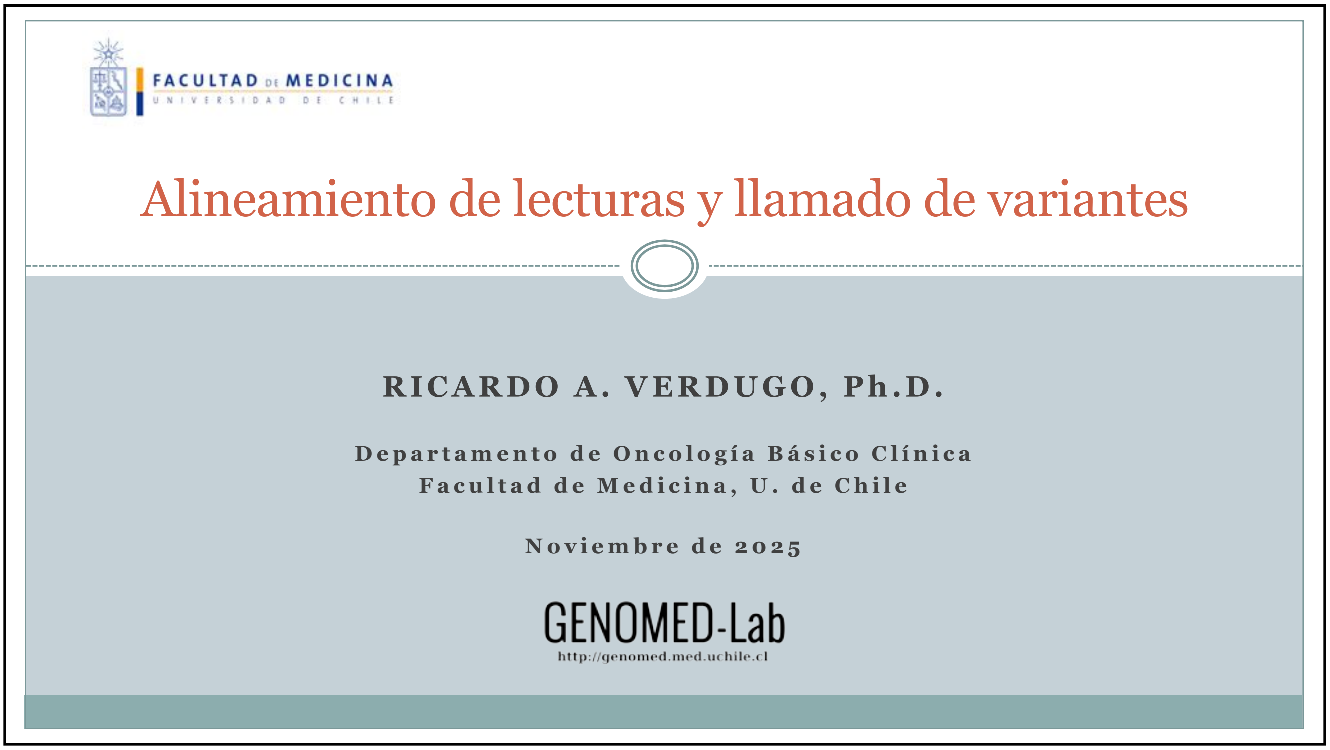


FACULTAD DE MEDICINA
UNIVERSIDAD DE CHILE

Alineamiento de lecturas y llamado de variantes

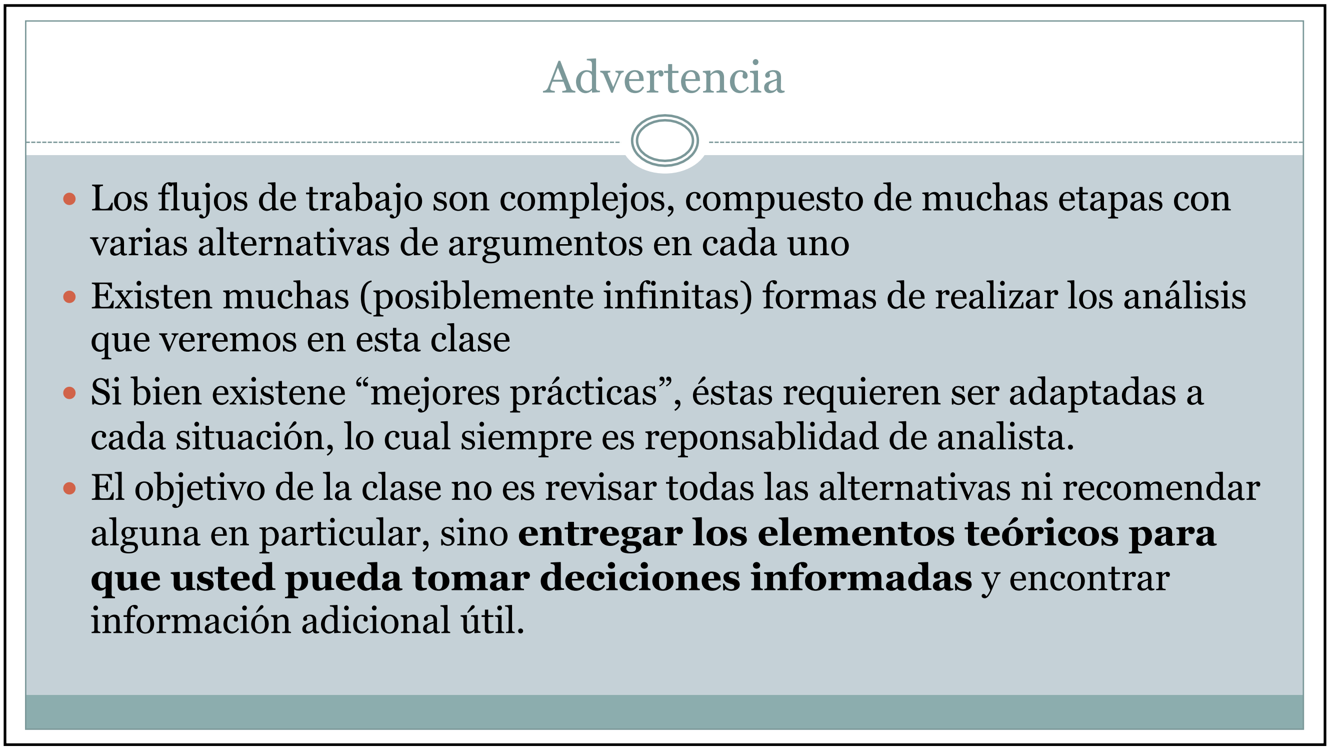
RICARDO A. VERDUGO, Ph.D.

Departamento de Oncología Básico Clínica
Facultad de Medicina, U. de Chile

Noviembre de 2025

GENOMED-Lab
http://genomed.med.uchile.cl

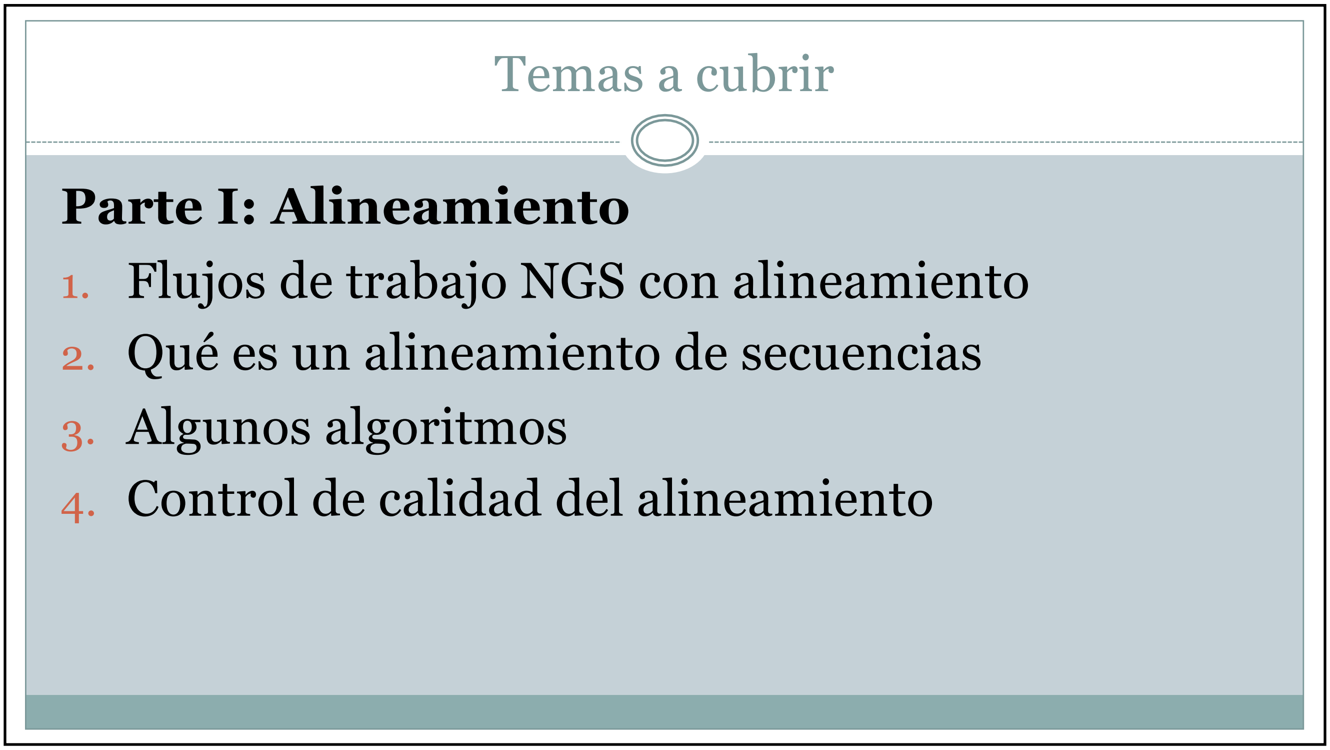
1



Advertencia

- Los flujos de trabajo son complejos, compuesto de muchas etapas con varias alternativas de argumentos en cada uno
- Existen muchas (posiblemente infinitas) formas de realizar los análisis que veremos en esta clase
- Si bien existene “mejores prácticas”, éstas requieren ser adaptadas a cada situación, lo cual siempre es reponsablidad de analista.
- El objetivo de la clase no es revisar todas las alternativas ni recomendar alguna en particular, sino **entregar los elementos teóricos para que usted pueda tomar deciciones informadas** y encontrar información adicional útil.

2

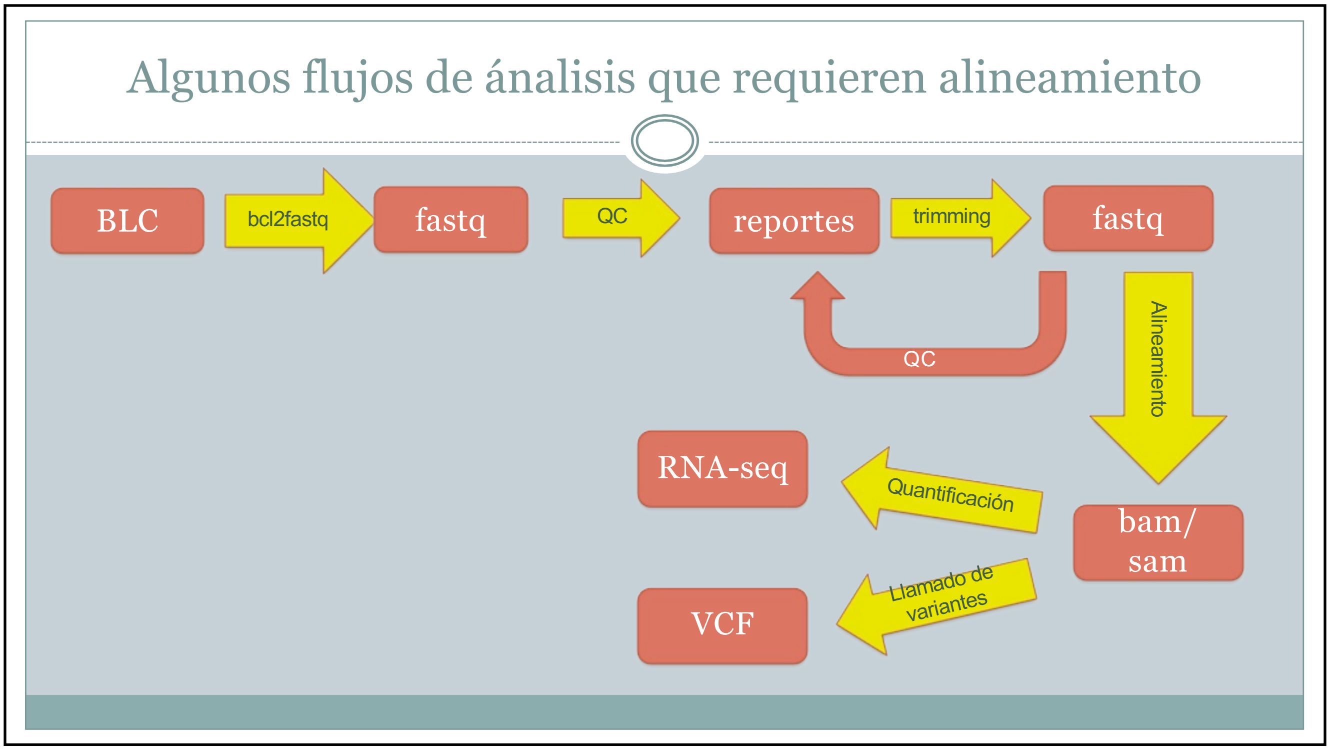


Temas a cubrir

Parte I: Alineamiento

1. Flujos de trabajo NGS con alineamiento
2. Qué es un alineamiento de secuencias
3. Algunos algoritmos
4. Control de calidad del alineamiento

3



Algunos flujos de ánalisis que requieren alineamiento

BLC

bcl2fastq

fastq

QC

reportes

trimming

fastq

QC

Alineamiento

bam/ sam

Quantificación

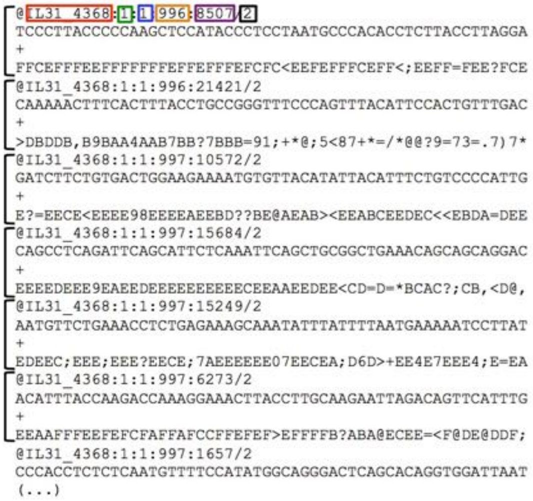
RNA-seq

Llamado de variantes

VCF

4

Raw sequencing data: Fastq format

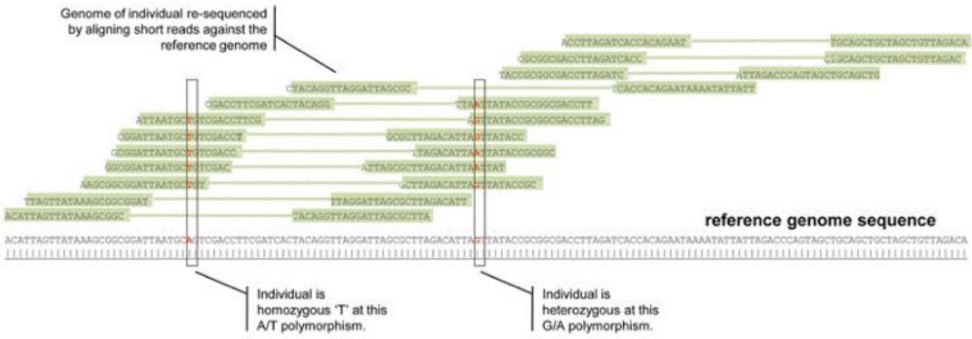


```
@IL31_4368:1:1:996:8507/2
TCCCTTACCCCCAAGCTCCATACCCTCCTAATGCCCACACCTCTTACCTTAGGA
+
FFCEFFFEEFFFFFFFFEFFEFFFEFCFC<EEFEFFFCEFF<;EEFF=FEE?FCE
@IL31_4368:1:1:996:21421/2
CAAAAACTTTCACTTTACCTGCCGGGTTTCCCAGTTTACATTCCACTGTTTGAC
+
>DBDDB,B9BAA4AAB7BB?7BBB=91;+*@;5<87+*=/*@@?9=73=.7)7*
@IL31_4368:1:1:997:10572/2
GATCTTCTGTGACTGGAAGAAAATGTGTTACATATTACATTTCTGTCCCCATTG
+
E?=EECE<EEEE98EEEEAEEBD??BE@AEAB><EEABCEEDEC<<EBDA=DEE
@IL31_4368:1:1:997:15684/2
CAGCCTCAGATTCAGCATTCTCAAATTCAGCTGCGGCTGAAACAGCAGCAGGAC
+
EEEEDEEE9EAEEDEEEEEEEEEEECEEAAEEDEE<CD=D=*BCAC?;CB,<D@,
@IL31_4368:1:1:997:15249/2
AATGTTCTGAAACCTCTGAGAAAGCAAATATTTATTTTAATGAAAAATCCTTAT
+
EDEEC;EEE;EEE?EECE;7AEEEEEE07EECEA;D6D>+EE4E7EEE4;E=EA
@IL31_4368:1:1:997:6273/2
ACATTTACCAAGACCAAAGGAAACTTACCTTGCAAGAATTAGACAGTTCATTTG
+
EEAAFFFEEFEFCFAFFAFCCFFEFEF>EFFFFB?ABA@ECEE=<F@DE@DDF;
@IL31_4368:1:1:997:1657/2
CCCACCTCTCTCAATGTTTTCCATATGGCAGGGACTCAGCACAGGTGGATTAAT
(...)
```

