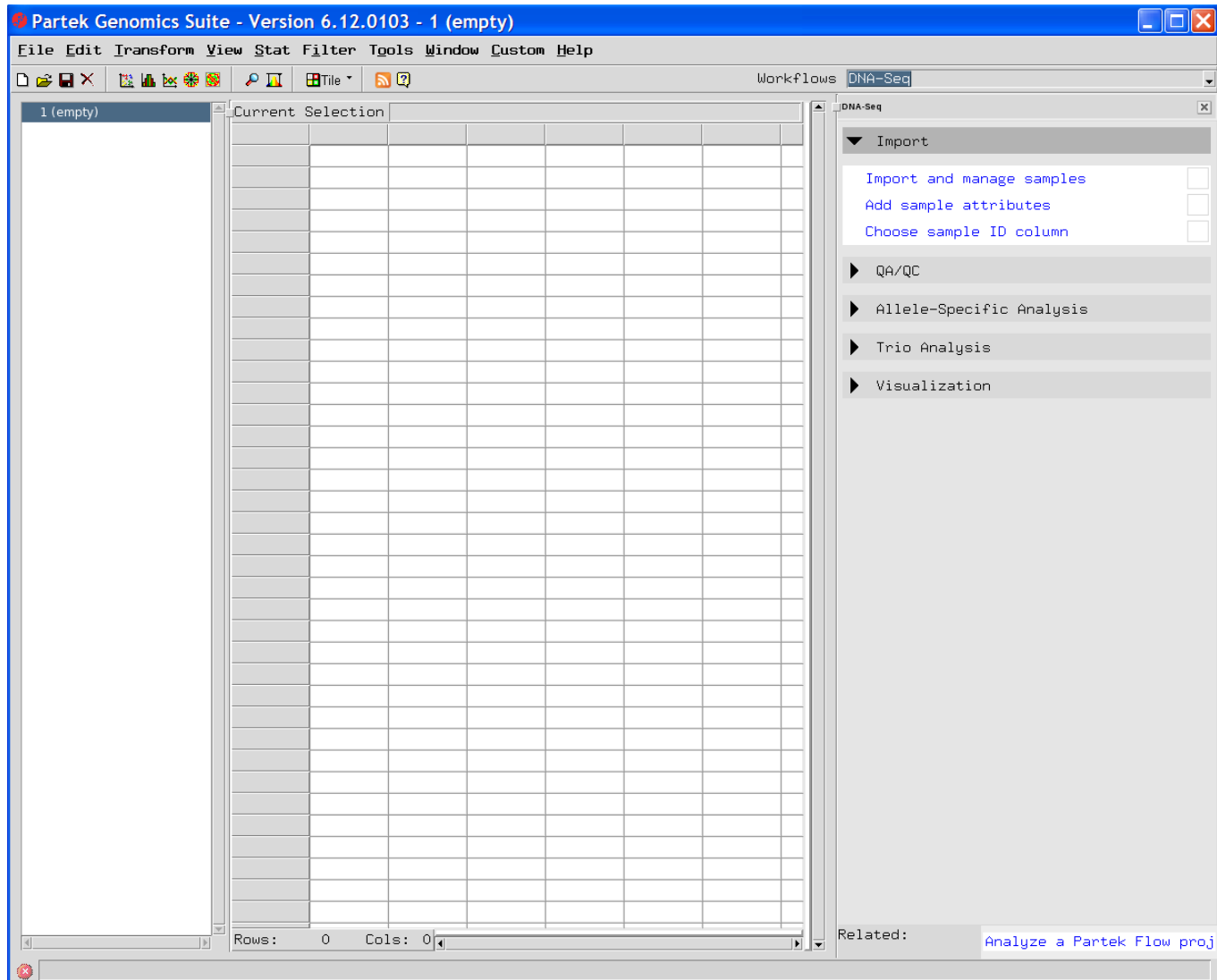
For NGS data, better run from powerful system (linux version Partek in dedicated server)

```
tork.ncifcrf.gov - PuTTY  
[yiming@fr-s-isp-partek ~]$ /bioinfoA/apps/Linux_x86_64/partekgs/bin/partek
```



Choose
Workflow
Of DNA-seq
For SNP
detection

Partek SNP calling procedure



Partek SNP calling procedure

The screenshot displays the Partek Genomics Suite interface, version 6.12.0103. The main window is titled "1 (empty)" and shows a grid for "Current Selection". A dialog box titled "Bam samples required" is overlaid on the grid, containing the following text:

Bam samples required

This step requires a spreadsheet associated with sequencing reads, but none were found.

