

Desarrollo de un pipeline
Bioinformático:

Análisis basado en un panel de
genes sobre cáncer de pulmón.

David Masip Galaso

Máster Universitario en Bioinformática y Bioestadística UOC - UB

Computación e Inteligencia artificial en problemas biológicos y clínicos

Consultora: Romina Astrid Rebrij

Contenido

- Objetivos
- Alcance y riesgos
- Fase analítica I y II
- Conclusiones



Objetivos

- Por qué cáncer de pulmón?
- Panel de genes específico
- Punto partida → Archivos en bruto “fastq”
- Plataforma utilizada → R y otros SW de visualización



Alcance y riesgos

- Uso de R como herramienta vehicular
- Librerías R y Bioconductor:
 - Lectura - ShortRead
 - Calidad - Fastq files Quality Check
 - Conversión - SAMtools
 - Alineamiento - Bowtie2
 - Anotaciones – Galaxy
 - Variaciones – Galaxy/UCSC Genome
- Fastq files + Index files
 - No FASTA – No SAM
 - Familiarización Bioconductor





Fase analítica I

- Origen Datos:
 - MiSeq Illumina
 - Tipo datos: Extensión .fastq
 - Longitud: 76 pb + índice (8pb)
- Control Calidad previo Illumina:
 - “Filtrado de calidad” → Alta calidad lecturas

Measure	Value
Filename	15-0991_S1_L001_R1_001_fastq
File type	Conventional base calls
Encoding	Sanger / Illumina 1.9
Total Sequences	2228315
Sequences flagged as poor quality	0
Sequence length	76
%GC	48

Lectura de datos

- Lectura de datos mediante Shortread

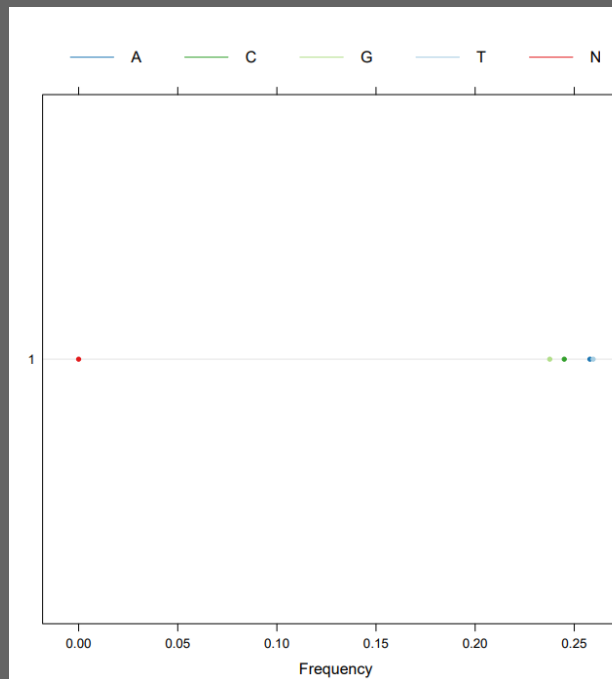
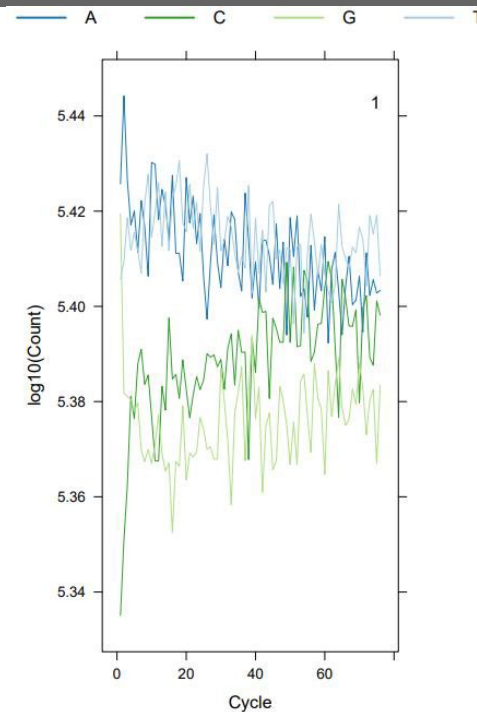
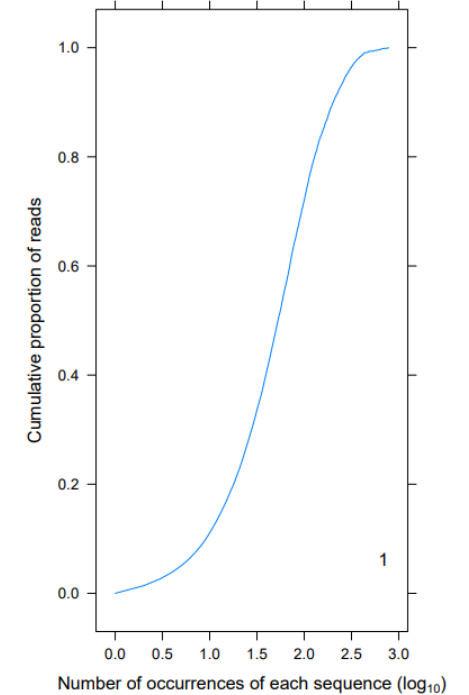
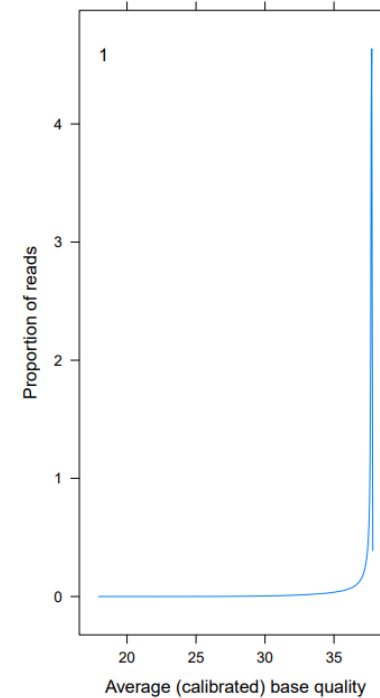
- Ejemplo:[1] 76