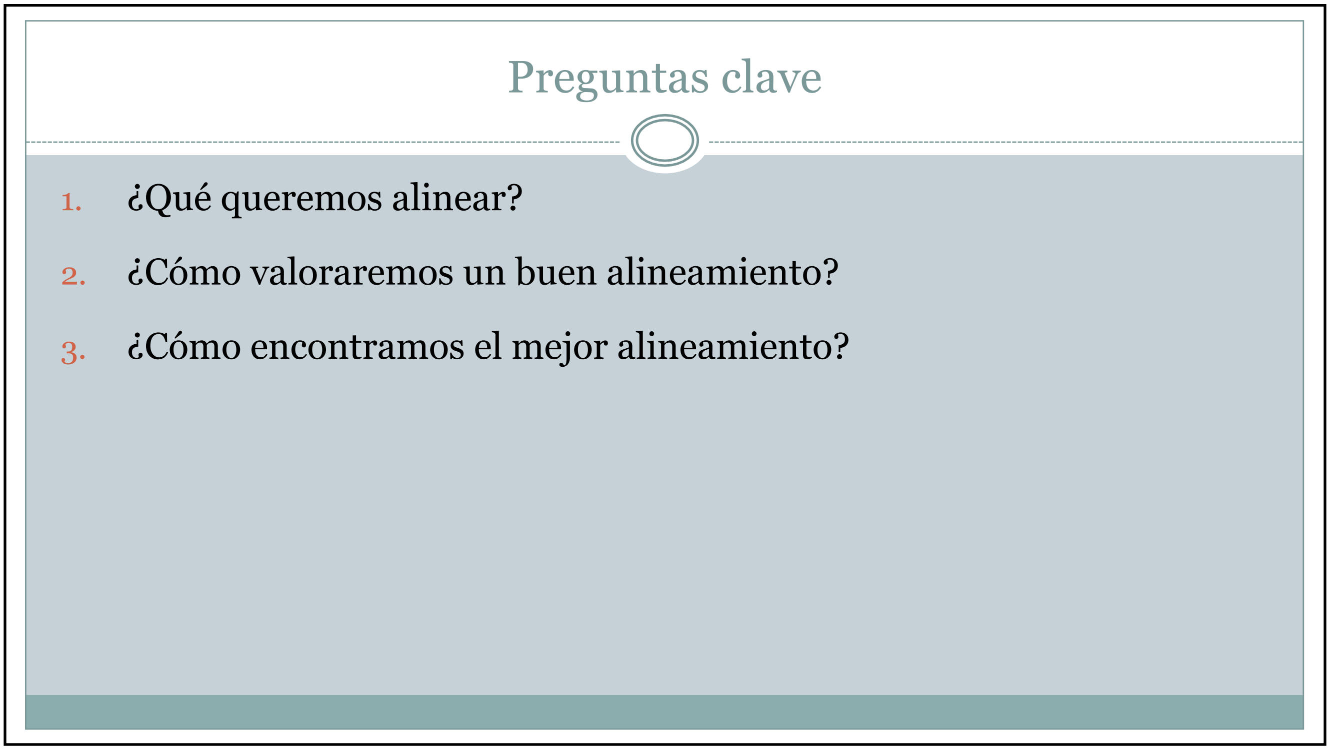
- 🟥 Instrument serial #
- 🟩 Lane
- 🟦 Swath
- 🟧 X coord
- 🟪 Y coord
- ⬛ Read direction

5

Alineamiento de lecturas de secuencias


Genome of individual re-sequenced by aligning short reads against the reference genome
reference genome sequence
ACATTAGTTATAAAGCGGCGGATTAATGCAGTCGACCTTCGATCACTACAGGTTAGGATTAGCGCTTAGACATTAGTTATACCGCGGCGACCTTAGATCACCACAGAATAAAATATTATTAGACCCAGTAGCTGCAGCTGCTAGCTGTTAGACA
Individual is homozygous 'T' at this A/T polymorphism.
Individual is heterozygous at this G/A polymorphism.

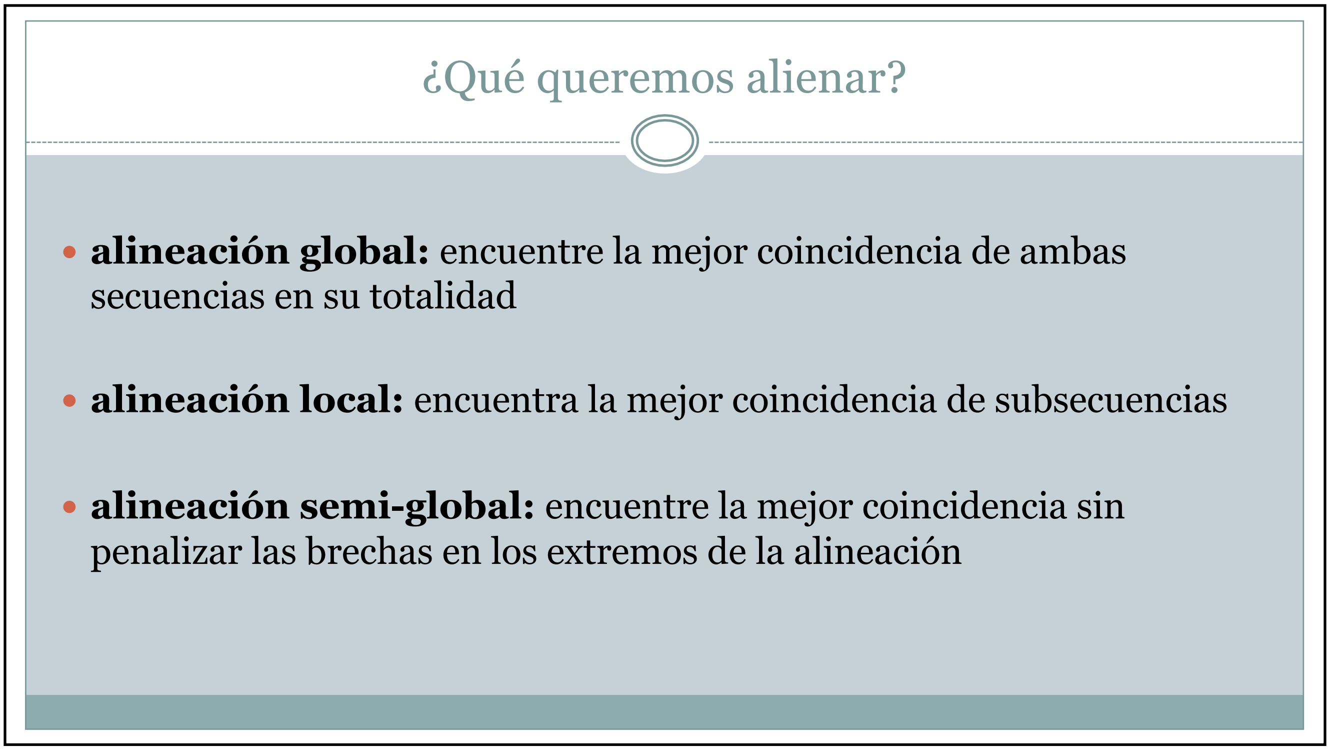
6



Preguntas clave

1. ¿Qué queremos alinear?
2. ¿Cómo valoraremos un buen alineamiento?
3. ¿Cómo encontramos el mejor alineamiento?

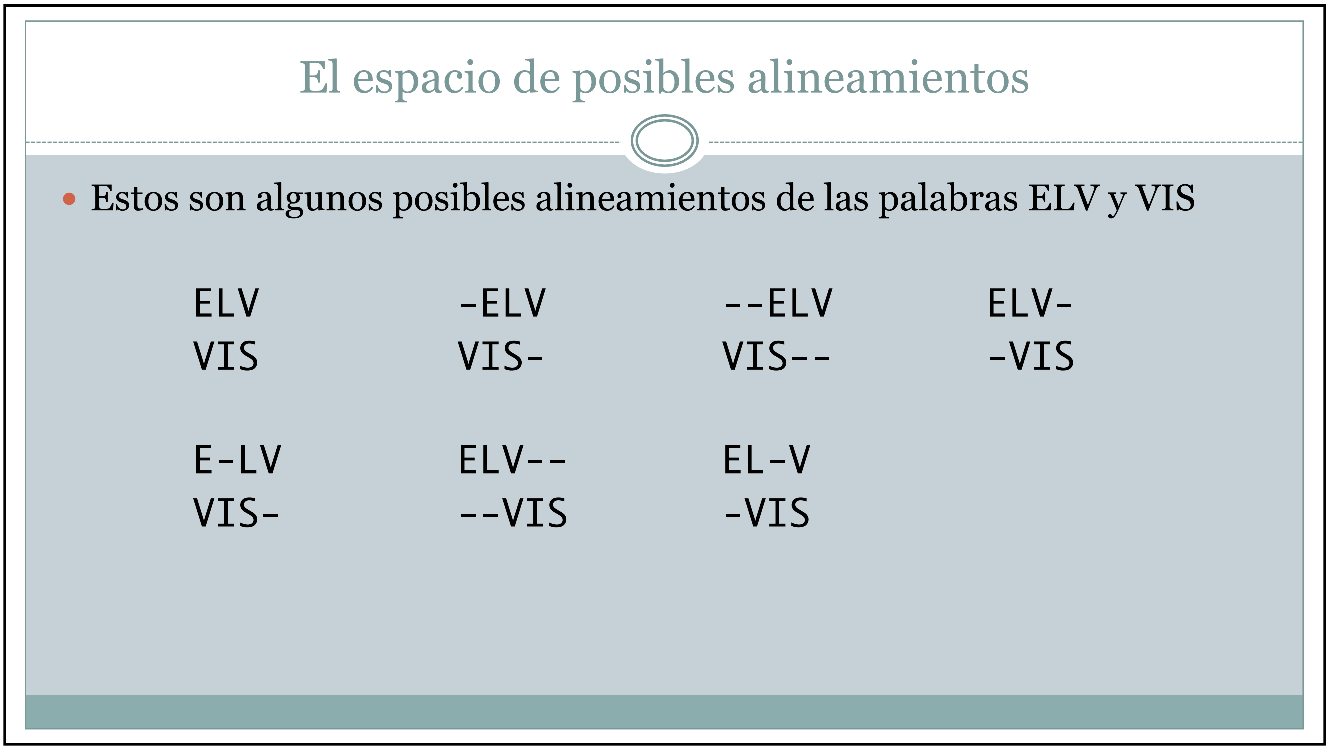
7



¿Qué queremos alienar?

- **alineación global:** encuentre la mejor coincidencia de ambas secuencias en su totalidad
- **alineación local:** encuentra la mejor coincidencia de subsecuencias
- **alineación semi-global:** encuentre la mejor coincidencia sin penalizar las brechas en los extremos de la alineación

8

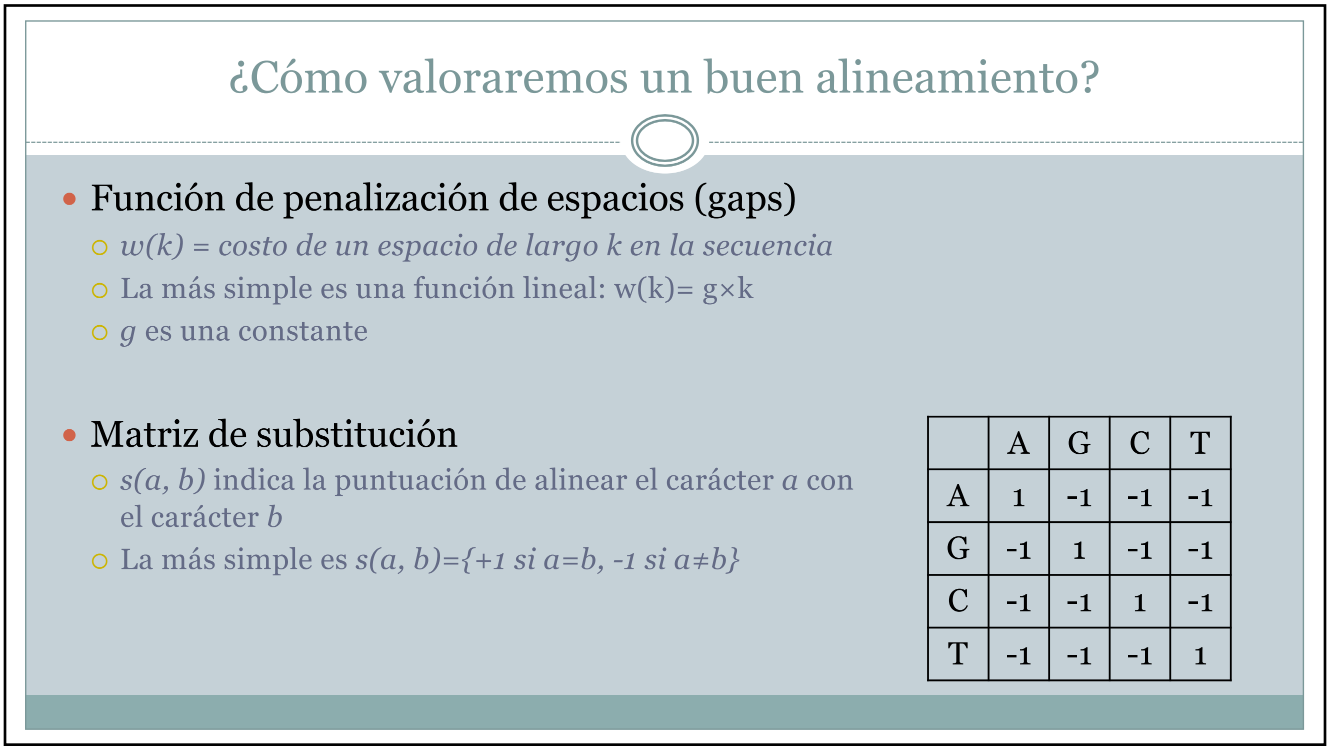


El espacio de posibles alineamientos

- Estos son algunos posibles alineamientos de las palabras ELV y VIS

```
ELV          -ELV         --ELV        ELV-
VIS          VIS-         VIS--        -VIS

E-LV         ELV--        EL-V
VIS-         --VIS        -VIS
```

9

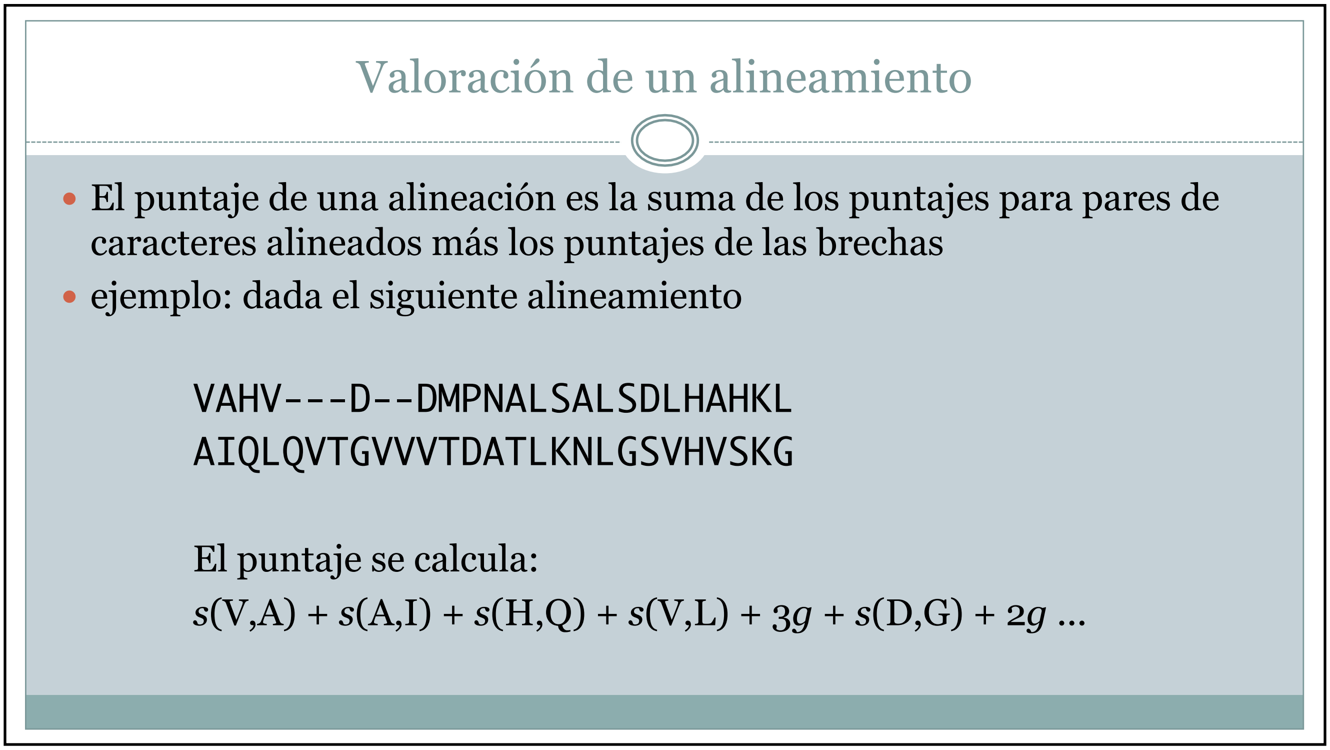


¿Cómo valoraremos un buen alineamiento?

