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

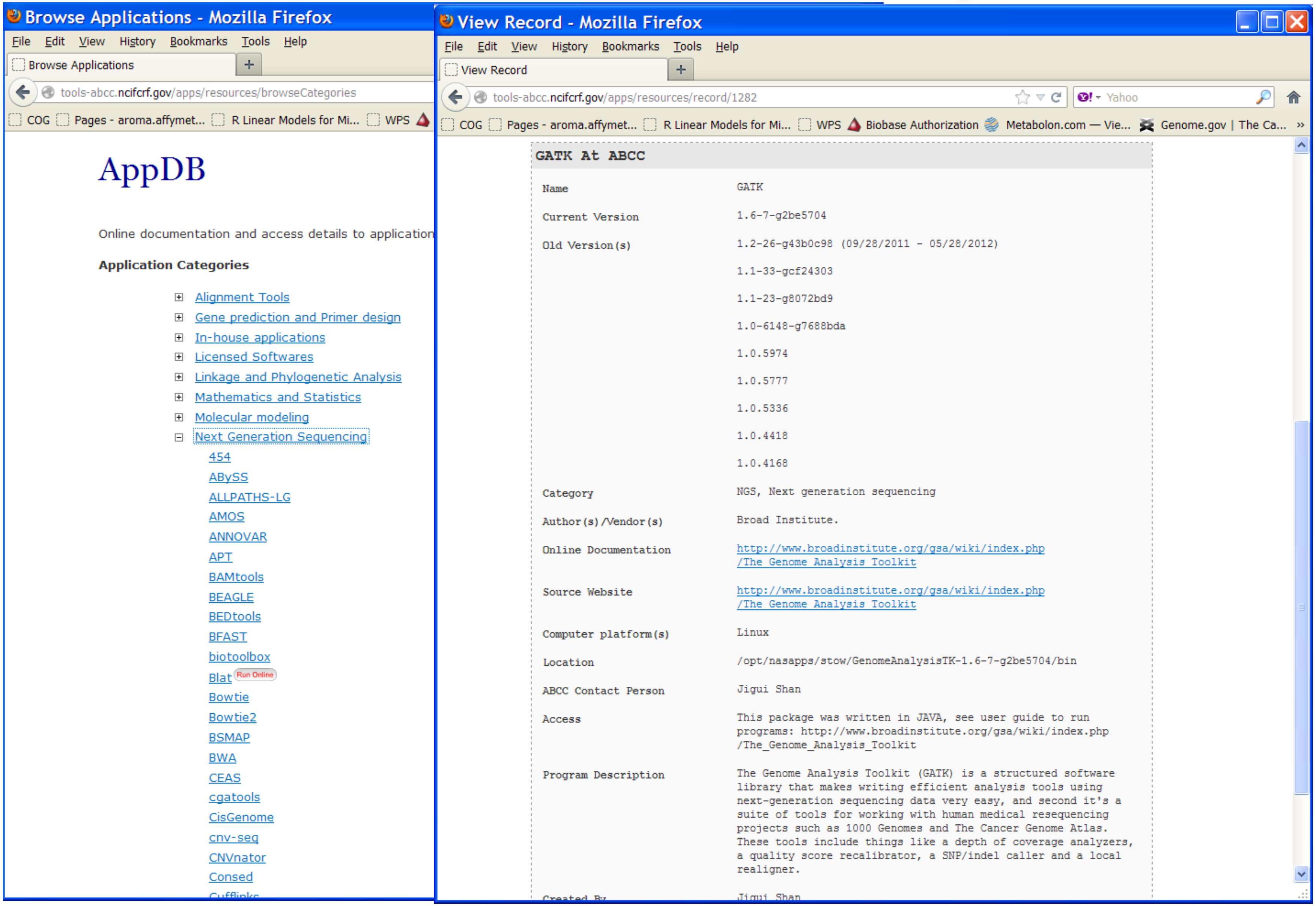
❖ Brief discussion of result files and some diagnosis plots, etc.

❖ 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



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 DePristo[1], Eric Banks[1], Ryan Poplin[1], Kiran V Garimella[1], Jared R Maguire[1], Christopher Hartl[1], Anthony A Philippakis[1–3], Guillermo del Angel[1], Manuel A Rivas[1,4], Matt Hanna[1], Aaron McKenna[1], Tim J Fennell[1], Andrew M Kernytsky[1], Andrey Y Sivachenko[1], Kristian Cibulskis[1], Stacey B Gabriel[1], David Altshuler[1,3,4] & Mark J Daly[1,3,4]

[1]Program in Medical and Population Genetics, Broad Institute of Harvard and MIT, Cambridge, Massachusetts, USA. [2]Brigham and Women's Hospital, Boston, Massachusetts, USA. [3]Harvard Medical School, Boston, Massachusetts, USA. [4]Center 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

CSH PRESS GENOME RESEARCH

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

Old Webpage-Best Practice v2

Best Practice Variant Detection with the GATK v2

From GSA

Contents

Introduction

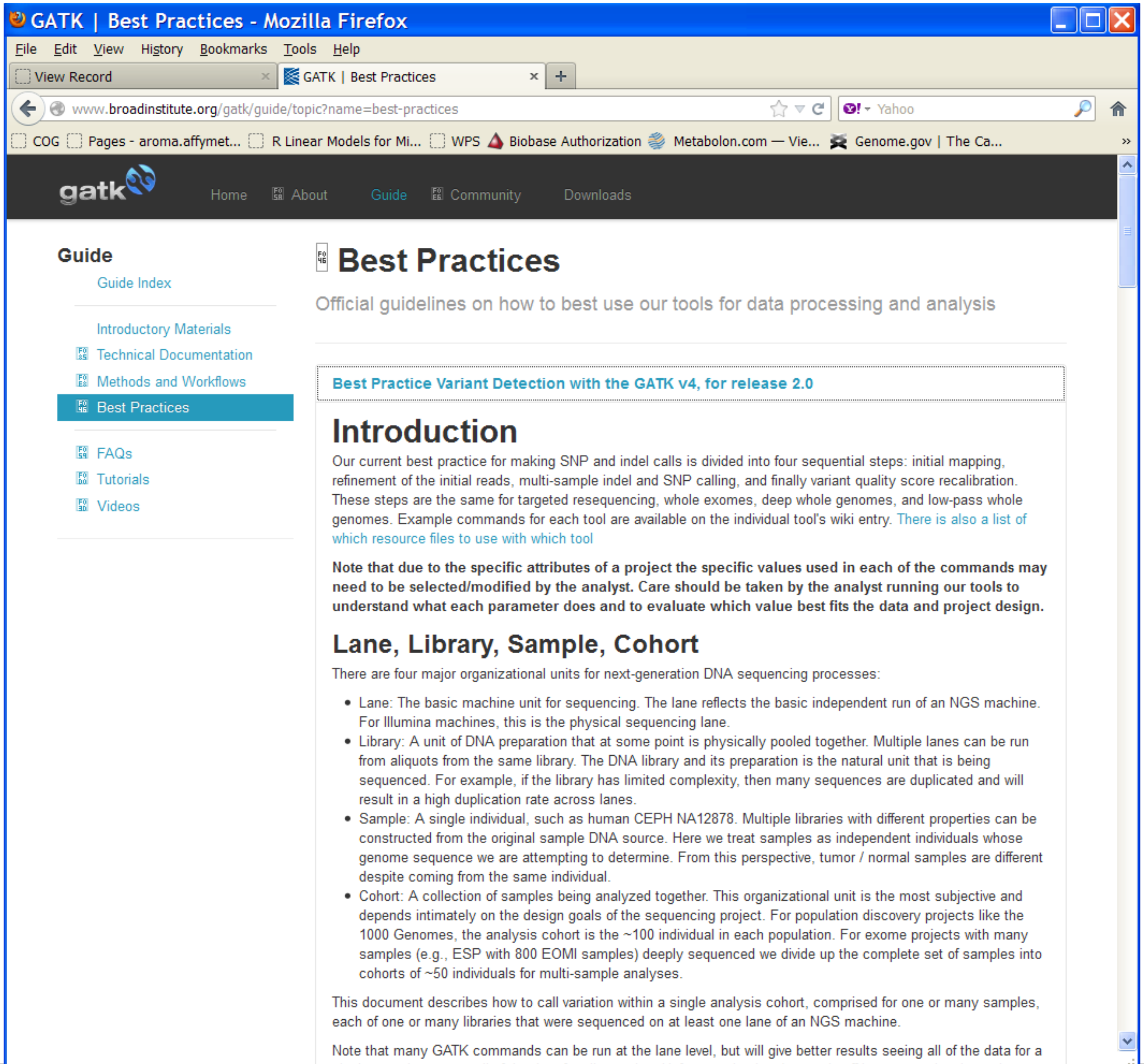
Old Webpage-Best Practice v3 (Up to GATK v1.6)

Best Practice Variant Detection with the GATK v3

From GSA

Contents

New Website-Best Practice v4-GATK v2.0



GATK | Best Practices - Mozilla Firefox

File Edit View History Bookmarks Tools Help

View Record | GATK | Best Practices

www.broadinstitute.org/gatk/guide/topic?name=best-practices

COG | Pages - aroma.affymet... | R Linear Models for Mi... | WPS | Biobase Authorization | Metabolon.com — Vie... | Genome.gov | The Ca...

gatk | Home | About | Guide | Community | Downloads

Guide

- Guide Index
- Introductory Materials
- Technical Documentation
- Methods and Workflows
- Best Practices
- FAQs
- Tutorials
- Videos

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

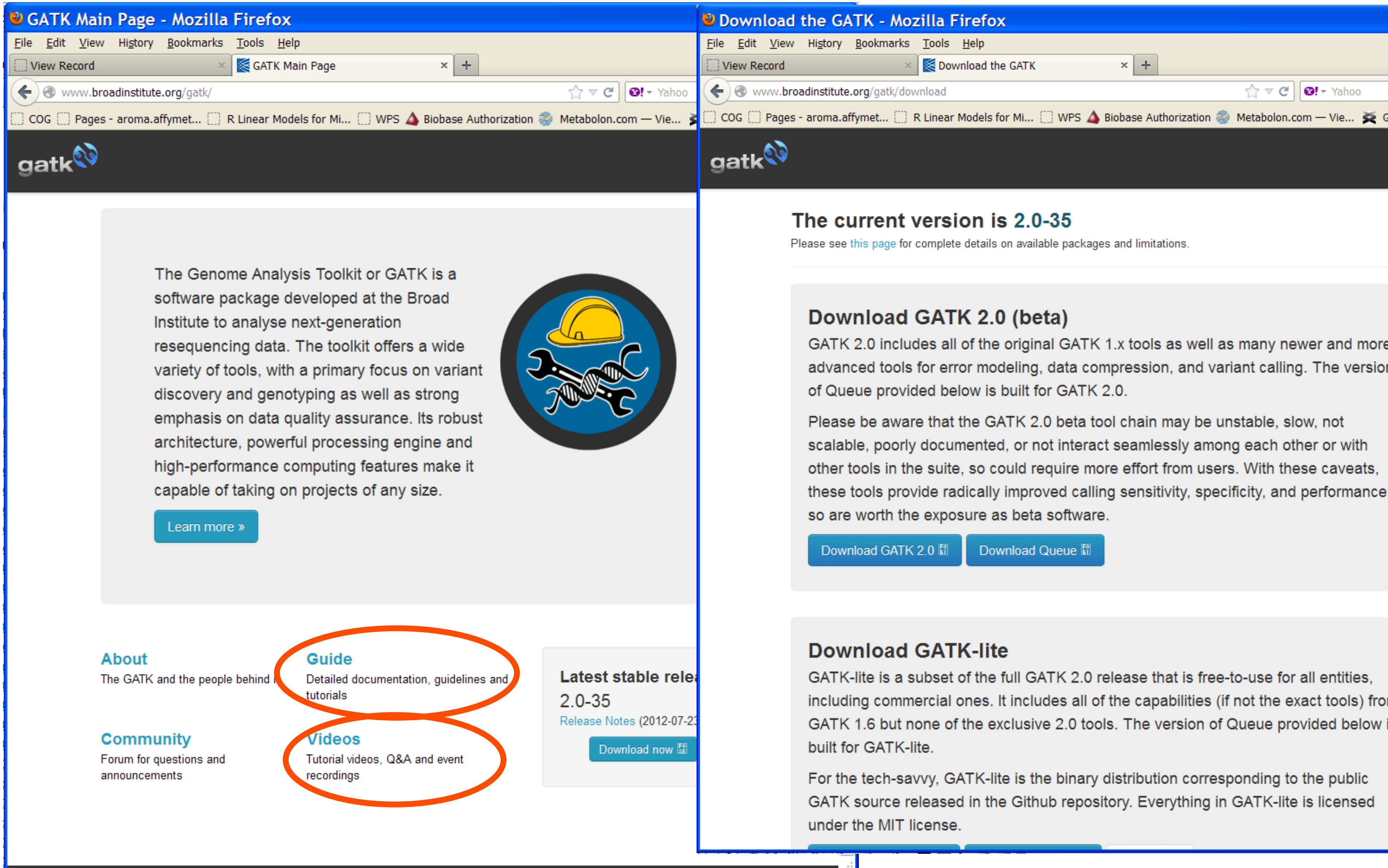
There are four major organizational units for next-generation DNA sequencing processes:

- 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.

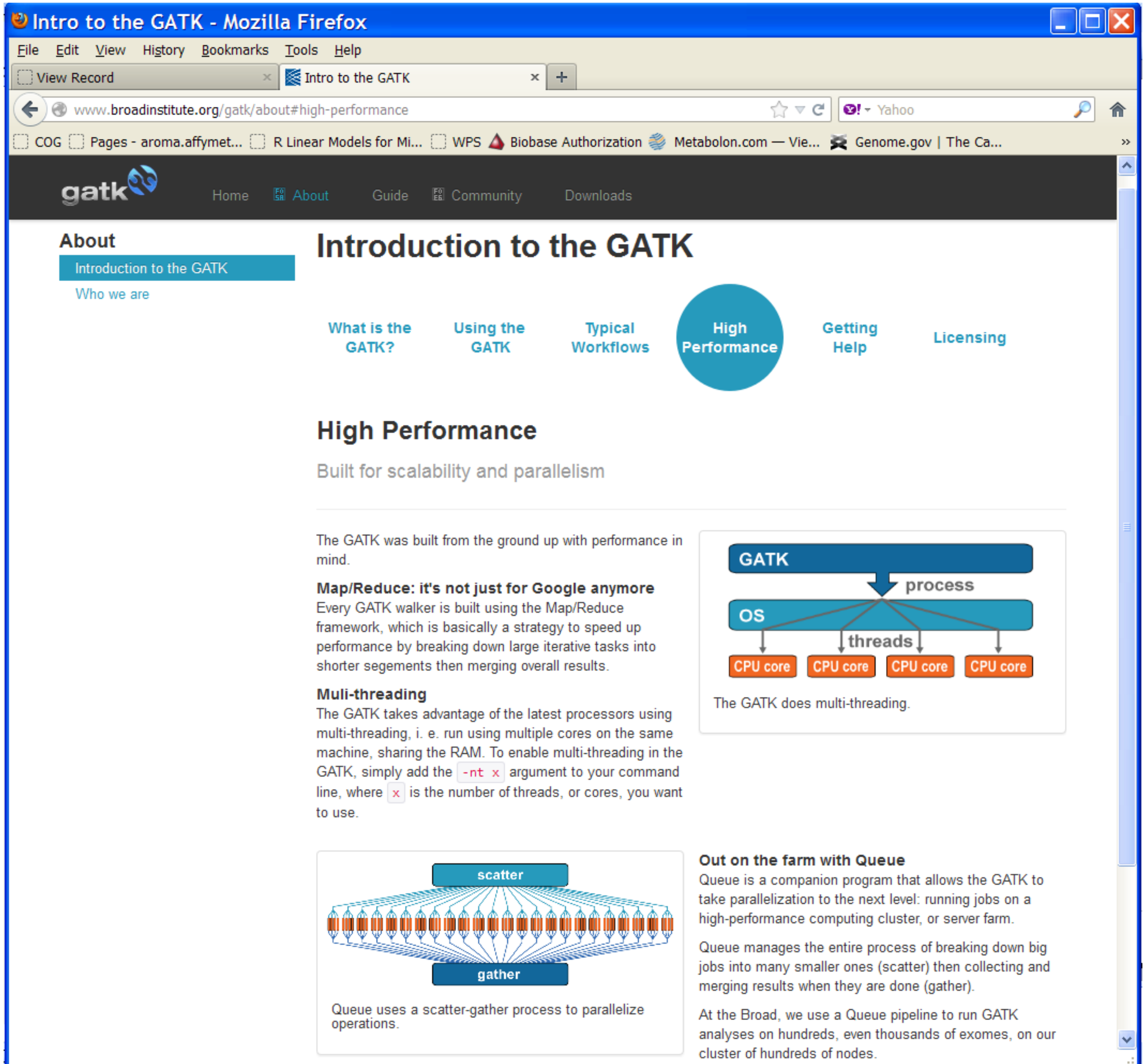
This document describes how to call variation within a single analysis cohort, comprised for one or many samples, each of one or many libraries that were sequenced on at least one lane of an NGS machine.

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

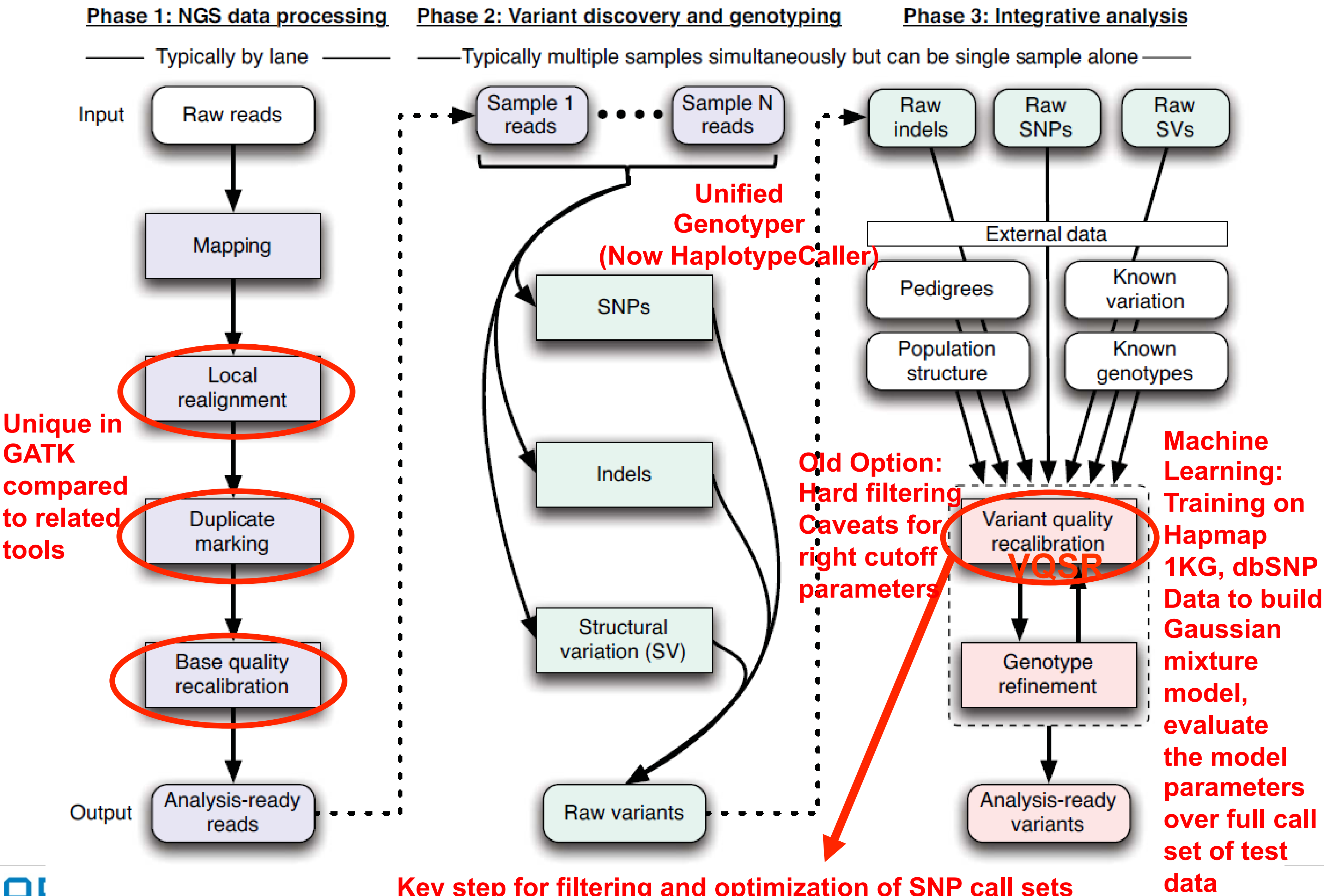
New GATK (v2.0) Website-Download and Guide



New GATK (v2.0) Website-Introduction



The Flagship Features of GATK WorkFlow



Key step for filtering and optimization of SNP call sets

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
GATCACAGGTCTATCACCCTATTAACCACTCACGGGAGCTCTCCATGCAT
TTGGTATTTTCGTCTGGGGGGTGTGCACGCGATAGCATTGCGAGACGCTG
GAGCCGGAGCACCCTATGTCGCAGTATCTGTCTTTGATTCCTGCCTCATT
CTATTATTTATCGCACCTACGTTCAATATTACAGGCGAACATACCTACTA
AAGTGTGTTAATTAATTAATGCTTGTAGGACATAATAATAACAATTGAAT
GTCTGCACAGCCGCTTTCCACACAGACATCATAACAAAAAATTTCCACCA
AACCCCCCCTCCCCCGCTTCTGGCCACAGCACTTAAACACATCTCTGC
CAAACCCCAAAAACAAAGAACCCTAACACCAGCCTAACCAGATTTCAAAT
TTTATCTTTAGGCGGTATGCACTTTTAACAGTCACCCCCCAACTAACACA
TTATTTTCCCCTCCCACTCCCATACTACTAATCTCATCAATACAACCCCC
GCCCATCCTACCCAGCACACACACACCGCTGCTAACCCCATACCCCGAAC
CAACCAAACCCCAAAGACACCCCCCACAGTTTATGTAGCTTACCTCCTCA
AAGCAATACACTGAAAATGTTTAGACGGGCTCACATCACCCCATAAACAA
ATAGGTTTGGTCCTAGCCTTTCTATTAGCTCTTAGTAAGATTACACATGC
AAGCATCCCCGTTCCAGTGAGTTCACCCTCTAAATCACCACGATCAAAAG
GGACAAGCATCAAGCACGCAGCAATGCAGCTCAAAACGCTTAGCCTAGCC
ACACCCCCACGGGAAACAGCAGTGATTAACCTTTAGCAATAAACGAAAGT
TTAACTAAGCTATACTAACCCCAGGGTTGGTCAATTTCGTGCCAGCCACC
GCGGTCACACGATTAACCCAAGTCAATAGAAGCCGGCGTAAAGAGTGTTT
TAGATCACCCCCTCCCCAATAAAGCTAAAACTCACCTGAGTTGTAAAAAA
CTCCAGTTGACACAAAATAGACTACGAAAGTGGCTTTAACATATCTGAAC
ACACAATAGCTAAGACCCAAACTGGGATTAGATACCCCACTATGCTTAGC
CCTAAACCTCAACAGTTAAATCAACAAAACTGCTCGCCAGAACACTACGA
GCCACAGCTTAAAACTCAAAGGACCTGGCGGTGCTTCATATCCCTCTAGA
GGAGCCTGTTCTGTAATCGATAAACCCCGATCAACCTCACCACCTCTTGC
TCAGCCTATATACCGCCATCTTCAGCAAACCCTGATGAAGGCTACAAAGT
AAGCGCAAGTACCCACGTAAAGACGTTAGGTCAAGGTGTAGCCCATGAGG
TGGCAAGAAATGGGCTACATTTTCTACCCCAGAAAACTACGATAGCCCTT
```

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:

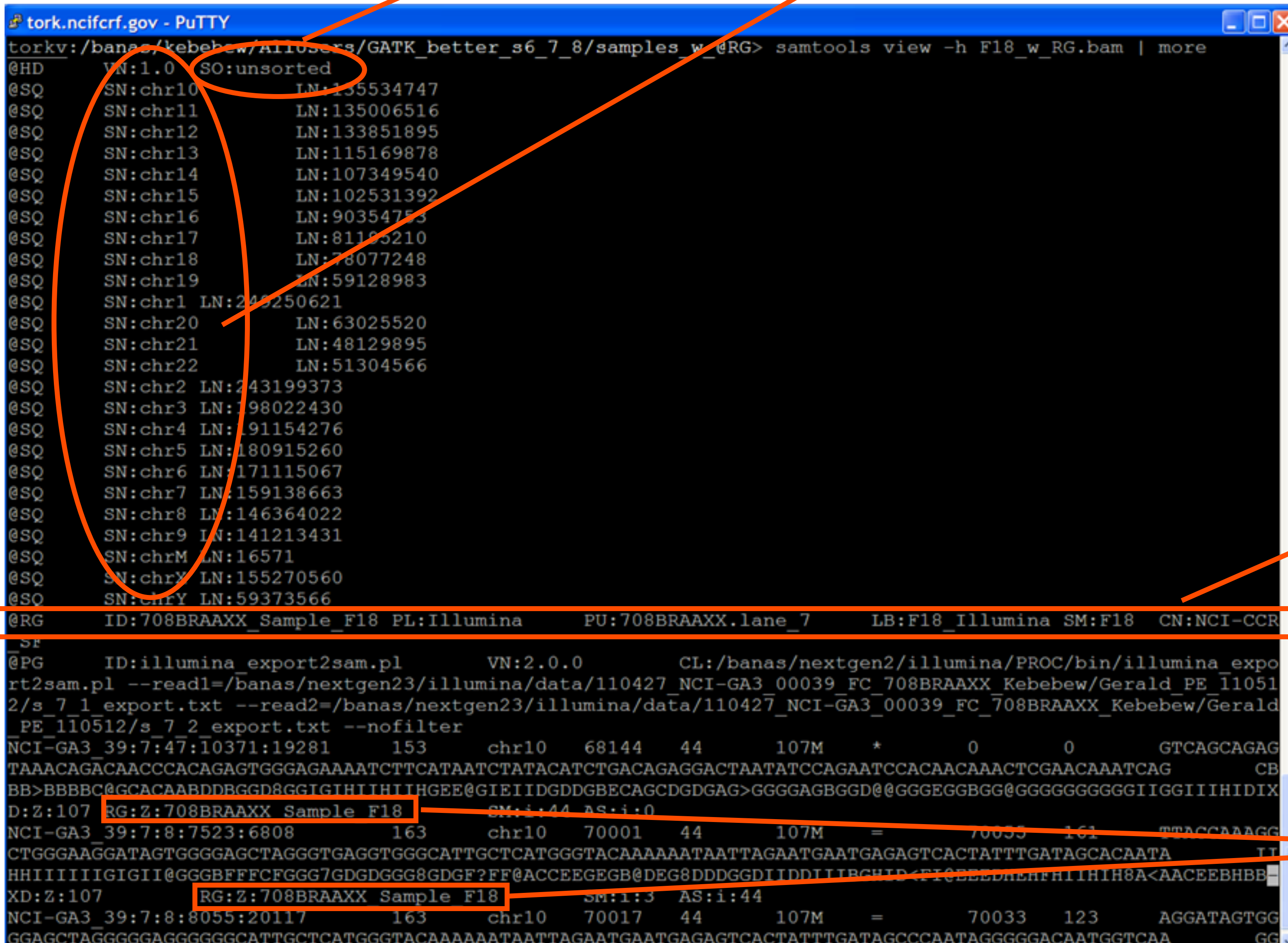
```
java –Xms4g –Xmx4g  -jar
/opt/nasapps/stow/picard-tools-1.70/AddOrReplaceReadGroups.jar
INPUT=/PathToBamFile/before.bam
OUTPUT=/PathToProcessedBamFile/samples_w_@RG/after.bam
RGID=708BRAAXX_Sample_F18
RGLB=F18_illumina
RGPL=Illumina
RGPU=708BRAAXX.lane_7
RGSM=F18
RGCN=NCI-CCR_SF
VALIDATION_STRINGENCY=SILENT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