Would you like to import samples now?

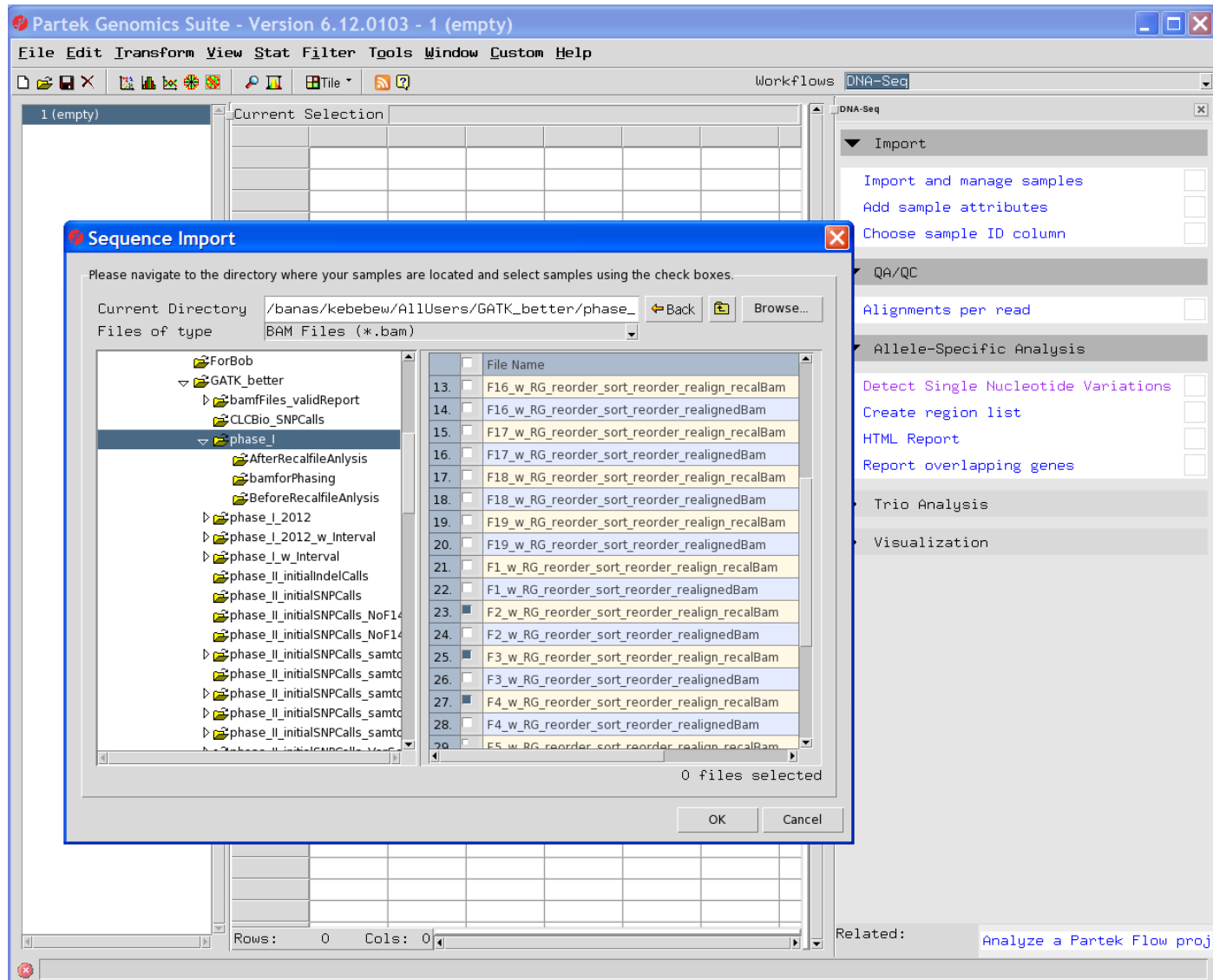
Yes No

The background interface includes a menu bar (File, Edit, Transform, View, Stat, Filter, Tools, Window, Custom, Help) and a toolbar. The "Workflows" dropdown is set to "DNA-Seq". The right-hand panel, titled "DNA-Seq", lists various analysis steps with checkboxes:

- Import
 - Import and manage samples
 - Add sample attributes
 - Choose sample ID column
- QA/QC
 - Alignments per read
- Allele-Specific Analysis
 - Detect Single Nucleotide Variations
 - Create region list
 - HTML Report
 - Report overlapping genes
- Trio Analysis
- Visualization

At the bottom right, a "Related:" section contains a link: "Analyze a Partek Flow proj". The status bar at the bottom indicates "Rows: 0 Cols: 0".

