UNU BIOLAC

Cursos Internacional .
Secuenciación y Análisis de Datos Genómicos para la Detección Microbiológica de Enfermedades Transmitidas por Alimentos y Aguas

RIF: G-20000101-1
INSTITUTO NACIONAL DE HIGIENE RAFAEL RANGEL

Secuenciación de ARN

Fernando Hernandez, IVIC

"RNA-Seq (named as an abbreviation of RNA sequencing) is a sequencing technique that uses next-generation sequencing (NGS) to reveal the presence and quantity of RNA in a biological sample, representing an **aggregated snapshot of the cells' dynamic pool of RNAs**, also known as transcriptome"

Wikipedia

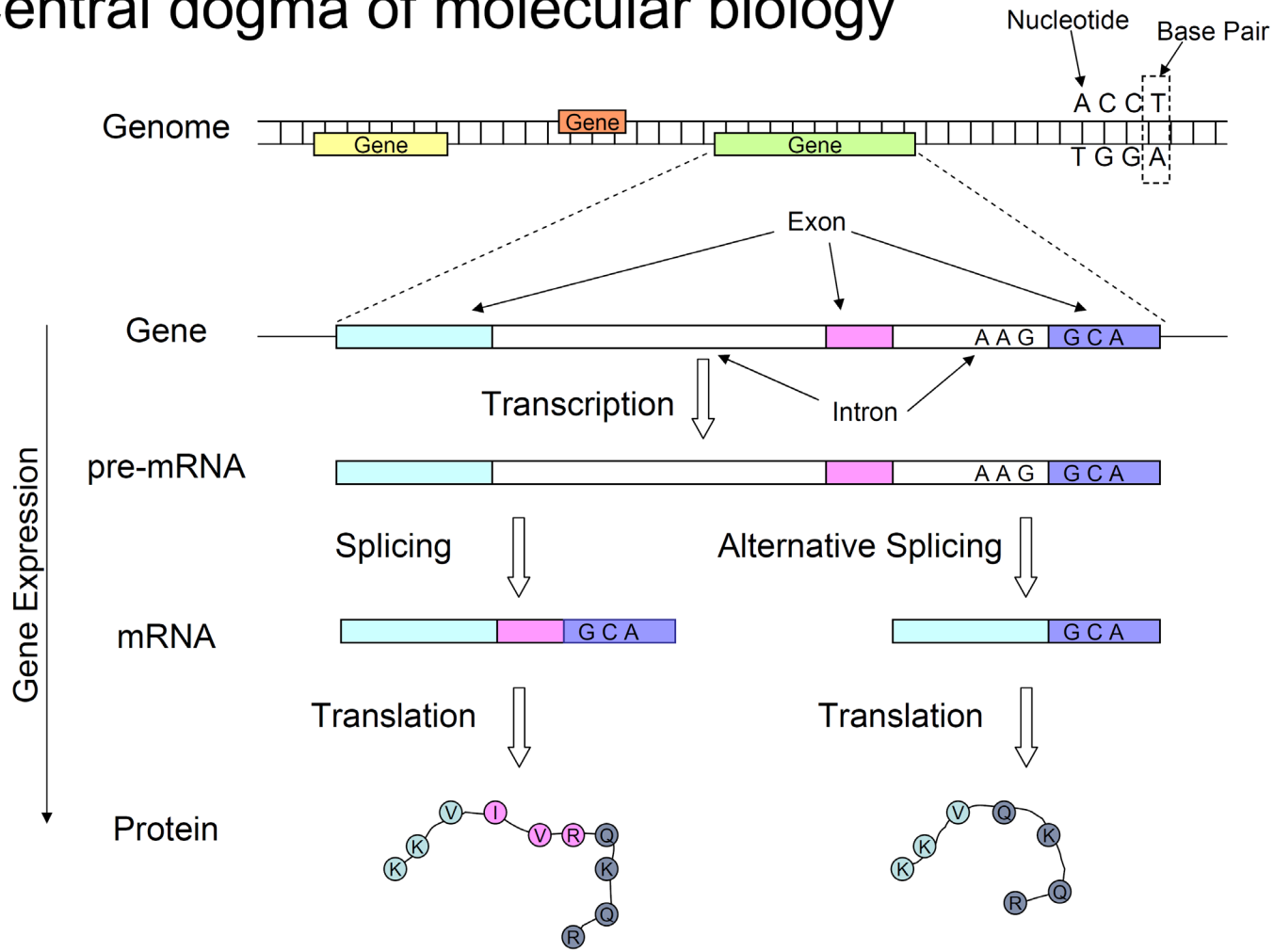
RNA-sequencing

- Ensamblaje de transcriptoma
- Expresión génica
- Análisis diferencial de la expresión

RNA-Seq (frente a microarreglos)

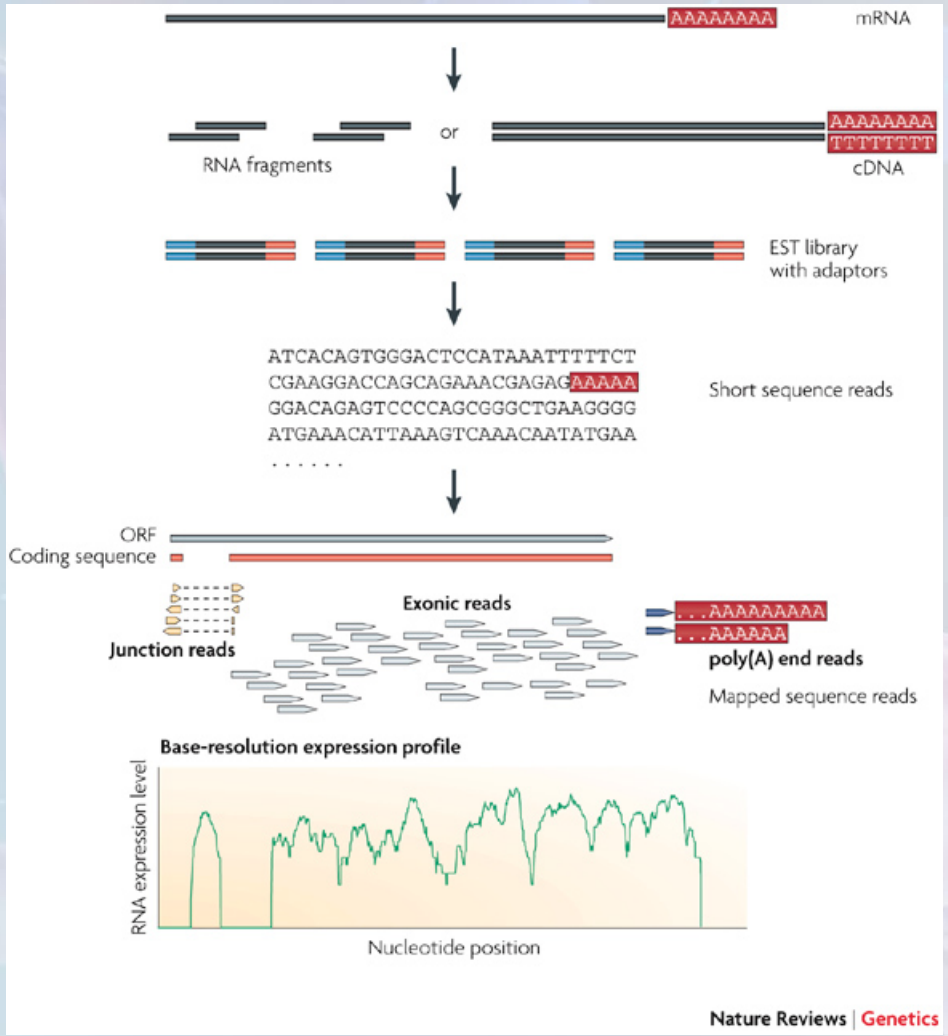
- Fuerte concordancia entre plataformas
- Mayor sensibilidad y rango dinámico
- Menor variación técnica
- Disponible para todas las especies
- Nuevas regiones transcritas
- Empalme alternativo
- Expresión específica de alelo
- Genes de fusión
- Mayor coste informático

Central dogma of molecular biology



92–94% of human genes undergo alternative splicing, 86% with a minor isoform frequency of 15% or more

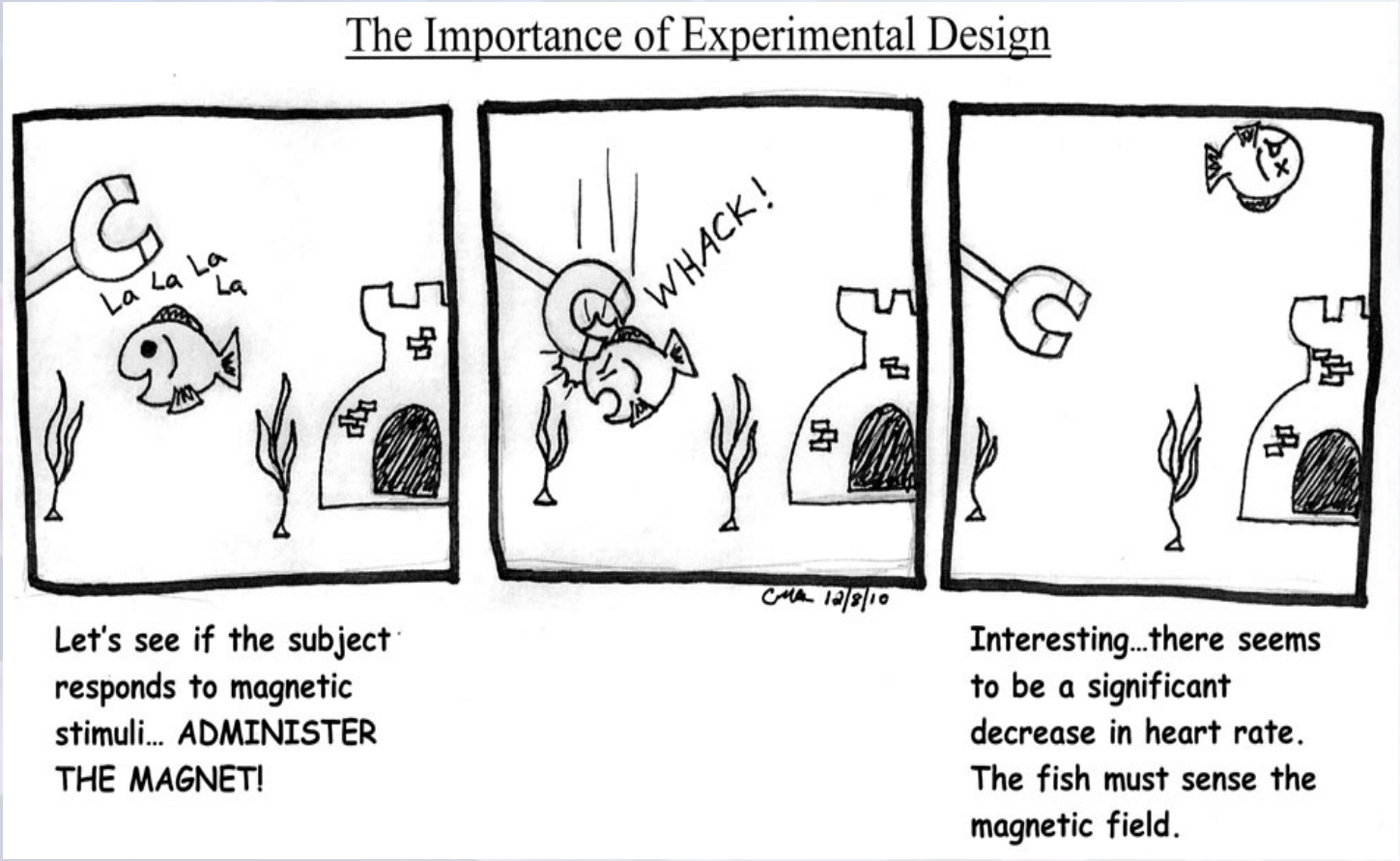
E.T. Wang, et al, Nature 456, 470-476 (2008)



1. Preparación de las librerías
2. Secuenciación
3. Análisis bioinformático

Wang, Z., et al. (2009), Nature Reviews Genetics, 10, 57–63

Diseñando el experimento ideal

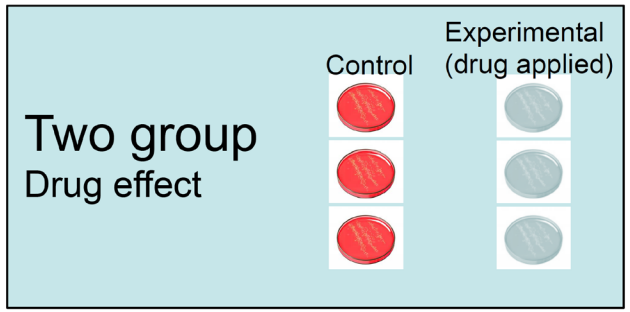


Diseñando el experimento ideal

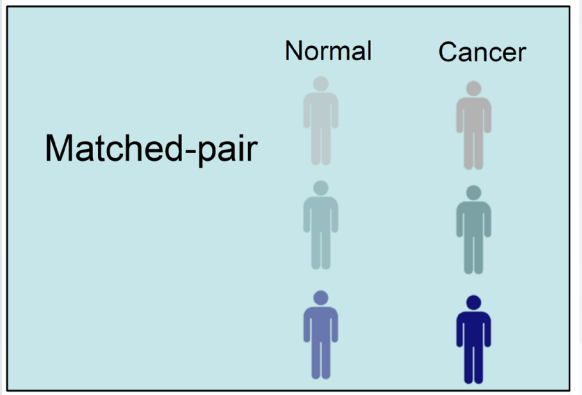
El diseño del experimento es el primer paso y obviamente es determinante para todos los análisis posteriores