- Función de penalización de espacios (gaps)
  - *w(k) = costo de un espacio de largo k en la secuencia*
  - La más simple es una función lineal: w(k)= g×k
  - *g* es una constante
- Matriz de substitución
  - *s(a, b)* indica la puntuación de alinear el carácter *a* con el carácter *b*
  - La más simple es *s(a, b)={+1 si a=b, -1 si a≠b}*

	A	G	C	T
A	1	-1	-1	-1
G	-1	1	-1	-1
C	-1	-1	1	-1
T	-1	-1	-1	1

10



Valoración de un alineamiento

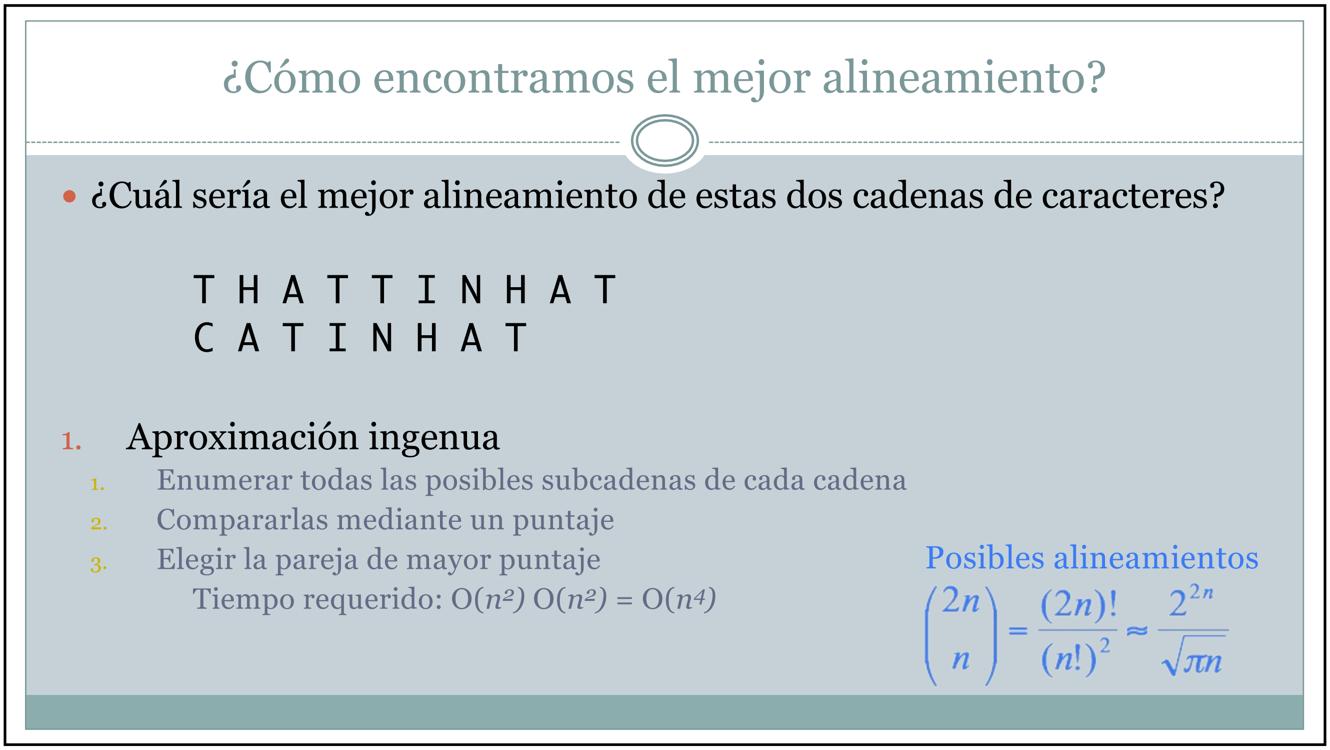
- El puntaje de una alineación es la suma de los puntajes para pares de caracteres alineados más los puntajes de las brechas
- ejemplo: dada el siguiente alineamiento

```
VAHV---D--DMPNALSALSDLHAHKL
AIQLQVTGVVVTDATLKNLGSVHVSKG
```

El puntaje se calcula:

$s(\mathrm{V,A}) + s(\mathrm{A,I}) + s(\mathrm{H,Q}) + s(\mathrm{V,L}) + 3g + s(\mathrm{D,G}) + 2g \ldots$

11



¿Cómo encontramos el mejor alineamiento?

- ¿Cuál sería el mejor alineamiento de estas dos cadenas de caracteres?

```
T H A T T I N H A T
C A T I N H A T
```

1. Aproximación ingenua
   1. Enumerar todas las posibles subcadenas de cada cadena
   2. Compararlas mediante un puntaje
   3. Elegir la pareja de mayor puntaje

      Tiempo requerido: $O(n^2)\, O(n^2) = O(n^4)$

Posibles alineamientos

$$\binom{2n}{n} = \frac{(2n)!}{(n!)^2} \approx \frac{2^{2n}}{\sqrt{\pi n}}$$

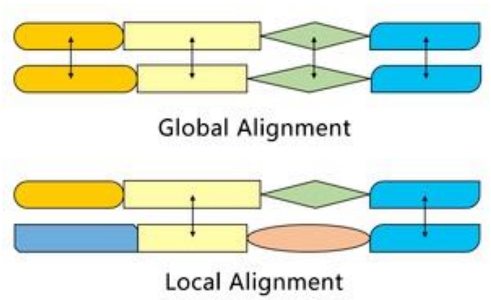
12

Solución por programación dinámica

- **Programación dinámica**: resuelva una instancia de un problema aprovechando las soluciones para subpartes del problema
  - reducir el problema de la mejor alineación de dos secuencias a la mejor alineación de todos los prefijos de las secuencias
  - Evitar recalcular las puntuaciones ya consideradas.
- Utilizado por primera vez en alineación por Needleman & Wunsch, Journal of Molecular Biology, 1970

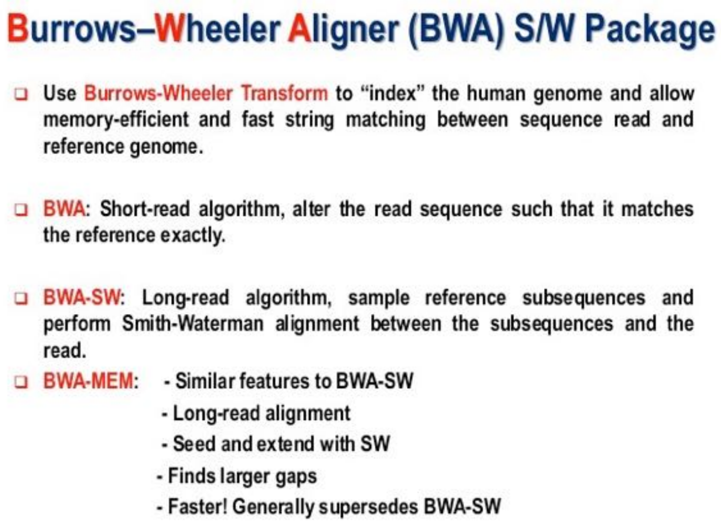
13

Alineamiento Global vs Local



14

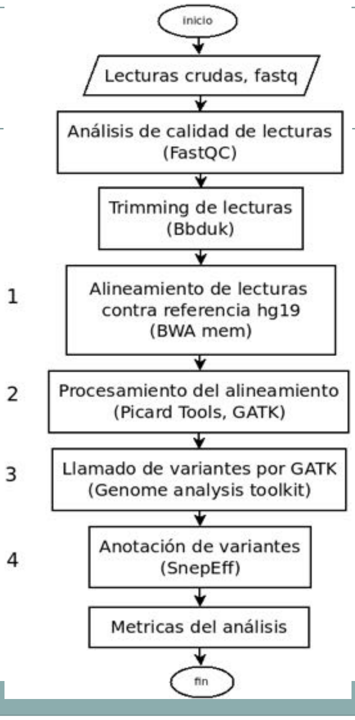
Alineador BWA



Burrows–Wheeler Aligner (BWA) S/W Package

- ❑ Use Burrows-Wheeler Transform to "index" the human genome and allow memory-efficient and fast string matching between sequence read and reference genome.
- ❑ BWA: Short-read algorithm, alter the read sequence such that it matches the reference exactly.
- ❑ BWA-SW: Long-read algorithm, sample reference subsequences and perform Smith-Waterman alignment between the subsequences and the read.
- ❑ BWA-MEM:
  - Similar features to BWA-SW
  - Long-read alignment
  - Seed and extend with SW
  - Finds larger gaps
  - Faster! Generally supersedes BWA-SW

15



inicio

Lecturas crudas, fastq

Análisis de calidad de lecturas (FastQC)

Trimming de lecturas (Bbduk)

1 Alineamiento de lecturas contra referencia hg19 (BWA mem)

2 Procesamiento del alineamiento (Picard Tools, GATK)

3 Llamado de variantes por GATK (Genome analysis toolkit)

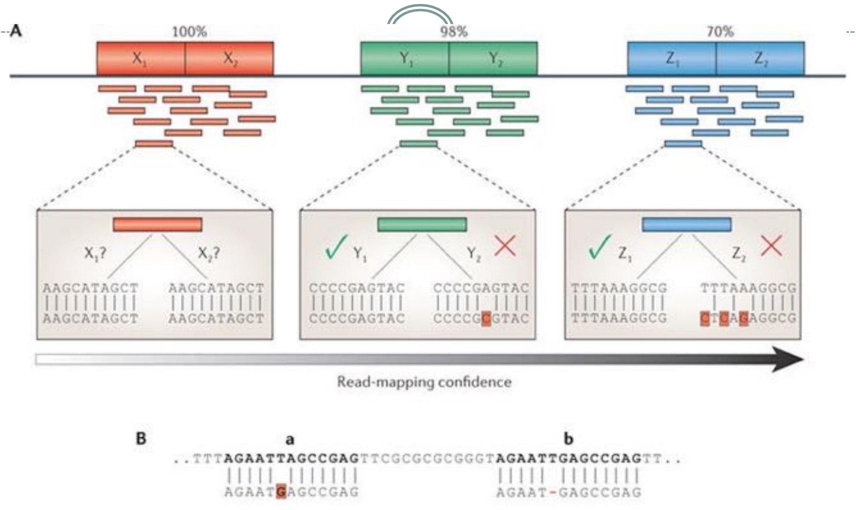
4 Anotación de variantes (SnepEff)

Metricas del análisis

fin

16

Alineamiento en Regiones Repetidas

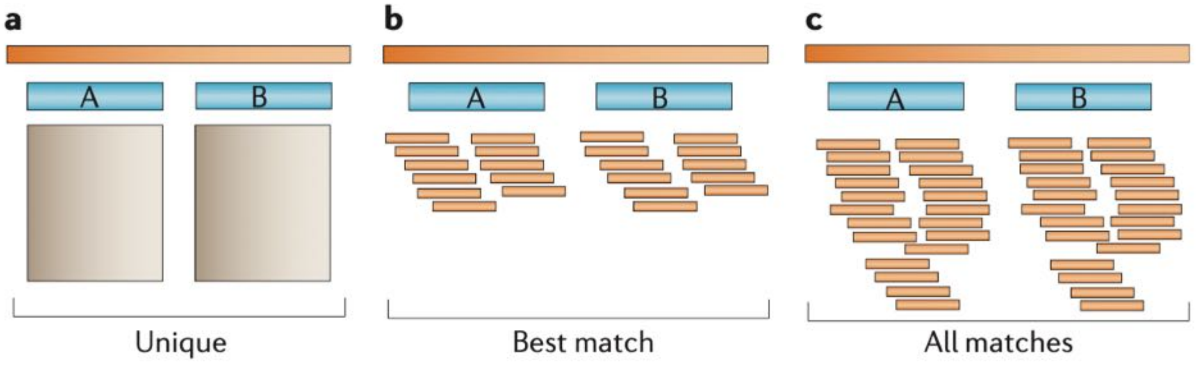


Nature Reviews | Genetics

Nature Reviews Genetics 13, 36-46 (January 2012)
doi:10.1038/nrg3117

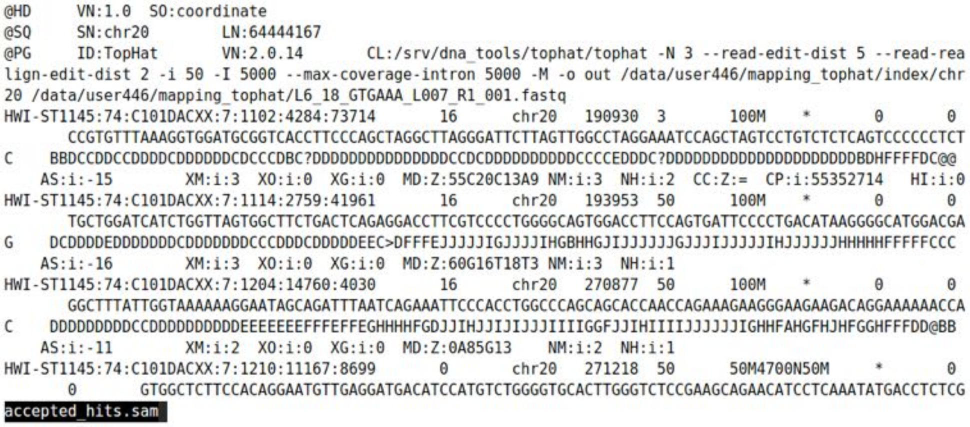
17

Estrategias de reporte para lecturas con más de un mapeo (multi-reads)



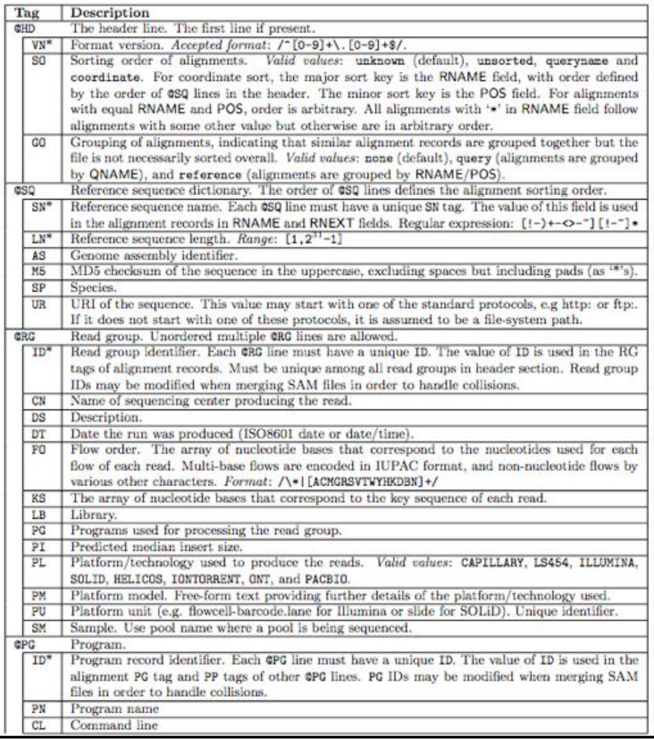
18

Formato Sequence Alignment Map (SAM)



```
@HD     VN:1.0  SO:coordinate
@SQ     SN:chr20        LN:64444167
@PG     ID:TopHat       VN:2.0.14       CL:/srv/dna_tools/tophat/tophat -N 3 --read-edit-dist 5 --read-rea
lign-edit-dist 2 -i 50 -I 5000 --max-coverage-intron 5000 -M -o out /data/user446/mapping_tophat/index/chr
20 /data/user446/mapping_tophat/L6_18_GTGAAA_L007_R1_001.fastq
HWI-ST1145:74:C101DACXX:7:1102:4284:73714       16      chr20   190930  3       100M    *       0       0
        CCGTGTTTAAAGGTGGATGCGGTCACCTTCCCAGCTAGGCTTAGGGATTCTTAGTTGGCCTAGGAAATCCAGCTAGTCCTGTCTCTCAGTCCCCCTCT
C    BBDCCDDCCDDDDCDDDDDDCDCCCDBC?DDDDDDDDDDDDDDDCCDCDDDDDDDDDDCCCCEDDDC?DDDDDDDDDDDDDDDDDDDDDBDHFFFFDC@@
    AS:i:-15        XM:i:3  XO:i:0  XG:i:0  MD:Z:55C20C13A9 NM:i:3  NH:i:2  CC:Z:=  CP:i:55352714   HI:i:0
HWI-ST1145:74:C101DACXX:7:1114:2759:41961       16      chr20   193953  50      100M    *       0       0
        TGCTGGATCATCTGGTTAGTGGCTTCTGACTCAGAGGACCTTCGTCCCCTGGGGCAGTGGACCTTCCAGTGATTCCCCTGACATAAGGGGCATGGACGA
G    DCDDDDEDDDDDDDCDDDDDDDCCCDDDCDDDDDEEC>DFFFEJJJJJIGJJJJIHGBHHGJIJJJJJJGJJJIJJJJJIHJJJJJJHHHHHFFFFFCCC
    AS:i:-16        XM:i:3  XO:i:0  XG:i:0  MD:Z:60G16T18T3 NM:i:3  NH:i:1
HWI-ST1145:74:C101DACXX:7:1204:14760:4030       16      chr20   270877  50      100M    *       0       0
        GGCTTTATTGGTAAAAAAGGAATAGCAGATTTAATCAGAAATTCCCACCTGGCCCAGCAGCACCAACCAGAAAGAAGGGAAGAAGACAGGAAAAAACCA
C    DDDDDDDDDCCDDDDDDDDDDDEEEEEEEFFFEFFEGHHHHFGDJJIHJJIJIJJJIIIIGGFJJIHIIIIJJJJJJIGHHFAHGFHJHFGGHFFFDD@BB
    AS:i:-11        XM:i:2  XO:i:0  XG:i:0  MD:Z:0A85G13    NM:i:2  NH:i:1
HWI-ST1145:74:C101DACXX:7:1210:11167:8699       0       chr20   271218  50      50M4700N50M     *       0
        0       GTGGCTCTTCCACAGGAATGTTGAGGATGACATCCATGTCTGGGGTGCACTTGGGTCTCCGAAGCAGAACATCCTCAAATATGACCTCTCG
accepted_hits.sam
```

