Variant calling in Exome-seq data using Varscan

Elodie Girard

U900 INSERM - Mines ParisTech - Institut Curie

ITMO Roscoff – 11/19/2013

institutCurie

Instituts thématiques Inserm
Institut national de la santé et de la recherche médicale

MINES ParisTech

Dataset

- Public data: exome sequenced by the International HapMap Project
- Single-end reads of 100bp, Illumina Genome Analyzer IIx
- RNA-seq data of this exome available (Pickrell *et al.*, Nature, 2010)

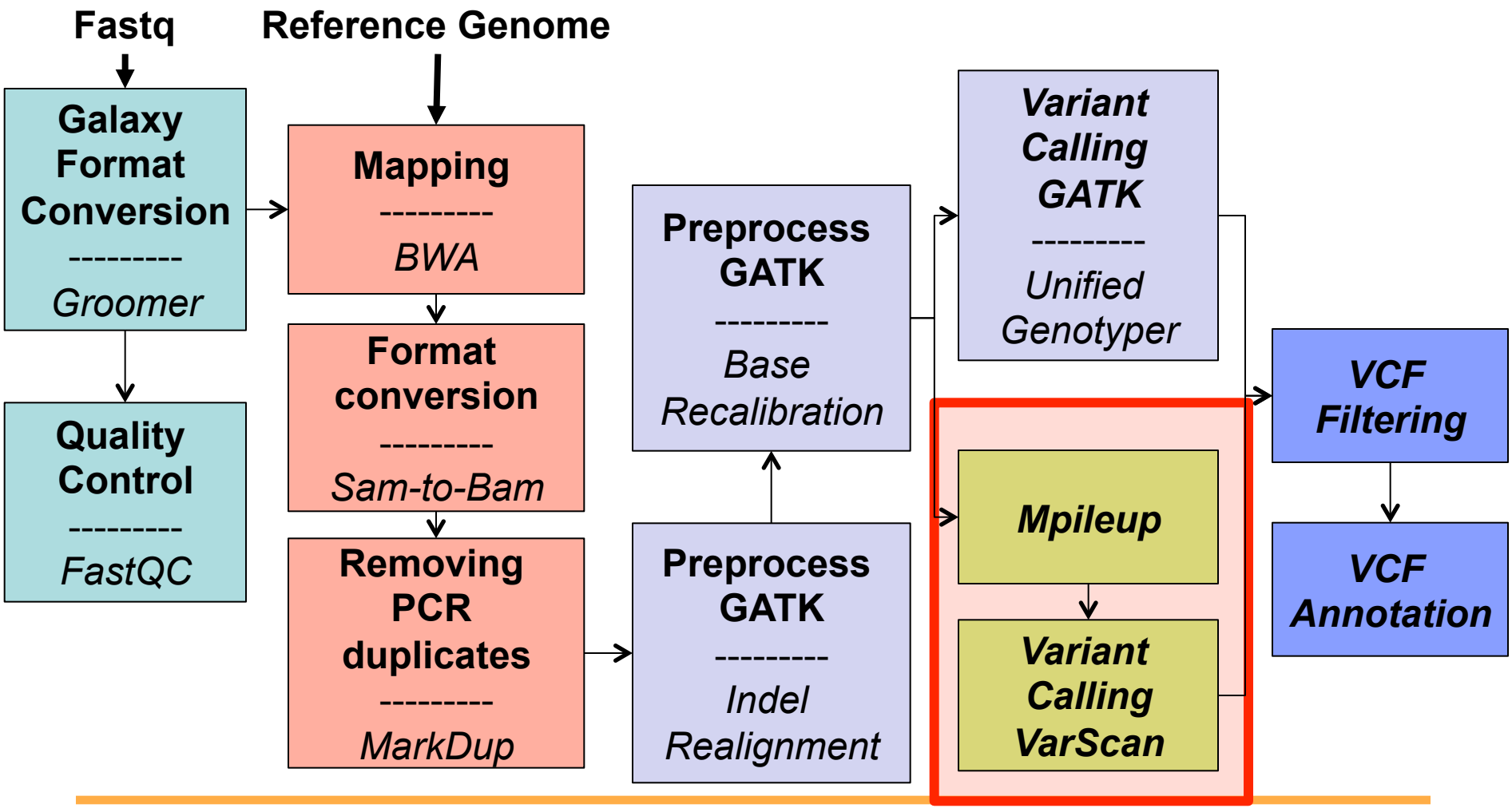
Objectives of the workshop:

- Variant calling, filtering and annotation in exome-seq data
- Observing the potential impact of these variants by looking at the corresponding RNA-seq data