CTTAAACTGATTTTACATGGTACATGAA

ACAAGGCAAATAACTGCGATTTTTTTCTT

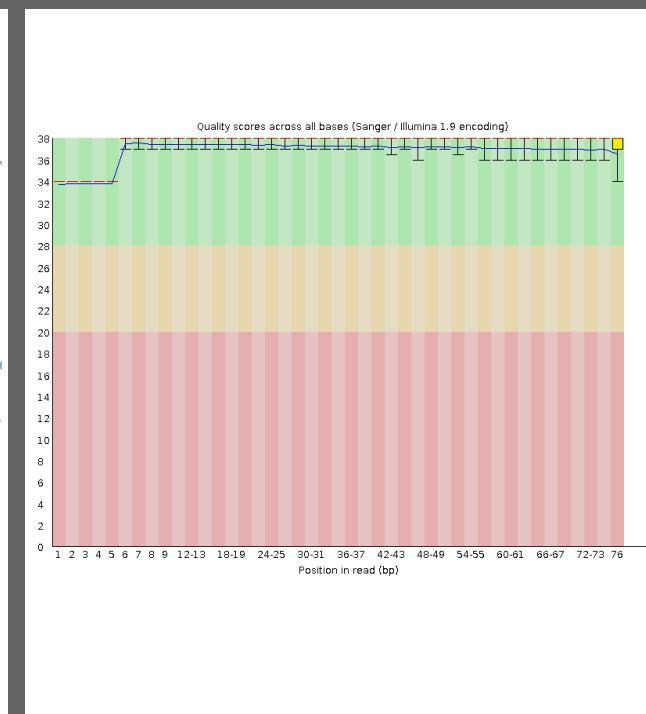
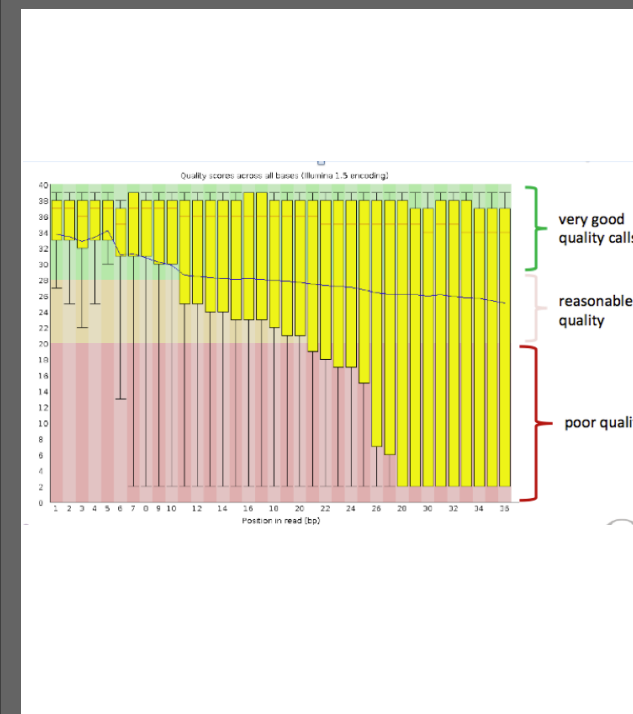
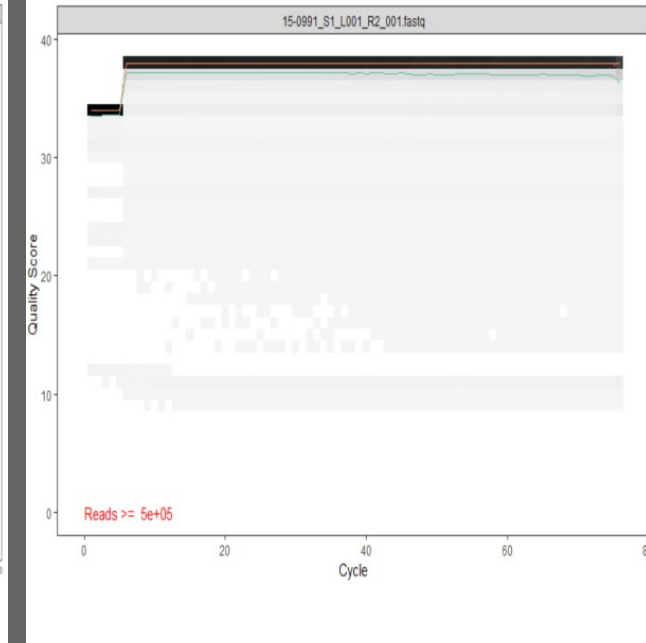
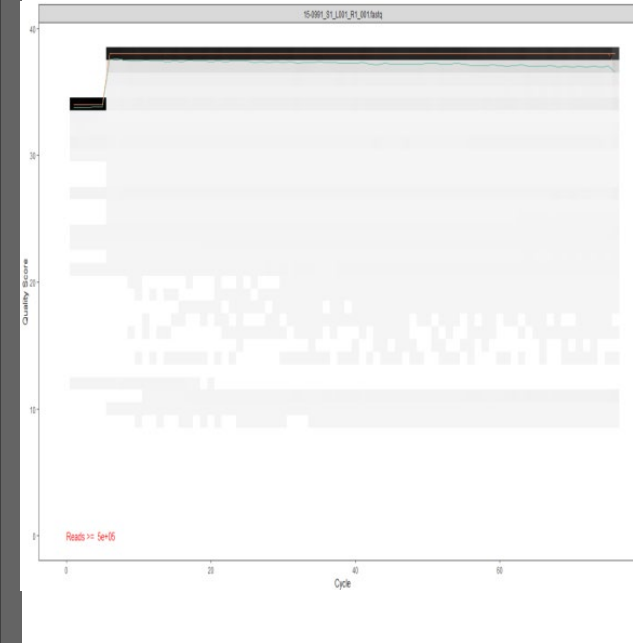
CCTTCTGCTCCTTCCCCT