Partek SNP calling procedure



Partek SNP calling procedure

The screenshot displays the Partek Genomics Suite interface. The main window shows a table with the following data:

1. Sample ID	2. Number of Alignmen
F10_w_RG_reorder_sort_reorder_realign_recalBam	20157378
F2_w_RG_reorder_sort_reorder_realign_recalBam	36727266
F3_w_RG_reorder_sort_reorder_realign_recalBam	38413215
F4_w_RG_reorder_sort_reorder_realign_recalBam	36535851
F9_w_RG_reorder_sort_reorder_realign_recalBam	21430001

The right-hand panel, titled 'DNA-Seq', contains a workflow menu with the following options:

- Import
 - Import and manage samples
 - Add sample attributes
 - Choose sample ID column
- QA/QC
 - Alignments per read
- Allele-Specific Analysis
 - Detect Single Nucleotide Variations
 - Create region list
 - HTML Report
 - Report overlapping genes
- Trio Analysis
- Visualization

At the bottom right, there is a 'Related:' section with a link: [Analyze a Partek Flow project](#)

Partek SNP calling procedure

The screenshot displays the Partek Genomics Suite interface. The main window title is "Partek Genomics Suite - Version 6.12.0103 - 1 (phase_l_w_Interval-01-25-2012_only5Samp)". The menu bar includes File, Edit, Transform, View, Stat, Filter, Tools, Window, Custom, and Help. The toolbar contains various icons for file operations and analysis. The main workspace shows a table with the following data:

1. Sample ID	2. Number of Alignments
F10_w_RG_reorder_sort_reorder_realign_recalBam	20157378
F2_w_RG_reorder_sort_reorder_realign_recalBam	36727266
F7_w_RG_reorder_sort_reorder_realign_recalBam	78417215

A dialog box titled "Detect Single Nucleotide Variations (SNVs)" is open, prompting the user to "Select a test to detect variants." Under the "Genotype likelihood test" section, two options are listed:

- ◆ Detect SNVs among samples
- ◆ Detect SNVs against the reference sequence

The "Detect SNVs against the reference sequence" option is selected. The dialog box has "OK" and "Cancel" buttons at the bottom.

On the right side, the "DNA-Seq" workflow panel is visible, showing a tree structure of analysis steps:

- Import
 - Import and manage samples
 - Add sample attributes
 - Choose sample ID column
- QA/QC
 - Alignments per read
- Allele-Specific Analysis
 - Detect Single Nucleotide Variations
 - Create region list
 - HTML Report
 - Report overlapping genes
- Trio Analysis
- Visualization

At the bottom right, there is a "Related:" section with a link to "Analyze a Partek Flow project".

Partek SNP calling procedure

The screenshot displays the Partek Genomics Suite interface. The main window shows a table with the following data:

1. Sample ID	2. Number of Alignments
F10_w_RG_reorder_sort_reorder_realign_recalBam	20157378
F2_w_RG_reorder_sort_reorder_realign_recalBam	36727266
F3_w_RG_reorder_sort_reorder_realign_recalBam	38413215
F4_w_RG_reorder_sort_reorder_realign_recalBam	36535851
F9_w_RG_reorder_sort_reorder_realign_recalBam	21430001

A dialog box titled "Detect Nucleotides that are Different from the Reference" is overlaid on the main window. The dialog contains the following text and options:

This procedure will detect locations in the DNA that are different from the reference.

Display loci that have a Log Odds Ratio greater than of differing from the reference.

Organism is diploid

Strand Search

- Include reads from both strands
- Include reads from positive strand only
- Include reads from negative strand only

Result file:

The background software interface includes a menu bar (File, Edit, Transform, View, Stat, Filter, Tools, Window, Custom, Help), a toolbar, and a right-hand panel with sections for "Import" (Import and manage samples, Add sample attributes, Choose sample ID column) and "QA/QC" (Comments per read, Sample-Specific Analysis, Detect Single Nucleotide Variations, Generate region list, Report, Detect overlapping genes). The status bar at the bottom indicates "Rows: 5 Cols: 2".