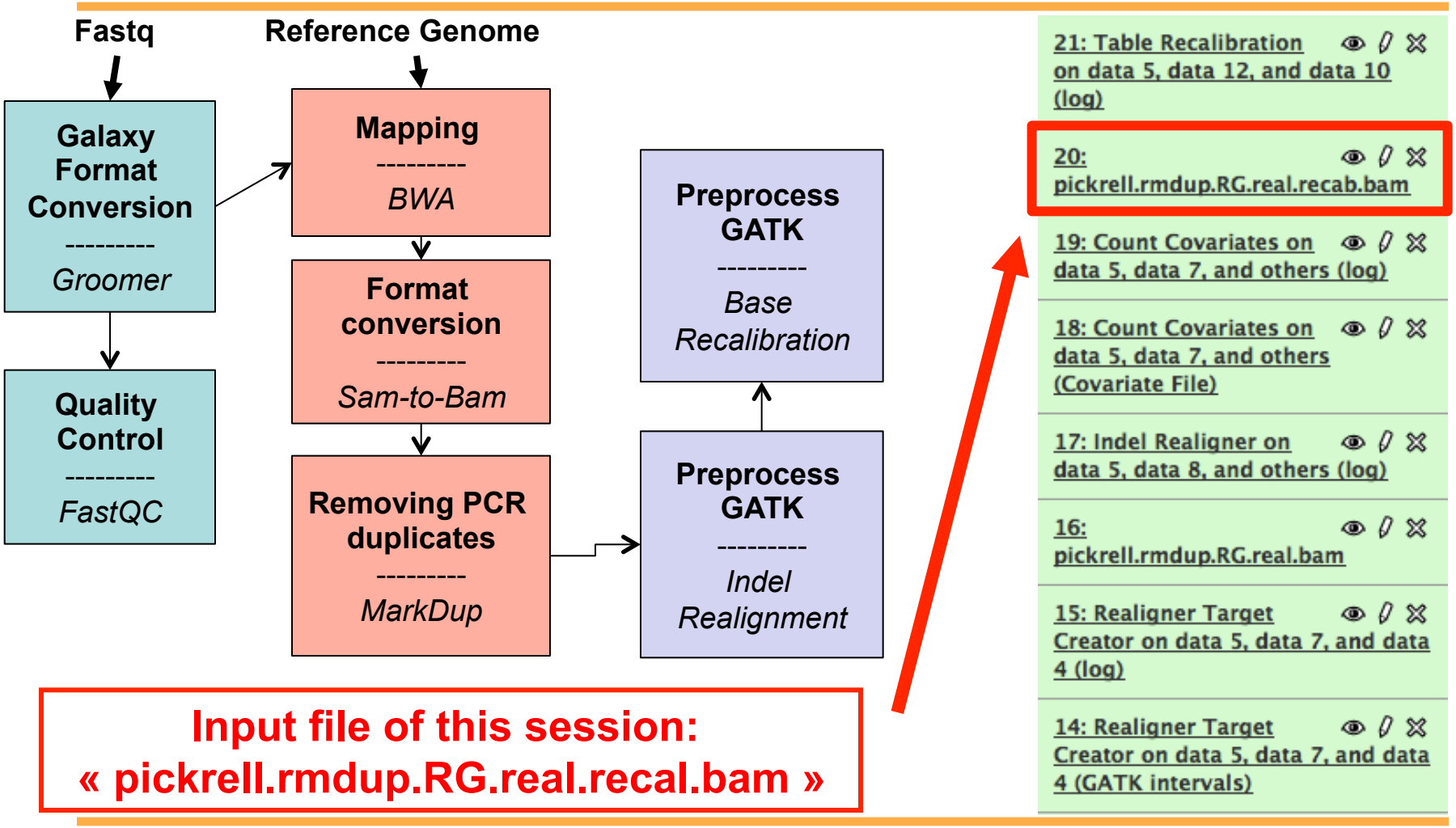
Objectif of this session :

➢ Variant detection using Varscan

Workflow


Fastq
Reference Genome
Galaxy Format Conversion
Groomer
Quality Control
FastQC
Mapping
BWA
Format conversion
Sam-to-Bam
Removing PCR duplicates
MarkDup
Preprocess GATK
Indel Realignment
Preprocess GATK
Base Recalibration
Variant Calling GATK
Unified Genotyper
Mpileup
Variant Calling VarScan
VCF Filtering
VCF Annotation

Galaxy: summary of the previous steps



Input file of this session:
« pickrell.rmdup.RG.real.recal.bam »

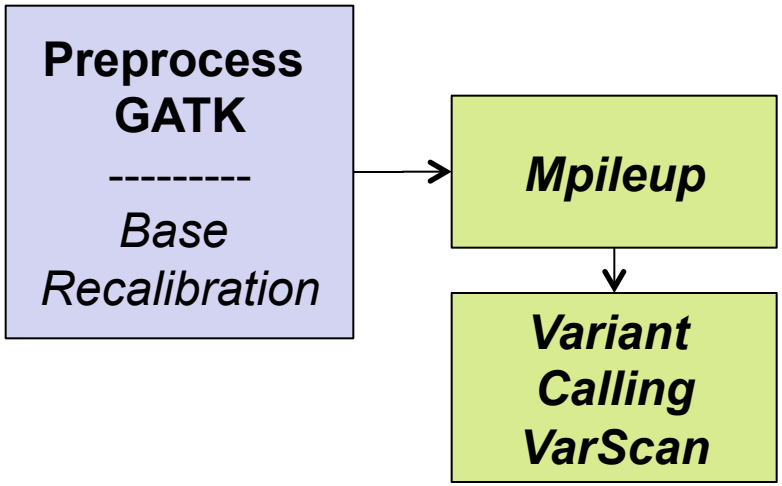
institutCurie

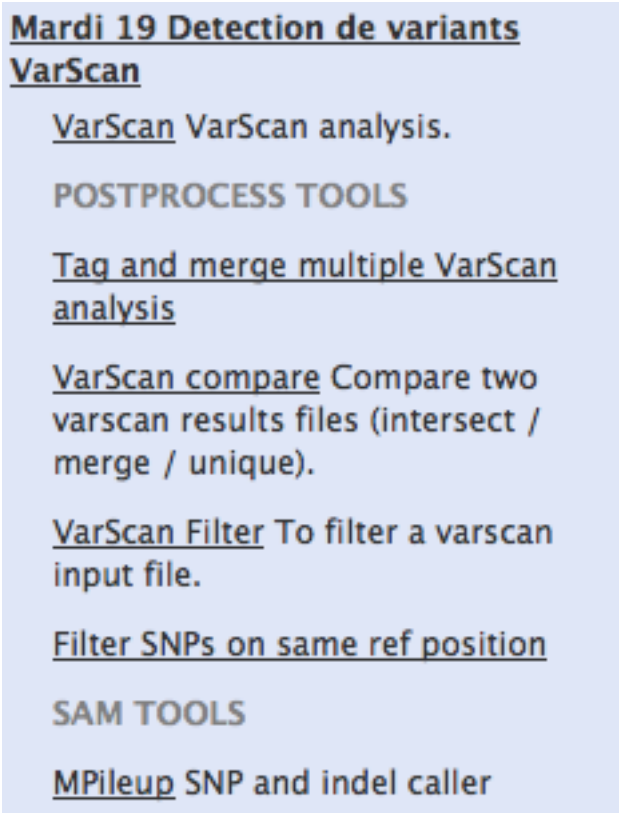
Instituts thématiques Inserm
Institut national de la santé et de la recherche médicale