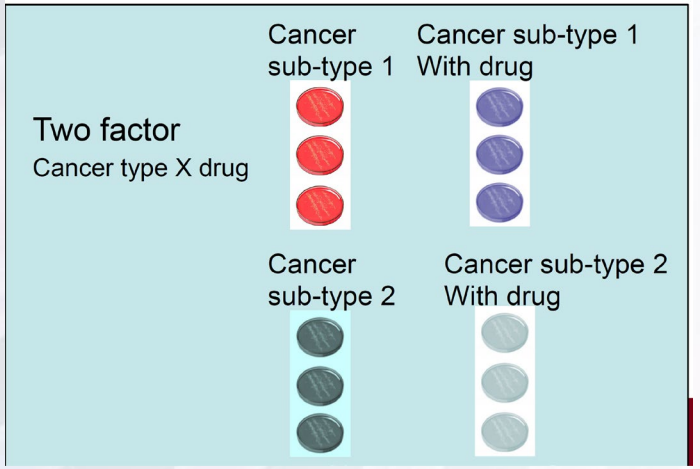
Tienes que evaluar todas las eventualidades y limitaciones de las tecnologías disponibles, diseñando el experimento de acuerdo con tus objetivos

Diseño simple



Diseño complejo





Diseñando el experimento ideal

¿Cuáles son mis objetivos?

– ¿**Ensamblaje** del transcriptoma?
– ¿Análisis de expresión diferencial?
– ¿Identificar **transcritos** raros?

¿Cuáles son las características de mi sistema?

– ¿Genoma **grande** y **complejo**?
– ¿Intrones y alto grado de empalme alternativo?
– ¿No hay **genoma** ni **transcriptoma** de referencia?

Diseñando el experimento ideal

Coverage: Cuantos reads son necesarios?

The coverage is defined as $C = (R_{length} \times R_{num}) / A_{length}$

R_{length} = length in nucleotides of the reads
R_{num} = number of sequenced reads
A_{length} = number of nucleotides of sequenced subject (genome, transcriptome, exome)

Diseñando el experimento ideal

For 12-15 samples prepared for bulk RNA Seq:

Library prep for bulk seq:
polyA selection
costs $250/samples. (so $3000 for 12 samples)
using the Illumina mRNA kit

Our sequencing runs are performed on the BU core's Illumina Nextseq as
75 bp paired-end reads in high output mode.
Cost is $3050 for a run (of up to 15 samples combined) and typically results in 400 million reads.

This typically results in the follow read depths:
33 million (M) reads per sample for 12 samples; or
22M reads/sample for 18 samples; or
26M reads/sample for 15 samples

3050$/15 muestras

CReM CENTER FOR REGENERATIVE MEDICINE
BOSTON UNIVERSITY
BOSTON MEDICAL

6600$/4 muestras

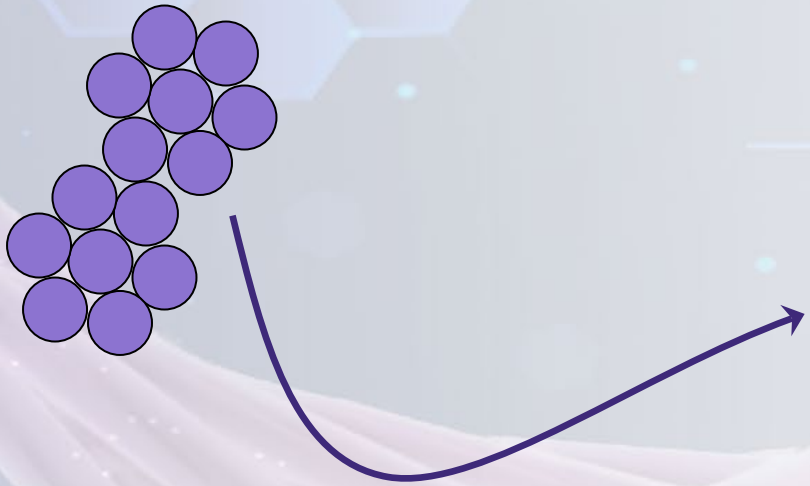
scRNA seq using the 10x Genomics system

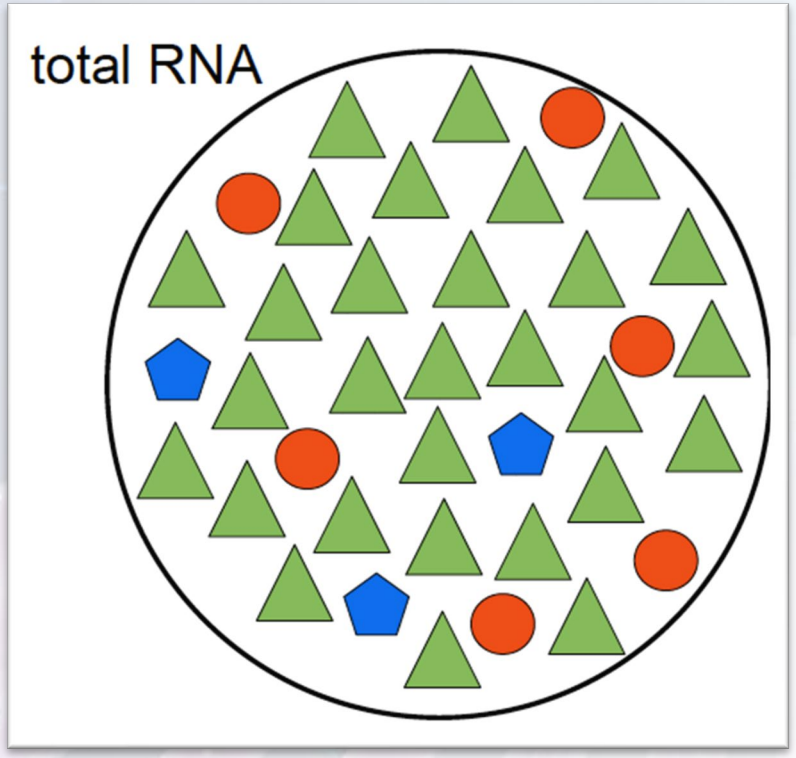
Costs based on our BU core http://www.bumc.bu.edu/singlecell/pricing/

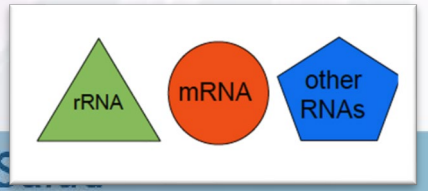
$6600 for 4 samples (can handle 4 at a time on the 10X machine)
we typically plan to capture 1000 cells per sample with library prep through the BU core.
We request all 4 samples run together in a single Illumina Nextseq run at a cost of $3050; requested as follows through the core:
Nextseq "high output" sequencing mode with "150 cycles" as paired-end reads:
(The paired end reads are at special settings for the 10X Genomics system: 26 bp for one read which covers the barcode and UMI, and the rest for the other paired end).
Prices are here: http://www.bumc.bu.edu/microarray/pricing/

This format typically yields more than 50,000 reads per cell, which is our target read depth).

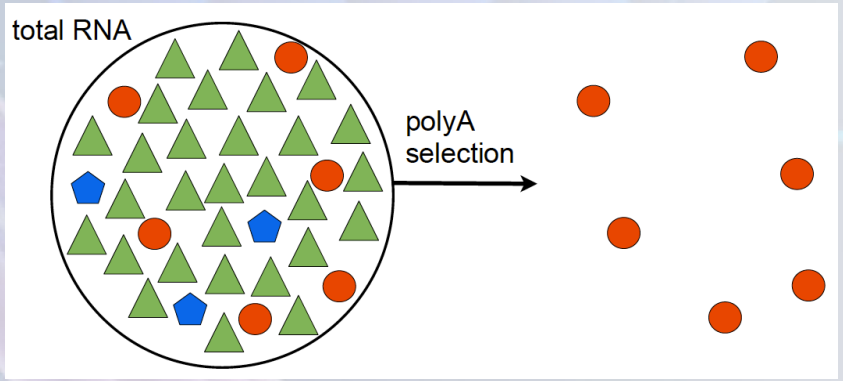
Preparación de las librerías





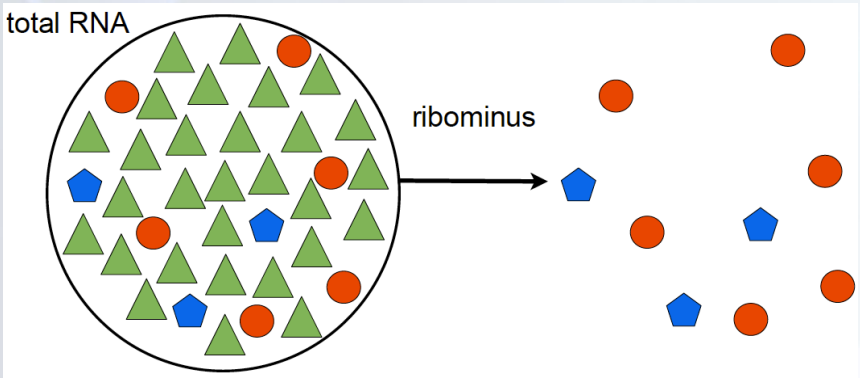


• Poli-A

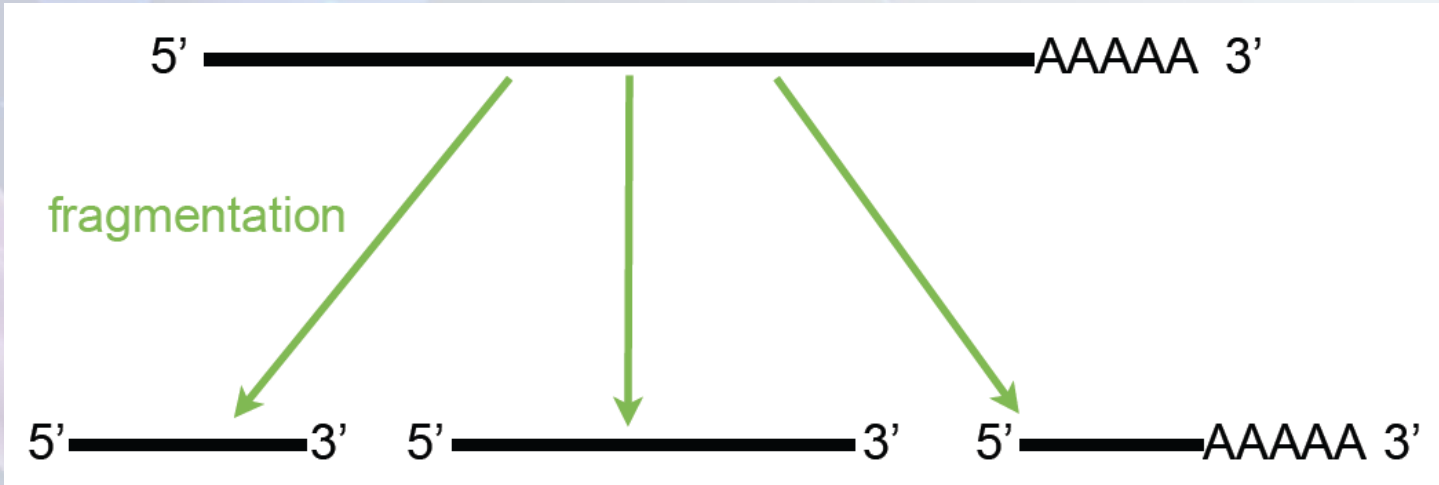


- Transcritos poli-A:
  - mRNAs
  - microRNAs inmaduros
  - snoRNAs

• Ribominus



- Transcritos no poli-A:
  - mRNAs
  - mRNAs de histonas
  - tRNAs
  - Otros RNAs pequeños