Partek SNP calling procedure

The screenshot displays the Partek Genomics Suite interface. A dialog box titled "Detect Nucleotides that are Different from the Reference" is open, providing instructions and configuration options for SNP calling. The dialog includes a text area for the Log Odds Ratio threshold, a checkbox for "Organism is diploid", a "Strand Search" section with three radio button options, and a "Result file" field with a "Browse..." button. The status bar at the bottom of the dialog indicates "Processing sequence 17 (9 of 25)".

Partek Genomics Suite - Version 6.12.0103 - 1 (phase_I_w_Interval-01-25-2012_only5Samp)

File Edit Transform View Stat Filter Tools Window Custom Help

Workflows DNA-Seq

1. Sample ID	2. Number of Alignmen
F10_w_RG_reorder_sort_reorder_realign_recalBam	20157378
F2_w_RG_reorder_sort_reorder_realign_recalBam	36727266

Detect Nucleotides that are Different from the Reference

This procedure will detect locations in the DNA that are different from the reference.

Display loci that have a Log Odds Ratio greater than of differing from the reference.

Organism is diploid [?](#)

Strand Search

Include reads from both strands. [?](#)

Include reads from positive strand only. [?](#)

Include reads from negative strand only. [?](#)

Result file

Processing sequence 17 (9 of 25)

Rows: 5 Cols: 2

Related: [Analyze a Partek Flow project](#)

Partek SNP calling procedure-SNP result file

Partek Genomics Suite - Version 6.12.0103 - 1/reference-snps (SNVsAgainstReference_2012.txt)

File Edit Transform View Stat Filter Tools Window Custom Help

Workflows DNA-Seq

1 (phase_1_w_Interval-01-25-2012) Current Selection 150

	3. sample ID	4. reference base	5. genotype call	6. total Non-Reference bases	7. total coverage at locus	8. non-reference average base qualities
1.	F3_w_RG_reorder_sort_recA	A	GG	146	150	31.5411
2.	F3_w_RG_reorder_sort_recA	A	GG	191	193	26.801
3.	F2_w_RG_reorder_sort_recG	G	AA	170	173	29.3824
4.	F3_w_RG_reorder_sort_recG	G	AA	175	176	30.5486
5.	F4_w_RG_reorder_sort_recG	G	AA	153	159	22.5229
6.	F2_w_RG_reorder_sort_recC	C	TT	170	174	29.4588
7.	F3_w_RG_reorder_sort_recC	C	TT	176	177	30.5795
8.	F4_w_RG_reorder_sort_recC	C	TT	152	156	22.4145
9.	F2_w_RG_reorder_sort_recC	C	CT	194	411	30.3711
10.	F3_w_RG_reorder_sort_recC	C	CT	210	411	31.5619
11.	F4_w_RG_reorder_sort_recC	C	CT	194	387	26.9897
12.	F2_w_RG_reorder_sort_recC	C	CC	146	147	31.6507
13.	F3_w_RG_reorder_sort_recC	C	CC	141	141	31.3688
14.	F4_w_RG_reorder_sort_recC	C	CC	141	141	25.8865
15.	F3_w_RG_reorder_sort_recT	T	CT	172	332	29
16.	F3_w_RG_reorder_sort_recA	A	AG	181	347	29.4696
17.	F3_w_RG_reorder_sort_recT	T	CT	181	353	32.0718
18.	F2_w_RG_reorder_sort_recT	T	CT	163	292	32.5031
19.	F3_w_RG_reorder_sort_recT	T	CT	172	344	34.8837
20.	F4_w_RG_reorder_sort_recT	T	CT	175	318	28.9771
21.	F3_w_RG_reorder_sort_recA	A	AG	179	359	33.5866
22.	F3_w_RG_reorder_sort_recG	G	AG	179	359	33.905
23.	F4_w_RG_reorder_sort_recG	G	AG	190	337	28.7316
24.	F4_w_RG_reorder_sort_recG	G	AG	167	319	31.6946
25.	F2_w_RG_reorder_sort_recA	A	GG	155	155	24.7484
26.	F3_w_RG_reorder_sort_recA	A	GG	153	153	26.8039
27.	F4_w_RG_reorder_sort_recG	G	AA	153	154	26.6144
28.	F3_w_RG_reorder_sort_recT	T	CC	133	134	32.3083
29.	F4_w_RG_reorder_sort_recC	C	TT	139	139	24.9496
30.	F2_w_RG_reorder_sort_recT	T	GG	138	138	29.2391
31.	F3_w_RG_reorder_sort_recT	T	CC	135	136	31.2593
32.	F2_w_RG_reorder_sort_recT	T	CC	337	338	32.5252
33.	F3_w_RG_reorder_sort_recT	T	CC	368	370	33.0897
34.	F4_w_RG_reorder_sort_recT	T	CC	350	353	25.86
35.	F9_w_RG_reorder_sort_recT	T	CC	175	180	31.0571

