CLASE 12

Llamado de variantes o SNP calling

DBT 792
GENÉTICA Y GENÓMICA EN PRODUCCIÓN ANIMAL

Profesor
Dr. José Gallardo

PLAN DE LA CLASE

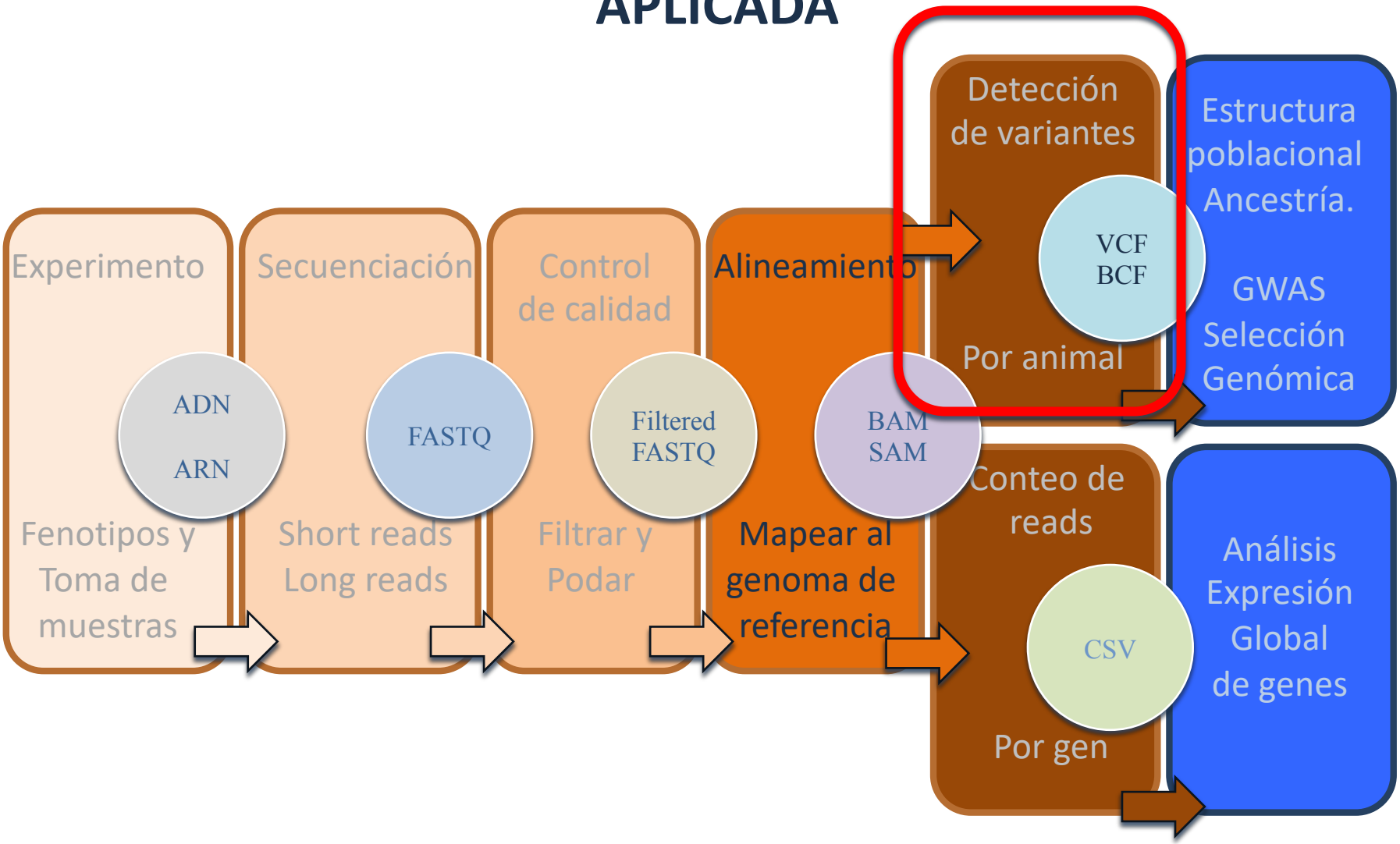
Introducción

- Repaso flujo de trabajo proyecto genómica aplicada.
- ¿Por qué y para qué buscar variantes?
- Formato de llamado de variantes .vcf
- Software GATK: Llamado de variantes
- Linux y vcftools: Para manipular y procesamiento básico de archivos .vcf