- 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 decleared as unsorted

chromosomes unordered



```
torkv:/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:81195210
@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
_SF
@PG     ID:illumina_export2sam.pl       VN:2.0.0        CL:/banas/nextgen2/illumina/PROC/bin/illumina_expo
rt2sam.pl --read1=/banas/nextgen23/illumina/data/110427_NCI-GA3_00039_FC_708BRAAXX_Kebebew/Gerald_PE_11051
2/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@GCACAABDDBGGD8GGIGIHIIHIIHGEE@GIEIIDGDDGBECAGCDGDGAG>GGGGAGBGGD@@GGGEGGBGG@GGGGGGGGGIIGGIIIHIDIX
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     TTACCAAAGG
CTGGGAAGGATAGTGGGGAGCTAGGGTGAGGTGGGCATTGCTCATGGGTACAAAAAAATAATTAGAATGAATGAGAGTCACTATTTGATAGCACAATA       II
HHIIIIIIGIGII@GGGBFFFCFGGG7GDGDGGG8GDGF?FF@ACCEEGEGB@DEG8DDDGGDIIDDIIIBGHID<FI@EEEDHEHFHIIHIH8A<AACEEBHBB
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
GGAGCTAGGGGGAGGGGGGCATTGCTCATGGGTACAAAAAAATAATTAGAATGAATGAGAGTCACTATTTGATAGCCCAATAGGGGGACAATGGTCAA       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

Header and reads in the bam file after reordering chromos

reads still decleared as unsorted

chromosomes re-ordered

tork.ncifcrf.gov - PuTTY

```
@HD     VN:1.0  SO:unsorted
@SQ     SN:chrM LN:16571        UR:file:/banas/nextgen2/illumina/PROC/RefGenomes/hg19/ordered/hg19_chrM_1st.fa  M5:d2ed829b8a1628d16cbeee88e88e3
9eb
@SQ     SN:chr1 LN:249250621    UR:file:/banas/nextgen2/illumina/PROC/RefGenomes/hg19/ordered/hg19_chrM_1st.fa  M5:1b22b98cdeb4a9304cb5d48026a85
128
@SQ     SN:chr2 LN:243199373    UR:file:/banas/nextgen2/illumina/PROC/RefGenomes/hg19/ordered/hg19_chrM_1st.fa  M5:a0d9851da00400dec1098a9255ac7
12e
@SQ     SN:chr3 LN:198022430    UR:file:/banas/nextgen2/illumina/PROC/RefGenomes/hg19/ordered/hg19_chrM_1st.fa  M5:641e4338fa8d52a5b781bd2a2c08d
3c3
@SQ     SN:chr4 LN:191154276    UR:file:/banas/nextgen2/illumina/PROC/RefGenomes/hg19/ordered/hg19_chrM_1st.fa  M5:23dccd106897542ad87d2765d28a1
9a1
@SQ     SN:chr5 LN:180915260    UR:file:/banas/nextgen2/illumina/PROC/RefGenomes/hg19/ordered/hg19_chrM_1st.fa  M5:0740173db9ffd264d728f32784845
cd1
@SQ     SN:chr6 LN:171115067    UR:file:/banas/nextgen2/illumina/PROC/RefGenomes/hg19/ordered/hg19_chrM_1st.fa  M5:1d3a93a248d92a729ee764823acbb
c6b
@SQ     SN:chr7 LN:159138663    UR:file:/banas/nextgen2/illumina/PROC/RefGenomes/hg19/ordered/hg19_chrM_1st.fa  M5:618366e953d6aaad97dbe4777c293
75e
@SQ     SN:chr8 LN:146364022    UR:file:/banas/nextgen2/illumina/PROC/RefGenomes/hg19/ordered/hg19_chrM_1st.fa  M5:96f514a9929e410c6651697bded59
aed
@SQ     SN:chr9 LN:141213431    UR:file:/banas/nextgen2/illumina/PROC/RefGenomes/hg19/ordered/hg19_chrM_1st.fa  M5:3e273117f15e0a400f01055d9f393
768
@SQ     SN:chr10        LN:135534747    UR:file:/banas/nextgen2/illumina/PROC/RefGenomes/hg19/ordered/hg19_chrM_1st.fa  M5:988c28e000e84c26d5523
59af1ea2e1d
@SQ     SN:chr11        LN:135006516    UR:file:/banas/nextgen2/illumina/PROC/RefGenomes/hg19/ordered/hg19_chrM_1st.fa  M5:98c59049a2df285c76ffb
1c6db8f8b96
@SQ     SN:chr12        LN:133851895    UR:file:/banas/nextgen2/illumina/PROC/RefGenomes/hg19/ordered/hg19_chrM_1st.fa  M5:51851ac0e1a115847ad36
449b0015864
--More--
```

tork.ncifcrf.gov - PuTTY

```
@RG     ID:708BRAAXX_Sample_F18 PL:Illumina     PU:708BRAAXX.lane_7     LB:F18_Illumina SM:F18  CN:NCI-CCR_SF
@PG     ID:illumina_export2sam.pl       VN:2.0.0        CL:/banas/nextgen2/illumina/PROC/bin/illumina_export2sam.pl --read1=/banas/nextgen23/ill
umina/data/110427_NCI-GA3_00039_FC_708BRAAXX_Kebebew/Gerald_PE_110512/s_7_1_export.txt --read2=/banas/nextgen23/illumina/data/110427_NCI-GA3_000
39_FC_708BRAAXX_Kebebew/Gerald_PE_110512/s_7_2_export.txt --nofilter
NCI-GA3_39:7:18:5713:9912       83      chrM    3       254     107M    =       16497   16387   TCACAGGTCTATCACCCTATTAACCACTCACGGGAGCTCTCCATGCAT
TTGGTATTTTCGTCTGGGGGGTATGCACGCGATAGCATTGCGAGACGCTGGAGCCGGAG     @B=@@A?===9.C4CCB@@@@B9?A=?7E>DCEDDEEDBGFGFFFHFFHHHHFFHHGEEEBGAHHHHGGDGE@GBGGHHH
HHHHHDHHHHHH@HGHHHHHHHHHHHH     XD:Z:70G36      RG:Z:708BRAAXX_Sample_F18       SM:i:503        AS:i:1017
NCI-GA3_39:7:26:13700:19023     99      chrM    4       254     107M    =       123     226     CACAGGTCTATCACCCTATTAACCACTCACGGGAGCTCTCCATGCATT
TGGTATTTTCGTCTGGGGGGTATGCACGCGATAGCATTGCGAGACGCTGGAGCCGGAGC     IIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIHIIIHIIIIIEHIIIIIIIIHIIFIIII=IFHHEHHHHGB
GGHCEGGGGEDDEEEDDE9D;A@@;@@     XD:Z:69G37      RG:Z:708BRAAXX_Sample_F18       SM:i:501        AS:i:958
NCI-GA3_39:7:34:13555:6376      163     chrM    4       254     107M    =       73      176     CACAGGTCTATCACCCTATTAACCACTCACGGGAGCTCTCCATGCATT
TGGTATTTTCGTCTGGGGGGTATGCACGCGATAGCATTGCGAGACGCTGGAGCCGGAGC     IIHIGGBGIGHIIIIIIGIIIIHIIIHIBEGGE8E@?A?AE8CEB;?BEAC;@ACCDDC@EED>EEFF<DCC@EDBDC1B
EEEEBEEEE9B@<BB8@=5B/@<@###     XD:Z:69G37      RG:Z:708BRAAXX_Sample_F18       SM:i:492        AS:i:927
NCI-GA3_39:7:25:5404:10475      83      chrM    13      254     107M    =       16519   16399   ATCCCCCTATTAACCACTCACGGGAGCTCTCCATGCATTTGGTATTTT
CGTCTGGGGGCTATGCACGCGATAGCATTGCGAGACGCTGGAGCCGGAGCACCCTATGT     97?'>==1====/??9?>D>EEEFDBEBDCDFBEFECHIGICHFIHIBHHFHGIHHHI>IIHFIBIIHIIGGHGHIGGGI
IHFIGIIIIIIIIIIHIIIGIIIIIIII    XD:Z:3A54G1G46  RG:Z:708BRAAXX_Sample_F18       SM:i:456        AS:i:1002
NCI-GA3_39:7:107:17432:11232    99      chrM    14      254     107M    =       64      157     TCACCCTATTAACCACTCACGGGAGCTCTCCATGCATTTGGTATTTTC
GTCTGGGGGGTATGCACGCGATAGCATTGCGAGACGCTGGAGCCGGAGCACCCTATGTC     IIIIIIIIIIIIIIIIIIIIIIIIIIHIIIIIIIEHIIGIIIIHGIIIIIHIIIHHHIIGICGIGHBHHIIIIGIIIFIIIIG
```

Although Reads already sorted in coordinate order by Samtools sort, but the @HD SO tag still as “unsorted”

Sort the reads in the bam files

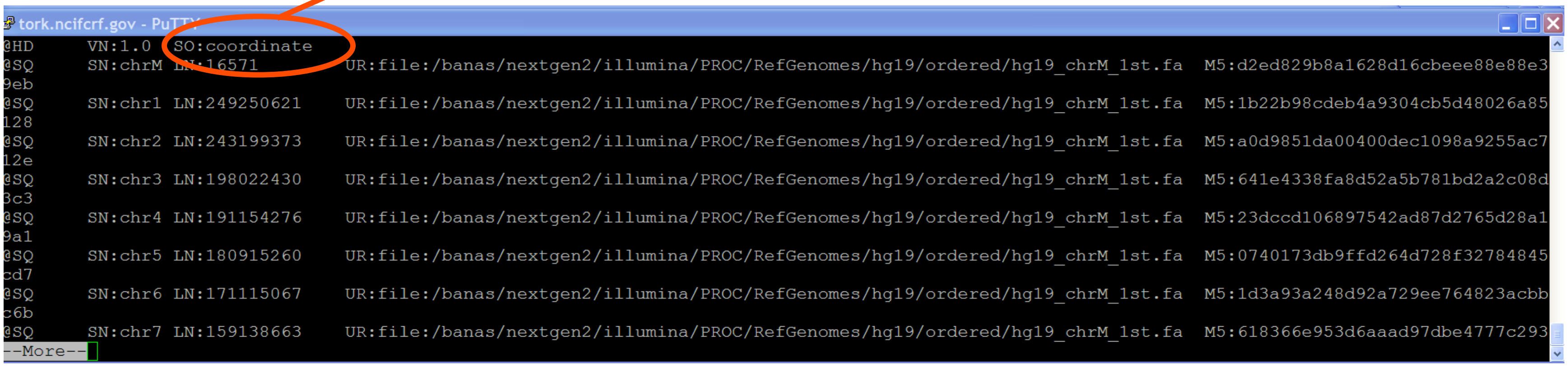
Typical Command Used:

```
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
```

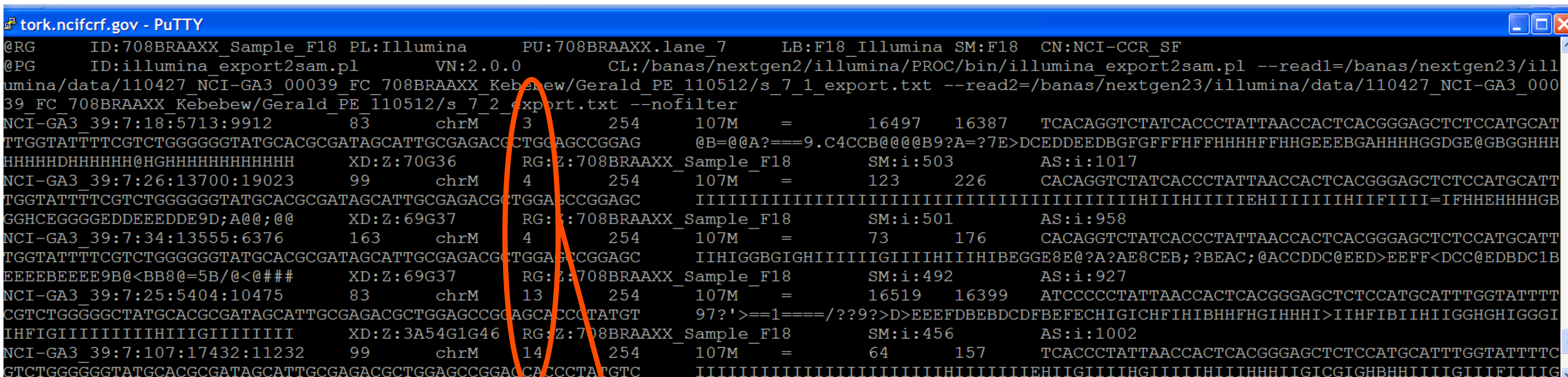
- Use picard SortSam to sort the alignment reads in the bam file
- Makes sure use SORT_ORDER=coordinate, which is required by many tools used in the GATK pipeline (e.g., MarkDuplicates)
- Use picard SortSam to make sure the header is declared as "coordinate" sorted, (samtools sort sorts the reads but 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

Header and reads in the bam file after sorting reads

Reads now declared as sorted in coordinate order
By using picard SortSam



```
@HD     VN:1.0  SO:coordinate
@SQ     SN:chrM LN:16571    UR:file:/banas/nextgen2/illumina/PROC/RefGenomes/hg19/ordered/hg19_chrM_1st.fa  M5:d2ed829b8a1628d16cbeee88e88e3
9eb
@SQ     SN:chr1 LN:249250621    UR:file:/banas/nextgen2/illumina/PROC/RefGenomes/hg19/ordered/hg19_chrM_1st.fa  M5:1b22b98cdeb4a9304cb5d48026a85
128
@SQ     SN:chr2 LN:243199373    UR:file:/banas/nextgen2/illumina/PROC/RefGenomes/hg19/ordered/hg19_chrM_1st.fa  M5:a0d9851da00400dec1098a9255ac7
12e
@SQ     SN:chr3 LN:198022430    UR:file:/banas/nextgen2/illumina/PROC/RefGenomes/hg19/ordered/hg19_chrM_1st.fa  M5:641e4338fa8d52a5b781bd2a2c08d
3c3
@SQ     SN:chr4 LN:191154276    UR:file:/banas/nextgen2/illumina/PROC/RefGenomes/hg19/ordered/hg19_chrM_1st.fa  M5:23dccd106897542ad87d2765d28a1
9a1
@SQ     SN:chr5 LN:180915260    UR:file:/banas/nextgen2/illumina/PROC/RefGenomes/hg19/ordered/hg19_chrM_1st.fa  M5:0740173db9ffd264d728f32784845
cd7
@SQ     SN:chr6 LN:171115067    UR:file:/banas/nextgen2/illumina/PROC/RefGenomes/hg19/ordered/hg19_chrM_1st.fa  M5:1d3a93a248d92a729ee764823acbb
c6b
@SQ     SN:chr7 LN:159138663    UR:file:/banas/nextgen2/illumina/PROC/RefGenomes/hg19/ordered/hg19_chrM_1st.fa  M5:618366e953d6aaad97dbe4777c293
--More--
```



```
@RG     ID:708BRAAXX_Sample_F18 PL:Illumina     PU:708BRAAXX.lane_7     LB:F18_Illumina SM:F18  CN:NCI-CCR_SF
@PG     ID:illumina_export2sam.pl       VN:2.0.0        CL:/banas/nextgen2/illumina/PROC/bin/illumina_export2sam.pl --read1=/banas/nextgen23/ill
umina/data/110427_NCI-GA3_00039_FC_708BRAAXX_Kebebew/Gerald_PE_110512/s_7_1_export.txt --read2=/banas/nextgen23/illumina/data/110427_NCI-GA3_000
39_FC_708BRAAXX_Kebebew/Gerald_PE_110512/s_7_2_export.txt --nofilter
NCI-GA3_39:7:18:5713:9912       83      chrM    3       254     107M    =       16497   16387   TCACAGGTCTATCACCCTATTAACCACTCACGGGAGCTCTCCATGCAT
TTGGTATTTTCGTCTGGGGGGTATGCACGCGATAGCATTGCGAGACGCTGGAGCCGGAG     @B=@@A?===9.C4CCB@@@@B9?A=?7E>DCEDDEEDBGFGFFFHFFHHHHFFHHGEEEBGAHHHHGGDGE@GBGGHHH
HHHHHDHHHHHH@HGHHHHHHHHHHHH     XD:Z:70G36      RG:Z:708BRAAXX_Sample_F18       SM:i:503        AS:i:1017
NCI-GA3_39:7:26:13700:19023     99      chrM    4       254     107M    =       123     226     CACAGGTCTATCACCCTATTAACCACTCACGGGAGCTCTCCATGCATT
TGGTATTTTCGTCTGGGGGGTATGCACGCGATAGCATTGCGAGACGCTGGAGCCGGAGC     IIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIHIIIHIIIIIEHIIIIIIIIHIIFIIII=IFHHEHHHHGB
GGHCEGGGGEDDEEEDDE9D;A@@;@@     XD:Z:69G37      RG:Z:708BRAAXX_Sample_F18       SM:i:501        AS:i:958
NCI-GA3_39:7:34:13555:6376      163     chrM    4       254     107M    =       73      176     CACAGGTCTATCACCCTATTAACCACTCACGGGAGCTCTCCATGCATT
TGGTATTTTCGTCTGGGGGGTATGCACGCGATAGCATTGCGAGACGCTGGAGCCGGAGC     IIHIGGBGIGHIIIIIIGIIIIHIIIIHIBEGGE8E@?A?AE8CEB;?BEAC;@ACCDDC@EED>EEFF<DCC@EDBDC1B
EEEEBEEEE9B@<BB8@=5B/@<@###     XD:Z:69G37      RG:Z:708BRAAXX_Sample_F18       SM:i:492        AS:i:927
NCI-GA3_39:7:25:5404:10475      83      chrM    13      254     107M    =       16519   16399   ATCCCCCTATTAACCACTCACGGGAGCTCTCCATGCATTTGGTATTTT
CGTCTGGGGGCTATGCACGCGATAGCATTGCGAGACGCTGGAGCCGGAGCACCCTATGT     97?'>==1====/??9?>D>EEEFDBEBDCDFBEFECHIGICHFIHIBHHFHGIHHHI>IIHFIBIIHIIGGHGHIGGI
IHFIGIIIIIIIIIHIIIGIIIIIIII     XD:Z:3A54G1G46  RG:Z:708BRAAXX_Sample_F18       SM:i:456        AS:i:1002
NCI-GA3_39:7:107:17432:11232    99      chrM    14      254     107M    =       64      157     TCACCCTATTAACCACTCACGGGAGCTCTCCATGCATTTGGTATTTTC
GTCTGGGGGGTATGCACGCGATAGCATTGCGAGACGCTGGAGCCGGAGCACCCTATGTC     IIIIIIIIIIIIIIIIIIIIIIIIIHIIIIIIIIEHIIGIIIIHGIIIIIHIIIHHHIIGICGIGHBHHIIIIGIIIFIIIIG
```

Reads all in coordinate order

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, c
hr20, 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, c
hr20, 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 -l 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 -l 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 -l 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

Used by next step

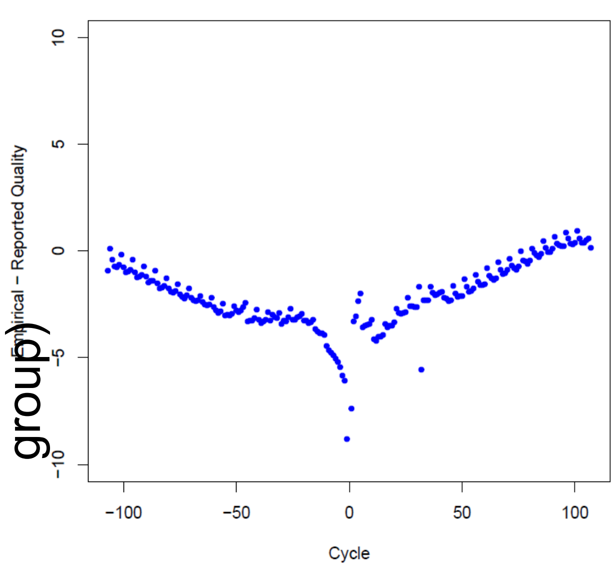
PrintReads (create recalibrated bam file)