Rows: 4062940 Cols: 15

DNA-Seq

- Import
 - Import and manage samples
 - Add sample attributes
 - Choose sample ID column
- QA/QC
 - Alignments per read
- Allele-Specific Analysis
 - Detect Single Nucleotide Variations
 - Create region list
 - HTML Report
 - Report overlapping genes
- Trio Analysis
- Visualization

Related: [Analyze a Partek Flow pro](#)

Partek SNP result file: not standard format (not in vcf format)

```
tork.ncifcrf.gov - PuTTY
torkv:/banas/kebew/AllUsers/PartekSNPs> more SNVsAgainstReference_2012.txt
position          log-odds ratio of SNP against reference sample ID      reference base genotype call  total Non-R
eference bases   total coverage at locus non-reference average base qualities  reference base qualities      non
-reference average mapping qualities  reference average mapping qualities  A      C      G      T
chr1.14907       1e+06    F3_w_RG_reorder_sort_reorder_realign_recalBam  A      GG      146    150    31.5411 365
4.9726 27.5    4      1      145    0
chr1.14930       1e+06    F3_w_RG_reorder_sort_reorder_realign_recalBam  A      GG      191    193    26.801 35.
5      56.2251 105    2      0      191    0
chr1.808922      1e+06    F2_w_RG_reorder_sort_reorder_realign_recalBam  G      AA      170    173    29.3824 251
40.247 53.6667 168    1      3      1
chr1.808922      1e+06    F3_w_RG_reorder_sort_reorder_realign_recalBam  G      AA      175    176    30.5486 291
59.246 44      175    0      1      0
chr1.808922      1e+06    F4_w_RG_reorder_sort_reorder_realign_recalBam  G      AA      153    159    22.5229 13.
3333 143.275 33      148    5      6      0
chr1.808928      1e+06    F2_w_RG_reorder_sort_reorder_realign_recalBam  C      TT      170    174    29.4588 22.
25      140.394 60.5    0      4      0      170
chr1.808928      1e+06    F3_w_RG_reorder_sort_reorder_realign_recalBam  C      TT      176    177    30.5795 381
57.966 44      0      1      0      176
chr1.808928      1e+06    F4_w_RG_reorder_sort_reorder_realign_recalBam  C      TT      152    156    22.4145 22.
75      144.342 29.25    0      4      2      150
chr1.1581096     1e+06    F2_w_RG_reorder_sort_reorder_realign_recalBam  c      CT      194    411    30.3711 31.
9309 70.1804 74.2949 0      217    2      192
chr1.1581096     1e+06    F3_w_RG_reorder_sort_reorder_realign_recalBam  c      CT      210    411    31.5619 32.
9005 82.9381 75.4627 0      201    5      205
chr1.1581096     1e+06    F4_w_RG_reorder_sort_reorder_realign_recalBam  c      CT      194    387    26.9897 26.
0259 76.8093 74.1917 1      193    3      190
```

Take-home Message

- Each tool is still evolving on its own pace
- Choosing the right tool not only depends upon which tool perform better, but also depends upon the user's capacity, e.g. running command line vs running interface/button click
- Input/output format shall be standardized for easy usage and performance cross-comparison/evaluation (e.g., vcf format for SNP result, bam files as input etc)
- Modularized tools are the best for integrative usage of the tools

Acknowledgements

ABCC

- Robert Stephens
- Yongmei Zhao
- Jigui Shan
- Jia Li
- Jack Chen

NCI/EOB

- Electron Kebebew
- Mei He

The Genome Institute at WashU

- Dan Koboldt (VarScan author)

Broad Institute GATK Team

- Ryan Poplin
- Mark Depristo
- Eric Banks

Email contact: myi@ncifcrf.gov