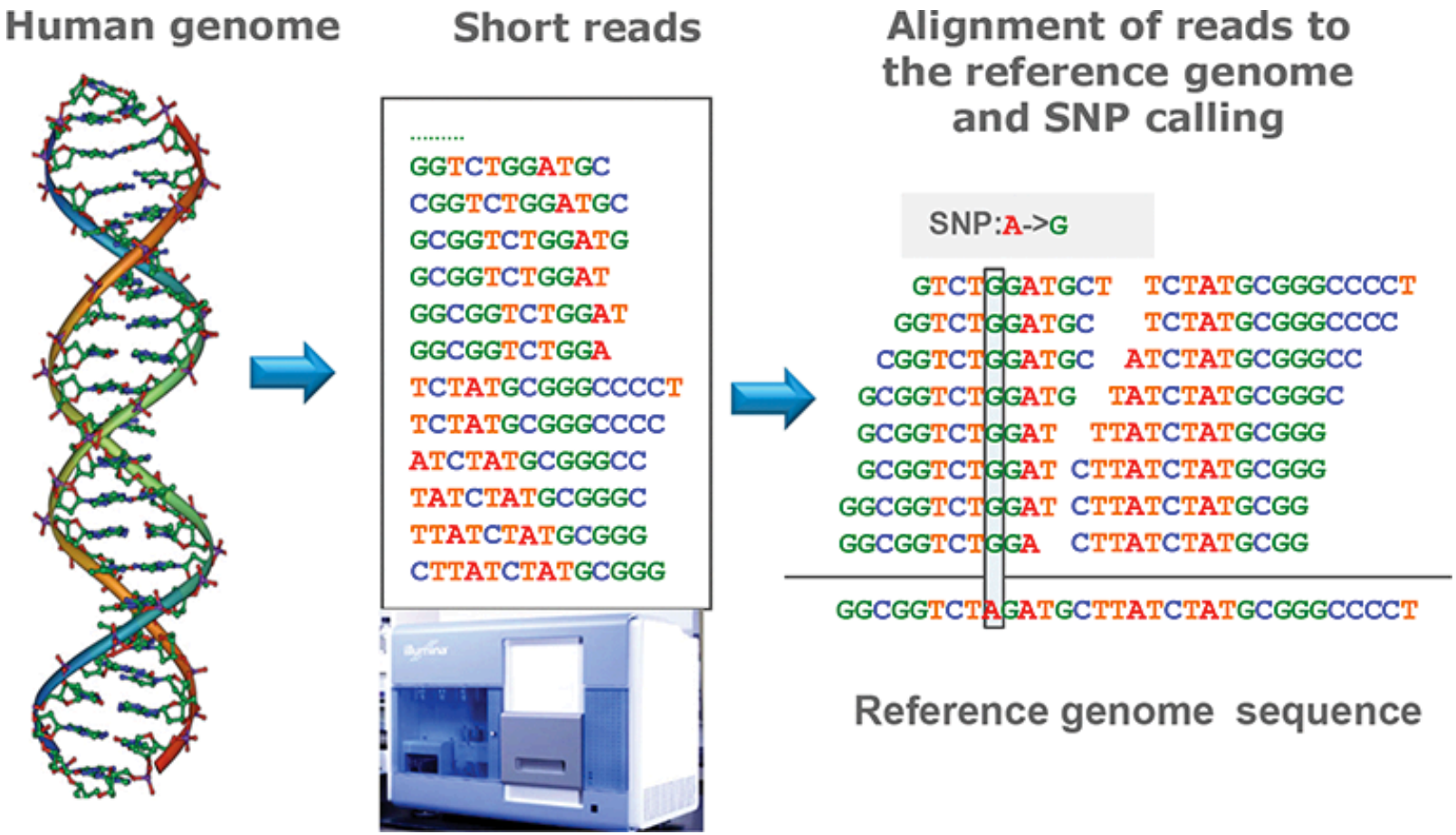
Práctica

- Linux para genómica: Llamado de variantes.

