Análisis Secundario I: Variant Calling

Sara Monzón

BU-ISCIII

Unidades Científico Técnicas – SGAFI-ISCIII

13-17 Junio 2022, 9ª Edición
Programa Formación Continua, ISCIII

Índice

- Dónde estamos
- ¿Qué es llamada a variantes?
- Problemas que nos encontramos
- Software de variant calling
- Formatos: vcf y bed
- Anotación y filtrado
- Ejemplos de llamada a variantes:
  - Cáncer
  - Trío
  - Bacterias

Dónde estamos

Fastq files

Preprocesamiento

Alineamiento

Postprocesamiento: Análisis de variantes

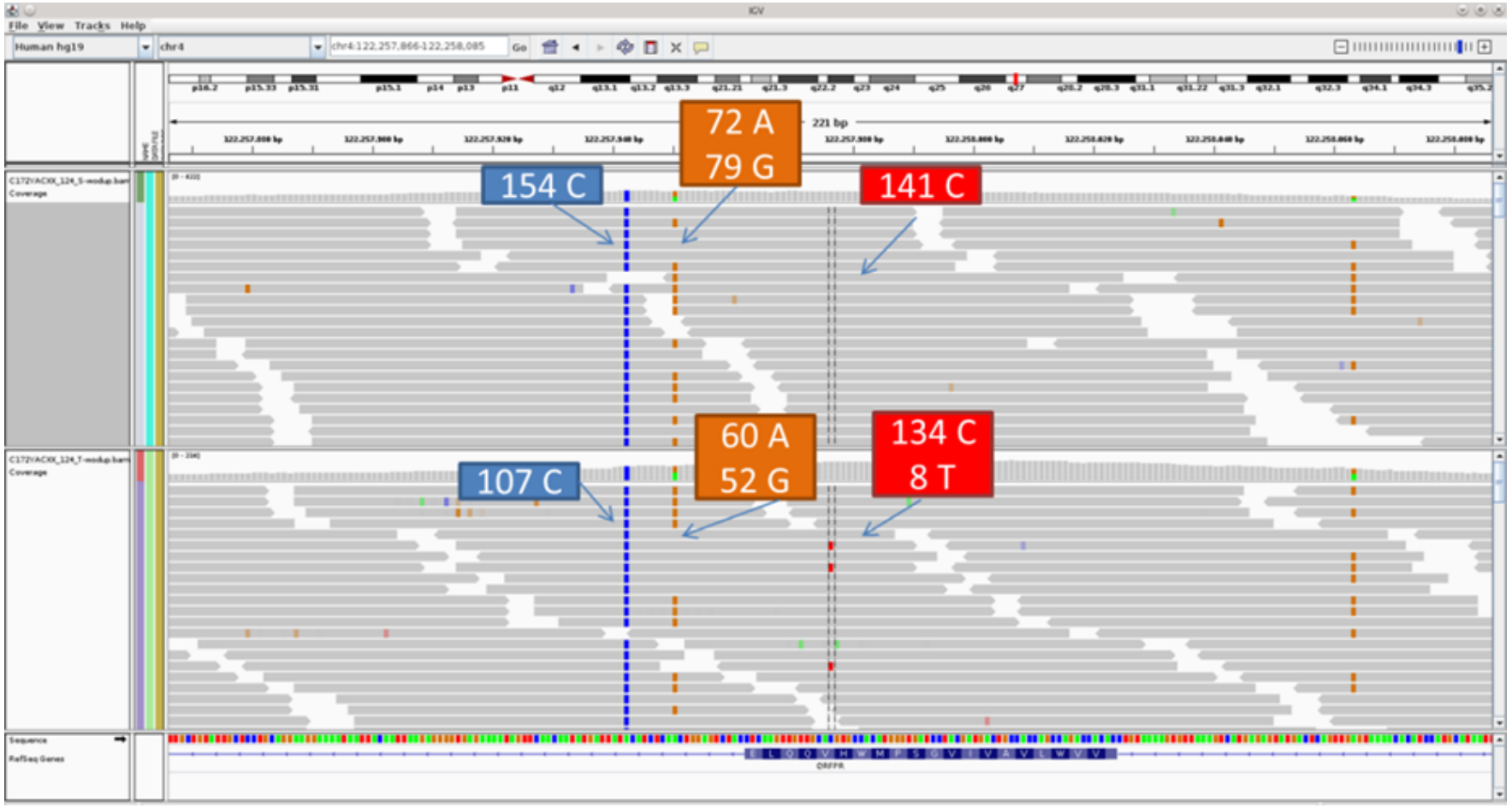
¿Qué es llamada a variantes?

- El concepto de la llamada a variantes es sencillo:

  - Encontrar posiciones en nuestras secuencias que sean diferentes a referencia –

- A partir de nuestras secuencias mapeadas en el genoma, se recorre cada columna del alineamiento y se cuentan cuántos alelos se encuentran y se comparan con la referencia.

¿Qué es la llamada a variantes?



Problemas que nos encontramos

- Preparación de la librería
- Errores en la secuenciación
- Errores de alineamiento
- Fiabilidad de la referencia

Problemas que nos encontramos

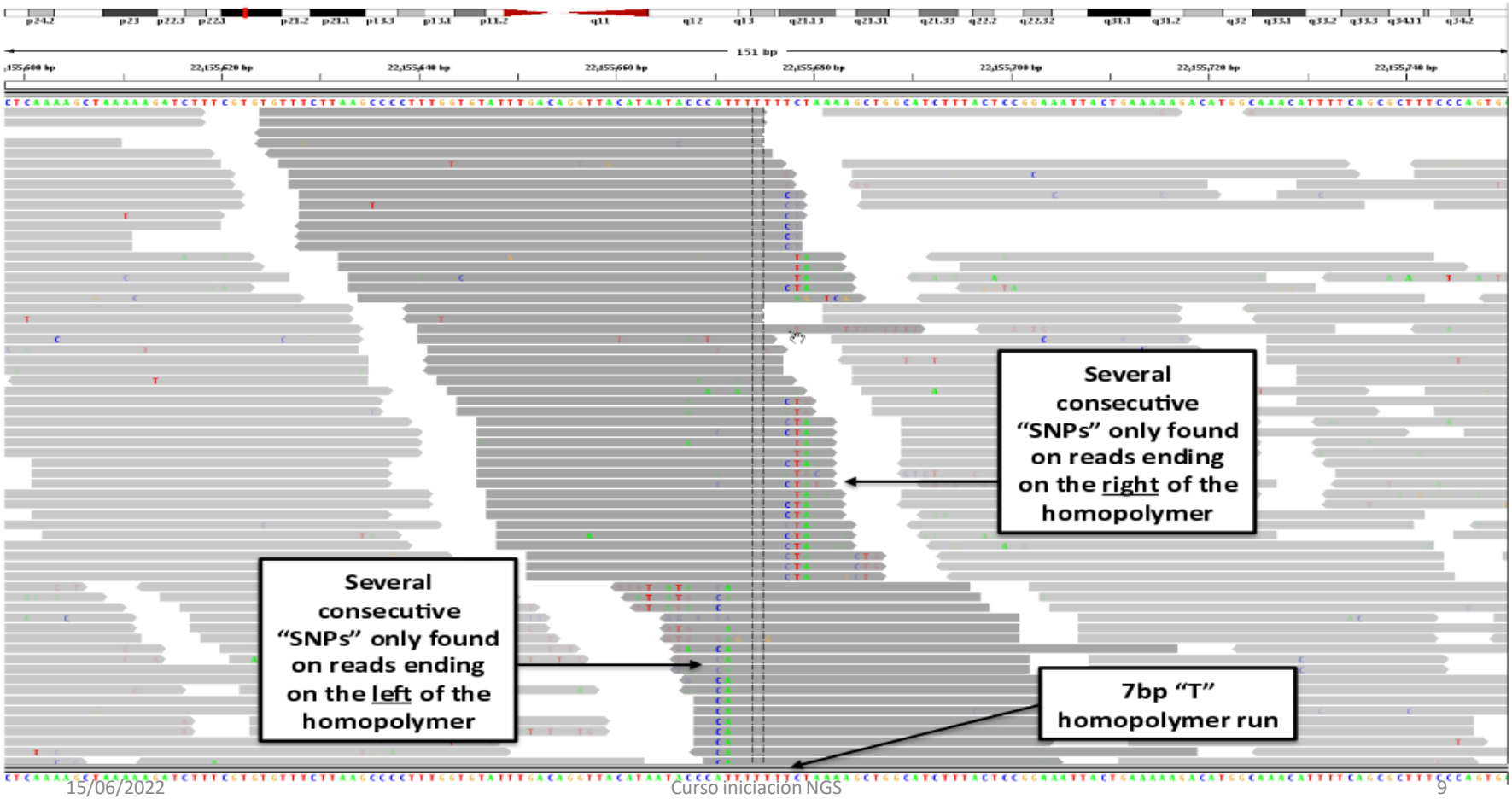
- Artefactos en la preparación de la librería
  - Mutaciones inducidas por PCR
    - Duplicados
    - Errores a final de la lectura
  - Contaminaciones

Problemas que nos encontramos

- Ratio de error asociado con la secuenciación.
- Soluciones:
  - Evaluación de Phred
  - Strand bias

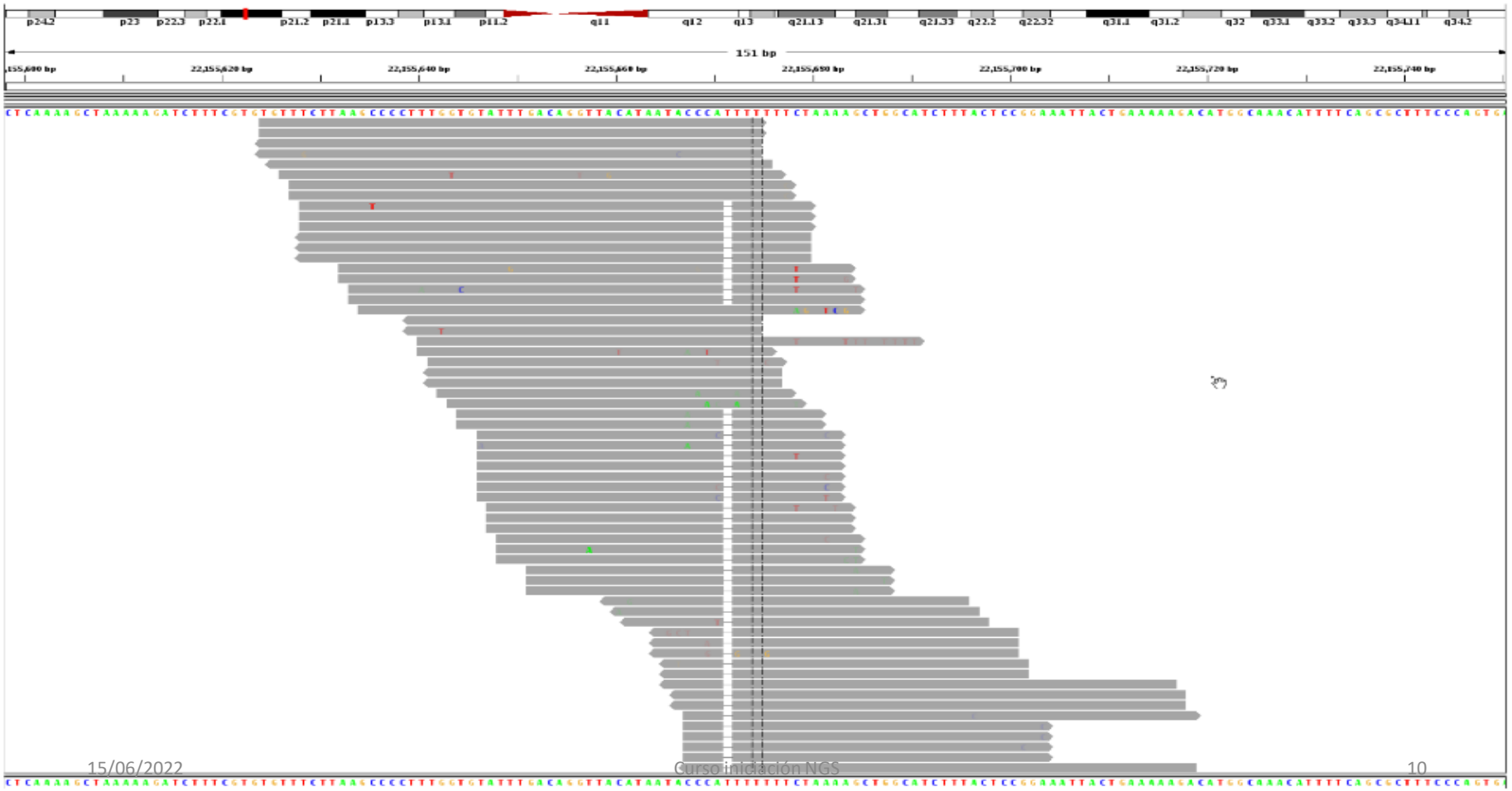
Problemas que nos encontramos

- Problemas de alineamiento



Problemas que nos encontramos

- Problemas de alineamiento



Problemas que nos encontramos

- Fiabilidad del genoma de referencia.
  - Ejemplo genoma humano:
    - Genoma obtenido de mezcla de 8 personas diferentes (Watson entre ellos)
    - Genoma haploide para individuo diploide.
    - Zonas de baja complejidad
    - Incompleto

Principales sofware de variant calling

- GATK
- Samtools
- Strelka
- MuTect
- JointSNVMix
- VarScan
- SolSNP
- Largo etc.

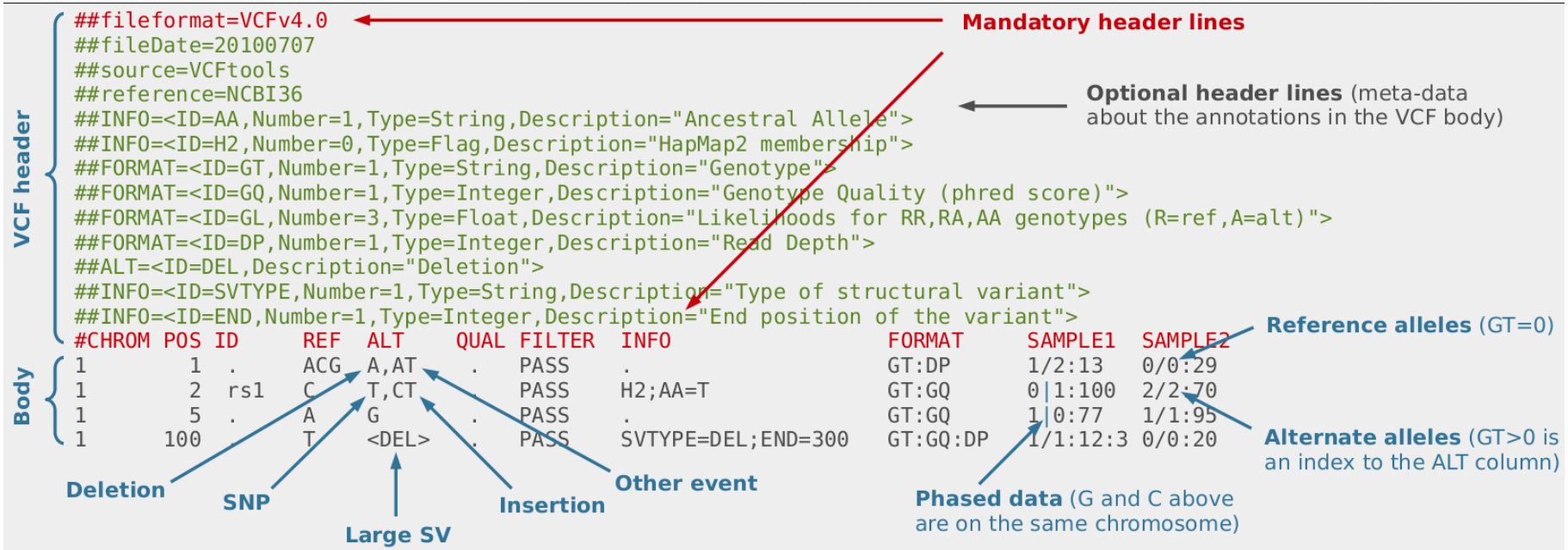
Llamada germinal en organismos diploides/haploides