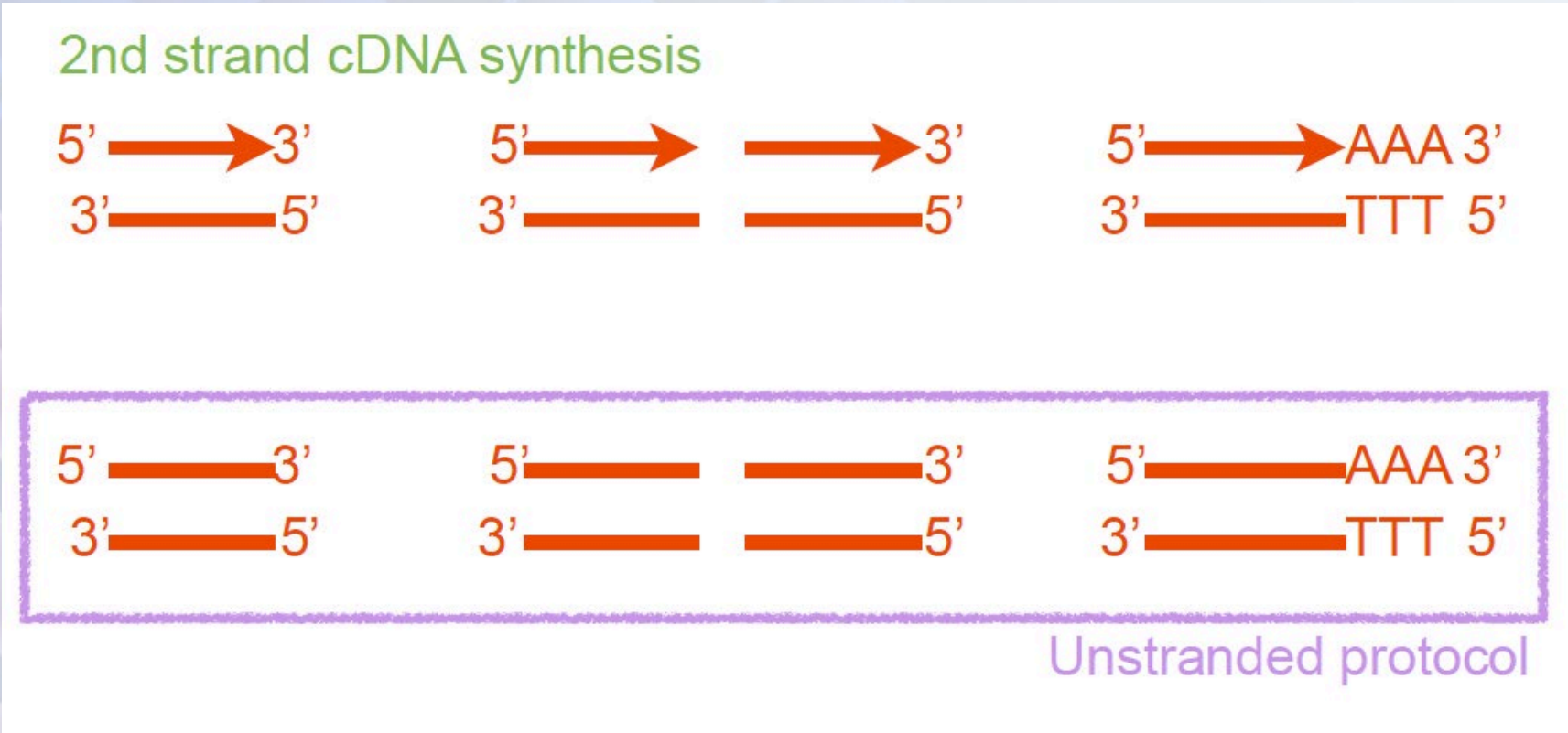

5’
AAAAA 3’
fragmentation
5’
3’
5’
3’
5’
AAAAA 3’

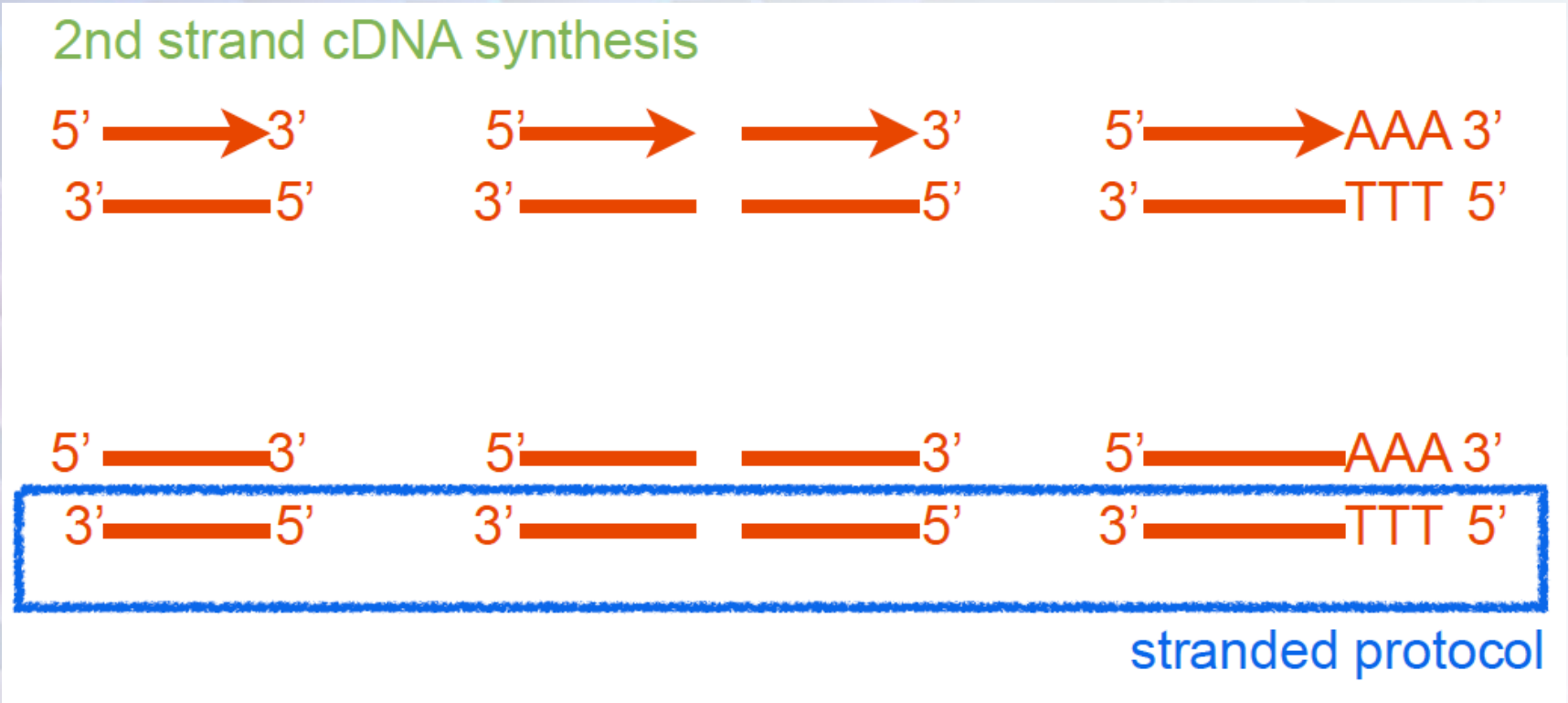
Transcriptome analysis by strand-specific sequencing of complementary DNA

Dmitri Parkhomchuk, Tatiana Borodina, Vyacheslav Amstislavskiy, Maria Banaru, Linda Hallen, Sylvia Krobitsch, Hans Lehrach and Alexey Soldatov*

1 strand cDNA synthesis

remove RNA strand


2nd strand cDNA synthesis
5’ 3’
3’ 5’
5’ 3’
3’ 5’
5’ AAA 3’
3’ TTT 5’
5’ 3’
3’ 5’
5’ 3’
3’ 5’
5’ AAA 3’
3’ TTT 5’
Unstranded protocol


2nd strand cDNA synthesis
5’ 3’
3’ 5’
5’ 3’
3’ 5’
5’ AAA 3’
3’ TTT 5’
5’ 3’
3’ 5’
5’ 3’
3’ 5’
5’ AAA 3’
3’ TTT 5’
stranded protocol

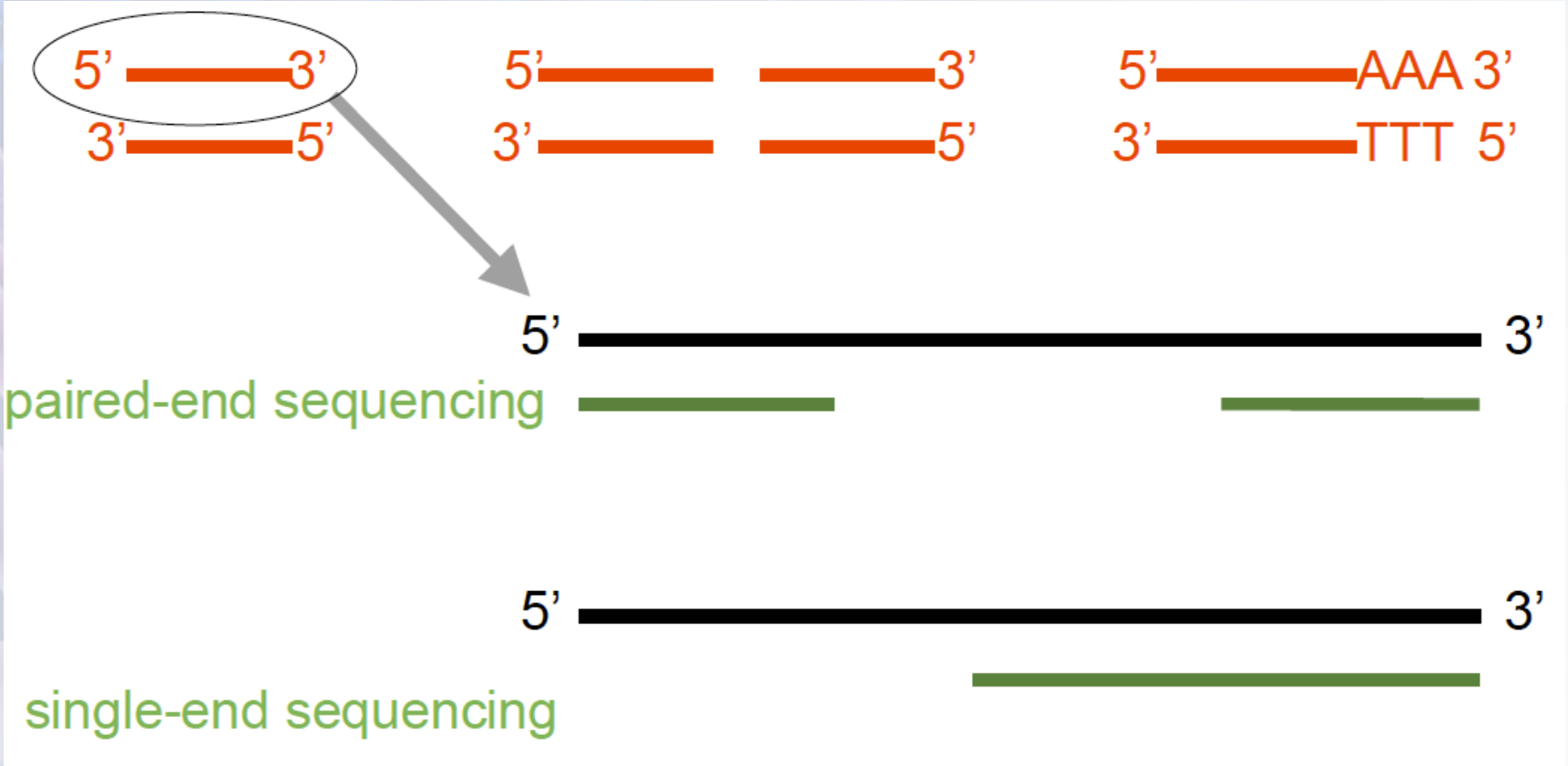
cDNA fragmentado

cDNA con adaptadores

Selección del tamaño

Amplificación (PCR)

Secuenciación



my_sequence.fastq

@HWI-BRUNOP16X_0001:1:1:1466:1018#0/1
AAGGAAGTGCTTGTCTGGCTAACACAGCNAGNCACGT
GAC
+
aVfbe`^^^_TTTSSdfffdfffabbZbbfebafbbbbb

my_sequence_1.fastq

@HWI-BRUNOP16X_0001:1:1:1278:989#0/1
NAAATTTCGAATTTCTGTGAAGTAAGCATCTTCTTTGTCAT
+
BJJGGKIINN^^^^^QQNTUQOOTTTRTOTY^^Y^\\^^^\

