Alineamiento de datos NGS

RICARDO A. VERDUGO, Ph.D.

Programa de Genética Humana, ICBM
Facultad de Medicina, U. de Chile

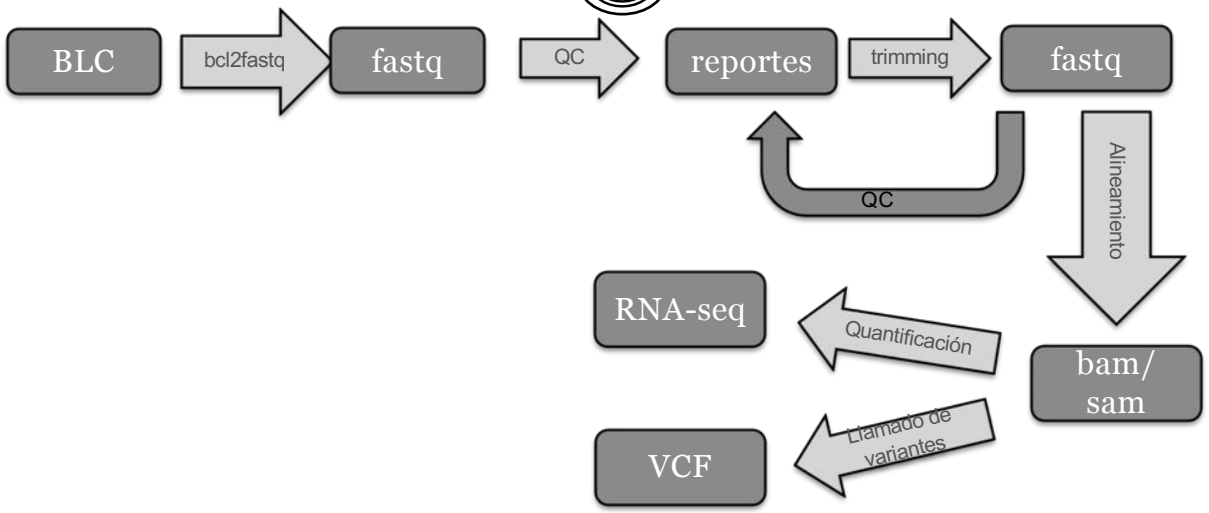
Abril 2019

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

Temas a cubrir

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

III. Standard NGS Analysis Pipeline

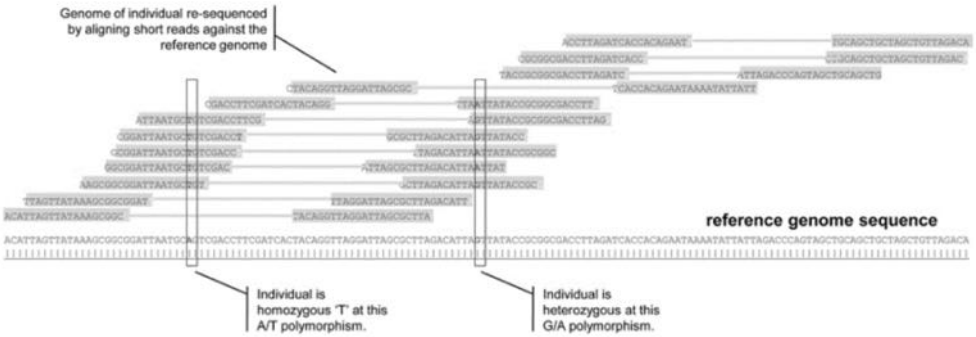


Raw sequencing data: Fastq format

```
@IL31_4368:1:1:996:3507/2
TCCCTTACCCCCAAGCTCCATACCCTCCTAATGCCCACACCTCTTACCTTAGGA
+
FFCEFFFEEFFFFFFFFEFFEFFFFEFCFC<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
+
EEEEDEEE9EAEEDEEEEEEEEEECEEAAEEDEE<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

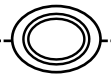
Alineamiento de lecturas de secuencias



Preguntas clave

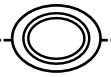
1. ¿Qué queremos alinear?
2. ¿Cómo valoraremos un buen alineamiento?
3. ¿Cómo encontramos el mejor alineamiento?

¿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

El espacio de posibles alineamientos



- Estos son algunos possible alineamineto de las palabras ELV y VIS

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