19

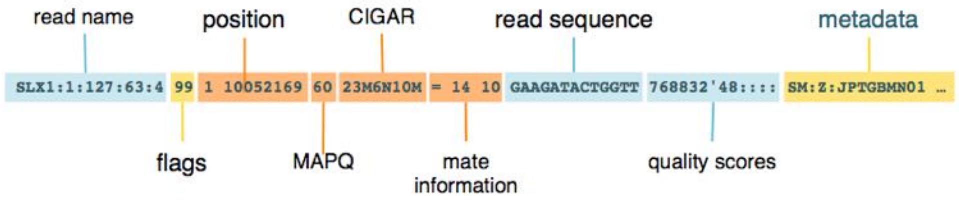


Tag	Description
@HD	The header line. The first line if present.
VN*	Format version. *Accepted format*: /^[0-9]+\.[0-9]+$/.
SO	Sorting order of alignments. *Valid values*: **unknown** (default), **unsorted**, **queryname** and **coordinate**. For coordinate sort, the major sort key is the RNAME field, with order defined by the order of @SQ lines in the header. The minor sort key is the POS field. For alignments with equal RNAME and POS, order is arbitrary. All alignments with '*' in RNAME field follow alignments with some other value but otherwise are in arbitrary order.
GO	Grouping of alignments, indicating that similar alignment records are grouped together but the file is not necessarily sorted overall. *Valid values*: **none** (default), **query** (alignments are grouped by QNAME), and **reference** (alignments are grouped by RNAME/POS).
@SQ	Reference sequence dictionary. The order of @SQ lines defines the alignment sorting order.
SN*	Reference sequence name. Each @SQ line must have a unique SN tag. The value of this field is used in the alignment records in RNAME and RNEXT fields. Regular expression: [!-)+-<>-~][!-~]*
LN*	Reference sequence length. *Range*: $[1,2^{31}-1]$
AS	Genome assembly identifier.
M5	MD5 checksum of the sequence in the uppercase, excluding spaces but including pads (as '*'s).
SP	Species.
UR	URI of the sequence. This value may start with one of the standard protocols, e.g http: or ftp:. If it does not start with one of these protocols, it is assumed to be a file-system path.
@RG	Read group. Unordered multiple @RG lines are allowed.
ID*	Read group identifier. Each @RG line must have a unique ID. The value of ID is used in the RG tags of alignment records. Must be unique among all read groups in header section. Read group IDs may be modified when merging SAM files in order to handle collisions.
CN	Name of sequencing center producing the read.
DS	Description.
DT	Date the run was produced (ISO8601 date or date/time).
FO	Flow order. The array of nucleotide bases that correspond to the nucleotides used for each flow of each read. Multi-base flows are encoded in IUPAC format, and non-nucleotide flows by various other characters. *Format*: /*|[ACMGRSVTWYHKDBN]+/
KS	The array of nucleotide bases that correspond to the key sequence of each read.
LB	Library.
PG	Programs used for processing the read group.
PI	Predicted median insert size.
PL	Platform/technology used to produce the reads. *Valid values*: **CAPILLARY**, **LS454**, **ILLUMINA**, **SOLID**, **HELICOS**, **IONTORRENT**, **ONT**, and **PACBIO**.
PM	Platform model. Free-form text providing further details of the platform/technology used.
PU	Platform unit (e.g. flowcell-barcode.lane for Illumina or slide for SOLiD). Unique identifier.
SM	Sample. Use pool name where a pool is being sequenced.
@PG	Program.
ID*	Program record identifier. Each @PG line must have a unique ID. The value of ID is used in the alignment PG tag and PP tags of other @PG lines. PG IDs may be modified when merging SAM files in order to handle collisions.
PN	Program name
CL	Command line

20

Archivo SAM: registros

HEADER containing metadata (sequence dictionary, read group definitions etc)
RECORDS containing structured read information (1 line per read record)



21

CIGAR

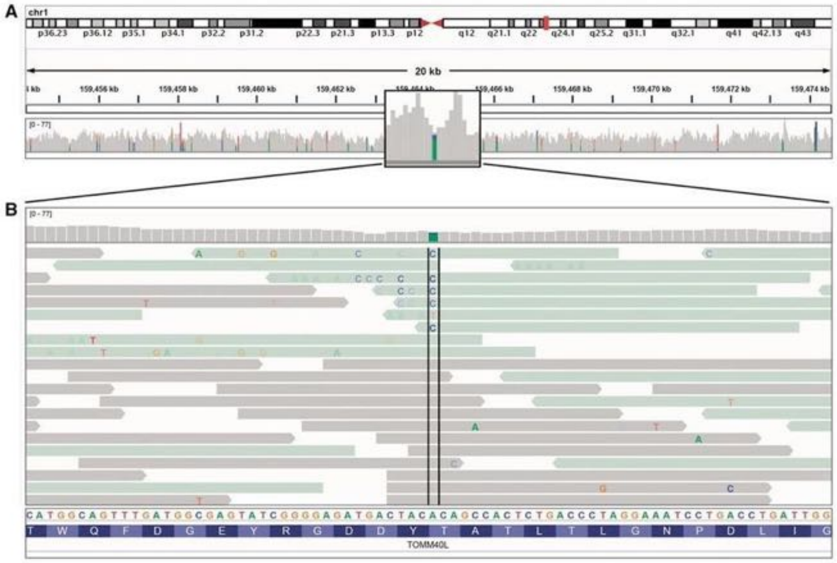
```
RefPos:      1  2  3  4  5  6  7  8  9 10 11 12 13 14 15 16 17 18 19
Reference:   C  C  A  T  A  C  T  G  A  A  C  T  G  A  C  T  A  A  C
Read:        ACTAGAATGGCT

Aligning these two:
RefPos:      1  2  3  4  5  6  7     8  9 10 11 12 13 14 15 16 17 18 19
Reference:   C  C  A  T  A  C  T     G  A  A  C  T  G  A  C  T  A  A  C
Read:                    A  C  T  A  G  A  A     T  G  G  C  T

With the alignment above, you get:
POS: 5 CIGAR: 3M1I3M1D5M
```

Cada operación está representada por un número seguido de una letra que especifica el tipo de evento:

- M representa coincidencias o diferencias,
- I inserciones en la lectura,
- D deleciones respecto a la referencia,
- N saltos (por ejemplo, intrones en datos de ARN-seq),
- S secuencias suavizadas (soft clipping),
- H secuencias recortadas (hard clipping).

22

Integrative Genomic Viewer


A
chr1
p36.23 p36.12 p35.1 p34.1 p32.2 p31.2 p22.3 p21.3 p13.3 p12 q12 q21.1 q22 q24.1 q25.2 q31.1 q32.1 q41 q42.13 q43
20 kb
159,456 kb 159,458 kb 159,460 kb 159,462 kb 159,464 kb 159,466 kb 159,468 kb 159,470 kb 159,472 kb 159,474 kb
[0 - 77]
B
[0 - 77]
CATGGCAGTTTGATGGCGAGTATCGGGGAGATGACTACACAGCCACTCTGACCCTAGGAAATCCTGACCTGATTGG
T W Q F D G E Y R G D D Y T A T L T L G N P D L I G
TOMM40L

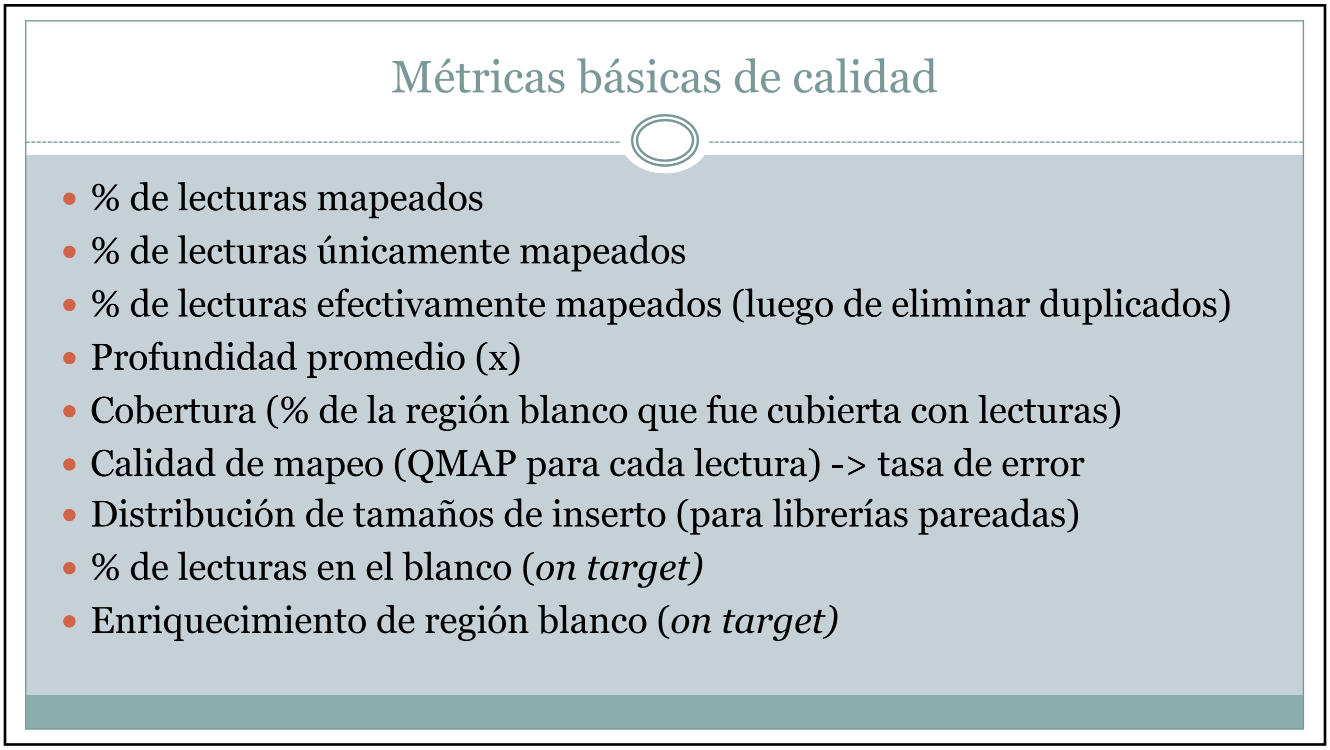
http://software.broadinstitute.org/software/igv/

23

¿Qué mirar en una visualización de alineamiento?

- Cobertura en la región de interés
- Posible sesgo de variantes entre R1 y R2 o por hebra (strand bias)
- Variantes que estén en los extremos
- INDELs en los extremos
- Sustituciones alrededor de los INDELs (realizar realineamiento local)

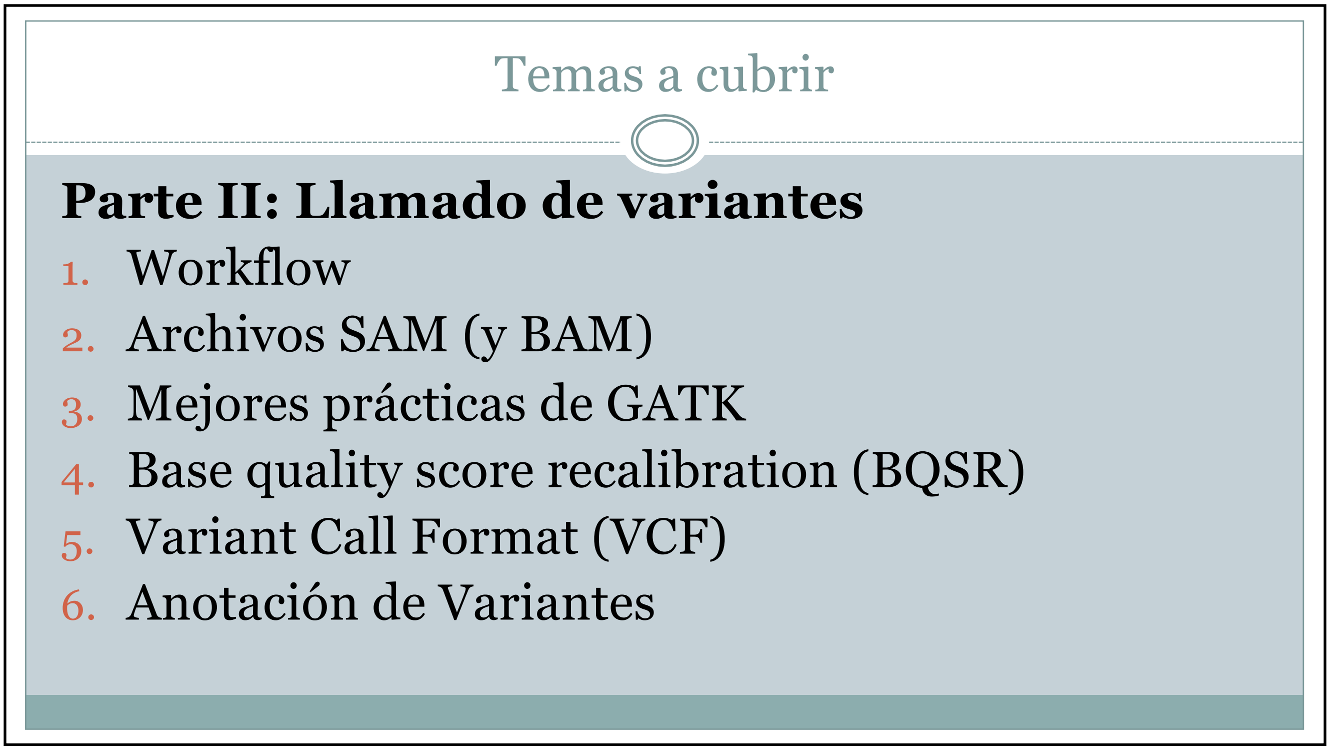
24



Métricas básicas de calidad

- % de lecturas mapeados
- % de lecturas únicamente mapeados
- % de lecturas efectivamente mapeados (luego de eliminar duplicados)
- Profundidad promedio (x)
- Cobertura (% de la región blanco que fue cubierta con lecturas)
- Calidad de mapeo (QMAP para cada lectura) -> tasa de error
- Distribución de tamaños de inserto (para librerías pareadas)
- % de lecturas en el blanco (*on target)*
- Enriquecimiento de región blanco (*on target)*

25



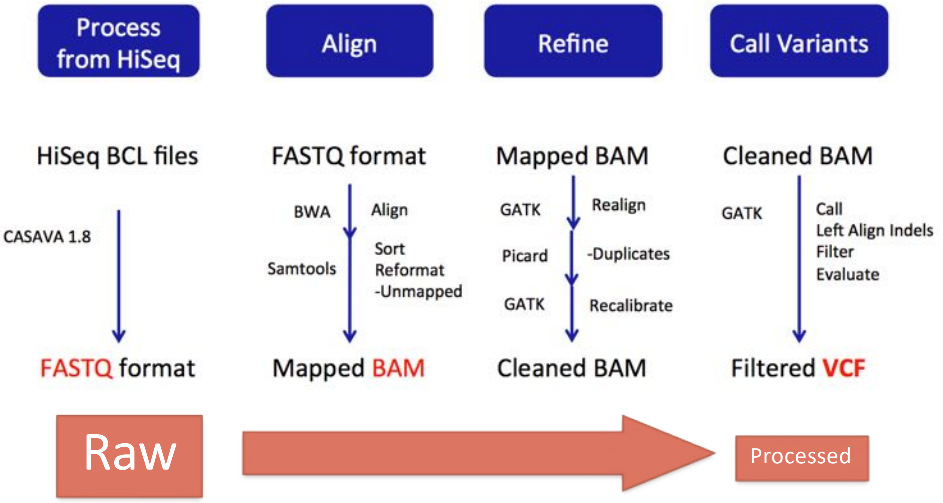
Temas a cubrir

Parte II: Llamado de variantes

1. Workflow
2. Archivos SAM (y BAM)
3. Mejores prácticas de GATK
4. Base quality score recalibration (BQSR)
5. Variant Call Format (VCF)
6. Anotación de Variantes

