

Bioinformática y análisis de datos ómicos

TEMA 7: IDENTIFICACIÓN DE ALTERACIONES GENÓMICAS



Ignacio Varela Egocheaga

DEPARTAMENTO DE BIOLOGÍA MOLECULAR

Este material se publica bajo la siguiente licencia:

[Creative Commons BY-NC-SA 4.0](https://creativecommons.org/licenses/by-nc-sa/4.0/)

Tema 7: Identificación de alteraciones genómicas

- 7.1 Estudios genómicos utilizando chips genéticos
- 7.2 Preparación de librerías genómicas
- 7.3 Alineamiento de secuencias
- 7.4 Identificación de sustituciones
- 7.5 Identificación de pequeñas inserciones y deleciones
- 7.6 Identificación de cambios en el número de copia
- 7.7 Identificación de grandes reorganizaciones genómicas
- 7.8 Anotación funcional de mutaciones

Tema 7: Identificación de alteraciones genómicas

7.1 Estudios genómicos utilizando chips genéticos

7.2 Preparación de librerías genómicas

7.3 Alineamiento de secuencias

7.4 Identificación de sustituciones

7.5 Identificación de pequeñas inserciones y deleciones

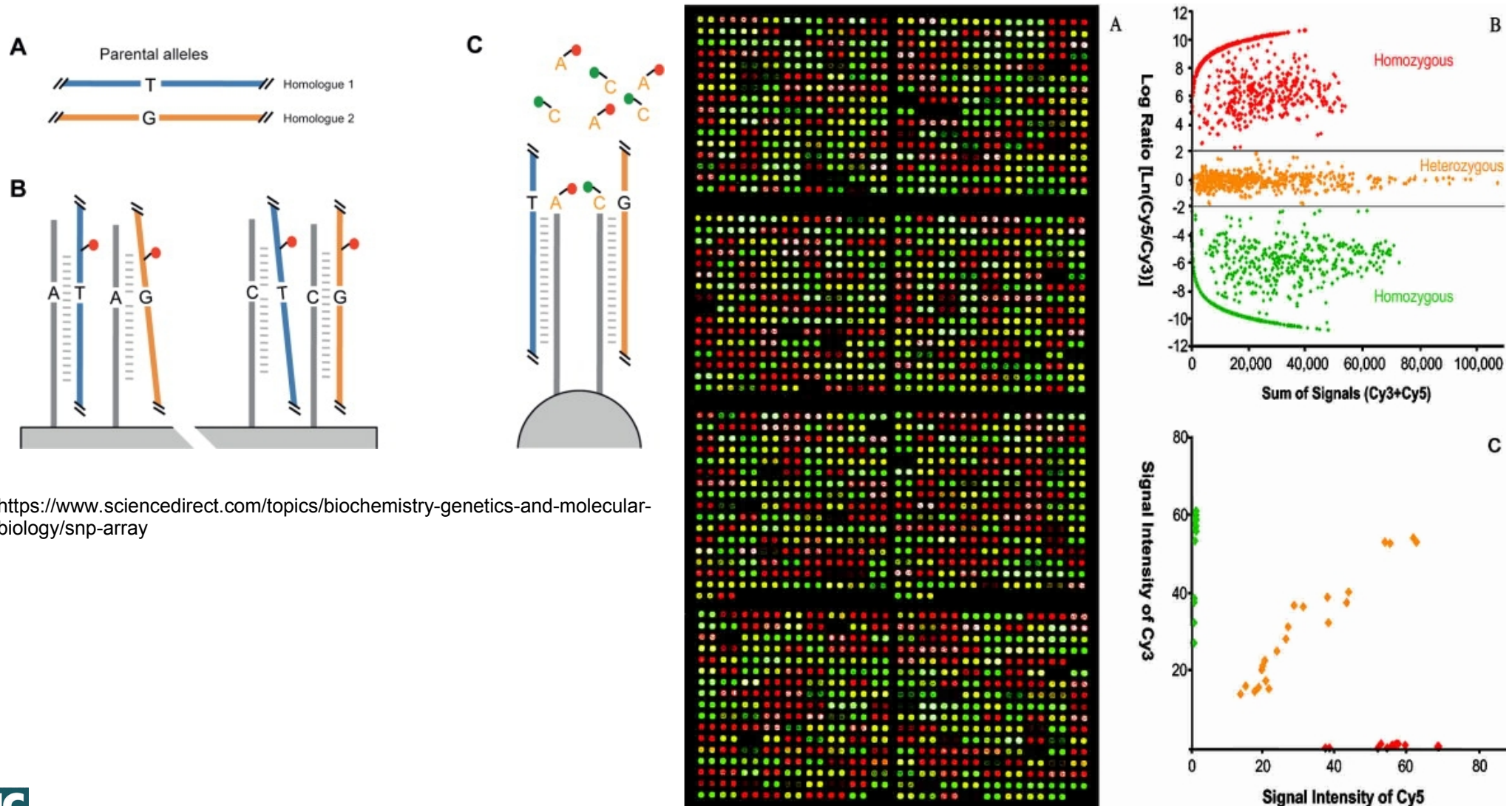
7.6 Identificación de cambios en el número de copia

7.7 Identificación de grandes reorganizaciones genómicas

7.8 Anotación funcional de mutaciones

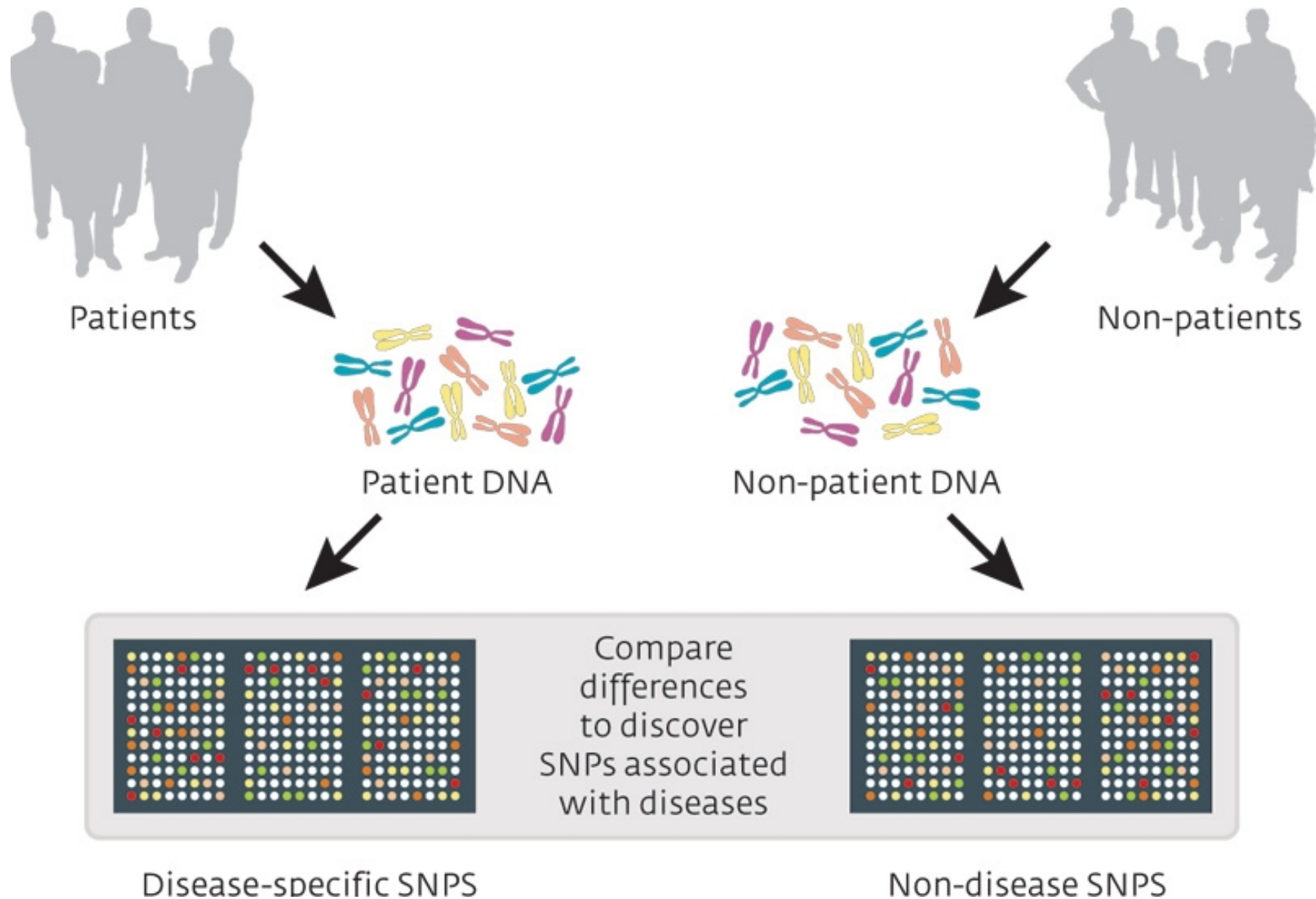
Chips de genotipado

En los chips de genotipado se unen covalentemente o bien sondas específicas de alelo o sondas que permiten la extensión de una base para identificar el alelo concreto. Sirven para identificar el genotipo de una muestra en una colección amplia de posiciones.



<https://www.sciencedirect.com/topics/biochemistry-genetics-and-molecular-biology/snp-array>

Estudios de asociación: GWAS



http://www.mpg.de/10680/Modern_psychiatry

Estudios de asociación: GWAS



GWAS Catalog

The NHGRI-EBI Catalog of human genome-wide association studies

Search the catalog



Examples: breast carcinoma, rs7329174, Yao, 2q37.1, HBS1L, 6:16000000-25000000

We now accept direct submissions of summary statistics for both published and pre-published/unpublished GWAS through our new [submission page](#)! See the [documentation](#) for detailed instructions.

Download

Download a full copy of the GWAS Catalog in spreadsheet format as well as current and older versions of the GWAS diagram in SVG format.

Summary statistics

Documentation and access to full summary statistics for GWAS Catalog studies where available.

Submit

Submit summary statistics to GWAS Catalog.

Documentation

Including FAQs, our curation process, training materials, related resources, a list of abbreviations and API documentation.

Diagram

Explore an interactive visualisation of all SNP-trait associations with genome-wide significance ($p \leq 5 \times 10^{-8}$).

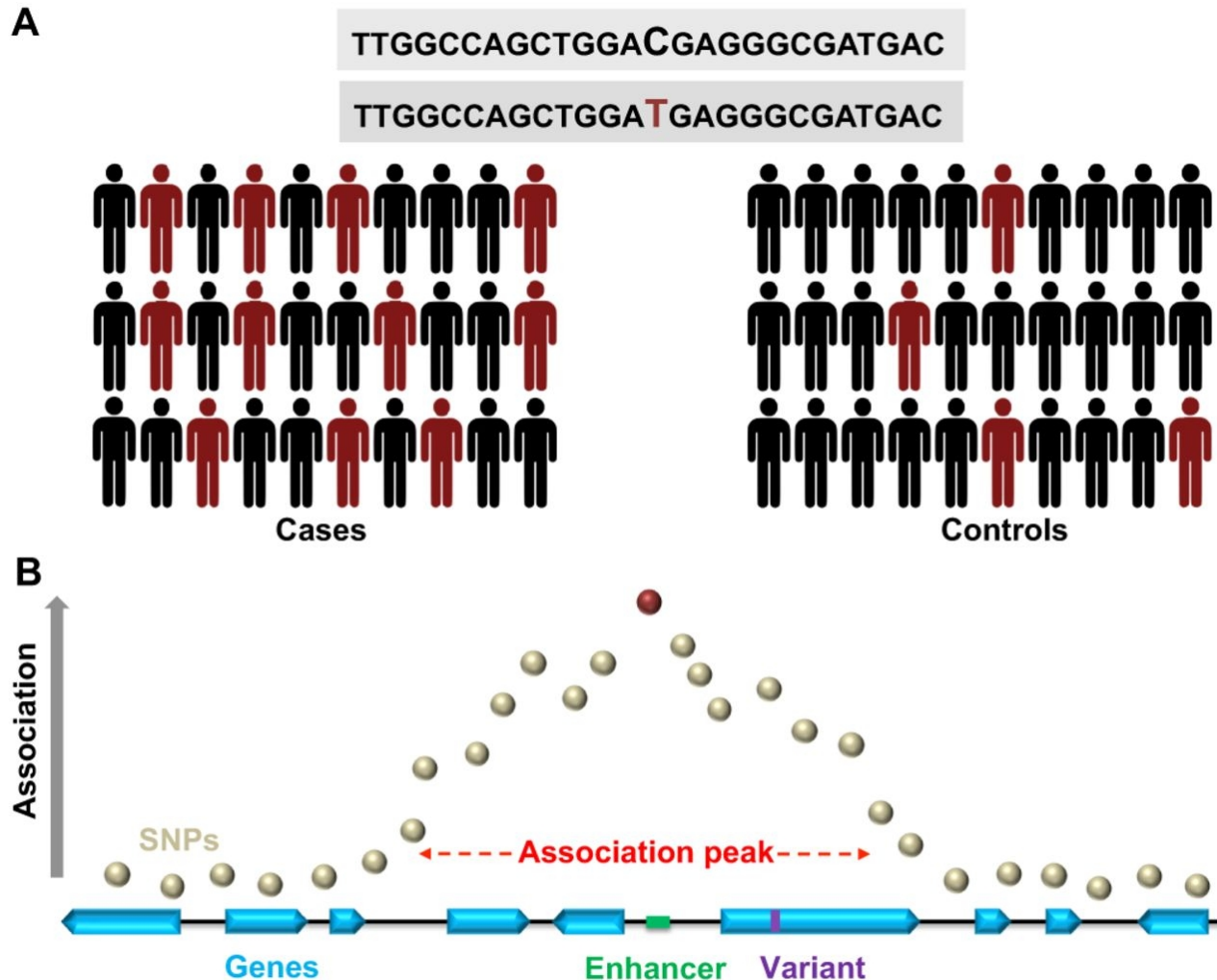
Ancestry

An introduction to our ancestry curation process.