Llamada somática en pares de control - tumor

Otros

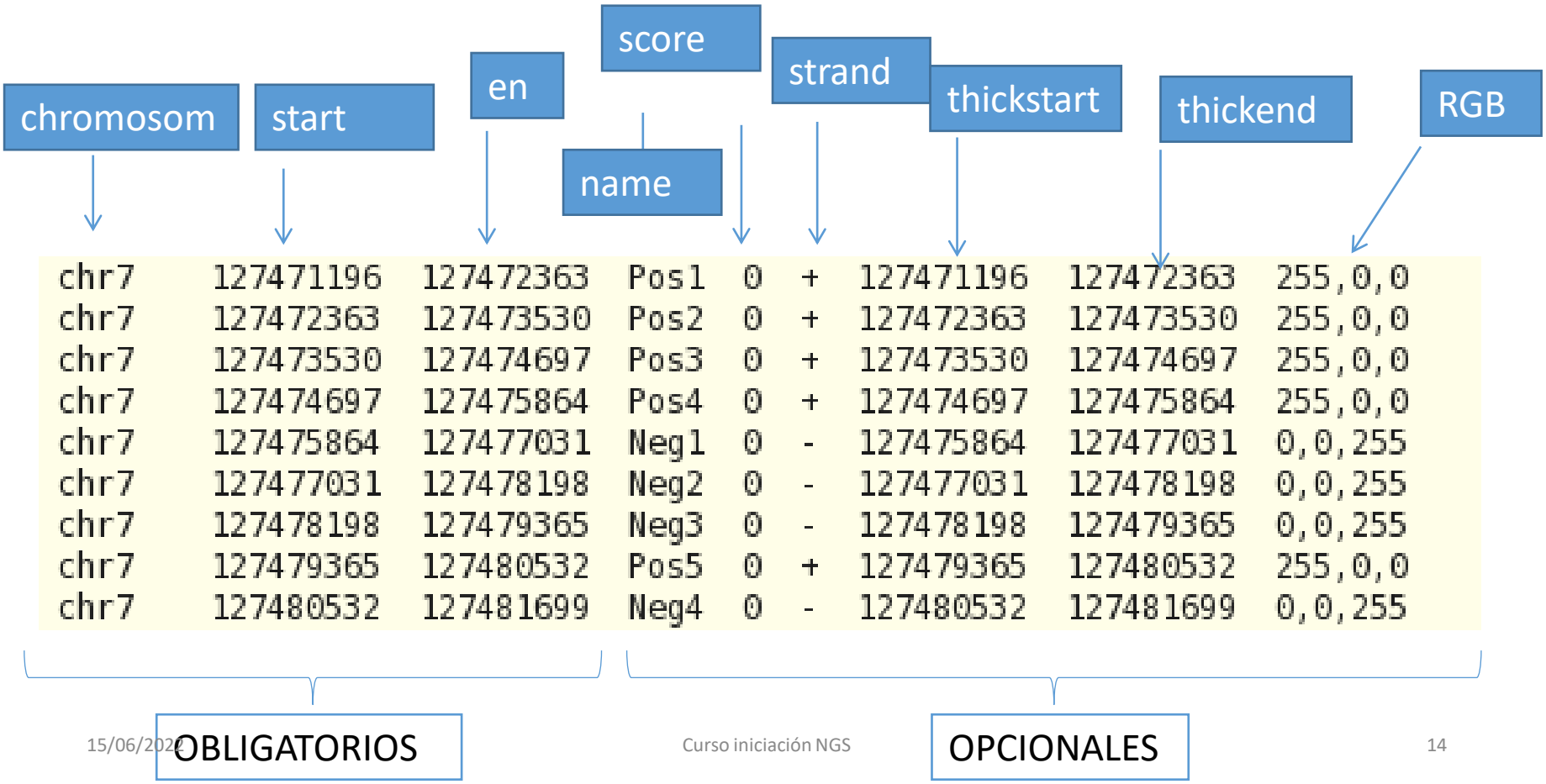
Formatos: vcf y bed

- Formato vcf



Formatos: vcf y bed

- Formato bed
  - Se utiliza para representar regiones y/o posiciones



Formatos: vcf y bed

- Formato bed
  - Se utiliza para representar regiones y/o posiciones
  - Consideraciones pare representar variantes.
  - Utiliza coordenadas 0-based para el inicio y 1-based para el final
  - De manera que la primera base del cromosoma 1 sería:

```
chr1    0    1    first_base
```

Formatos: vcf y bed

- Ejemplo de formato bed de variantes:

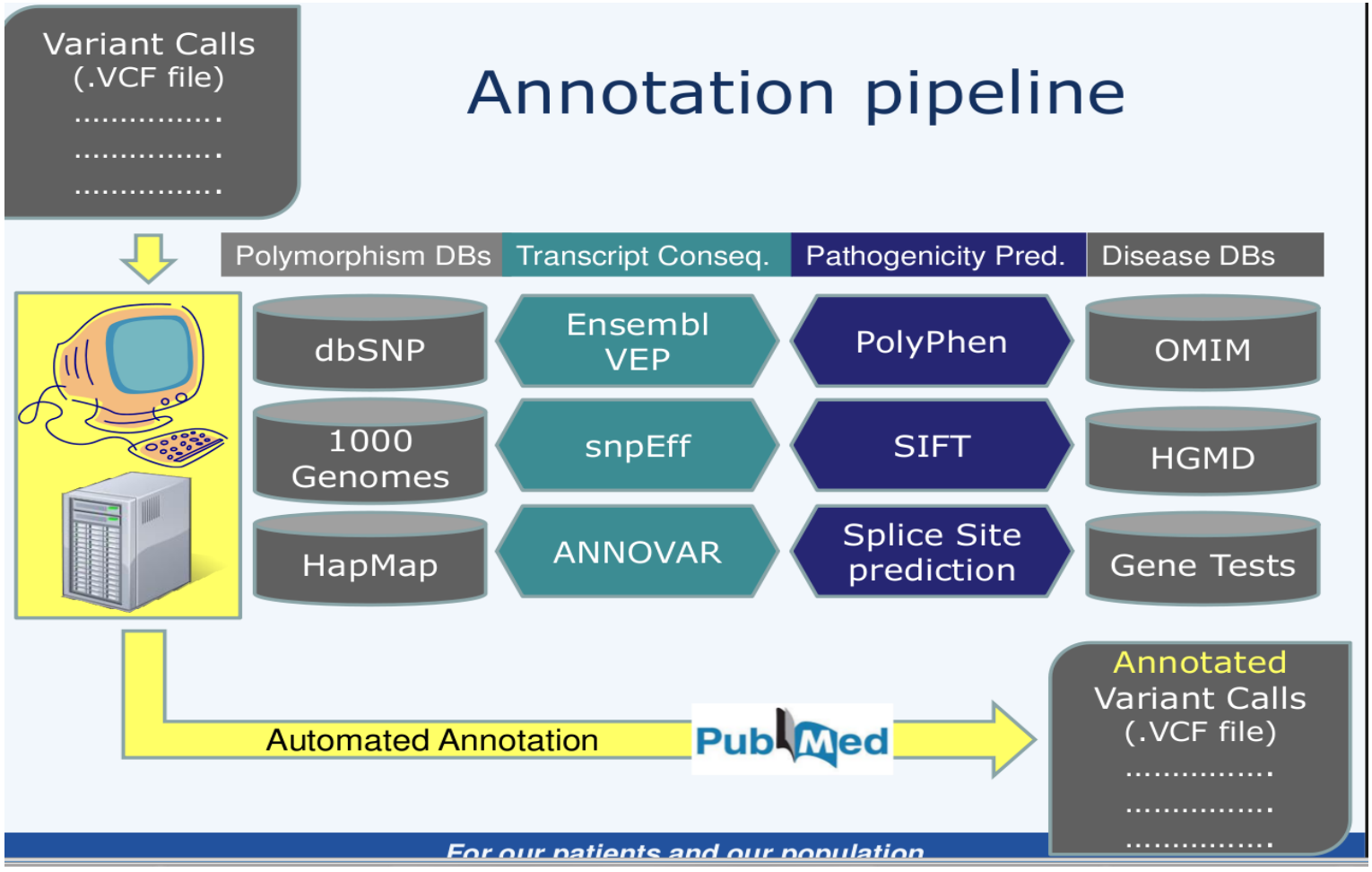
chr1	100154496	100154497	A	G
chr1	100182982	100182983	C	T
chr1	100195206	100195207	C	A
chr1	1002596	1002597	C	A
chr1	100343384	100343385	G	T
chr1	10041131	10041132	C	A
chr1	100575981	100575982	G	T
chr1	100621863	100621864	G	T
chr1	100672062	100672063	C	T
chr1	10067673	10067674	G	T
chr1	100733834	100733835	G	T
chr1	101007160	101007161	G	T
chr1	101186145	101186146	G	T
chr1	101376658	101376659	C	A
chr1	101379322	101379323	C	A
chr1	101490740	101490741	G	T
chr1	10161234	10161235	G	A
chr1	101705323	101705324	C	A
chr1	101705774	101705775	C	A
chr1	10179467	10179468	C	A
chr1	10197177	10197178	C	G
chr1	10197185	10197186	G	T

Esta es la posición donde se encuentra la variante

Admite prácticamente de todo
En este caso Alelo alternativo y Alelo referencia

Obligatorios

Anotación y Filtrado



Anotación y filtrado

- Anotación:
  - A nivel de gen: se anota gen y “feature” según la base de datos refgene (variante tipo missense, frameshit, intron, etc.)
  - Anotación de variantes no sinónimas: dbNSFP

    - SLR
    - SIFT
    - Polyphen2_HDIV
    - Polyphen2_HVAR
    - LRT
    - Mutation Taster
    - Mutation Assesor
    - FATHMM_score
    - CADD_score
    - GERP++_NR
    - GERP++_RS
    - PhyloP100way_vertebrate
    - 29way_logOdds

  - A nivel funcional: pseudogenes,UniprotFeature,etc.
  - A nivel de enfermedad: anotación de enfermedad asociada con ese gen en OMIM

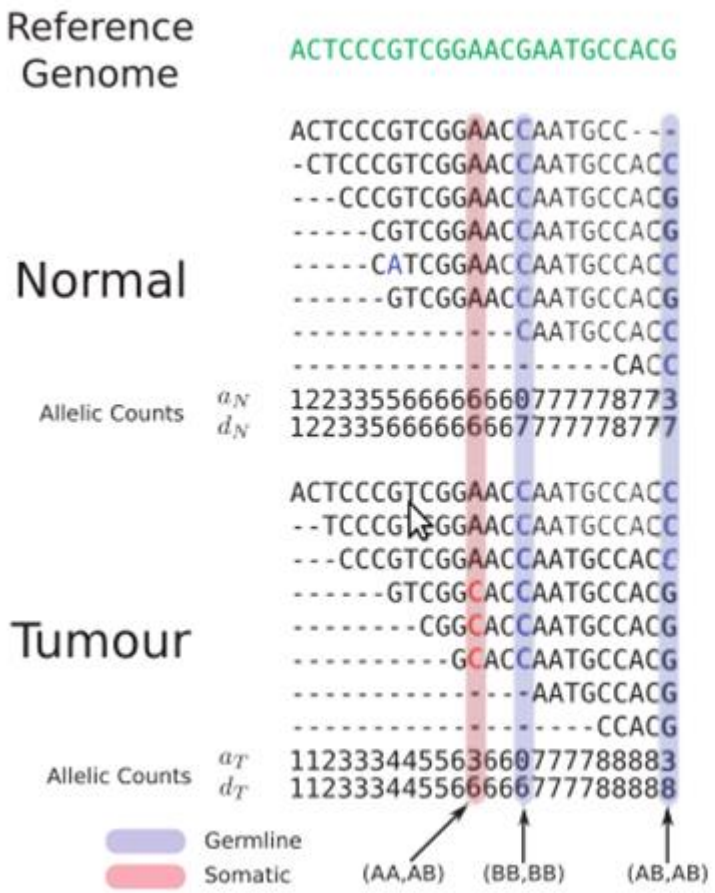
Ejemplo de variant calling: Cáncer

- Software específico para comparaciones tumor-control

	Samtools	GATK	VarScan 2	Somatic Sniper	JointSNVmix	Strelka	LoFreq	MuTec	Shimmer	EBCalling	Virmid
Publication	Li et al	McKenna et al	Koboldt et al	Larson et al	Roth et al	Saunders et al	Wilm et al	Cibulskis et al	Hansen et al	Shiraishi et al	Kim et al
Year	2009	2010	2012	2012	2012	2012	2012	2013	2013	2013	2013
Model	Bayesian	Bayesian	Fisher test		Prob	Bayesian	Binomial	Bayesian		Bayesian	Prob
Programming language	C	java	Java,perl	C	python	perl	C,python	java	perl	Perl,c,R	Bayesian
Paired sample	No	No	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Realignment	No	Yes	No	No	No	Yes	No	Yes	No	No	No

Ejemplo de variant calling: Cáncer

- Teoría de la llamada a variantes en cáncer



Joint Genotype Probabilities

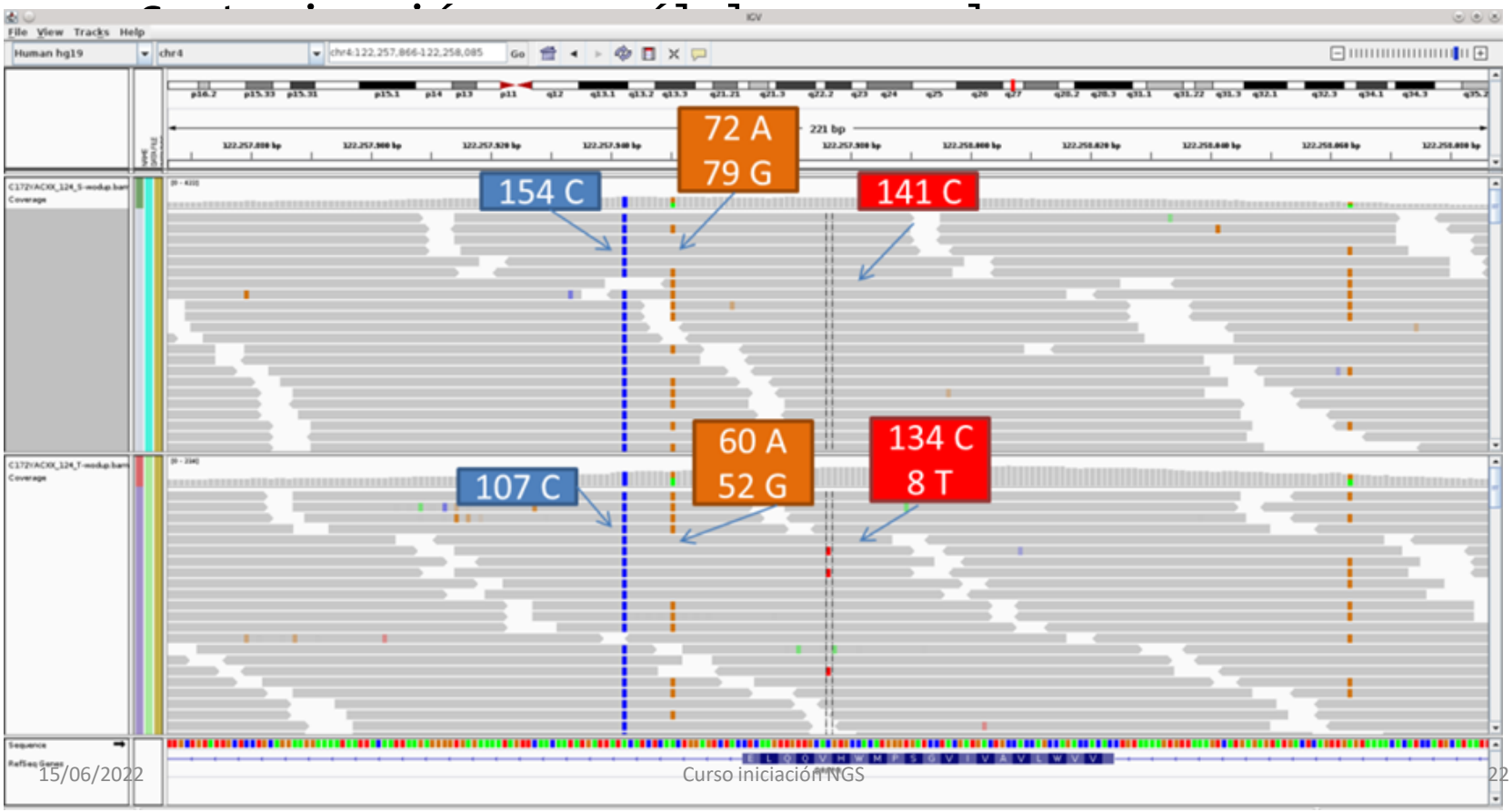
$P(G^i_{(g_N,g_T)} = 1)$

$g_N \backslash g_T$	AA	AB	BB
AA	0.01	0.95	0.00
AB	0.00	0.04	0.00
BB	0.00	0.00	0.00

Ejemplo de variant calling: Cáncer

Ejemplo de variant calling: Cáncer

- Problemas añadidos a la llamada a variantes
  - Heterogeneidad tumoral

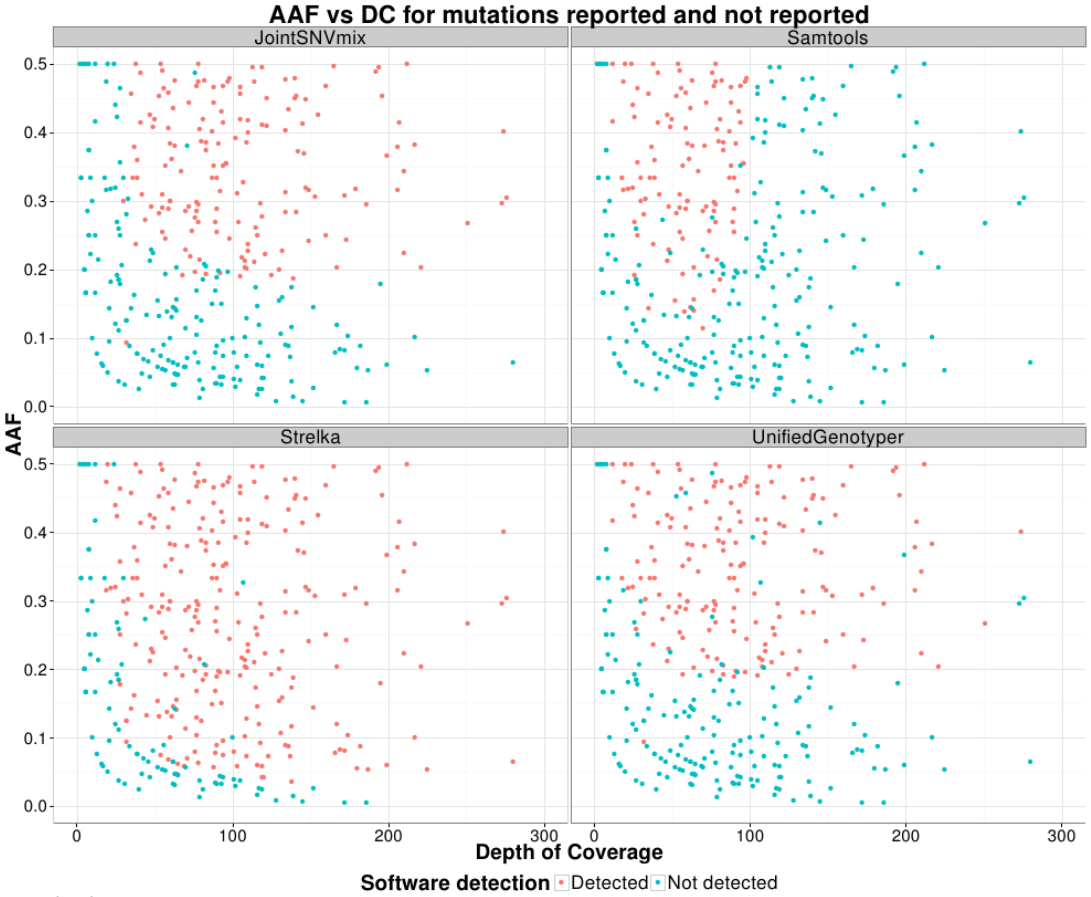


Ejemplo de variant calling: Cáncer

- Comparativa de distintos software de llamada a variantes en cáncer
  - Evaluar la mejor opción
  - Plantearse seleccionar la intersección de varios.
  - Caracterizar su comportamiento frente a cobertura y frecuencia del alelo alternativo.