26

NGS data files



Process from HiSeq

HiSeq BCL files

CASAVA 1.8

FASTQ format

Align

FASTQ format

BWA — Align

Samtools — Sort, Reformat, -Unmapped

Mapped BAM

Refine

Mapped BAM

GATK — Realign

Picard — -Duplicates

GATK — Recalibrate

Cleaned BAM

Call Variants

Cleaned BAM

GATK — Call, Left Align Indels, Filter, Evaluate

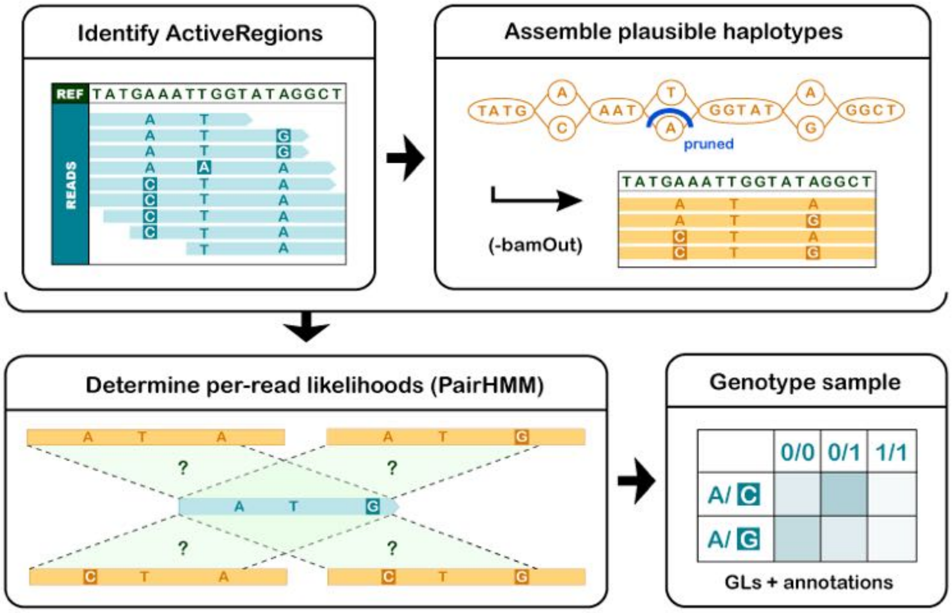
Filtered VCF

Raw

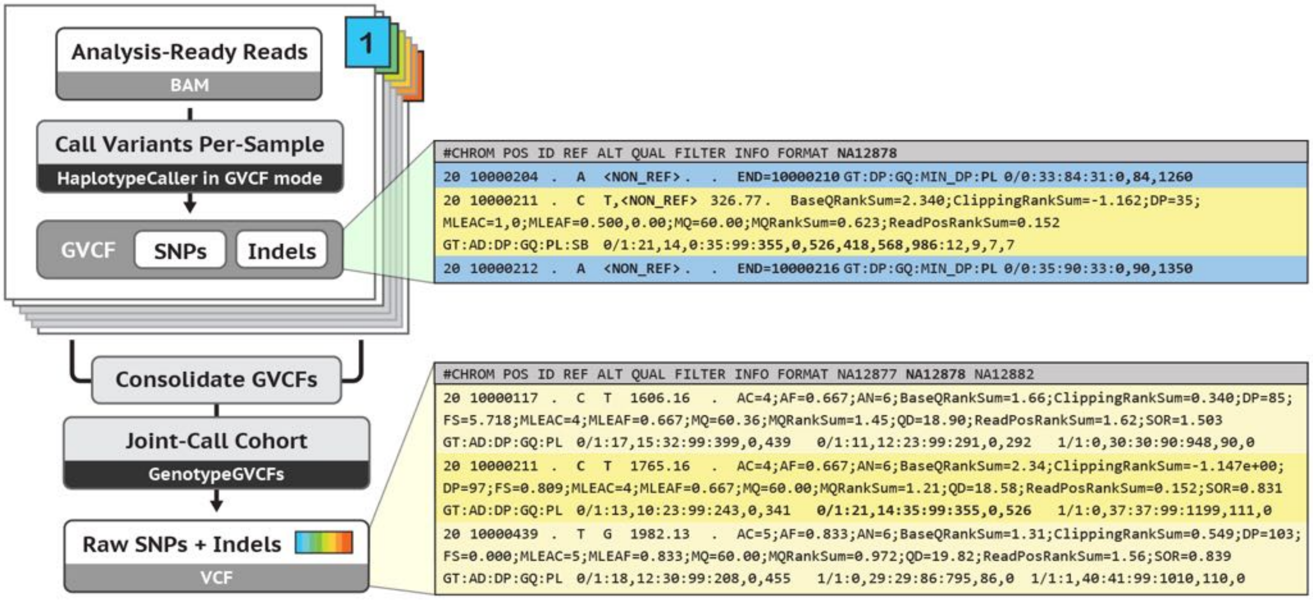
Processed

27

Llamado de Variantes: GATK HaplotypeCaller



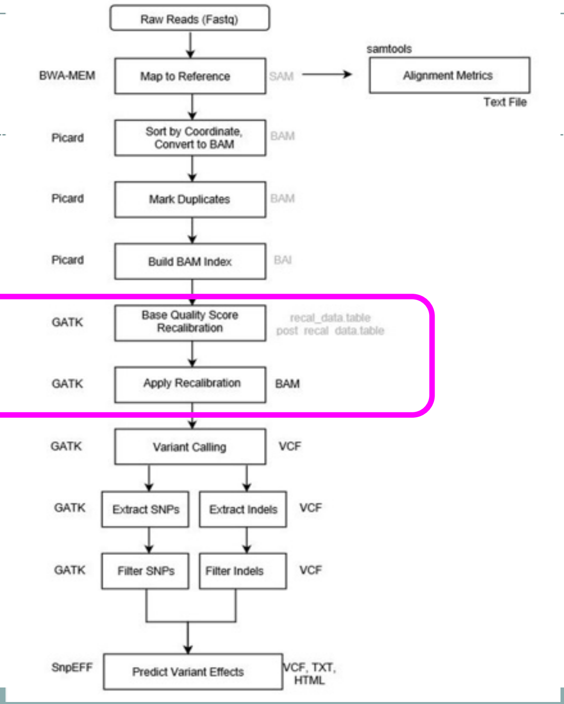
28



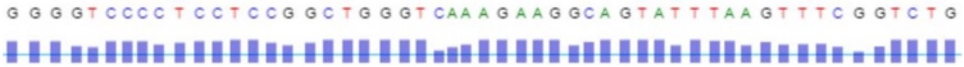
```
#CHROM POS ID REF ALT QUAL FILTER INFO FORMAT NA12878
20 10000204  .   A   <NON_REF> .   .   END=10000210 GT:DP:GQ:MIN_DP:PL 0/0:33:84:31:0,84,1260
20 10000211  .   C   T,<NON_REF> 326.77 .   BaseQRankSum=2.340;ClippingRankSum=-1.162;DP=35;
MLEAC=1,0;MLEAF=0.500,0.00;MQ=60.00;MQRankSum=0.623;ReadPosRankSum=0.152
GT:AD:DP:GQ:PL:SB  0/1:21,14,0:35:99:355,0,526,418,568,986:12,9,7,7
20 10000212  .   A   <NON_REF> .   .   END=10000216 GT:DP:GQ:MIN_DP:PL 0/0:35:90:33:0,90,1350
```

```
#CHROM POS ID REF ALT QUAL FILTER INFO FORMAT NA12877 NA12878 NA12882
20 10000117  .   C   T   1606.16   .   AC=4;AF=0.667;AN=6;BaseQRankSum=1.66;ClippingRankSum=0.340;DP=85;
FS=5.718;MLEAC=4;MLEAF=0.667;MQ=60.36;MQRankSum=1.45;QD=18.90;ReadPosRankSum=1.62;SOR=1.503
GT:AD:DP:GQ:PL  0/1:17,15:32:99:399,0,439   0/1:11,12:23:99:291,0,292   1/1:0,30:30:90:948,90,0
20 10000211  .   C   T   1765.16   .   AC=4;AF=0.667;AN=6;BaseQRankSum=2.34;ClippingRankSum=-1.147e+00;
DP=97;FS=0.809;MLEAC=4;MLEAF=0.667;MQ=60.00;MQRankSum=1.21;QD=18.58;ReadPosRankSum=0.152;SOR=0.831
GT:AD:DP:GQ:PL  0/1:13,10:23:99:243,0,341   0/1:21,14:35:99:355,0,526   1/1:0,37:37:99:1199,111,0
20 10000439  .   T   G   1982.13   .   AC=5;AF=0.833;AN=6;BaseQRankSum=1.31;ClippingRankSum=0.549;DP=103;
FS=0.000;MLEAC=5;MLEAF=0.833;MQ=60.00;MQRankSum=0.972;QD=19.82;ReadPosRankSum=1.56;SOR=0.839
GT:AD:DP:GQ:PL  0/1:18,12:30:99:208,0,455   1/1:0,29:29:86:795,86,0  1/1:1,40:41:99:1010,110,0
```

29



30



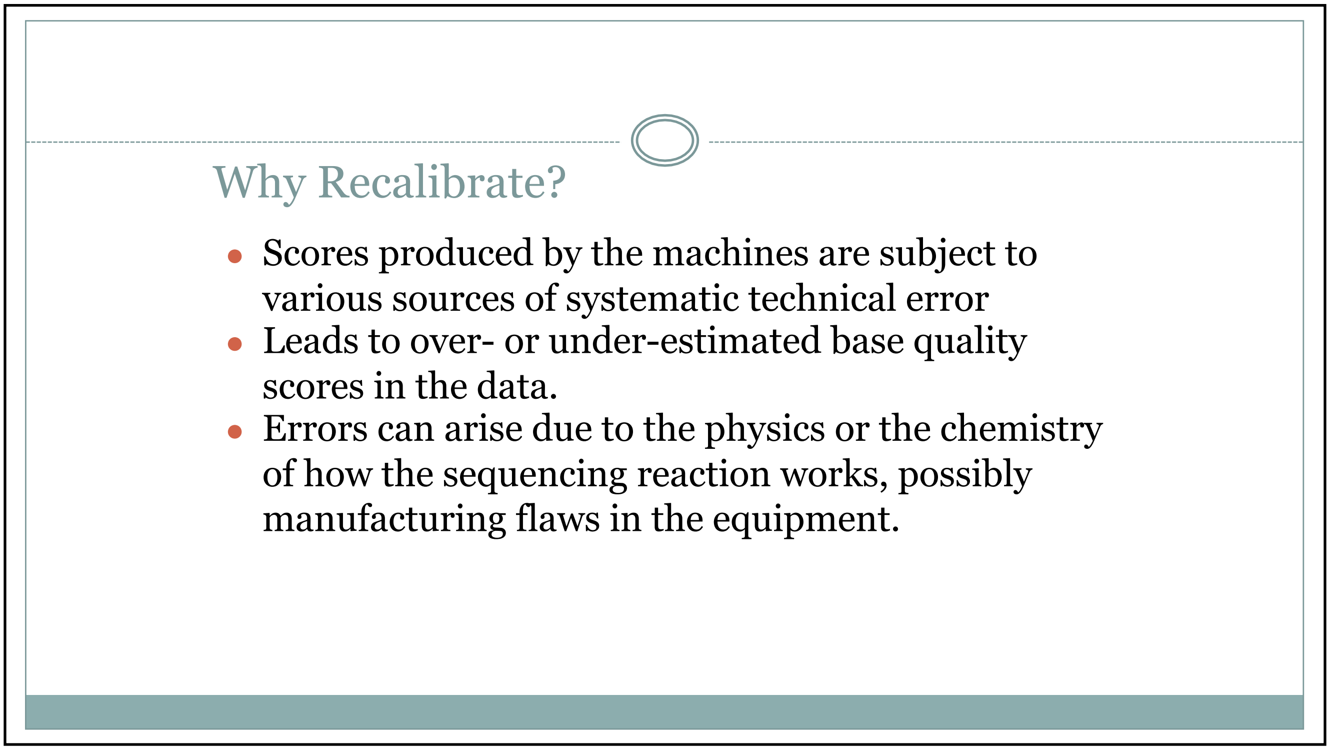
Individual quality scores (blue bars) for each position in a single read. The horizontal blue line represents the Q20 phred score value

31

Why do we care about quality scores so much?

- Variant calling algorithms rely on the quality score assigned to the individual base calls
- Tells us how much we can trust that particular observation to inform us about the biological truth of the site
- If we have a basecall that has a low quality score, that means we're not sure we actually read that A correctly, and it could actually be something else
- So we won't trust it as much as other base calls that have higher qualities
- We use that score to weigh the evidence that we have for or against a variant existing at a particular site

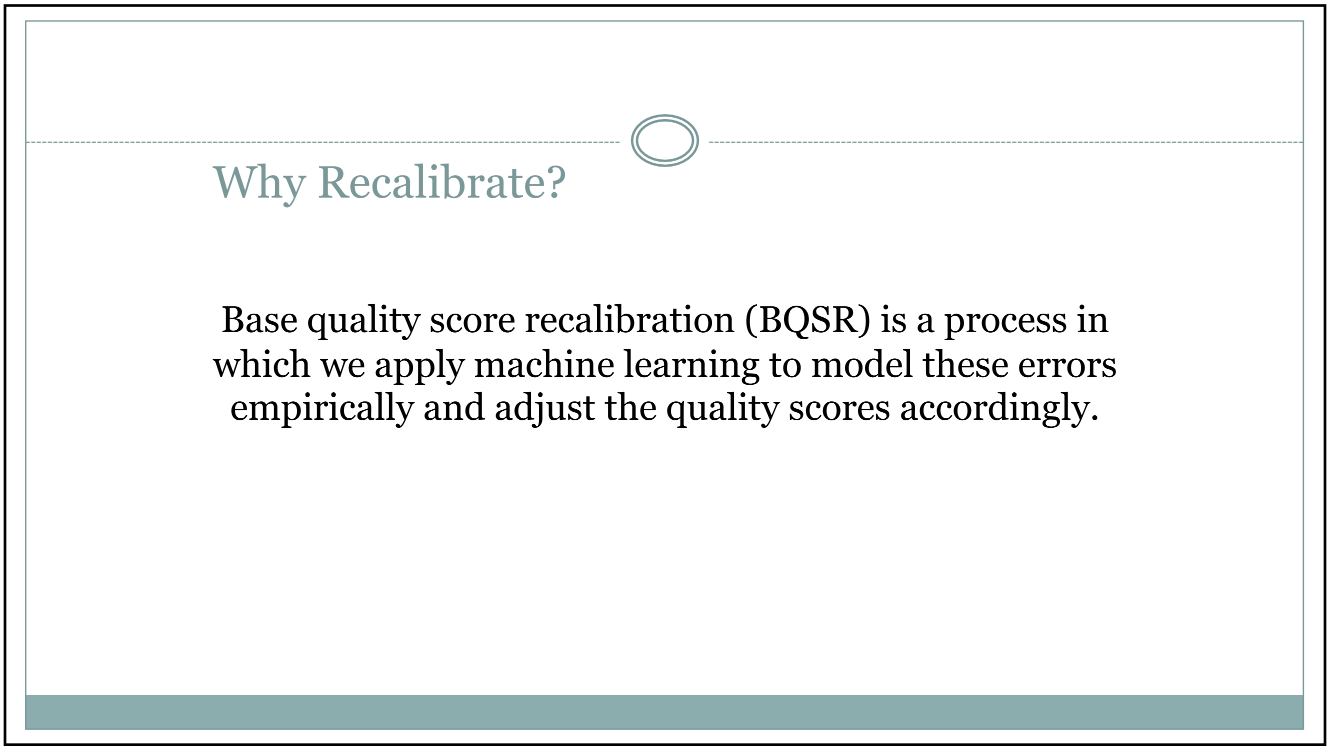
32



Why Recalibrate?

- Scores produced by the machines are subject to various sources of systematic technical error
- Leads to over- or under-estimated base quality scores in the data.
- Errors can arise due to the physics or the chemistry of how the sequencing reaction works, possibly manufacturing flaws in the equipment.