```
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
```

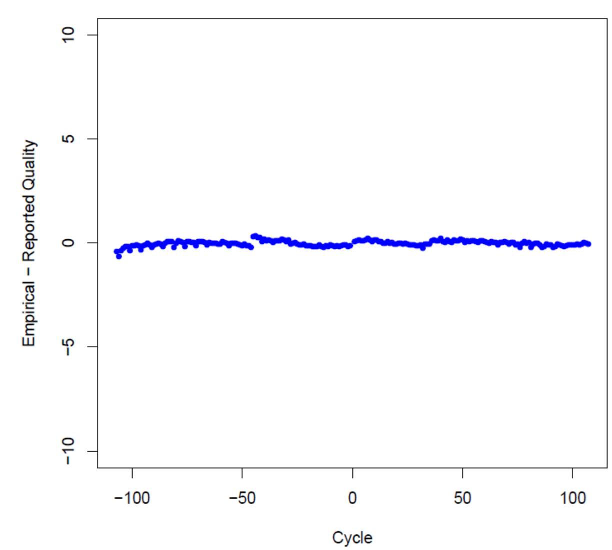
Benefit of Base Quality Score Recalibration

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

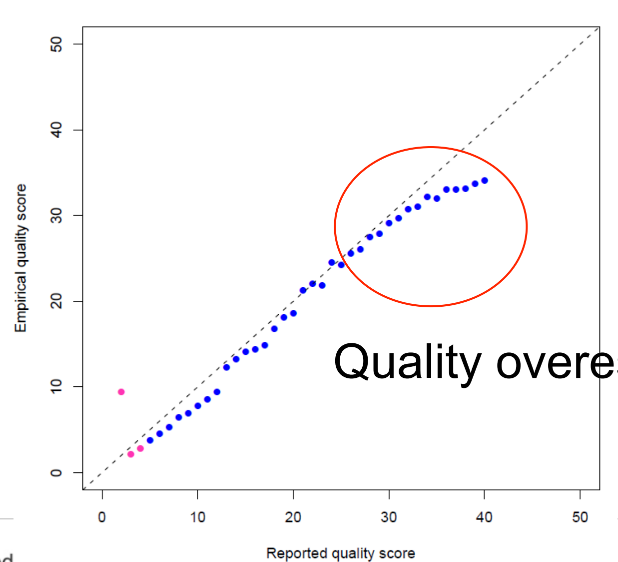
Original



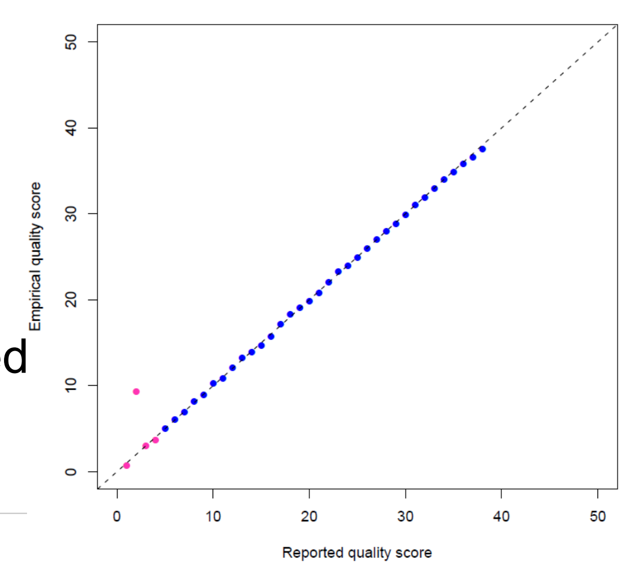
After



Reported vs Empirical Quality



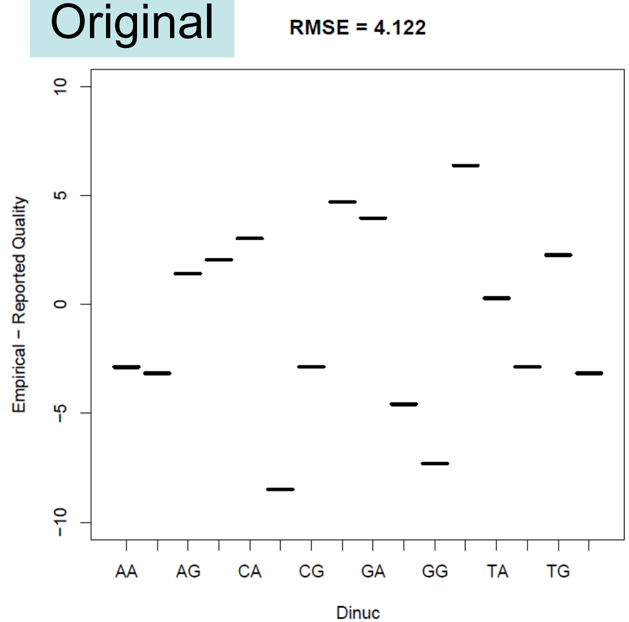
Quality overestimated



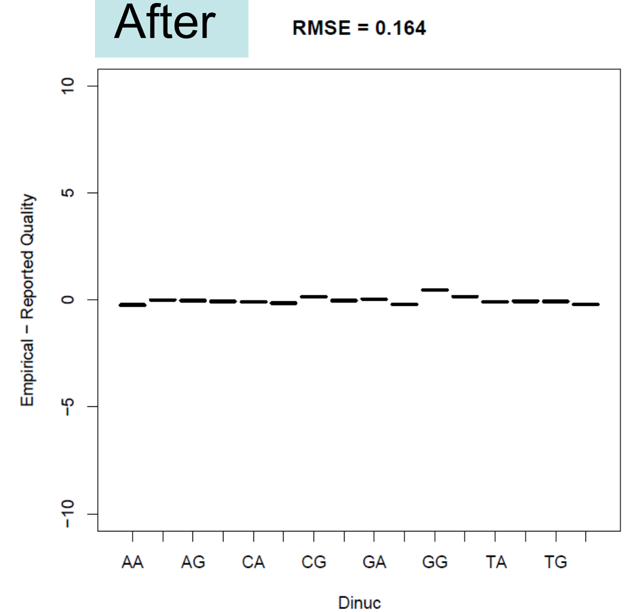
Benefit of Base Quality Score Recalibration

Residual Error by Dinucleotide

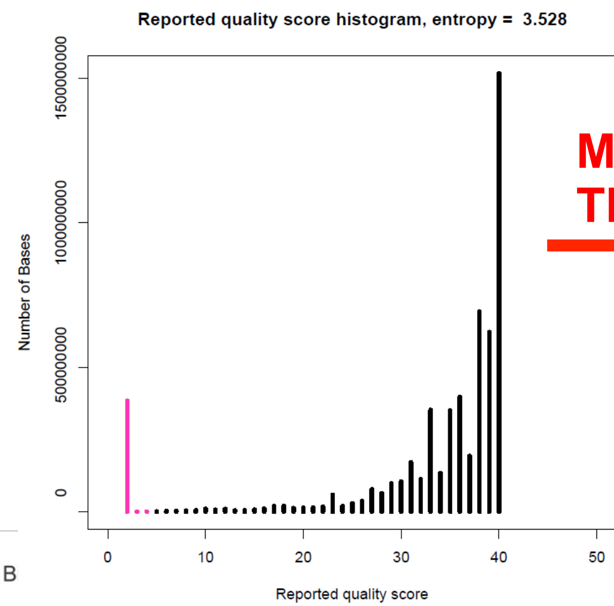
Original



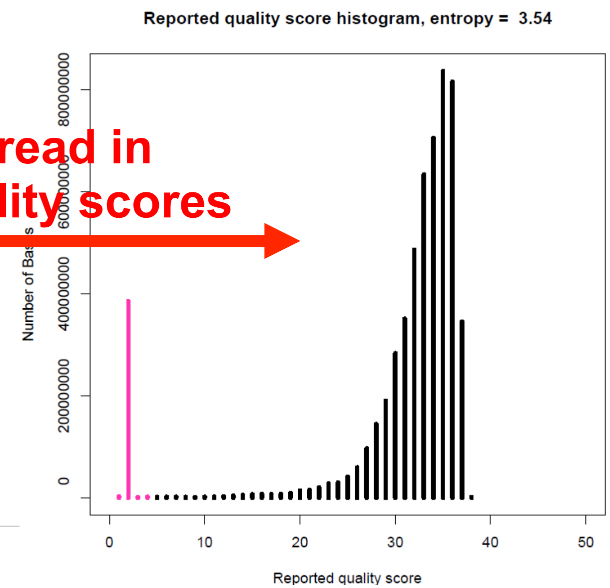
After



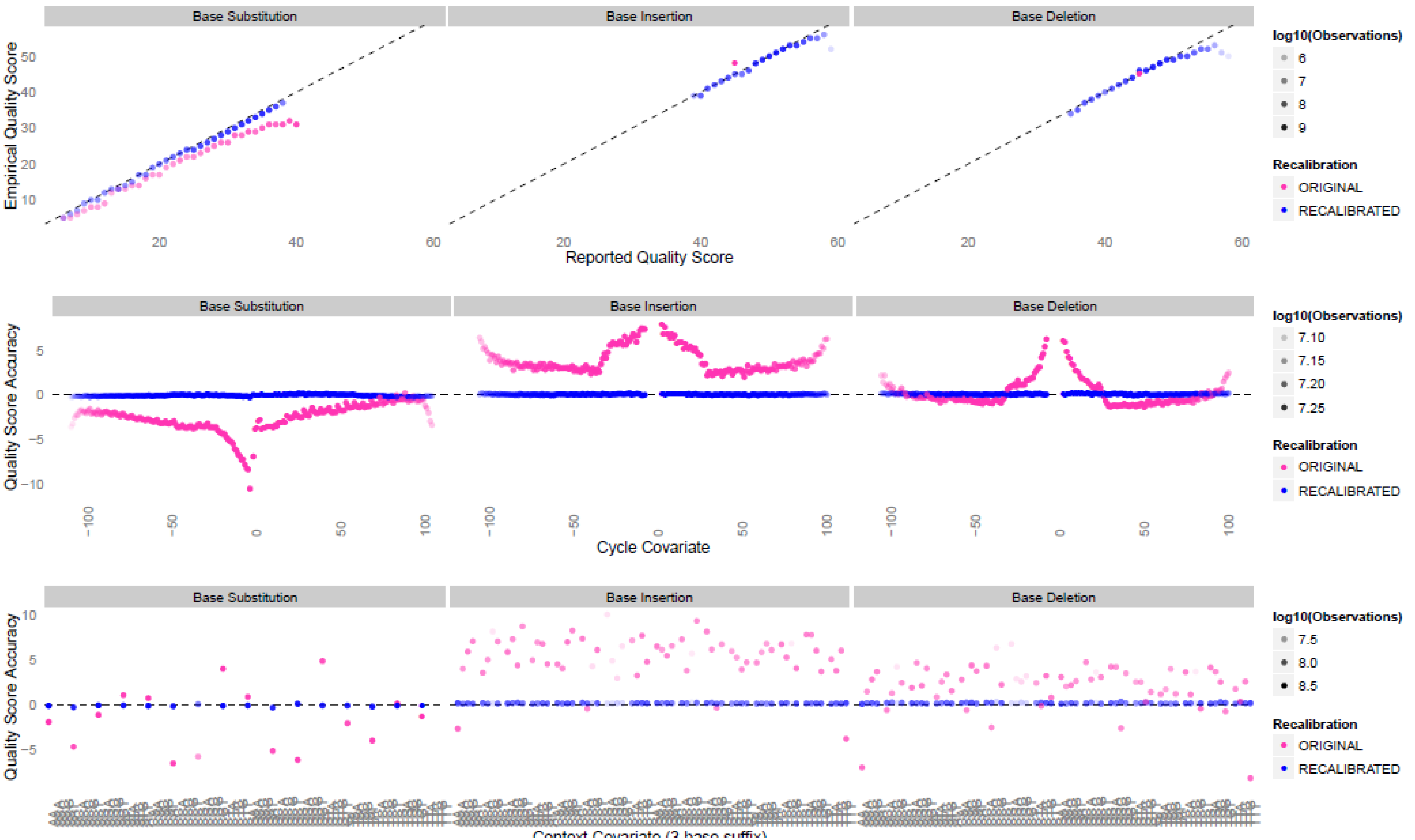
Reported Quality Scores Distribution



More spread in
The quality scores



Example Plots of Base Quality Score Recalibration For GATK Version2.0



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, HyplotyperCaller (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

HyplotypeCaller (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_SelSNP.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_SelSNP.VarRecal
-tranchesFile /Path/phase_II_initialSNPCalls/phase_III/GATK_UG_AllSamples_SelSNP.tranches
-rscriptFile /Path/phase_II_initialSNPCalls/phase_III/GATK_UG_AllSamples_SelSNP.R
-mode SNP
```

Used by next step

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_SelSNP.vcf
-ts_filter_level 99.0
-tranchesFile /Path/phase_II_initialSNPCalls/phase_III/GATK_UG_AllSamples_SelSNP.tranches
-recalFile /Path/phase_II_initialSNPCalls/phase_III/GATK_UG_AllSamples_SelSNP.VarRecal
-o /Path/phase_II_initialSNPCalls/phase_III/GATK_UG_AllSamples_SelSNP.recalibrated.filtered.099.vcf
-mode SNP
```

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 lis
ted">
##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 w
here 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 o
rder as listed">
##INFO=<ID=AF,Number=.,Type=Float,Description="Allele Frequency, for each ALT allele, in the same order as lis
ted">
##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 ba
se 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_EXPERIMENTA
L ignoreSNPAlleles=false output_all_callable_bases=false genotype=false out=org.broadinstitute.sting.gatk.io.s
tubs.VCFWriterStub NO_HEADER=org.broadinstitute.sting.gatk.io.stubs.VCFWriterStub sites_only=org.broadinstitut
e.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;S
B=-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%)
```


General resource: http://vcftools.sourceforge.net/specs.html

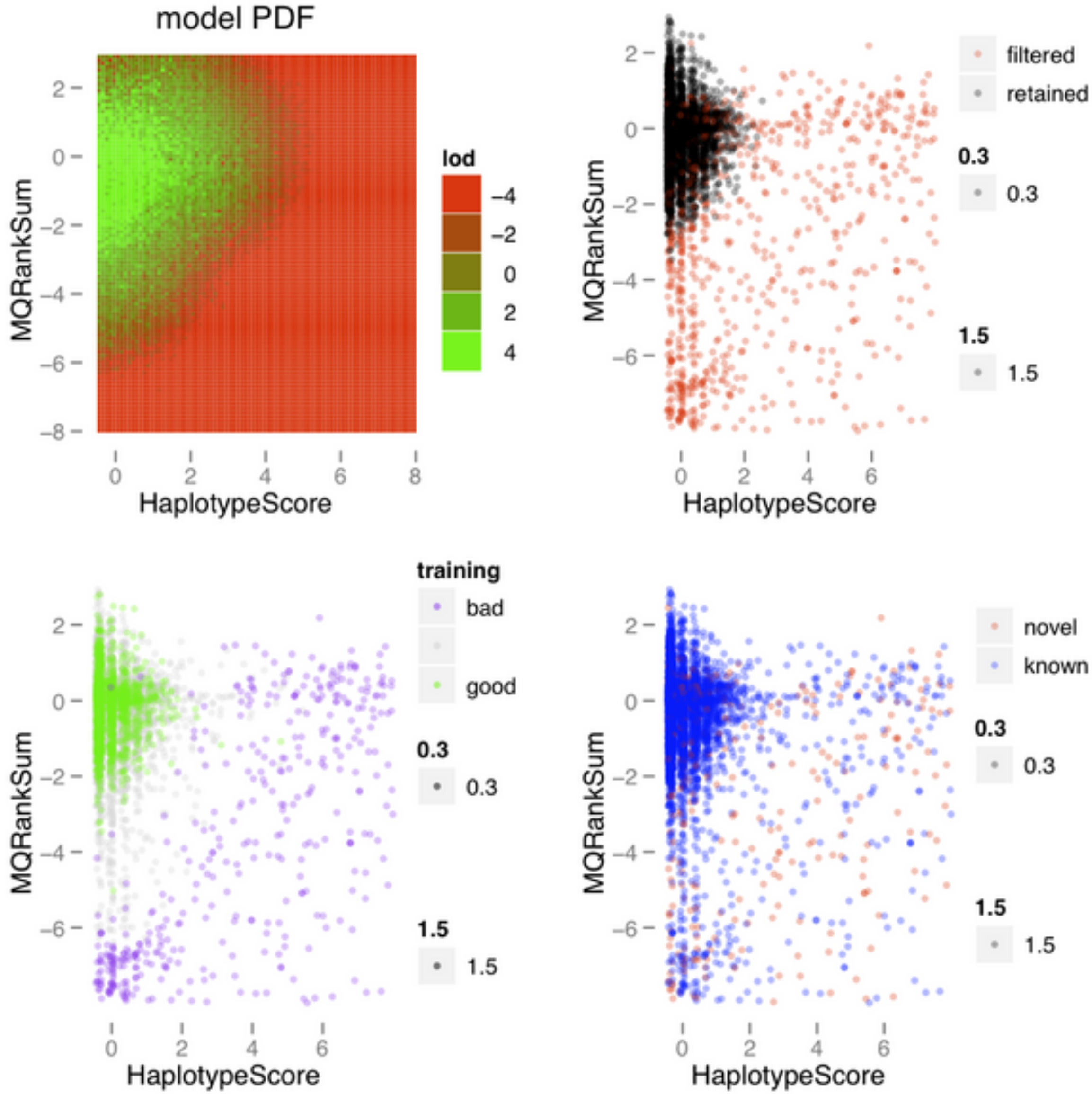
VCF File QC

- **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?

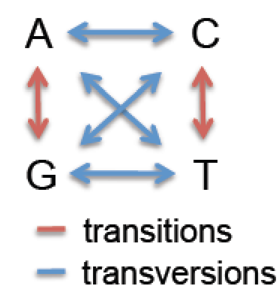
- Often we have genotype chip data that indicates the hom-ref, het, hom-var status at millions of sites
- Good SNP calls should be >99.5% consistent these chip results, and >99% of the variable sites should be found
- The chip sites are in the better parts of the genome, and so are not representative 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 has Ti/Tv around 0.5
- Ti/Tv is a good metric for assessing SNP call quality



Mark Depristo (Broad GATK Team)

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

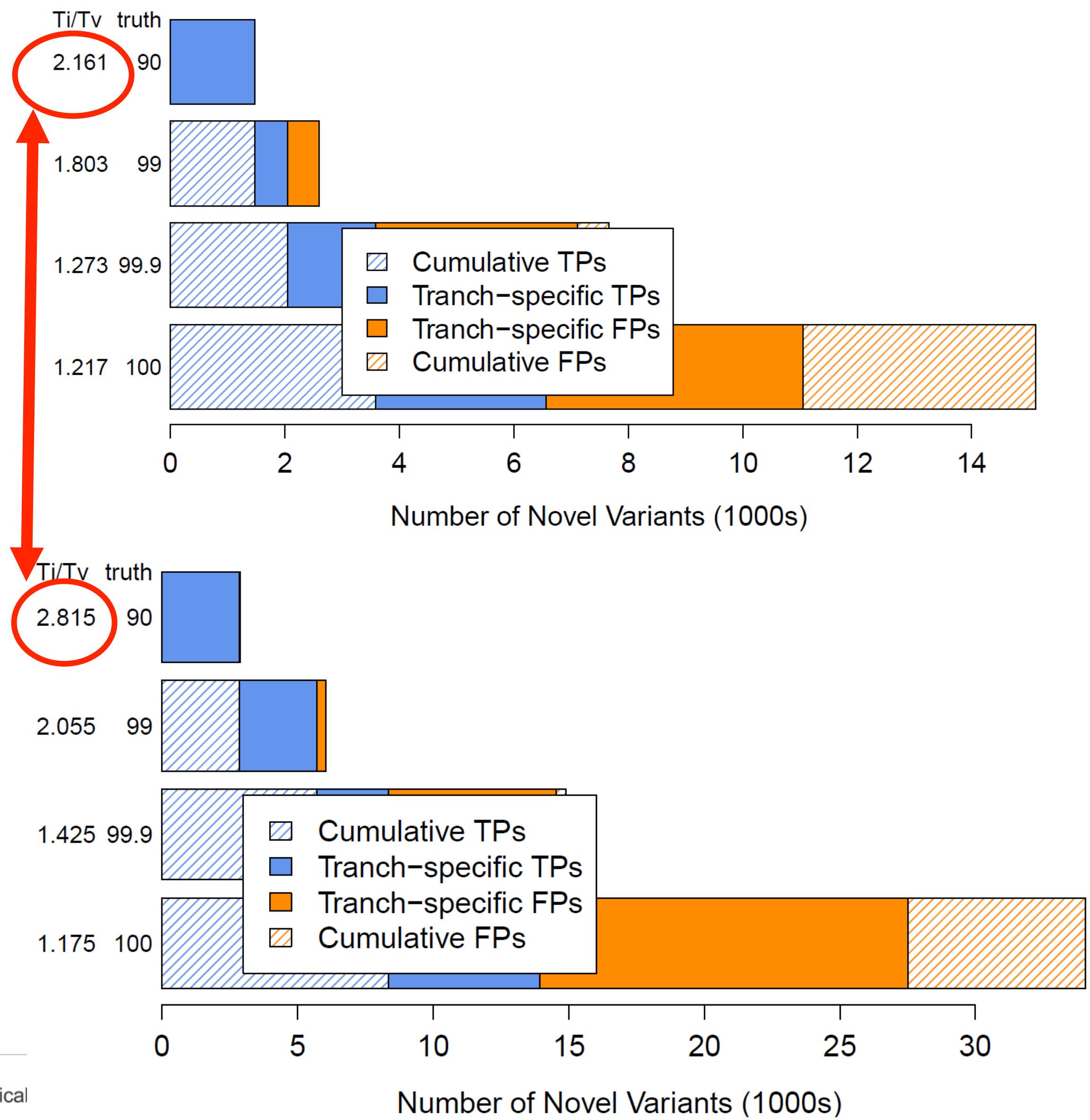
SNP call set with 19-sample dataset

Note: All SNP calls within target interval

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

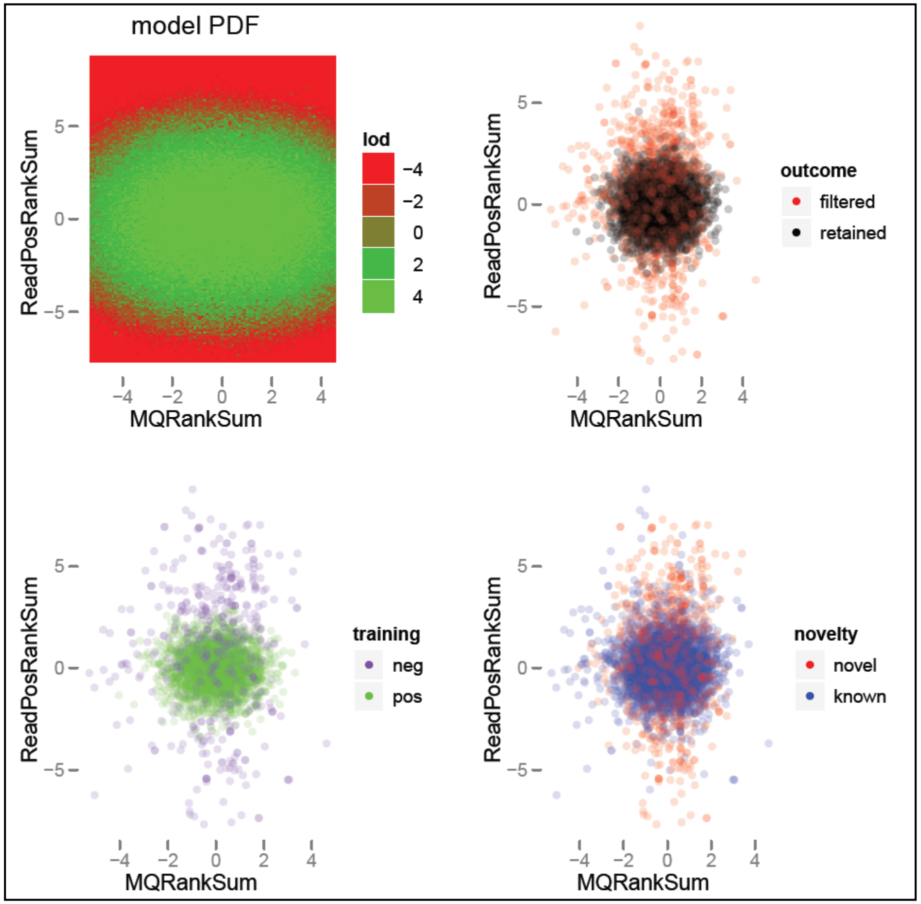
SNP callset with only 3 samples

SNP callset with all 19 samples

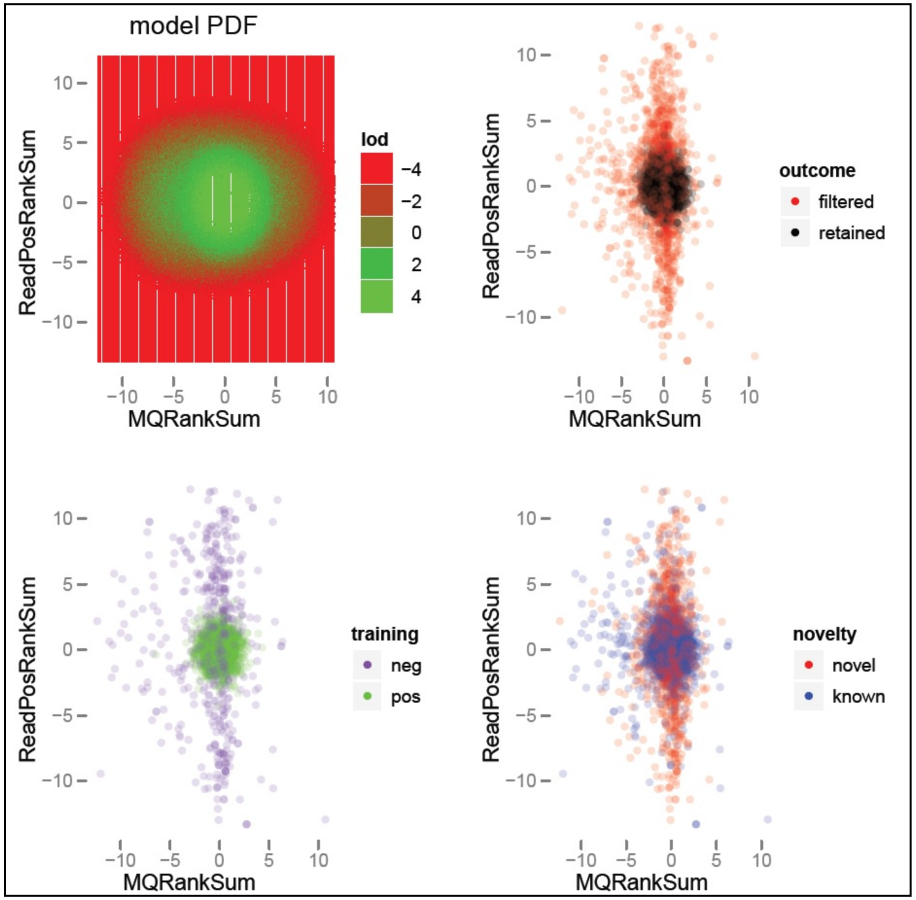


VQSR: Pair-wise combination of annotations used in modeling 2D projection of the Gaussian mixture model is shown (ReadPosRankSum vs MQRankSum)

3-sample dataset



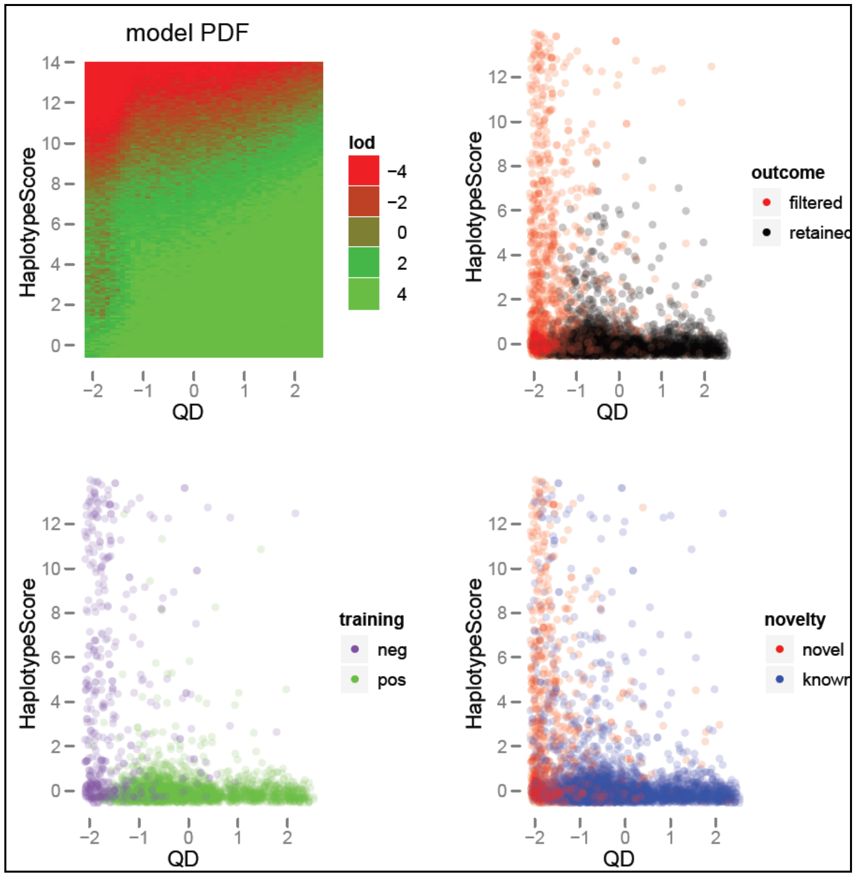
19-sample dataset



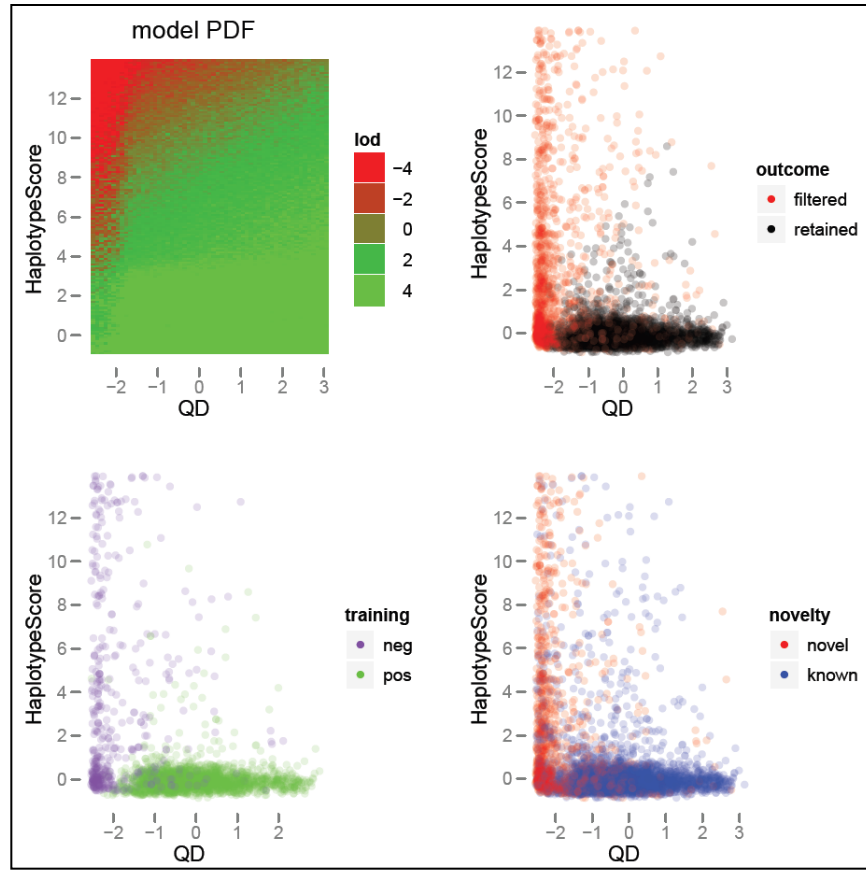
Clustered more tightly in 19-sample data indicates improved separation