Ejemplo de variant calling: Cáncer

Características de mutaciones detectadas en rojo y no detectadas en azul.



Frecuencia del alelo alternativo en función de la cobertura.

Ejemplo de variant calling: Cáncer

Análisis de la cobertura en función de la sensibilidad

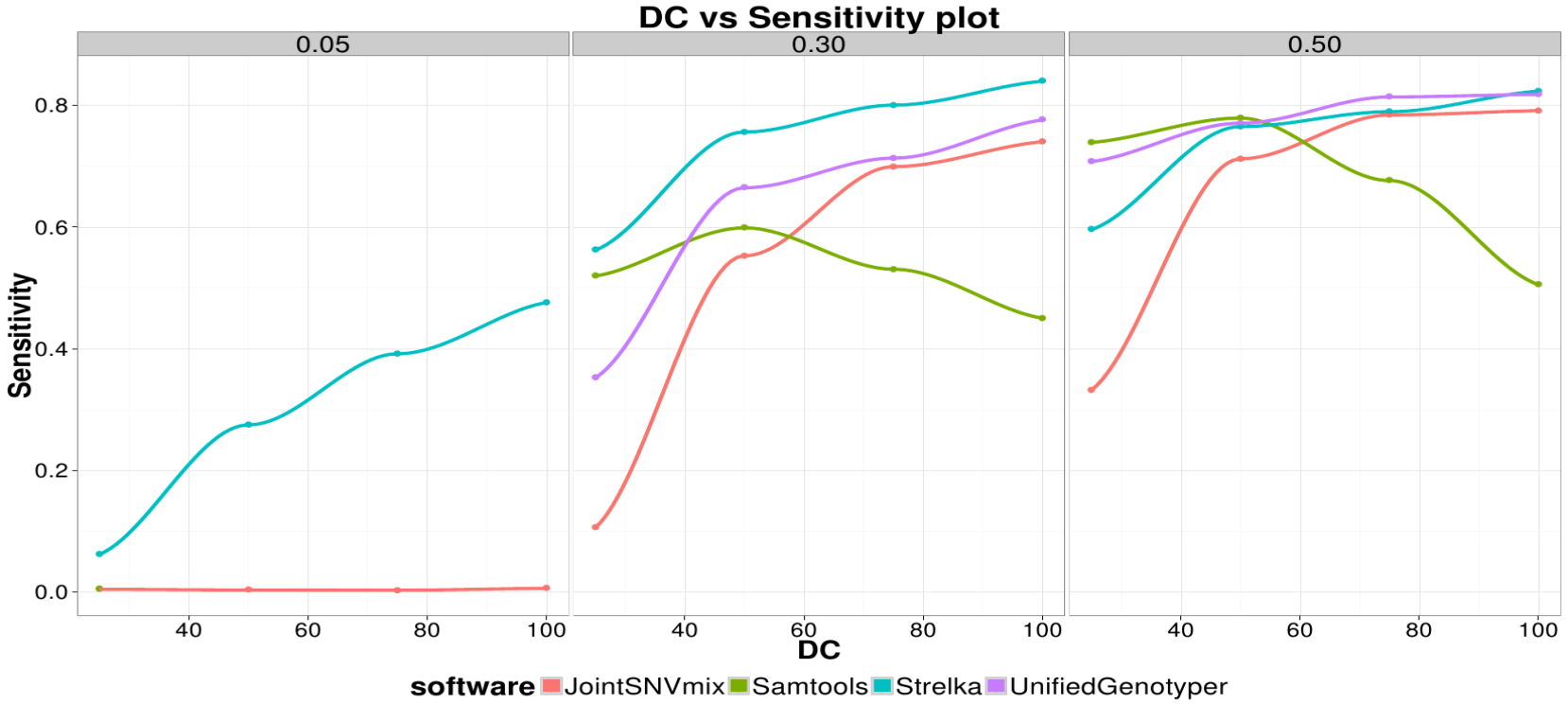
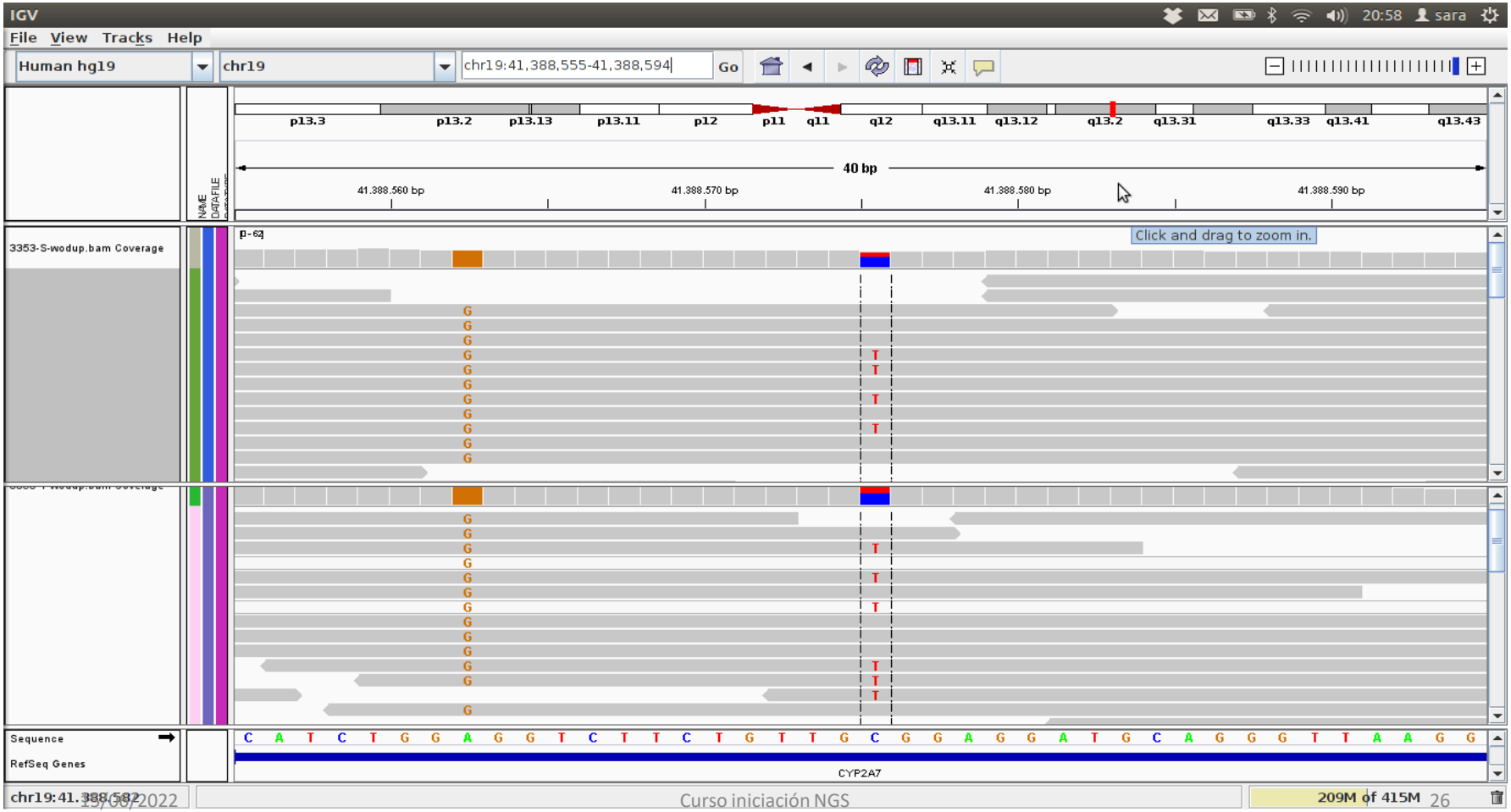


Figure.- Sensitivity as a function of sequencing depth of coverage. 0,50; 0.30; 0,05 three different AAF for mutations in the virtual tumors.

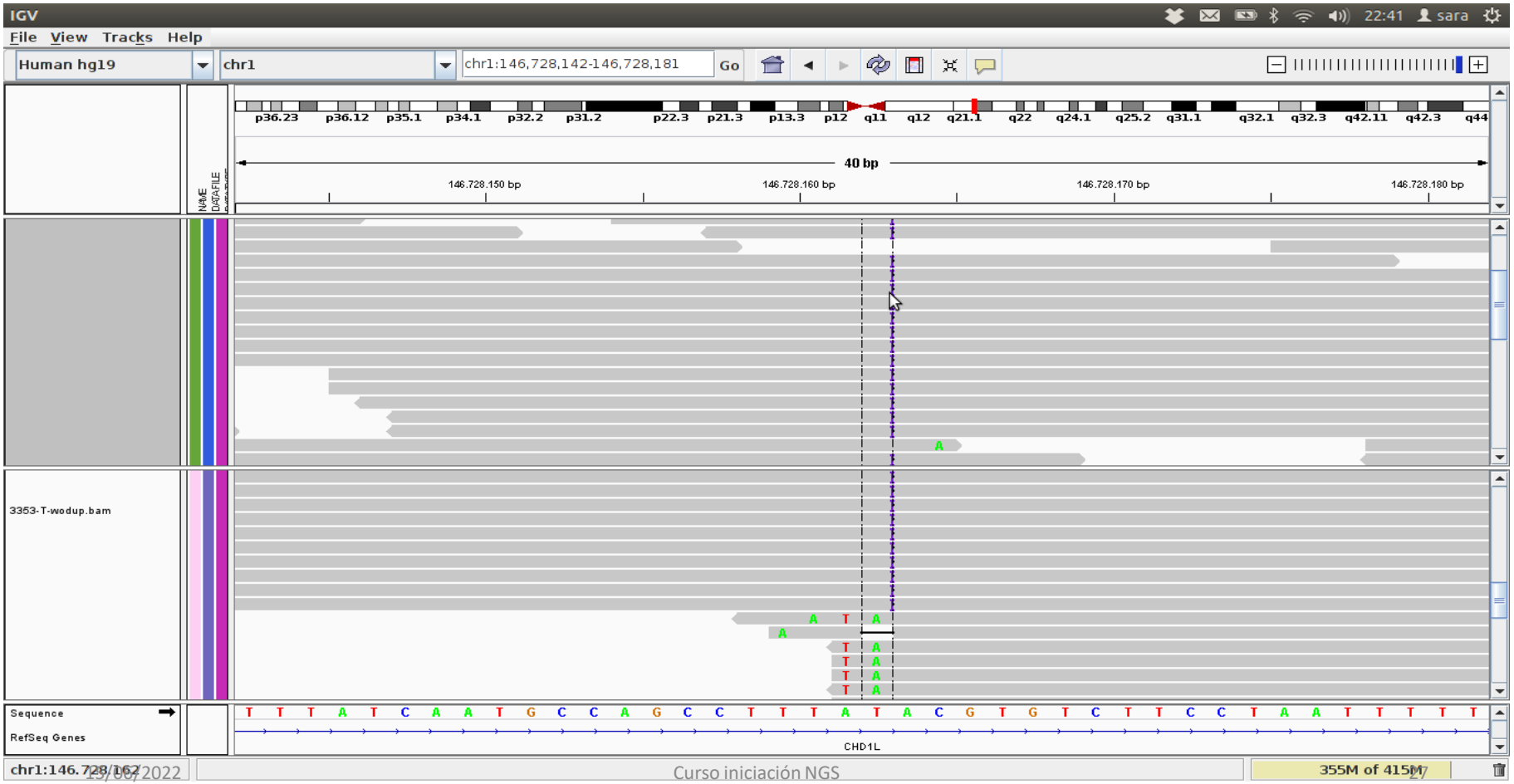
Ejemplo de variant calling: Cáncer

- Problemas de no detectar verdaderas mutaciones somáticas



Ejemplo de variant calling: Cáncer

- Problemas de realineamiento



Ejemplo de variant calling: Cáncer

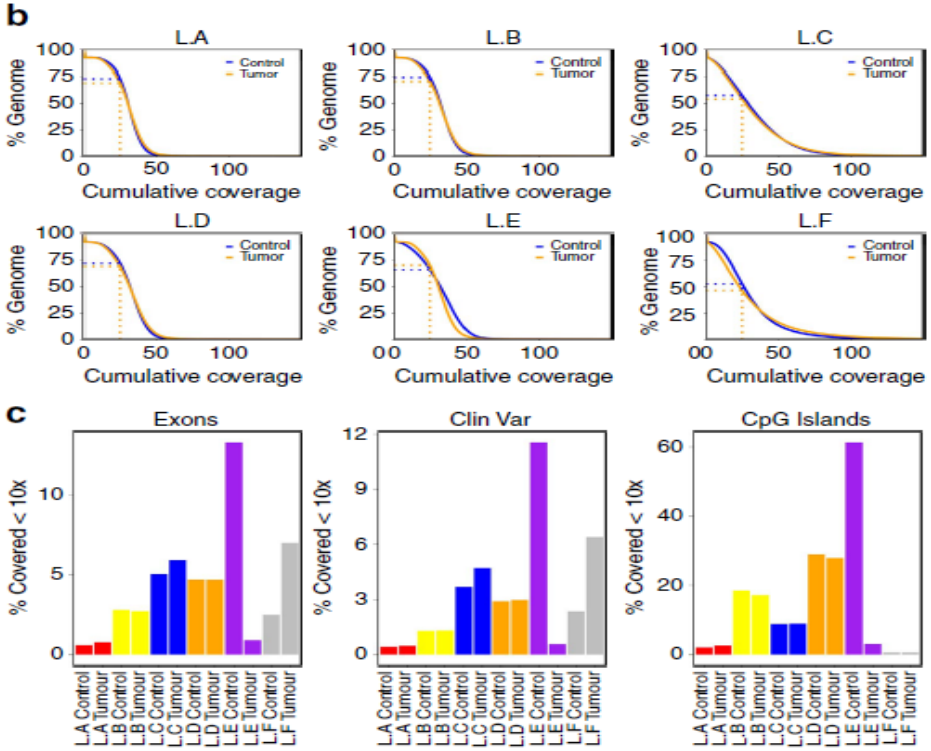
- Distinta llamada dependiendo del software que se utilice

Samtools	GATK	Strelka	JointSNVMix
13_48941648_C/T	6_33161869_G/C	1_17571938_C/T	1_146728162_T/A
22_50312951_A/G	6_36569717_C/T	1_22148740_C/T	11_110036661_T/C
5_72374150_C/T	9_33798052_C/T	6_2623313_T/C	13_48941648_C/T
10_127434402_C/T	9_33798633_C/T	6_123130685_C/T	2_198267775_G/A
2_54152643_C/A	9_99702631_T/C	6_136664990_A/T	20_22542571_A/G
6_32627812_C/G	9_140773503_C/G	10_46999030_C/T	22_50312951_A/G
20_26062036_A/G	12_392976_A/T	10_86237317_GT	3_46751098_A/G
17_79104962_G/A	12_131649643_G/A	10_127434402_C/T	
	13_110407574_C/T	11_117789327_T/C	
	17_38175428_A/C	13_48941648_C/T	
	19_54378472_G/T	18_29709508_T/A	
	19_56687467_C/G	22_28146770_A/G	
	22_21045452_C/T	22_50312951_A/G	
	22_50312951_A/G		

Ejemplo de variant calling: Cáncer

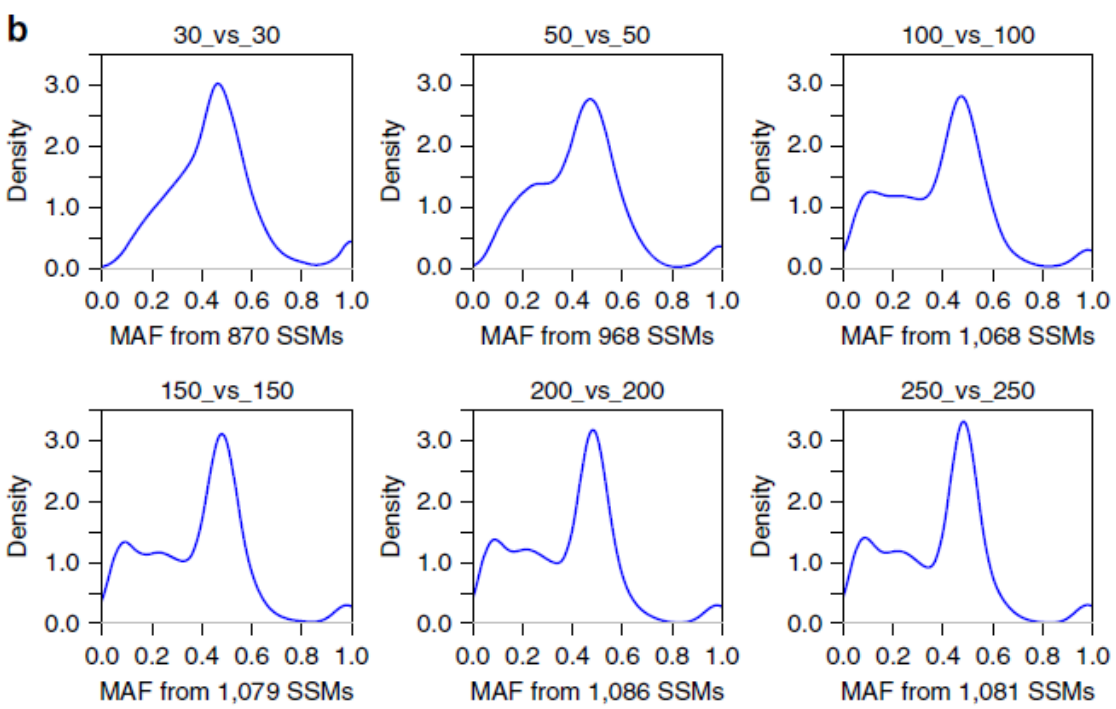
Table 1 | Summary of medulloblastoma tumour-normal pair library construction and sequencing.