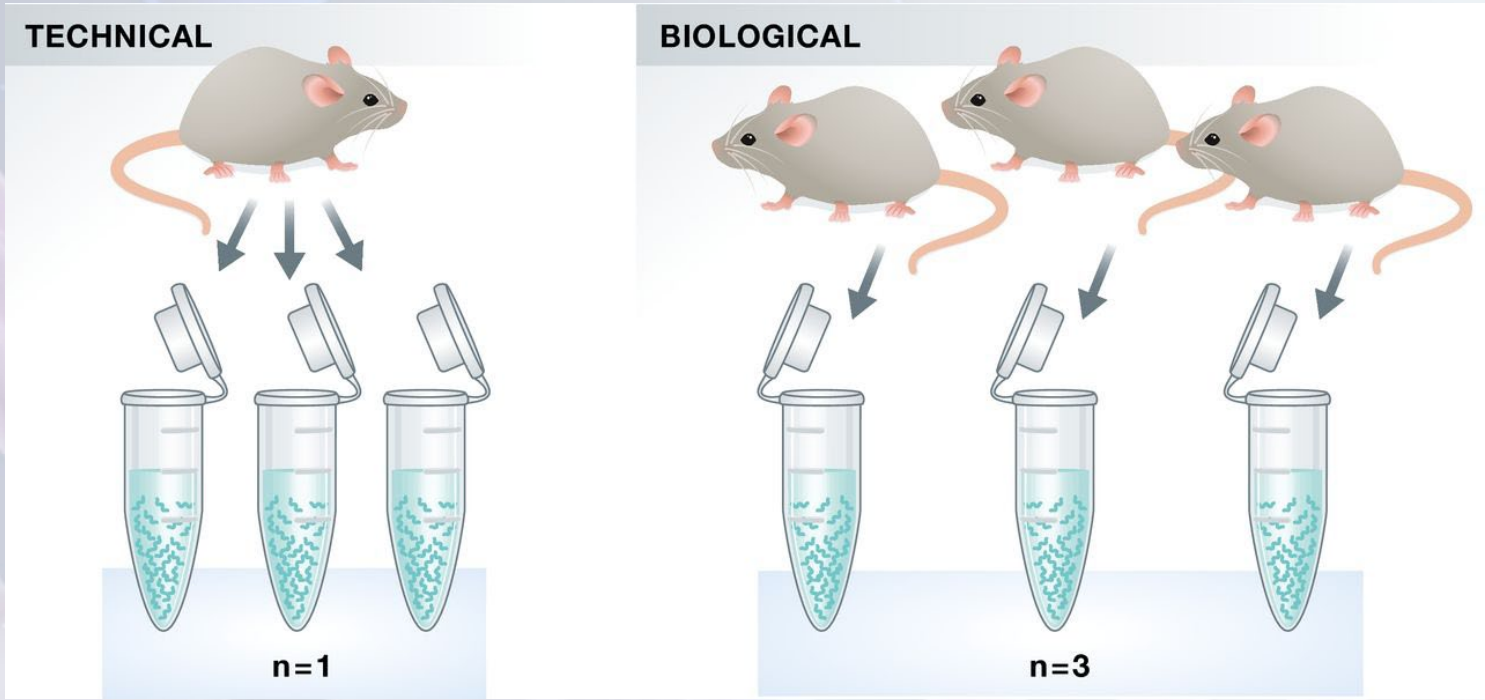
my_sequence_2.fastq

@HWI-BRUNOP16X_0001:1:1:1278:989#0/2
AACCCACACAGGAGAGCAGCCTTACAGATGCAAATACTGTG
+
]K___fffffggghgegggggdgggggfgggggeggggghh

SE

PE

Réplicas: ¿las necesito?



Técnica

Misma muestra biológica – mismas condiciones

Preparación y secuenciación → Replica técnica 1

Preparación y secuenciación → Replica técnica 2

Mide variación técnica

Biológica

Preparación y secuenciación

Replica biológica 1

Preparación y secuenciación

Replica biológica 2

Diferente muestra biológica – mismas condiciones

Mide variación biológica

- **Réplicas técnicas**

No son necesarias: baja variación técnica

- Minimizar los efectos por lotes
- Aleatorizar el orden de las muestras

- **Réplicas biológicas**

No son necesarias para el ensamblaje del transcriptoma.
Esenciales para el análisis diferencial de la expresión.
Difícil de estimar

- 3+ para líneas celulares
- 5+ para líneas endogámicas
- 20+ para muestras humanas

Controlar los batch effects

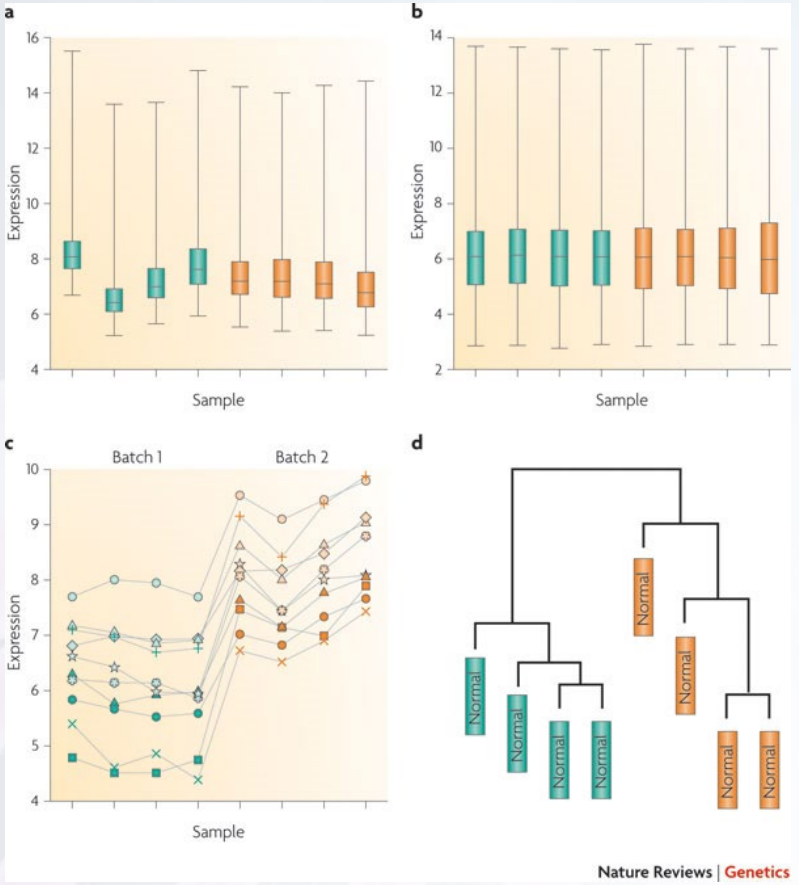
Published: 14 September 2010

Tackling the widespread and critical impact of batch effects in high-throughput data

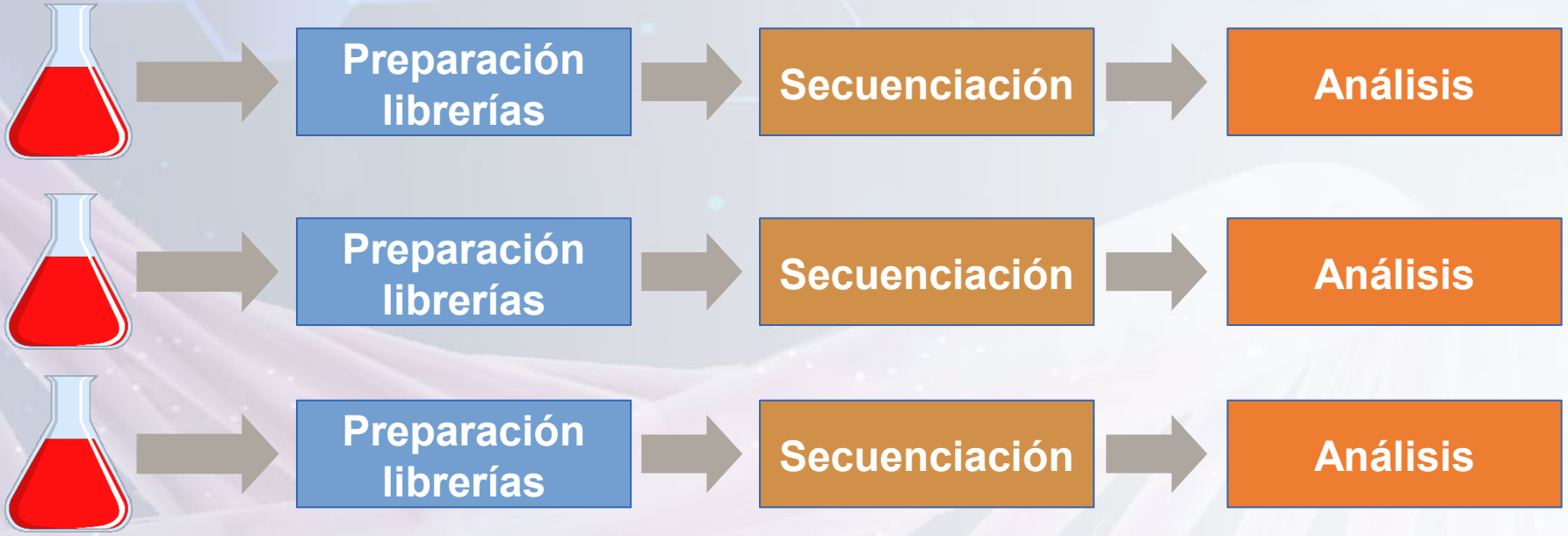
Jeffrey T. Leek, Robert B. Scharpf, Héctor Corrada Bravo, David Simcha, Benjamin Langmead, W. Evan Johnson, Donald Geman, Keith Baggerly & Rafael A. Irizarry

Nature Reviews Genetics **11**, 733–739 (2010) | Cite this article

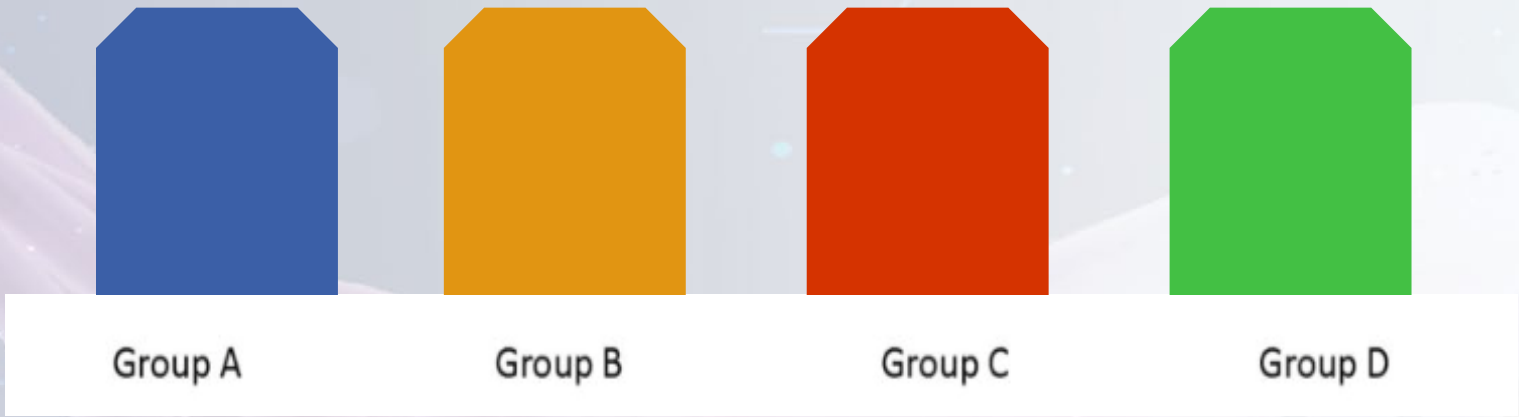
Los **batch effect** son subgrupos de mediciones que tienen un comportamiento cualitativamente diferente en todas las condiciones y no están relacionados con las variables biológicas o científicas de un estudio