¡Actualmente la base de datos recoge los resultados de cerca de 4800 artículos con más de 216.000 asociaciones!

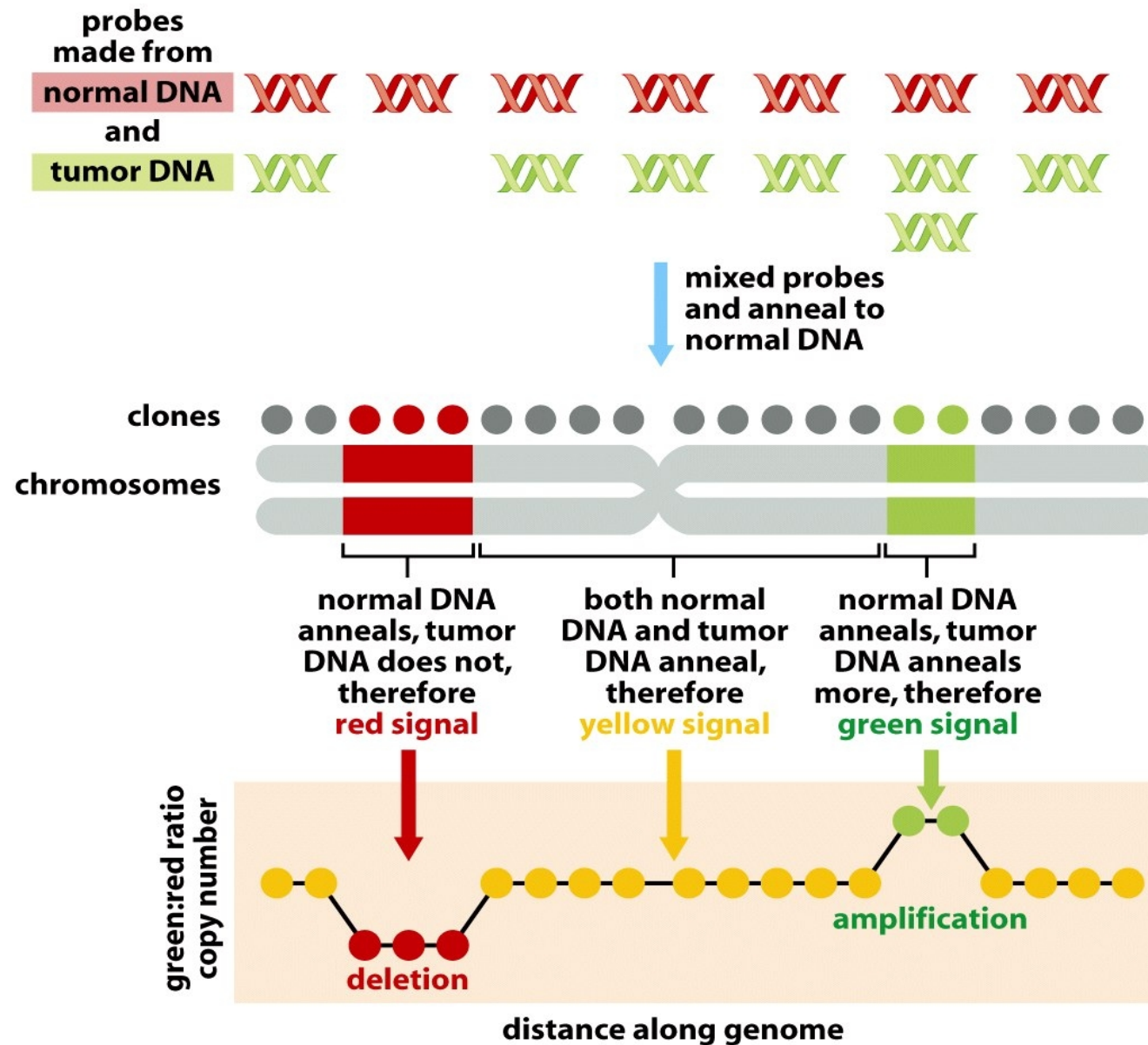
Estudios de asociación: GWAS

La asociación de un polimorfismo con una enfermedad no implica que ese polimorfismo sea la causa funcional de la predisposición a esa enfermedad.



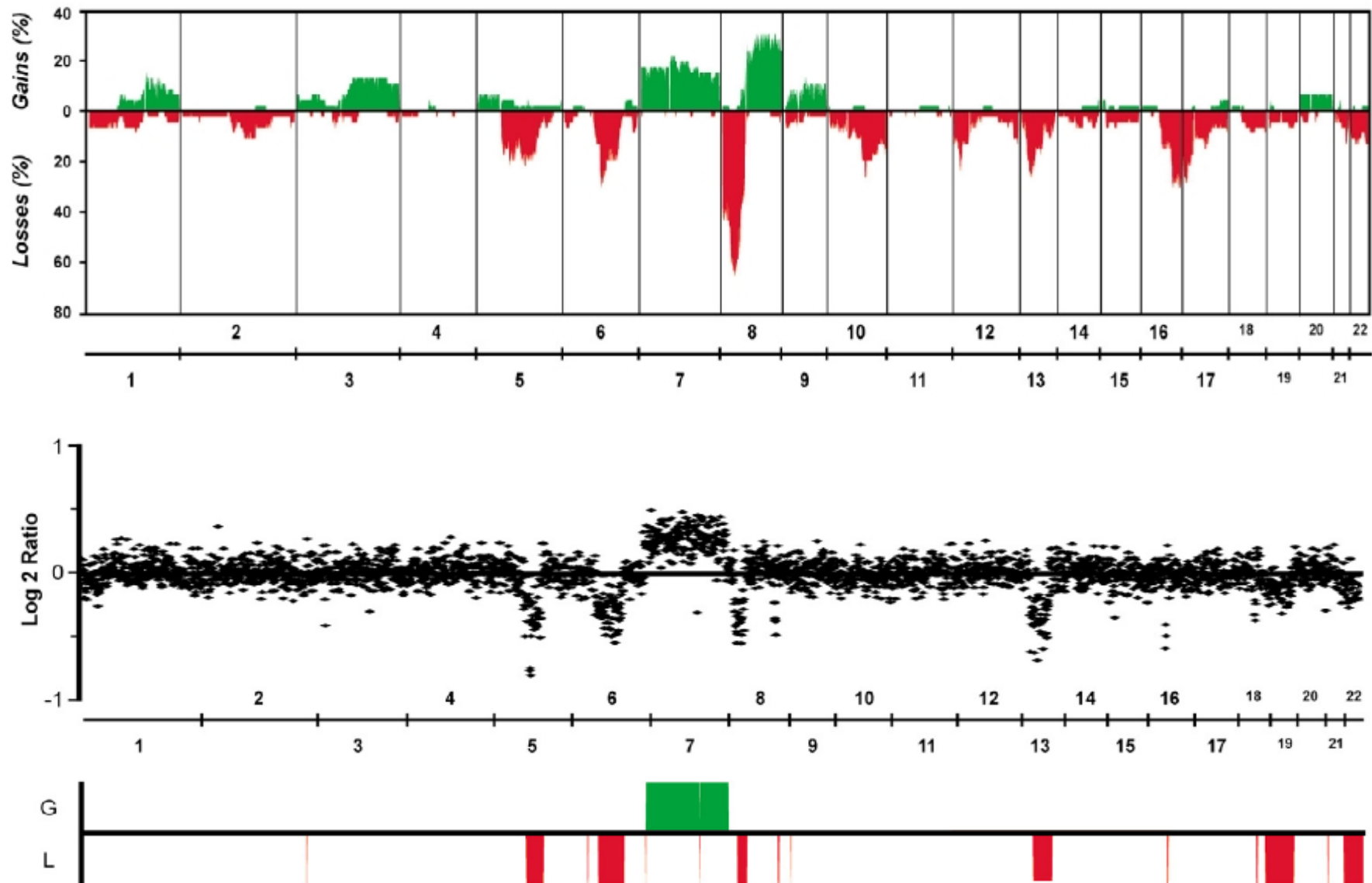
<https://dmm.biologists.org/content/10/2/77>

Hibridación comparativa de genoma (CGH)



Robert Weinberg. *The Biology of Cancer*. Garland Science(2007)

Hibridación comparativa de genoma (CGH)



Ribeiro et al. Molecular Cancer 5:33 (2006)

Tema 7: Identificación de alteraciones genómicas

7.1 Estudios genómicos utilizando chips genéticos

7.2 Preparación de librerías genómicas

7.3 Alineamiento de secuencias

7.4 Identificación de sustituciones

7.5 Identificación de pequeñas inserciones y deleciones

7.6 Identificación de cambios en el número de copia

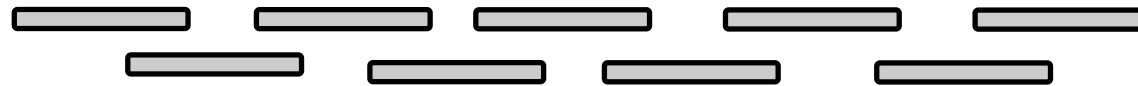
7.7 Identificación de grandes reorganizaciones genómicas