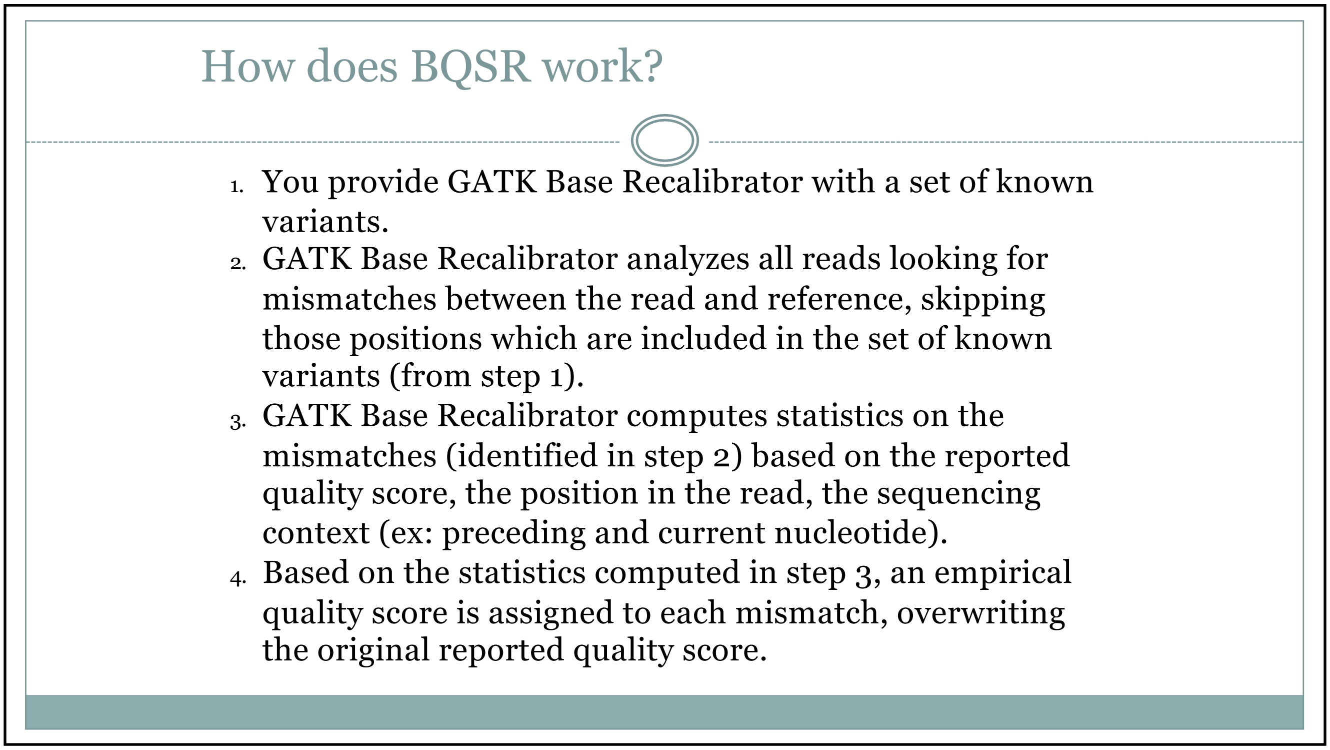
33



Why Recalibrate?

Base quality score recalibration (BQSR) is a process in which we apply machine learning to model these errors empirically and adjust the quality scores accordingly.

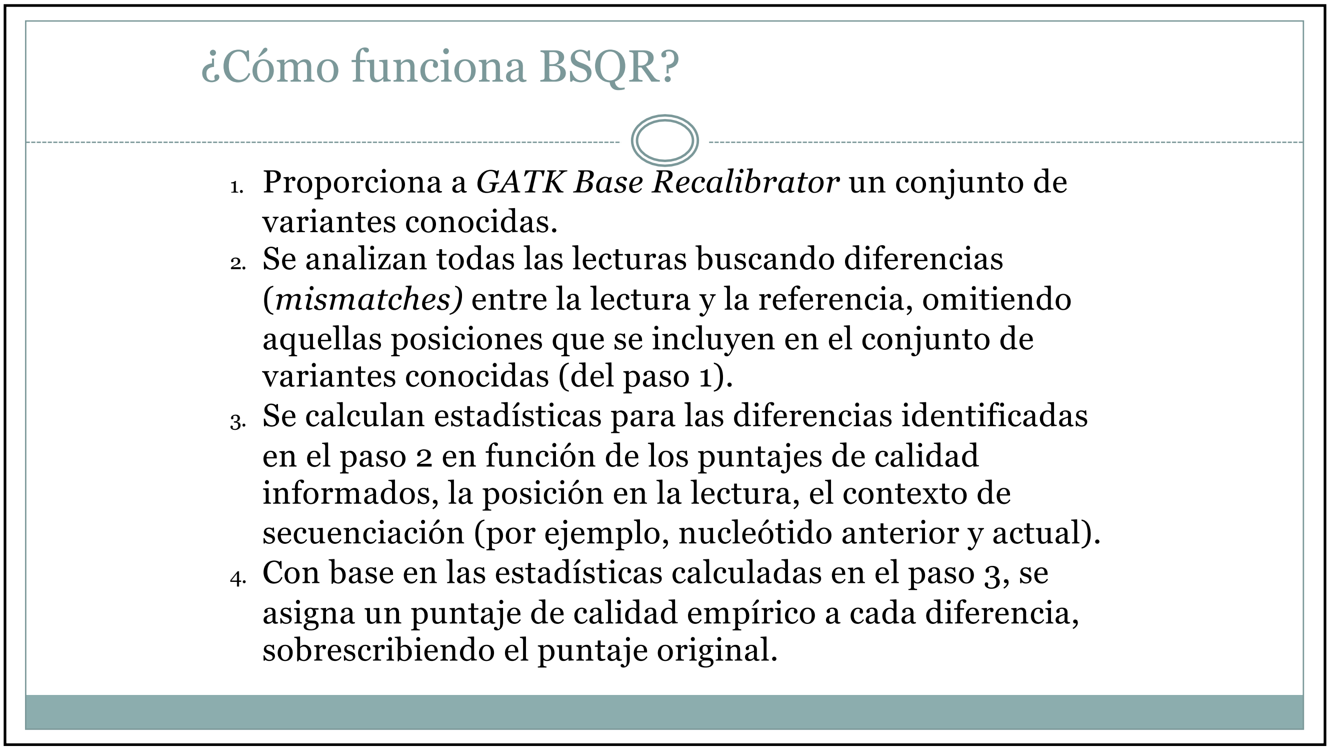
34



How does BQSR work?

1. You provide GATK Base Recalibrator with a set of known variants.
2. GATK Base Recalibrator analyzes all reads looking for mismatches between the read and reference, skipping those positions which are included in the set of known variants (from step 1).
3. GATK Base Recalibrator computes statistics on the mismatches (identified in step 2) based on the reported quality score, the position in the read, the sequencing context (ex: preceding and current nucleotide).
4. Based on the statistics computed in step 3, an empirical quality score is assigned to each mismatch, overwriting the original reported quality score.

35



¿Cómo funciona BSQR?

1. Proporciona a *GATK Base Recalibrator* un conjunto de variantes conocidas.
2. Se analizan todas las lecturas buscando diferencias (*mismatches)* entre la lectura y la referencia, omitiendo aquellas posiciones que se incluyen en el conjunto de variantes conocidas (del paso 1).
3. Se calculan estadísticas para las diferencias identificadas en el paso 2 en función de los puntajes de calidad informados, la posición en la lectura, el contexto de secuenciación (por ejemplo, nucleótido anterior y actual).
4. Con base en las estadísticas calculadas en el paso 3, se asigna un puntaje de calidad empírico a cada diferencia, sobrescribiendo el puntaje original.

36

Variant annotations are the “features” of the model

VCF record for an A/G SNP at 22:49582364

```
22 49582364              .        A        G        198.96  .
   AC=3;
   AF=0.50;
   AN=6;
   DP=87;
   MLEAC=3;
   MLEAF=0.50;   INFO field
   MQ=71.31;
   MQ0=22;
   QD=2.29;
   SB=-31.76
   GT:DP:GQ    0/1:12:99    0/1:11:89    0/1:28:37
```

AC	No. chromosomes carrying alt allele	MLEAF	Max likelihood AF
AN	Total no. of chromosomes	MQ	RMS MAPQ of all reads
AF	Allele frequency	MQ0	No. of MAPQ 0 reads at locus
DP	Depth of coverage	QD	QUAL score over depth
MLEAC	Max likelihood AC		

Note that VQSR will only look at INFO annotations;

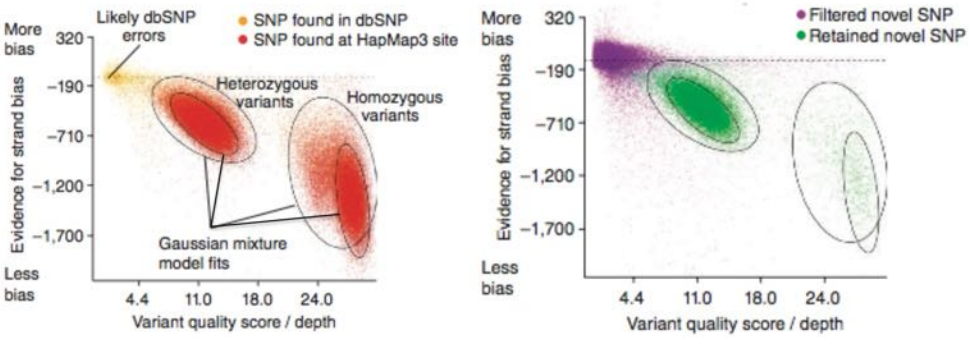
37

Two steps: (1) train a model then (2) apply to callset

Basic idea: training on high-confidence known sites to determine the probability that other sites are true

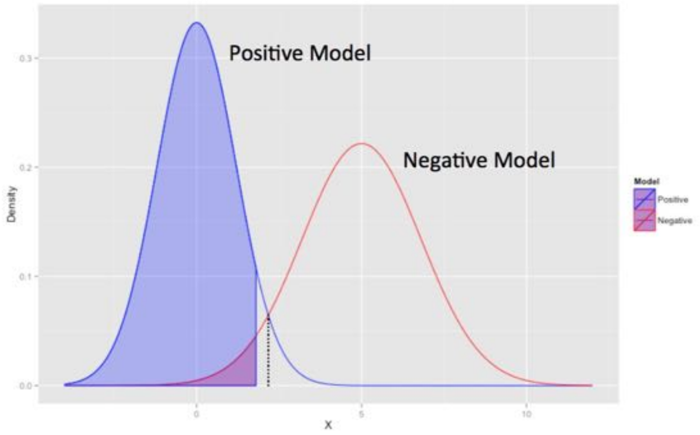
(1) Train model using HapMap

(2) Apply model to callset



38

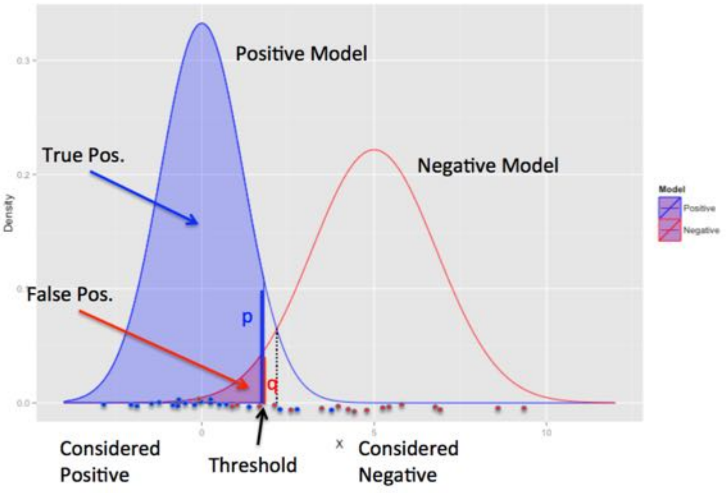
There are in fact two components to the model



- A **negative model** is also built during training
- It represents the probability of variants to be **false positives**

39

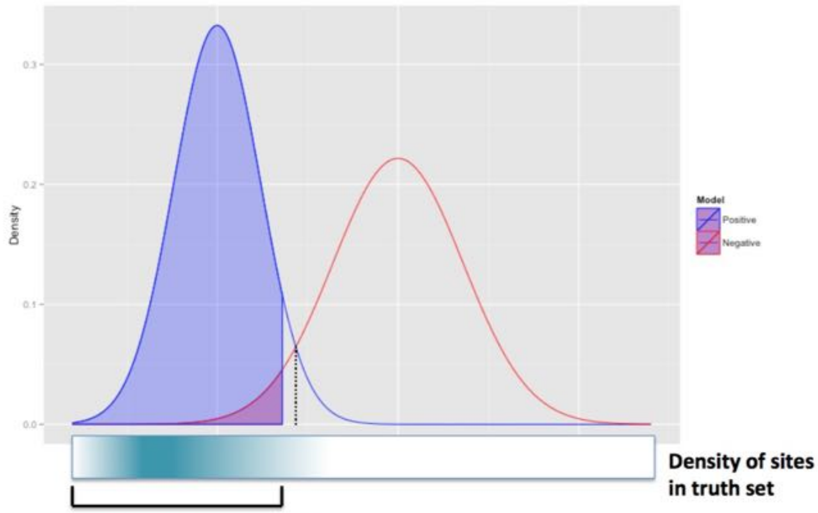
The VQSLOD threshold is a tradeoff between TP and FP



VQSLOD(x) = Log(p(x)/q(x)) (VQSLOD is distinct from QUAL!)

40

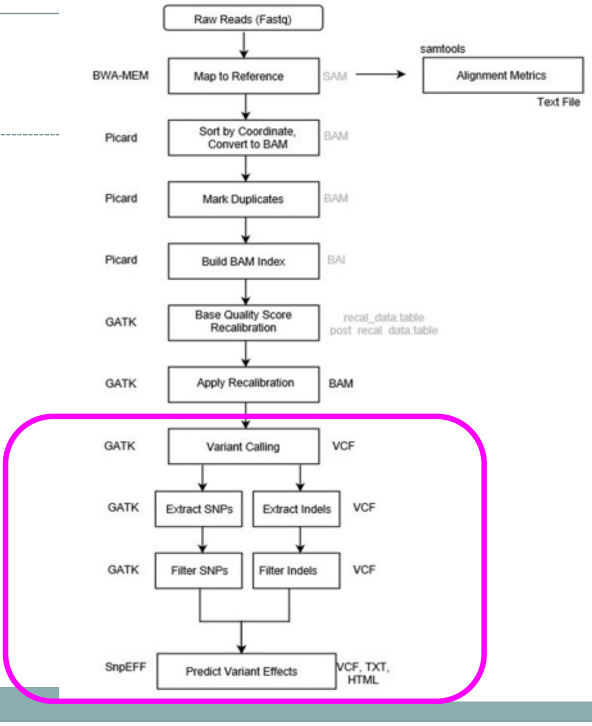
We set the threshold based on **sensitivity to truth data**



Density of sites in truth set

What threshold do we need to set to capture X % of the sites in the truth set?