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

¿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

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(V,A) + s(A,I) + s(H,Q) + s(V,L) + 3g + s(D,G) + 2g ...

¿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: $O(n^2)$ $O(n^2) = O(n^4)$

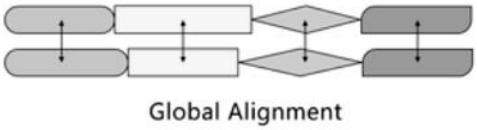
Posibles alineamientos

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

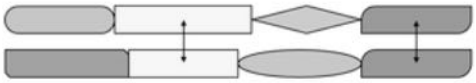
Solución por programación dinámica

- **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

Alineamiento Global vs Local



Global Alignment



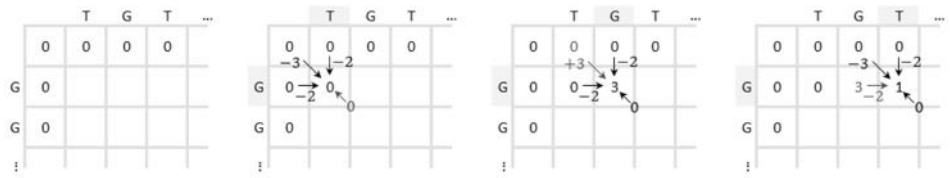
Local Alignment

Algoritmo de Smith y Waterman

Ejemplo: TGTTACGG vs GGTTGACTA

Matriz de substitución: $s(a_i, b_j) = \begin{cases} +3, & a_i = b_j \\ -3, & a_i \neq b_j \end{cases}$

Penalización por espacios: $W_k = kW_1, W_1 = 2$



https://en.wikipedia.org/wiki/Smith-Waterman_algorithm

Algoritmo de Smith y Waterman

$$H_{ij} = \max \begin{cases} H_{i-1,j-1} + s(a_i, b_j), \\ \max_{k \geq 1}\{H_{i-k,j} - W_k\}, \\ \max_{l \geq 1}\{H_{i,j-l} - W_l\}, \\ 0 \end{cases}$$