7.8 Anotación funcional de mutaciones

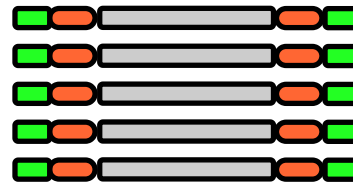
Librerías de genoma completo

DNA

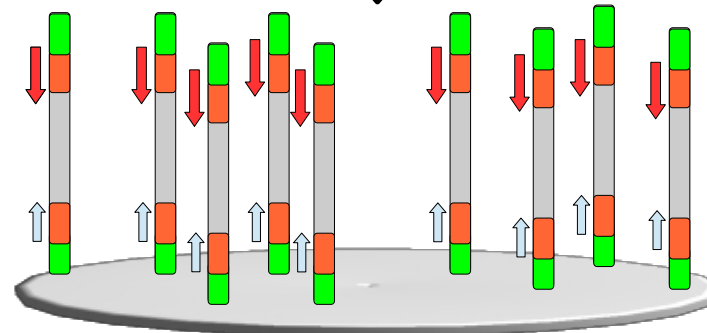
Fragmentación



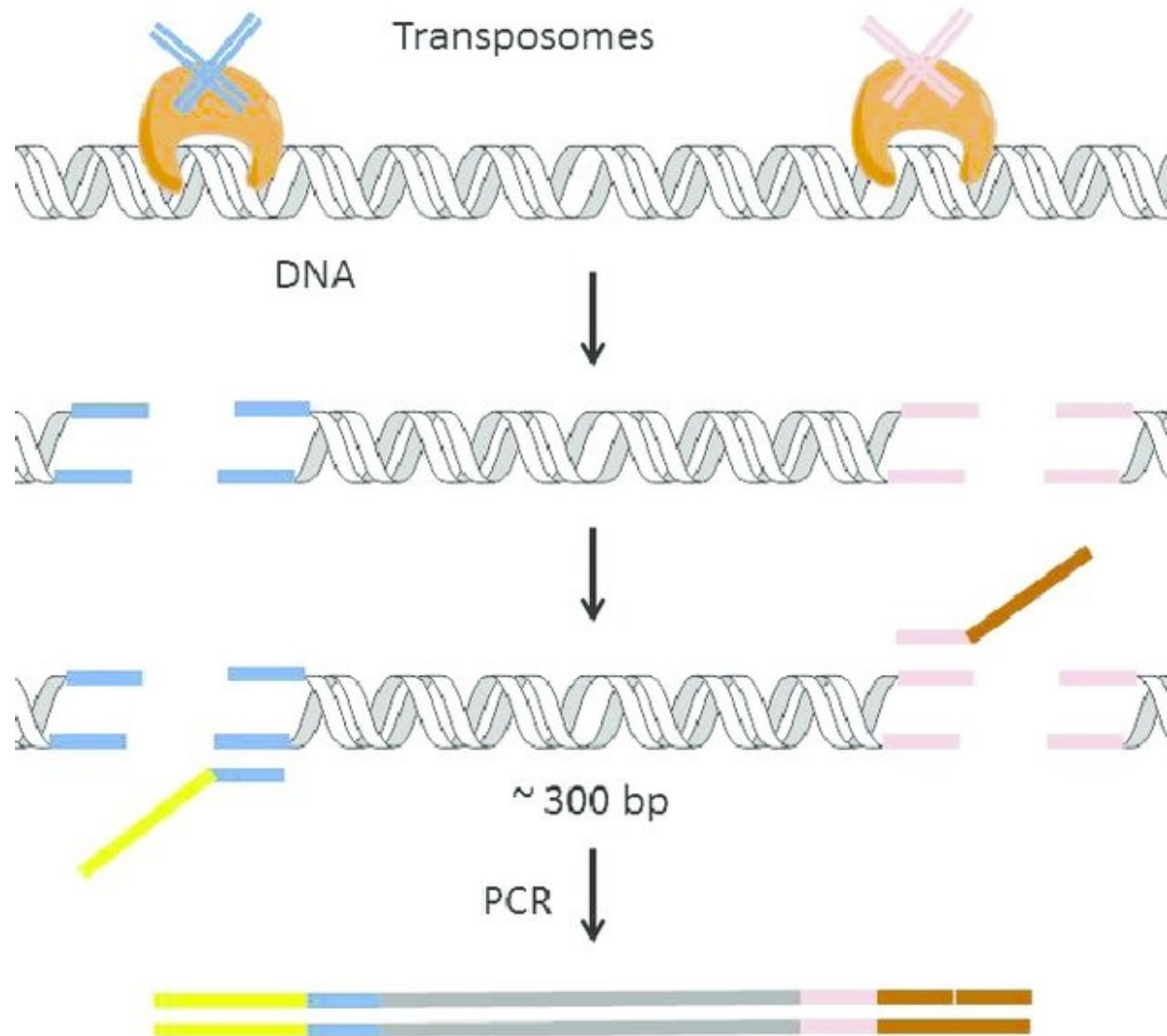
Ligación de
adaptadores



Formación de
clústers

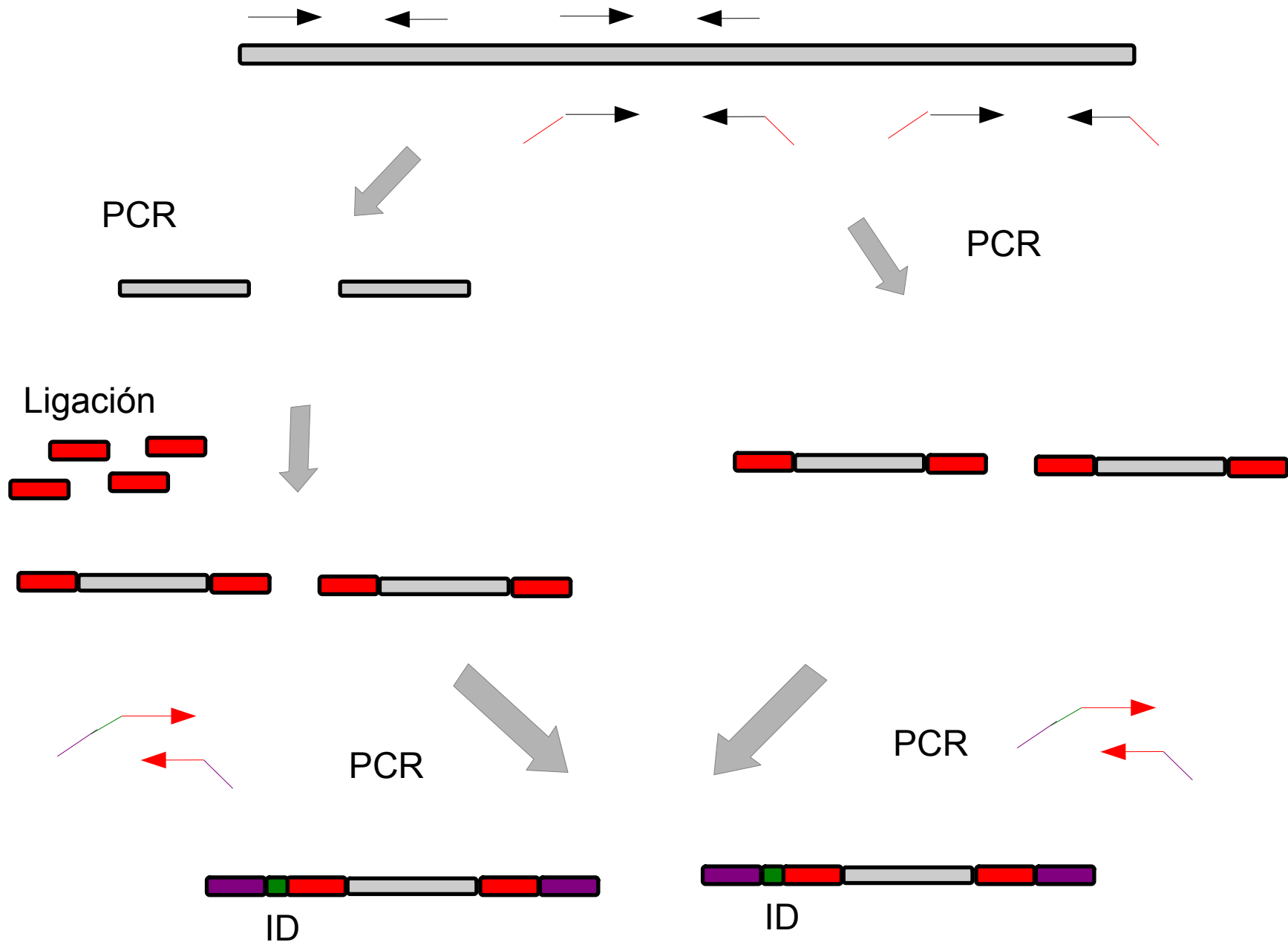


Variante mediante el uso de una trasposasa



Head SR *et al.* Biotechniques 56:61

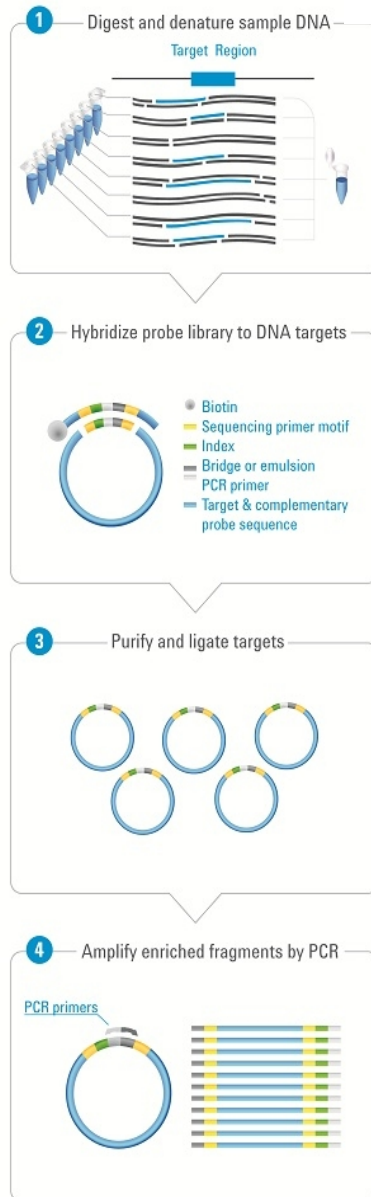
Variante utilizando PCR



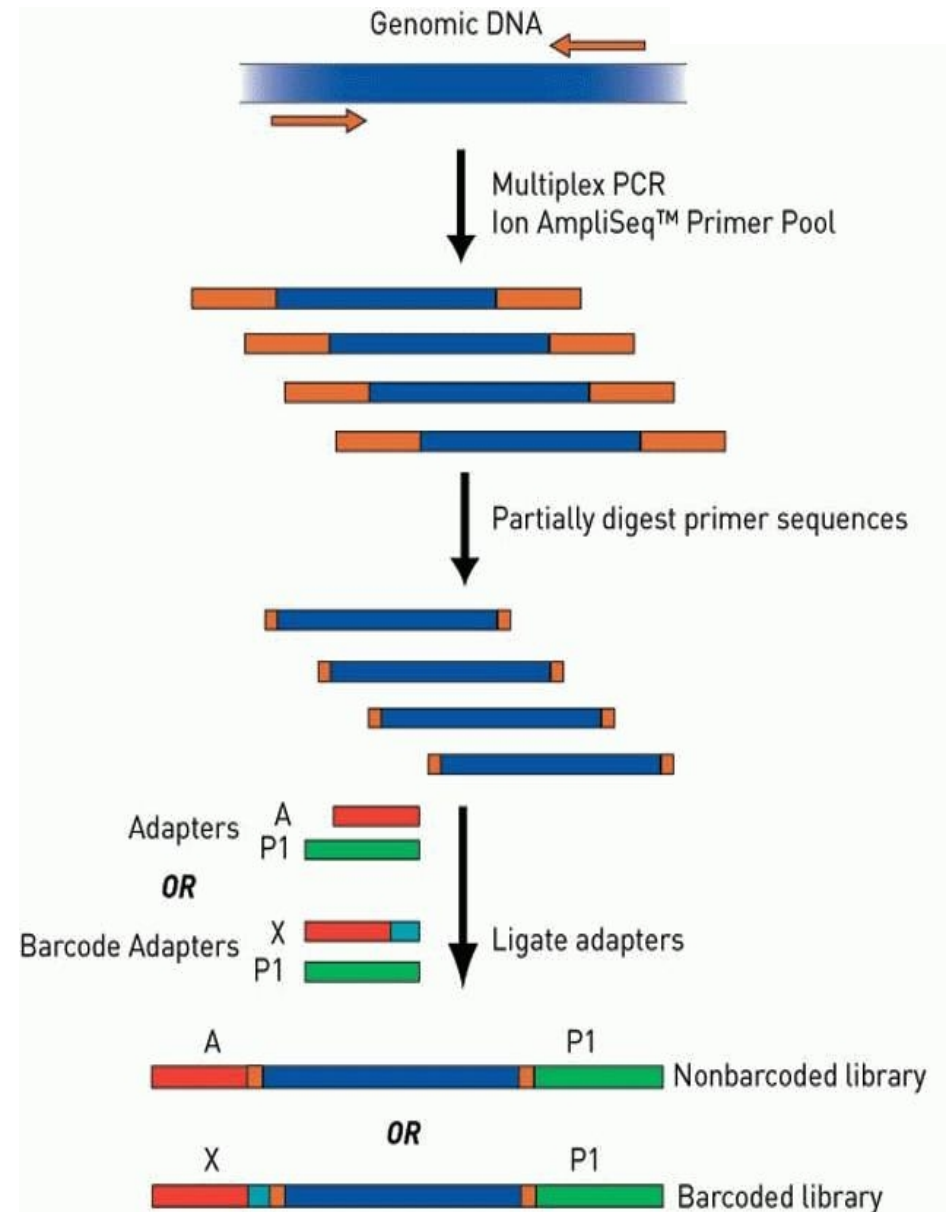
! Problemas con la conservación de la estequiometría