Library	Starting DNA (μg)	Fragment Size (bp)	Size selection	Library protocol	PCR cycles	Sequencing machine	Chemistry (Illumina)	Depth (×) control:tumour
L.A	4	~400	2% Agarose gel	KapaBio	0	HiSeq 2500 HiSeq 2000 MiSeq	V1 (RR) V3 V2	29.6 : 40.5
L.B	1	~400	2% Agarose gel, Invitrogen E-gel	TrueSeq DNA	10			
L.C	2.5	~500	2% Agarose gel	NEBNext	12			
L.D	1	~550	Agarose gel	TrueSeq DNA	10			
L.E	2.8	~620	1.5% Agarose gel pippin	NEBNext	0			
L.F	1	~400	AMPureXP beads	NEBDNA	10			
L.G	1	~350	AMPureXP beads	TrueSeq DNA PCR-Free	0			
L.H	0.5	~175	AMPureXP beads	SureSelect WGS	10			



Alioto et al. Nature Communications 2015

Ejemplo de variant calling: Cáncer



Contrariamente a lo que se piensa identificar variantes somáticas de datos WGS es todavía un gran reto.

Alioto et al. Nature Communications 2015

Ejemplo de variant calling: Cáncer

Table 3 | Summary of accuracy measures.

SSM calls	Aligner	SSM Detection Software	TP	FP	FN	P	R	F1
MB.GOLD	BWA, GEM	Curated	1,255 (8)	0	0	1.00	1.00	1.00
MB.A	BWA	In-house	775 (0)	147	480	0.84	0.62	0.71
MB.B	BWA	samtools, Varscan	788 (1)	12	467	0.99	0.63	0.77
MB.C	GEM	samtools, bcftools	766 (3)	1,025	489	0.43	0.61	0.50
MB.D	n.a.	SMuFin	737 (4)	1,086	518	0.41	0.59	0.48
MB.E	BWA	SomaticSniper	750 (4)	229	505	0.77	0.60	0.67
MB.F	BWA	Strelka	884 (2)	165	371	0.84	0.70	0.77
MB.G	BWA	Caveman, Picnic	899 (3)	140	356	0.87	0.72	0.78
MB.H	Novoalign	MuTect	947 (3)	6,296	308	0.13	**0.76**	0.22
MB.I	BWA	samtools	879 (7)	129	376	0.87	0.70	0.78
MB.J	None, BWA	SGA + freebayes	856 (1)	62	399	0.93	0.68	**0.79**
MB.K	BWA	Atlas2-snp	945 (8)	7,923	310	0.11	0.75	0.19
MB.L1	BWA	MuTect, Strelka	385 (0)	3	870	**0.99**	0.31	0.47
MB.L2	BWA	MuTect, Strelka	900 (1)	253	355	0.78	0.72	0.75
MB.M	BWA mem	samtools, GATK + MuTect	937 (4)	1,695	318	0.36	0.75	0.48
MB.N	BWA	Strelka	847 (1)	289	408	0.75	0.68	0.71
MB.O	BWA	MuTect	944 (3)	272	311	0.78	0.75	0.76
MB.P	BWA	Sidron	833 (3)	256	422	0.77	0.66	0.71
MB.Q	BWA	qSNP + GATK	842 (2)	25	413	0.97	0.67	**0.79**
SIM calls								
MB.GOLD	BWA, GEM	Curated	337 (10)	0	0	1.00	1.00	1.00
MB.A	BWA	In-house	16 (0)	63	321	0.20	0.05	0.08
MB.B	BWA	GATK SomaticIndelDetector, Varscan	167 (0)	20	173	0.89	0.49	0.63
MB.C	GEM	samtools, bcftools	103 (0)	26	236	0.80	0.30	0.44
MB.D	none	SMuFin	29 (0)	25	308	0.54	0.09	0.15
MB.F	BWA	Strelka	147 (8)	12	193	0.93	0.43	0.58
MB.G	BWA	Pindel	189 (2)	82	152	0.70	0.55	0.61
MB.H	Novoalign	VarScan2	55 (0)	248	282	0.18	0.16	0.17
MB.I	BWA	Platypus	271 (7)	224	70	0.55	**0.79**	**0.65**
MB.J	None	SGA	90 (1)	34	249	0.72	0.26	0.38
MB.K	BWA	Atlas2-indel	268 (6)	444	72	0.38	0.79	0.51
MB.L1	BWA	Strelka	64 (1)	3	273	**0.96**	0.19	0.32
MB.L2	BWA	Strelka	130 (3)	13	210	0.91	0.38	0.53
MB.N	BWA	Strelka	128 (6)	16	209	0.89	0.38	0.53
MB.O	BWA	GATK SomaticIndelDetector	140 (1)	47	197	0.75	0.42	0.53
MB.P	BWA	bcftools, PolyFilter	37 (0)	57	301	0.39	0.11	0.17
MB.Q	BWA	Pindel	100 (2)	61	237	0.63	0.30	0.40

F1, F1 score; FN, false negative; FP, false positives; P, precision; R, recall; TP, true positives.
Shown are the evaluation results with respect to the medulloblastoma Gold Set (Tier 3). Shown are the number of true calls (TP) with additional Tier 4 calls in parentheses, the number of FP, the number of FN, P, R and F1. The submissions with the best precision, recall and F1 score are in bold.

Alioto et al Nature Communications 2015

- Llamada a variantes con diferentes pipelines y datos de diferentes librerías da lugar a un bajo consenso.
- Checklist para estudios WGS en cáncer:
  - Preparar librería PCR-free
  - Tumor coverage 100x
  - Control coverage close to tumor coverage (+/-10%)
  - Reference genome hs37d5 o GRCh38
  - Combinación de alineador/variant caller optimo
  - Combinar varios llamadores de mutaciones
  - Permitir mutaciones en zonas repetidas o cerca de repeticiones.
  - Filtrado por calidad de mapado, strand bias, positional bias, presencia de soft-clipping

Ejemplo de variant calling: TRIOS

- Formato: fastq
- Plataforma: HiSeq Illumina, 2x101
- Carreras: 1
- Lanes: 2 y 8
- Kit Enriquecimiento: TruSeq Exome Enrichment Kit (2011)

Pedigrí	Sexo	Afectado
Padre	V	N
Madre	M	N
Hijo	V	S

Ejemplo de variant calling: TRIOS

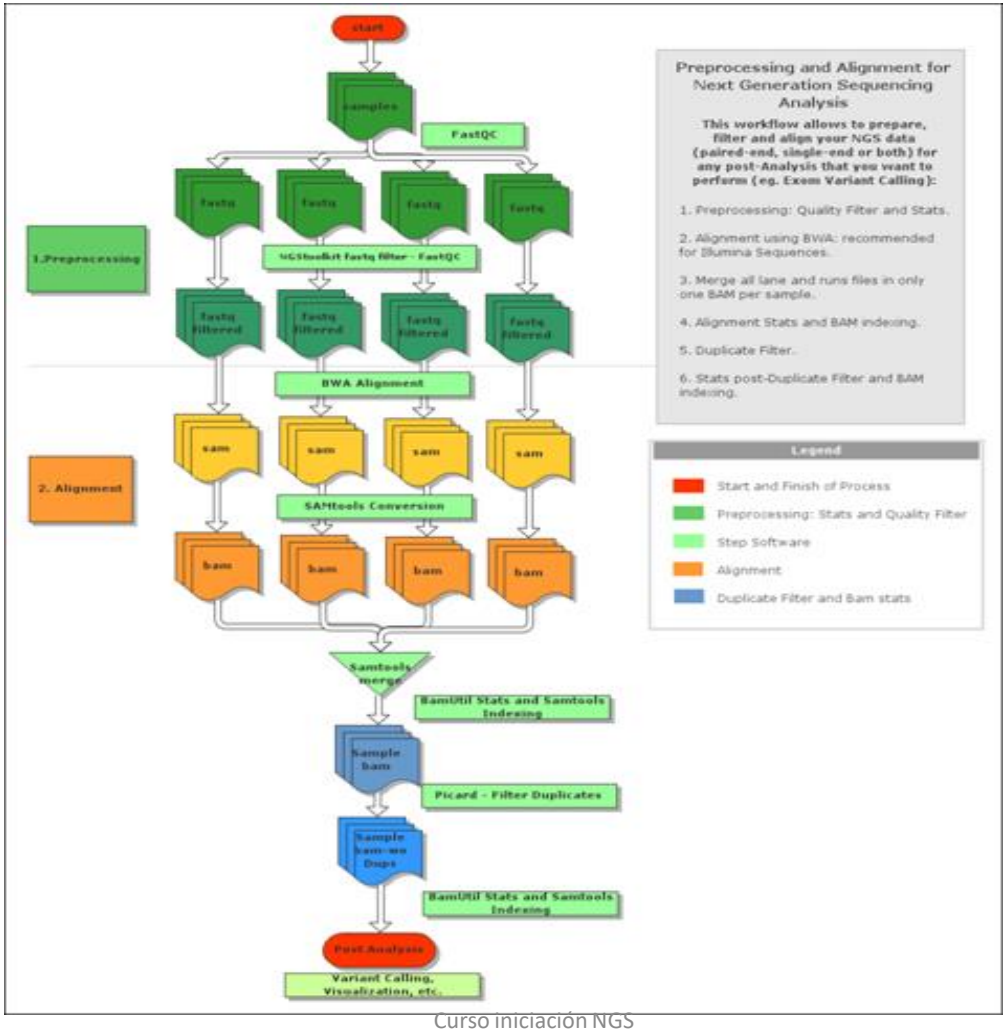
- Referencia:hg19/GRCh37 (versión de los 1000 genomas)
- Preprocesado:

  exome_pipeline.v2.0
- Alineamiento:

  exome_pipeline.v2.0 (BWA)
- Post-Análisis:

  GATK 3.5

Esquema análisis



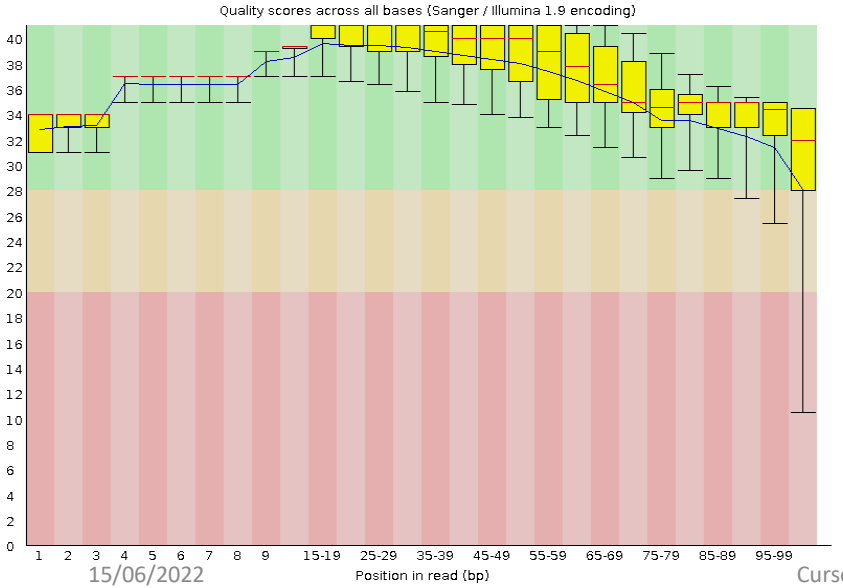
Preprocesamiento y QC

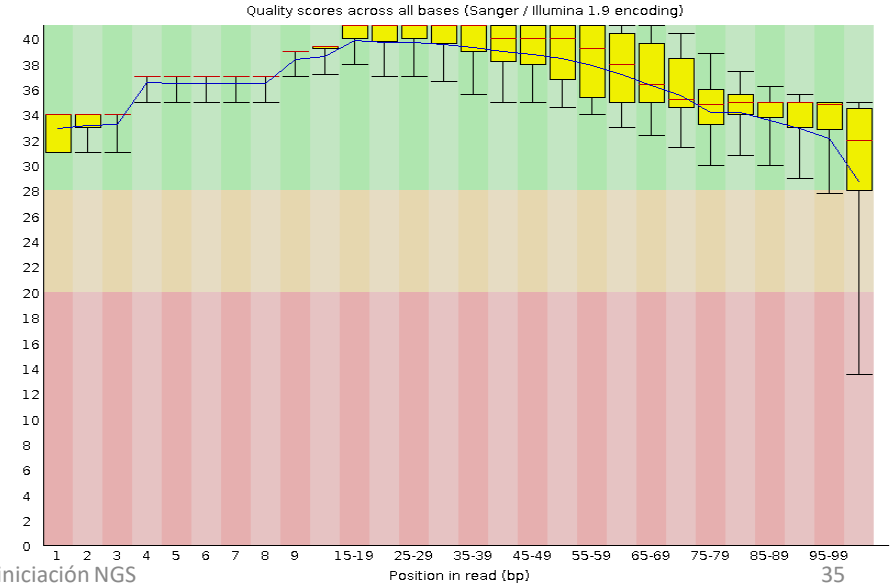
FASTQC

70% de los nucleótidos de la read > 20 calidad phred

No trimming

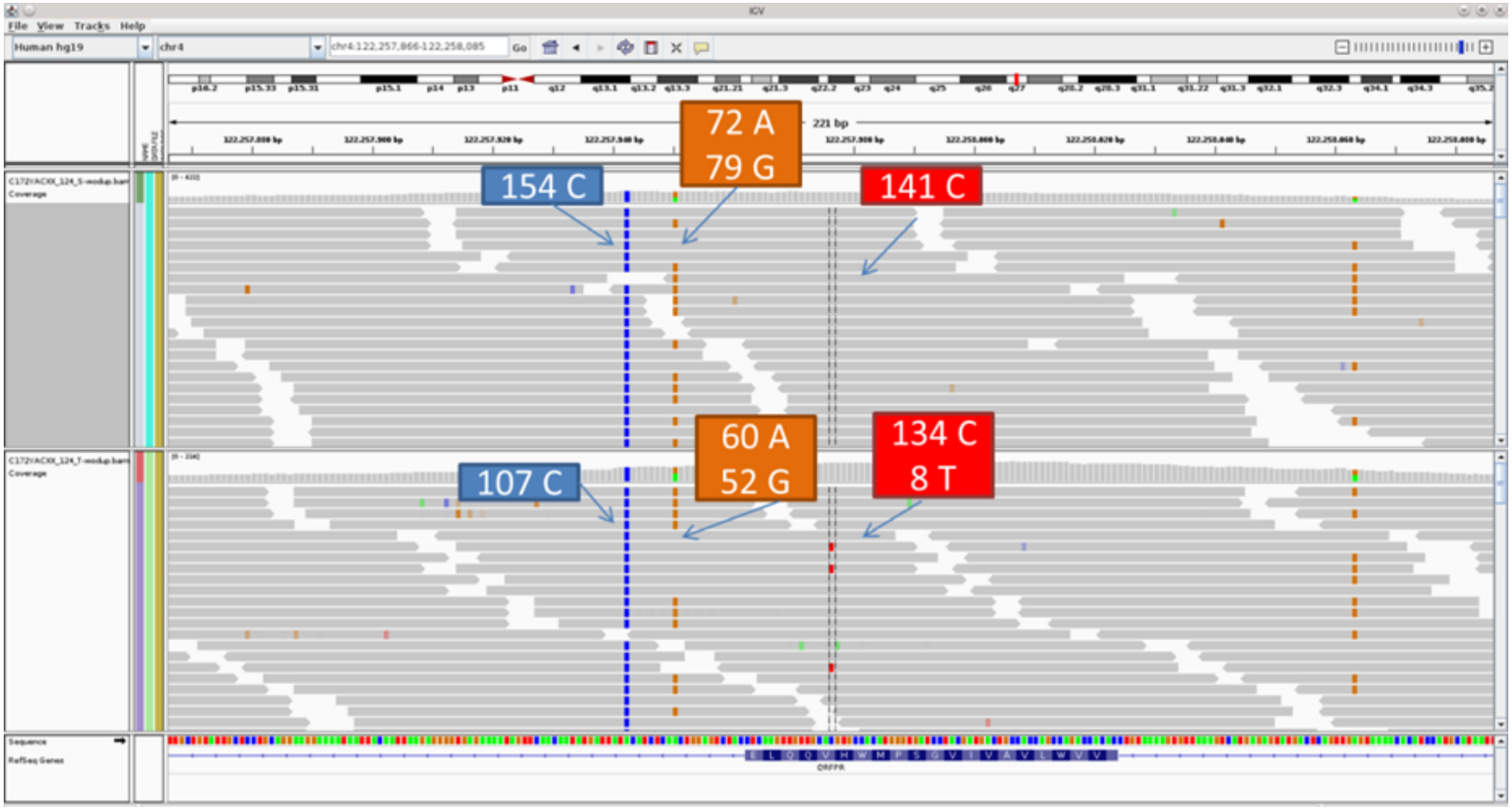
Sample	Sample_ND0011	Sample_ND0012	Sample_ND0013
Pre-Filter			
Sequence length	101	101	101
Total Sequences	60120812	47392782	38753830
#Total Duplicate Percentage	50.85	48.50	46.03
%GC	51	51	51
Post-Filter			
Sequence length	101	101	101
Total Sequences	57748786	45713509	37383698
#Total Duplicate Percentage	50.52	48.18	45.67
%GC	51	51	51