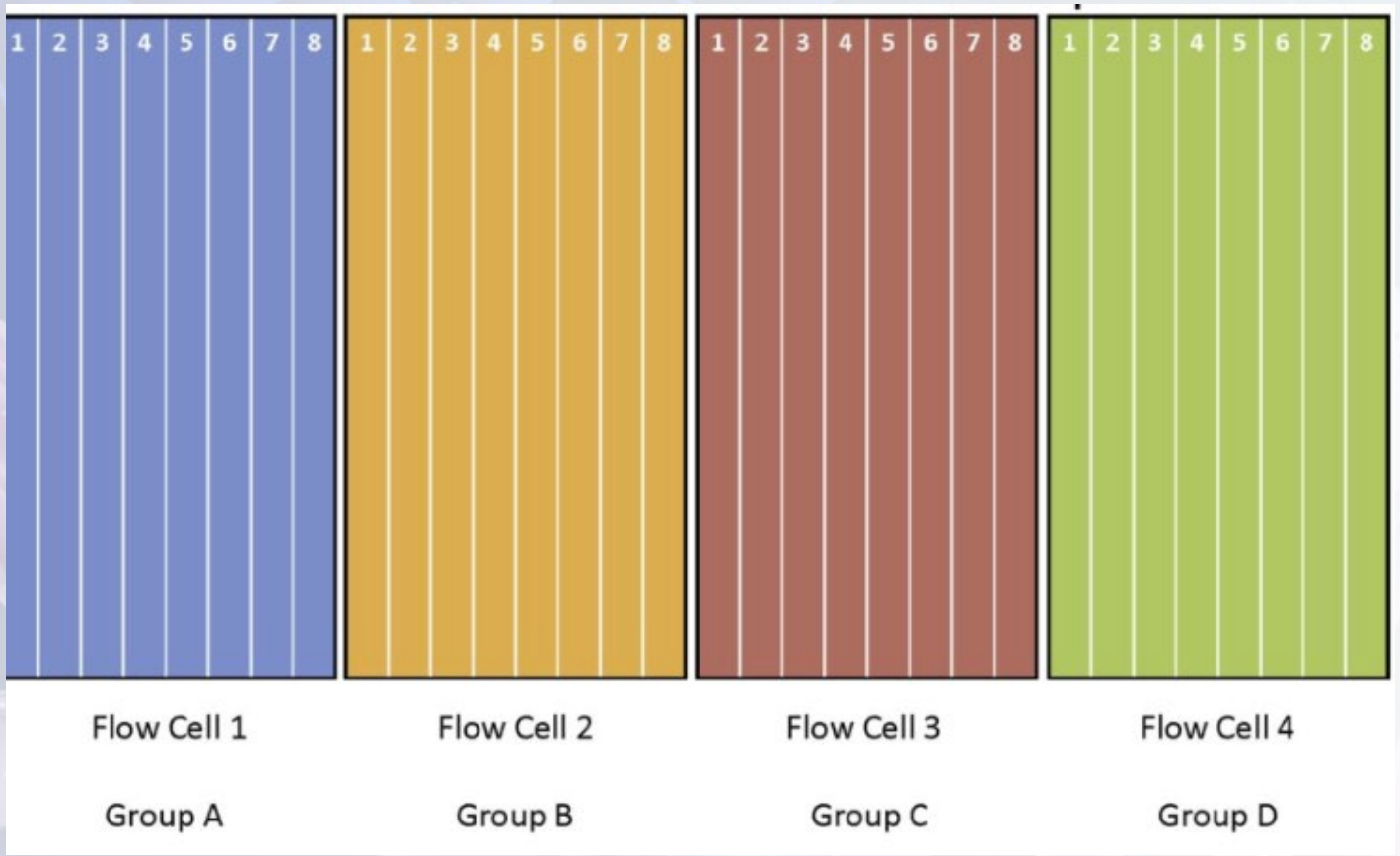
Controlar los batch effects



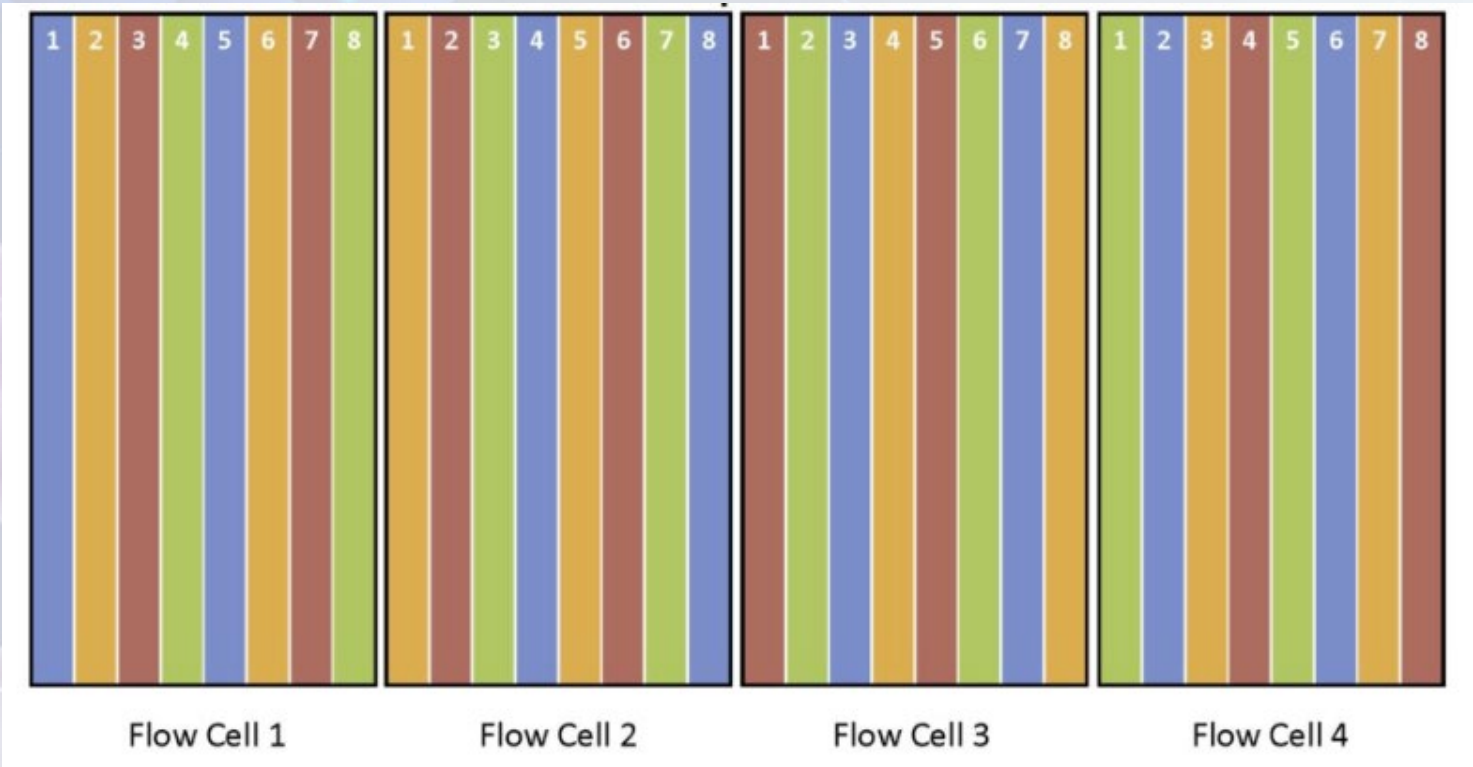
Ejemplo de diseño experimental



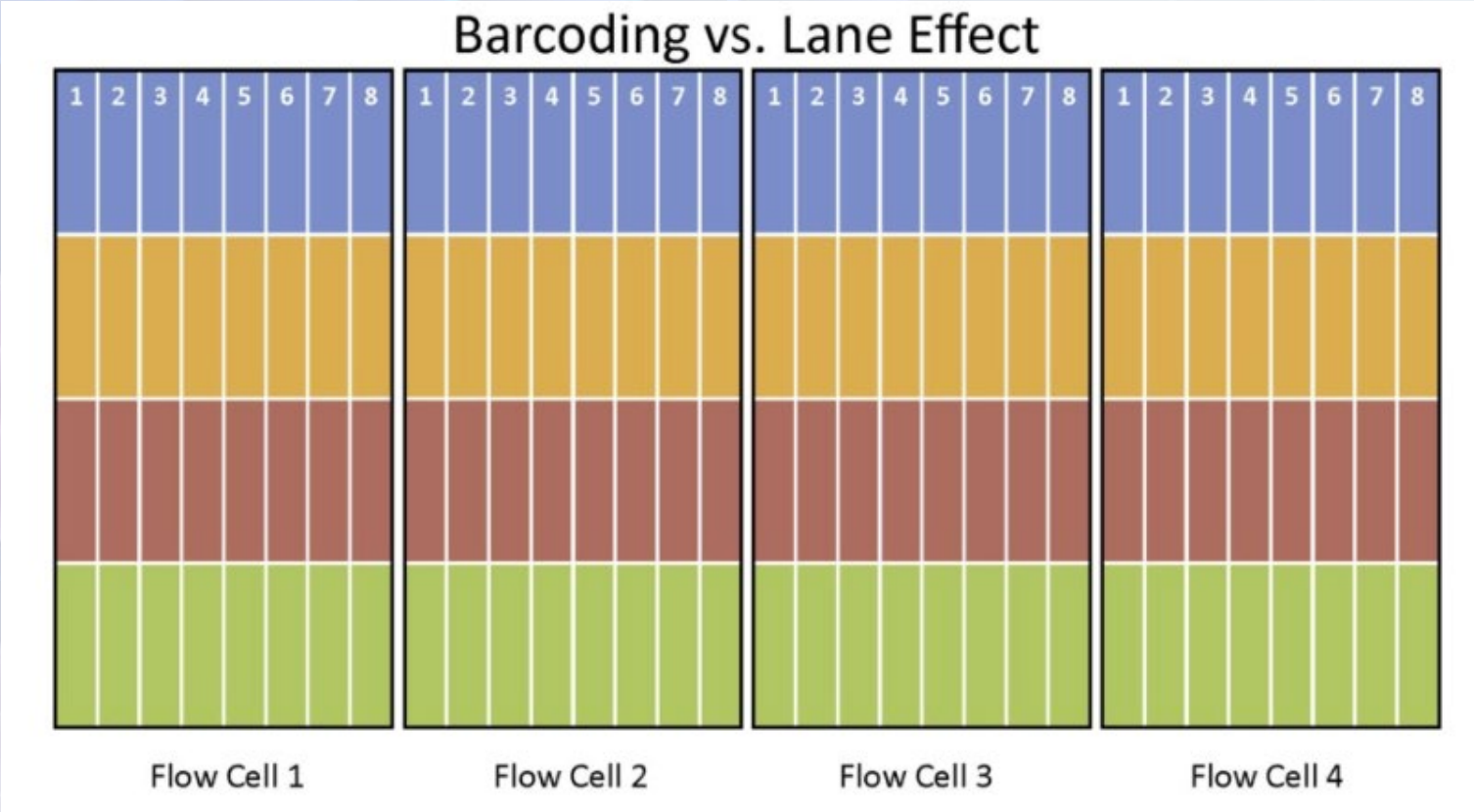
Ejemplo de diseño experimental



Ejemplo de diseño experimental

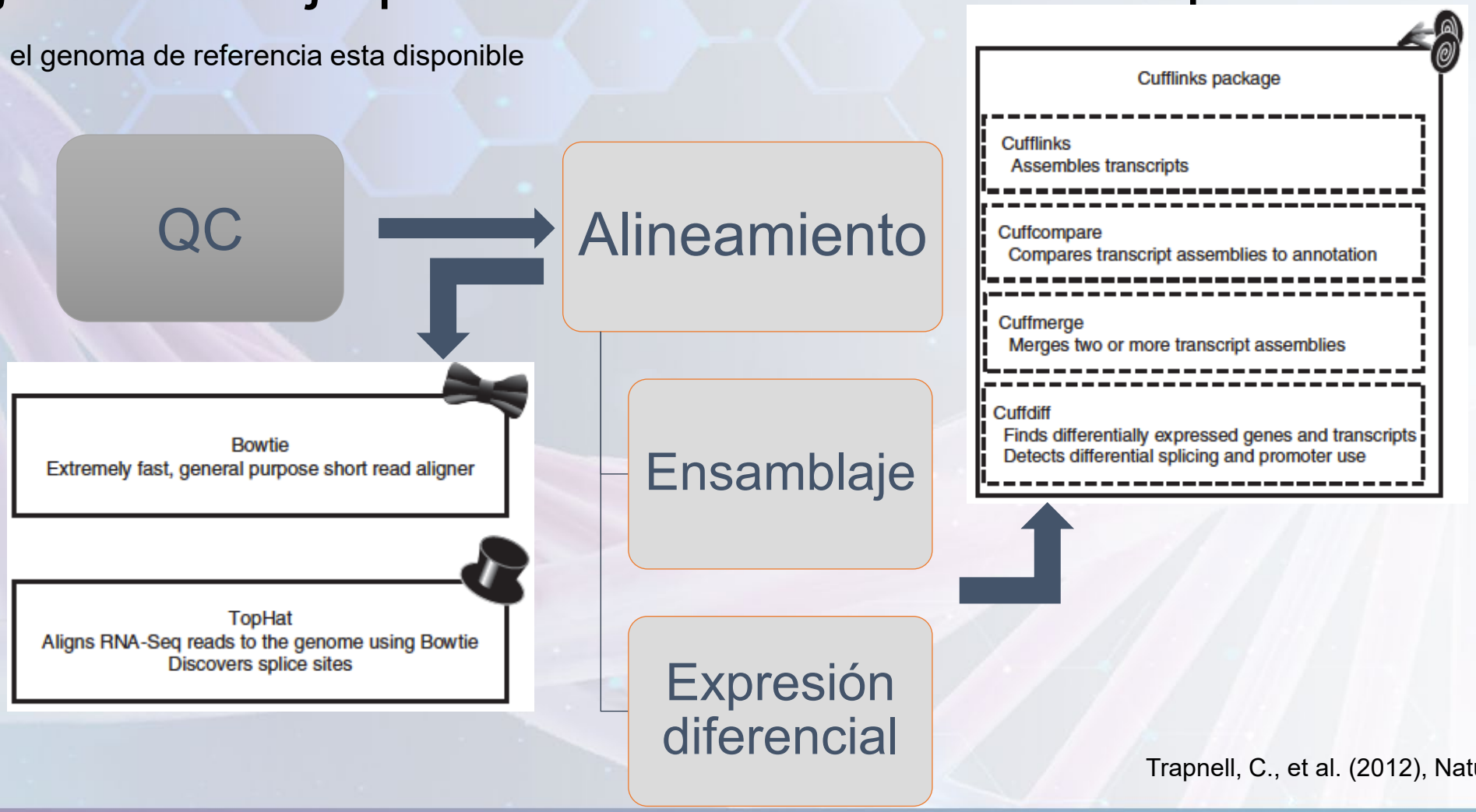


Ejemplo de diseño experimental

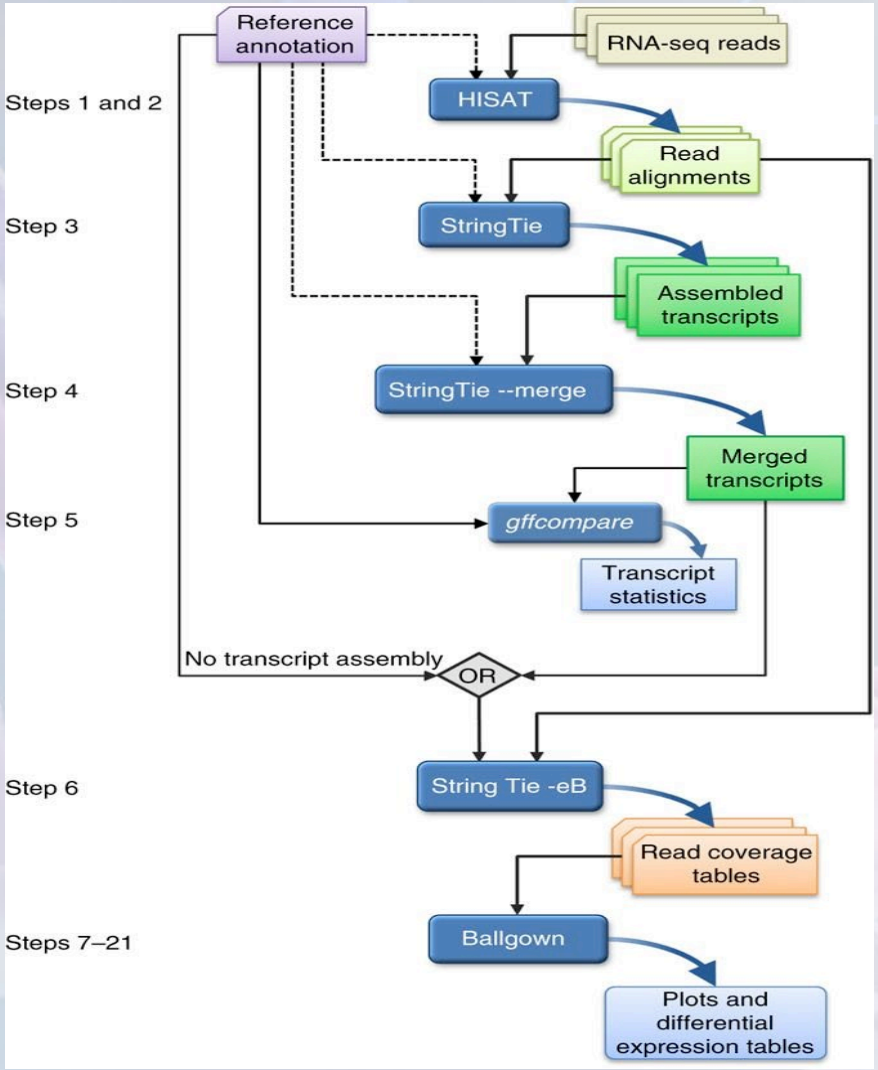


Flujo de trabajo para el análisis de RNA-seq*

* Si el genoma de referencia esta disponible



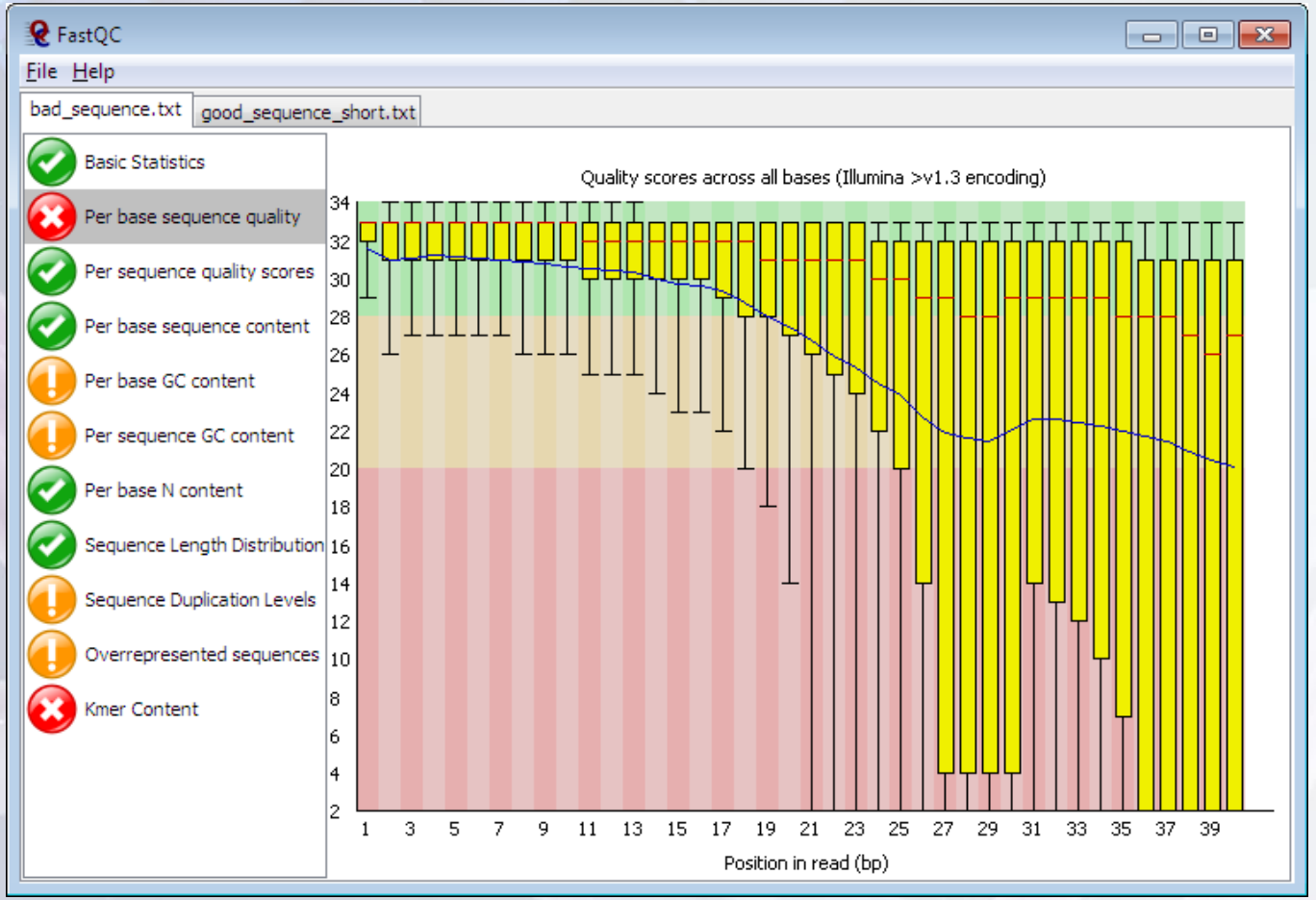
Trapnell, C., et al. (2012), Nature protocols, 7(3), 562–578.



Actualizado!!!!

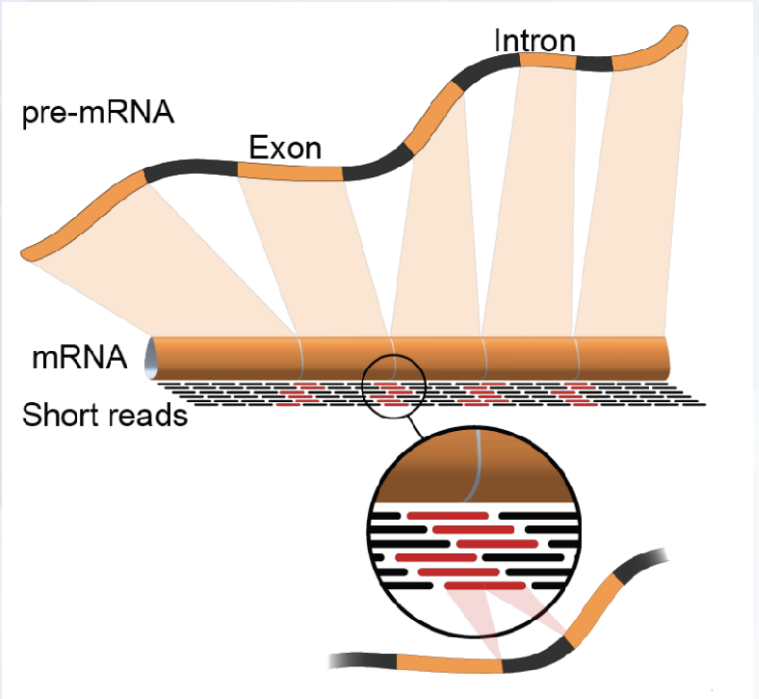
Pertea, M., et al. (2016), Nature protocols

FastQC


FastQC
File Help
bad_sequence.txt
good_sequence_short.txt
Basic Statistics
Per base sequence quality
Per sequence quality scores
Per base sequence content
Per base GC content
Per sequence GC content
Per base N content
Sequence Length Distribution
Sequence Duplication Levels
Overrepresented sequences
Kmer Content
Quality scores across all bases (Illumina >v1.3 encoding)
34
32
30
28
26
24
22
20
18
16
14
12
10
8
6
4
2
1 3 5 7 9 11 13 15 17 19 21 23 25 27 29 31 33 35 37 39
Position in read (bp)

Alineamiento

Class	Category	Package
Read mapping		
Unspliced aligners[a]	Seed methods	Short-read mapping package (SHRiMP)[41]
		Stampy[39]
	Burrows-Wheeler transform methods	Bowtie[43]
		BWA[44]
Spliced aligners	Exon-first methods	MapSplice[52]
		SpliceMap[50]
		TopHat[51]
	Seed-extend methods	GSNAP[53]
		OPAL MA[54]



Garber, M., et al. (2011), Nature Methods, 8(6), 469–477.

Los alineamientos se reportan como SAM

La mayoría de los alineadores utilizan su propio formato para generar los alineamientos.

Por lo tanto, es difícil realizar comparaciones y análisis posteriores.

Para resolver este problema, Li et al. han sugerido un formato de archivo estandarizado: el formato **Sequence Alignment Map** (SAM)

SAM es actualmente el formato estándar para los resultados de alineación.

SAMtools es un conjunto de programas para interactuar con datos de secuenciación de alto rendimiento. (**http://www.htslib.org/**)