Kits comerciales multiplex

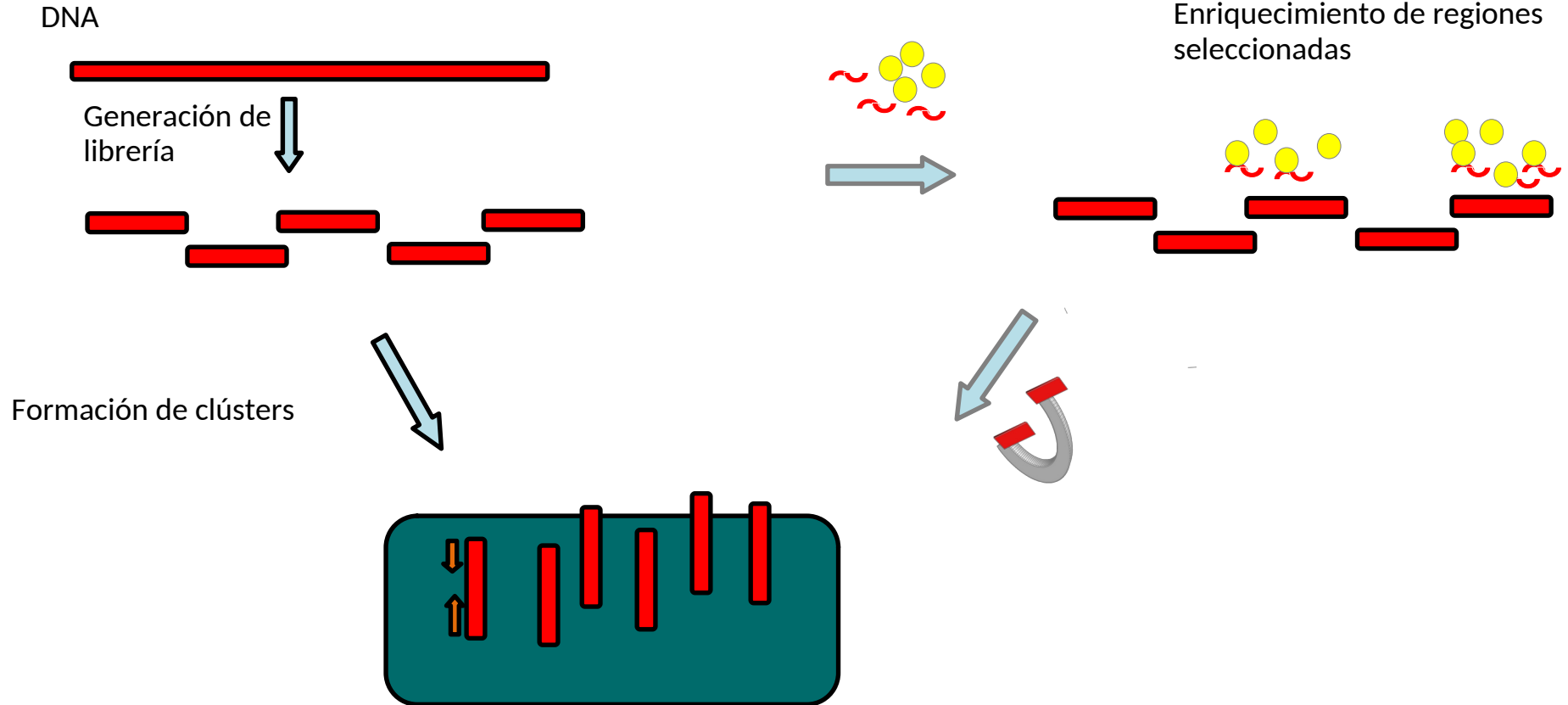
Agilent: Haloplex



Life Techn: Ampliseq



Enriquecimiento con sondas



! Problemas con la conservación de la estequiometría

Tema 7: Identificación de alteraciones genómicas

7.1 Estudios genómicos utilizando chips genéticos

7.2 Preparación de librerías genómicas

7.3 Alineamiento de secuencias

7.4 Identificación de sustituciones

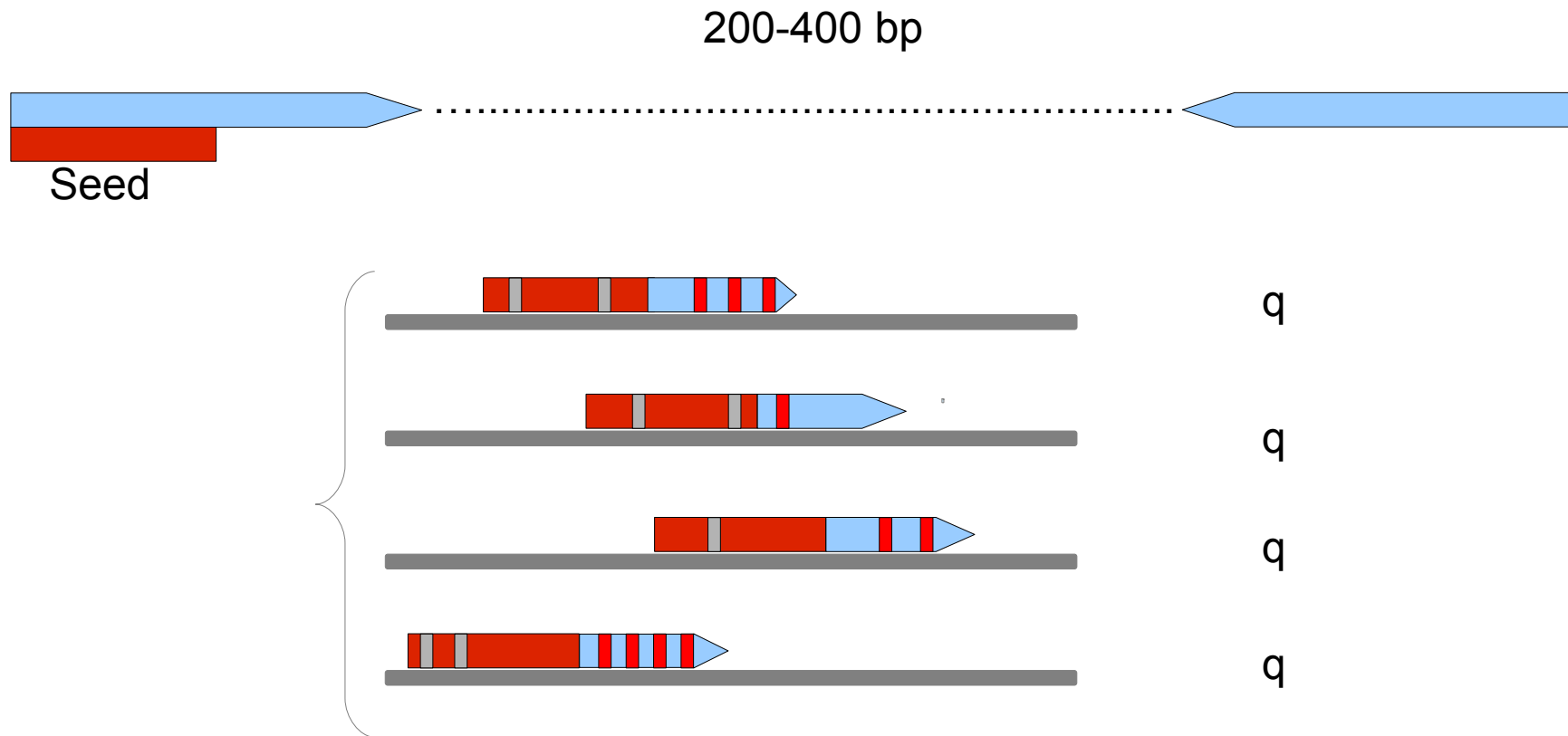
7.5 Identificación de pequeñas inserciones y deleciones

7.6 Identificación de cambios en el número de copia

7.7 Identificación de grandes reorganizaciones genómicas

7.8 Anotación funcional de mutaciones

Alineamiento frente a un genoma de referencia



$$\text{Al. Score} = \text{SeqQ} + \text{nhits} + \text{diffOtherhit} > \text{Threshold}$$

Problemas en el alineamiento de secuencias

1- Tenemos que localizar pequeñas secuencias (<100 bp) en un genoma de referencia muy grande (genoma humano 3×10^9 bp).

2- Los genomas eucariotas contienen secuencias con un alto grado de similitud: Duplicaciones segmentales, secuencias repetitivas, pseudogenes, familias génicas, etc.

3- La secuencia de cualquier individuo contiene de media 3×10^6 mutaciones en comparación al genoma de referencia. Algunas veces hay grandes reorganizaciones genómicas como es el caso del locus de las inmunoglobulinas.

Indexado del genoma para mayor rapidez

Software

Open Access

Ultrafast and memory-efficient alignment of short DNA sequences to the human genome

Ben Langmead, Cole Trapnell, Mihai Pop and Steven L Salzberg

Langmead B et al. Genome Biology 2009, 10:R25

Transformación de Burrows-Wheeler

Burrows Wheeler Transform

- Block-sorting compression scheme [Burrows et al, 1994]
- Rearranges the positions of the data such that the few distinct values under the same previous context are grouped together in position.

tennessee*	tennessee*
ennessee*t	*tennessee
nnessee*te	ssee*tenne
nessee*ten	e*tennesse
essee*tenn	nnessee*te
ssee*tenne	nessee*ten
see*tennes	essee*tenn
ee*tenness	see*tennes
e*tennesse	ee*tenness
*tennessee	ennessee*t

An example of the Burrows-Wheeler transform.

$\text{bwt}(\text{tennessee}^*) = \text{t}^*\text{sennese}$. The matrix on the right is obtained by sorting the rows of the left matrix in right-to-left lexicographic order. * denotes end of the data block and can be considered as the smallest symbol.

Programas como BWA o Bowtie utilizan este indexado para alinear eficientemente las secuencias frente al genoma de referencia.

Formato SAM de alineamiento

Encabezamiento

Tag	Description
@HD	The header line. The first line if present.
VN*	Format version. <i>Accepted format:</i> / [^] [0-9]+\ [.] [0-9]+\$/.
SO	Sorting order of alignments. <i>Valid values:</i> 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.
@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 PNEXT fields. Regular expression: [!-)+-<>-~][!-~]*
LN*	Reference sequence length. <i>Range:</i> [1,2 ²⁹ -1]
AS	Genome assembly identifier.
M5	MD5 checksum of the sequence in the uppercase, with gaps and spaces removed.
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.

```
@HD VN:1.5 G0:none SO:coordinate
@SQ SN:Chr1 LN:249250621
@SQ SN:Chr2 LN:243199373
@SQ SN:Chr3 LN:198022430
@SQ SN:Chr4 LN:191154276
@SQ SN:Chr5 LN:180915260
@SQ SN:Chr6 LN:171115067
@SQ SN:Chr7 LN:159138663
@SQ SN:Chr8 LN:146364022
@SQ SN:Chr9 LN:141213431
```

Formato SAM de alineamiento

Información de cada lectura en una línea

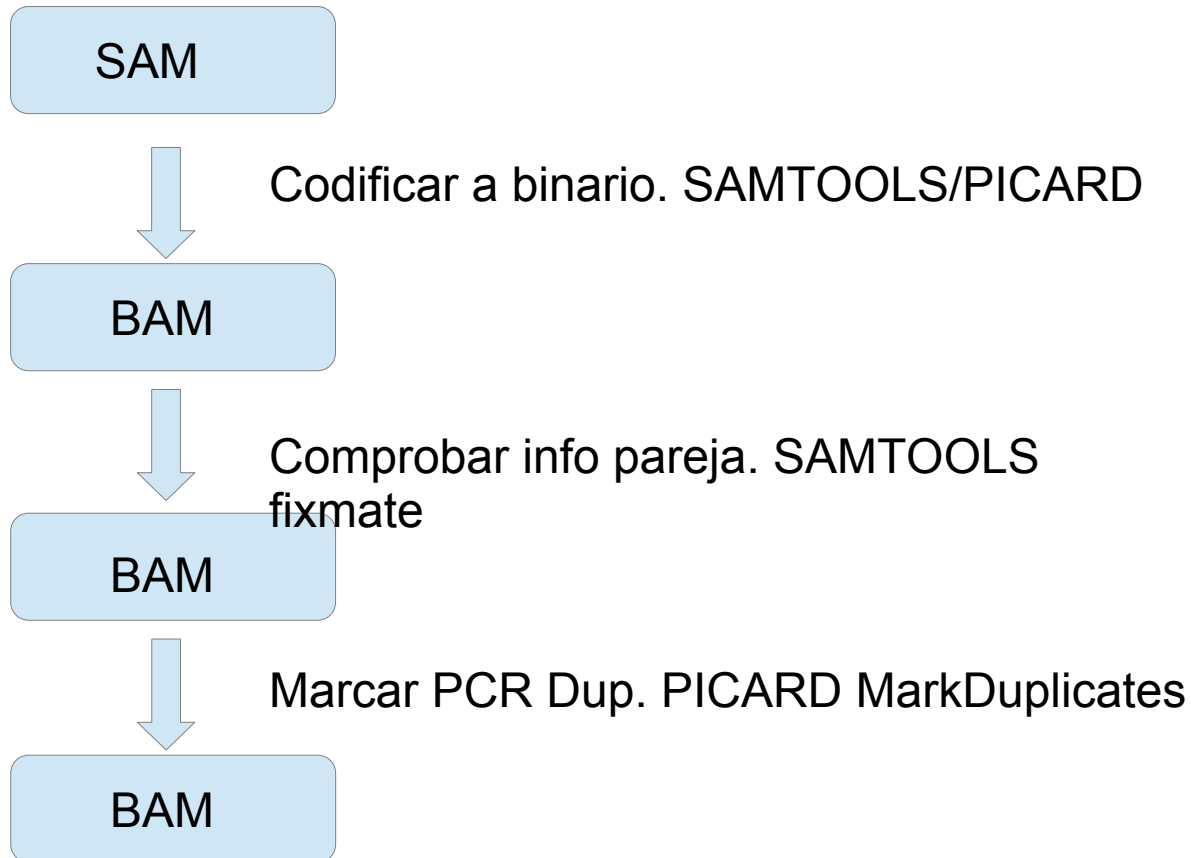
Col	Field	Type	Regexp/Range	Brief description
1	QNAME	String	[!-?A-~]{1,255}	Query template NAME
2	FLAG	Int	[0,2 ¹⁶ -1]	bitwise FLAG
3	RNAME	String	* [!-()+-<>-~] [!-~]*	Reference sequence NAME
4	POS	Int	[0,2 ²⁹ -1]	1-based leftmost mapping POSition
5	MAPQ	Int	[0,2 ⁸ -1]	MAPping Quality
6	CIGAR	String	* ([0-9]+[MIDNSHPX=])+	CIGAR string
7	RNEXT	String	* = [!-()+-<>-~] [!-~]*	Ref. name of the mate/next segment
8	PNEXT	Int	[0,2 ²⁹ -1]	Position of the mate/next segment
9	TLEN	Int	[-2 ²⁹ +1,2 ²⁹ -1]	observed Template LENgth
10	SEQ	String	* [A-Za-z=.] +	segment SEQUENCE
11	QUAL	String	[!-~]+	ASCII of Phred-scaled base QUALity+33

```

M03913:60:000000000-D46KB:1:1101:10044:17328 73 Chr1 1466216 0 71S31M48S = 1466216 0 ATCGGAAGAGCGGTTTCAGCAGGAA
TATGCCGCTTCTGCTTGAAGAAAAAATAAGATAATCAATCACCCGACACAAAAACCTAGGAAAATATCAGGAAAATAAAA EEEEEEEFFCGGGGGGGGGGGHHHHGGGGGGGGHHHHHHHH
HHGGGGEE/><////////11111111111111<1</.-----/.....</00/01//10111100///01000 MD:Z:31 RG:Z:Nacho NM:i:0 AS:i:31 XS:i:29 RX:Z:TCAGC

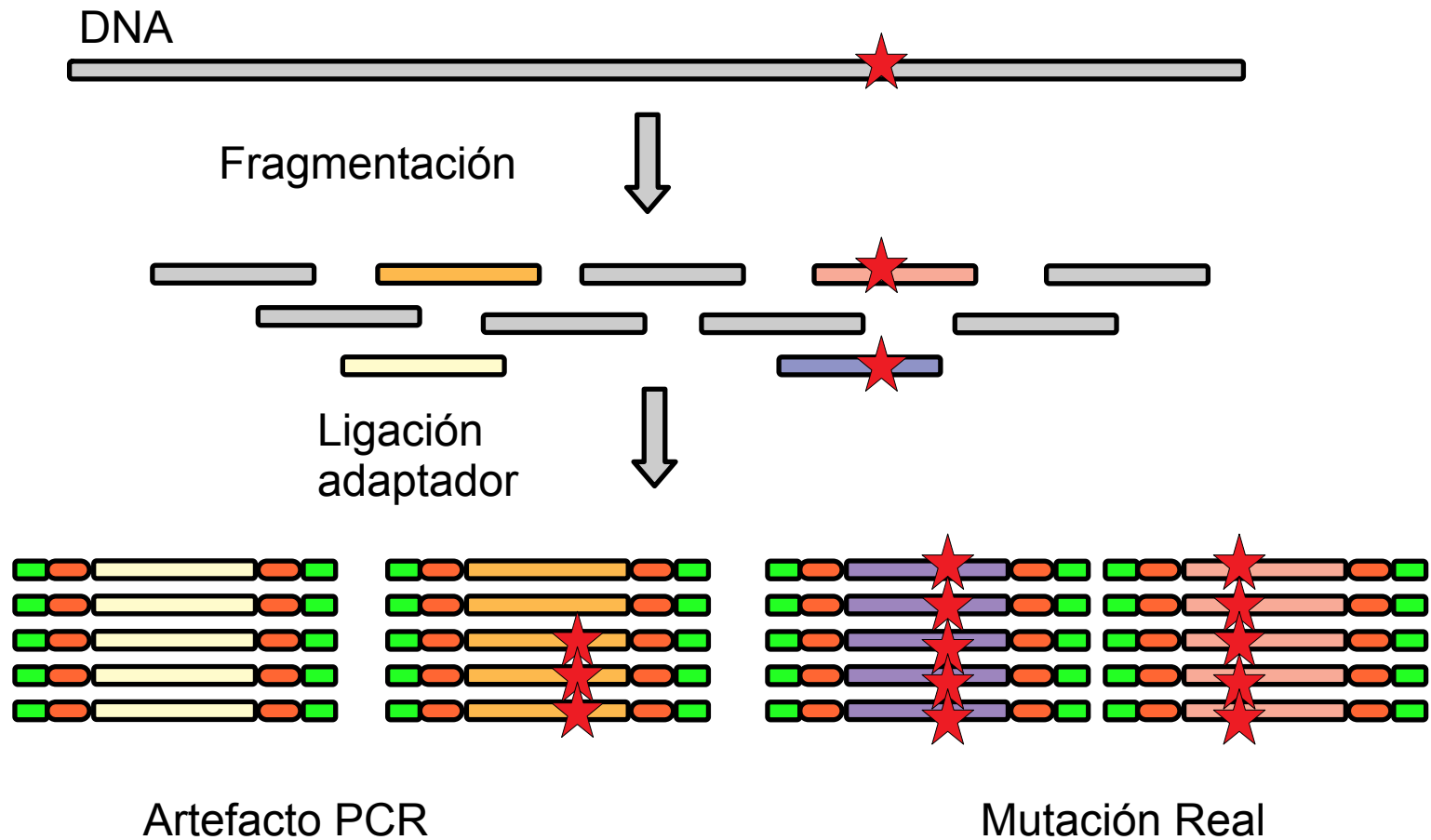
```