MINES ParisTech

Tools

All the tools are in the left panel: **Mardi 19 Détection de variants VarScan**





Pileup

- **Pileup format:** describes the base-pair information at each position

Reference base

Base qualities

chr12	112888238	A	108	.$, $.,,.,,..,,,.,..,.,..,.,,,.,,,.,..,,..,,.,,.,.,.,,,.,.,,,.,.,.,.,,.,,.,.,., ,,,.,.,.,.,,,.,.,.,.,,,,,.,,.,.,..,,,,,,.,.,,,,^F,	=4 =???????@??@?@@ @=@ ??@ ? @? ? ??< ? ??@???????? ? @??? ? ??@?? A???@@ ?@@???AB????= ? @ @@??@@? @ A 00
chr12	112888239	C	108	. $t,.,,.T,tT,.,T.,.,t.tTtt.tTT,t.T,tTt.tT,T,,.tT tTttt.,.,Tt.ttt.,T,.,.tT,,T,T,.tT,,t,TttTtT,T.,t t,,T,T,tttt^F.^F,	936 78??6??6 45<875? ??? ?@6 @6???<? =6 66= ? ???? 6=7??=???<8@7=7=? 77?7? 8 ??78?7????? 8 <8??88 9?8 ?0048

Number of reads covering the site (total depth)

Read bases:

. / , = match on forward/reverse strand

ACGTN / acgtn = mismatch on forward/reverse strand

`-\+[0-9]+[ACGTNacgtn]+‘ indicates an indel

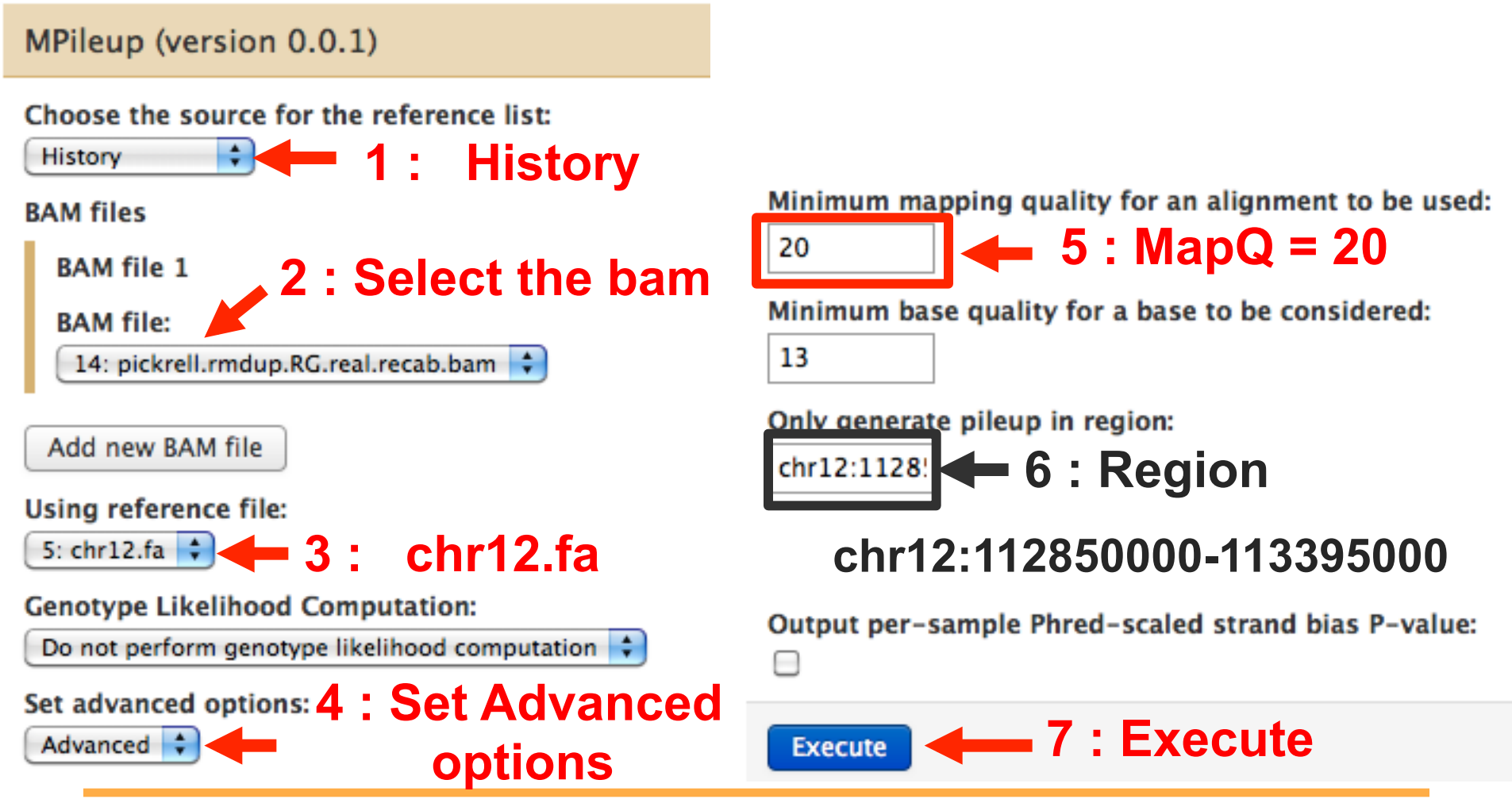
institut Curie

Instituts thématiques Inserm

Institut national de la santé et de la recherche médicale

MINES ParisTech

Mpileup



institutCurie

Instituts thématiques Inserm Institut national de la santé et de la recherche médicale