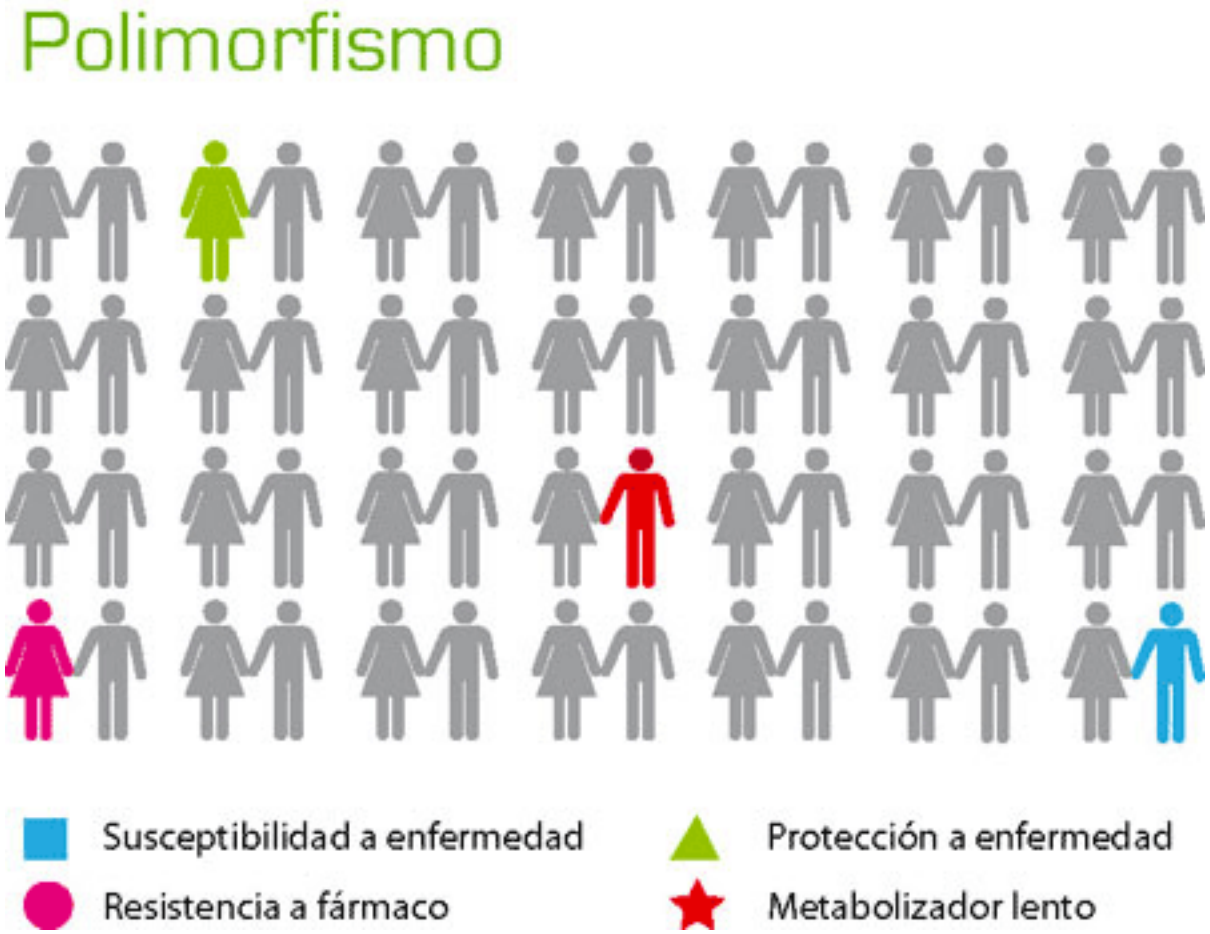
FLUJO DE TRABAJO DE UN PROYECTO DE GENÓMICA APLICADA



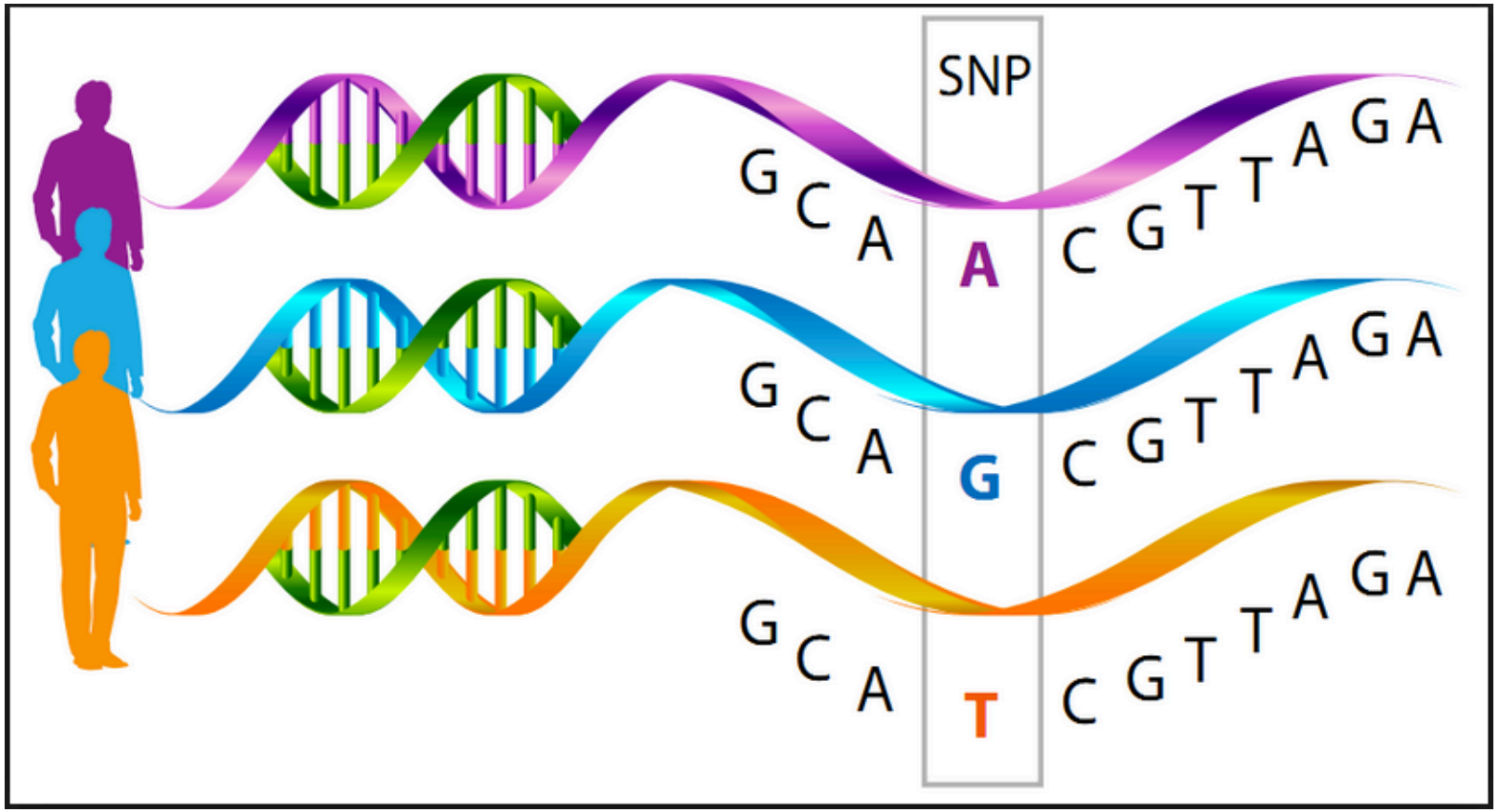
RESUMEN DEL FLUJO DE TRABAJO PARA LLAMAR VARIANTES



¿POR QUÉ ES IMPORTANTES EL POLIMORFISMO?



¿QUÉ VARIANTES PUEDO ENCONTRAR EN EL ADN?



¿QUÉ VARIANTES PUEDO ENCONTRAR EN EL ADN?