- Análisis de calidad [1]
- Distribución lecturas [2]
- Frecuencia de llamadas por base [3]



Control de calidad

- Contaminación y distribución
 - Anverso y Reverso [1][2]
- Gráfico de caja QC [3]
- Gráfico de caja Real [4]





Fase analítica II

- Transformación e inspección de BAM files
- Preparación de datos y observación de opciones:
 - BLAST/BLAT → Poderosos pero inadecuados
 - BWA/Bowtie2 → SW más usado para NGS
- Mapeo en un genoma de referencia
- Visualización de anotaciones y variants mediante navegador genómico (IGV)

Preparación de datos y Genoma de referencia

- Uso SAM/BAM files para alineamiento
 - Conversión con Galaxy → Rápido y eficaz
- BAM File
 - Sección encabezado
 - Sección de alineación (11 campos obl)
- Preparación Genoma de referencia
 - GRCh37.hg19
 - Alineamiento chr10

@SQ SN:chr10 LN:135534747

QNAME: M01621:321:000000000-BV2JC:1:2108:23966:23449

FLAG: 16

RNAME: chr10

POS: 1739232

MAPQ: 42

CIGAR: 76M

MRNM: *

MPOS: 0

ISIZE: 0

SEQ:

CTTCTTCTGATGTGGCCCAAGCCAAGATTGGACACCCCTGATCTAAAGGTTTTCAITTTCTGTTTCT
CTC

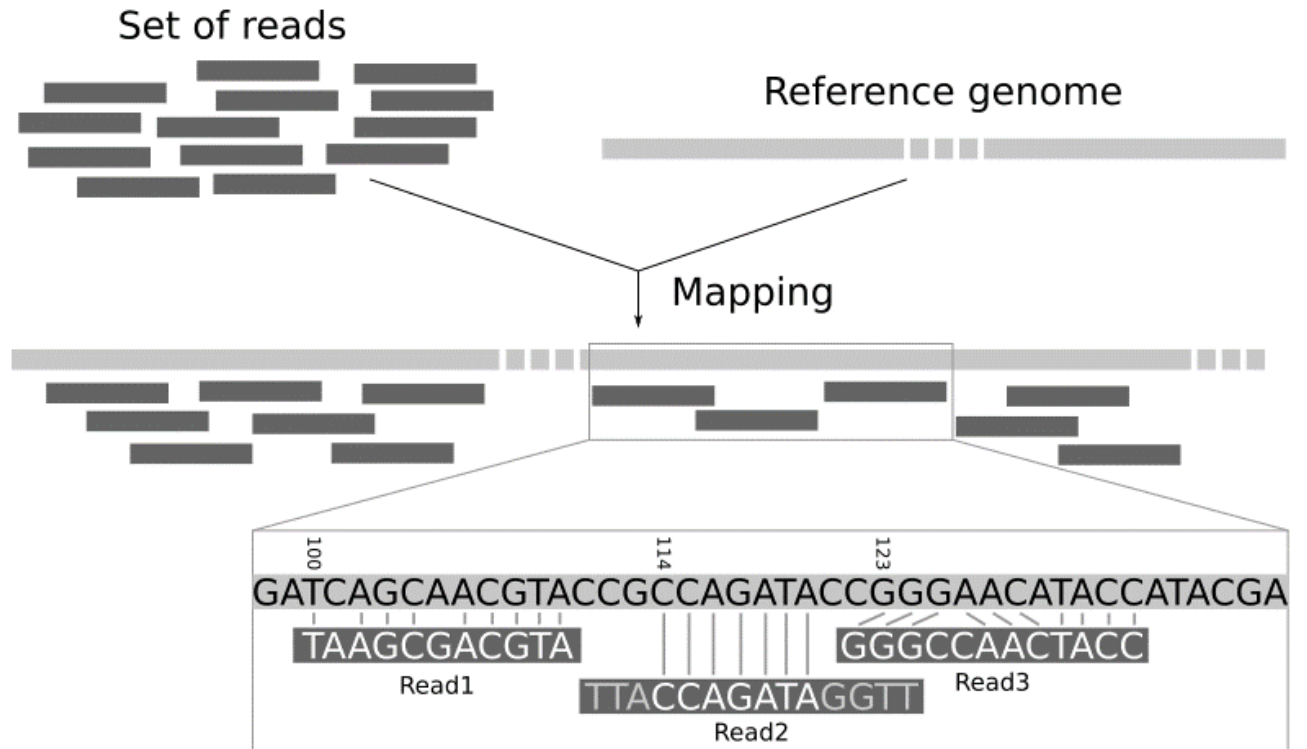
QUAL:

FGFGGG
GGGGGCCCCC

Col	Field	Type	Brief Description
1	QNAME	String	Query template NAME
2	FLAG	Integer	bitwise FLAG
3	RNAME	String	References sequence NAME
4	POS	Integer	1- based leftmost mapping POSition
5	MAPQ	Integer	MAPping Quality
6	CIGAR	String	CIGAR String
7	RNEXT	String	Ref. name of the mate/next read
8	PNEXT	Integer	Position of the mate/next read
9	TLEN	Integer	observed Template LENgth
10	SEQ	String	segment SEquence
11	QUAL	String	ASCII of Phred-scaled base QUALity+33

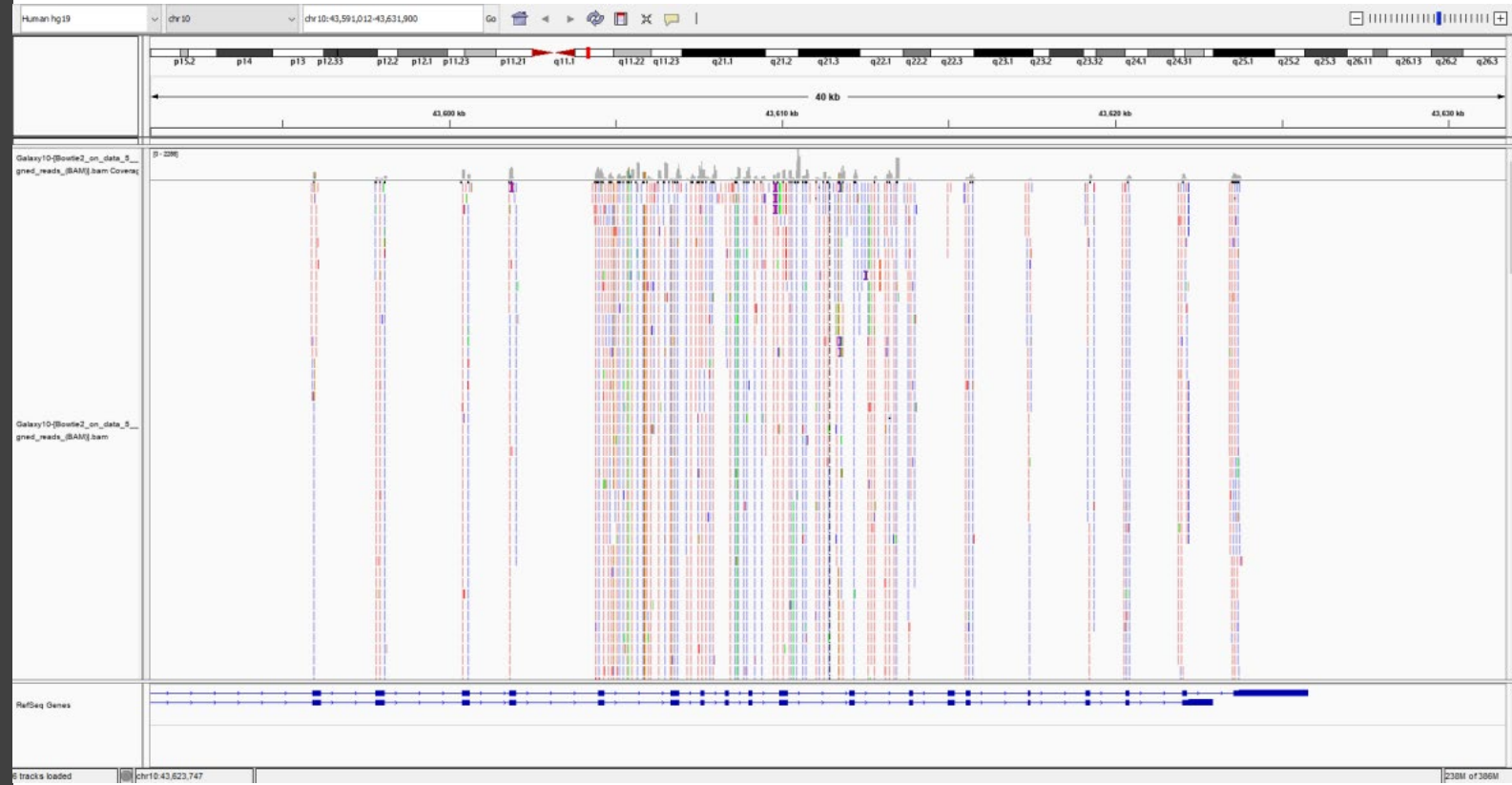
Mapa leído en Genoma de referencia

- Utilidades del mapping:
 - Alineación con genoma de referencia.
 - Alinea la lectura, permite ver desajustes, indeles y recortes de fragmentos cortos.
 - Verificación de estadísticas de mapeo.
- Reacción en cadena de polimerasa (PCR).
- Errores de secuenciación.
- Errores de asignación.