Formato SAM

Un archivo SAM consta de dos partes:

- Encabezado
  - Contiene metadatos (fuente de las lecturas, genoma de referencia, alineador, etc.)
  - Las líneas de encabezado necesariamente comienzan con “@”.
  - Los campos de encabezado tienen códigos estandarizados de dos letras para facilitar el análisis

- Sección de alineación
  - Una tabla separada por tabulaciones con al menos 11 columnas
  - Cada línea describe un alineamiento

@HD VN:1.4 SO:coordinate
@SQ SN:CHROMOSOME_I LN:15072423
@SQ SN:CHROMOSOME_II LN:15279345
@SQ SN:CHROMOSOME_III LN:13783700
@SQ SN:CHROMOSOME_IV LN:17493793
@SQ SN:CHROMOSOME_V LN:20924149
@SQ SN:CHROMOSOME_X LN:17718866
@SQ SN:CHROMOSOME_MtDNA LN:13794
@SQ SN:sensor_piRNA_mjls144 LN:1663
@CO user command line: /Users/berkyurekahmetcan/Desktop/Data_analysis/STAR-master/bin/MacOSX_x86_64/STAR --runThreadN 4 --genomeDir /Users/berkyurekahmetcan/Desktop/Data_analysis/pichip/totalRNA/data/reference/ --readFilesIn "/Users/berkyurekahmetcan/Dropbox (ericmiskalab)/cambridge-UK/NGS/piChIP/totalRNA/outfilterMultimapper_500/elution/wt1/elution_wt_1.fastq.gz" --readFilesCommand "gunzip -c" --outSAMtype BAM SortedByCoordinate --outMultimapperOrder Random --outFilterMultimapNmax 500 --alignIntronMax 1
L180:540:HTMV2BCXY:1:1112:6120:12935 0 CHROMOSOME_I 3745 255 49M * 0 0 T AGAGGGTTAGACCCAAAATTCAGCCCGCGAAGGCATGACGTCAGCGCG GGGGGGIIIIIGIIIIIIIIIIGIIIIIIIIIIIII IIIIIIIIIIII NH:i:1 HI:i:1 AS:i:44 nM:i:2
L180:540:HTMV2BCXY:1:1113:17520:75431 0 CHROMOSOME_I 3745 255 49M * 0 0 T AGAGGGTTAGACCCAAAATTCAGCCCGCGAAGGCATGACGTCAGCGCG GGGGGIIIIIGIIIIIIIIIIGIIIIIIIIGIIII IIIIIIIGIIII NH:i:1 HI:i:1 AS:i:44 nM:i:2
L180:540:HTMV2BCXY:1:2111:6429:77948 0 CHROMOSOME_I 3745 255 49M * 0 0 T AGAGGGTTAGACCCAAAATTCAGCCCGCGAAGGCATGACGTCAGCGCG GGGGGIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIII IIIIII NH:i:1 HI:i:1 AS:i:44 nM:i:2
L180:540:HTMV2BCXY:1:2202:18496:19290 0 CHROMOSOME_I 3745 255 49M * 0 0 T AGAGGGTTAGACCCAAAATTCAGCCCGCGAAGGCATGACGTCAGCGCG GGGGGIIIIIIIIIIIIIIIIGIIIIIIIGIIII