Indel examples

wild-type sequence

ATCTTCAGCCATAAAAGATGAAGTT

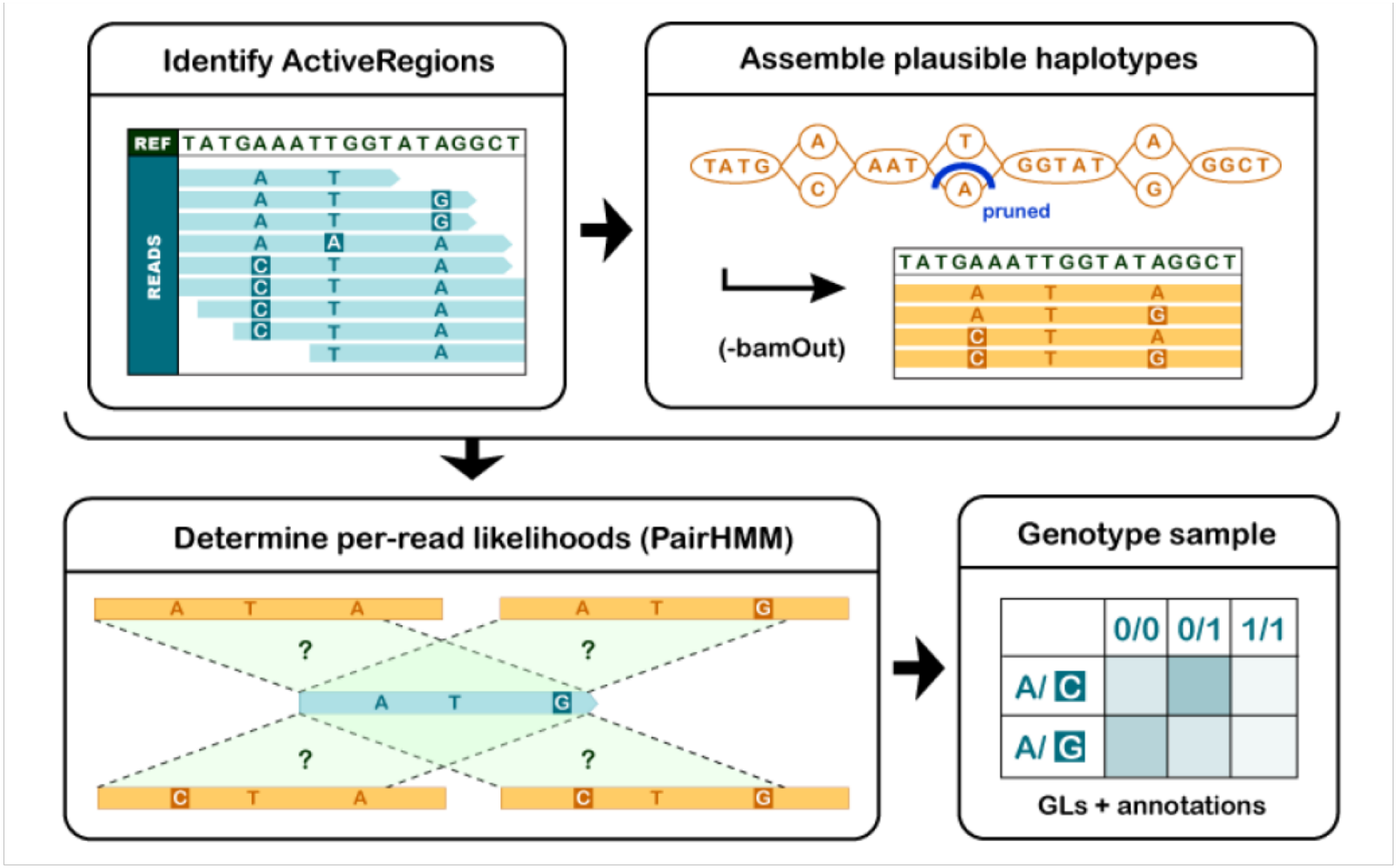