Alineamiento

- Versión 0.6.2 de BWA



Postprocesamiento y QC alineamiento

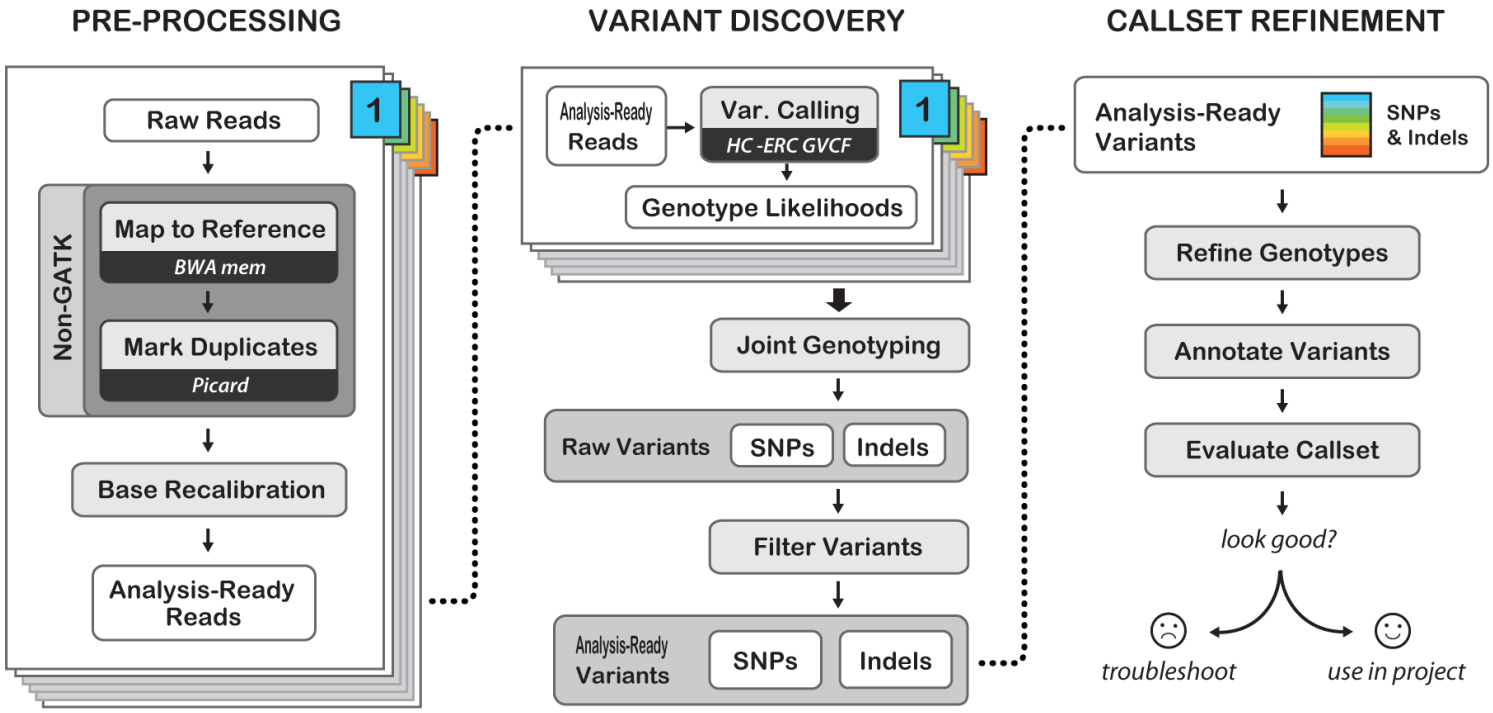
- Filtrado de duplicados: Picard
- Análisis de la calidad del BAM
- Recalibración de variantes
- Realineamiento

} GATK

Workflow de "Best Practices" instaurado por GATK. La documentación completa del framework se puede encontrar (http://www.broadinstitute.org/gatk/guide/best-practices, febrero 2014)

Sample	Target Specificity	Target Enrichment	Mean Coverage	SD Coverage	5X	10X	20X	30X
Padre	0.77	39.40	90.92	79.87	95%	93%	90%	85%
Madre	0.78	39.85	72.68	64.91	93%	92%	87%	81%
Hijo	0.78	39.78	60.46	53.12	93%	90%	84%	75%

Variant Calling



Best Practices for Germline SNPs and Indels in Whole Genomes and Exomes - June 2016

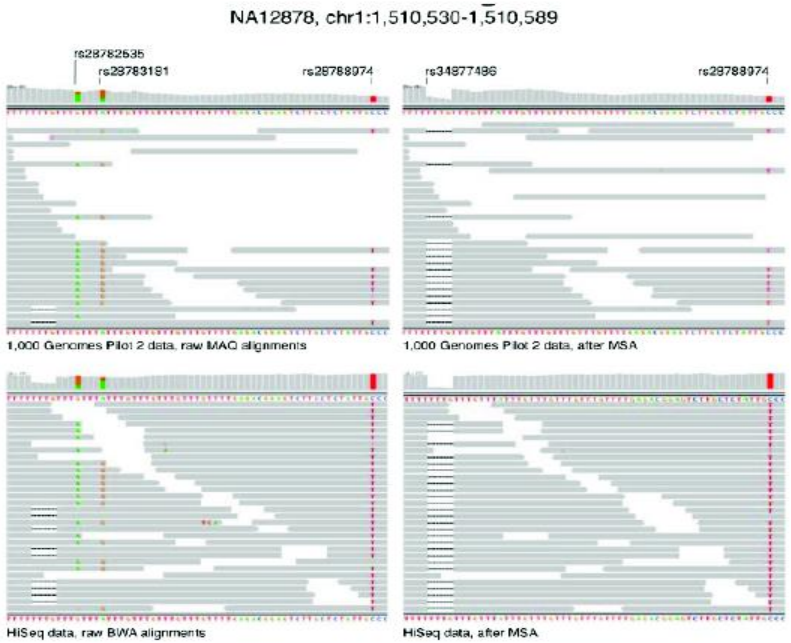
Variant Calling: Realineamiento

Realineamiento local de múltiples secuencias

Proporciona un alineamiento consistente entre todas las lecturas. Se identifican las regiones susceptibles de realineamiento, si:

- Al menos una lectura contiene un indel
- Existe un *cluster* de bases *mismatch*
- Existe un indel conocido

Before



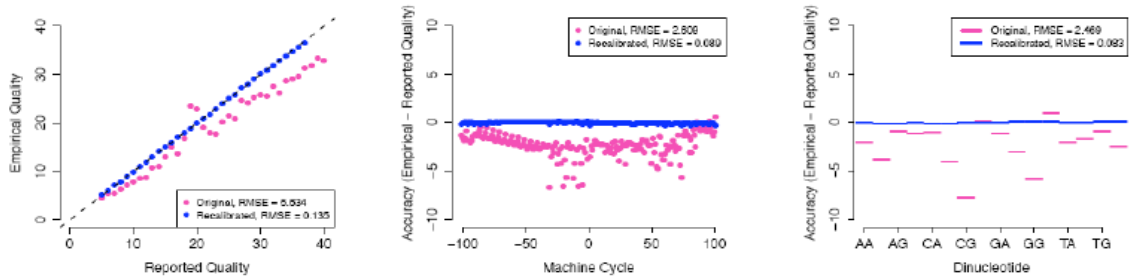
After

Variant Calling: Realineamiento

Score de calidad de una base

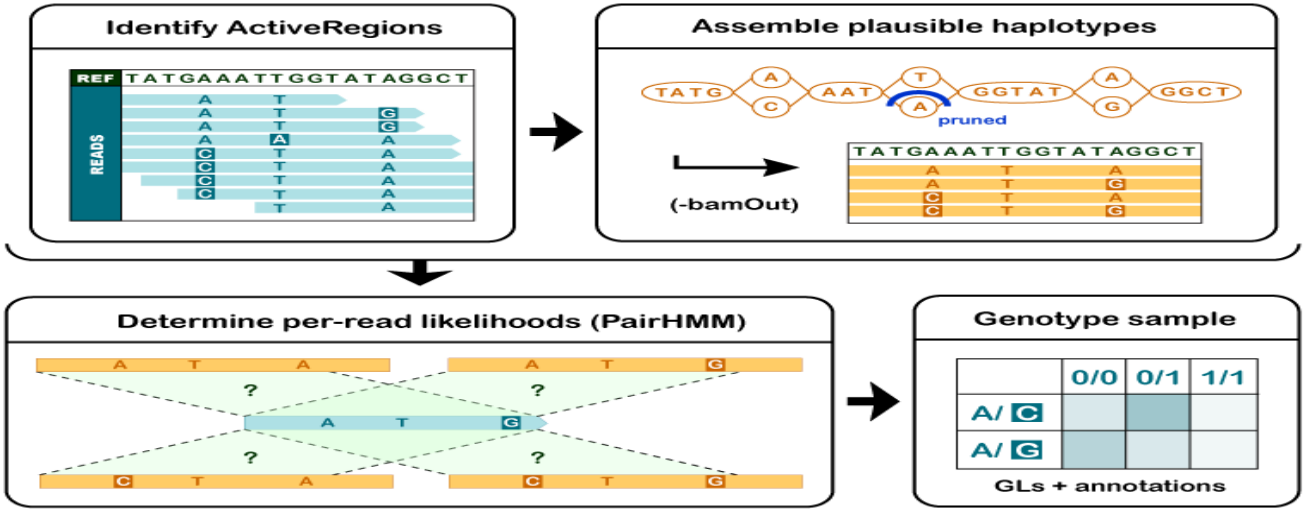
Probabilidad de que la base determinada sea verdadera (y no un error de secuenciación).

- En escala *Phred*. $Q = -10 \cdot \log_{10} P$
- Se códifica en ASCII (normalmente Q+33)
- Se estima de un modo muy inexacto porque sigue un esquema de correlación complejo entre la tecnología de secuenciación, el ciclo de máquina y el contexto de secuencia.



Los errores de mapado y la inexactitud de los scores de calidad se propagan a la etapa de identificación de variantes y genotipado

Variant Calling: HaplotypeCaller



- Determina si una región es potencialmente variable
- Construye un ensamblado de Bruijn de la región.
- Los "paths" en el grafo son haplotipos potenciales que tienen que ser evaluados.
- Se calcula los likelihoods de los haplotipos dados los datos usando un modelo PairHMM.
- Determina si hay alguna variante entre los haplotipos más probables.
- Calcula la distribución de la frecuencia alélica para determinar el contaje de alelos más probables y emite una variante si se da el caso.
- Si se emite una variante se calcula el genotipo para cada muestra.

Variant Calling: Filtrado 1

- MQ0 >= 4 && ((MQ0 / (1.0 * DP)) > 0.1): las variantes que cumplen esta regla se marcan con el filtro HARD_TO_VALIDATE. MQ0: Total Mapping Quality Zero Reads
- DP <5: LowCoverage.
- QUAL <30.0: VeryLowQual
- QUAL >30.0 && QUAL <50.0: LowQual
- QD <1.5: LowQD. QD: Variant Confidence/Quality by Depth
- SB >-10.0: StrandBias
- FS >60.0: p-value StrandBias. FS: Phred-scaled p-value using Fisher's exact test to detect strand bias
- HaplotypeScore >13.0: HaplotypeScore: Consistency of the site with at most two segregating haplotypes

Variant Calling: Refinamiento de Genotipos

- **PhaseByTransmission:** técnica estadística para determinar en el niño, cuándo es posible, qué alelo proviene del padre y cuál de la madre. Por consenso se pone primero el alelo de la madre y luego el del padre. Madre 1/0 padre 0/0, niño 1(madre) | 0(padre).
- **ReadBackedPhasing:** determina la presencia de haplotipos en cada muestra, no entre ellas. Busca grupos de SNPs que se encuentran en el mismo cromosoma. Al correr este Walker en el vcf se marcan los genotipos con la | en vez de con / cuando se ha conseguido determinar un haplotipo.

Original VCF

#CHROM	POS	ID	REF	ALT	QUAL	FILTER	INFO	FORMAT	MOTHER	FATHER	CHILD
1	10109	.	A	T	99	PASS	.	GT:PL	0/0:0,50,200	0/0:0,40,200	0/1:30,0,200
1	10147	.	C	A	99	PASS	.	GT:PL	**0/1**:0,30,200	**0/0**:0,50,200	**0/1**:200,40,0
1	10150	.	C	T	99	PASS	.	GT:PL	0/1:0,40,200	0/1:30,0,200	1/1:200,50,0

Phased VCF

#CHROM	POS	ID	REF	ALT	QUAL	FILTER	INFO	FORMAT	MOTHER	FATHER	CHILD
1	10109	.	A	T	99	PASS	.	GT:PL:TP	0|0:0,50,200:10	0|0:0,40,200:10	0|0:30,0,200:10
1	10147	.	C	A	99	PASS	.	GT:PL:TP	**1|0**:0,30,200:10	**0|0**:0,50,200:10	**1|0**:200,40,0:10
1	10150	.	C	T	99	PASS	.	GT:PL:TP	1|0:0,40,200:10	1|0:30,0,200:10	1|1:200,50,0:10

The convention is:
Allele From Mother | Allele From Father

Anotación y Filtrado

- Para el proceso de anotación y filtrado se utiliza el software KGGSeq. Se establecen los siguientes parámetros de filtrado además los ya establecidos en GATK:
    - Ignora missing genotypes.
    - Ignora genotipos de baja calidad < 4.
    - Ignora genotipos debido a que la fracción de los reads que lleva el alelo alternativo es >=0.05 en un genotipo homocigoto 0/0 y que es <=0.25 en un genotipo heterocigoto 1/0 y que es <=0.75 en un homocigoto 1/1.
    - Ignora genotipos debido a que el segundo Phred Scaled likelihood (PL) es < 20.
    - Ignora variantes con calidad de secuenciación < 50.0
    - Ignora variantes debido a que ningún sujeto tiene genotipos válidos después del QC.

Anotación y filtrado

- Anotación:
  - A nivel de gen: se anota gen y “feature” según la base de datos refgene (variante tipo missense, frameshit, intron, etc.)
  - Anotación de variantes no sinónimas: dbNSFP

    - SLR
    - SIFT
    - Polyphen2_HDIV
    - Polyphen2_HVAR
    - LRT
    - Mutation Taster
    - Mutation Assesor
    - FATHMM_score
    - CADD_score
    - GERP++_NR
    - GERP++_RS
    - PhyloP100way_vertebrate
    - 29way_logOdds

  - A nivel funcional: pseudogenes,UniprotFeature,etc.
  - A nivel de enfermedad: anotación de enfermedad asociada con ese gen en OMIM

Estadísticas del filtrado

Variants	Raw	HardFiltering*	GenotypeRefinement	Quality Filtering
SNPs	294018	189031	188910	177660
INDELs	40677	27695	26646	

*Siendo este número aquellas variantes marcadas como PASS después del filtrado.

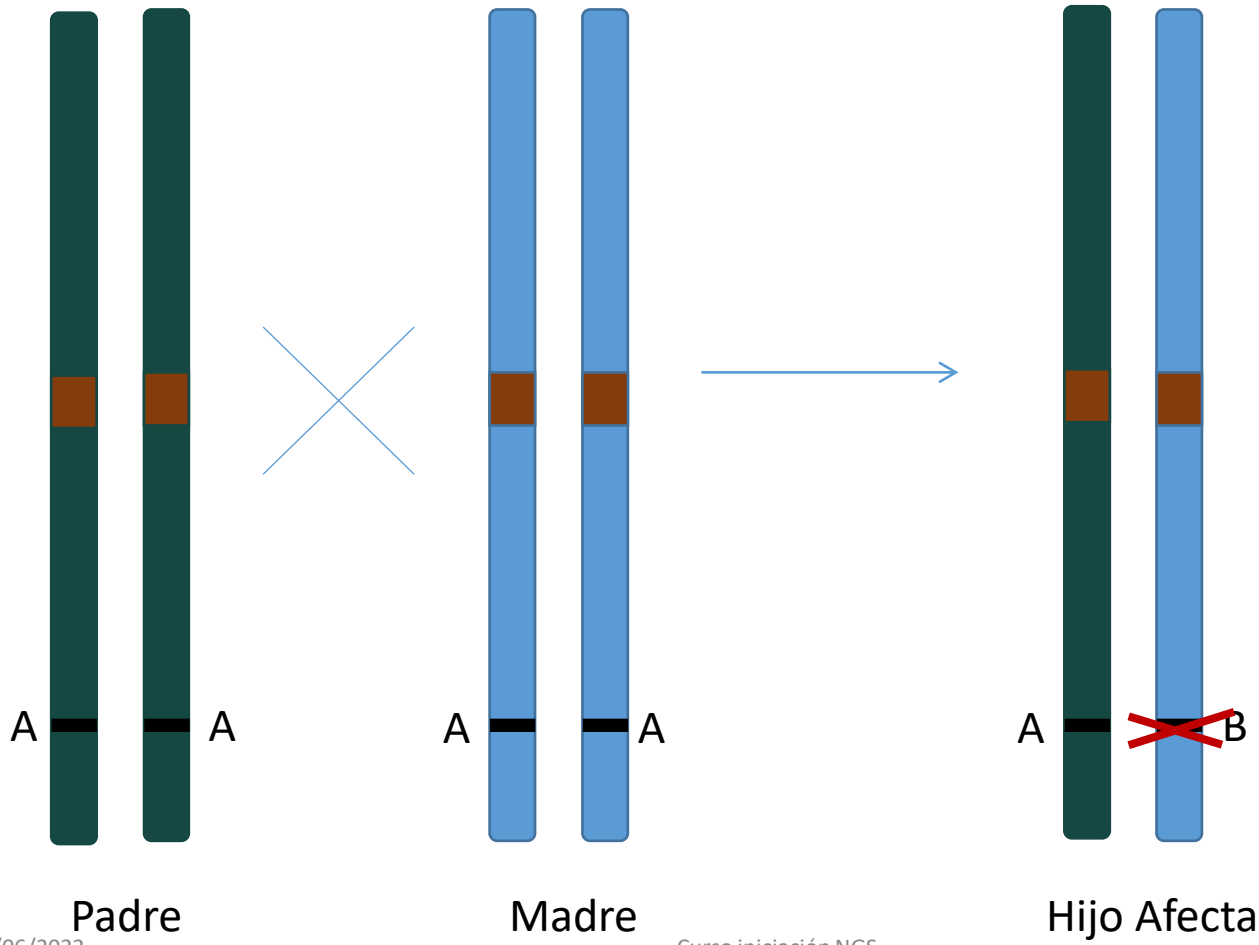
Modelos de Enfermedad

- Modelo de novo
- Modelo double-hit gene
- Modelo Recesivo
- Modelo dominante

Seleccionamos estos dos como los más probables en nuestro caso.