Procesado del alineamiento



Procesado del alineamiento

Marcado duplicados PCR



Tema 7: Identificación de alteraciones genómicas

7.1 Estudios genómicos utilizando chips genéticos

7.2 Preparación de librerías genómicas

7.3 Alineamiento de secuencias

7.4 Identificación de sustituciones

7.5 Identificación de pequeñas inserciones y deleciones

7.6 Identificación de cambios en el número de copia

7.7 Identificación de grandes reorganizaciones genómicas

7.8 Anotación funcional de mutaciones

Problemas en la identificación de mutaciones

1- Durante la generación de librerías, diversas reacciones enzimáticas pueden producir daños/mutaciones en el DNA. Sobre todo la sonicación y los pasos de PCR.

2- La secuenciación se basa en síntesis de DNA que tiene sus propias tasas de error.

3- Cualquier error en el paso de alineamiento puede afectar a la identificación de mutaciones. Sobre todo en regiones con un alto grado de homología (familias génicas)

4- En el caso de muestras con alta inestabilidad genética (por ejemplo en muestras de cáncer), la presencia de aneuploidías, grandes alteraciones genéticas o de DNA de células normales puede dificultar mucho la identificación de mutaciones somáticas.

Identificación de sustituciones

The diagram consists of 40 rows of text, each containing a sequence of characters and symbols. The rows are color-coded: yellow, blue, and red. The sequence of characters and symbols is as follows:

- Row 1: t a t a c a w o w o a C a a t a t t w o a t w o
- Row 2: t a t a c a w o w o a a C a a t a t t w o a t w o
- Row 3: t a t a c a w o w o a C a a t a t t w o a t w o
- Row 4: t a t a c a w o w o a a C a a t a t t w o a t w o
- Row 5: t a t a c a w o w o a a C a a t a t t w o a t w o
- Row 6: t a t a c a w o w o a a C a a t a t t w o a t w o
- Row 7: t a t a c a w o w o a C a a t a t t w o a t w o
- Row 8: t a t a c a w o w o a a C a a t a t t w o a t w o
- Row 9: t a t a c a w o w o a a C a a t a t t w o a t w o
- Row 10: t a t a c a w o w o a a C a a t a t t w o a t w o
- Row 11: t a t a c a w o w o a a C a a t a t t w o a t w o
- Row 12: t a t a c a w o w o a a C a a t a t t w o a t w o
- Row 13: t a t a c a w o w o a a C a a t a t t w o a t w o
- Row 14: t a t a c a w o w o a a C a a t a t t w o a t w o
- Row 15: t a t a c a w o w o a a C a a t a t t w o a t w o
- Row 16: t a t a c a w o w o a a C a a t a t t w o a t w o
- Row 17: t a t a c a w o w o a a C a a t a t t w o a t w o
- Row 18: t a t a c a w o w o a a C a a t a t t w o a t w o
- Row 19: t a t a c a w o w o a a C a a t a t t w o a t w o
- Row 20: t a t a c a w o w o a a C a a t a t t w o a t w o
- Row 21: t a t a c a w o w o a a C a a t a t t w o a t w o
- Row 22: t a t a c a w o w o a a C a a t a t t w o a t w o
- Row 23: t a t a c a w o w o a a C a a t a t t w o a t w o
- Row 24: t a t a c a w o w o a a C a a t a t t w o a t w o
- Row 25: t a t a c a w o w o a a C a a t a t t w o a t w o
- Row 26: t a t a c a w o w o a a C a a t a t t w o a t w o
- Row 27: t a t a c a w o w o a a C a a t a t t w o a t w o
- Row 28: t a t a c a w o w o a a C a a t a t t w o a t w o
- Row 29: t a t a c a w o w o a a C a a t a t t w o a t w o
- Row 30: t a t a c a w o w o a a C a a t a t t w o a t w o
- Row 31: t a t a c a w o w o a a C a a t a t t w o a t w o
- Row 32: t a t a c a w o w o a a C a a t a t t w o a t w o
- Row 33: t a t a c a w o w o a a C a a t a t t w o a t w o
- Row 34: t a t a c a w o w o a a C a a t a t t w o a t w o
- Row 35: t a t a c a w o w o a a C a a t a t t w o a t w o
- Row 36: t a t a c a w o w o a a C a a t a t t w o a t w o
- Row 37: t a t a c a w o w o a a C a a t a t t w o a t w o
- Row 38: t a t a c a w o w o a a C a a t a t t w o a t w o
- Row 39: t a t a c a w o w o a a C a a t a t t w o a t w o
- Row 40: t a t a c a w o w o a a C a a t a t t w o a t w o

Identificación de sustituciones

The diagram consists of a 20x20 grid of small yellow rectangles, each containing a black dot. The rectangles are arranged in a grid pattern, with some rows and columns highlighted in blue. The diagram is labeled 'Figure 1' at the top left.

Identificación de sustituciones

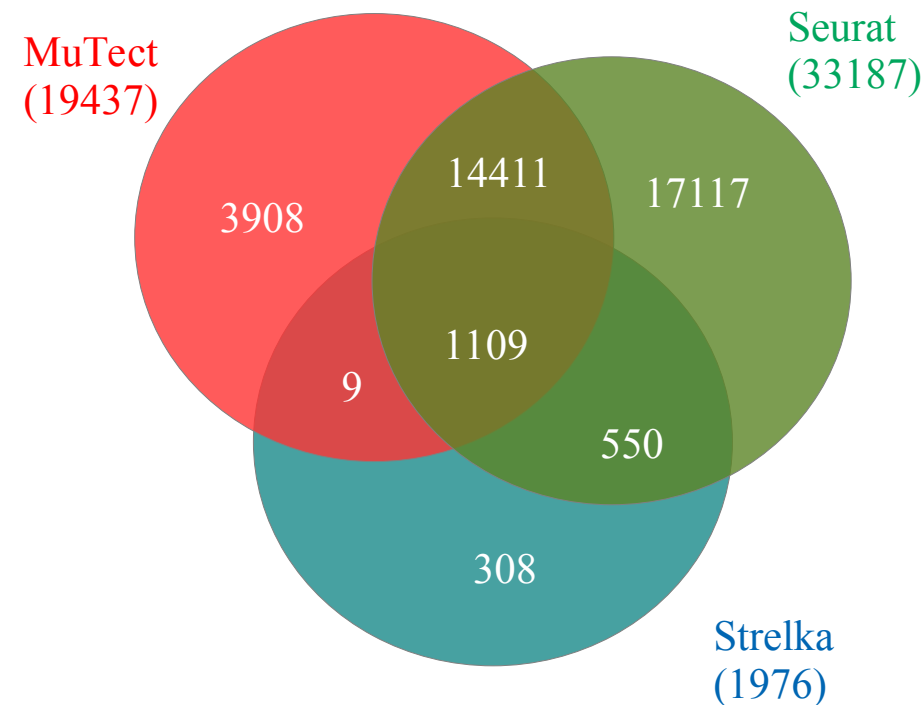
Variant caller	Type of variant	Single-sample mode	Type of core algorithm
BAYSIC [48]	SNV	No	Machine learning (ensemble caller)
CaVEMan [34]	SNV	No	Joint genotype analysis
deepSNV [38]	SNV	No	Allele frequency analysis
EBCall [37]	SNV, indel	No	Allele frequency analysis
FaSD-somatic [31]	SNV	Yes	Joint genotype analysis
FreeBayes [44]	SNV, indel	Yes	Haplotype analysis
HapMuC [42]	SNV, indel	Yes	Haplotype analysis
JointSNVMix2 [30]	SNV	No	Joint genotype analysis
LocHap [43]	SNV, indel	No	Haplotype analysis
LoFreq [36]	SNV, indel	Yes	Allele frequency analysis
LoLoPicker [39]	SNV	No	Allele frequency analysis
MutationSeq [45]	SNV	No	Machine learning
MuSE [40]	SNV	No	Markov chain model
MuTect [35]	SNV	Yes	Allele frequency analysis
SAMtools [8]	SNV, indel	Yes	Joint genotype analysis
Platypus [41]	SNV, indel, SV	Yes	Haplotype analysis
qSNP [24]	SNV	No	Heuristic threshold
RADIA [26]	SNV	No	Heuristic threshold
Seurat [33]	SNV, indel, SV	No	Joint genotype analysis
Shimmer [25]	SNV, indel	No	Heuristic threshold
SNooPer [47]	SNV, indel	Yes	Machine learning
SNVSniffer [32]	SNV, indel	Yes	Joint genotype analysis
SOAPsnv [27]	SNV	No	Heuristic threshold
SomaticSeq [46]	SNV	No	Machine learning (ensemble caller)
SomaticSniper [28]	SNV	No	Joint genotype analysis
Strelka [17]	SNV, indel	No	Allele frequency analysis
TVC [97]	SNV, indel, SV	Yes	Ion Torrent specific
VarDict [18]	SNV, indel, SV	Yes	Heuristic threshold
VarScan2 [9]	SNV, indel	Yes	Heuristic threshold
Virmid [29]	SNV	No	Joint genotype analysis

ICGC Study Highlights Challenges in Consistent Somatic Variant Calling, Tips for Improvement

Dec 10, 2015 | [staff reporter](#)

NEW YORK (GenomeWeb) – A new report this week has added to a growing recognition of the challenges faced in calling somatic mutations in cancer genome sequences, despite continued advances in next-gen sequencing technologies and informatics.

In a study [appearing yesterday](#) in *Nature Communications*, a team from the International Cancer Genome Consortium published results both identifying weaknesses and laying the groundwork for more accurate characterization of somatic mutations in the ICGC's ongoing cancer genome sequencing efforts.