MINES ParisTech

VarScan

- Mutation caller written in **Java** (no installation required) working with **Pileup files** of Targeted, **Exome**, and Whole-Genome sequencing data

- **Multi-platforms**: Illumina, SOLiD, Life/PGM, Roche/454

- Detection of different kinds of variants (SNVs/Indels) :
  - Germline variants in individual samples
  - Multi-sample variants **shared or private** in multi-sample datasets

- VarScan specificity is to be able to work with **Tumor/Normal pairs**:
  - Somatic and germline mutation, LOH events in tumor-normal pairs
  - Somatic copy number alterations (CNAs) in tumor-normal exome data

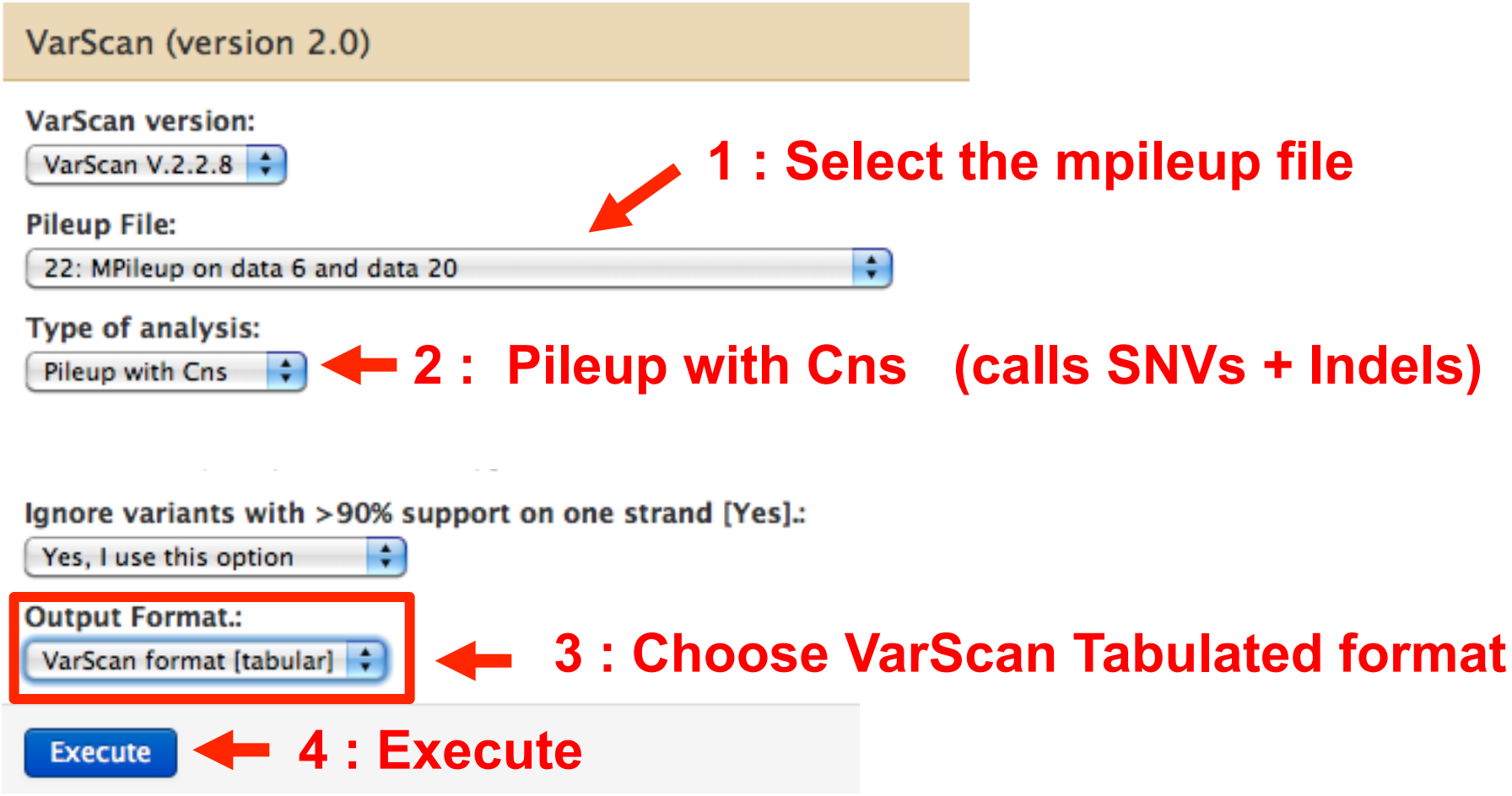
VarScan

- Most published variant callers use **Bayesian statistics** (a probabilistic framework) to detect variants and assess confidence in them (*e.g.:* GATK)

- VarScan uses a robust **heuristic/statistic** approach to call variants that meet desired thresholds for read depth, base quality, variant allele frequency, and statistical significance

- In Stead *et* al. (2013), they compared 3 different **somatic callers** : MuTect, Strelka, VarScan2
  - **VarScan2 performed best** overall with sequencing depths of 100x, 250x, 500x and 1000x required to accurately identify variants present at 10%, 5%, 2.5% and 1% respectively

VarScan


VarScan (version 2.0)
VarScan version:
VarScan V.2.2.8
1 : Select the mpileup file
Pileup File:
22: MPileup on data 6 and data 20
Type of analysis:
Pileup with Cns
2 : Pileup with Cns (calls SNVs + Indels)
Ignore variants with >90% support on one strand [Yes].:
Yes, I use this option
Output Format.:
VarScan format [tabular]
3 : Choose VarScan Tabulated format
Execute
4 : Execute