Anotaciones

- Uso de anotaciones con datos biológicos relevantes
- Ayuda a científicos:
 - Entendimiento
 - Estructura
 - Funcionamiento
- Parte principal:
 - Localización de genes y proteínas



Anotaciones

- Entrada/Muestra
 - Ensamble de alta calidad > 90%
 - SW utilizado dependerá:
 - Tipo de datos utilizados
 - Objetivo de anotación
 - Recursos disponibles

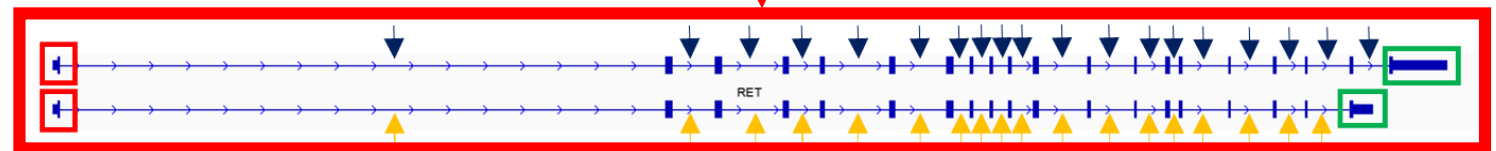
```
2228315 reads; of these:  
 2228315 (100.00%) were unpaired; of these:  
   4428 (0.20%) aligned 0 times  
 2139611 (96.02%) aligned exactly 1 time  
   84276 (3.78%) aligned >1 times  
99.80% overall alignment rate
```

- Fuentes de error más frecuentes:
 - No filtrar las regiones en el genoma que no contienen genes.
 - Fallo al elegir los programas computacionales.
 - Los datos de referencia contienen errores.
 - Se utiliza un genoma de referencia con anotaciones erróneas.
 - Aplicaciones Predicción de genes Predicción de funciones de genes.
- Anotaciones → Aplicaciones en investigación:
 - Desarrollo de hipótesis
 - Análisis de genómica comparativa
 - Medio importante para anotación de otros genomas
- Alineación → Infinidad de posibilidades:
 - Lookseq, IGV, Jbrowse, Genome Workbench

The IGV logo consists of the letters 'IGV' in a bold, white, sans-serif font, centered within a solid orange square.

Anotaciones

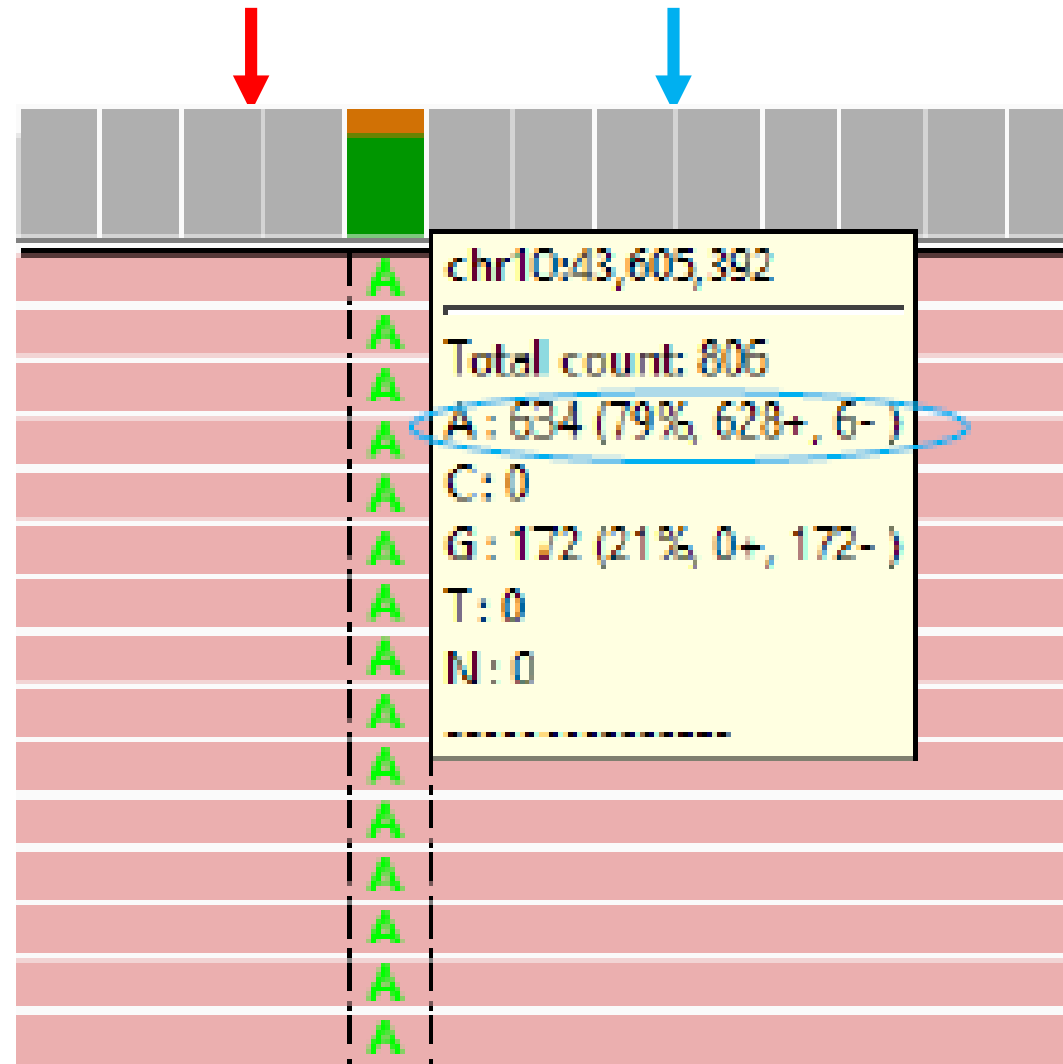
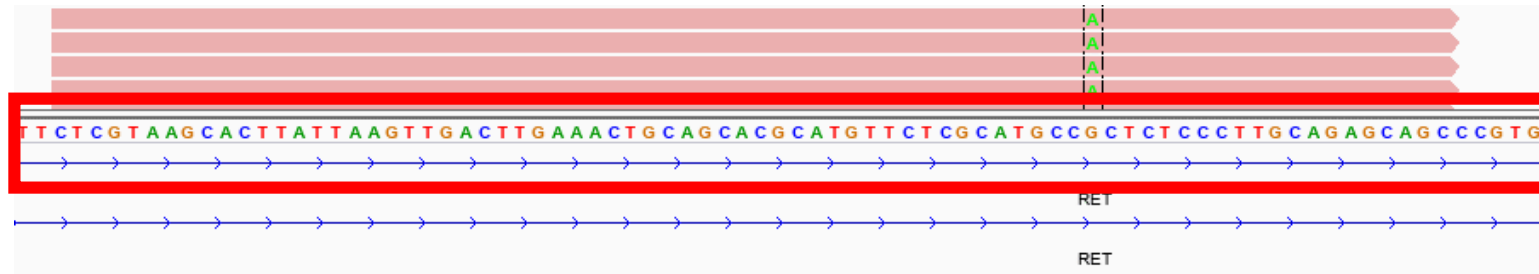
- Carga datos: BAM + BAI
- IGV: Parte superior muestra extensión cromosoma entera
- Encontrar foco secuencia: chr10 – posición
- Parte inferior – Referencia con Gen/Genes asociados: RET
 - 2 posibles variaciones
 - 5' UTR en la parte izquierda (roja)
 - 3' UTR en la parte derecha (verde)
 - Diferencia de exones e Intrones en 2 variaciones del Gen RET