Tema 7: Identificación de alteraciones genómicas

7.1 Estudios genómicos utilizando chips genéticos

7.2 Preparación de librerías genómicas

7.3 Alineamiento de secuencias

7.4 Identificación de sustituciones

7.5 Identificación de pequeñas inserciones y deleciones

7.6 Identificación de cambios en el número de copia

7.7 Identificación de grandes reorganizaciones genómicas

7.8 Anotación funcional de mutaciones

Identificación de sustituciones

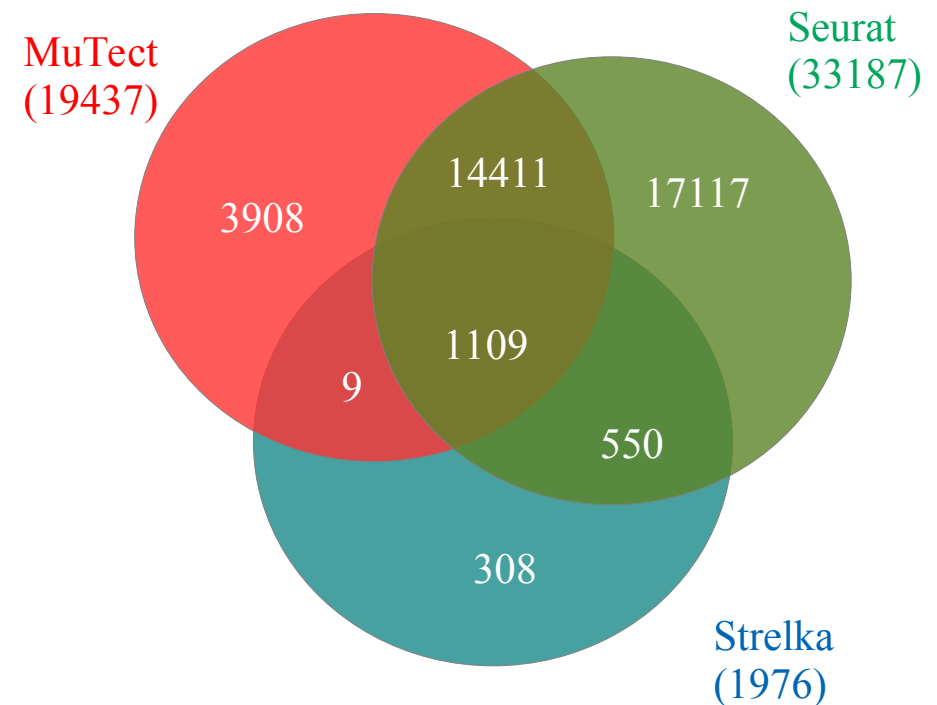
Variant caller	Type of variant	Single-sample mode	Type of core algorithm
BAYSIC [48]	SNV	No	Machine learning (ensemble caller)
CaVEMan [34]	SNV	No	Joint genotype analysis
deepSNV [38]	SNV	No	Allele frequency analysis
EBCall [37]	SNV, indel	No	Allele frequency analysis
FaSD-somatic [31]	SNV	Yes	Joint genotype analysis
FreeBayes [44]	SNV, indel	Yes	Haplotype analysis
HapMuC [42]	SNV, indel	Yes	Haplotype analysis
JointSNVMix2 [30]	SNV	No	Joint genotype analysis
LocHap [43]	SNV, indel	No	Haplotype analysis
LoFreq [36]	SNV, indel	Yes	Allele frequency analysis
LoLoPicker [39]	SNV	No	Allele frequency analysis
MutationSeq [45]	SNV	No	Machine learning
MuSE [40]	SNV	No	Markov chain model
MuTect [35]	SNV	Yes	Allele frequency analysis
SAMtools [8]	SNV, indel	Yes	Joint genotype analysis
Platypus [41]	SNV, indel, SV	Yes	Haplotype analysis
qSNP [24]	SNV	No	Heuristic threshold
RADIA [26]	SNV	No	Heuristic threshold
Seurat [33]	SNV, indel, SV	No	Joint genotype analysis
Shimmer [25]	SNV, indel	No	Heuristic threshold
SNooPer [47]	SNV, indel	Yes	Machine learning
SNVSniffer [32]	SNV, indel	Yes	Joint genotype analysis
SOAPsnv [27]	SNV	No	Heuristic threshold
SomaticSeq [46]	SNV	No	Machine learning (ensemble caller)
SomaticSniper [28]	SNV	No	Joint genotype analysis
Strelka [17]	SNV, indel	No	Allele frequency analysis
TVC [97]	SNV, indel, SV	Yes	Ion Torrent specific
VarDict [18]	SNV, indel, SV	Yes	Heuristic threshold
VarScan2 [9]	SNV, indel	Yes	Heuristic threshold
Virmid [29]	SNV	No	Joint genotype analysis

ICGC Study Highlights Challenges in Consistent Somatic Variant Calling, Tips for Improvement

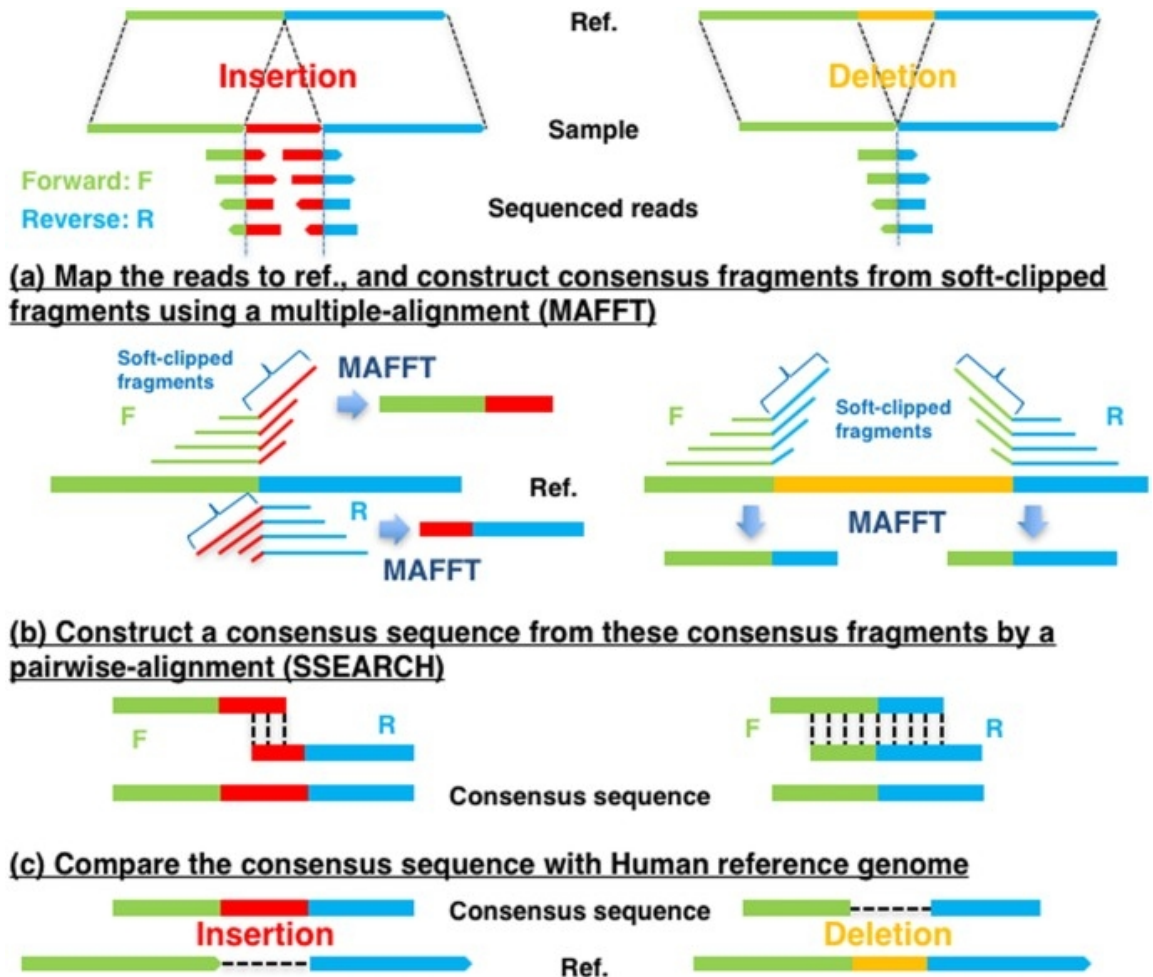
Dec 10, 2015 | [staff reporter](#)

NEW YORK (GenomeWeb) – A new report this week has added to a growing recognition of the challenges faced in calling somatic mutations in cancer genome sequences, despite continued advances in next-gen sequencing technologies and informatics.

In a study [appearing yesterday](#) in *Nature Communications*, a team from the International Cancer Genome Consortium published results both identifying weaknesses and laying the groundwork for more accurate characterization of somatic mutations in the ICGC's ongoing cancer genome sequencing efforts.



Identificación de indels



Shigemizu D et al. Sci Rep (2018) 8:5608

Programas como **IMSindel**, **PINDEL** o **DINDEL** utilizan realineamientos y ensamblaje para identificar indels que no haya encontrado el alineador

Tema 7: Identificación de alteraciones genómicas

7.1 Estudios genómicos utilizando chips genéticos

7.2 Preparación de librerías genómicas

7.3 Alineamiento de secuencias

7.4 Identificación de sustituciones

7.5 Identificación de pequeñas inserciones y deleciones

7.6 Identificación de cambios en el número de copia

7.7 Identificación de grandes reorganizaciones genómicas

7.8 Anotación funcional de mutaciones

Alteraciones en el número de copia

CNV

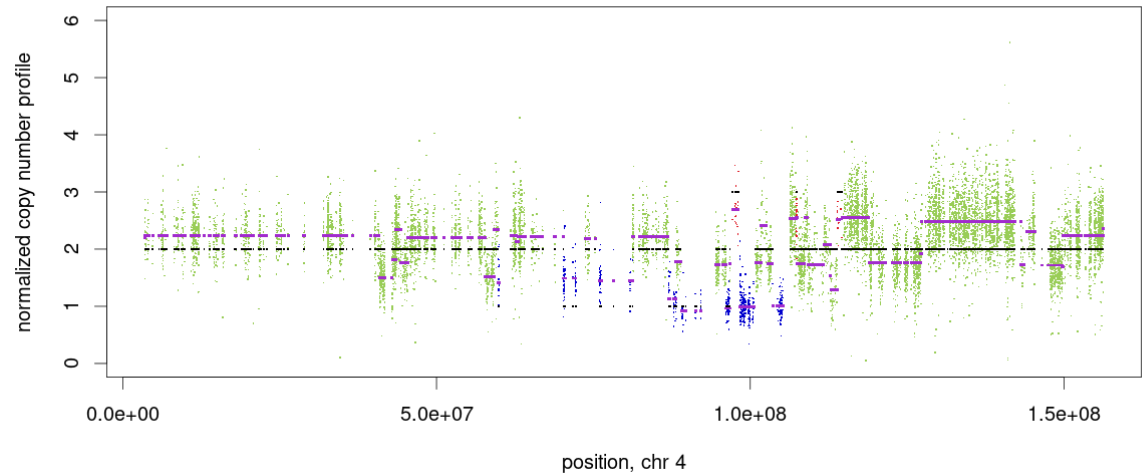
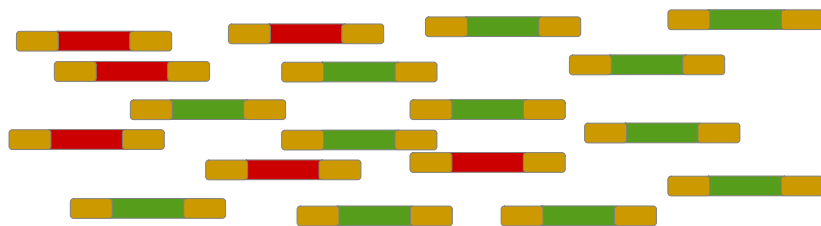
Cromosoma 1



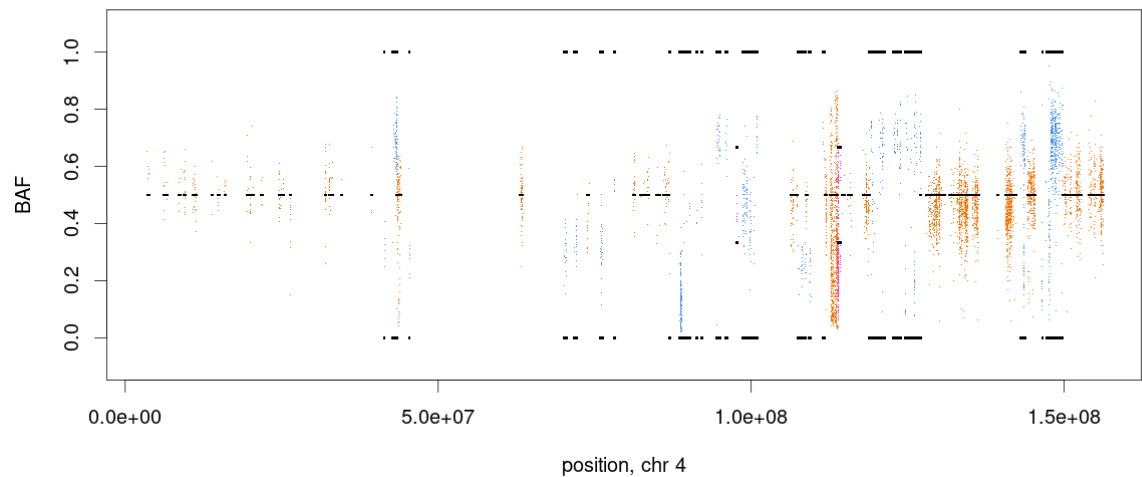
Cromosoma 2



Library Generation

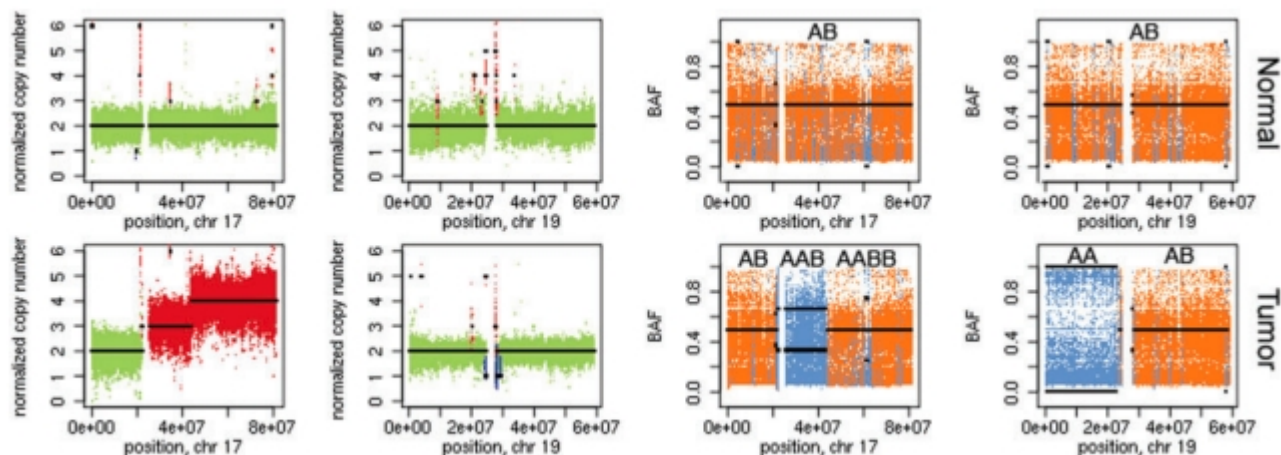


LOH



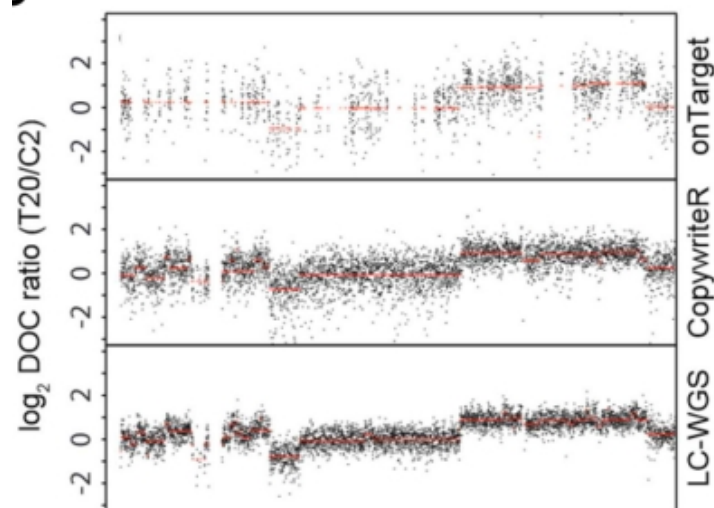
Problemas con sesgos por enriquecimiento