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



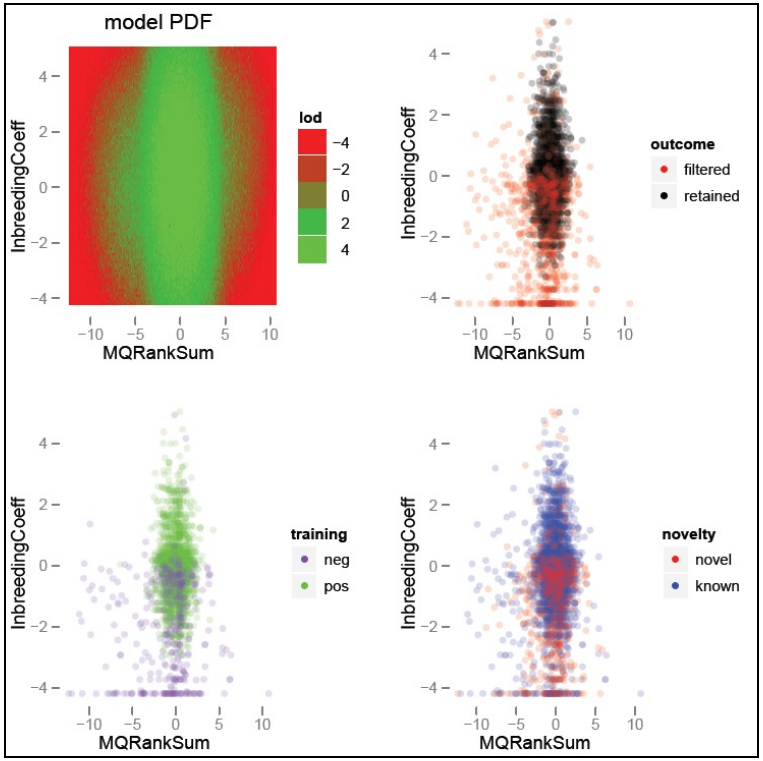
SNP call set with all 19 samples



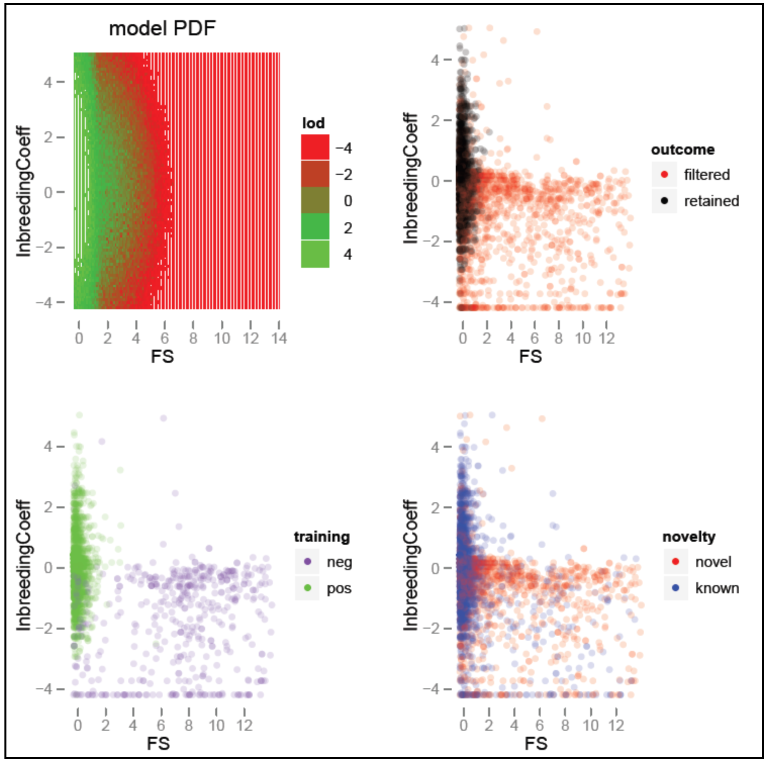
Many annotations used to help modeling

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

InbreedingCoeff vs MQRankSum

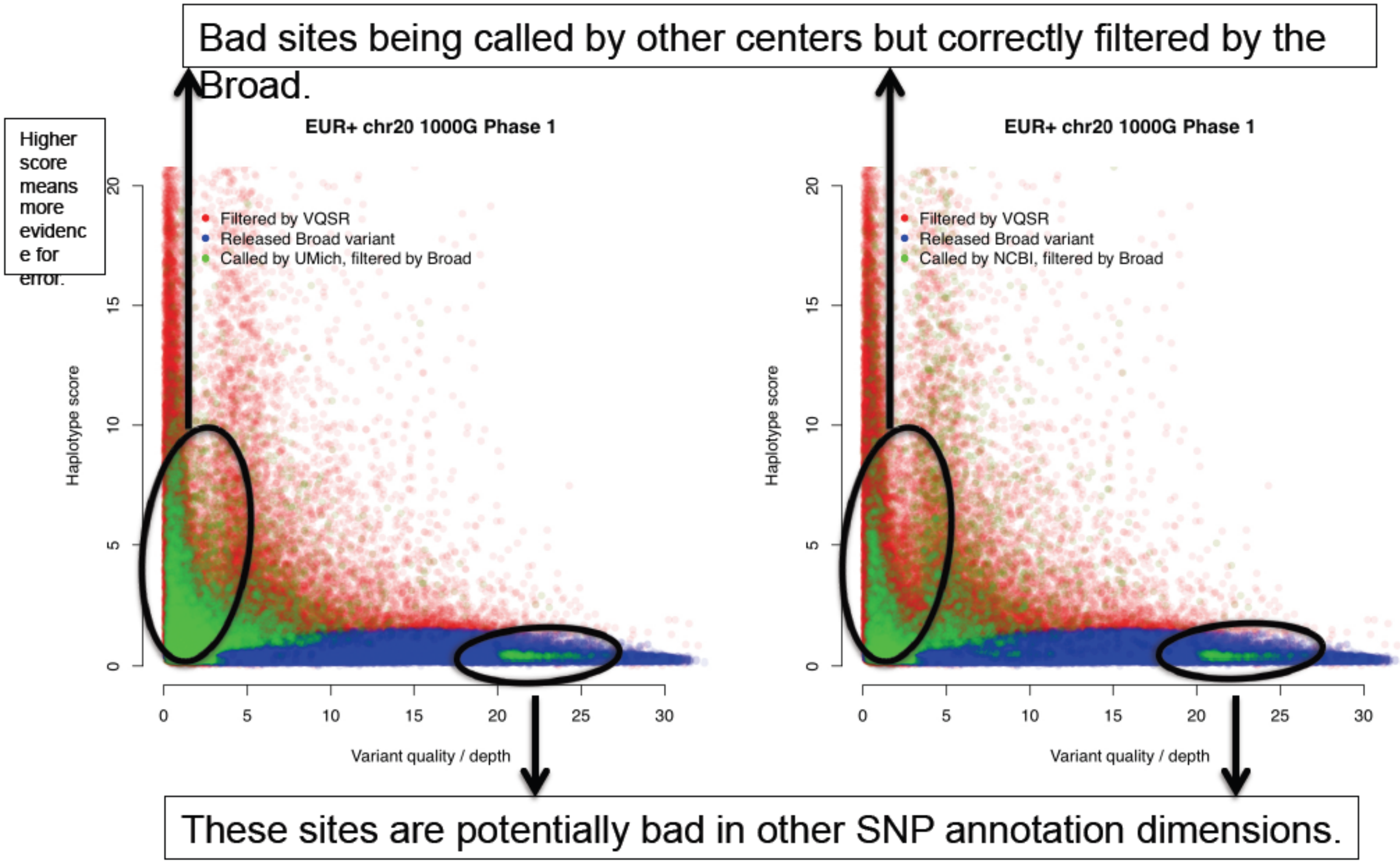


InbreedingCoeff vs FS



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

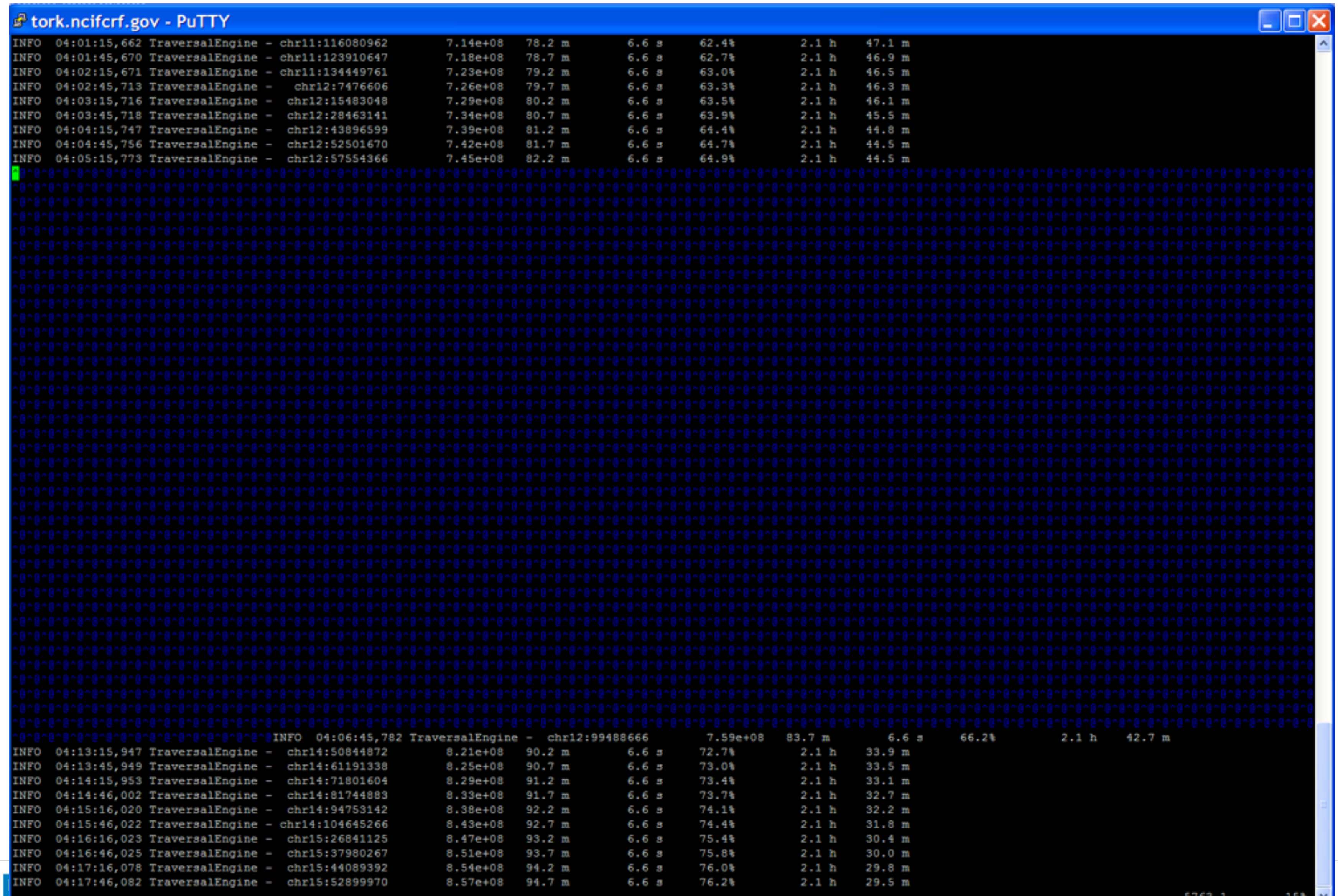
Variants with bad Haplotype Scores often exhibit good Ti/Tv ratios and are included in other centers' callsets, but are likely FPs



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



IT infrastructure and system stability is critical for NGS

Occurence: GATK Phase II Unified Genotyper (UG) for SNP/Indel Call For individual samples

Observation:

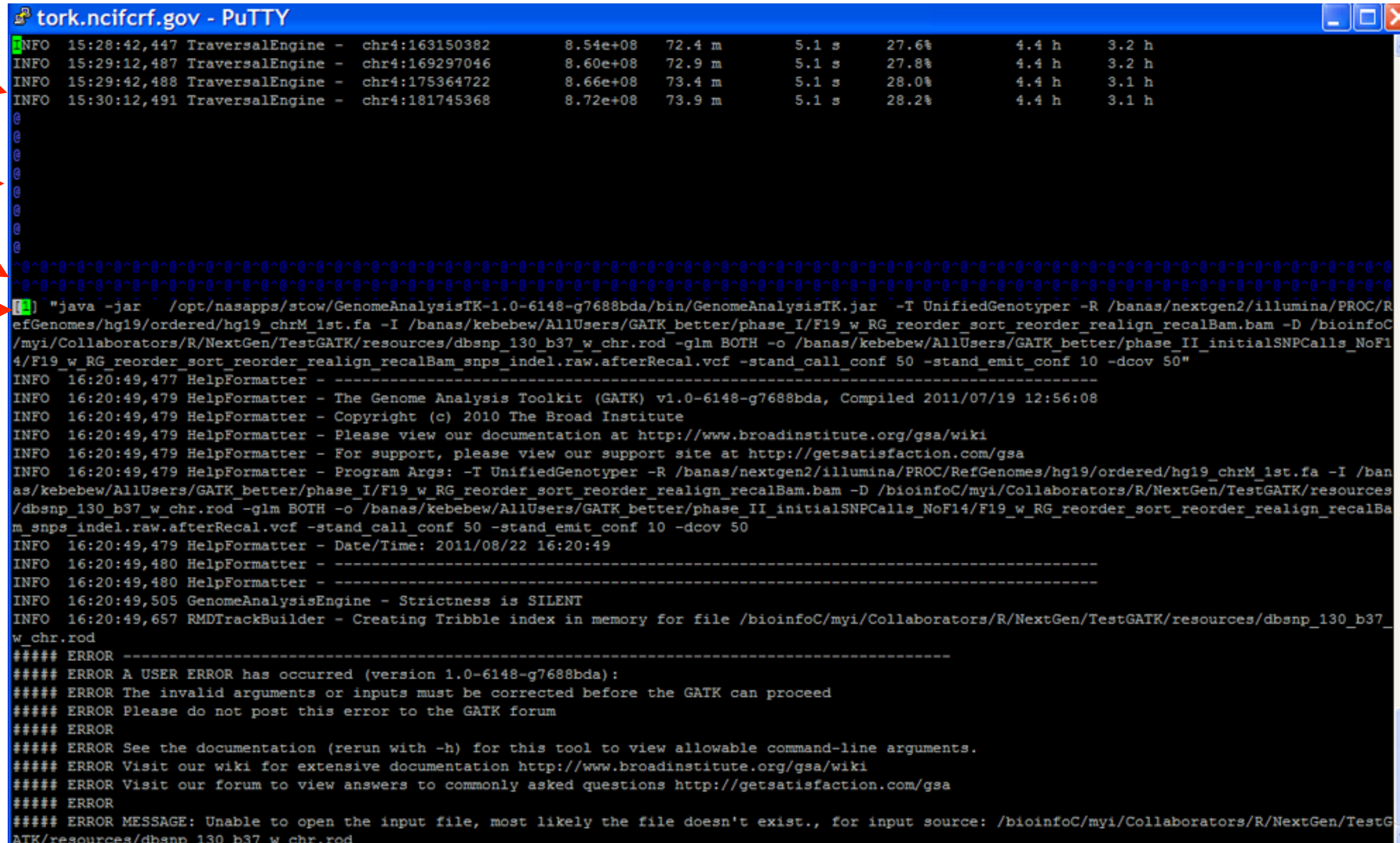
1. The output vcf files for this set of data used to have ~88M-129M, one sample bam file has only 24M
2. Also Missing many samples' vcf files (F12-F19).
3. For one sample: got ERROR MESSAGE: unable to open a supporting file, which has been used for every call

Unfinished log for UG call of one sample

These Alien Characters show only as one line for each

Call UG on the next sample

Got Error message, However, this file has been used as support file for every UG call



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

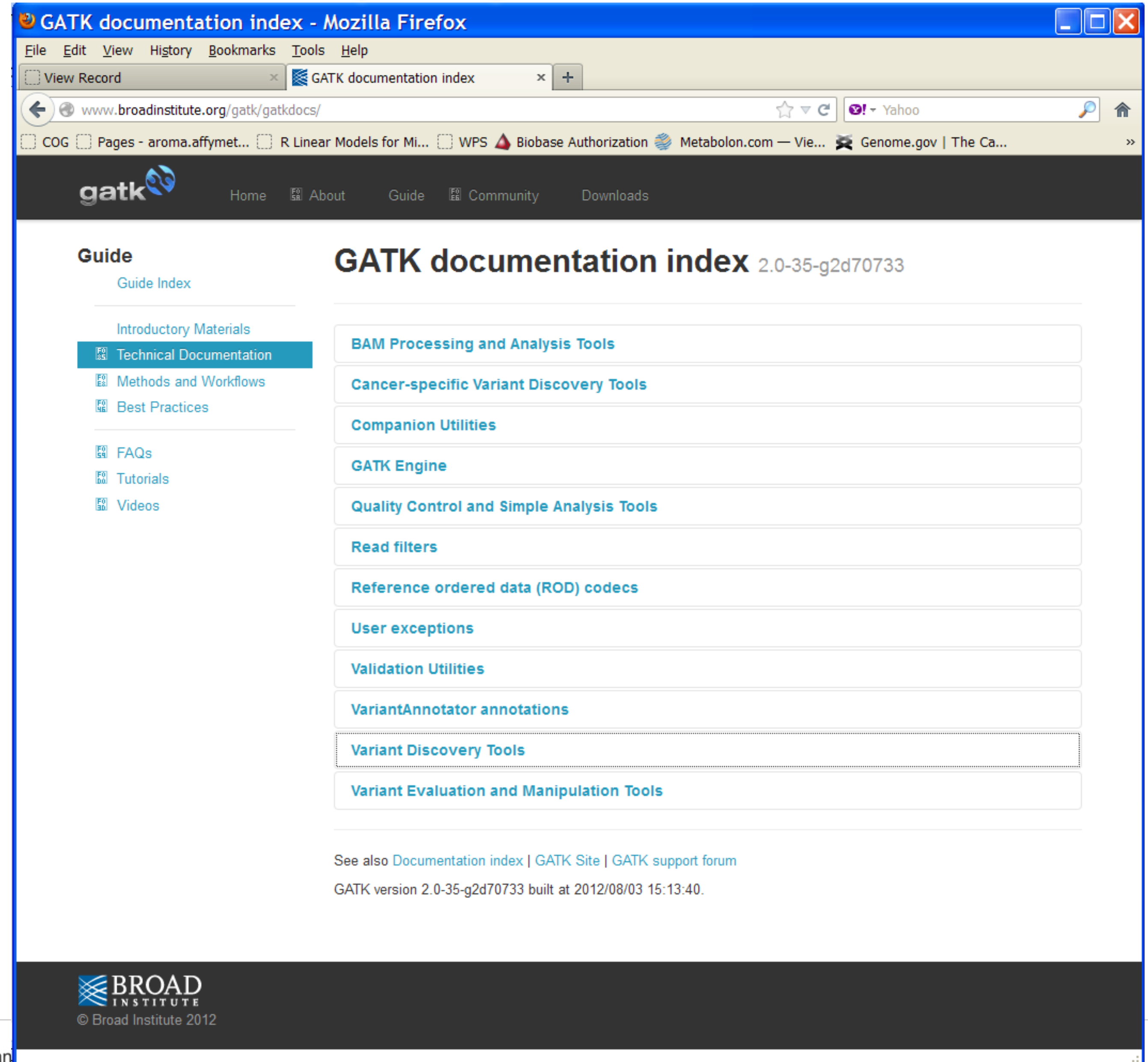
# samples	Center	Total # variants	dbSNP % (129)	# knowns	Known ti/tv	# novels	Novel ti/tv	Includes genotype refinement?
1004	Broad	**765,365**	24.82	190,000	2.36	575,365	2.37	No
1004	BC	**733,155**	25.34	185,787	2.37	547,368	2.32	No
1004	Sanger	**728,374**	25.31	184,341	2.36	544,033	2.36	No
1004	UMich	**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)

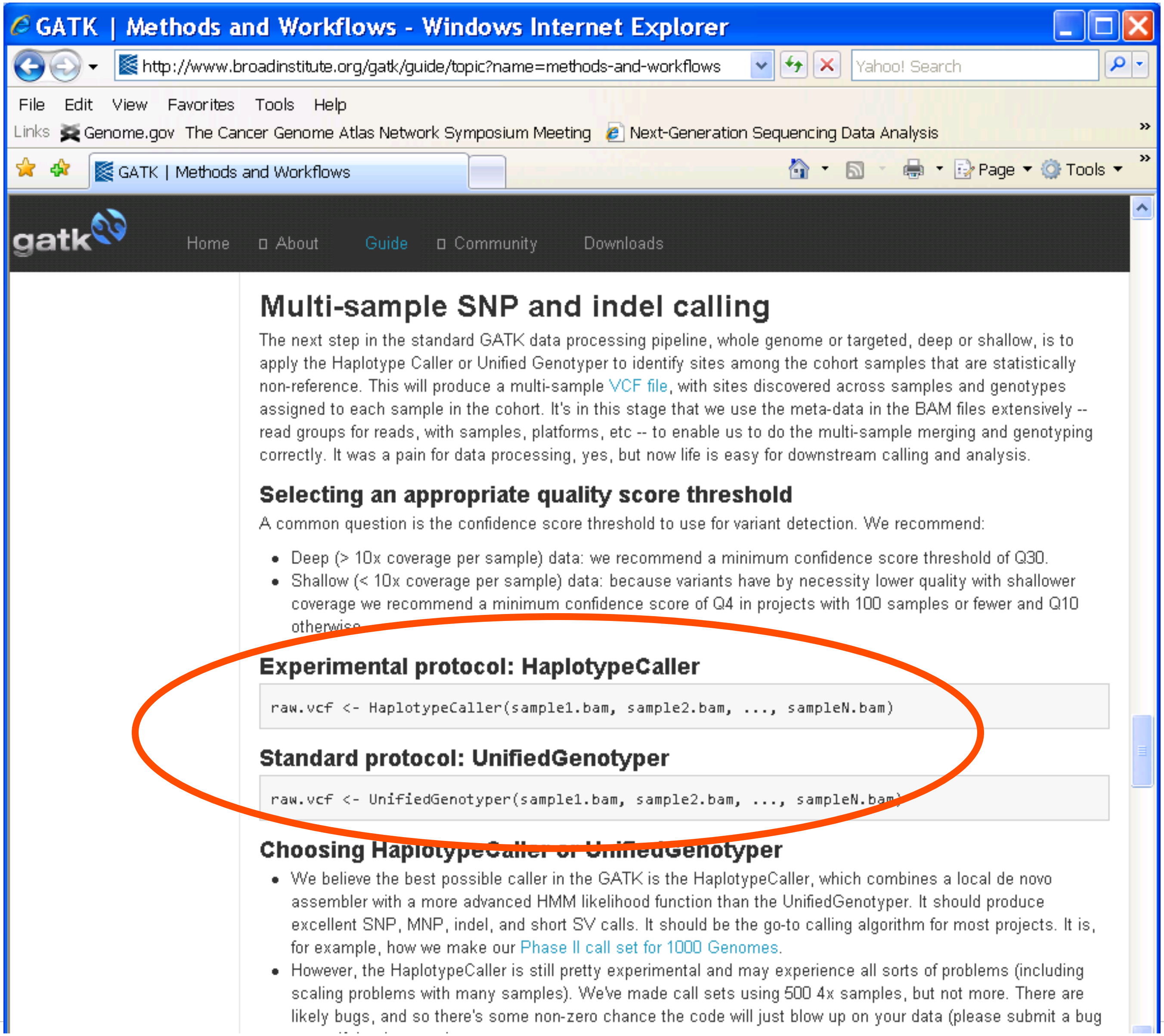
Summary of GATK Testing

- 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 Shortgun Experiment and hand filtering is recommended for exome indels.

New GATK (v2.0) Website-Documentation



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

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	Haplotype Caller
Propose Haplotypes	Look for coincident events in the read data. Must be seen at least 5 times.	Local de novo assembly via DeBruijn graphs. Paths through graph are weighted by number reads which support each kmer.
Evaluate Haplotypes	Pair HMM evaluates the likelihood of the proposed event with the reference. Affine gap penalties based on homopolymer context.	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.
Assign Genotypes	“Exact model” from Heng Li chooses optimal configuration.	No change from UG.