$$(1 \leq i \leq n, 1 \leq j \leq m)$$

Initialize the scoring matrix

		T	G	T	T	A	C	G	G
	0	0	0	0	0	0	0	0	0
G	0								
G	0								
T	0								
T	0								
G	0								
A	0								
C	0								
T	0								
A	0								

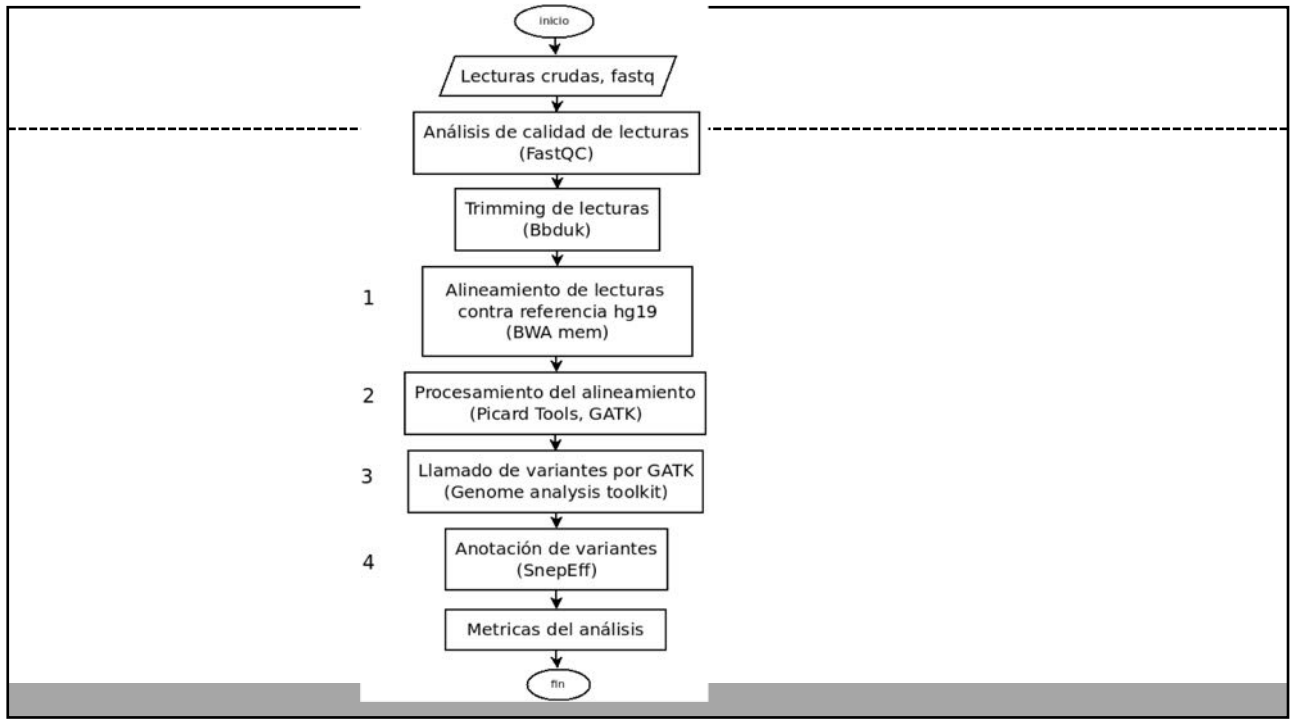
Substitution matrix: $S(a_i, b_j) = \begin{cases} +3, & a_i = b_j \\ -3, & a_i \neq b_j \end{cases}$

Gap penalty: $W_k = kW_1$, $W_1 = 2$

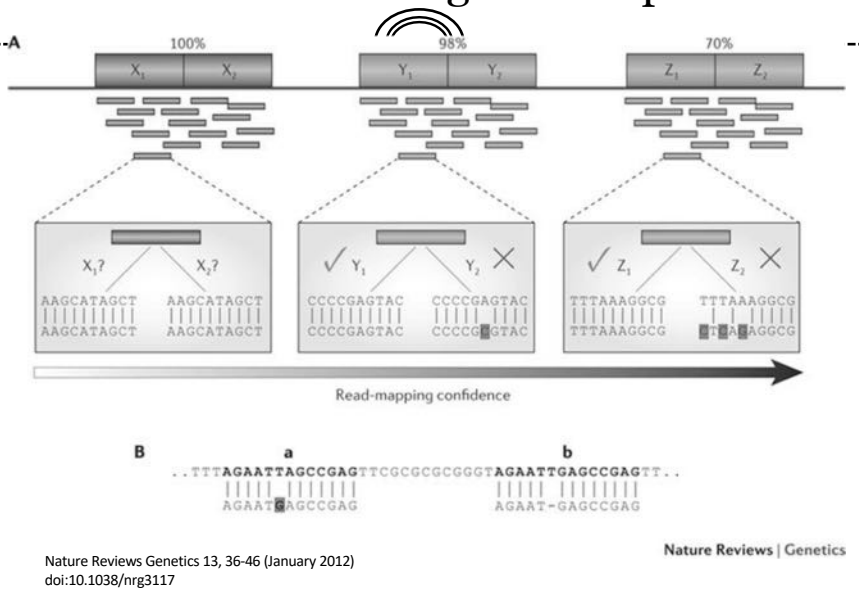
Alineador BWA

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

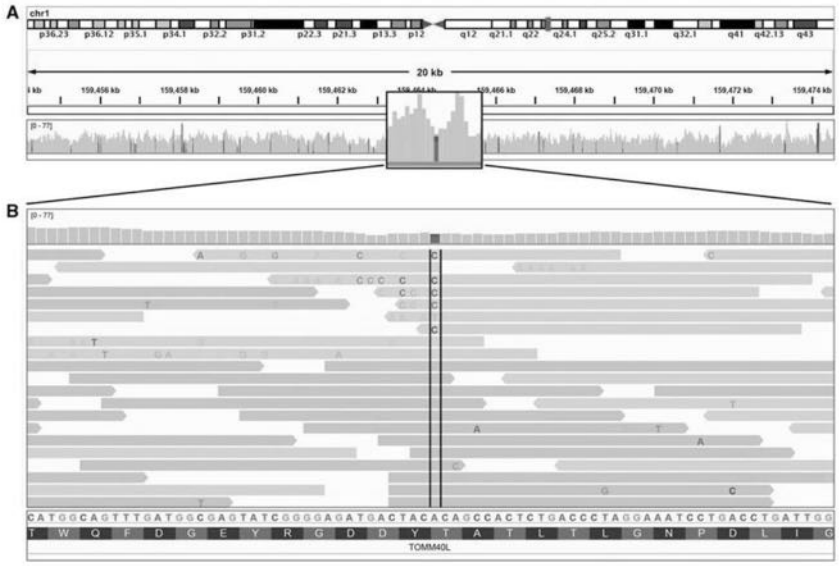


Alineamiento en Regiones Repetidas



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

Integrative Genomic Viewer



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

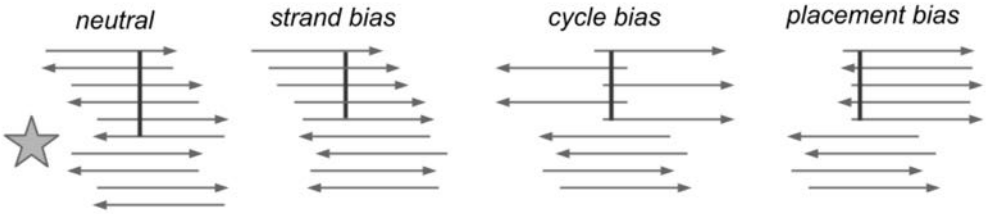
¿Qué mirar en una visualización de alineamiento?

- Cobertura en la reunión 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 realineamineto local)

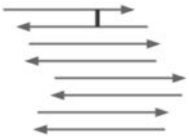
Métricas básicas de calidad

- % 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)*

Sesgos de variantes detectadas



allele imbalance



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