SAMtools

SAMtools es un conjunto de herramientas sencillas útiles para:

- convertir entre SAM y BAM
- SAM: un archivo de texto legible por humanos
- BAM: una versión binaria de un archivo SAM, adecuada para un procesamiento rápido

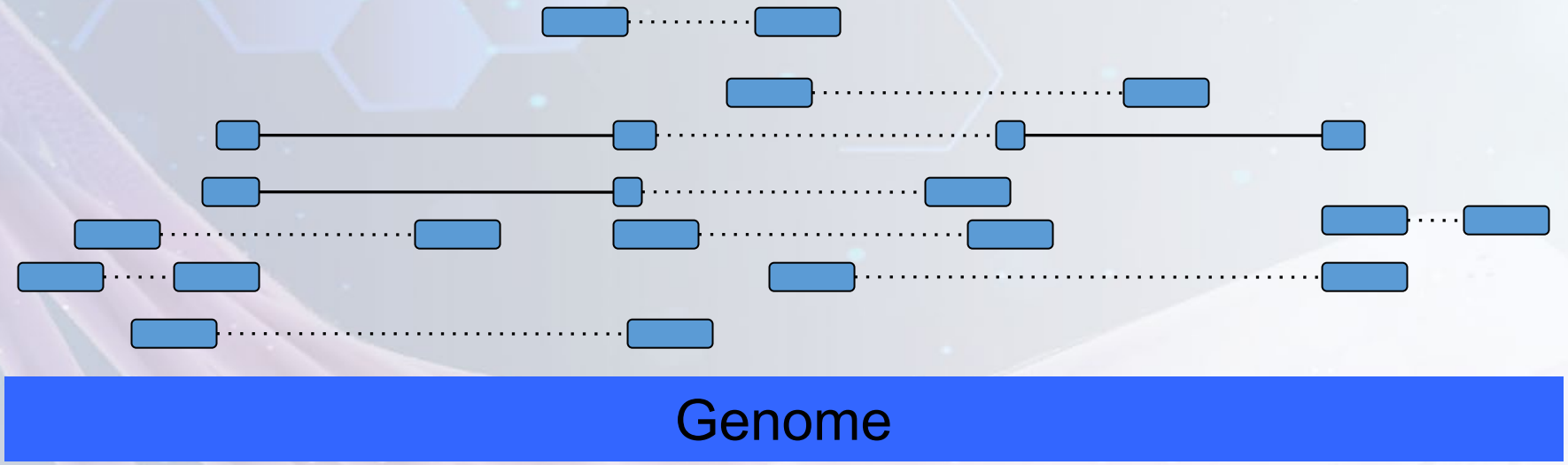
- ordenar y fusionar archivos SAM

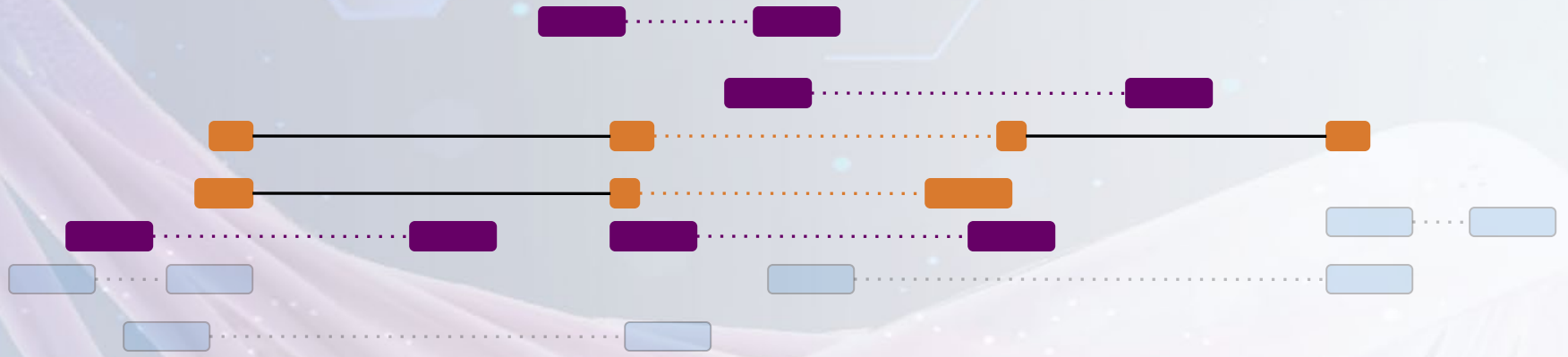
- Indexar archivos SAM y FASTA para un acceso rápido

- ver alineaciones (“tview”)

- producir un “acumulado”, es decir, un archivo que muestra:
- cobertura local
- desajustes y llamadas de consenso
- indels

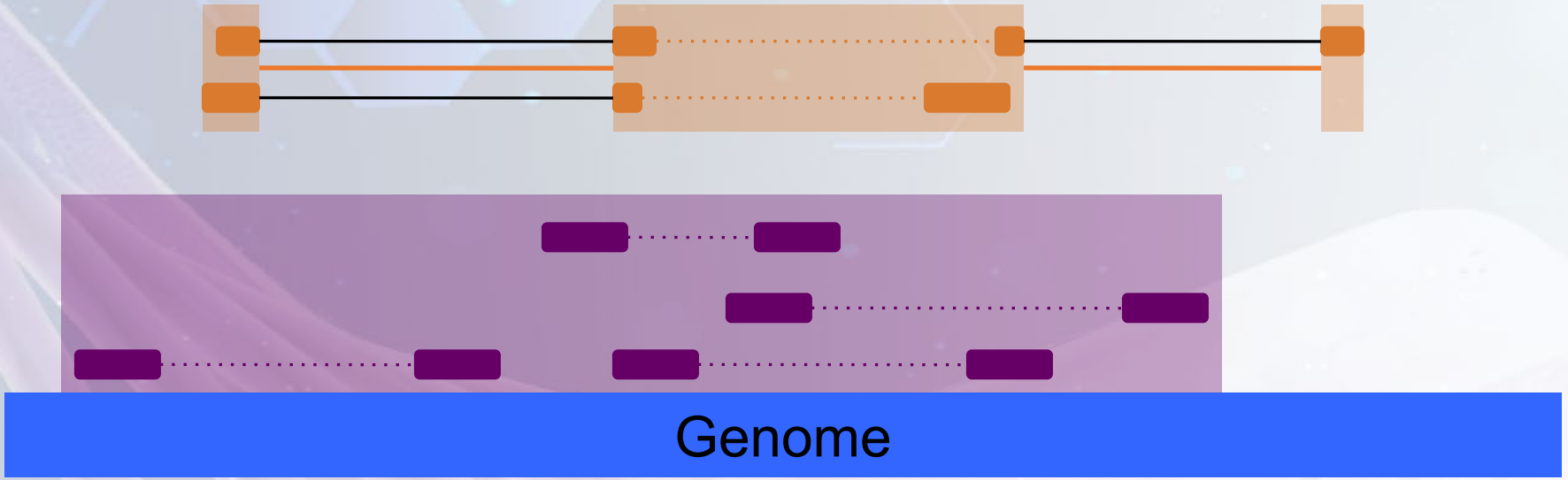
STEP 1: Identify fragments that cannot have originated from the same transcript





..... paired reads
—— spliced reads

STEP 2: Connect ‘incompatible’ fragments into directed graphs

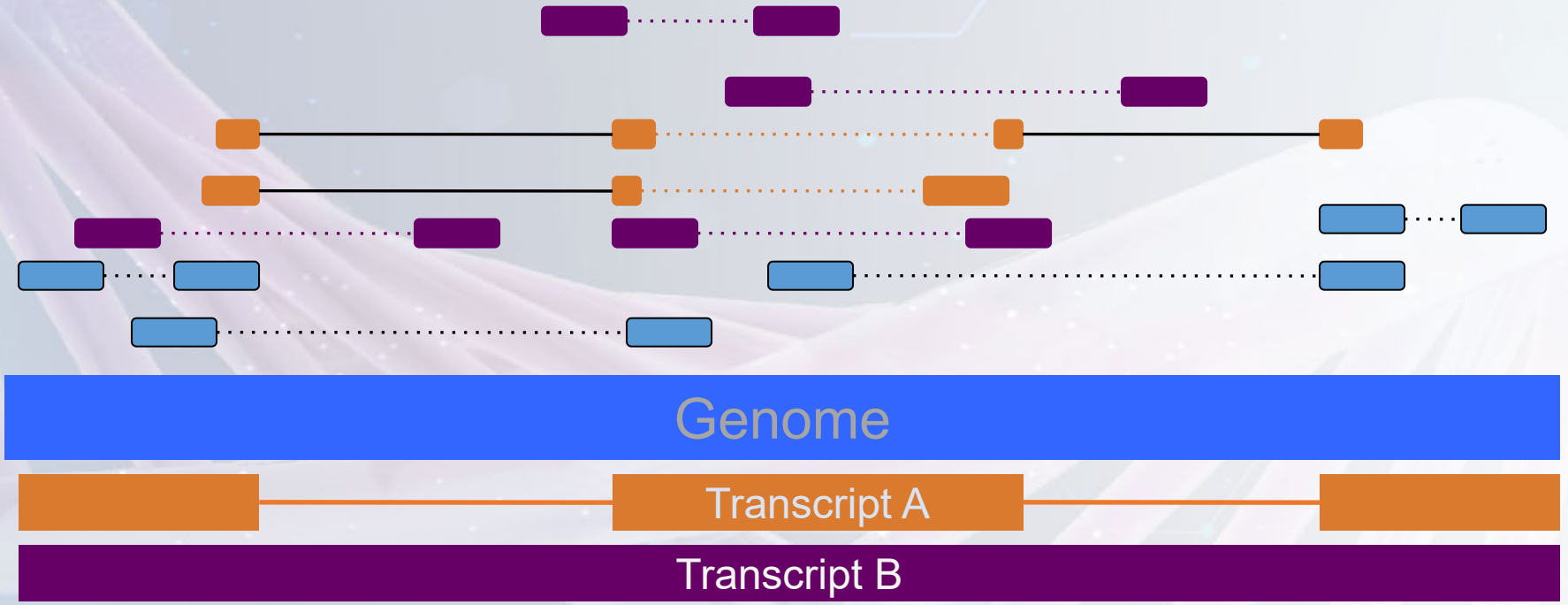


..... paired reads

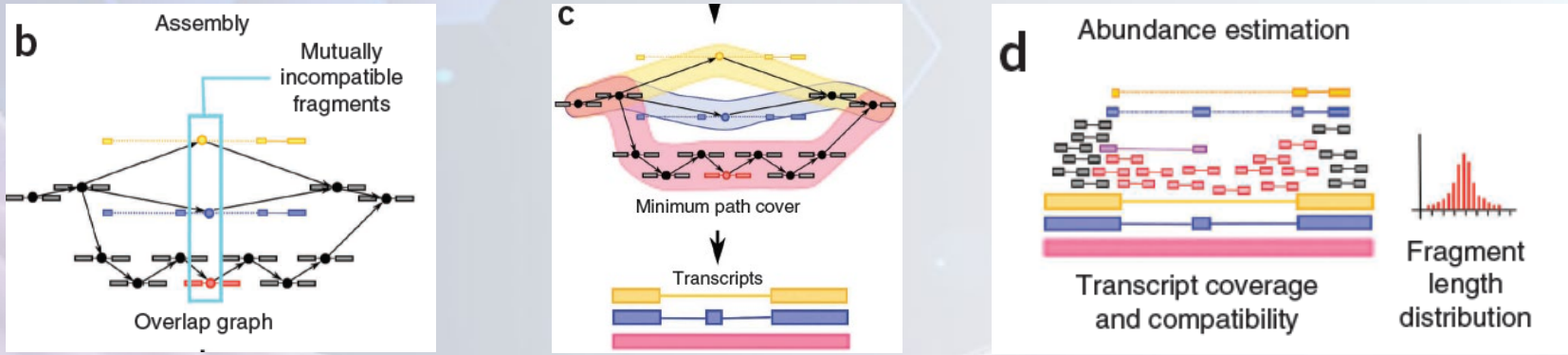
—— spliced reads

STEP 3: Assemble transcripts

STEP 4: Quantify transcript expression



Cufflinks



Transcript abundance is estimated in FPKMs (Fragments Per Kilobase of exon per Million fragments mapped)