Por defecto, el string está representado de 5' a 3', pero si se selecciona la señal, se gira hacia el reverso.

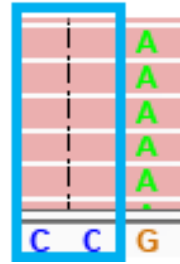
Anotaciones

- Visualización → Zoom in
 - Genoma referencia vs lecturas
- Alineaciones referenciadas
 - Amplitud barra superior
- Eventos interesantes → Color
 - Caso particular → Calidades de mayor intensidad a menor
 - Heterocigoto SNP → Secuencia que afecta a una sola base o polimorfismo de nucleótido simple



Anotaciones

- Intensidad de color → mayor intensidad → mayor calidad
- Misma lectura → Mapping distinta calidad. Ejemplo 42 vs 0
- Inserciones
- Delecciones



Two panels showing read alignment details. The left panel shows a read with a mapping quality of 42, and the right panel shows a read with a mapping quality of 0. Both panels include fields for Hap name, Dist, Read name, Read length, Mapping, Reference span, Cigar, Clipping, and various quality metrics (XG, NM, XM, XN, XO, AS, YT, Hidden tags, Location, Base).

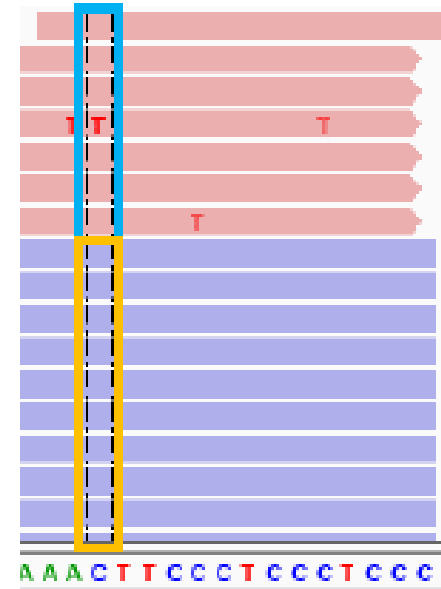
Two panels showing read alignment details. The left panel shows a read with a mapping quality of 42, and the right panel shows a read with a mapping quality of 0. Both panels include fields for Hap name, Dist, Read name, Read length, Mapping, Reference span, Cigar, Clipping, and various quality metrics (XG, NM, XM, XN, XO, AS, YT, Hidden tags, Location, Base). Arrows point from the mapping quality fields to the corresponding read details.

A diagram showing read alignment with insertion and deletion annotations. The left panel shows a read with a mapping quality of 42, and the right panel shows a read with a mapping quality of 0. Both panels include fields for Hap name, Dist, Read name, Read length, Mapping, Reference span, Cigar, Clipping, and various quality metrics (XG, NM, XM, XN, XO, AS, YT, Hidden tags, Location, Base). The left panel shows an insertion of 5 bases (CCGAT) and the right panel shows a deletion of 2 bases (CT).



Anotaciones

- SNP potencial sin llamadas a umbral de representación y genoma de ref.
- Evitar falsos positivos → Color lecturas: Anverso y Reverso
- Ejemplo: No hay SNP en reverso pero si posible Heterocito SNP → True → Espera Timina en reverso. Al ser inexistente → FALSO POSITIVO
- SNP más destacados → Galaxy



	1	2	3	4	5	6
chr10	1739232		c	1	^K,	E
chr10	1739233		t	1	,	G
chr10	1739234		t	1	,	F

dónde:

Column	Definition
1	Chromosome
2	Position (1-based)
3	Reference base at that position
4	Coverage (# reads aligning over that position)
5	Bases within reads where (see Galaxy wiki for more info)
6	Quality values (phred33 scale, see Galaxy wiki for more)

Variantes

- Eliminación variantes fuera regiones codificantes.
- Visualización lecturas para eliminar posibles errores de secuenciación.
- Detección SNP tras mapeo de lecturas frente genoma ref.
- Obtención archivo .vcf con variants y archive BAM/BAI con lecturas alineadas → Permite conocer número lecturas respaldan cada SNP

- Chr 10 → Longitud aprox 128mb → 1,3M. Se reduce para mostrar detalle

- Filtrar variantes que no se vayan a mostrar, debido a errores de secuenciación por ejemplo, fuera zona codificante, etc.