Ryan Poplin (Broad GATK Team)

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

Caller	Variant Sensitivity (strict)	Genotype Concordance (strict)	Variant Sensitivity (strict)	Genotype Concordance (strict)
	Mullikin		Mills	
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)

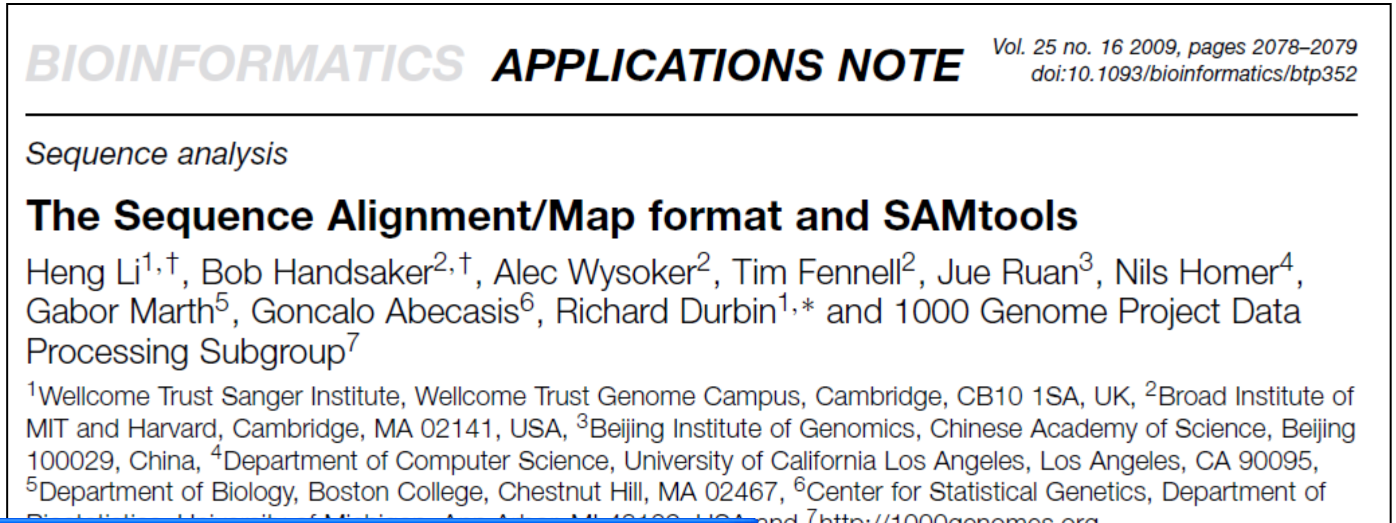
- Input data is NA12878 b37+decoy WGS HiSeq high coverage
- Sites chosen to be very difficult (het) but high confidence in being real (require family transmission)
- Evaluation sets
  - Mullikin Fosmids and Mills et al, GR, 2011 (2x hit, double center)
  - Large events (> 15 bp), largest is 106bp (which 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

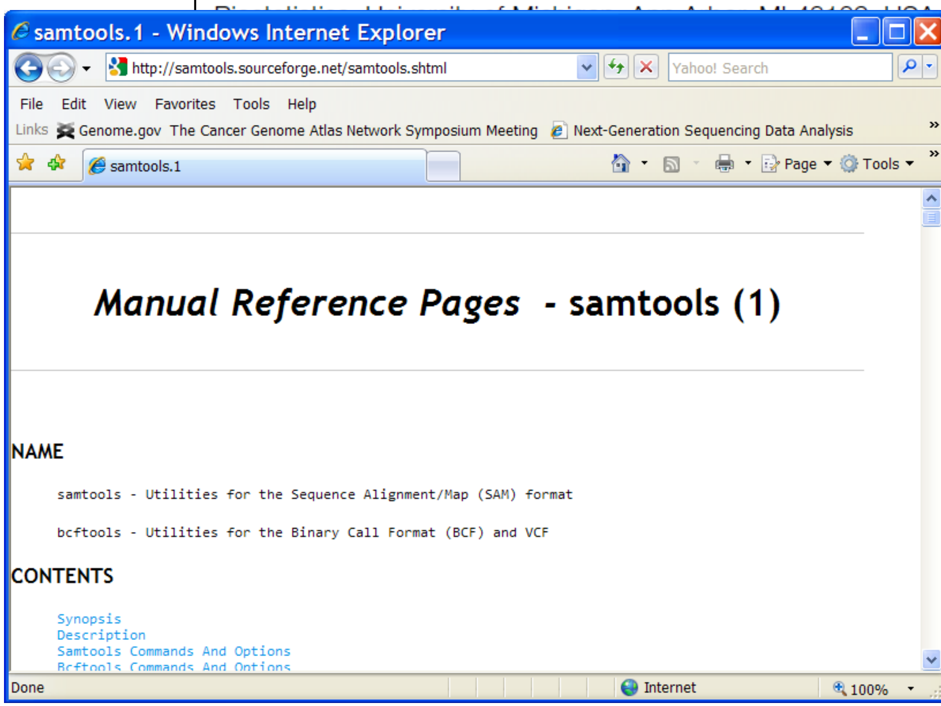
Vol. 25 no. 16 2009, pages 2078–2079
doi:10.1093/bioinformatics/btp352

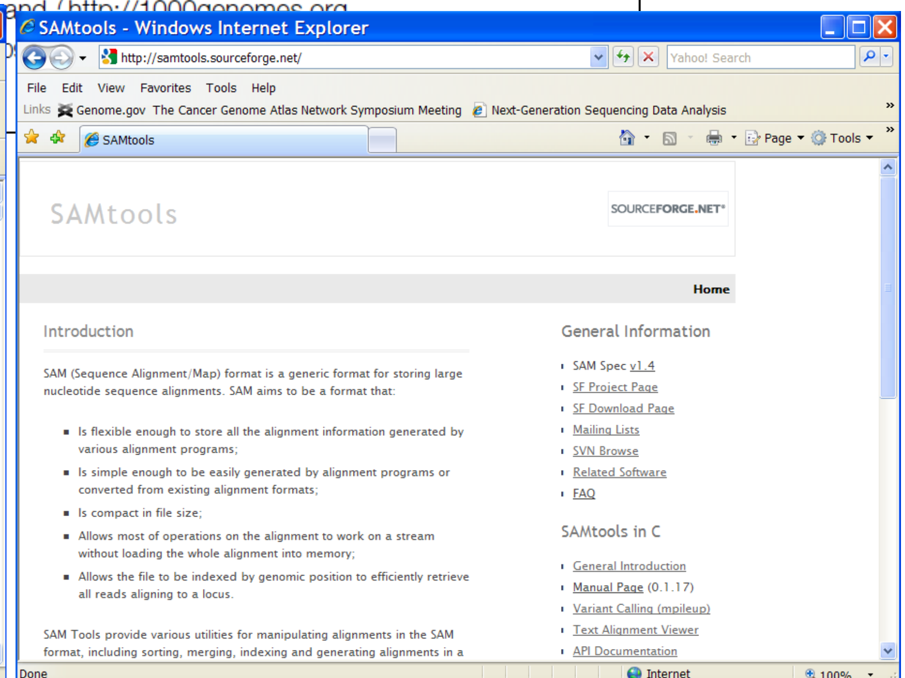
Sequence analysis

The Sequence Alignment/Map format and SAMtools

Heng Li[1,†], Bob Handsaker[2,†], Alec Wysoker[2], Tim Fennell[2], Jue Ruan[3], Nils Homer[4], Gabor Marth[5], Goncalo Abecasis[6], Richard Durbin[1,*] and 1000 Genome Project Data Processing Subgroup[7]

[1]Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Cambridge, CB10 1SA, UK, [2]Broad Institute of MIT and Harvard, Cambridge, MA 02141, USA, [3]Beijing Institute of Genomics, Chinese Academy of Science, Beijing 100029, China, [4]Department of Computer Science, University of California Los Angeles, Los Angeles, CA 90095, [5]Department of Biology, Boston College, Chestnut Hill, MA 02467, [6]Center for Statistical Genetics, Department of





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 -bcvg - > 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
```

Or

bcftools/varFilter filtered by –d 10:

```
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
```

Or

No Filtering:

```
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
```

Or

No Filtering:

```
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

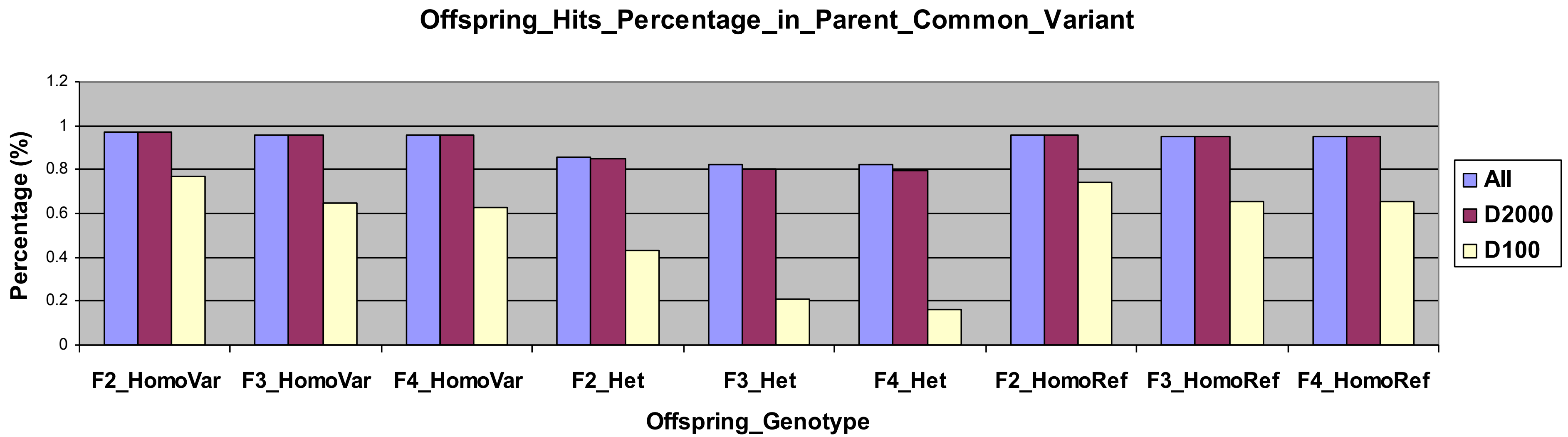
Select out only SNPs:

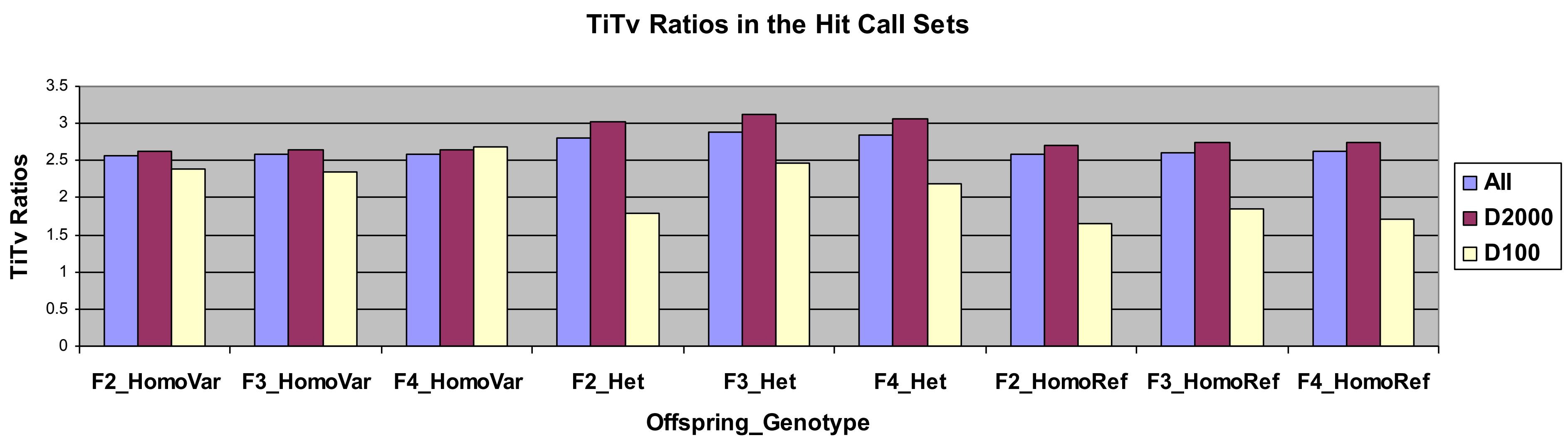
```
java -jar  /Path/GenomeAnalysisTK-1.6-7-g2be5704/bin/GenomeAnalysisTK.jar
-T SelectVariants -R /Path/hg19_chrM_1st.fa
-L /Path/Exome_Target_Region_SureSelect_AllExon_50mb_Bedfiles/029720_D_BED_20101013.bed
--variant /Path/samtools_mpileup_AllSamples_snps.raw.fild10.vcf
-selectType SNP
-o /Path/samtools_mpileup_AllSamples_snps.raw.fild10_SelSNP.vcf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

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!



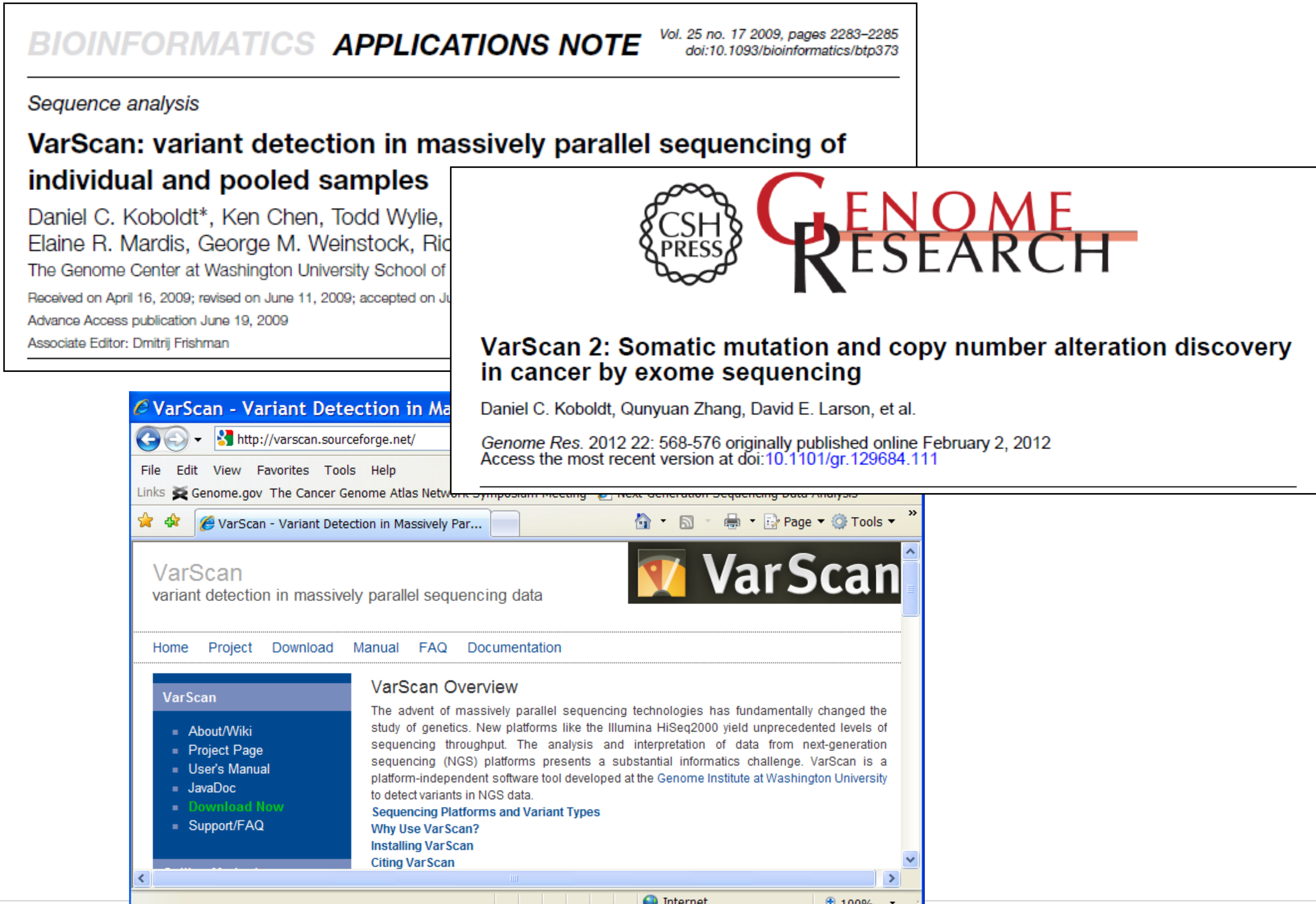


All: All variant call; D2000: Filtered by –D 2000; D100: Filtered by –D 100

Tool By Tool Highlighting Major Aspects of Practical Usage

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

VarScan: A Variant Discovery Tool from WashU



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