3 bp deletion

ATCTTCAGCCAAAGATGAAGTT

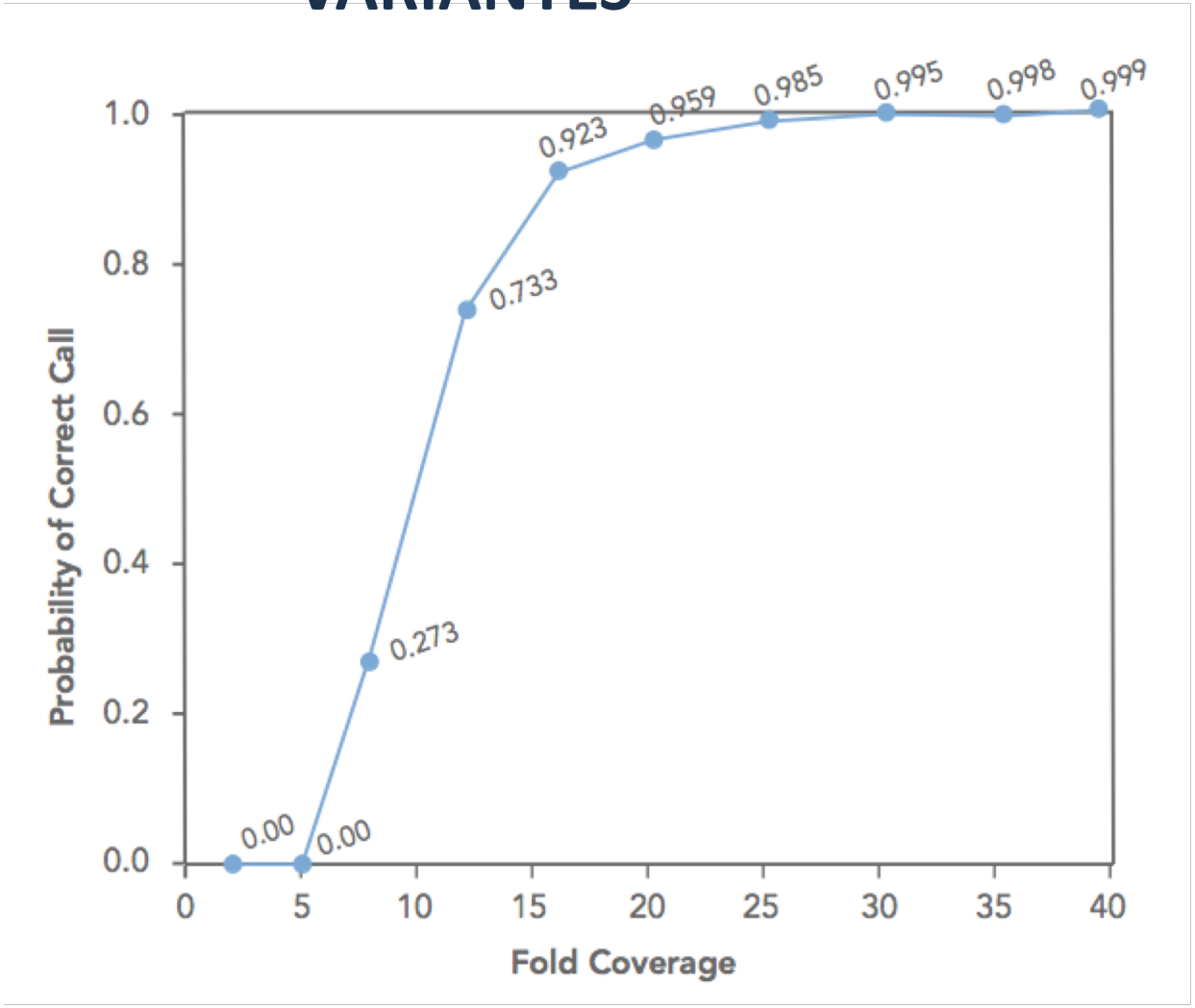
4 bp insertion (orange)