Modelos de enfermedad

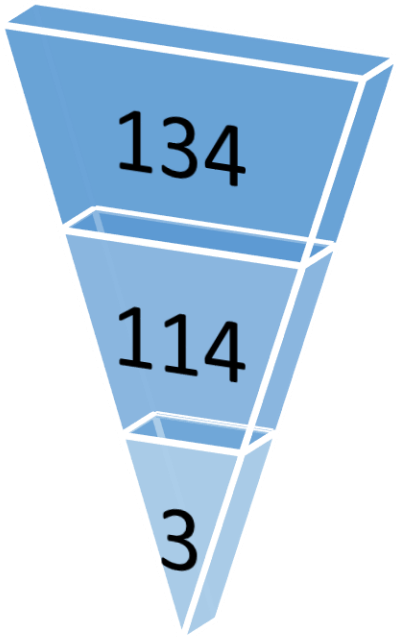
- Modelo de novo



Modelo de novo

- Filtros ad-hoc
  - Primera aproximación:
    - FILTER = PASS
    - Genotipo =

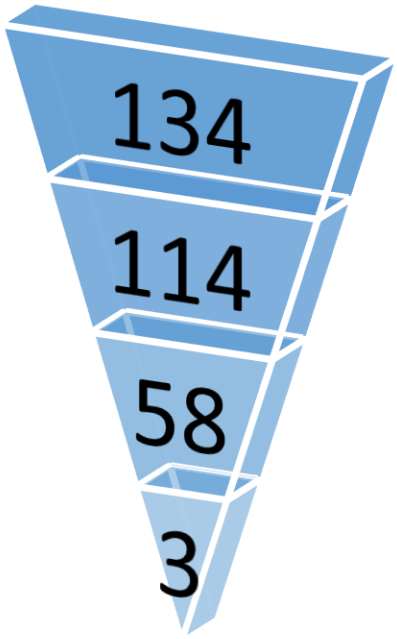
Padre	0/0
Madre	0/0
Hijo	0/1 o 1/0



Modelo de novo

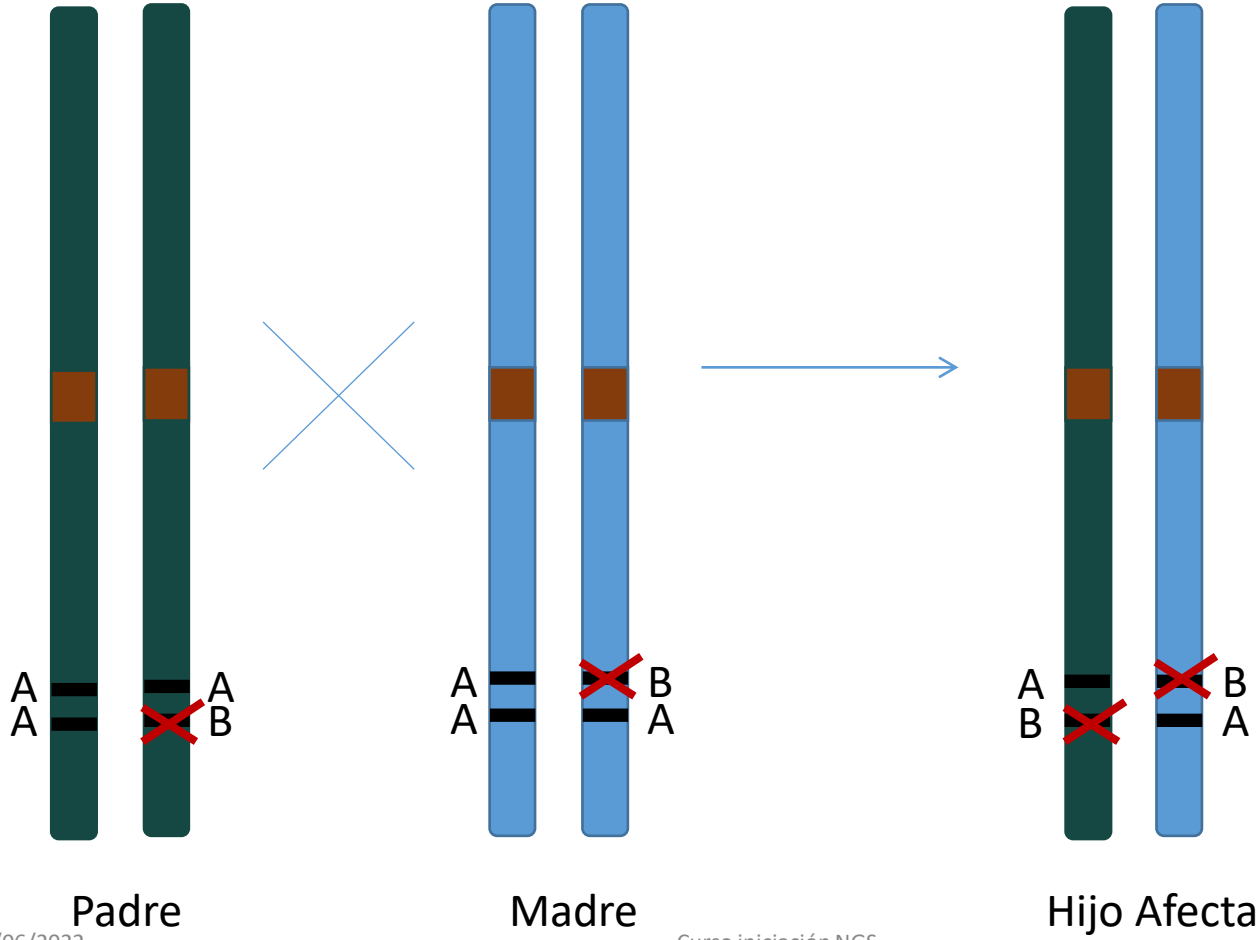
- Filtros ad-hoc
  - Segunda aproximación:
    - FILTER = PASS
    - Genotipo =

Padre	x
Madre	x
Hijo	!0/0



Modelos de enfermedad

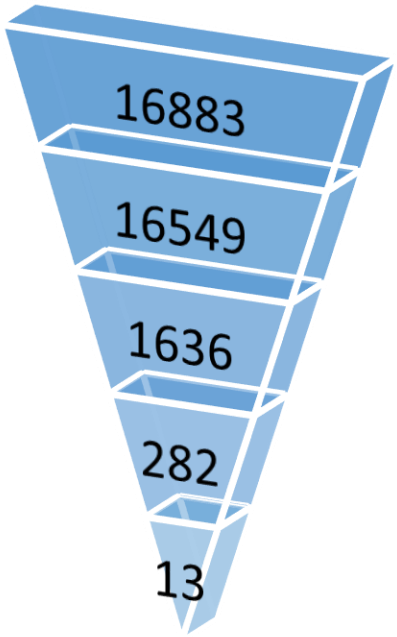
- Modelo double-hit gene



Modelo double-hit gene

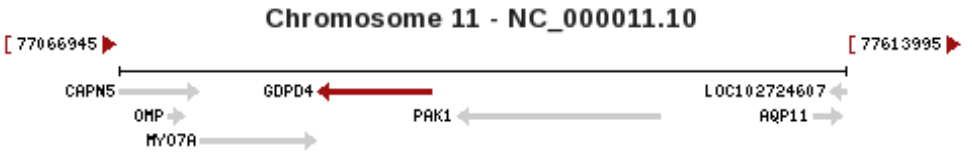
- Filtros ad-hoc
  - Primera aproximación:
    - FILTER = PASS
    - Filtro por missense,frameshift, splicing o stop-gain
    - Mutation taster: A o D
    - Genotipo =

Padre	0/1
Madre	0/1
Hijo	1/1



Modelo double-hit gene

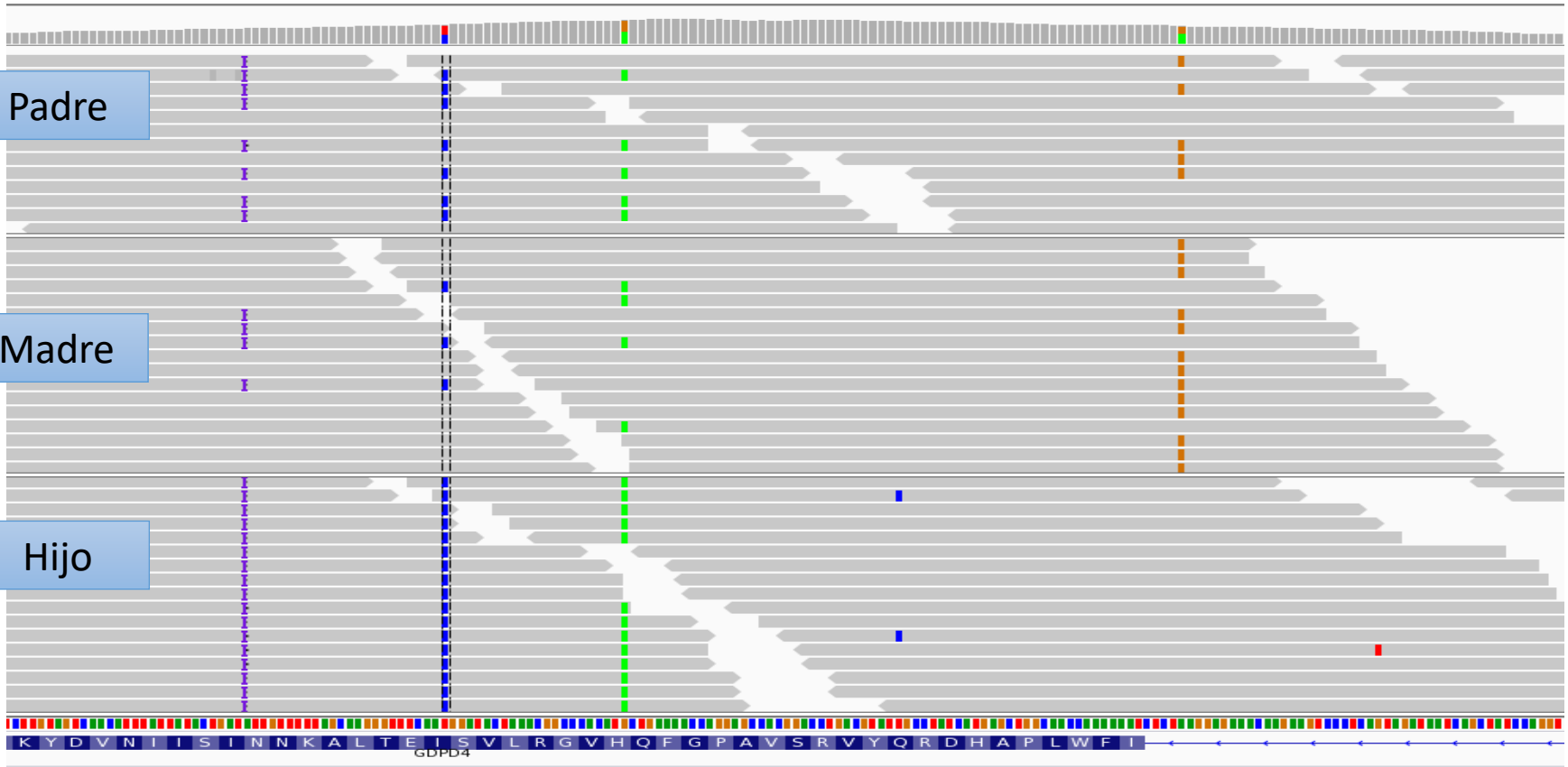
GDPD4



- Glycerophosphodiester Phosphodiesterase domain-containing 4
- Proteína de membrana
- Relacionada con el metabolismo de glicerofosfolípidos.
- Relacionado mutaciones en este gen con el síndrome del shock tóxico (TSS).
- Variantes vistas en el gen:
  - Delección patogénica en el cromosoma 11 71680927-7794394
  - Relacionado con retraso en el desarrollo y fenotipos morfológicos significativos.

Modelo double-hit gene

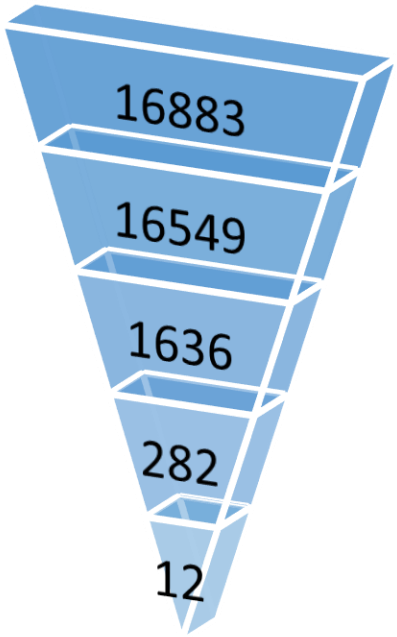
GDPD4



Modelo double-hit gene

- Filtros ad-hoc
  - Segunda aproximación:
    - FILTER = PASS
    - Filtro por missense,frameshift, splicing o stop-gain
    - Mutation taster: A o D
    - Genotipo =

Padre	1/0	0/0
Madre	0/0	1/0
Hijo	0/1	0/1



Modelo double-hit gene

NBPF1

Chromosome	StartPosition	Reference AlternativeAllele	rsID	MostImportantFeatureGene	MostImportantGeneFeature	GeneDescription
1	16890484	G/C	rs12117084 (suspected)	NBPF1	missense (cys -> ser)	neuroblastoma breakpoint family, member 1 (Approved)
1	16909134	G/A	.	NBPF1	missense	

Mutación en el 5% de los reads
En madre e hijo. Posibles errores de alineamiento por tratarse de genes duplicados.

- Gen de la familia “breakpoint” de neuroblastoma. Docenas de genes duplicados localizados en duplicaciones segmentales en el cromosoma 1.
- Cambios en el número de copia se ha relacionado con enfermedades del desarrollo y neurogenéticas como microcefalia, macrocephalia, autismo, retraso mental, neuroblastoma, enfermedades del corazón congénitas, etc.

Conclusiones

- Resultados dependientes del filtrado.
- Muchas variantes. Importancia del número de casos para poder acotar.
- Resultados probables pero difícil demostración funcional.
- Necesario ayuda de la parte clínica para poder unir las mutaciones resultantes con el fenotipo causante de la enfermedad.
- Necesario establecer protocolos y estándares de calidad en todos los puntos del proceso.

Creación de estándares

College of American Pathologists' Laboratory Standards for Next-Generation Sequencing Clinical Tests

Nazneen Aziz, PhD; Qin Zhao, PhD; Lynn Bry, MD, PhD; Denise K. Driscoll, MS, MT(ASCP)SBB; Birgit Funke, PhD; Jane S. Gibson, PhD; Wayne W. Grody, MD; Madhuri R. Hegde, PhD; Gerald A. Hoeltge, MD; Debra G. B. Leonard, MD, PhD; Jason D. Merker, MD, PhD; Rakesh Nagarajan, MD, PhD; Linda A. Palicki, MT(ASCP); Ryan S. Robetorye, MD; Iris Schrijver, MD; Karen E. Weck, MD; Karl V. Voelkerding, MD

- Recomendaciones en documentación, trazabilidad, validación, almacenamiento,…
  - Extracción de ADN
  - Preparación de librerías
  - Referencias y versiones
  - Pipeline bioinformático de análisis

Creación de estándares

Interpretación de variantes.

Table 5

Rules for Combining Criteria to Classify Sequence Variants

Pathogenic

1. 1 Very Strong (PVS1) *AND*
   - a. ≥1 Strong (PS1–PS4) *OR*
   - b. ≥2 Moderate (PM1–PM6) *OR*
   - c. 1 Moderate (PM1–PM6) and 1 Supporting (PP1–PP5) *OR*
   - d. ≥2 Supporting (PP1–PP5)
2. ≥2 Strong (PS1–PS4) *OR*
3. 1 Strong (PS1–PS4) *AND*
   - a. ≥3 Moderate (PM1–PM6) *OR*
   - b. 2 Moderate (PM1–PM6) *AND* ≥2 Supporting (PP1–PP5) *OR*
   - c. 1 Moderate (PM1–PM6) *AND* ≥4 Supporting (PP1–PP5)

Likely Pathogenic

1. 1 Very Strong (PVS1) *AND* 1 Moderate (PM1–PM6) *OR*
2. 1 Strong (PS1–PS4) *AND* 1–2 Moderate (PM1–PM6) *OR*
3. 1 Strong (PS1–PS4) *AND* ≥2 Supporting (PP1–PP5) *OR*
4. ≥3 Moderate (PM1–PM6) *OR*
5. 2 Moderate (PM1–PM6) *AND* ≥2 Supporting (PP1–PP5) *OR*
6. 1 Moderate (PM1–PM6) *AND* ≥4 Supporting (PP1–PP5)

Benign

1. 1 Stand-Alone (BA1) *OR*
2. ≥2 Strong (BS1–BS4)

Likely Benign

1. 1 Strong (BS1–BS4) and 1 Supporting (BP1–BP7) *OR*
2. ≥2 Supporting (BP1–BP7)

* Variants should be classified as Uncertain Significance if other criteria are unmet or the criteria for benign and pathogenic are contradictory.