VarScan VCF Format

- 2 types: **VCF** and **Tabulated** Formats - **VarScan Specific**
  - ➢ VCF output might not work with some filtering or annotating tools

- **VarScan VCF format**: classic VCF header (#) but specific variant lines

#CHROM	POS	ID	REF	ALT	QUAL	FILTER	INFO	FORMAT	GENO
chr12	250239	.	A	G	20	PASS	ADP=104;WT=0; **HET=1**;HOM=0; NC=0	GT:GQ:SDP:DP:RD:AD: FREQ: PVAL:RBQ:ABQ: RDF:RDR:ADF:ADR	0/1:153:111:104:61:43: 41,35%: 4,5644E-16:38:32: 48:13:36:7

ADP = Average per-sample depth of bases with Phred score = 20
WT = Number of samples called reference (wild-type)
HET = Number of samples called heterozygous-variant
HOM = Number of samples called homozygous-variant
NC = Number of samples not called

➔ Useful in multi-sample studies

VarScan VCF Format

- 2 types: **VCF** and **Tabulated** Formats - **VarScan Specific**
  - ➢ VCF output might not work with some filtering or annotating tools

- **VarScan VCF format**: classic VCF header (#) but specific variant lines

#CHROM	POS	ID	REF	ALT	QUAL	FILTER	INFO	FORMAT	GENO
chr12	250239	.	A	G	20	PASS	ADP=104;WT=0;HET=1;HOM=0;NC=0	**GT**:GQ:SDP:**DP**:**RD:AD**:**FREQ**:	**0/1**:153:111:**104**:**61:43**:**41,35%**:
								PVAL:RBQ:ABQ:RDF:RDR:ADF:ADR	4,5644E-16:38:32:48:13:36:7

GT=Genotype (1/1: Homozygous ; 0/1 : Heterozygous) / GQ= Genotype Quality
SDP= Raw Read Depth as reported by SAMtools
DP= Quality Read Depth of bases with Phred score >= 20
RD= Depth of reference-supporting bases
AD= Depth of variant-supporting bases
FREQ= Variant allele frequency

VarScan VCF Format

- 2 types: **VCF** and **Tabulated** Formats - **VarScan Specific**
  - ➢ VCF output might not work with some filtering or annotating tools

- **VarScan VCF format**: classic VCF header (#) but specific variant lines

#CHROM	POS	ID	REF	ALT	QUAL	FILTER	INFO	FORMAT	GENO
chr12	250239	.	A	G	20	PASS	ADP=104;WT=0; HET=1;HOM=0; NC=0	GT:GQ:SDP:DP:RD:AD: FREQ: PVAL:RBQ:ABQ: **RDF:RDR:ADF:ADR**	0/1:153:111:104:61:43: 41,35%: 4,5644E-16:38:32: **48:13:36:7**

PVAL= P-value from Fisher's Exact Test (**not computed here : default value)**
RBQ= Average quality of reference-supporting bases
ABQ= Average quality of variant-supporting bases
RDF / RDR= Depth of reference-supporting bases on forward/reverse strand
ADF / ADR = Depth of variant-supporting bases on forward/reverse strand

VarScan VCF Format

- 2 types: **VCF** and **Tabulated** Formats - **VarScan Specific**
  - ➢ VCF output might not work with some filtering or annotating tools

- **VarScan VCF format**: classic VCF header (#) but specific variant lines

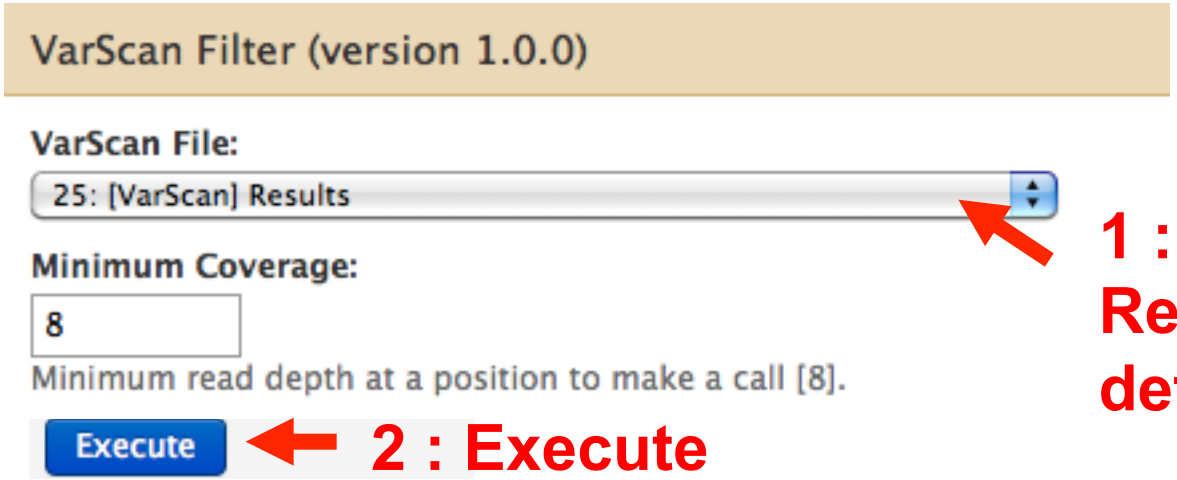
#CHROM	POS	ID	REF	ALT	QUAL	FILTER	INFO	FORMAT	GENO
chr12	250239	.	A	G	20	PASS	ADP=104;WT=0; HET=1;HOM=0; NC=0	GT:GQ:SDP:DP:RD:AD: FREQ: PVAL:RBQ:ABQ: **RDF:RDR:ADF:ADR**	0/1:153:111:104:61:43: 41,35%: 4,5644E-16:38:32: **48:13:36:7**