```
torkv:~> 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]
```

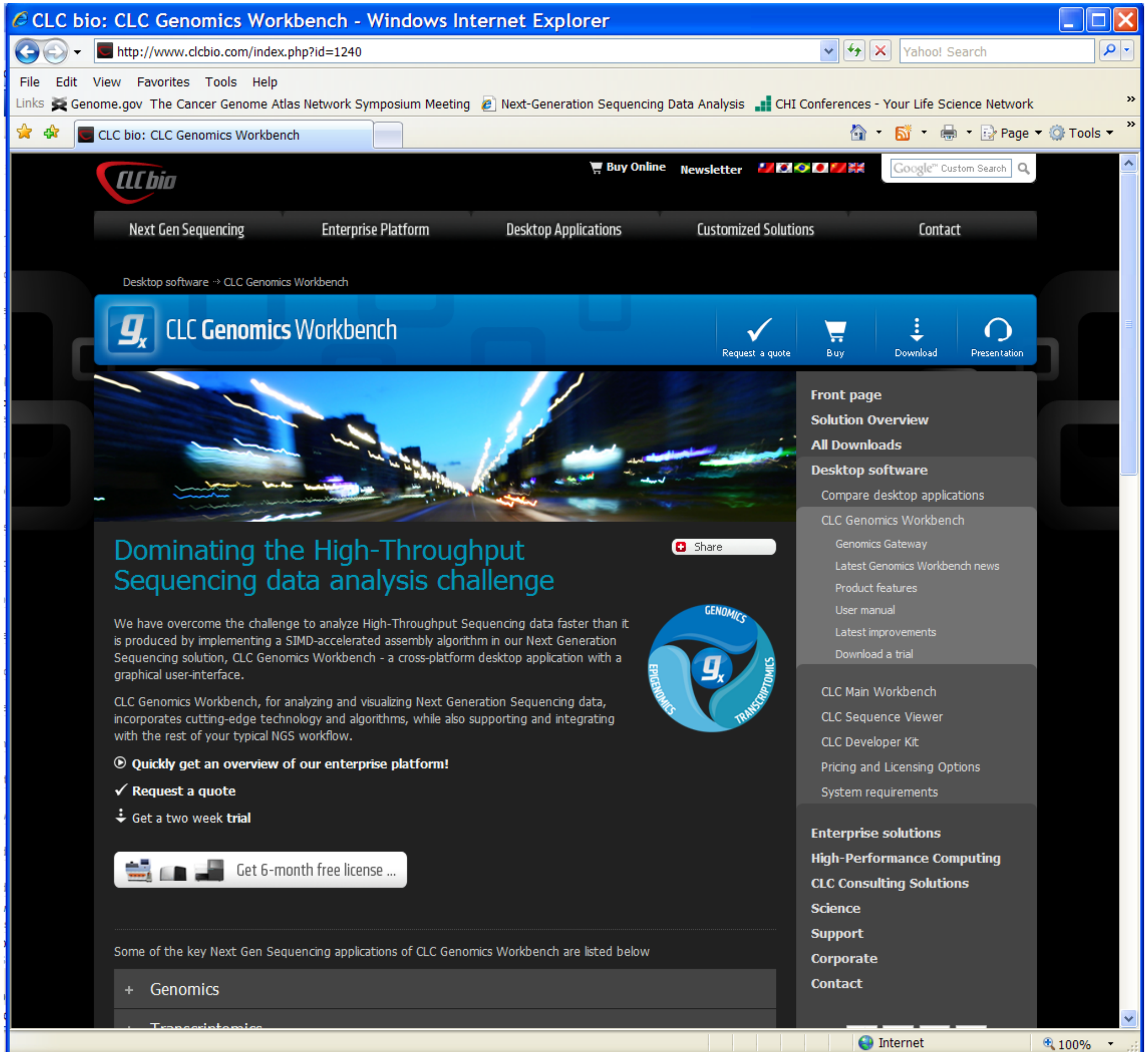
Tool By Tool Highlighting Major Aspects of Practical Usage

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

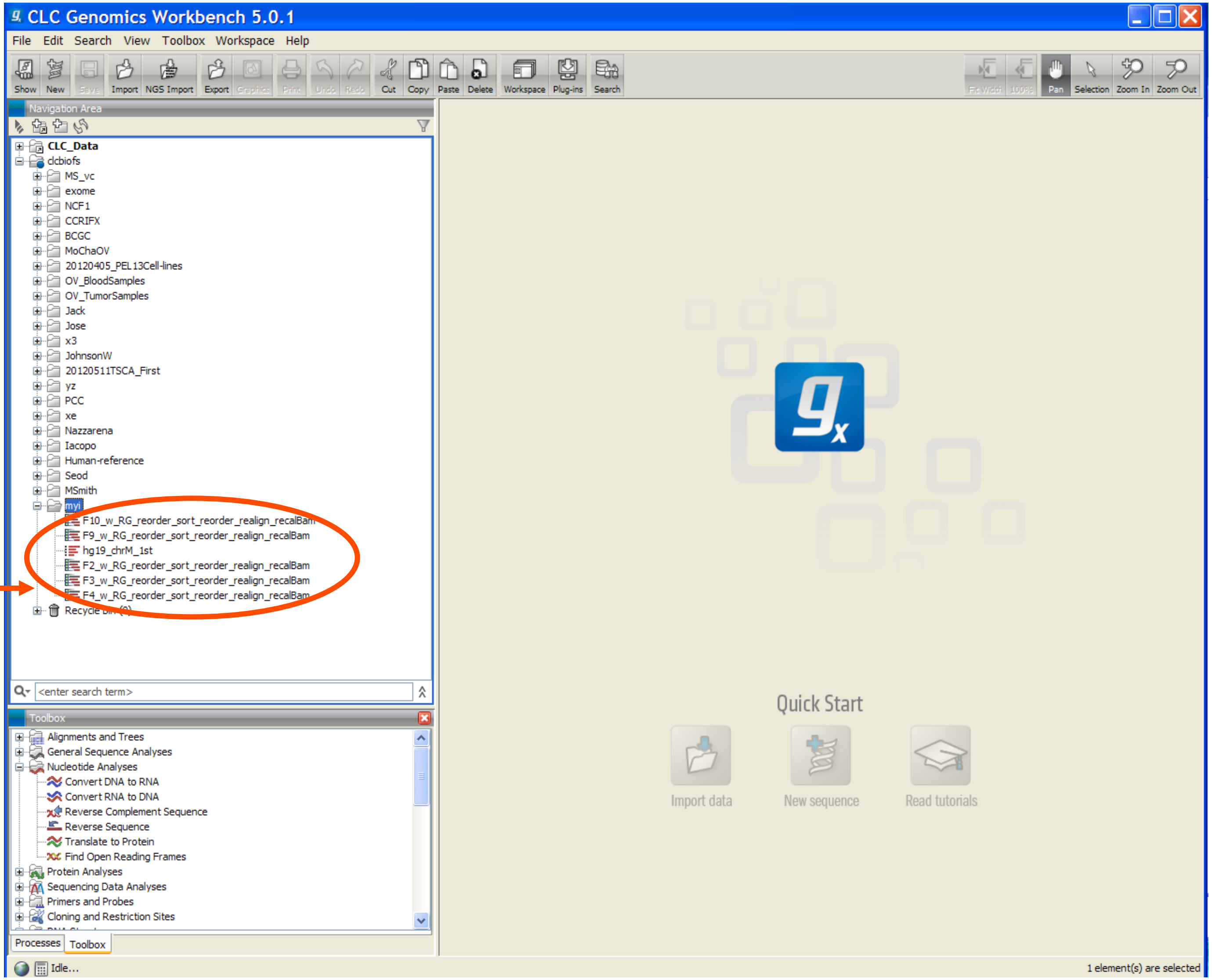
NGS-based SNP Discovery Tools

- 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

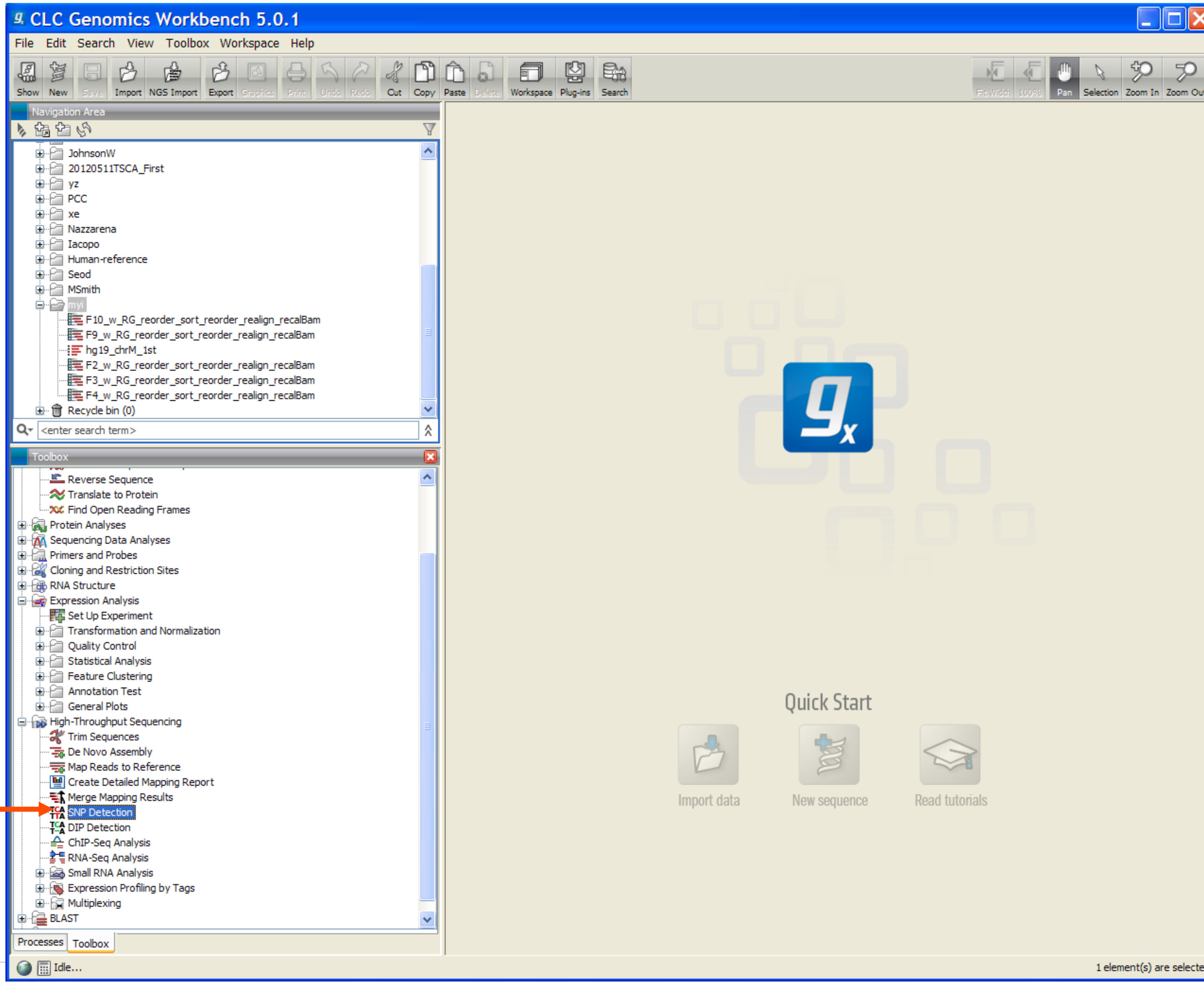


CLCbio Genomic Workbench Interface



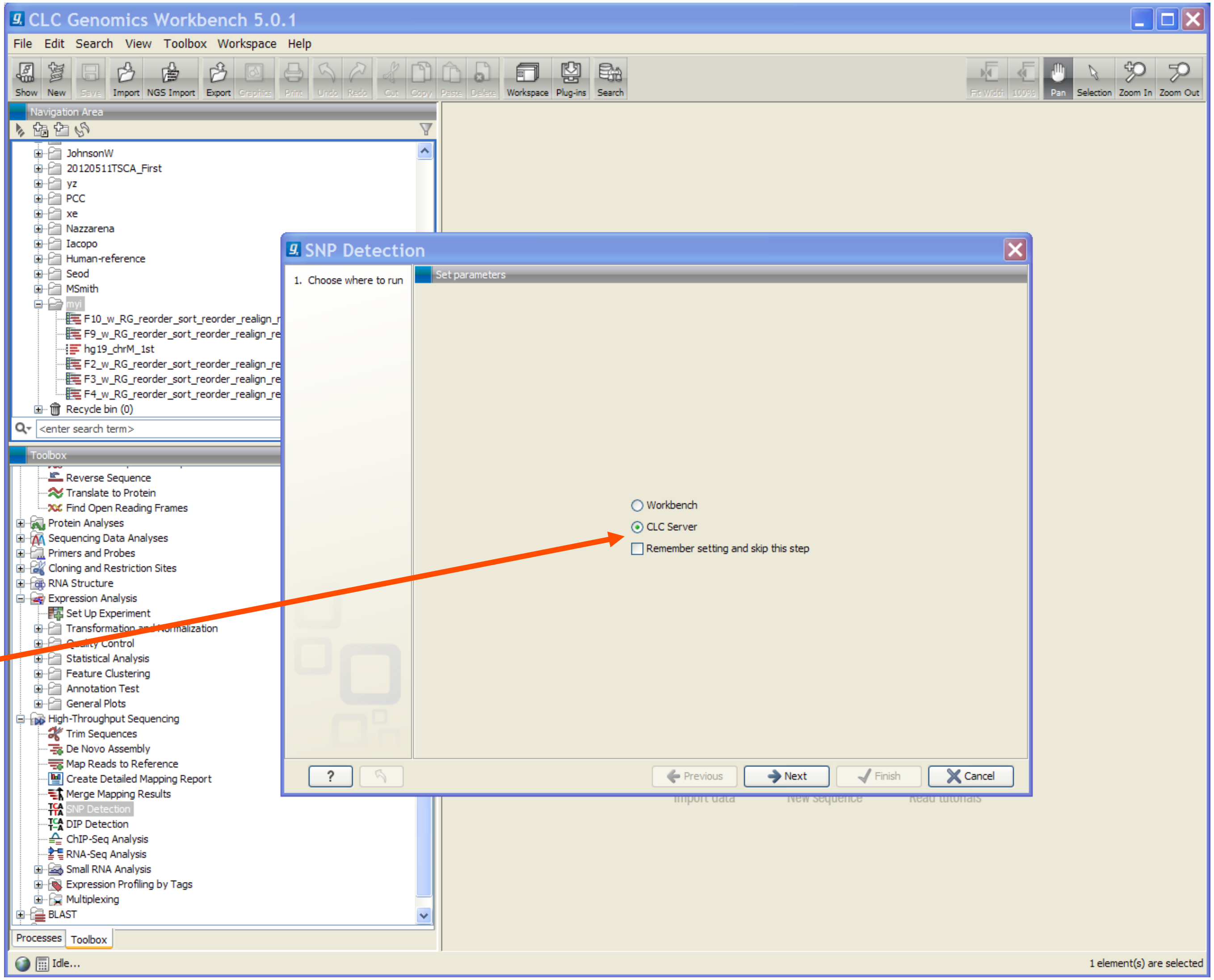
Need first to
import
Bam files
And reference
Into server

CLCbio SNP calling procedure



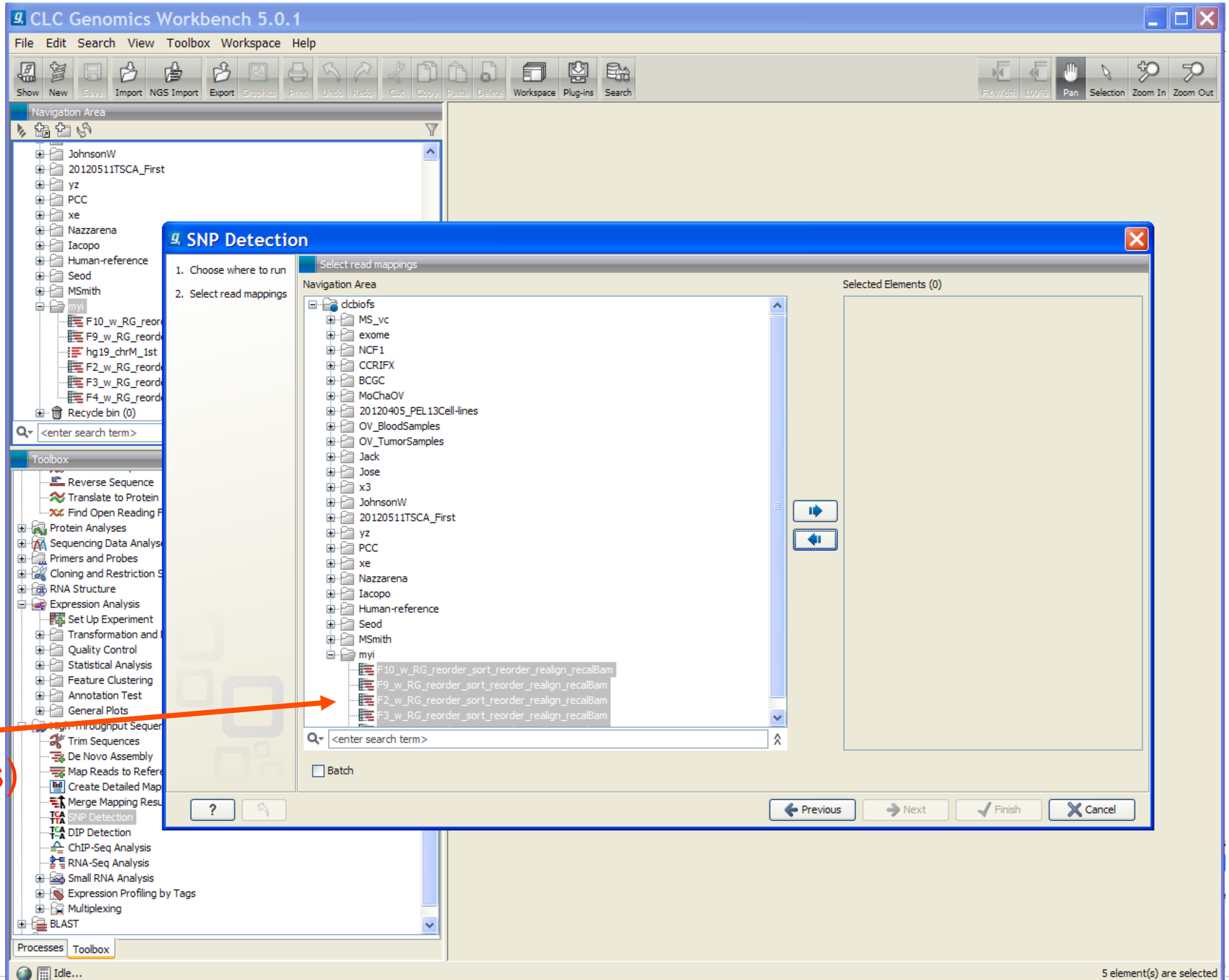
SNP detection module

CLCbio SNP calling procedure



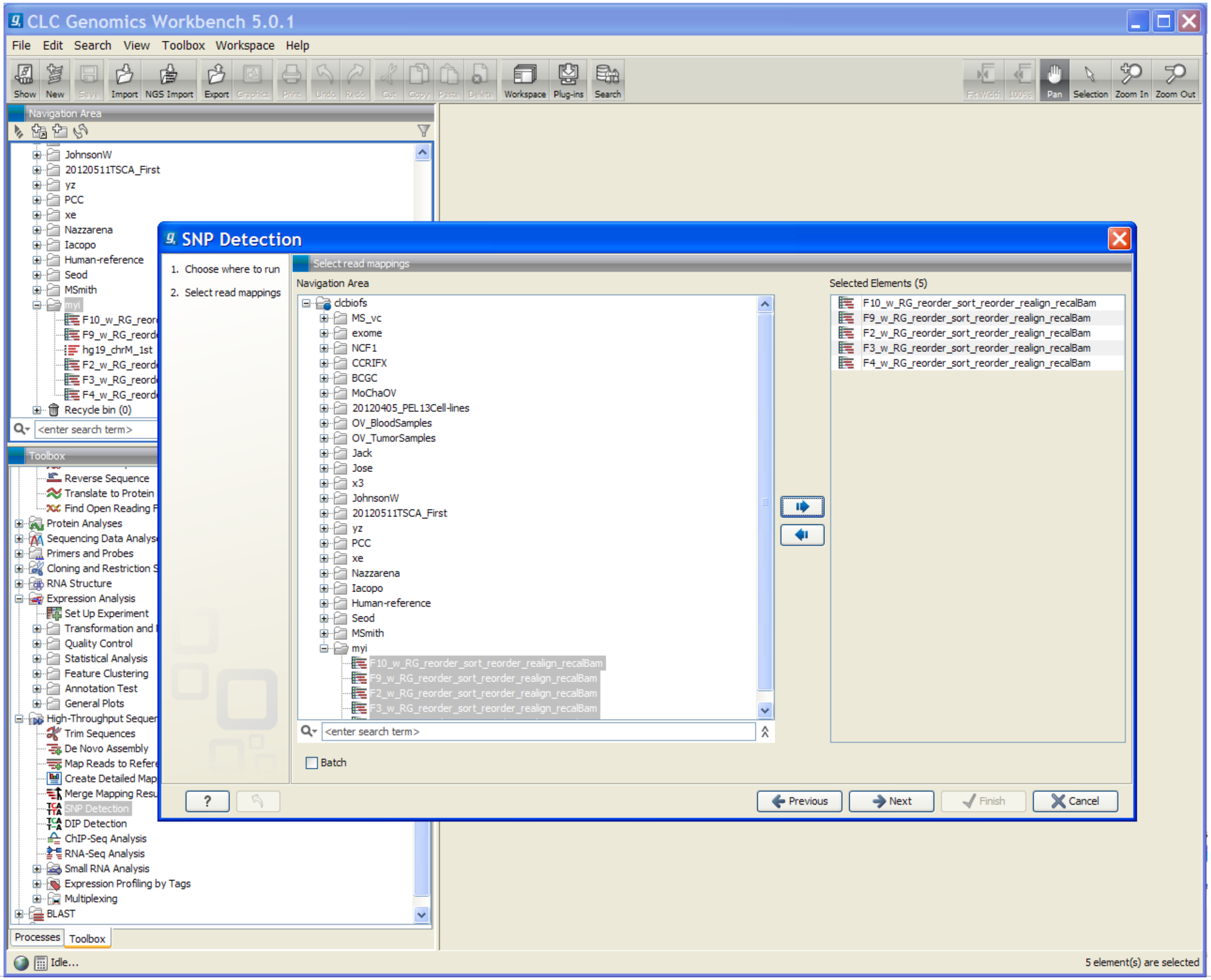
Use Server with more Computing power

CLCbio SNP calling procedure

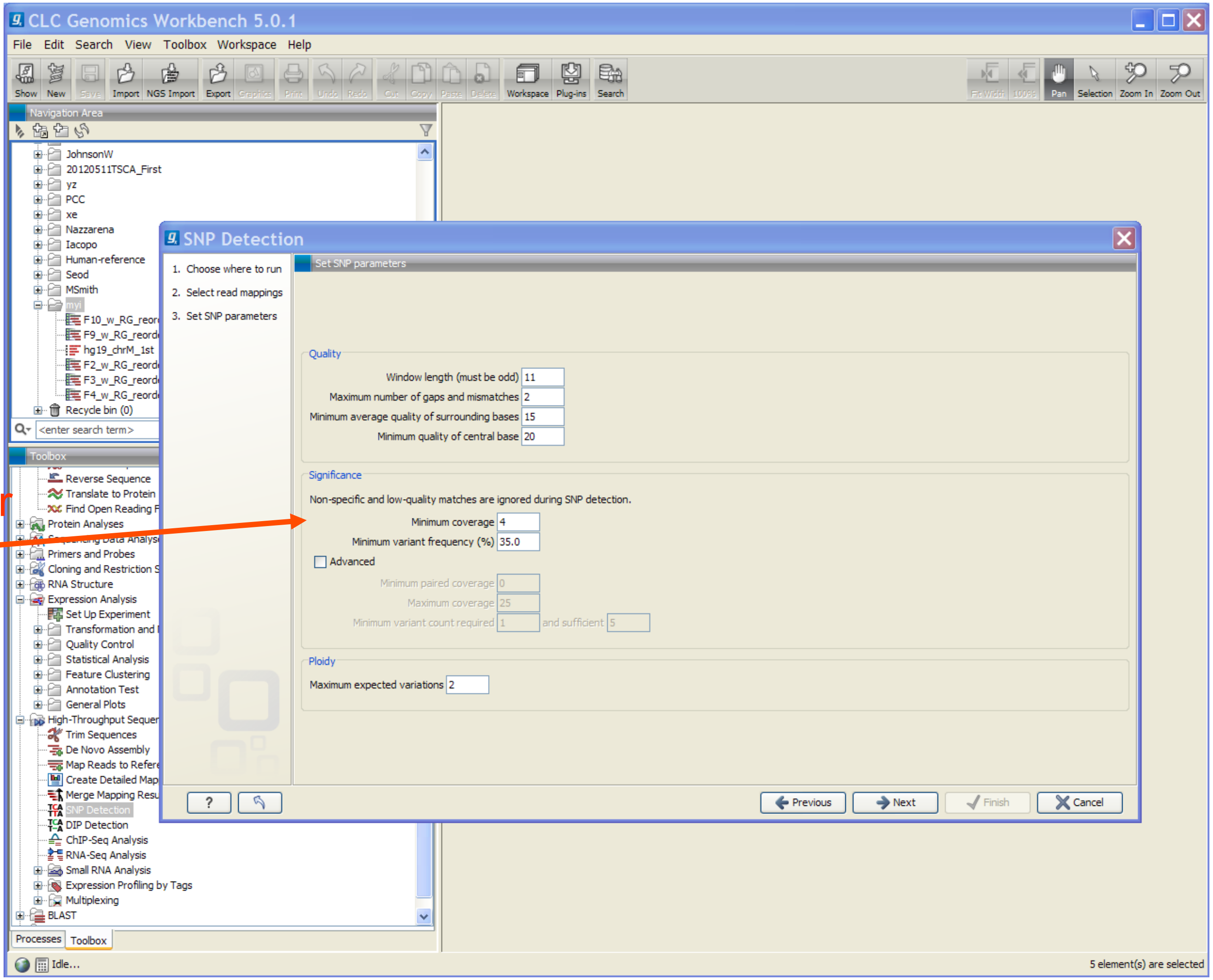


Select Samples' Bam file(s) For SNP calling

CLCbio SNP calling procedure

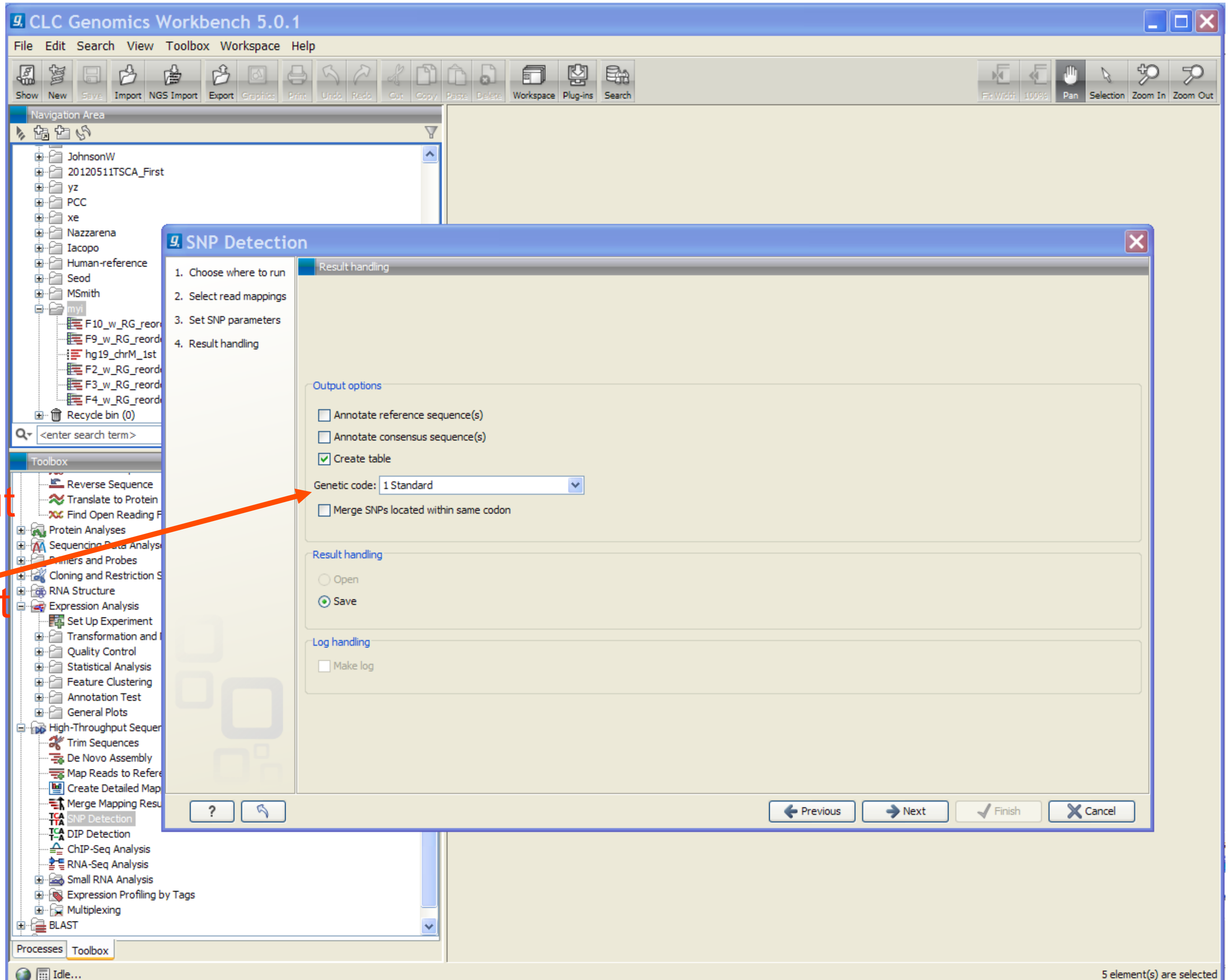


CLCbio SNP calling procedure



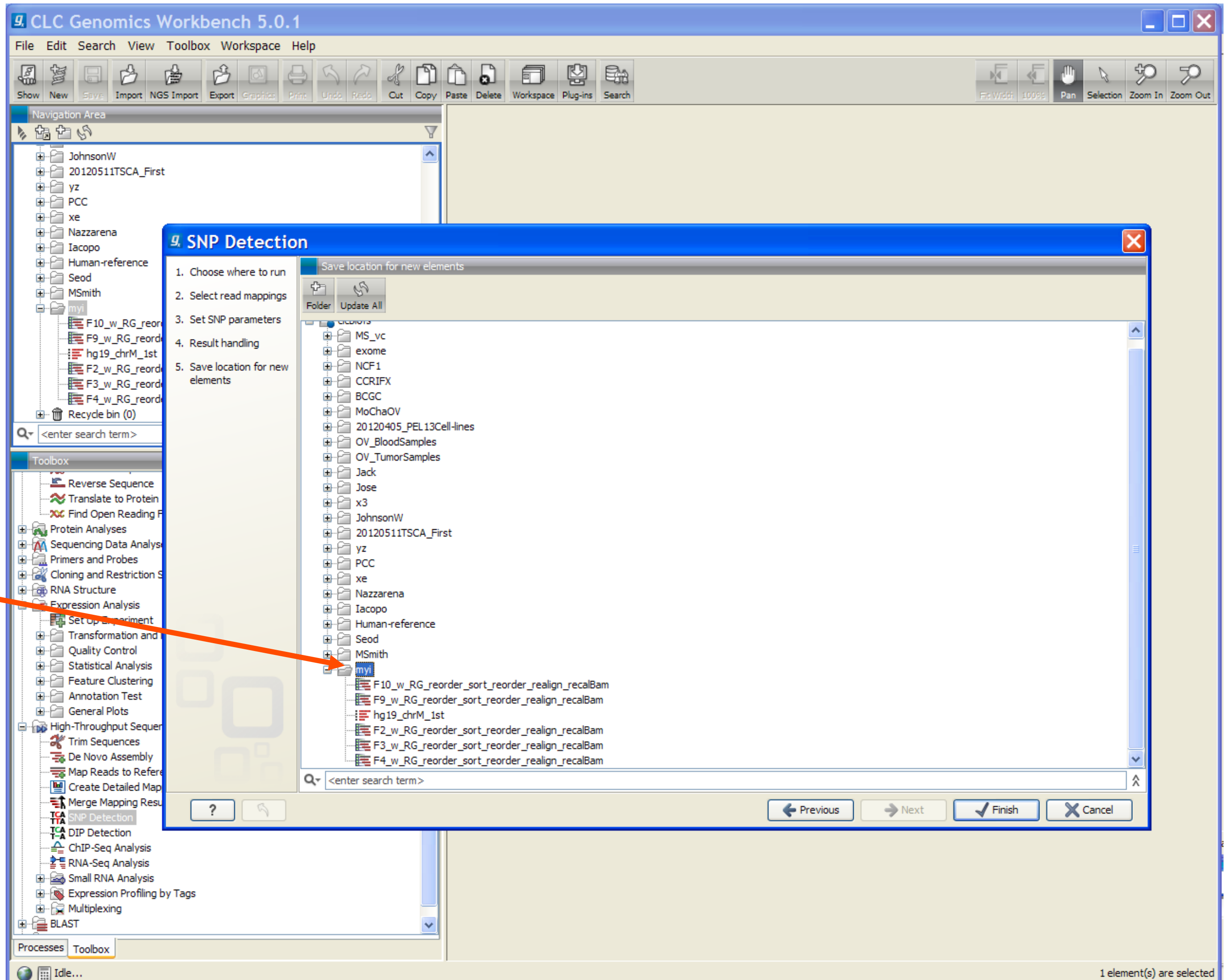
Parameter For SNP detection

CLCbio SNP calling procedure



Still output
Not in
Vcf format

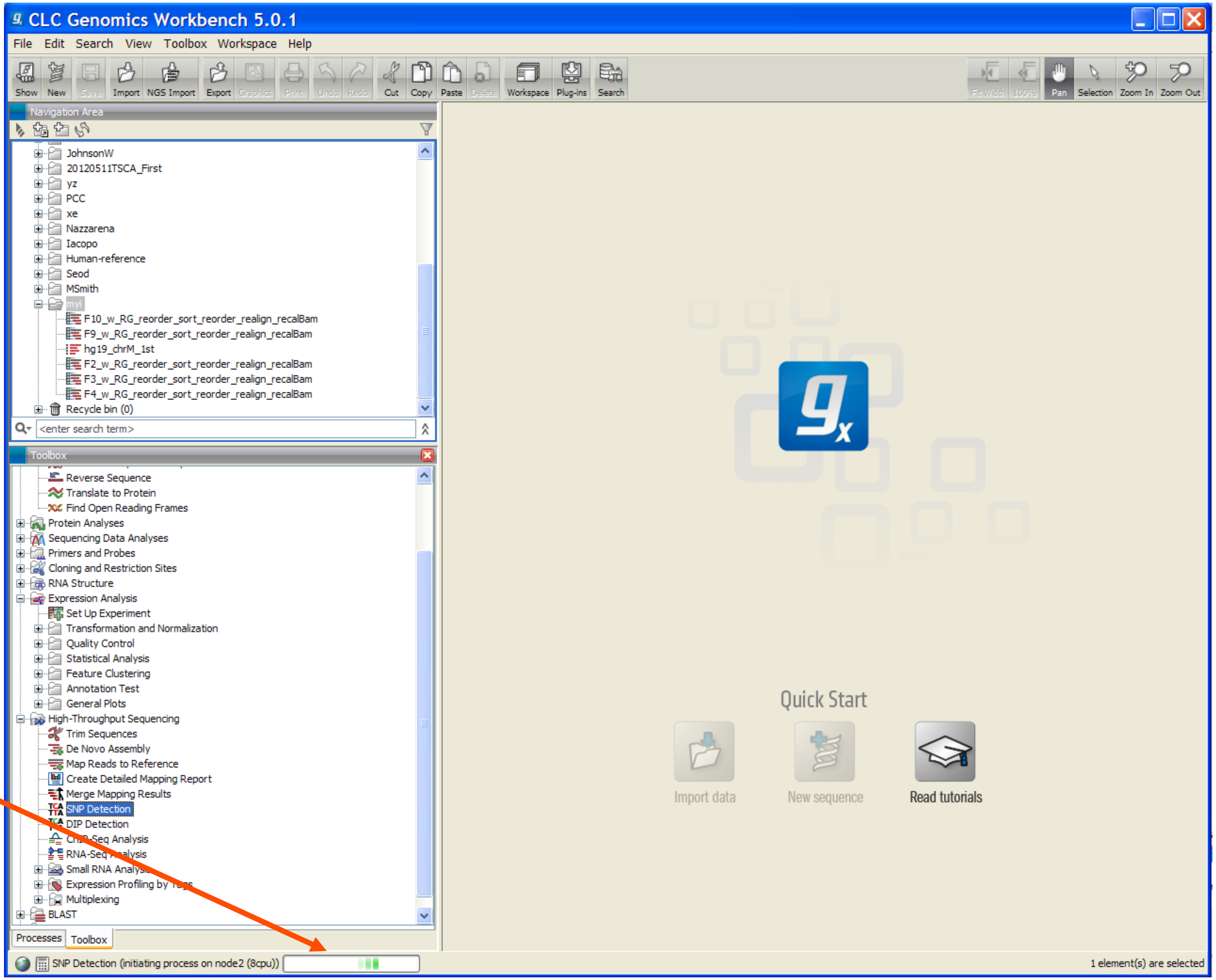
CLCbio SNP calling procedure



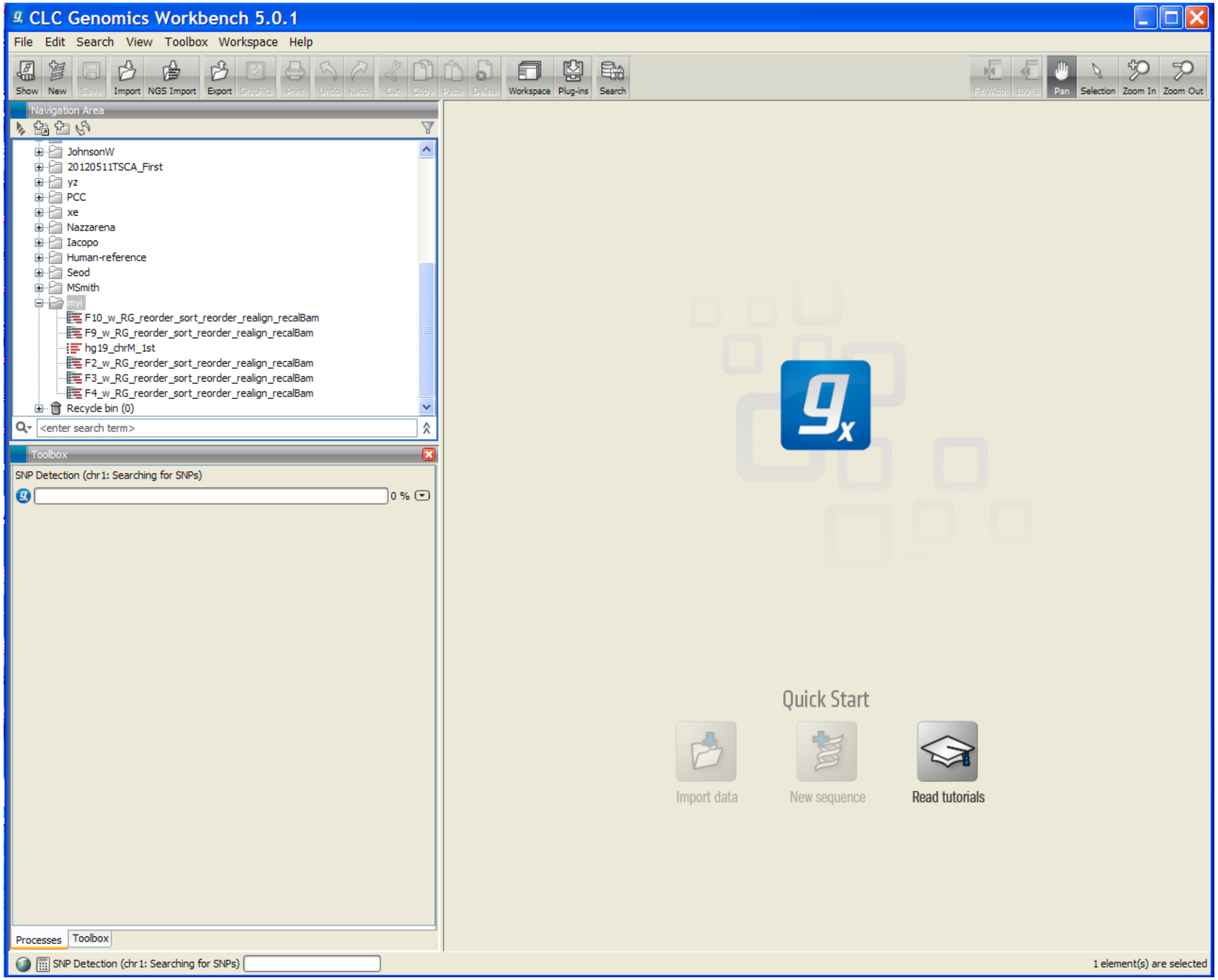
Choose location to save result

CLCbio SNP calling procedure

Progress

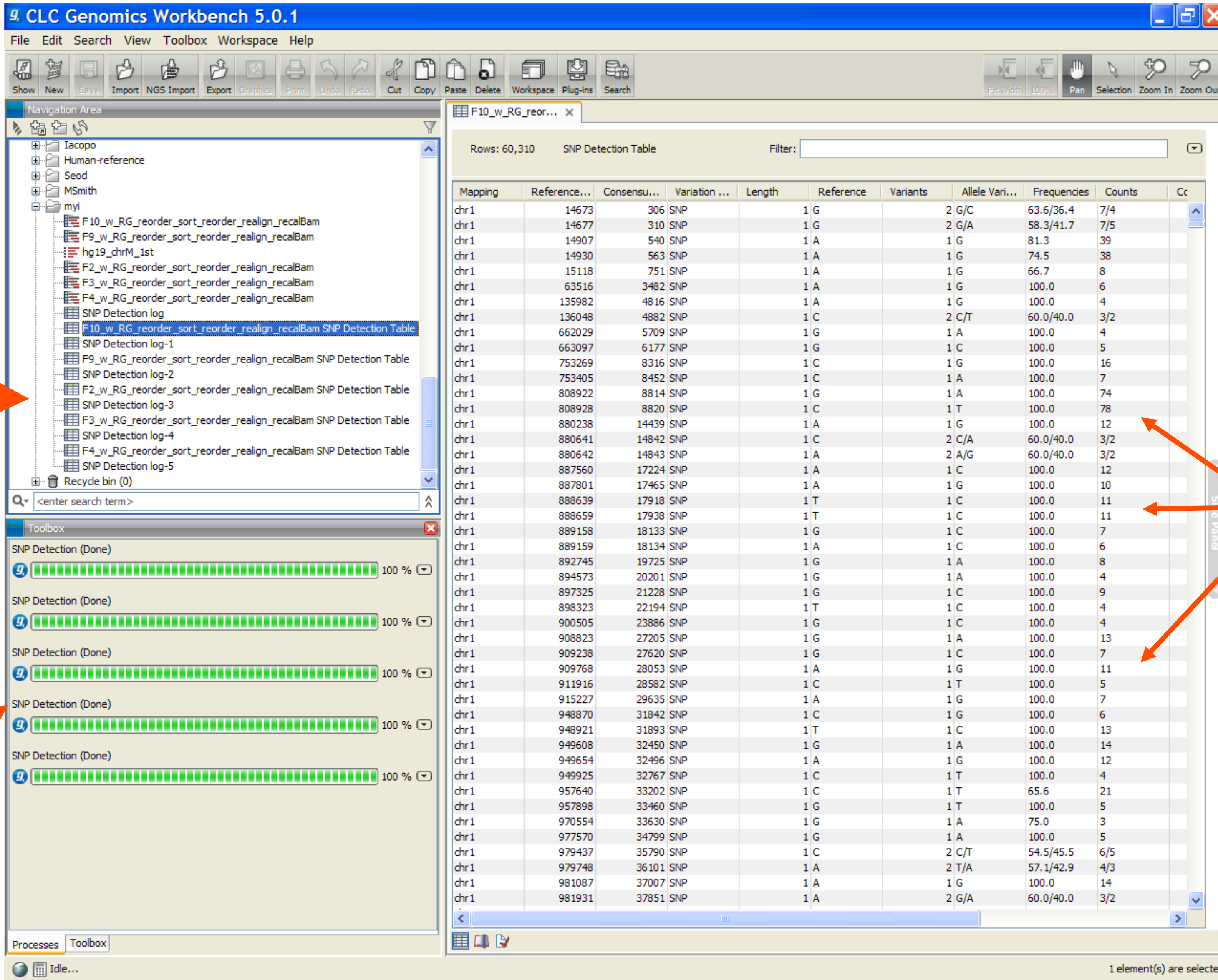


CLCbio SNP calling procedure



CLCbio SNP calling procedure

SNP result files & log file



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	
chr1	63516	3482	SNP	1	A	1	G	100.0	6	
chr1	135982	4816	SNP	1	A	1	G	100.0	4	
chr1	136048	4882	SNP	1	C	2	C/T	60.0/40.0	3/2	
chr1	662029	5709	SNP	1	G	1	A	100.0	4	
chr1	663097	6177	SNP	1	G	1	C	100.0	5	
chr1	753269	8316	SNP	1	C	1	G	100.0	16	
chr1	753405	8452	SNP	1	C	1	A	100.0	7	
chr1	808922	8814	SNP	1	G	1	A	100.0	74	
chr1	808928	8820	SNP	1	C	1	T	100.0	78	
chr1	880238	14439	SNP	1	A	1	G	100.0	12	
chr1	880641	14842	SNP	1	C	2	C/A	60.0/40.0	3/2	
chr1	880642	14843	SNP	1	A	2	A/G	60.0/40.0	3/2	
chr1	887560	17224	SNP	1	A	1	C	100.0	12	
chr1	887801	17465	SNP	1	A	1	G	100.0	10	
chr1	888639	17918	SNP	1	T	1	C	100.0	11	
chr1	888659	17938	SNP	1	T	1	C	100.0	11	
chr1	889158	18133	SNP	1	G	1	C	100.0	7	
chr1	889159	18134	SNP	1	A	1	C	100.0	6	
chr1	892745	19725	SNP	1	G	1	A	100.0	8	
chr1	894573	20201	SNP	1	G	1	A	100.0	4	
chr1	897325	21228	SNP	1	G	1	C	100.0	9	
chr1	898323	22194	SNP	1	T	1	C	100.0	4	
chr1	900505	23886	SNP	1	G	1	C	100.0	4	
chr1	908823	27205	SNP	1	G	1	A	100.0	13	
chr1	909238	27620	SNP	1	G	1	C	100.0	7	
chr1	909768	28053	SNP	1	A	1	G	100.0	11	
chr1	911916	28582	SNP	1	C	1	T	100.0	5	
chr1	915227	29635	SNP	1	A	1	G	100.0	7	
chr1	948870	31842	SNP	1	C	1	G	100.0	6	
chr1	948921	31893	SNP	1	T	1	C	100.0	13	
chr1	949608	32450	SNP	1	G	1	A	100.0	14	
chr1	949654	32496	SNP	1	A	1	G	100.0	12	
chr1	949925	32767	SNP	1	C	1	T	100.0	4	
chr1	957640	33202	SNP	1	C	1	T	65.6	21	
chr1	957898	33460	SNP	1	G	1	T	100.0	5	
chr1	970554	33630	SNP	1	G	1	A	75.0	3	
chr1	977570	34799	SNP	1	G	1	A	100.0	5	
chr1	979437	35790	SNP	1	C	2	C/T	54.5/45.5	6/5	
chr1	979748	36101	SNP	1	A	2	T/A	57.1/42.9	4/3	
chr1	981087	37007	SNP	1	A	1	G	100.0	14	
chr1	981931	37851	SNP	1	A	2	G/A	60.0/40.0	3/2	

SNP result File content

Progress

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 absolte 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 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

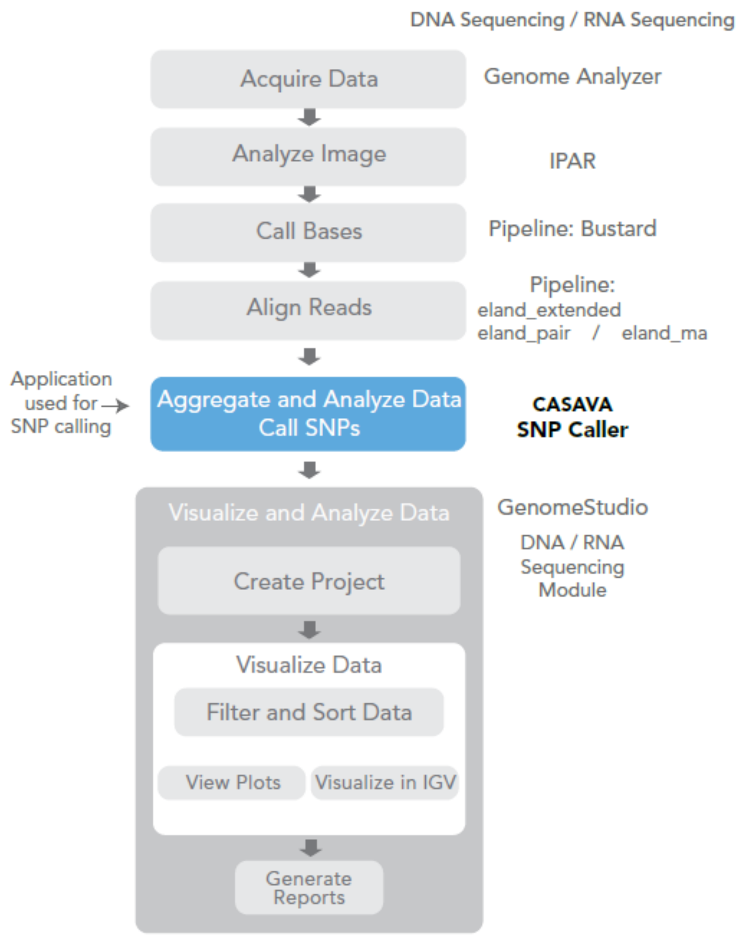
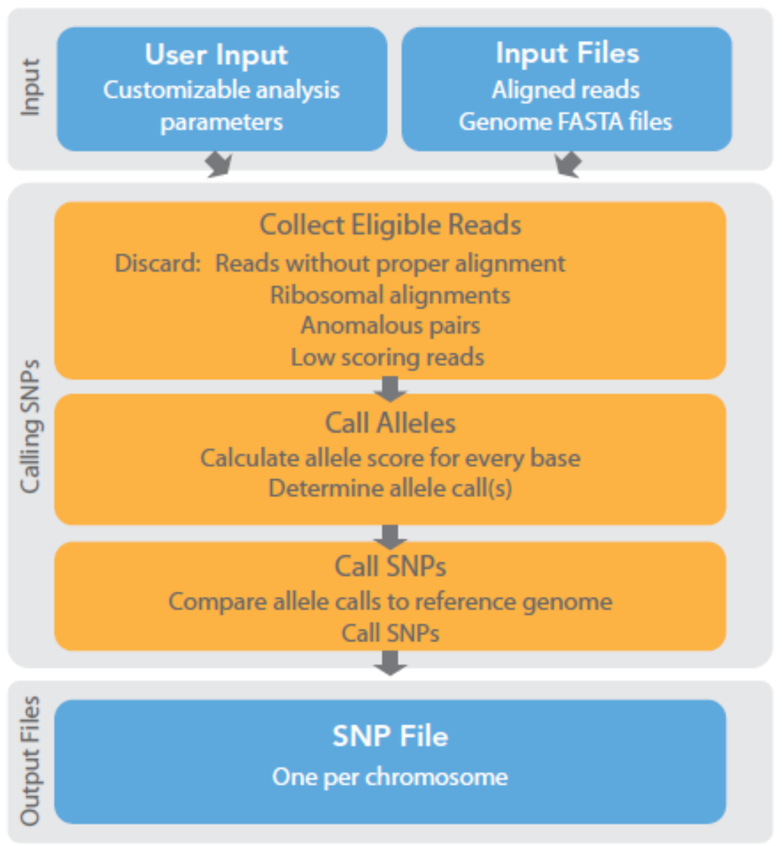


Figure 2: SNP Calling Workflow



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