- Creación Pileup:
 - Filtrado 6 columnas
 - Sin reporte posiciones mayor cobertura 30
 - Reporte sólo variantes

Variantes

- FreeBayes → Evalúa probabilidad de cada genotipo posible para cada pos genoma ref.
- Exportación vcf para obtención variantes en zona analizada.
- Evaluación variantes:
 - SNP Rojos → cambios en los aminoácidos.
 - SNP Verde → variantes sinónimas.
 - SNP Azul → regiones traducidas o empalmadas.
 - SNP Negro → Regiones de intrones.

UCSC Genome Browser

chr10:43,358,938-43,967,937 609

Scale chr10: 43,500,000

Ch37/hg19) Assembly

hg19 43,800,000 43,900,000

Viewer on Human Feb. 2009 (GRCh37/hg19) Assembly

5,384 5 bp

Variant Information Panel:

Chr: chr10
Position: 43605392
ID: .
Reference: G*
Alternate: A
Qual: 13870.6
Type: SNP
Is Filtered Out: No

Alleles:

Alternate Alleles: A
Allele Count: 1
Total # Alleles: 2
Allele Frequency: 0.5

Variant Attributes

Allele Frequency: 0.5
CIGAR: 1X
ODDS: 214.644
PAIRED: 0
SAP: 1323.76
EPP: 774.969
Number of Samples with Data: 1
SAR: 6
SRF: 0
NUMALT: 1
Depth: 803
PRO: 0
EPPR: 374.332
ABP: 577.71
RPL: 559
QA: 23658

SNP List:

rs115025059	rs55300933	rs3004256	rs3123	rs79198826	rs2253260	rs7905676	rs4564528
rs76590777	rs2488289	rs7074380	rs4949	rs79529756	rs2253099	rs10409	rs2492405
rs4948691	rs2796568	rs7097557	rs73264	rs55725459	rs2252975	rs1126487	rs4559639
rs2796554	rs1625738	rs3123727	rs731	rs77130662	rs1879311	rs7077713	rs4551711
rs4949839	rs1775083	rs3121320	rs7277	rs2681910	rs2252761	rs4597022	rs9422365
rs149496396	rs61843489	rs2995398	rs1064	rs71533058	rs2252758	rs7900207	rs7069214
rs11657726	rs2795521	rs12411293	rs1743	rs1915149	rs2460565	rs5784597	rs2265887
rs2744879	rs1775084	rs61845272	rs1743	rs2503842	rs2252744	rs7092125	rs3006387
rs788286	rs1775085	rs10900290	rs312	rs2681912	rs2252458	rs12220230	rs2265953
rs788285	rs1774205	rs72779317	rs706	rs70939486	rs7094716	rs76889847	rs2492404
rs2796553	rs1774204	rs78855838	rs1734	rs79518300	rs2460564	rs72785266	rs2612797
rs61843371	rs2487916	rs2995399	rs618	rs76714963	rs2493661	rs75015989	rs2492403
rs1085783	rs35759751	rs61845274	rs769	rs7100218	rs2249345	rs7918726	rs2261665
rs2744878	rs2488290	rs11819429	rs618	rs2251416	rs2249332	rs4424628	rs7073767

Variantes

- Filtrado y evaluación del impacto de variantes → Análisis localización SNP
- Según impacto funcional:
 - SNPs sinónimos: no causan alteración de la secuencia de proteína codificada por ese gen.
 - SNPs missense que sí alteran la secuencia de proteína.
 - SNPS que producen la ganancia de un codón STOP.
 - SNPs que producen la pérdida de un codón de inicio.
 - SNPs en una región de secuencias repetitivas.
 - SNPs o en una región no codificante.

- Las anotaciones por dbSNP:
 - RET (NM_020630): synonymous_variant S (TCC) --> S (TCG)
 - RET (NM_020975): synonymous_variant S (TCC) --> S (TCG)

 - RET (uc010qez.1) synonymous_variant S (TCC) → S (TCA) → STOP
 - RET (uc010qez.1) synonymous_variant S (TCC) → S (TCG) → Cysteina [15]
 - RET (uc001jal.3) synonymous_variant S (TCC) → S (TCA) → STOP
 - RET (uc001jal.3) synonymous_variant S (TCC) → S (TCG) → Cysteina[15]
 - RET (uc001jak.1) synonymous_variant S (TCC) → S (TCA) → STOP
 - RET (uc001jak.1) synonymous_variant S (TCC) → S (TCG) → Cysteina[15]

Validación y Efectos Biológicos

- Comparación variantes obtenidas con anotadas en bdd dbSNP.
- BBDD dbSNP → archivo público gratuito.
 - Variación genética dentro y entre especies
- Caso Particular → Variaciones en intrones. Rs2251674
- Variante → Relación con cancer tiroides de origen folicular y polimorfismo en RET.
- Asociación haplotípica → Susceptibilidad sobre cancer tiroides