PVAL= P-value from Fisher's Exact Test (**not computed here : default value)**
RBQ= Average quality of reference-supporting bases
ABQ= Average quality of variant-supporting bases
RDF / RDR= Depth of reference-supporting bases on forward/reverse strand
ADF / ADR = Depth of variant-supporting bases on forward/reverse strand

VarScan Tabulated Format

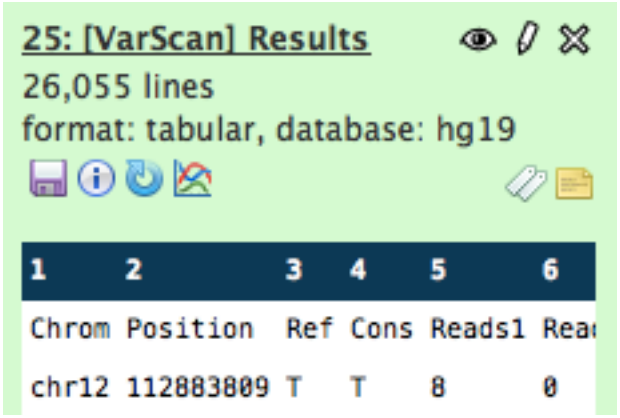
- 2 types: **VCF** and **Tabulated** Formats - **VarScan Specific**
  - Tabulated output works with other VarScan Tools
  - By default on Galaxy, VarScan outputs a line for each base covered by the selected minimum coverage **even if there is no alternative variant**

→ **Pre-process**: use « **VarScan Filter** » to keep only variants

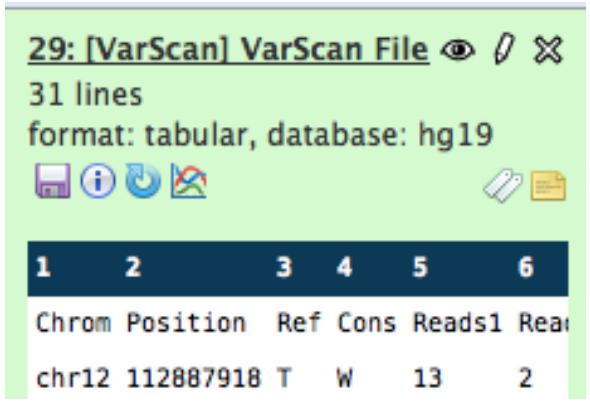


1 : Select the VarScan Results and leave the default parameters

VarScan Tabulated Format



VarScan Filter

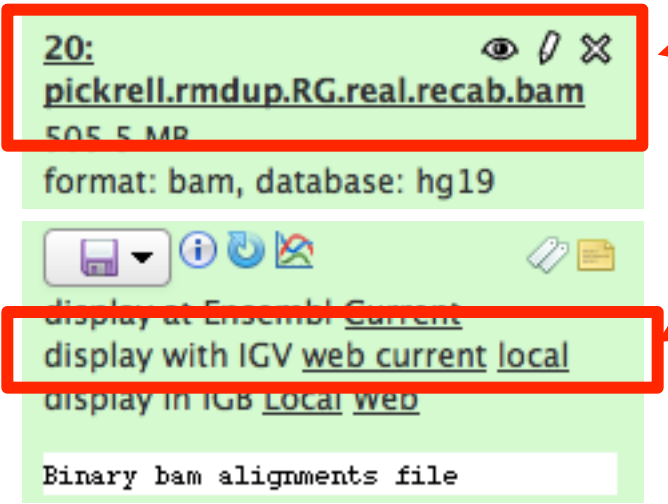


Chrom	Position	Ref	Cons	Reads1	Reads2	VarFreq	Strands 1	Strands 2	Qual1	Qual2	Pvalue	Map Qual1	Map Qual2	R1 +	R1 -	R2 +	Rs2 -	Alt
chr12	113348849	C	Y	31	30	**49.18%**	2	2	27	27	0.98	1	1	19	12	25	5	**T**
chr12	113354329	G	R	72	2	**2.70%**	2	2	31	26	0.98	1	1	48	24	1	1	**A**
chr12	113357193	G	A	2	72	**97.30%**	1	2	28	24	0.98	1	1	2	0	45	27	**A**
chr12	113357209	G	A	0	77	**100%**	0	2	0	29	0.98	0	1	0	0	51	26	**A**

Cons : Consensus Genotype of Variant Called (IUPAC code):

M -> A or C	Y -> C or T	D -> A or G or T	W -> A or T	V -> A or C or G
R -> A or G	K -> G or T	B -> C or G or T	S -> C or G	H -> A or C or T