```
torkv:/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 tha
 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 aligment score threshold to NUMBER (default 90)
        --QVCutoffSingle=NUMBER    - Sets the single-read aligment 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 snps.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:/is1/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=/is1/p
rojects/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

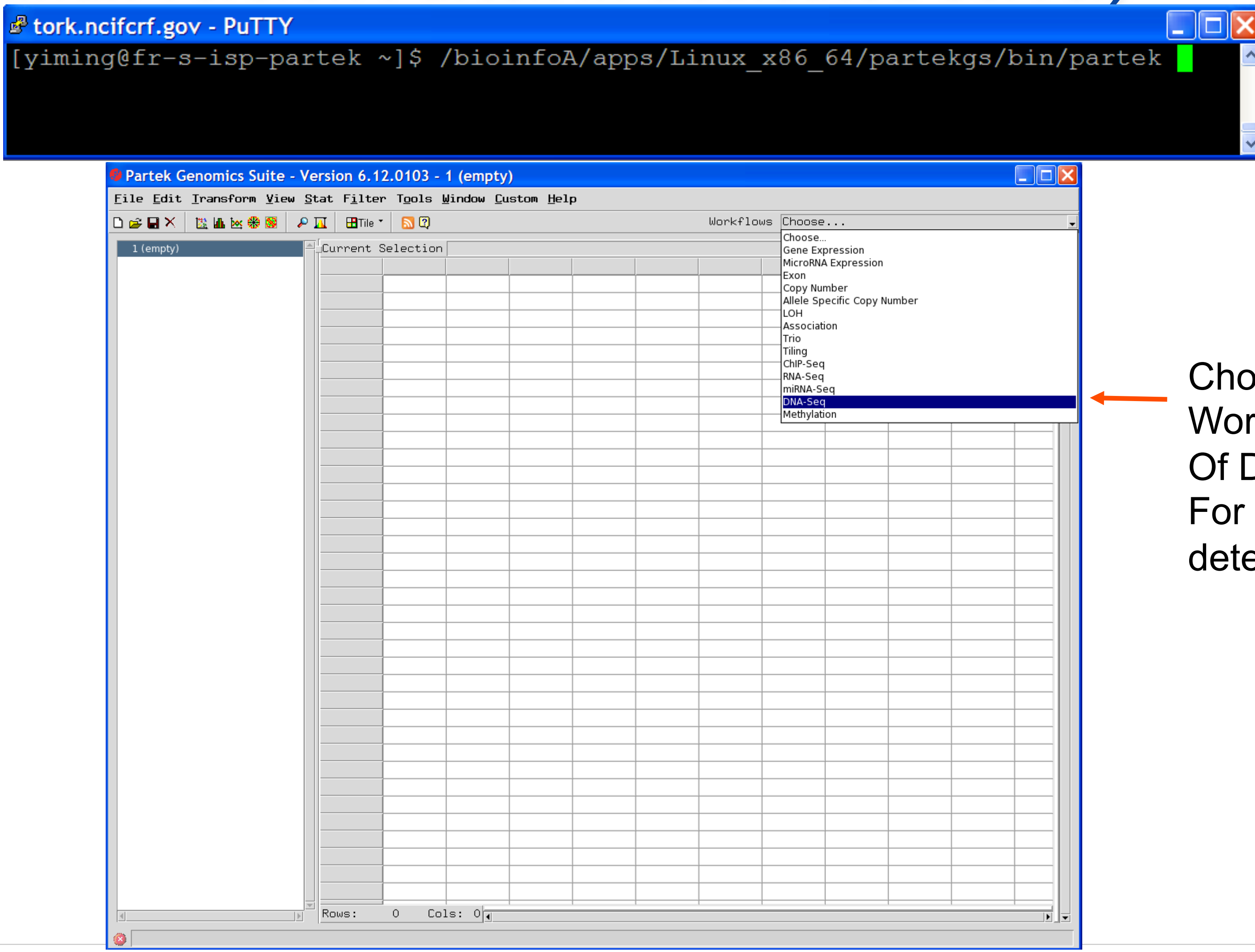
Tool By Tool Highlighting Major Aspects of Practical Usage

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

Partek solution for NGS data analysis

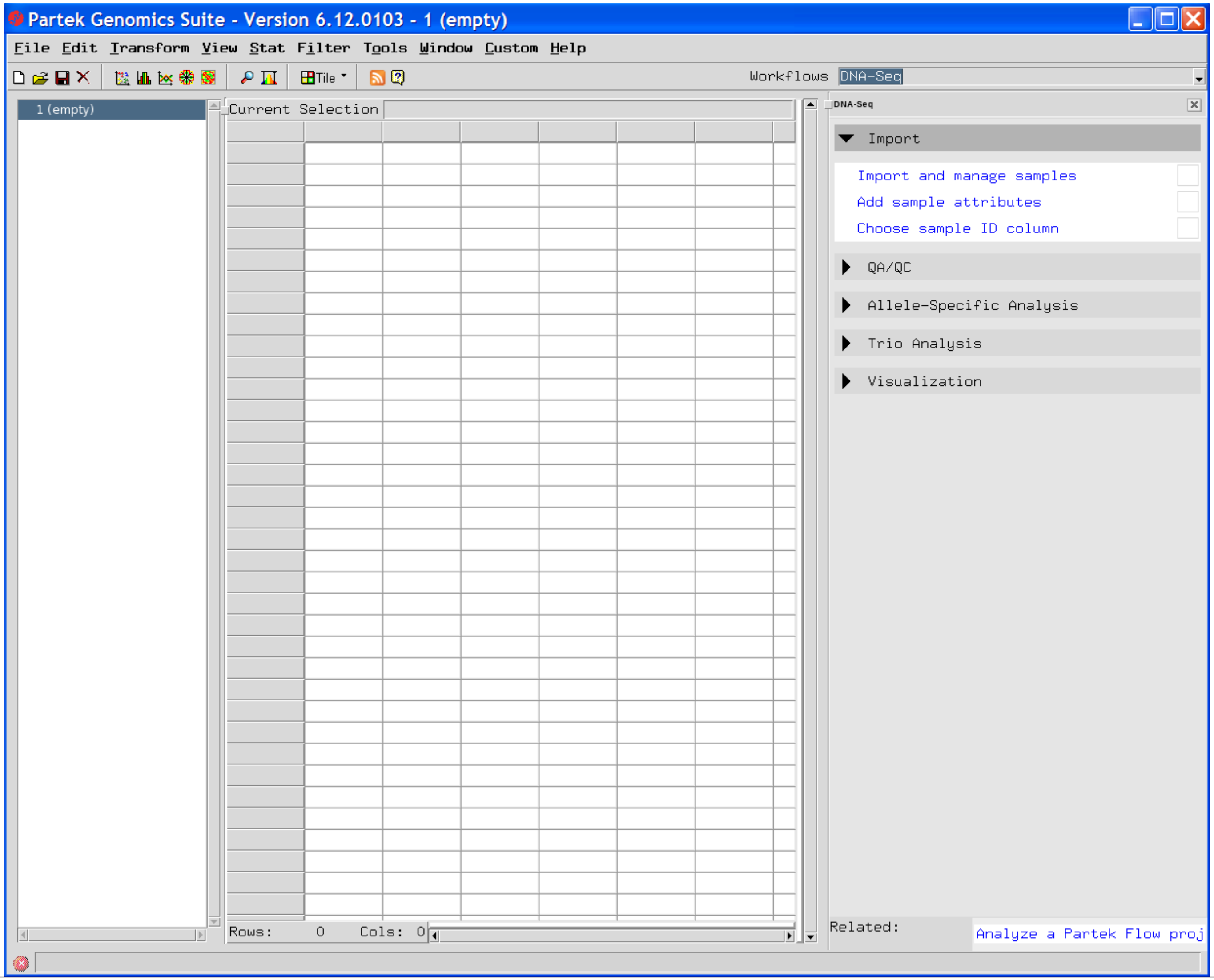
Partek Flow uses external SNP detection methods (e.g, samtools), Genomic suite has its own SNP detection method