Alignment between genome (hg19 chr10:43615133-43616133, + strand; 1001 bp) and dbSNP sequence (rs1800863; 1001 bp)
ID (including gaps) 99.9%, coverage (of both) 100.0%

```
43615133 CGACCTCATCTCATTTGCCTG5CAGATCTCACAGGGGATGCAGTATCTGGCCGAGATGAAGGTGCGTGCATATG5CTCTGCACCCAGCCAGCCCCGgcca 43615232
|||||
00000001 CGACCTCATCTCATTTGCCTG5CAGATCTCACAGGGGATGCAGTATCTGGCCGAGATGAAGGTGCGTGCATATG5CTCTGCACCCAGCCAGCCCCGGCCA 00000100
|||||

43615233 ggccacaccctgacccaccacgccccctgccaccacaccctggcctgccactccccaccatgccacactctagcccaccatgccccctgccatggcatgc 43615332
+++++
00000101 GGCCACACCCTGACCCACCACGCCCTGCCACCCACACCCTGGCCTGCCACTCCCCACCATGCCACACTCTAGCCCACCATGCCCTGCCATGGCATGC 00000200
|||||

43615333 catgctatggctcaccacgccccctgccatgtcacaccctgactccaccacgccccctgccatgccacaccCCCGCCAGGTCTCACAGGCCGCTACCCG 43615432
+++++
00000201 CATGCTATGGCTACCCACGCCCTGCCATGTACACCCTGACTCCACCACGCCCTGCCATGCCACACCCCCGCCAGGTCTCACAGGCCGCTACCCG 00000300
|||||

43615433 GGCCACACACCACCCCTCTGCTGGTACACCAG5CTGAGCCAGTGACCCTGCTGCCTGGCCATGGCCTGACGACTCGTGCTATTTTTCTCACAGCTCG 43615532
|||||
00000301 GGCCACACACCACCCCTCTGCTGGTACACCAG5CTGAGCCAGTGACCCTGCTGCCTGGCCATGGCCTGACGACTCGTGCTATTTTTCTCACAGCTCG 00000400
|||||

43615533 TTCATCGGGACTTGGCAGCCAGAAACATCCTGGTAGCTGAGGGGCGGAAGATGAAGATTTCGGATTTCCGGCTGTCCCAGATGTTTATGAAGAGGATTC 43615632
|||||
00000401 TTCATCGGGACTTGGCAGCCAGAAACATCCTGGTAGCTGAGGGGCGGAAGATGAAGATTTCGGATTTCCGGCTGTCCCAGATGTTTATGAAGAGGATTC 00000500
|||||

43615633 C 43615633
00000501 V 00000501

43615634 TACGTGAAGAGGAGCCAGGTGCCAGTCCCAGG5ATGAGGCGGGGCTCCCAGGATCCCAGGTGCACCATGGGGCAGGCAGTGCCTTGGGAAGCCTAGG 43615733
|||||
00000502 TACGTGAAGAGGAGCCAGGTGCCAGTCCCAGG5ATGAGGCGGGGCTCCCAGGATCCCAGGTGCACCATGGGGCAGGCAGTGCCTTGGGAAGCCTAGG 00000601
|||||

43615734 AAAGATACCGAAGATTAGTGGAGCTCTAAGCTTTTTATAGCCCTCACCCAAATCTTTCTGACCTGGGTCCCCAAGGACCCAATTAGAACTCCGCTCAG 43615833
|||||
00000602 AAAGATACCGAAGATTAGTGGAGCTCTAAGCTTTTTATAGCCCTCACCCAAATCTTTCTGACCTGGGTCCCCAAGGACCCAATTAGAACTCCGCTCAG 00000701
|||||

43615834 CCTCTGCCATGTCCTTCTCCTCCAGG5CCTCCAGG5CACCCCTCCCTGGCAGCATACTGACCCGAGGCCCTTGCCGCACTTTTCAGAGGCCACCTCATGC 43615933
|||||
00000702 CCTCTGCCATGTCCTTCTCCTCCAGG5CCTCCAGG5CACCCCTCCCTGGCAGCATACTGACCCGAGGCCCTTGCCGCACTTTTCAGAGGCCACCTCATGC 00000801
|||||

43615934 TGCGGAACTAACAGTCTCTTCTGCAGAATAAAGGTCACCGTTCTGATATGACCTTAGCTCTTTCTCAAAGAAGGGTGGGATGAAATTAGCAGGATCGT 43616033
|||||
00000802 TGCGGAACTAACAGTCTCTTCTGCAGAATAAAGGTCACCGTTCTGATATGACCTTAGCTCTTTCTCAAAGAAGGGTGGGATGAAATTAGCAGGATCGT 00000901
|||||

43616034 CATTCTTTGCAAAAAGGAATGAACTGCTTTACAAGTGAGGCTTCTCCCGCACAGGG5CCTTG5ACTG5GCTG5GTGAGTTTAGAGGCATAGGAACCCC 43616133
|||||
00000902 CATTCTTTGCAAAAAGGAATGAACTGCTTTACAAGTGAGGCTTCTCCCGCACAGGG5CCTTG5ACTG5GCTG5GTGAGTTTAGAGGCATAGGAACCCC 00001001
|||||
```

Figura 63. Alineación entre el genoma (hg19 chr10: 43604892-43605892, cadena +; 1001 pb) y la secuencia dbSNP (rs2251674; 1001 pb)

Conclusiones

1. Representación Pipeline a partir datos tipo Fastq.
2. Posibilidad de guía útil genérica para cualquier tipo de estudio
3. Representación Gráfica gracias a SW: R, IGV, Genome Browser