41



Raw Reads (Fastq)

BWA-MEM — Map to Reference — SAM → samtools — Alignment Metrics — Text File

Picard — Sort by Coordinate, Convert to BAM — BAM

Picard — Mark Duplicates — BAM

Picard — Build BAM Index — BAI

GATK — Base Quality Score Recalibration — recal_data.table, post_recal_data.table

GATK — Apply Recalibration — BAM

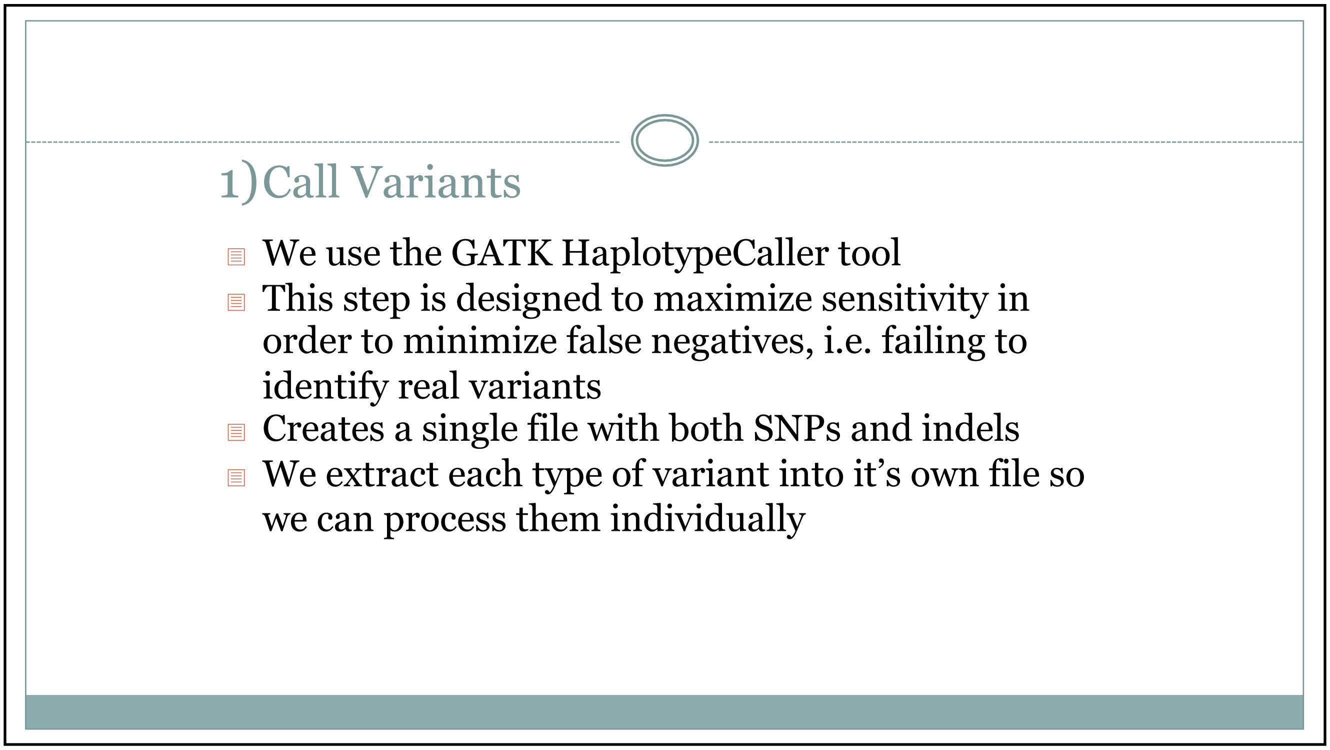
GATK — Variant Calling — VCF

GATK — Extract SNPs, Extract Indels — VCF

GATK — Filter SNPs, Filter Indels — VCF

SnpEFF — Predict Variant Effects — VCF, TXT, HTML

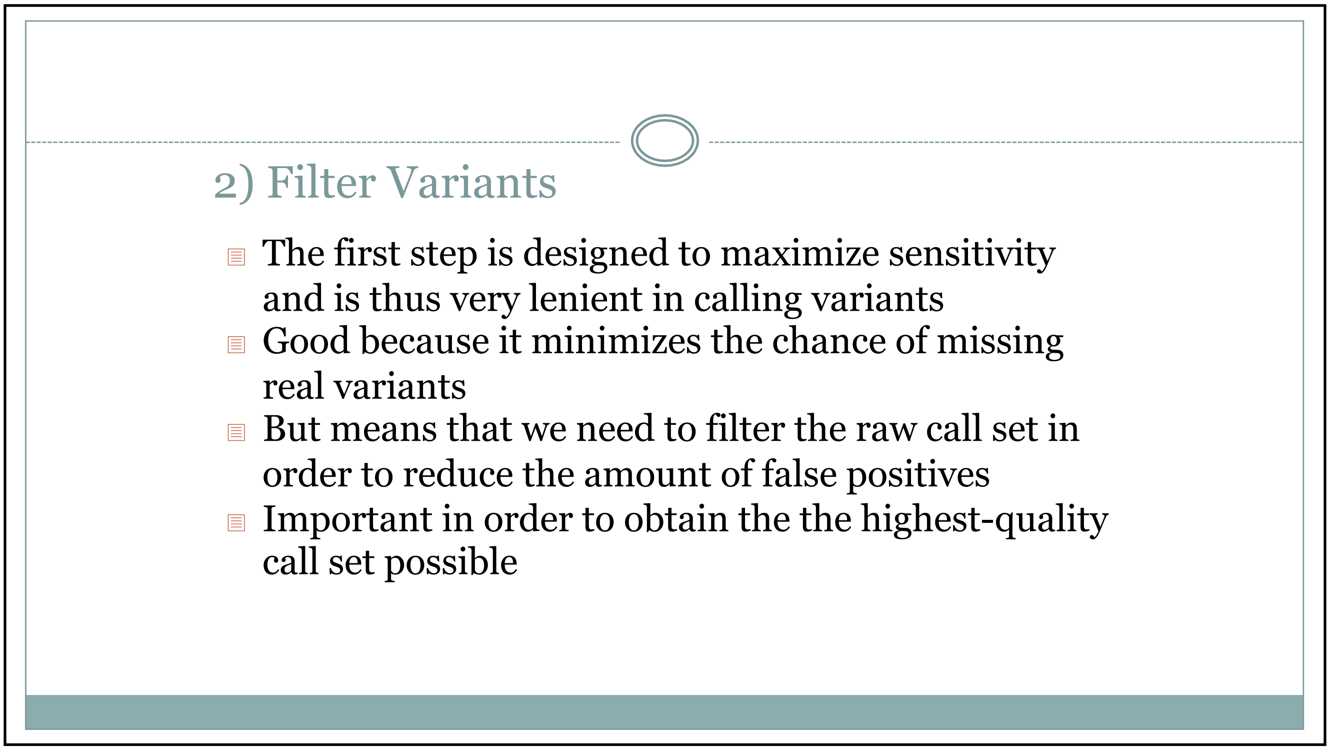
42



1)Call Variants

- We use the GATK HaplotypeCaller tool
- This step is designed to maximize sensitivity in order to minimize false negatives, i.e. failing to identify real variants
- Creates a single file with both SNPs and indels
- We extract each type of variant into it's own file so we can process them individually

43

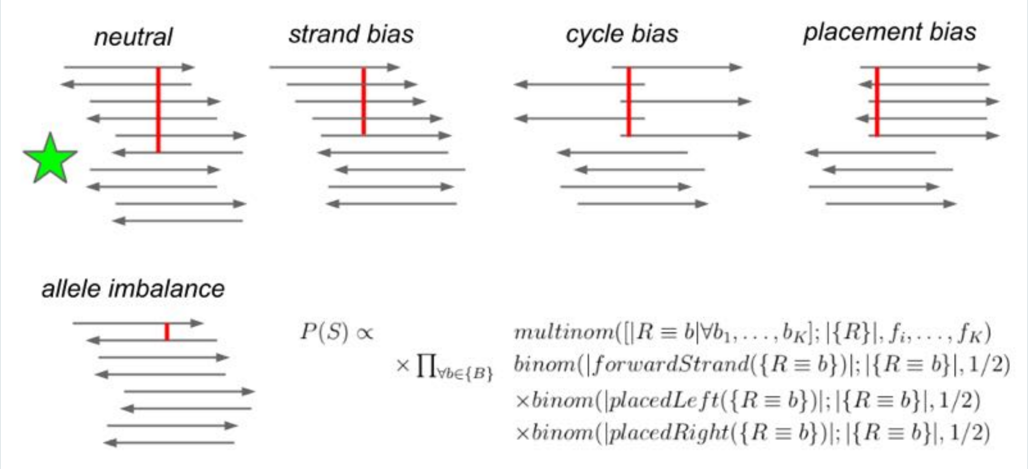


2) Filter Variants

- The first step is designed to maximize sensitivity and is thus very lenient in calling variants
- Good because it minimizes the chance of missing real variants
- But means that we need to filter the raw call set in order to reduce the amount of false positives
- Important in order to obtain the the highest-quality call set possible

44

Sesgos de variantes detectadas

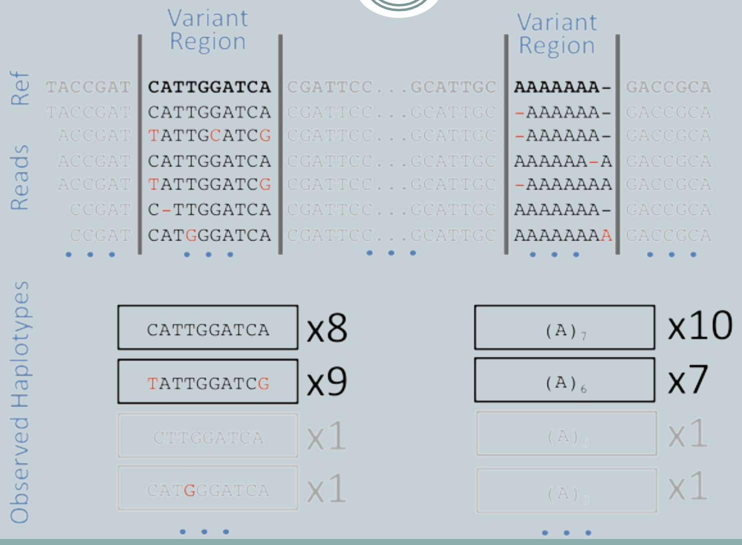


$$P(S) \propto \quad multinom([|R \equiv b|\forall b_1,\ldots,b_K];|\{R\}|,f_i,\ldots,f_K)$$
$$\times \prod_{\forall b \in \{B\}} binom(|forwardStrand(\{R \equiv b\})|;|\{R \equiv b\}|,1/2)$$
$$\times binom(|placedLeft(\{R \equiv b\})|;|\{R \equiv b\}|,1/2)$$
$$\times binom(|placedRight(\{R \equiv b\})|;|\{R \equiv b\}|,1/2)$$

https://ekg.github.io/2015/12/08/How-to-freebayes

45

Freebayes – Llamador de variantes basado en haplotipos


Variant Region
Variant Region
Ref
Reads
Observed Haplotypes
TACCGAT CATTGGATCA CGATTCC...GCATTGC AAAAAAA- GACCGCA
TACCGAT CATTGGATCA CGATTCC...GCATTGC -AAAAAA- GACCGCA
ACCGAT TATTGCATCG CGATTCC...GCATTGC -AAAAAA- GACCGCA
ACCGAT CATTGGATCA CGATTCC...GCATTGC AAAAAA-A GACCGCA
ACCGAT TATTGGATCG CGATTCC...GCATTGC -AAAAAAA GACCGCA
CCGAT C-TTGGATCA CGATTCC...GCATTGC AAAAAAA- GACCGCA
CCGAT CATGGGATCA CGATTCC...GCATTGC AAAAAAAA GACCGCA
CATTGGATCA x8
TATTGGATCG x9
CTTGGATCA x1
CATGGGATCA x1
(A)7 x10
(A)6 x7
(A) x1
(A) x1

https://github.com/freebayes/freebayes

46

3) Annotation

- We use SnpEff
- Annotates and predicts the effects of variants on genes
  - Codon changes
  - Amino acid changes
  - Genomic region
  - Functional effect (silent, missense)
- SnpEff has pre-built databases for thousands of genomes

47

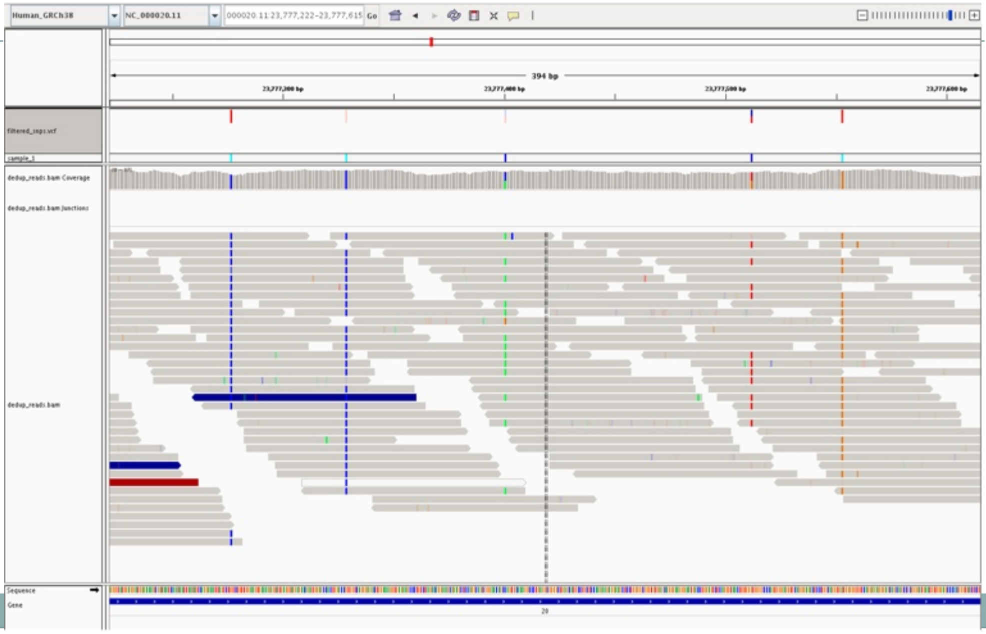
Archivo VCF

```
##fileformat=VCFv4.0
##fileDate=20090805
##source=myImputationProgramV3.1
##reference=1000GenomesPilot-NCBI36
##phasing=partial
##INFO=<ID=NS,Number=1,Type=Integer,Description="Number of Samples With Data">
##INFO=<ID=DP,Number=1,Type=Integer,Description="Total Depth">
##INFO=<ID=AF,Number=.,Type=Float,Description="Allele Frequency">
##INFO=<ID=AA,Number=1,Type=String,Description="Ancestral Allele">
##INFO=<ID=DB,Number=0,Type=Flag,Description="dbSNP membership, build 129">
##INFO=<ID=H2,Number=0,Type=Flag,Description="HapMap2 membership">
##FILTER=<ID=q10,Description="Quality below 10">
##FILTER=<ID=s50,Description="Less than 50% of samples have data">
##FORMAT=<ID=GT,Number=1,Type=String,Description="Genotype">
##FORMAT=<ID=GQ,Number=1,Type=Integer,Description="Genotype Quality">
##FORMAT=<ID=DP,Number=1,Type=Integer,Description="Read Depth">
##FORMAT=<ID=HQ,Number=2,Type=Integer,Description="Haplotype Quality">
#CHROM POS     ID        REF ALT    QUAL FILTER INFO                              FORMAT      NA00001        NA00002        NA00003
20     14370   rs6054257 G      A       29   PASS   NS=3;DP=14;AF=0.5;DB;H2           GT:GQ:DP:HQ 0|0:48:1:51,51 1|0:48:8:51,51 1/1:43:5:.,.
20     17330   .         T      A       3    q10    NS=3;DP=11;AF=0.017               GT:GQ:DP:HQ 0|0:49:3:58,50 0|1:3:5:65,3   0/0:41:3
20     1110696 rs6040355 A      G,T     67   PASS   NS=2;DP=10;AF=0.333,0.667;AA=T;DB GT:GQ:DP:HQ 1|2:21:6:23,27 2|1:2:0:18,2   2/2:35:4
20     1230237 .         T      .       47   PASS   NS=3;DP=13;AA=T                   GT:GQ:DP:HQ 0|0:54:7:56,60 0|0:48:4:51,51 0/0:61:2
20     1234567 microsat1 GTCT   G,GTACT 50   PASS   NS=3;DP=9;AA=G                    GT:GQ:DP    0/1:35:4       0/2:17:2       1/1:40:3
```

http://www.internationalgenome.org/wiki/Analysis/vcf4.0/

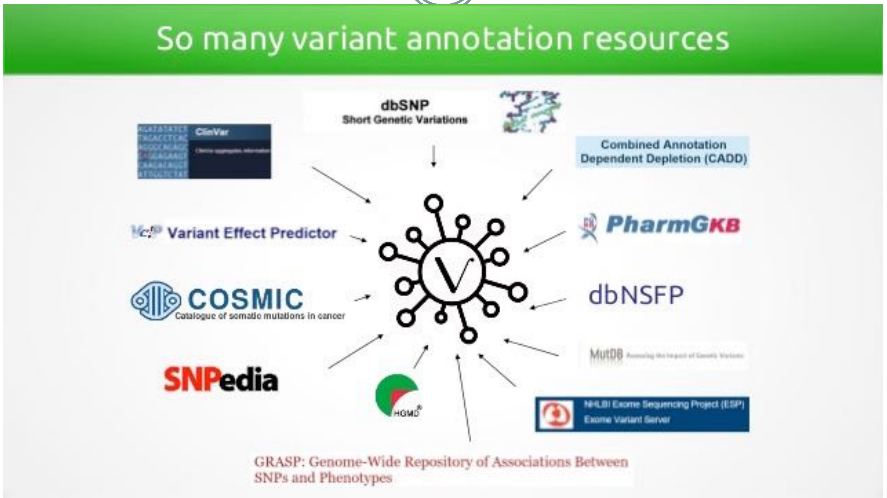
48

4) Visualization - IGV

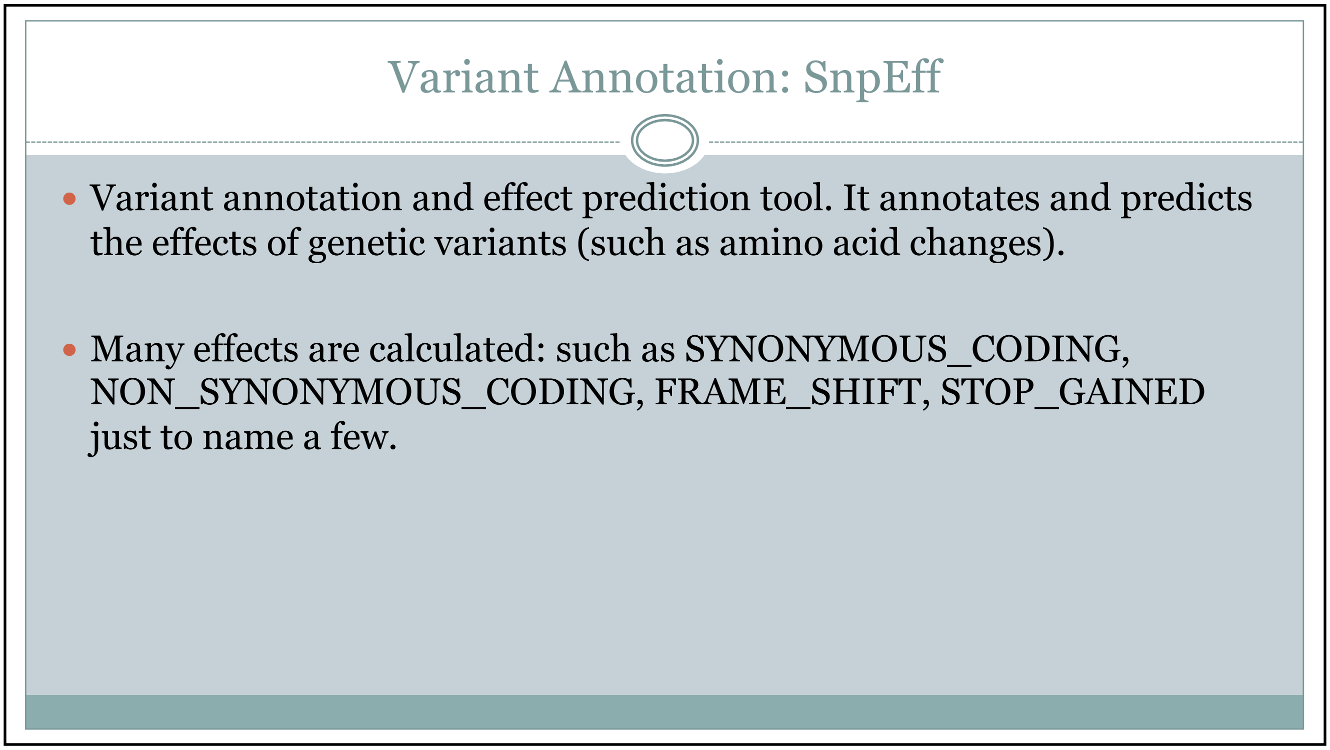


49

Variant Annotation


So many variant annotation resources
dbSNP
Short Genetic Variations
ClinVar
Combined Annotation Dependent Depletion (CADD)
Variant Effect Predictor
PharmGKB
COSMIC
Catalogue of somatic mutations in cancer
dbNSFP
MutDB
SNPedia
HGMD
NHLBI Exome Sequencing Project (ESP)
Exome Variant Server
GRASP: Genome-Wide Repository of Associations Between SNPs and Phenotypes