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

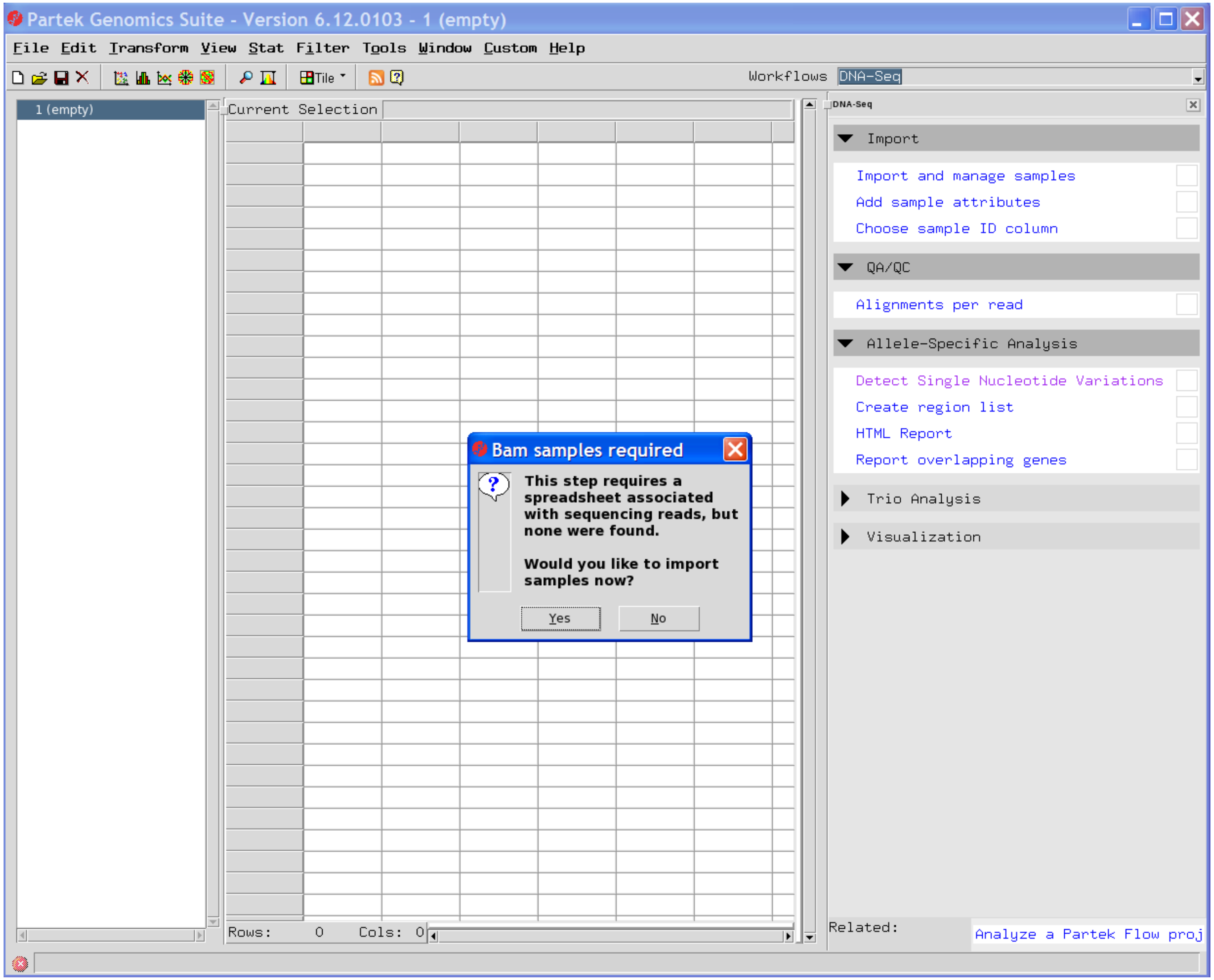


Choose Workflow Of DNA-seq For SNP detection

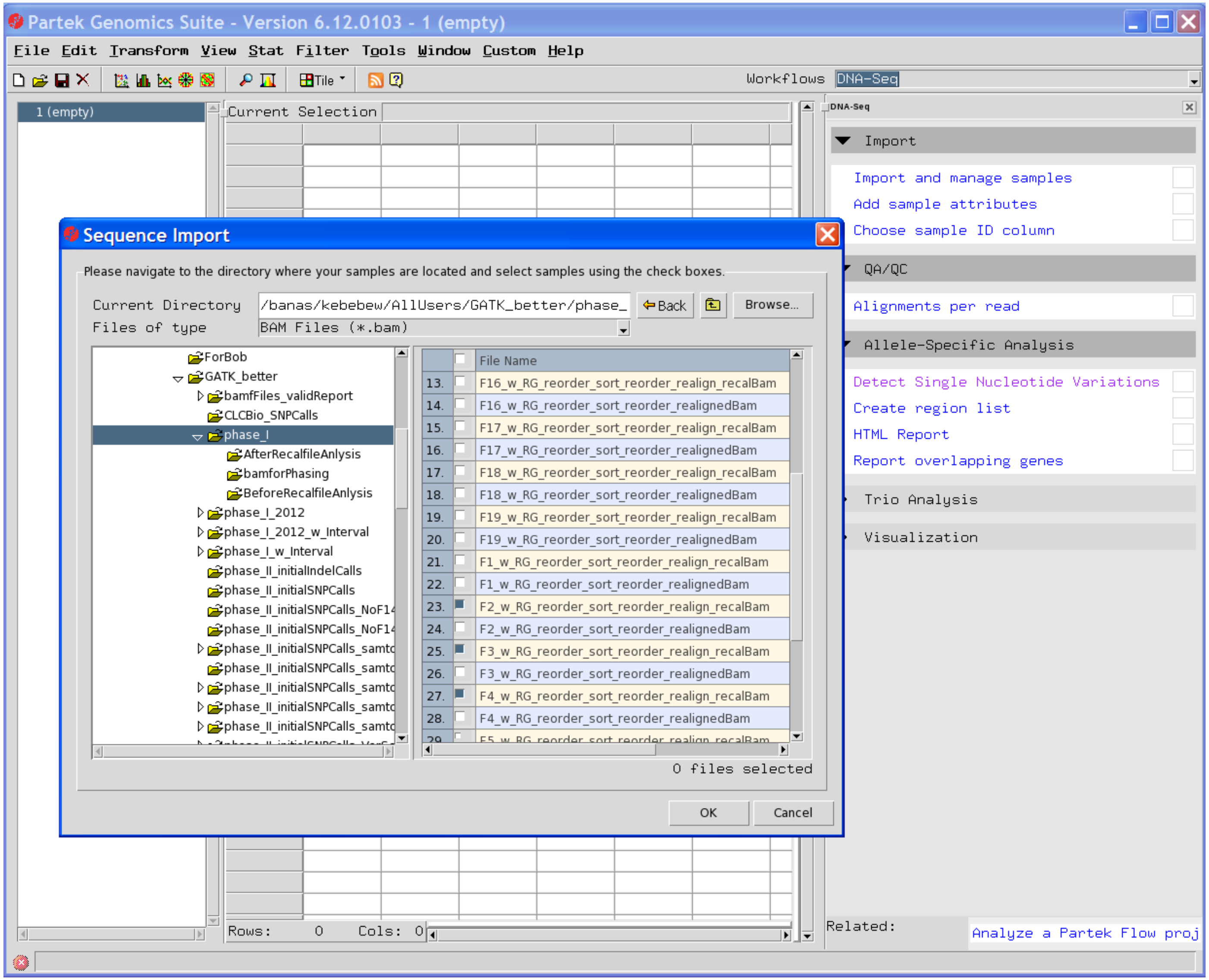
Partek SNP calling procedure



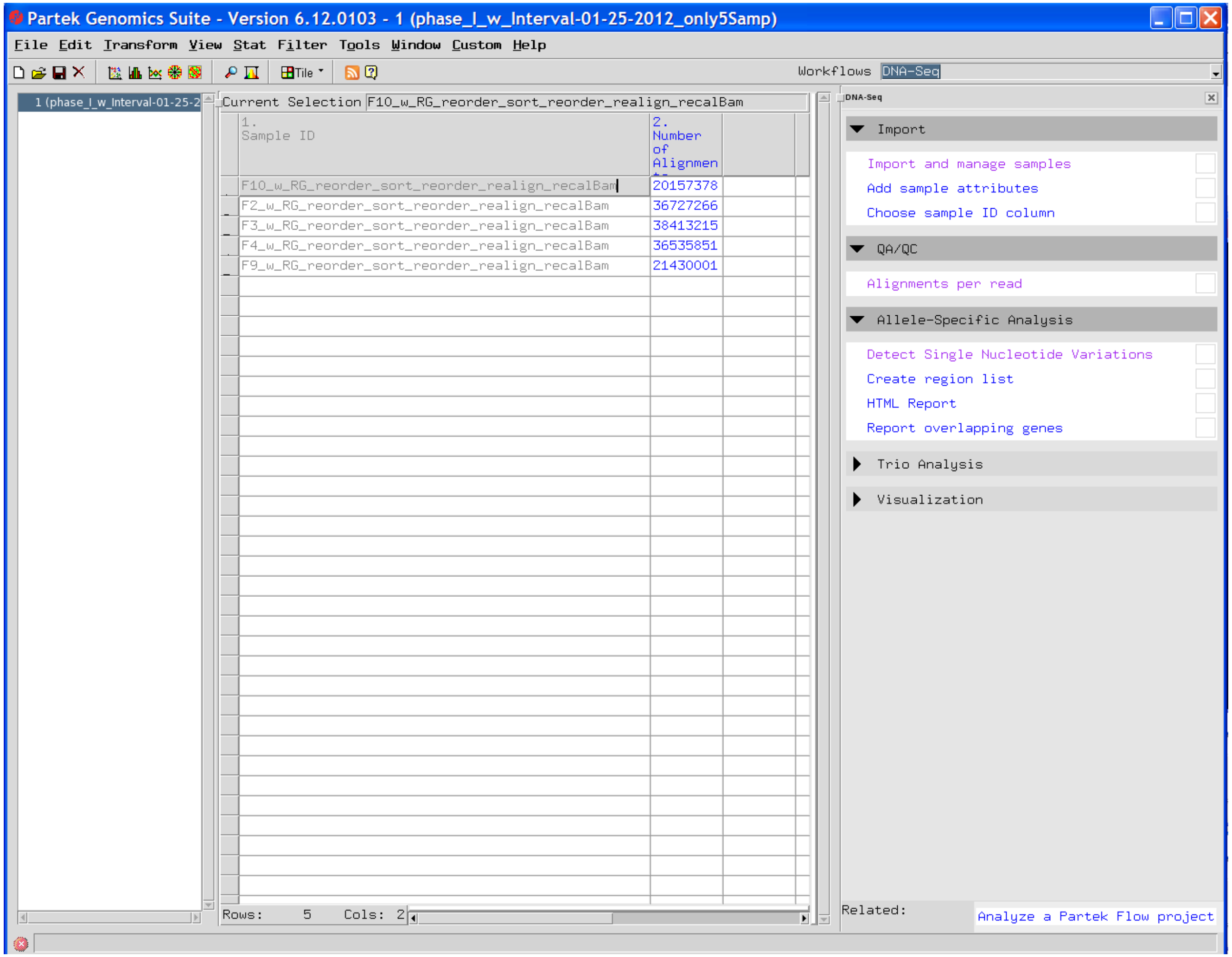
Partek SNP calling procedure



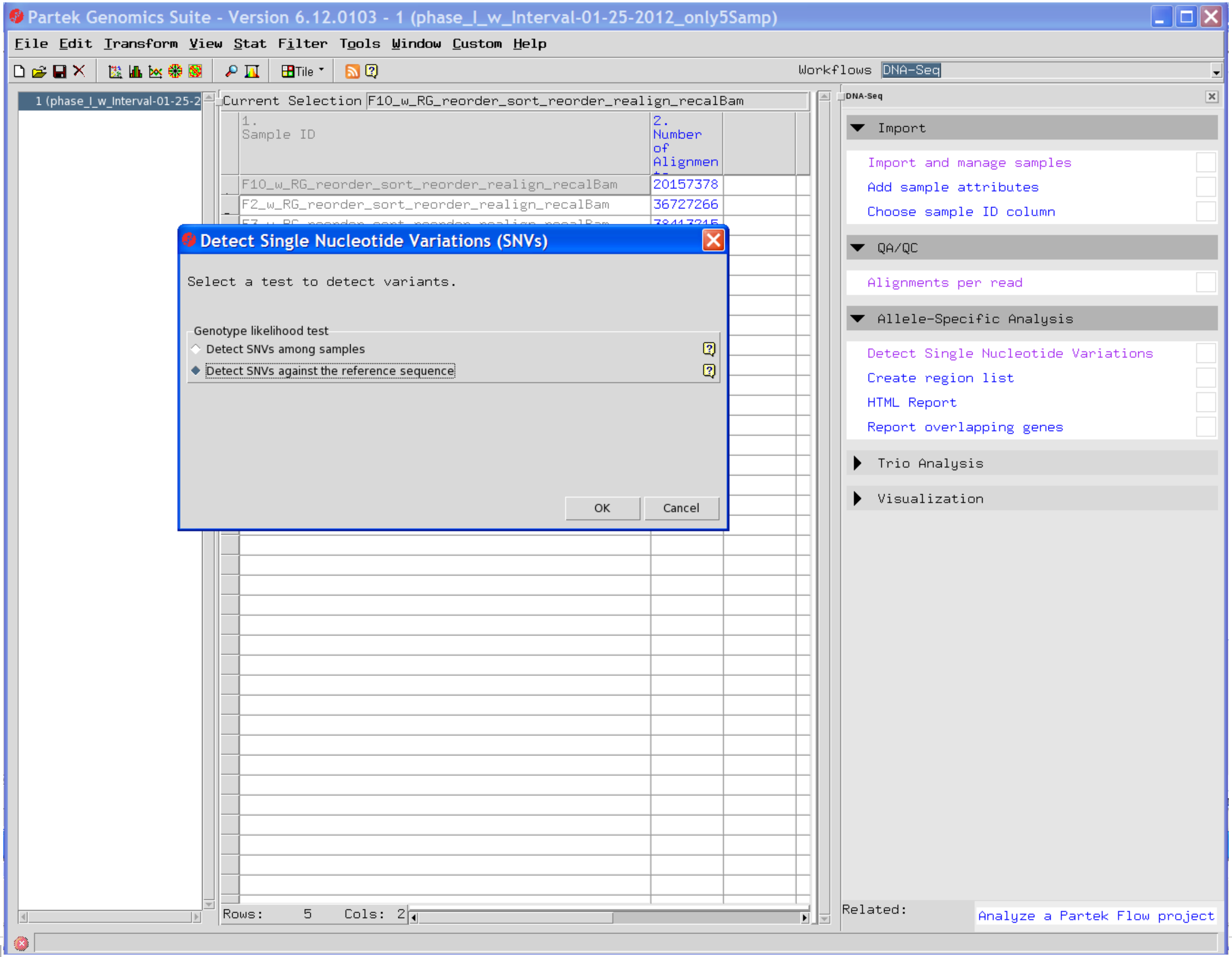
Partek SNP calling procedure



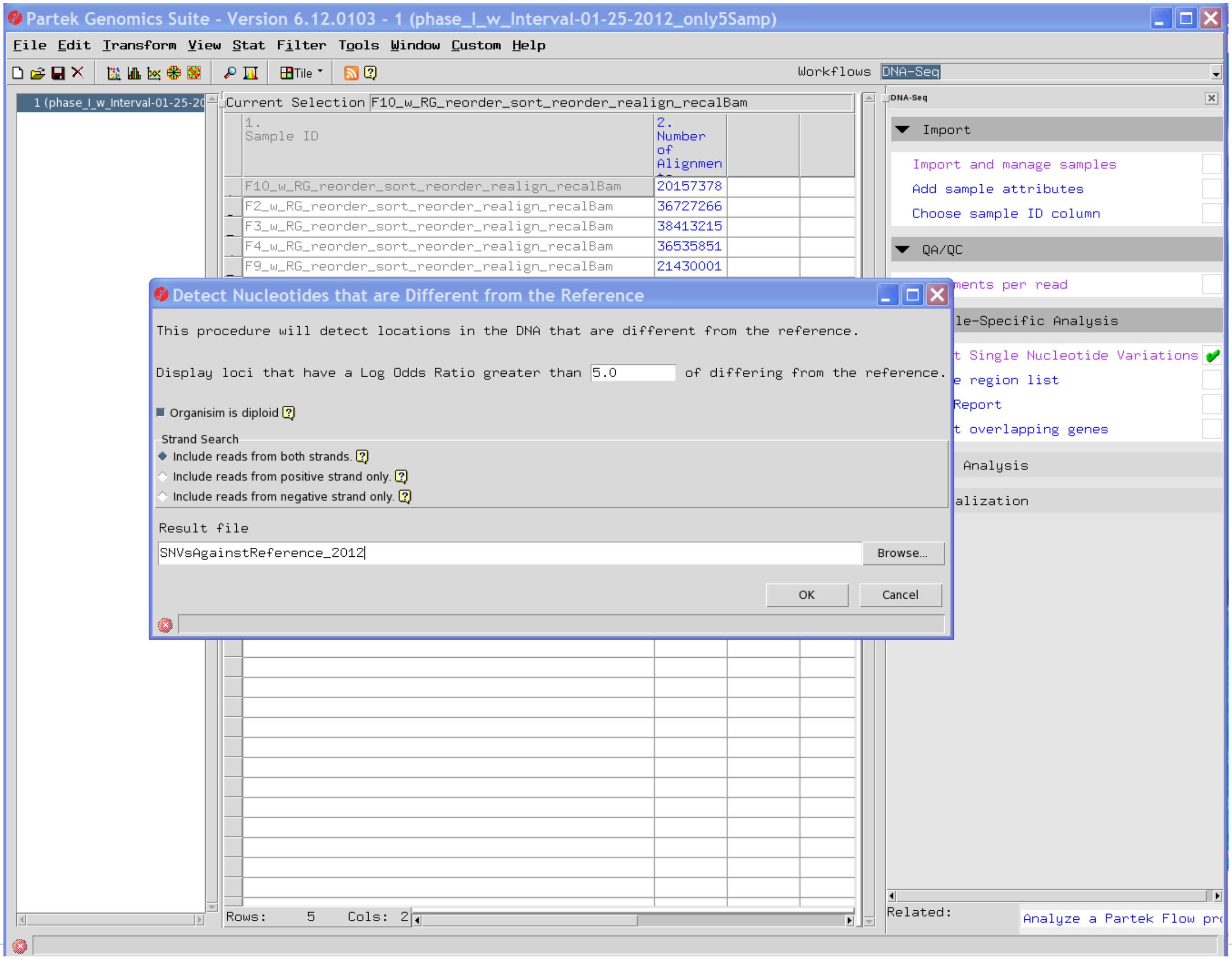
Partek SNP calling procedure



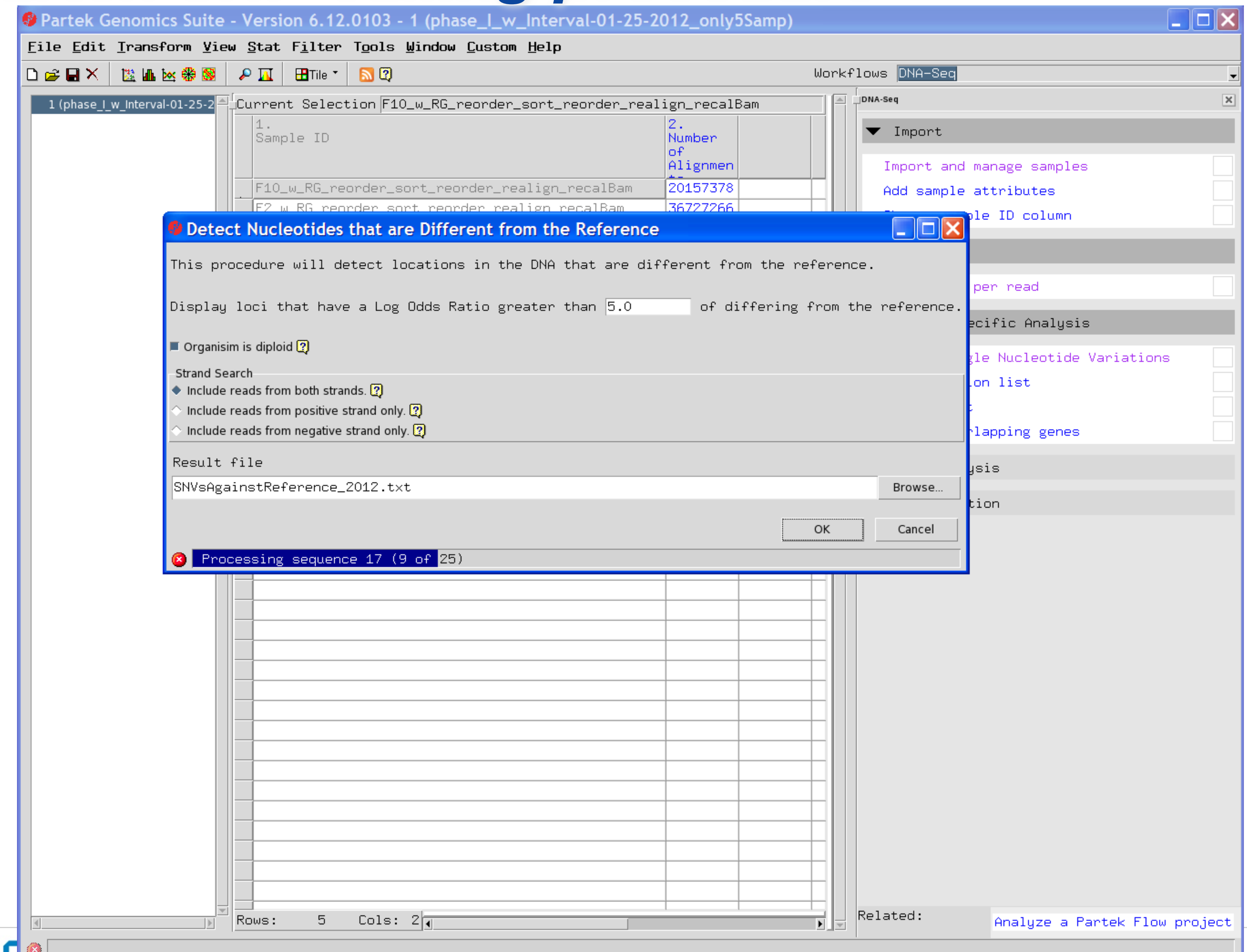
Partek SNP calling procedure



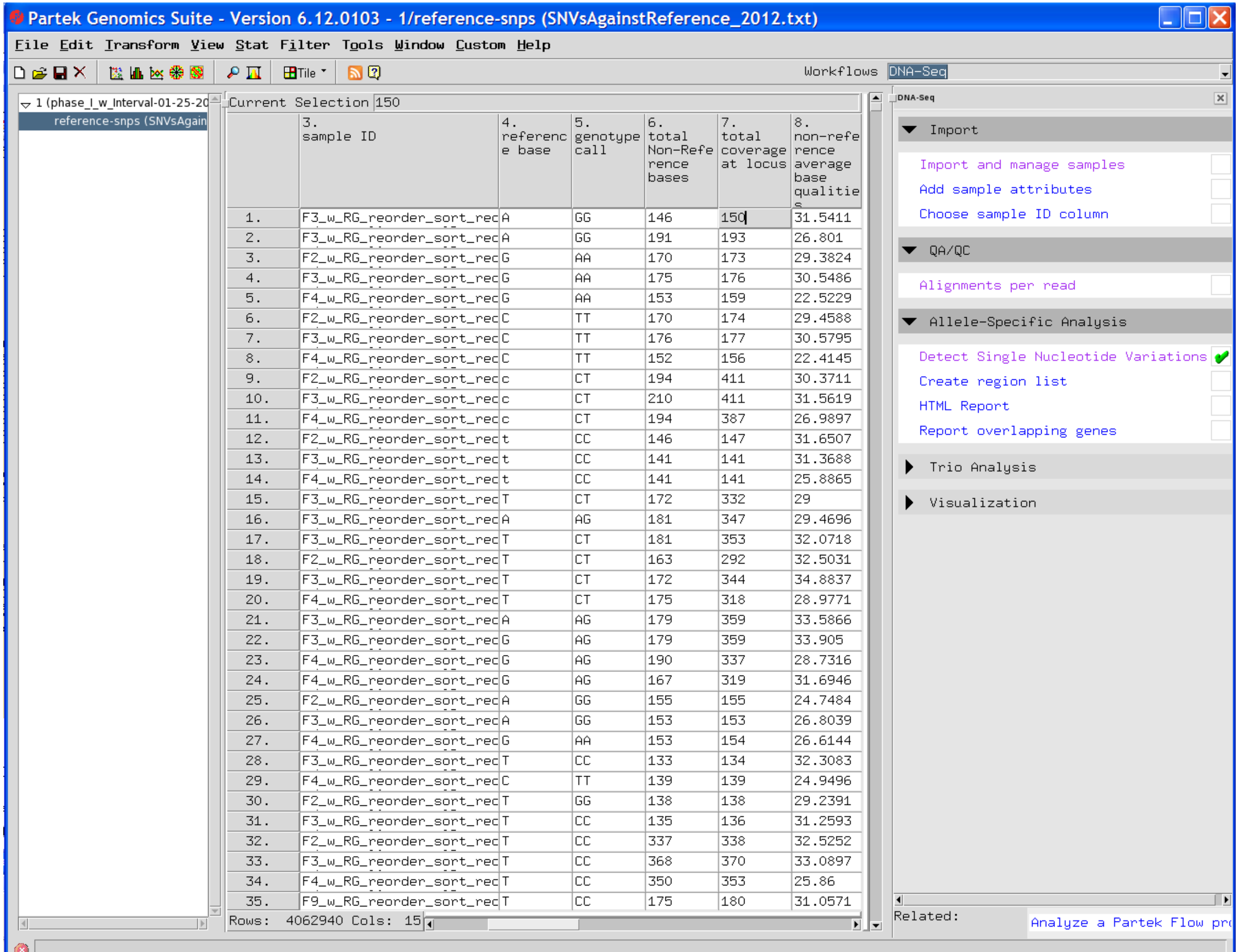
Partek SNP calling procedure



Partek SNP calling procedure



Partek SNP calling procedure-SNP result file



	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_rec	A	GG	146	150	31.5411
2.	F3_w_RG_reorder_sort_rec	A	GG	191	193	26.801
3.	F2_w_RG_reorder_sort_rec	G	AA	170	173	29.3824
4.	F3_w_RG_reorder_sort_rec	G	AA	175	176	30.5486
5.	F4_w_RG_reorder_sort_rec	G	AA	153	159	22.5229
6.	F2_w_RG_reorder_sort_rec	C	TT	170	174	29.4588
7.	F3_w_RG_reorder_sort_rec	C	TT	176	177	30.5795
8.	F4_w_RG_reorder_sort_rec	C	TT	152	156	22.4145
9.	F2_w_RG_reorder_sort_rec	c	CT	194	411	30.3711
10.	F3_w_RG_reorder_sort_rec	c	CT	210	411	31.5619
11.	F4_w_RG_reorder_sort_rec	c	CT	194	387	26.9897
12.	F2_w_RG_reorder_sort_rec	t	CC	146	147	31.6507
13.	F3_w_RG_reorder_sort_rec	t	CC	141	141	31.3688
14.	F4_w_RG_reorder_sort_rec	t	CC	141	141	25.8865
15.	F3_w_RG_reorder_sort_rec	T	CT	172	332	29
16.	F3_w_RG_reorder_sort_rec	A	AG	181	347	29.4696
17.	F3_w_RG_reorder_sort_rec	T	CT	181	353	32.0718
18.	F2_w_RG_reorder_sort_rec	T	CT	163	292	32.5031
19.	F3_w_RG_reorder_sort_rec	T	CT	172	344	34.8837
20.	F4_w_RG_reorder_sort_rec	T	CT	175	318	28.9771
21.	F3_w_RG_reorder_sort_rec	A	AG	179	359	33.5866
22.	F3_w_RG_reorder_sort_rec	G	AG	179	359	33.905
23.	F4_w_RG_reorder_sort_rec	G	AG	190	337	28.7316
24.	F4_w_RG_reorder_sort_rec	G	AG	167	319	31.6946
25.	F2_w_RG_reorder_sort_rec	A	GG	155	155	24.7484
26.	F3_w_RG_reorder_sort_rec	A	GG	153	153	26.8039
27.	F4_w_RG_reorder_sort_rec	G	AA	153	154	26.6144
28.	F3_w_RG_reorder_sort_rec	T	CC	133	134	32.3083
29.	F4_w_RG_reorder_sort_rec	C	TT	139	139	24.9496
30.	F2_w_RG_reorder_sort_rec	T	GG	138	138	29.2391
31.	F3_w_RG_reorder_sort_rec	T	CC	135	136	31.2593
32.	F2_w_RG_reorder_sort_rec	T	CC	337	338	32.5252
33.	F3_w_RG_reorder_sort_rec	T	CC	368	370	33.0897
34.	F4_w_RG_reorder_sort_rec	T	CC	350	353	25.86
35.	F9_w_RG_reorder_sort_rec	T	CC	175	180	31.0571

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

tork.ncifcrf.gov - PuTTY

```
torkv:/banas/kebebew/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