Uso de muestras control (*FreeC-Control*)



Boeva V et al .Bioinformatics 28:423–425

Uso de secuencias no enriquecidas (*CopywriteR*)



Kuilman T et al . Genome Biology 16:49

Tema 7: Identificación de alteraciones genómicas

7.1 Preparación de librerías genómicas

7.2 Alineamiento de secuencias

7.3 Identificación de sustituciones

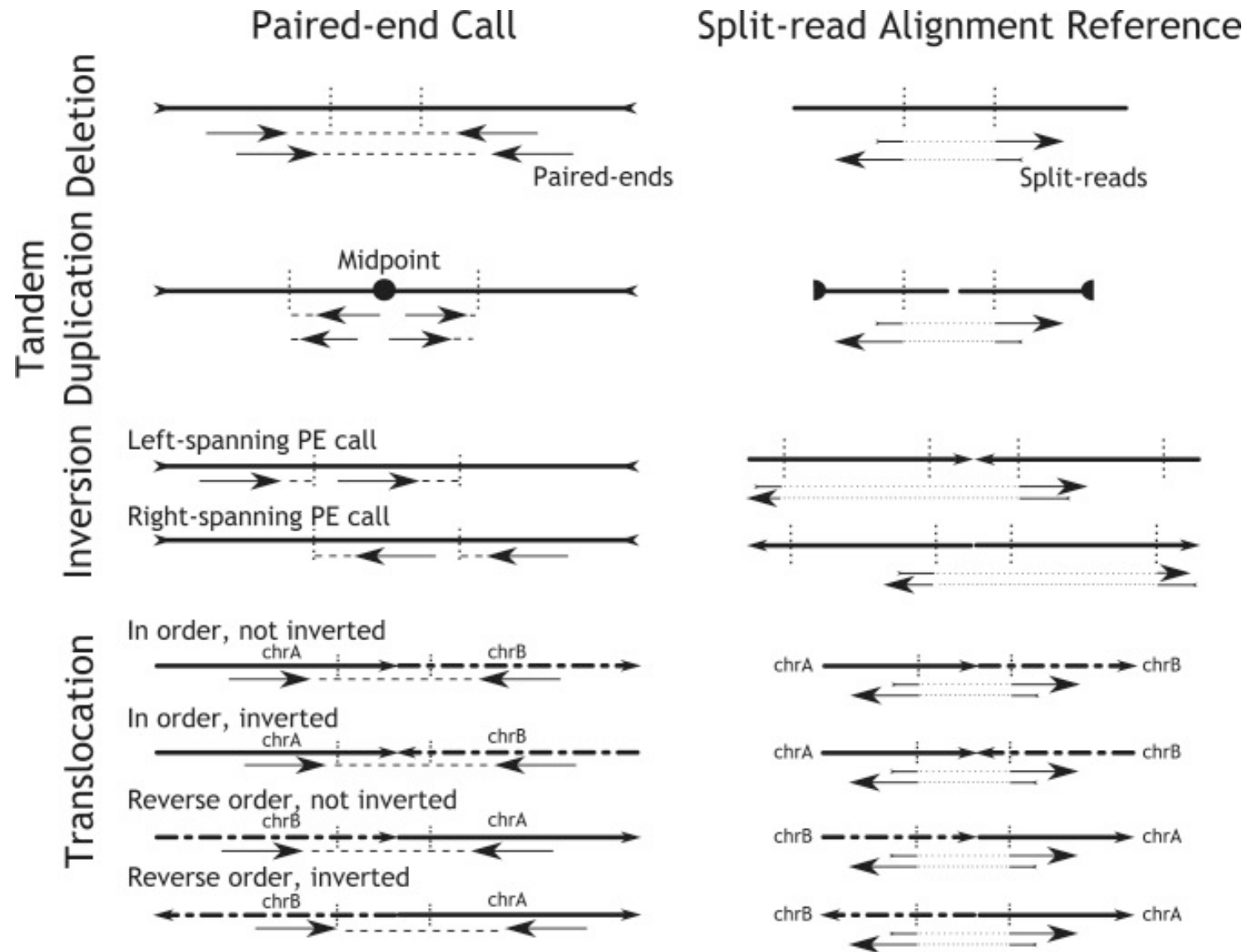
7.4 Identificación de pequeñas inserciones y deleciones

7.5 Identificación de cambios en el número de copia

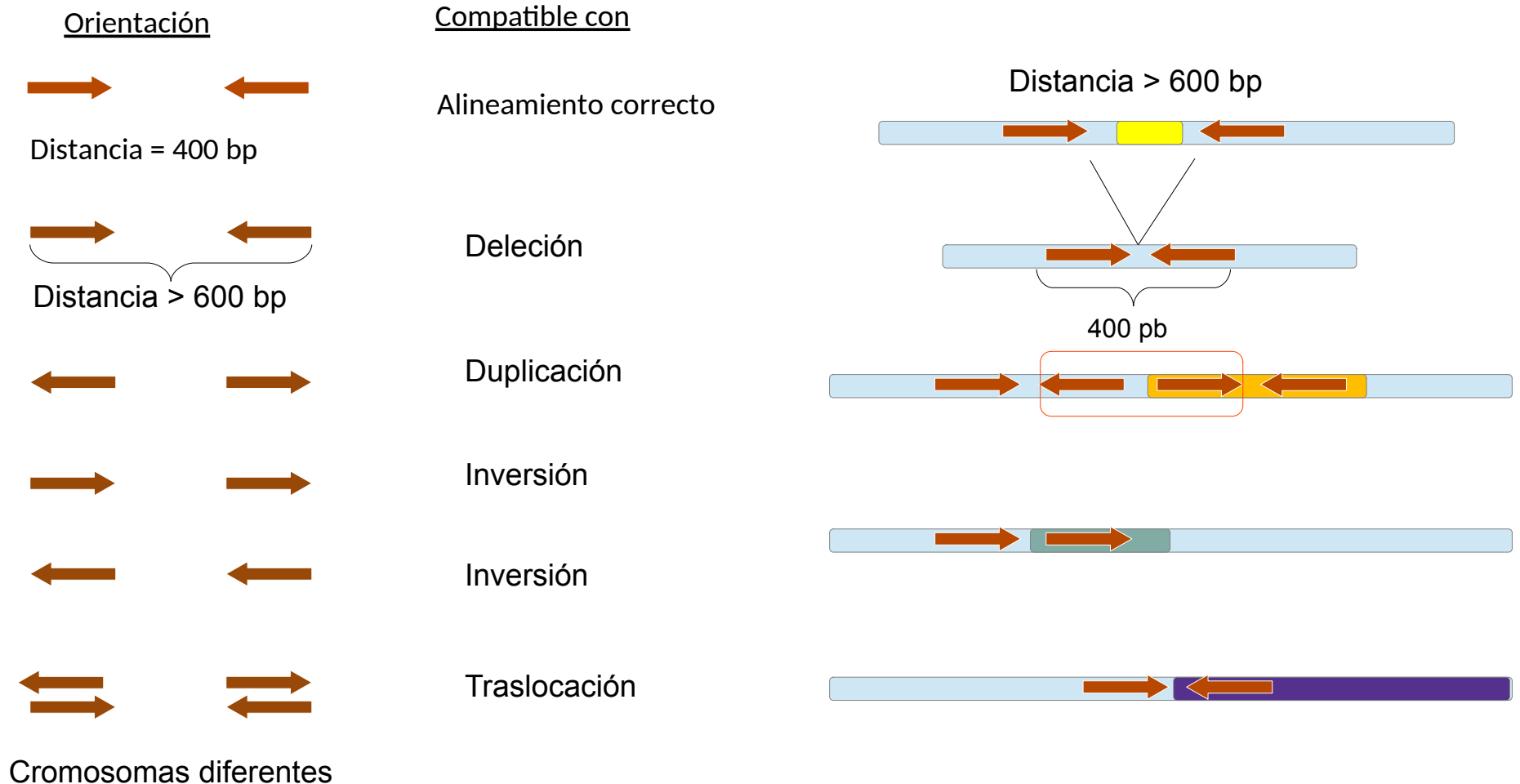
7.6 Identificación de grandes reorganizaciones genómicas

7.7 Anotación funcional de mutaciones

Lecturas pareadas discordantes y lecturas cortadas



Lecturas pareadas discordantes y lecturas cortadas



Programas como **BreakDancer** o **Delly** utilizan estas estrategias para identificar reorganizaciones genómicas en datos de NGS.

Tema 7: Identificación de alteraciones genómicas

7.1 Estudios genómicos utilizando chips genéticos

7.2 Preparación de librerías genómicas

7.3 Alineamiento de secuencias

7.4 Identificación de sustituciones

7.5 Identificación de pequeñas inserciones y deleciones

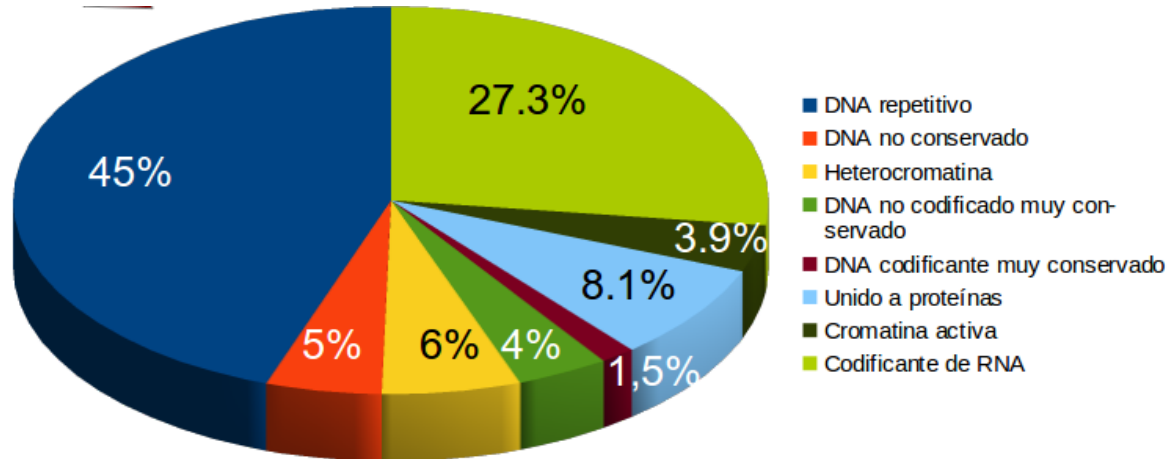
7.6 Identificación de cambios en el número de copia

7.7 Identificación de grandes reorganizaciones genómicas

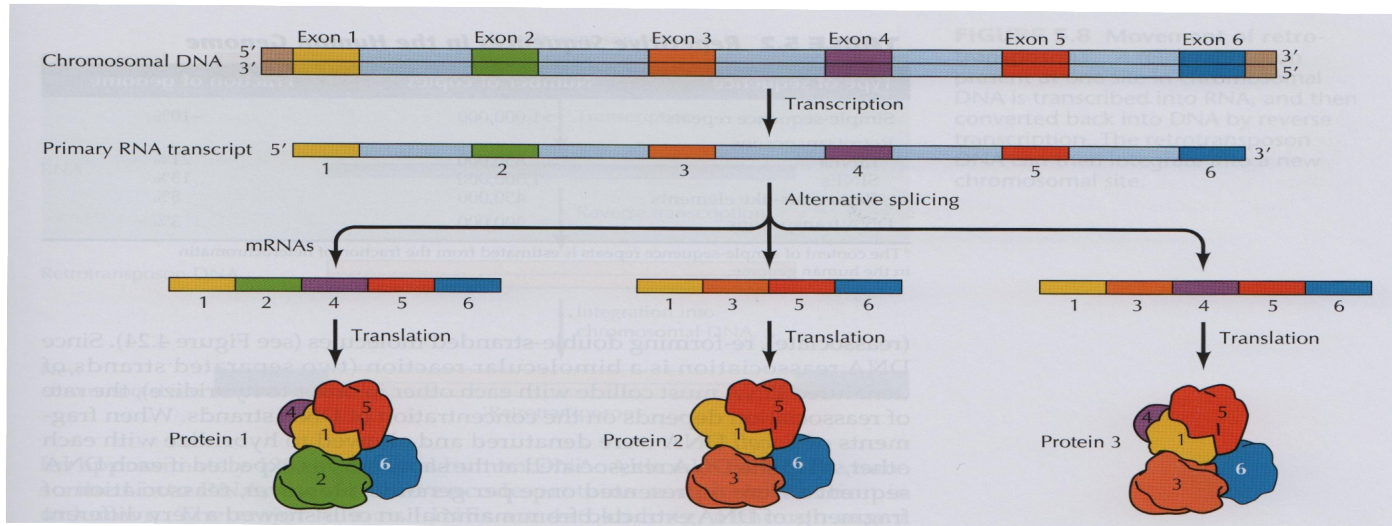
7.8 Anotación funcional de mutaciones

DNA/RNA en las células humanas

Mayoría de DNA no codificante



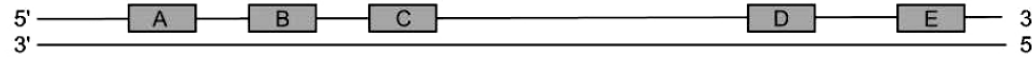
Los GENES son las regiones del DNA que producen RNA



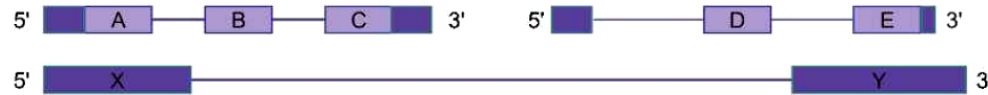
Distintos modelos de genes

¿Cuántos genes hay aquí?