ATCTTCAGCCATATGTGAAAGATGAAGTT

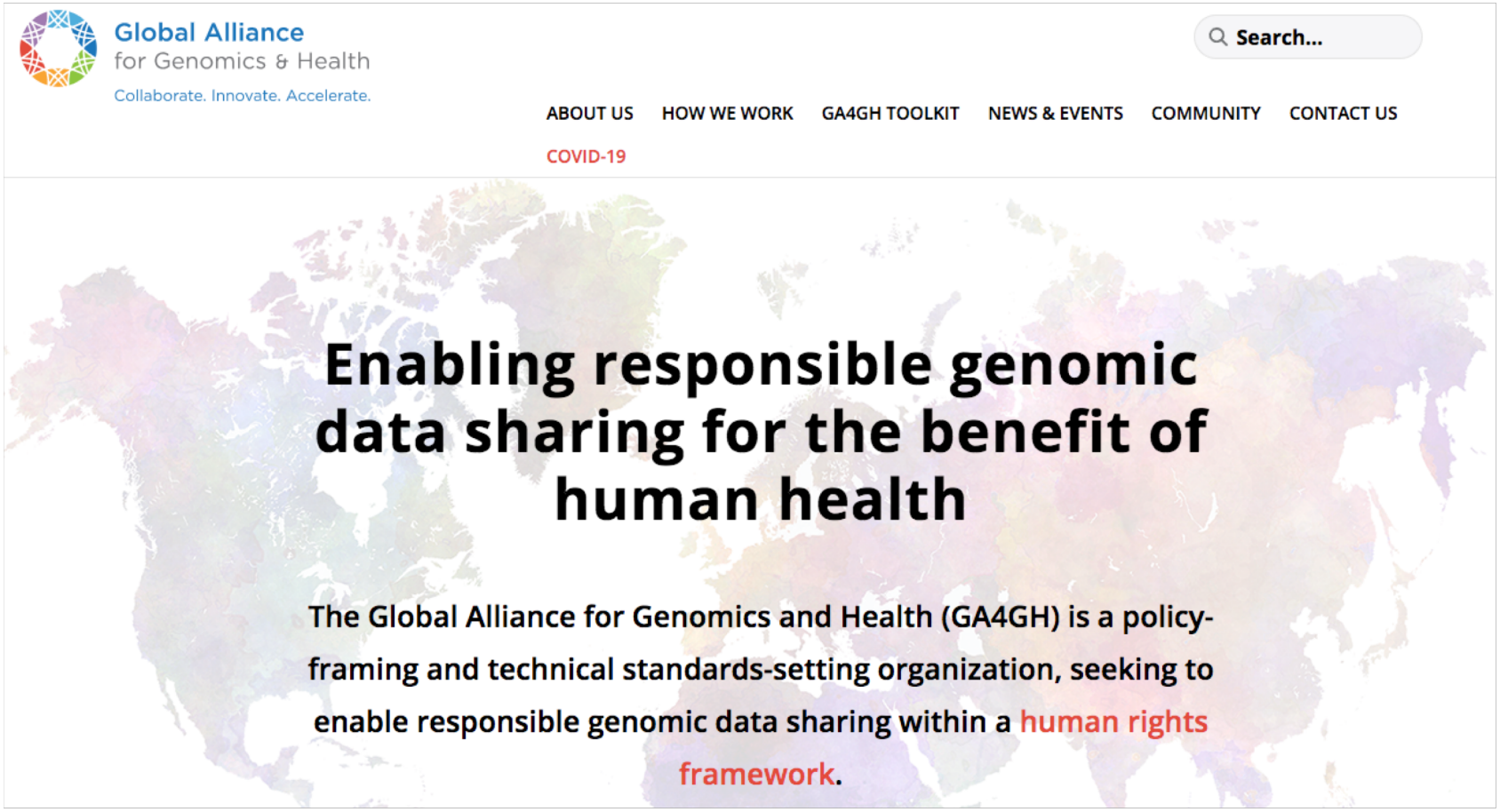
GATK: SOFTWARE PARA LLAMADO DE VARIANTES



LA IMPORTANCIA DE LA COVERTURA PARA DESCUBRIR VARIANTES



The Variant Call Format (VCF)
Version 4.1 – 4.2 – 4.3



The Variant Call Format (VCF) Version 4.1 Specification ##FORMAT

##FORMAT=<ID=AD,
Number=R,Type=Integer,Description="Allelic depths for the ref and alt alleles in the order listed">
##FORMAT=<ID=DP,
Number=1,Type=Integer,Description="Approximate read depth (reads with MQ=255 or with bad mates are filtered)">
##FORMAT=<ID=GQ,
Number=1,Type=Integer,Description="Genotype Quality">
##FORMAT=<ID=GT,
Number=1,Type=String,Description="Genotype">
##FORMAT=<ID=PL,
Number=G,Type=Integer,Description="Normalized, Phred-scaled likelihoods for genotypes as defined in the VCF specification">

#CHROM	POS	ID	REF	ALT	QUAL	FILTER	INFO	FORMAT	FR07958834	FR07958842	FR
1	155	.	G	A	3936.94	PASS	.	GT:AD:DP:GQ:PL	0/0:16,0:16:39:0,39,585		0/
1	249	.	A	C	1046.42	PASS	.	GT:AD:DP:GQ:PL	0/0:16,0:16:48:0,48,720		0/
1	290	.	A	G	4564.38	PASS	.	GT:AD:DP:GQ:PL	0/0:20,0:20:60:0,60,880		0/
1	345	.	T	C	6285.38	PASS	.	GT:AD:DP:GQ:PL	0/0:27,0:27:72:0,72,1080		
1	366	.	A	T	7984.14	PASS	.	GT:AD:DP:GQ:PL	0/1:15,11:26:99:395,0,586		
1	369	.	C	T	6171.75	PASS	.	GT:AD:DP:GQ:PL	0/0:25,0:25:63:0,63,945		0/
1	416	.	A	T	16059.6	PASS	.	GT:AD:DP:GQ:PL	0/1:17,13:30:99:439,0,659		

The Variant Call Format (VCF) Version 4.1 Specification ## INFO

##INFO=<ID=AC,
Number=A,Type=Integer,Description="Allele count in genotypes, for each ALT allele, in the same order as listed">
##INFO=<ID=AF,
Number=A,Type=Float,Description="Allele Frequency, for each ALT allele, in the same order as listed">
##INFO=<ID=AN,
Number=1,Type=Integer,Description="Total number of alleles in called genotypes">
##INFO=<ID=BaseQRankSum,
Number=1,Type=Float,Description="Z-score from Wilcoxon rank sum test of Alt Vs. Ref base qualities">
##INFO=<ID=DP,
Number=1,Type=Integer,Description="Approximate read depth; some reads may have been filtered">
##INFO=<ID=ExcessHet,
Number=1,Type=Float,Description="Phred-scaled p-value for exact test of excess heterozygosity">
##INFO=<ID=FS,
Number=1,Type=Float,Description="Phred-scaled p-value using Fisher's exact test to detect strand bias">
##INFO=<ID=InbreedingCoeff,
Number=1,Type=Float,Description="Inbreeding coefficient as estimated from the genotype likelihoods per-sample when compared against the Hardy-Weinberg expectation">
##INFO=<ID=MLEAC,
Number=A,Type=Integer,Description="Maximum likelihood expectation (MLE) for the allele counts (not necessarily the same as the AC), for each ALT allele, in the same order as listed”>