Standards and Guidelines for the Interpretation of sequence variants. American College of Medical Genetics and Genomics. Association for Molecular Pathology. 2015

Ejemplo de variant calling: Bacterias

- Identificación de Brotes de origen alimentario, “Crisis del Pepino”

2011

Mayo
- 24 Primera muerte en Alemania
- 26 Alemania acusa a los pepinos españoles
- 30 Prohibición de importaciones de verduras de España y Alemania
- 31 Laboratorios alemanes desmienten oficialmente que los pepinos españoles sean el foco de infección

Junio
- 10 Resolución de la crisis

Causado por la toxi-infección de Escherichia coli enterohemorrágica (EHEC) (*Escherichia coli* O104:H4)

Muerte: 32 personas en Alemania, 1 Suecia y 1 Francia y 2263 infectados en 12 países de Europa.

Crisis Política y Económica Europa: Alto impacto en la Economía Europea, mayor afectación en la Española

Secuenciación Genoma

华大基因 BGI

Universitätsklinikum Hamburg-Eppendorf

Ejemplo de variant calling: Bacterias

- Identificación de Brotes de origen alimentario, “Crisis del Pepino”

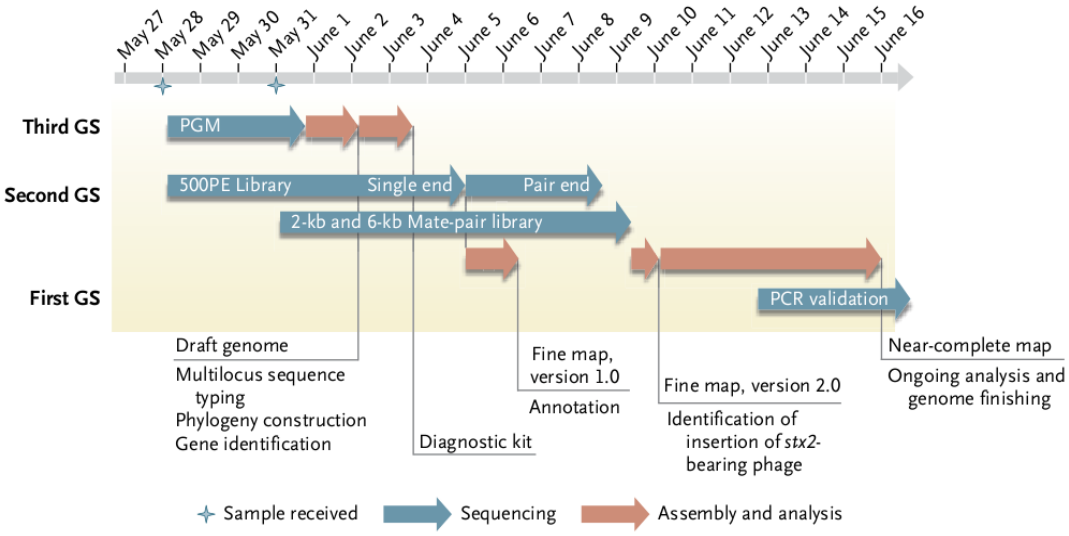


Figure 1. Timeline of the Open-Source Genomics Program.

After receiving the first batch of DNA samples on May 28, 2011, sequencing runs with the use of the Ion Torrent Personal Genome Machine (PGM) and Illumina (small-insert library) were initiated simultaneously. On May 31, the second batch of DNA was received and used for Illumina large-insert sequencing. An assembly of the Ion Torrent reads was released on June 2, which enabled subsequent analyses (multilocus sequence typing, phylogenetic analysis, and genome comparisons). Errors in the Ion Torrent data were corrected with the use of later Illumina data, and a high-quality draft genome sequence was created. GS denotes generation of sequencing technology. The symbols at May 28 and May 31 in the timeline indicate the arrival of DNA samples.

Rohde et al NEJM 2011, 365:718-24

Ejemplo de variant calling: Bacterias

- Identificación de Brotes de origen alimentario, “Crisis del Pepino”

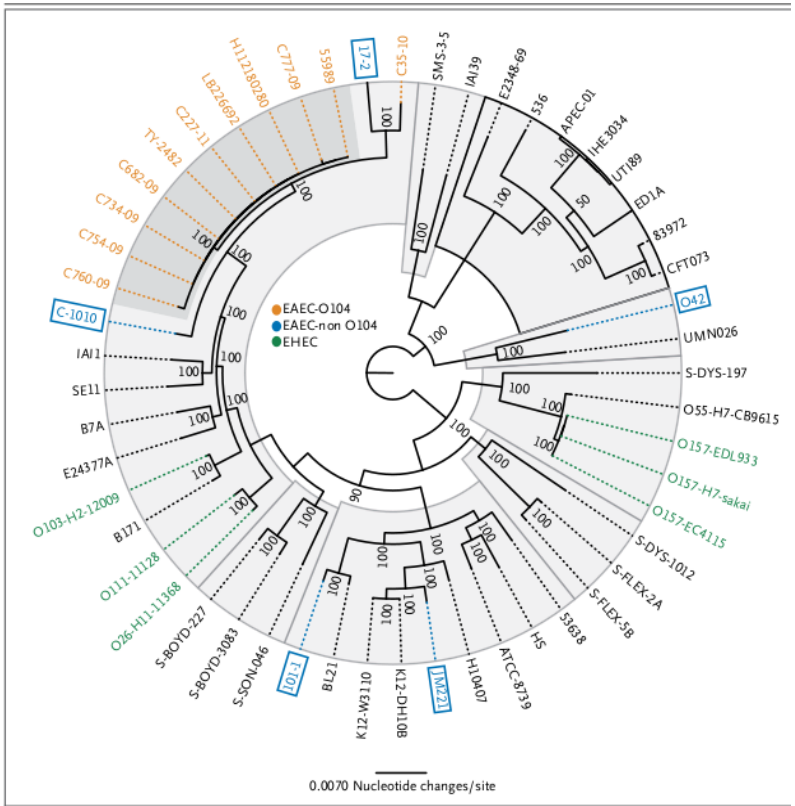
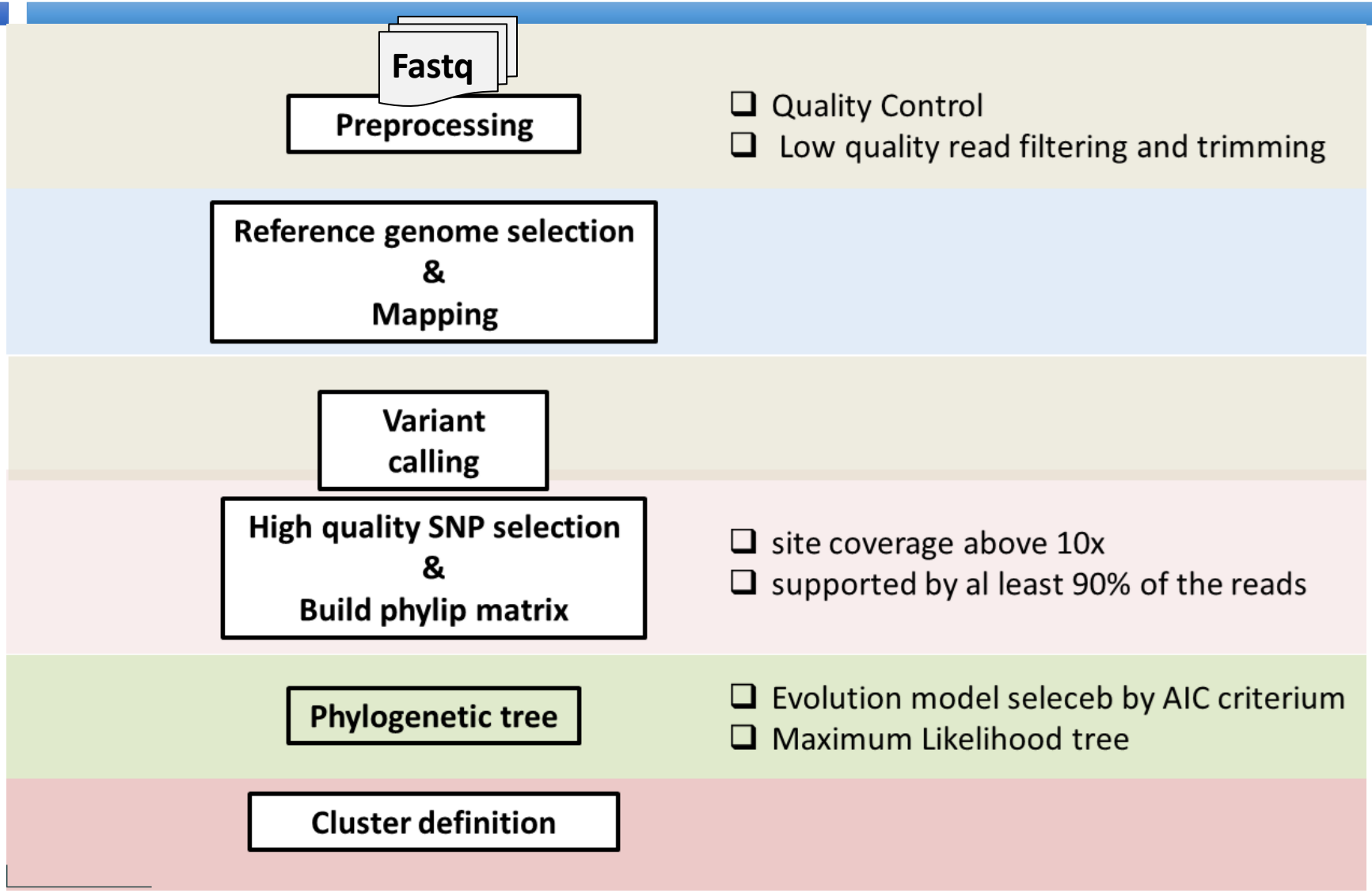


Figure 2. Phylogenetic Comparisons of 53 *Escherichia coli* and Shigella Isolates.

Genomic sequences were compared with the use of 100 bootstrap calculations, as described by Sahl et al.[35] The species-based phylogeny was inferred with the use of 2.56 Mbp of the conserved core genome. The O104:H4 isolates are shown in orange, the reference enteroaggregative *E. coli* (EAEC) isolates in blue, and the enterohemorrhagic *E. coli* isolates in green. (The classification of the other strains is shown in Fig. 4 and Table 4 in the Supplementary Appendix.) The O104:H4 isolates cluster into a single clade (dark gray); in contrast, the reference EAEC isolates are extremely divergent and are represented throughout the phylogeny.

Rohde et al NEJM 2011, 365:718-24

Pipeline variant calling: Bacterias



Software disponible

- CFSAN SNP Pipeline

Extracción de SNPs de alta calidad de aislados relacionados

http://snppipeline.readthedocs.io/en/latest/

- GATK, modo haploide
- Samtools
- Varscan
- Snippy

Identificación de variantes haploides y construcción de filogenia usando core genome SNPs

http://github.com/tseemann/snippy

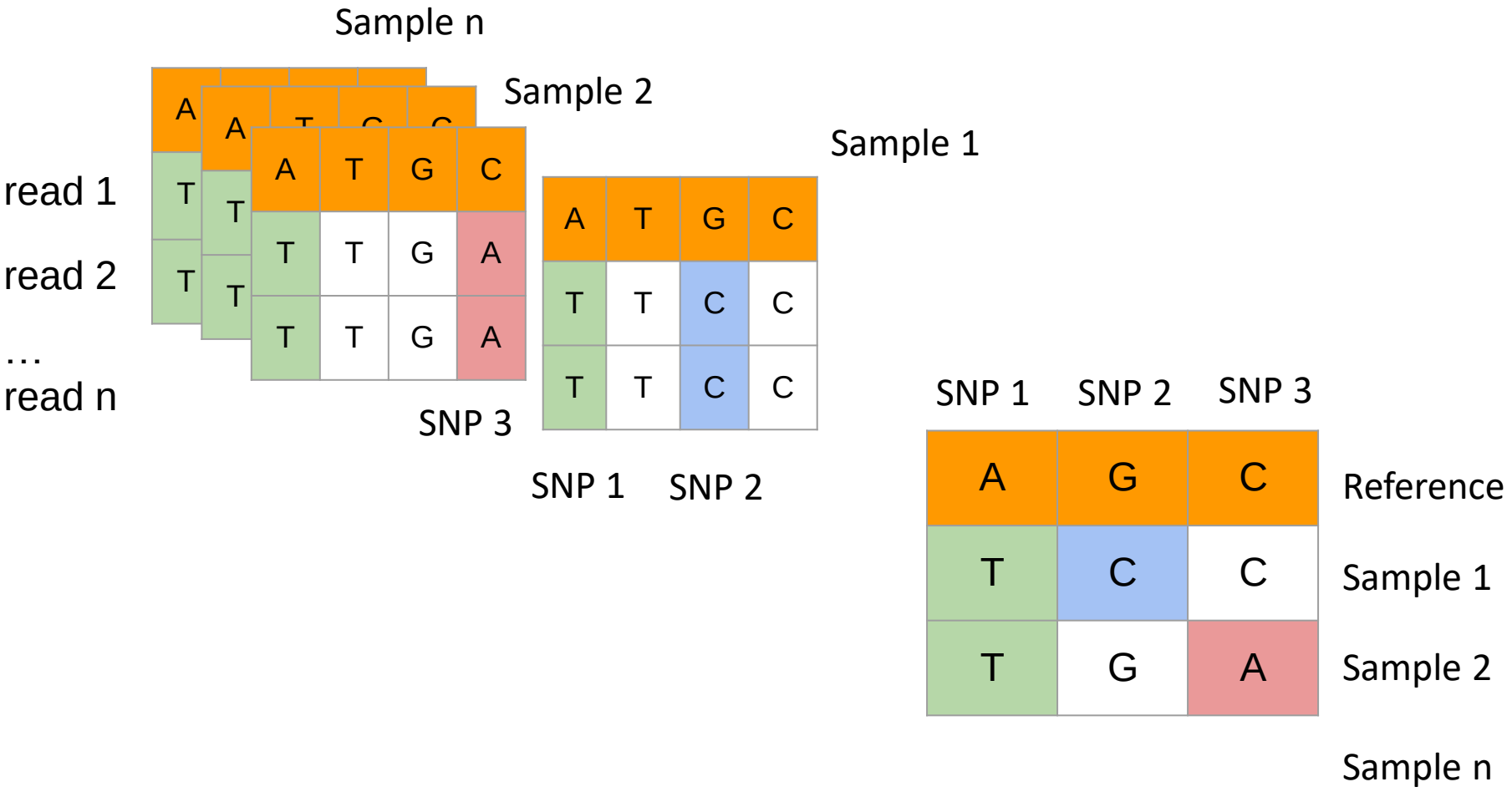
- Live-SET

High-quality SNPs para crear filogenia para investigación de brotes

https://github.com/lskatz/lyve-SET

- WGS-Outbraker

Generación de matriz de SNPs

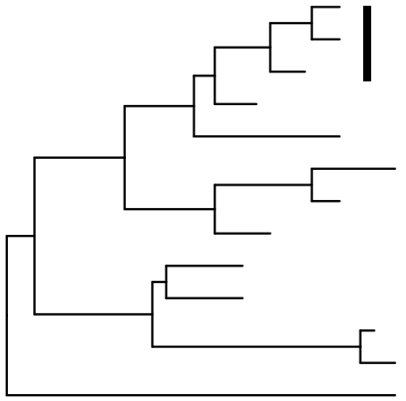


Generación de matriz de SNPs

SNP matrix

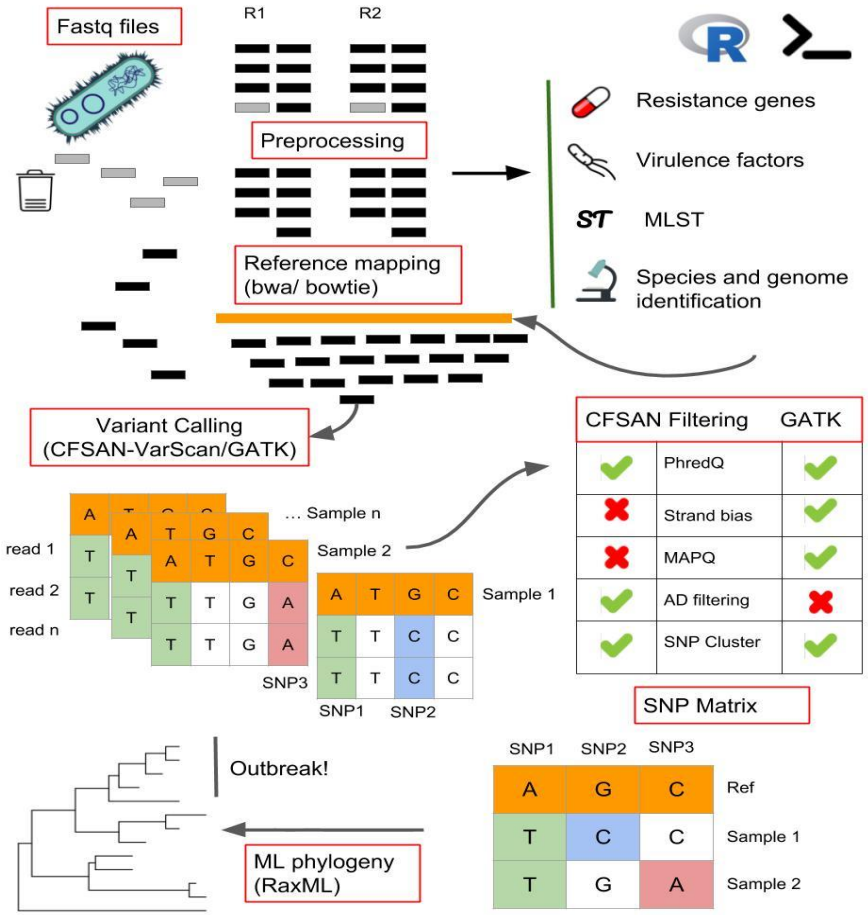
SNP 1	SNP 2	SNP 3	
A	G	C	Reference
T	C	C	Sample 1
T	G	A	Sample 2
			Sample n

Phylogeny



Outbreak!