Trapnell, C., et al. (2010), Nature biotechnology, 28(5), 511–515

• *De novo* transcriptome assembly

- **Requirements**:
  - Deep sequencing and/or longer reads
  - Thorough quality control
  - Large memory/Multiple processors
  - Patience

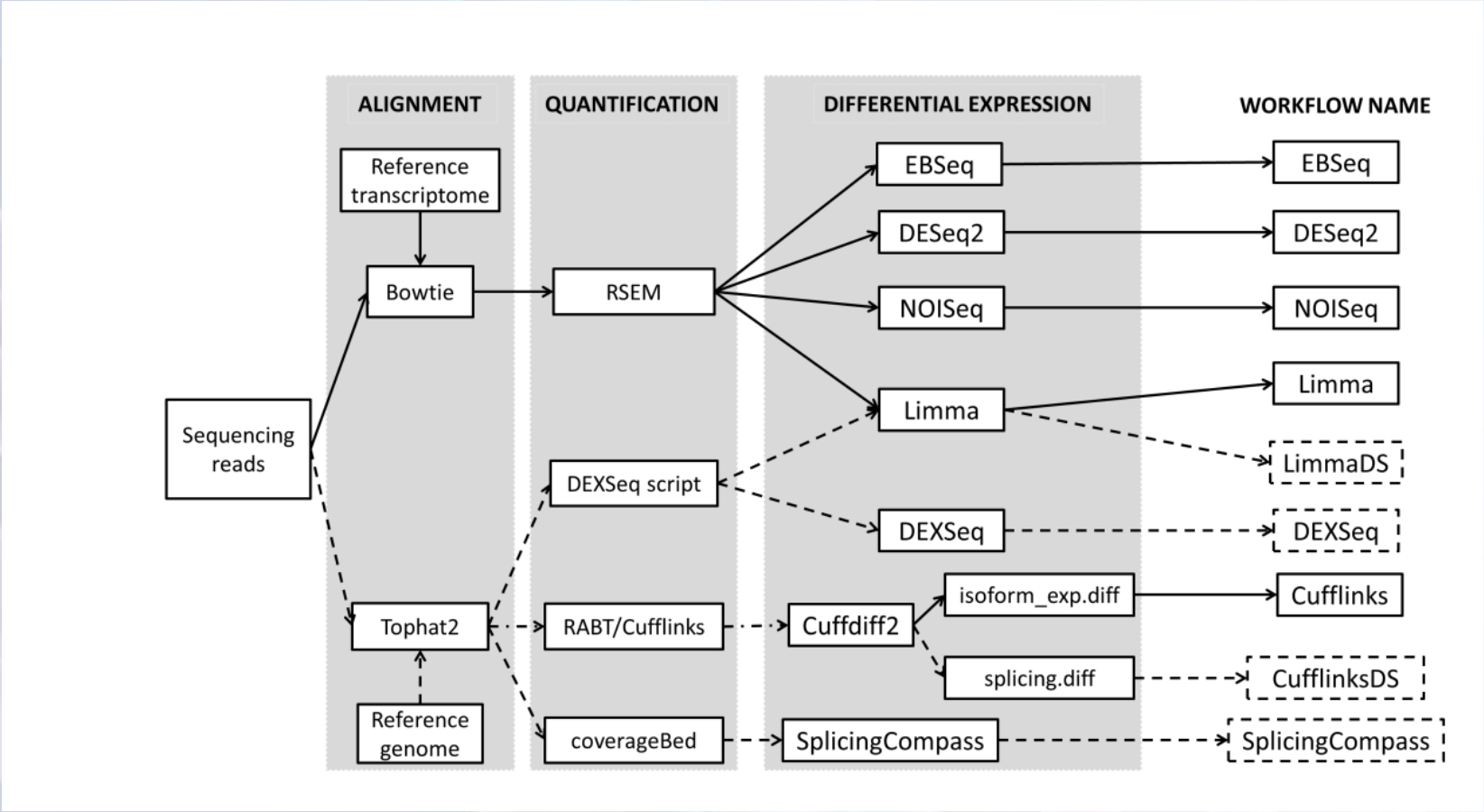
- **Tools**:
  - Velvet/Oases: http://www.ebi.ac.uk/~zerbino/oases/
  - Trinity: http://trinityrnaseq.sourceforge.net/
  - Trans-ABySS: http://www.bcgsc.ca/platform/bioinfo/software/trans-abyss
  - MIRA, CLC etc.
  - HISAT, STAR, StringTie

Análisis diferencial de la expresión

Utilizar **pruebas estadísticas** para decidir si una diferencia observada en el recuento de lecturas es significativa.

¿Qué **genes/isoformas** se expresan a diferentes niveles en diferentes condiciones?

Herramientas de expresión diferencial

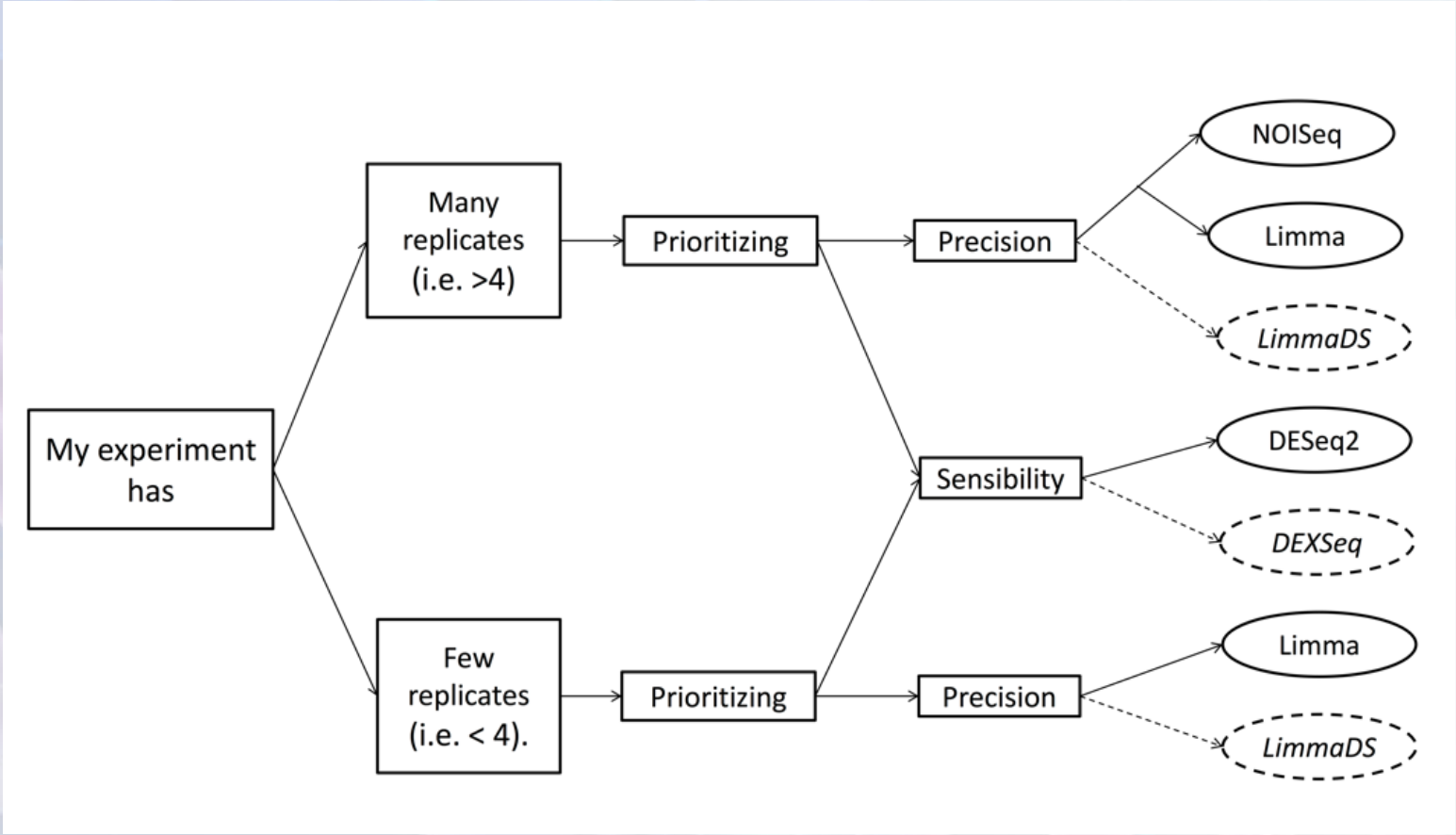


Briefings in Bioinformatics, Volume 20, Issue 2, March 2019, Pages 471–481

Diferentes herramientas producen resultados diferentes

Briefings in Bioinformatics, Volume 20, Issue 2, March 2019, Pages 471–481

Esquema de selección de flujos de trabajo.



Briefings in Bioinformatics, Volume 20, Issue 2, March 2019, Pages 471–481

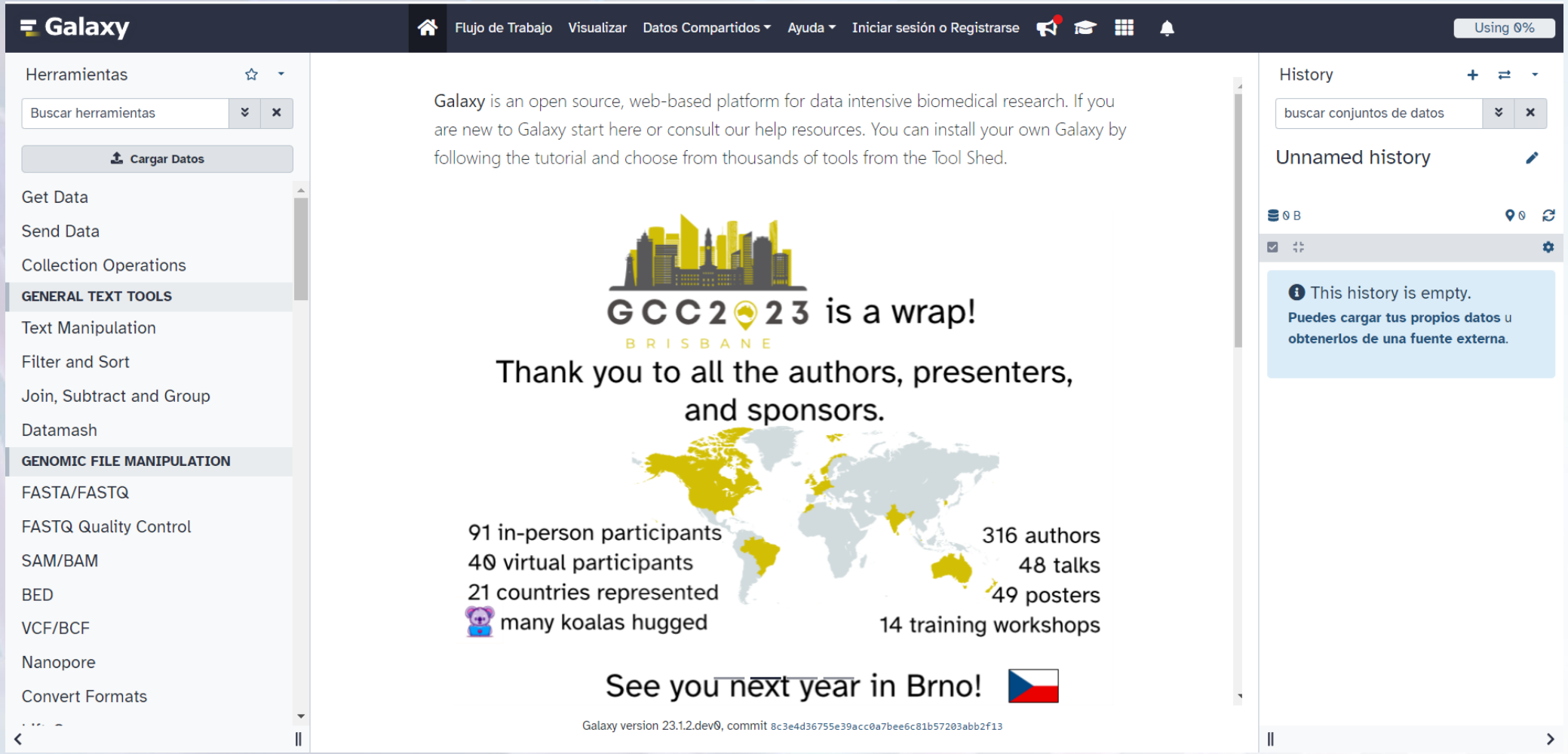
¿Cómo elijo la herramienta adecuada?

Comprender los requisitos de cada herramienta
p.ej. MMSEQ requiere alinear con el transcriptoma

Identificar cómo se comportan de manera diferente las herramientas y elija una en consecuencia
p.ej. Cufflinks (primero el mapeo) versus ABySS (primero el ensamblaje)

Elegir herramientas de uso común porque hay ayuda en línea

..o herramientas implementadas por alguien en el laboratorio/instituto
Porque siempre puedes molestarlo cuando no funciona.



https://usegalaxy.org/