Variants visualization with IGV



1: Click on the name to open the options

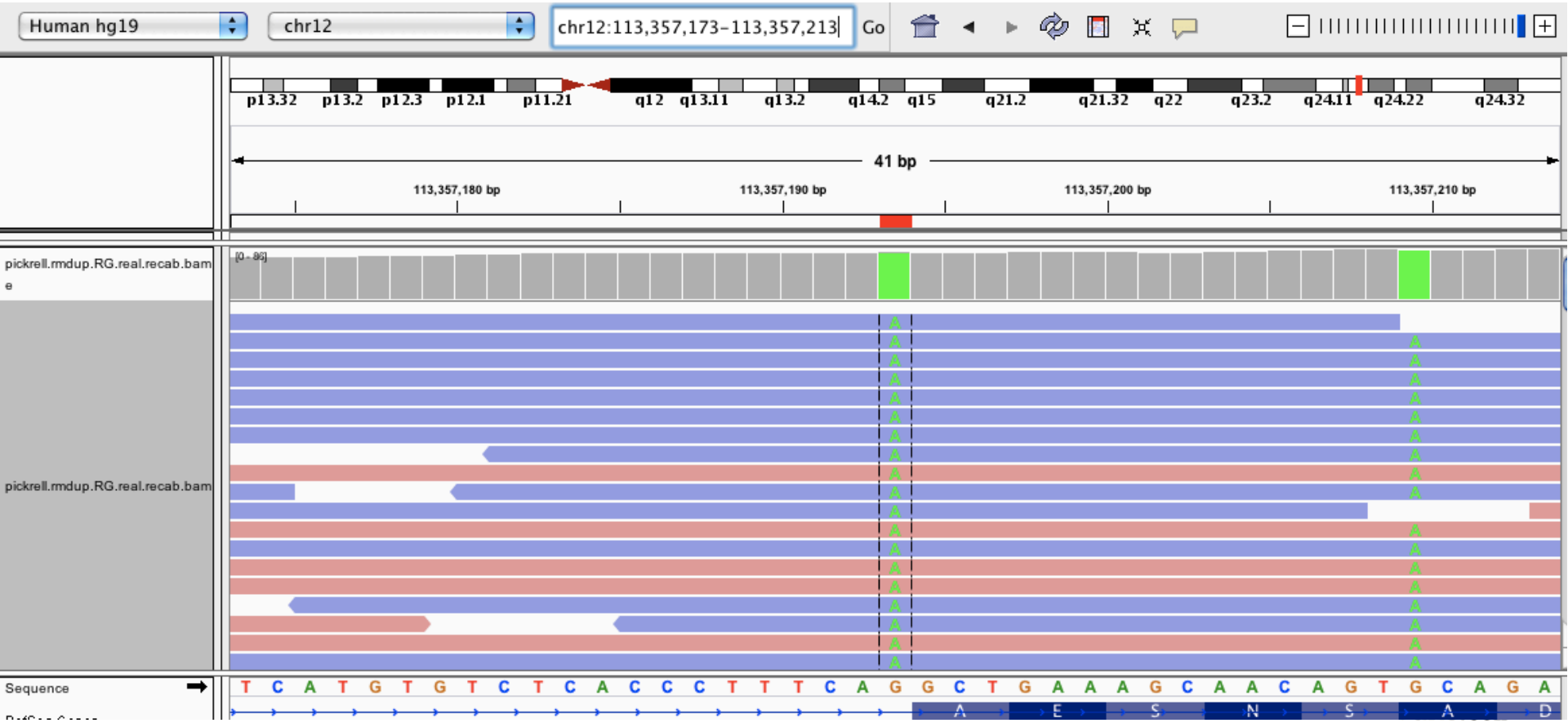
2: Click on « Local » to open it on IGV (if already open, otherwise « Web current »)

- Look at those two variants:

Chrom	Position	Ref	Cons	Reads 1	Reads 2	VarFreq	Strands 1	Strands 2	Qual 1	Qual2	Pval	Map Qual1	Map Qual2	R1 +	R1 -	R2 +	Rs2 -	Alt
chr12	113357193	G	A	2	72	97.30%	1	2	28	24	0.98	1	1	2	0	45	27	A
chr12	112888239	C	Y	52	56	51.85%	2	2	24	28	0.98	1	1	20	32	24	32	T