50

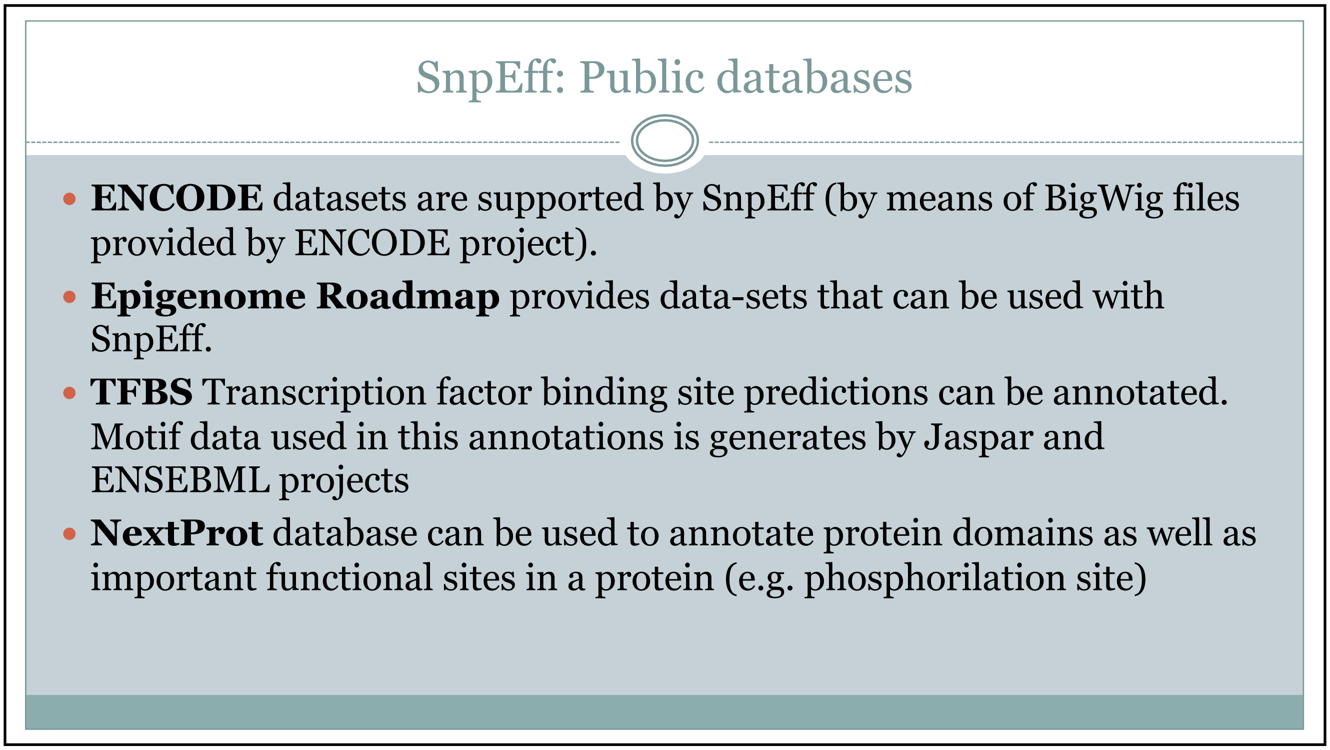


Variant Annotation: SnpEff

- Variant annotation and effect prediction tool. It annotates and predicts the effects of genetic variants (such as amino acid changes).

- Many effects are calculated: such as SYNONYMOUS_CODING, NON_SYNONYMOUS_CODING, FRAME_SHIFT, STOP_GAINED just to name a few.

51



SnpEff: Public databases

- **ENCODE** datasets are supported by SnpEff (by means of BigWig files provided by ENCODE project).
- **Epigenome Roadmap** provides data-sets that can be used with SnpEff.
- **TFBS** Transcription factor binding site predictions can be annotated. Motif data used in this annotations is generates by Jaspar and ENSEBML projects
- **NextProt** database can be used to annotate protein domains as well as important functional sites in a protein (e.g. phosphorilation site)

52

CADD - Combined Annotation Dependent Depletion

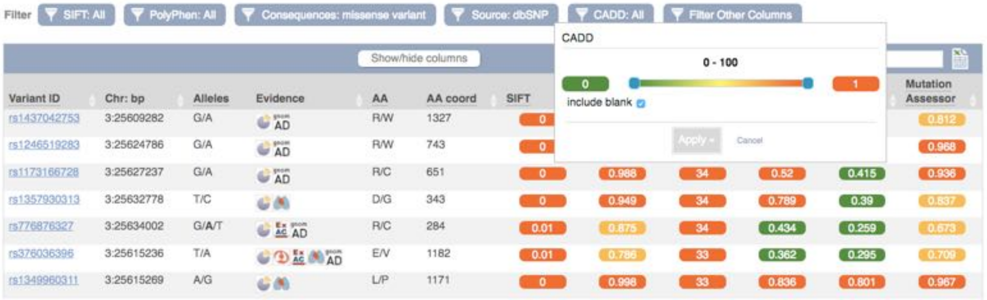
- Puntaje que integra múltiples anotaciones en una métrica contrastando variantes que sobrevivieron a la selección natural con mutaciones simuladas

- Generalmente, se considera que:
  - Puntuaciones CADD mayores o iguales a 20 son sugestivas de un efecto patogénico importante, ya que corresponden al 1% más dañino de todas las variantes posibles en el genoma humano.
  - Puntuaciones entre 10 y 20 indican variantes con un probable efecto funcional o moderadamente perjudicial.
  - Valores por debajo de 10 suelen considerarse variantes benignas o poco dañinas.

https://cadd.gs.washington.edu/

53

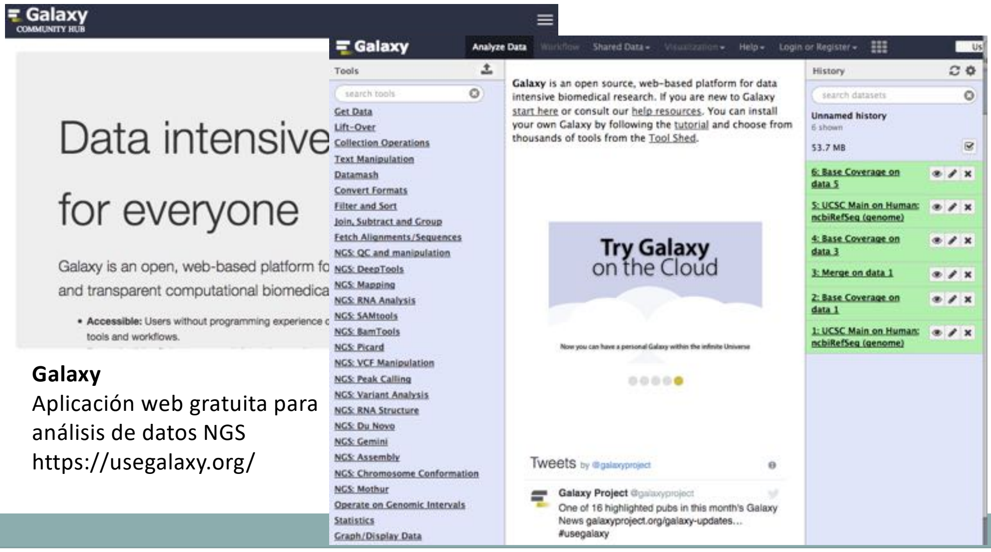
Puntajes CADD



Variant ID	Chr: bp	Alleles	Evidence	AA	AA coord	SIFT					Mutation Assessor
rs1437042753	3:25609282	G/A	gnomAD	R/W	1327	0					0.812
rs1246519283	3:25624786	G/A	gnomAD	R/W	743	0					0.968
rs1173166728	3:25627237	G/A	gnomAD	R/C	651	0	0.988	34	0.52	0.415	0.936
rs1357930313	3:25632778	T/C		D/G	343	0	0.949	34	0.789	0.39	0.837
rs776876327	3:25634002	G/A/T	ExAC gnomAD	R/C	284	0.01	0.875	34	0.434	0.259	0.673
rs376036396	3:25615236	T/A	ExAC gnomAD	E/V	1182	0.01	0.786	33	0.362	0.295	0.709
rs1349960311	3:25615269	A/G		L/P	1171	0	0.998	33	0.836	0.801	0.967

54

Software Libre



Galaxy COMMUNITY HUB

Data intensive

for everyone

Galaxy is an open, web-based platform fo
and transparent computational biomedica

- **Accessible:** Users without programming experience c
tools and workflows.

Get Data
Lift-Over
Collection Operations
Text Manipulation
Datamash
Convert Formats
Filter and Sort
Join, Subtract and Group
Fetch Alignments/Sequences
NGS: QC and manipulation
NGS: DeepTools
NGS: Mapping
NGS: RNA Analysis
NGS: SAMtools
NGS: BamTools
NGS: Picard
NGS: VCF Manipulation
NGS: Peak Calling
NGS: Variant Analysis
NGS: RNA Structure
NGS: Du Novo
NGS: Gemini
NGS: Assembly
NGS: Chromosome Conformation
NGS: Mothur
Operate on Genomic Intervals
Statistics
Graph/Display Data

Galaxy is an open source, web-based platform for data intensive biomedical research. If you are new to Galaxy start here or consult our help resources. You can install your own Galaxy by following the tutorial and choose from thousands of tools from the Tool Shed.

Try Galaxy on the Cloud

Now you can have a personal Galaxy within the infinite Universe

Tweets by @galaxyproject

Galaxy Project @galaxyproject
One of 16 highlighted pubs in this month's Galaxy News galaxyproject.org/galaxy-updates…
#usegalaxy

Unnamed history
6 shown
53.7 MB

6: Base Coverage on data 5
5: UCSC Main on Human: ncbiRefSeq (genome)
4: Base Coverage on data 3
3: Merge on data 1
2: Base Coverage on data 1
1: UCSC Main on Human: ncbiRefSeq (genome)

Galaxy
Aplicación web gratuita para análisis de datos NGS
https://usegalaxy.org/

55

Resources in NGS data analysis

Public forums:

SEQanswers
the next generation sequencing community

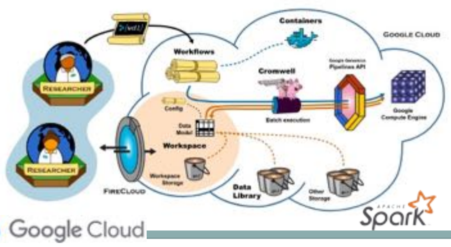
Bioconductor
OPEN SOURCE SOFTWARE FOR BIOINFORMATICS

BioBits: A Blog about bioinformatics

Biostars
— BIOINFORMATICS EXPLAINED —

BROAD INSTITUTE **FireCloud**

An open platform for secure and scalable analysis on the cloud



Google Cloud

Azure Databricks

Featured Notebooks

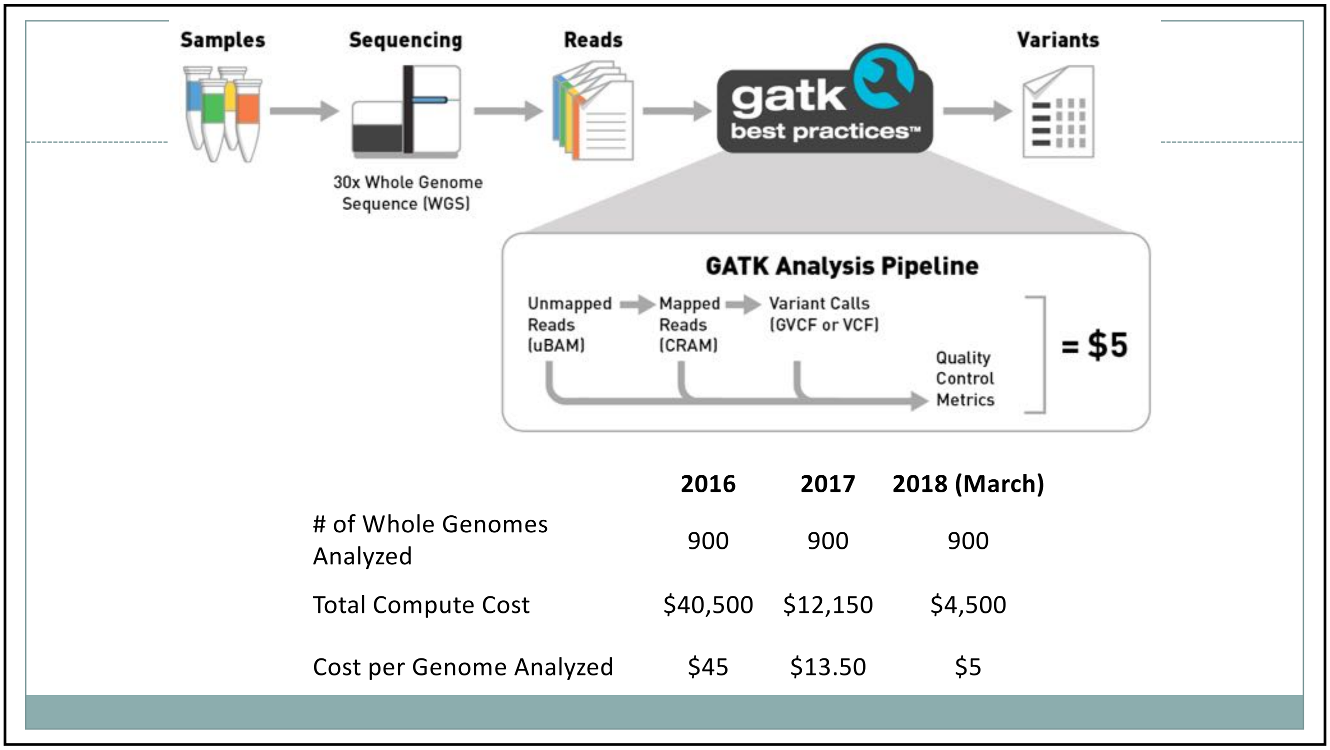
Introduction to Apache Spark on Databricks
Databricks for Data Scientists
Introduction to Structured Streaming

New
Notebook
Job
Cluster
Table
Library

Documentation
Databricks Guide
Python, R, Scala, SQL
Importing Data

Open Recent
iris
Tutorial Test
mynotebook
Introduction to Apache Spark on Databricks
2018-01-25 - Azure Blob Storage Import Example Notebook

56



Samples

Sequencing

Reads

gatk best practices™

Variants

30x Whole Genome Sequence (WGS)

GATK Analysis Pipeline

Unmapped Reads (uBAM)

Mapped Reads (CRAM)

Variant Calls (GVCF or VCF)

Quality Control Metrics

= $5

	2016	2017	2018 (March)
# of Whole Genomes Analyzed	900	900	900
Total Compute Cost	$40,500	$12,150	$4,500
Cost per Genome Analyzed	$45	$13.50	$5

57