WGS-Outbreaker



GMI, http://www.globalmicrobialidentifier.org/about-gmi

1. **Proficiency Testing for bacterial WGS, 2012**

   an end-user survey of current capabilities, requirements and priorities

2. **Proficiency Test Pilot, 2014**

   Wet lab and Dry lab

   Escherichia coli, Staphilococus aureus and Salmonella typhimurium

3. **Full Proficiency Test, 2015**

   Escherichia coli, Staphilococus aureus and Salmonella tiphimurium

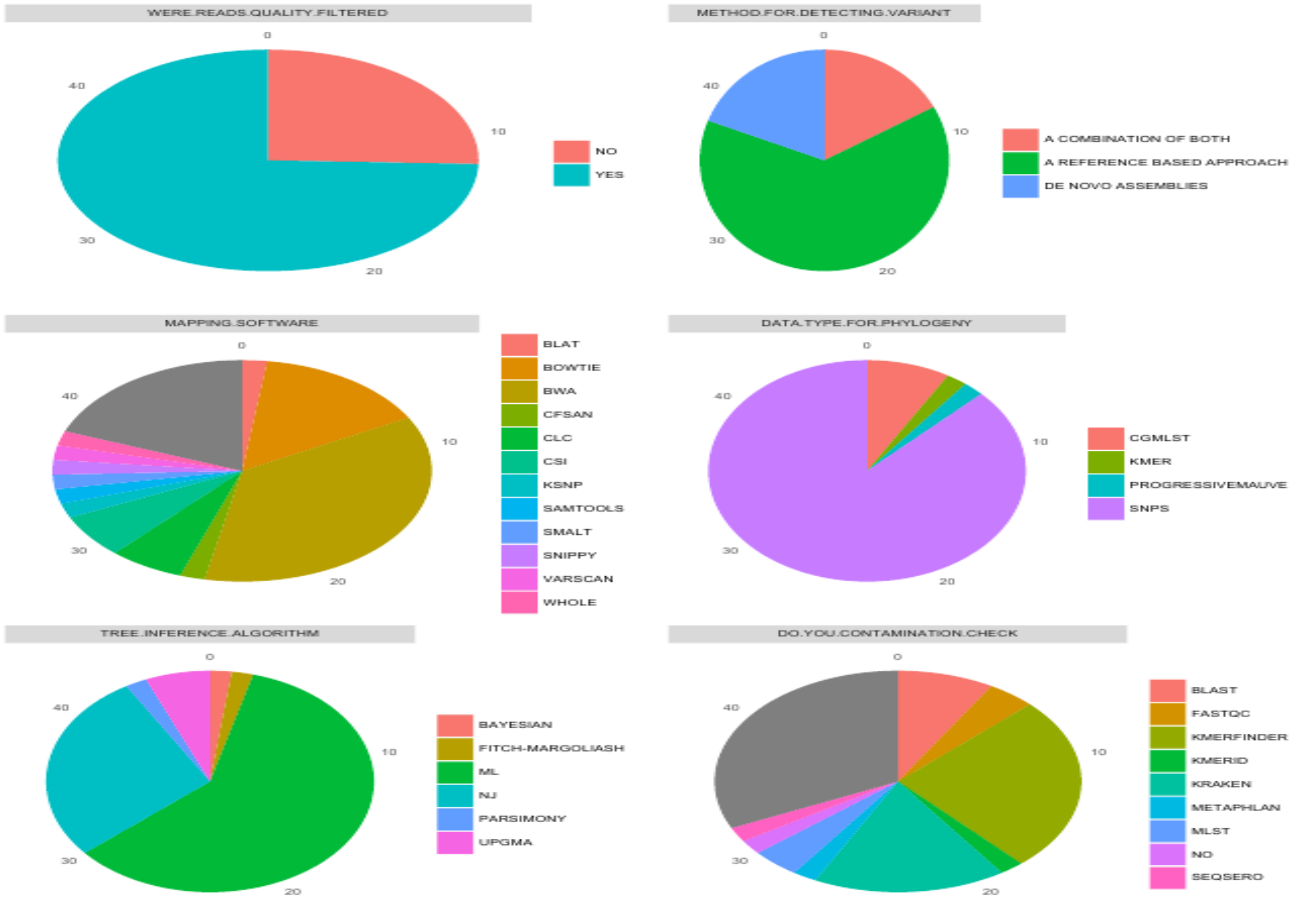
4. **Full Proficiency Test, 2016**

   Wet lab and Dry lab

   Campylobacter coli and C. jejuni, Listeria monocytogenes and klebsiella pneumoniae

GMI, Full Proficiency Test, 2015, Dry lab, Diversidad de métodos usados en el análisis

Figure 1. Pie charts illustrating the diversity of methods and practices employed for detecting variant from WGS data.



GMI, Full Proficiency Test, 2015, Dry lab,

- Número de SNPs reportado por cada laboratorio parcipante

Lab	EC	SA	ST
GMI02	25731	1383	8968
GMI04	25731	1383	8968
GMI06	43264	6226	5822
GMI10	13083	1797	12902
GMI14	14687	NA	1431
GMI26	92831	6164	31044
GMI39	52590	2672	16034
GMI42	9460	NA	12884
GMI43	38532	4163	16562
GMI46	63273	2341	9958
GMI48	67034	2063	14080
GMI58	79231	NA	19656
GMI59	23561	2715	14199
GMI13	9276	1628	8746
GMI16	55473	2122	13630
GMI21	5187829	2837196	5090636
GMI22	33416	1597	13066
GMI27	33664	2130	13297
GMI30	607217	11881	12733
GMI31	NA	NA	4141
GMI32	14667	25949	28164
GMI33	71822	5420	21668
GMI35	6706	1334	NA
GMI37	73355	2897	14294
GMI40	45725	2033	11180
GMI44	35039	1836	9446
GMI45	5183821	2836332	5088344
GMI47	20707	1805	12198
GMI50	84	NA	1300
GMI51	35521	NA	10042
GMI55	NA	1644	9102
GMI61	NA	NA	24
GMI63	NA	2834703	5077509
GMI7	21731	1673	9192
GMI8	15972	1851	12979

GMI, Full Proficiency Test, 2015, Dry lab, Detección de clusters

Escherichia coli

Lab	Cluster1	Cluster2
GMI02	TRUE	TRUE
GMI04	TRUE	TRUE
GMI10	TRUE	TRUE
GMI17	FALSE	FALSE
GMI26	TRUE	TRUE
GMI34	TRUE	TRUE
GMI39	TRUE	TRUE
GMI42	TRUE	TRUE
GMI43	TRUE	TRUE
GMI48	TRUE	TRUE
GMI58	TRUE	TRUE
GMI59	TRUE	TRUE
GMI13	TRUE	TRUE
GMI15	TRUE	FALSE
GMI16	TRUE	TRUE
GMI21	TRUE	TRUE
GMI22	TRUE	TRUE
GMI24	TRUE	TRUE
GMI27	TRUE	TRUE
GMI30	TRUE	FALSE
GMI32	TRUE	TRUE
GMI33	TRUE	TRUE
GMI35	TRUE	TRUE
GMI38	TRUE	TRUE
GMI40	TRUE	TRUE
GMI44	TRUE	TRUE
GMI45	TRUE	TRUE
GMI47	TRUE	TRUE
GMI50	TRUE	TRUE
GMI51	TRUE	TRUE
GMI7	TRUE	TRUE
GMI8	TRUE	TRUE

Staphilococcus aureus

Lab	Cluster1	Cluster2	Cluster3
GMI02	TRUE	TRUE	TRUE
GMI04	TRUE	TRUE	TRUE
GMI06	TRUE	TRUE	TRUE
GMI10	TRUE	TRUE	TRUE
GMI17	FALSE	FALSE	FALSE
GMI26	TRUE	TRUE	TRUE
GMI34	TRUE	TRUE	TRUE
GMI39	TRUE	TRUE	TRUE
GMI43	TRUE	TRUE	TRUE
GMI48	TRUE	TRUE	TRUE
GMI59	TRUE	TRUE	TRUE
GMI13	TRUE	TRUE	TRUE
GMI15	TRUE	TRUE	TRUE
GMI16	TRUE	TRUE	TRUE
GMI21	TRUE	TRUE	TRUE
GMI22	TRUE	TRUE	TRUE
GMI24	TRUE	TRUE	TRUE
GMI27	TRUE	TRUE	TRUE
GMI30	TRUE	FALSE	TRUE
GMI32	TRUE	TRUE	TRUE
GMI33	TRUE	TRUE	TRUE
GMI35	FALSE	TRUE	TRUE
GMI37	TRUE	TRUE	TRUE
GMI38	TRUE	TRUE	TRUE
GMI40	TRUE	TRUE	TRUE
GMI44	TRUE	TRUE	TRUE
GMI45	TRUE	TRUE	TRUE
GMI47	TRUE	TRUE	TRUE
GMI7	TRUE	TRUE	TRUE
GMI8	TRUE	TRUE	TRUE

Salmonella typhimurium

Lab	Cluster1	Cluster2
GMI02	TRUE	FALSE
GMI04	TRUE	FALSE
GMI06	TRUE	TRUE
GMI10	TRUE	TRUE
GMI14	TRUE	TRUE
GMI17	FALSE	FALSE
GMI26	TRUE	TRUE
GMI34	TRUE	TRUE
GMI39	TRUE	TRUE
GMI43	TRUE	FALSE
GMI46	TRUE	TRUE
GMI48	TRUE	TRUE
GMI59	TRUE	TRUE
GMI13	TRUE	TRUE
GMI15	TRUE	TRUE
GMI16	TRUE	TRUE
GMI21	TRUE	TRUE
GMI22	TRUE	TRUE
GMI24	TRUE	TRUE
GMI27	TRUE	TRUE
GMI28	TRUE	TRUE
GMI30	TRUE	TRUE
GMI31	TRUE	TRUE
GMI32	TRUE	TRUE
GMI33	TRUE	TRUE
GMI37	TRUE	TRUE
GMI38	TRUE	TRUE
GMI40	TRUE	TRUE
GMI44	TRUE	TRUE
GMI45	TRUE	TRUE
GMI47	TRUE	TRUE
GMI55	TRUE	TRUE
GMI63	TRUE	FALSE
GMI7	TRUE	TRUE
GMI8	TRUE	TRUE

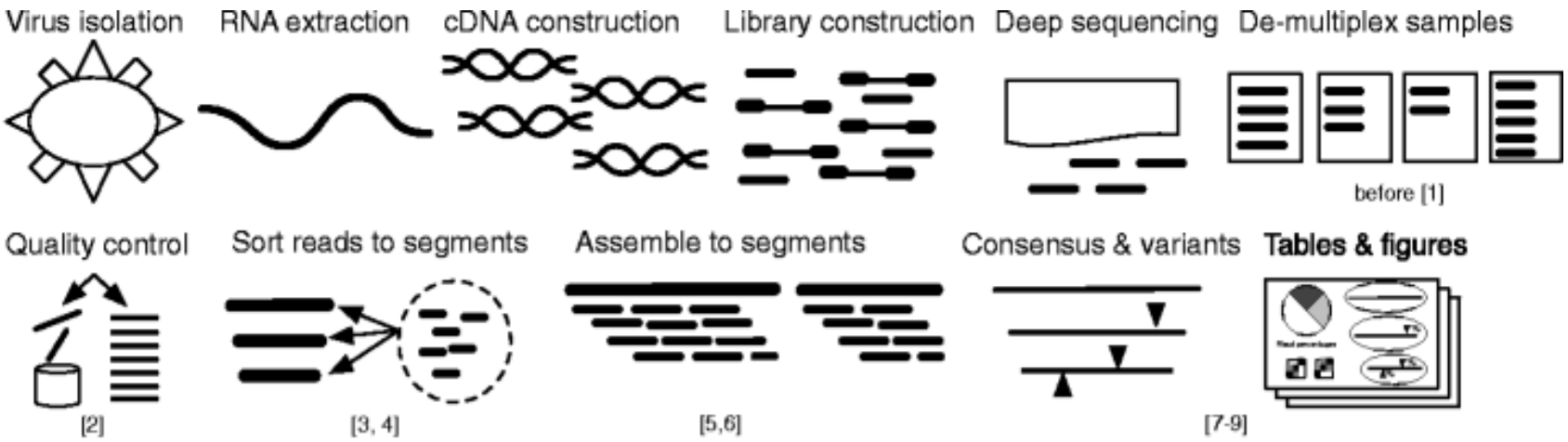
GMI, Full Proficiency Test, 2015, Dry lab, Conclusiones

- 2015 GMI PT highlight the diversity of bioinformatics tools that are being employed around the world to analyze WGS data of bacteria that are of importance to public health and food safety.
- These methods do not produce the same data objects (variant positions and SNP matrices) from which phylogenetic trees (topologies) are inferred.
- However, the topologies clustered samples quite similarly (>93% participants clustered samples correctly).
- A vast majority of labs would reach similar conclusions.
- Individual centers will be able to define sensible thresholds for determining clusters of isolates.
- A standardized approach will likely emerge within which thresholds will be decided upon that will facilitate congruence among center-specific pipelines in the conclusions that are reached.

Ejemplo de llamada a variantes: Virus

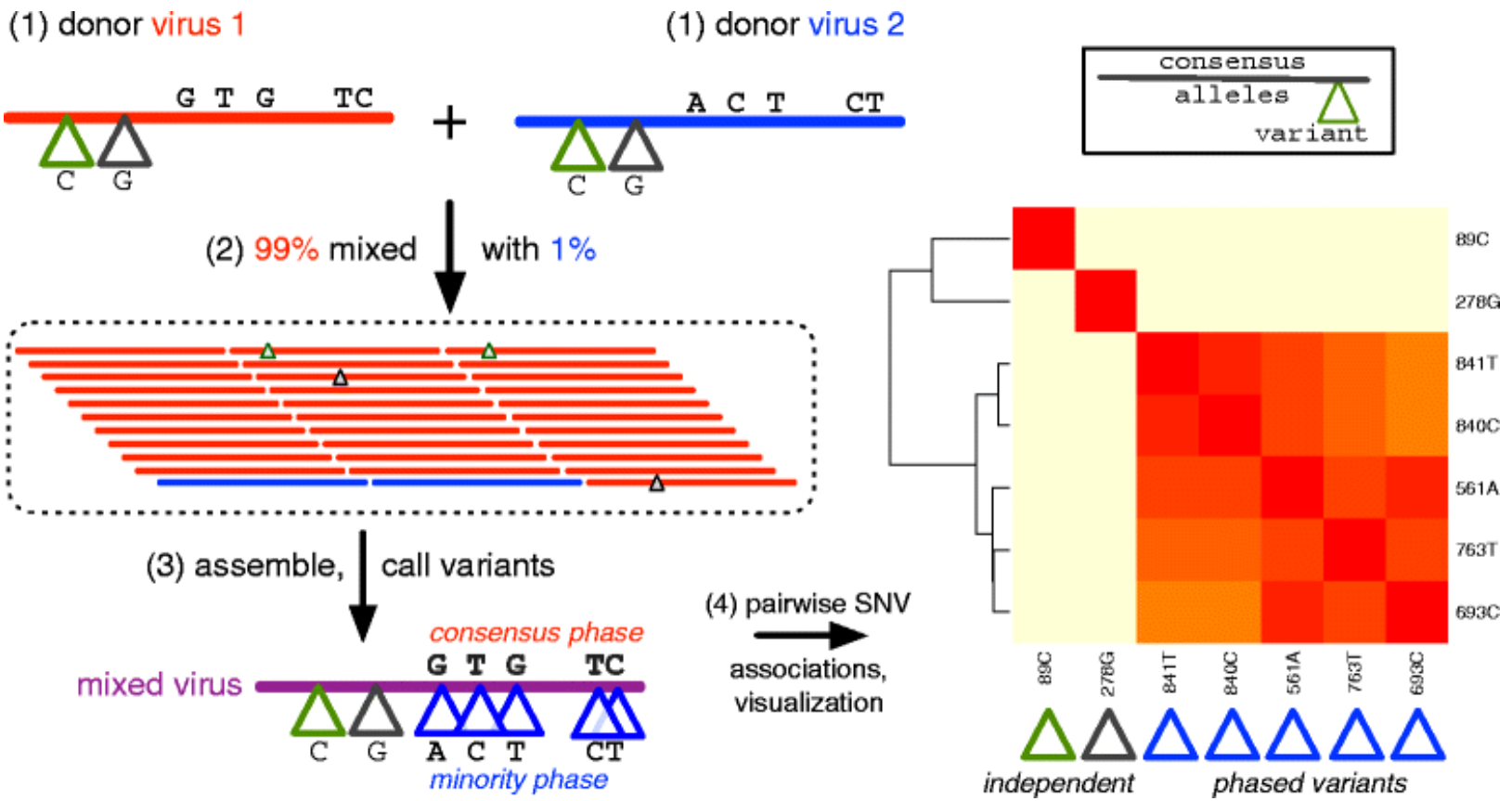
IRMA: Iterative Refinement Meta-Assembler

A



Shepard et al BMC Genomics 2016, **17**:708

Ejemplo de llamada a variantes: Virus



Shepard et al BMC Genomics 2016, **17**:708

¿Preguntas?