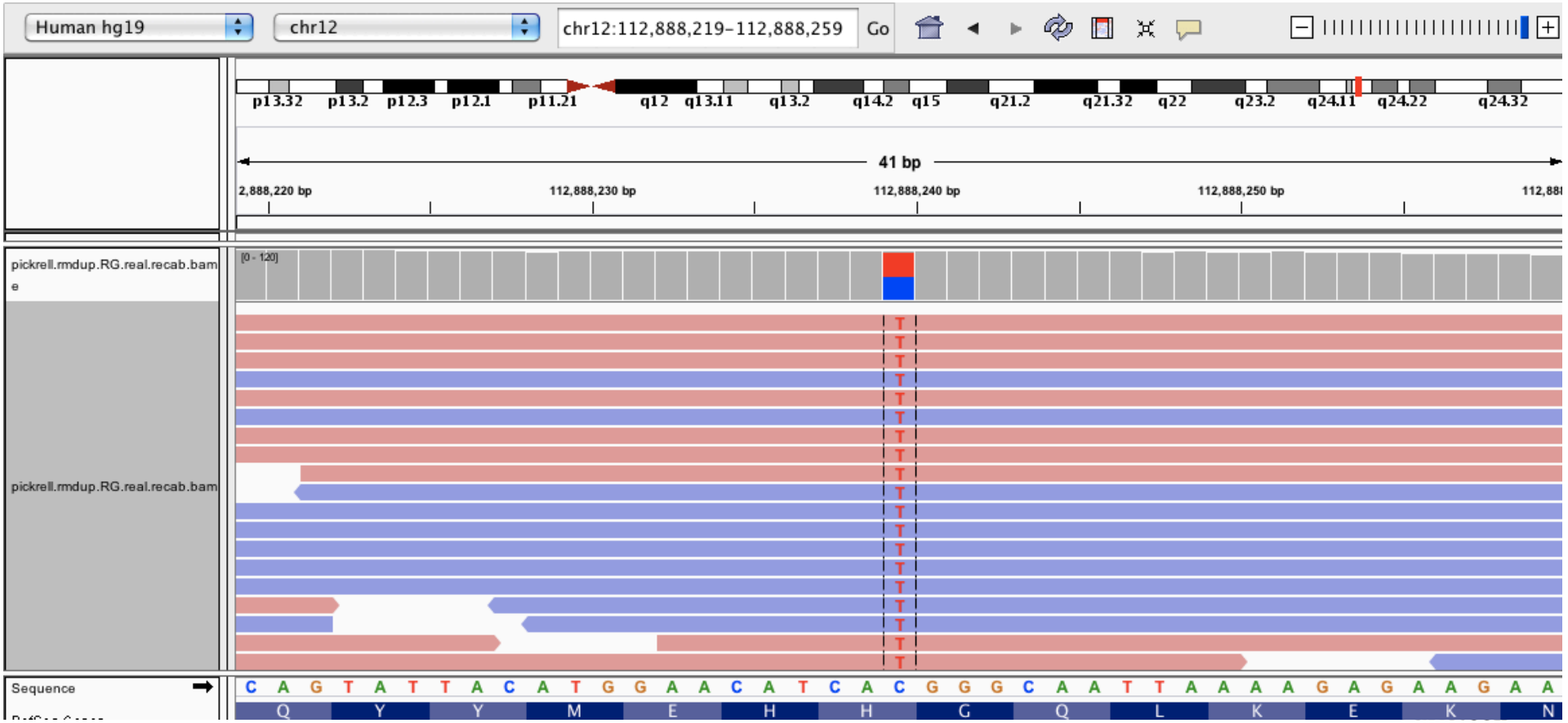
Variants visualization with IGV

Chrom	Position	Ref	Cons	Reads 1	Reads 2	VarFreq	Strands 1	Strands 2	Qual 1	Qual2	Pval	Map Qual1	Map Qual2	R1 +	R1 -	R2 +	Rs2 -	Alt
chr12	113357193	G	A	2	72	97.30%	1	2	28	24	0.98	1	1	2	0	45	27	A

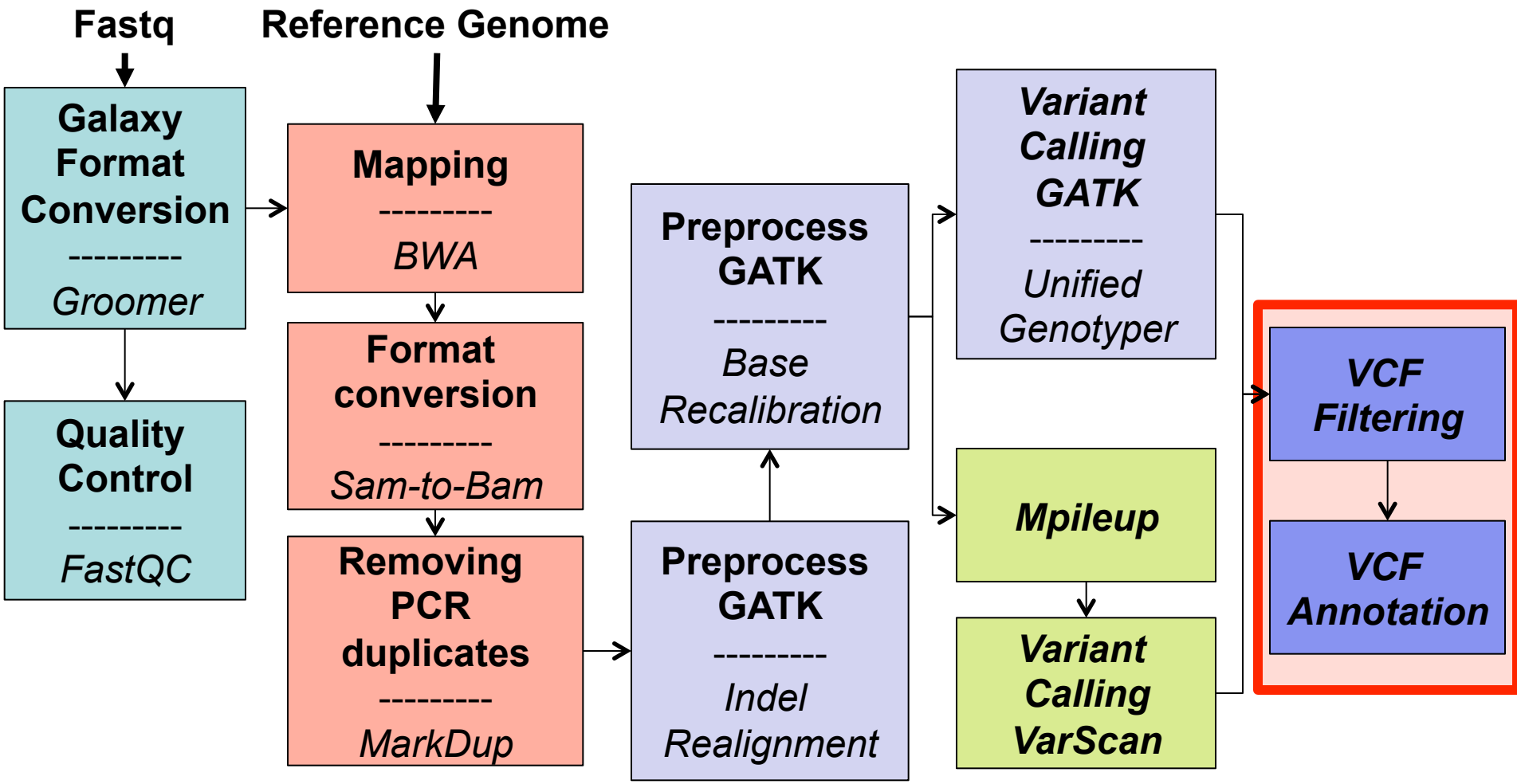


Variants visualization with IGV

Chrom	Position	Ref	Cons	Reads 1	Reads 2	VarFreq	Strands 1	Strands 2	Qual 1	Qual2	Pval	Map Qual1	Map Qual2	R1 +	R1 -	R2 +	Rs2 -	Alt
chr12	112888239	C	Y	52	56	51.85%	2	2	24	28	0.98	1	1	20	32	24	32	T



Next Step


Fastq
Reference Genome
Galaxy Format Conversion

Groomer
Quality Control

FastQC
Mapping

BWA
Format conversion

Sam-to-Bam
Removing PCR duplicates

MarkDup
Preprocess GATK

Indel Realignment
Preprocess GATK

Base Recalibration
Variant Calling GATK

Unified Genotyper
Mpileup
Variant Calling VarScan
VCF Filtering
VCF Annotation