The Variant Call Format (VCF) Version 4.1 Specification ## INFO

```
AC=2;AF=1.00;AN=2;DP=9;ExcessHet=3.0103;FS=0.000;MLEAC=2;MLEAF=1.00;MQ=60.00;QD=26.42;SOR=4.615 G
AC=2;AF=1.00;AN=2;DP=8;ExcessHet=3.0103;FS=0.000;MLEAC=2;MLEAF=1.00;MQ=60.00;QD=28.26;SOR=1.179 G
AC=2;AF=1.00;AN=2;DP=13;ExcessHet=3.0103;FS=0.000;MLEAC=2;MLEAF=1.00;MQ=60.00;QD=35.67;SOR=2.147
AC=2;AF=1.00;AN=2;DP=16;ExcessHet=3.0103;FS=0.000;MLEAC=2;MLEAF=1.00;MQ=60.00;QD=31.27;SOR=2.049
AC=2;AF=1.00;AN=2;DP=15;ExcessHet=3.0103;FS=0.000;MLEAC=2;MLEAF=1.00;MQ=60.00;QD=28.18;SOR=2.448
AC=2;AF=1.00;AN=2;DP=14;ExcessHet=3.0103;FS=0.000;MLEAC=2;MLEAF=1.00;MQ=60.00;QD=26.96;SOR=1.765
AC=2;AF=1.00;AN=2;DP=7;ExcessHet=3.0103;FS=0.000;MLEAC=2;MLEAF=1.00;MQ=60.00;QD=31.71;SOR=4.174 G
AC=2;AF=1.00;AN=2;DP=45;ExcessHet=3.0103;FS=0.000;MLEAC=2;MLEAF=1.00;MQ=60.00;QD=27.39;SOR=1.954
AC=2;AF=1.00;AN=2;DP=44;ExcessHet=3.0103;FS=0.000;MLEAC=2;MLEAF=1.00;MQ=60.00;QD=22.48;SOR=1.954
AC=2;AF=1.00;AN=2;DP=18;ExcessHet=3.0103;FS=0.000;MLEAC=2;MLEAF=1.00;MQ=60.00;QD=34.75;SOR=0.941
AC=2;AF=1.00;AN=2;DP=44;ExcessHet=3.0103;FS=0.000;MLEAC=2;MLEAF=1.00;MQ=60.00;QD=28.17;SOR=1.005
AC=2;AF=1.00;AN=2;BaseQRankSum=-1.085;DP=56;ExcessHet=3.0103;FS=0.000;MLEAC=2;MLEAF=1.00;MQ=60.00
AC=2;AF=1.00;AN=2;DP=35;ExcessHet=3.0103;FS=0.000;MLEAC=2;MLEAF=1.00;MQ=60.00;QD=30.08;SOR=1.751
AC=2;AF=1.00;AN=2;DP=41;ExcessHet=3.0103;FS=0.000;MLEAC=2;MLEAF=1.00;MQ=60.00;QD=29.00;SOR=1.179
AC=2;AF=1.00;AN=2;BaseQRankSum=-0.398;DP=59;ExcessHet=3.0103;FS=3.163;MLEAC=2;MLEAF=1.00;MQ=60.00
AC=2;AF=1.00;AN=2;DP=37;ExcessHet=3.0103;FS=0.000;MLEAC=2;MLEAF=1.00;MQ=60.00;QD=25.15;SOR=1.002
AC=2;AF=1.00;AN=2;DP=30;ExcessHet=3.0103;FS=0.000;MLEAC=2;MLEAF=1.00;MQ=60.00;QD=34.18;SOR=1.528
AC=2;AF=1.00;AN=2;DP=16;ExcessHet=3.0103;FS=0.000;MLEAC=2;MLEAF=1.00;MQ=60.00;QD=36.27;SOR=1.244
AC=2;AF=1.00;AN=2;DP=20;ExcessHet=3.0103;FS=0.000;MLEAC=2;MLEAF=1.00;MQ=60.00;QD=29.33;SOR=2.093
AC=2;AF=1.00;AN=2;DP=19;ExcessHet=3.0103;FS=0.000;MLEAC=2;MLEAF=1.00;MQ=60.00;QD=32.52;SOR=2.712
```

Software vcftools

VCFtools

A set of tools written in Perl and C++ for working with VCF files.

Home

Documentation

Download ZIP

VCFtools es un software diseñado para trabajar con archivos VCF, como los generados por el Proyecto 1000 Genomas. El objetivo de VCFtools es proporcionar métodos de fácil acceso para trabajar con datos complejos de variación genética en forma de archivos VCF.

Filtra variantes específicas:
Comparar archivos
Resumir variantes
Convierta a diferentes tipos de archivos
Validar y fusionar archivos
Crea intersecciones y subconjuntos de variantes.

OBJETIVOS DEL TRABAJO PRÁCTICO

Esta práctica tiene como propósito:

Realizar el llamado de variantes de una secuencia de Salmo salar obtenida por secuenciación de extremos emparejados.