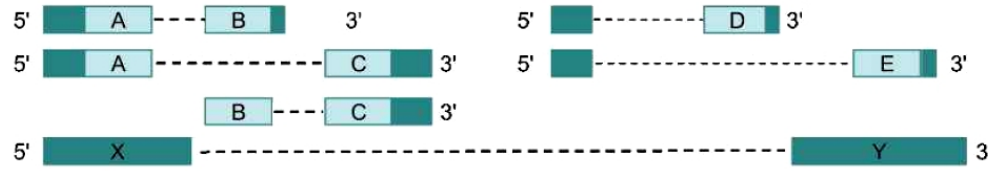
DNA genómico {



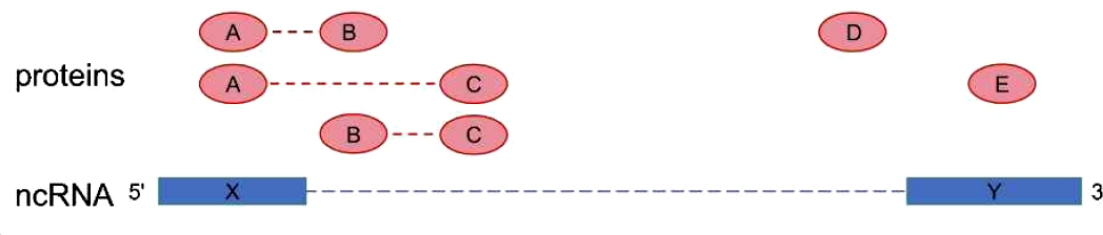
RNAs primarios {



RNAs maduros {



Productos funcionales {



Distintas bases de datos de anotación génica

Distintas iniciativas de anotación de genes

-Vega : <http://vega.sanger.ac.uk/index.html>



-RefSeq : <http://www.ncbi.nlm.nih.gov/RefSeq/>

-Hugo : <http://www.genenames.org/>

-Ensembl : <http://www.ensembl.org/index.html>

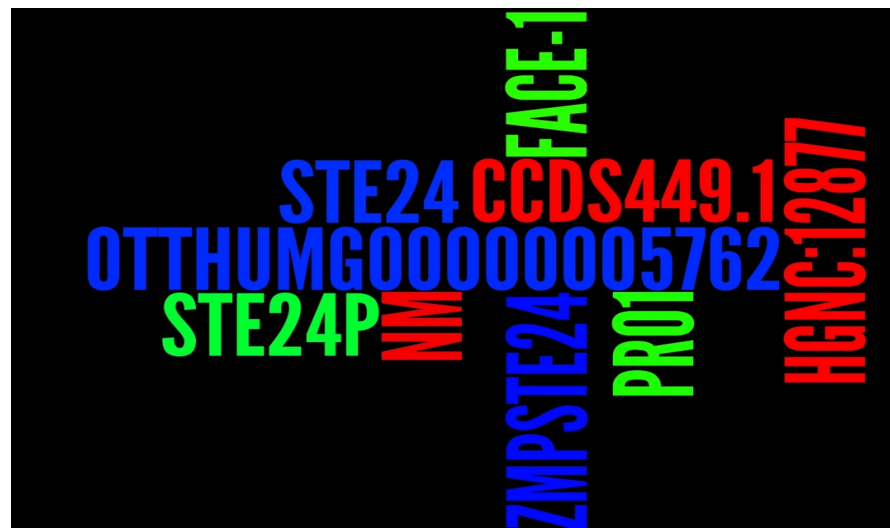


-Encode : <http://genome.ucsc.edu/ENCODE/>



-Havana : <http://www.sanger.ac.uk/research/projects/vertebrategenome/havana/>

Confusión de anotaciones



Herramientas como **Variant Effect Predictor (VEP)** o **SnEff** permiten identificar las consecuencias funcionales de las mutaciones.

Formato VCF

Encabezamiento empezando con

Metadata header

A VCF file starts with lines of metadata that begin with `##`. Some key components of this section include:

- **gdcWorkflow:** Information on the pipelines that were used by the GDC to generate the VCF file. Annotated VCF files contain two *gdcWorkflow* lines, one that reports the variant calling process and one that reports the variant annotation process.
- **INDIVIDUAL:** information about the study participant (`case`), including:
 - *NAME*: Submitter ID (barcode) associated with the participant
 - *ID*: GDC case UUID
- **SAMPLE:** sample information, including:
 - *ID*: NORMAL or TUMOR
 - *NAME*: Submitter ID (barcode) of the aliquot
 - *ALIQOT_ID*: GDC aliquot UUID
 - *BAM_ID*: The UUID for the BAM file used to produce the VCF
- **INFO:** Format of *additional information* fields
 - NOTE: GDC Annotated VCFs may contain multiple INFO lines. The last INFO line contains information about annotation fields generated by the [Somatic Annotation Workflow](#) (see [GDC INFO Fields](#) below).
- **FILTER:** Description of filters that have been applied to the variants
- **FORMAT:** Description of genotype fields
- **reference:** The reference genome used to generate the VCF file (GRCh38.d1.vd1.fa)
- **contig:** A list of IDs for the contiguous DNA sequences that appear in the reference genome used to produce VCF files
 - NOTE: Annotated VCFs include contig information for autosomes, sex chromosomes, and mitochondrial DNA. Unplaced, unlocalized, human decoy, and viral genome sequences are not included.
- **VEP:** the VEP command used by the [Somatic Annotation Workflow](#) to generate the annotated VCF file.

```
##fileformat=VCFv4.2
##FILTER=<ID=LowQual,Description="Low quality">
##FORMAT=<ID=AD,Number=R,Type=Integer,Description="Allelic depths for the ref and alt alleles in the order listed">
##FORMAT=<ID=DP,Number=1,Type=Integer,Description="Approximate read depth (reads with MQ=255 or with bad mates are filtered)">
##FORMAT=<ID=GQ,Number=1,Type=Integer,Description="Genotype Quality">
##FORMAT=<ID=GT,Number=1,Type=String,Description="Genotype">
```

Formato VCF

Encabezamiento normal empezando con

1. **CHROM:** The chromosome on which the variant is located
2. **POS:** The position of the variant on the chromosome. Refers to the first position if the variant includes more than one base
3. **ID:** A unique identifier for the variant; usually a dbSNP rs number if applicable
4. **REF:** The base(s) exhibited by the reference genome at the variant's position
5. **ALT:** The alternate allele(s), comma-separated if there are more than one
6. **QUAL:** Not populated
7. **FILTER:** The names of the filters that have flagged this variant. The types of filters used will depend on the variant caller used.
8. **INFO:** Additional information about the variant. This includes the annotation applied by the VEP.
9. **FORMAT:** The format of the sample genotype data in the next two columns. This includes descriptions of the colon-separated values.
10. **NORMAL:** Colon-separated values that describe the normal sample
11. **TUMOR:** Colon-separated values that describe the tumor sample

```
##contig=<ID=Chr17,length=81195210>
##contig=<ID=Chr18,length=78077248>
##contig=<ID=Chr19,length=59128983>
##contig=<ID=Chr20,length=63025520>
##contig=<ID=Chr21,length=48129895>
##contig=<ID=Chr22,length=51304566>
##contig=<ID=ChrX,length=155270560>
##contig=<ID=ChrY,length=59373566>
##contig=<ID=ChrMT,length=16569>
##source=HaplotypCaller
#CHROM POS ID REF ALT QUAL FILTER INFO FORMAT B15-11194.IonCode_0105 B15-12759.IonCode_0101 B15-14049.IonCode_0103 B15-14218.IonCode_0106 B15-18204.IonCode_0125 B15-19
370.IonCode_0107 B15-22959.IonCode_0104 B15-23608.IonCode_0102 B15-24439.IonCode_0126 B15-9874.IonCode_0108 B19-10031.IonCode_0127 B20-13716.IonCode_0128 B20-16469.IonCode_0132 B20-21576.Io
nCode_0131 B21-16319.IonCode_0129 B21-19492.IonCode_0130 Placa1R_25 Placa1R_26 Placa1R_27 Placa1R_28 Placa1R_29 Placa1R_30 Placa1R_31 Placa1R_32 Placa1_1
Placa1_10 Placa1_11 Placa1_12 Placa1_13 Placa1_14 Placa1_15 Placa1_16 Placa1_17 Placa1_18 Placa1_19 Placa1_2 Placa1_20
Placa1_21 Placa1_22 Placa1_23 Placa1_24 Placa1_3 Placa1_4 Placa1_5 Placa1_6 Placa1_7 Placa1_8 Placa1_9 Placa2_1
Placa2_10 Placa2_11 Placa2_12 Placa2_13 Placa2_14 Placa2_15 Placa2_16 Placa2_17 Placa2_18 Placa2_19 Placa2_2 Placa2_20
Placa2_21 Placa2_22 Placa2_23 Placa2_24 Placa2_25 Placa2_26 Placa2_27 Placa2_28 Placa2_29 Placa2_3 Placa2_30 Placa2_31
Placa2_32 Placa2_4 Placa2_5 Placa2_6 Placa2_7 Placa2_8 Placa2_9
Chr1 22404949 . G A 64.55 . GT:AD:DP:GQ:PL ./.:.....:0,0,0 ./.:.....:0,0,0 ./.:.....:0,0,0 ./.:.....:0,0,0 ./.:.....:0,0,0 ./.:.....:0,0,0
0 ./.:.....:0,0,0 ./.:.....:0,0,0 ./.:.....:0,0,0 ./.:.....:0,0,0 ./.:.....:0,0,0 ./.:.....:0,0,0 ./.:.....:0,0,0 ./.:.....:0,0,0 0/0:1,0:1:3:0,3,32 0/0:2,0:2:6:0,6,60 0/0:
1,0:1:3:0,3,32 0/0:1,0:1:3:0,3,32 ./.:.....:0,0,0 0/0:2,0:2:6:0,6,61 0/0:1,0:1:3:0,3,32 0/0:1,0:1:3:0,3,32 ./.:.....:0,0,0 0/0:1,0:1:3:0,3,25 0/0:1,0:1:3:0,3,32
./.:.....:0,0,0 0/0:2,0:2:6:0,6,49 0/0:3,0:3:9:0,9,79 0/0:1,0:1:3:0,3,27 0/0:1,0:1:3:0,3,32 ./.:.....:0,0,0 0/0:1,0:1:3:0,3,35 ./.:.....:0,0,0
0/0:2,0:2:6:0,6,61 0/0:1,0:1:3:0,3,27 0/0:1,0:1:3:0,3,30 0/0:2,0:2:6:0,6,52 0/0:2,0:2:6:0,6,53 ./.:.....:0,0,0 ./.:.....:0,0,0 ./.:.....:0,0,0 0/0:1,0:1:3:0,3,36 ./.:...
.:0,0,0 ./.:.....:0,0,0 0/0:2,0:2:6:0,6,58 ./.:.....:0,0,0 0/0:1,0:1:3:0,3,24 ./.:.....:0,0,0 0/0:1,0:1:3:0,3,32 ./.:.....:0,0,0 0/0:1,0:1:3:0,3,32 0/0:1,0:1:3:0,3,32 0/0:
1,0:1:3:0,3,32 0/0:1,0:1:3:0,3,32 0/0:4,0:4:12:0,12,119 0/0:4,0:5:15:0,15,156 0/0:1,0:1:3:0,3,31 0/0:2,0:2:6:0,6,59 0/0:1,0:1:3:0,3,36 0/0:2,0:2:6:0,6,61 0/0:1,0:1:
3:0,3,27 0/0:1,0:1:3:0,3,32 0/0:2,0:2:6:0,6,63 0/0:4,0:4:12:0,12,116 0/0:1,0:1:3:0,3,32 0/0:3,0:3:9:0,9,94 1/1:0:3:3:9:82,9,0 0/0:6,0:6:18:0,18,158 0/0:2,0:2:6:0,6
```

El formato VCF tiene una versión binaria comprimida que se denomina **BCF**

Bioinformática y análisis de datos ómicos

