

# Procesamiento Masivo de Datos

---

Alex Di Genova

August 21, 2023

Universidad de O'higgins

Bienvenida curso de PMD

Planificación curso PMD

# **Bienvenida curso de PMD**

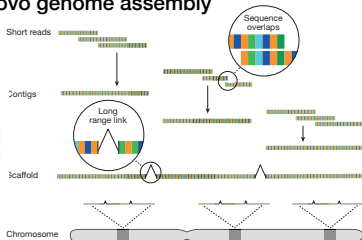
---

Alex Di Genova

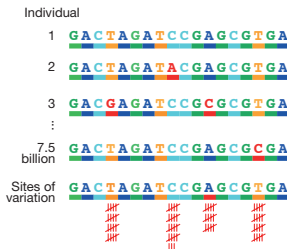
- 2003–2008 Ingeniero en Bioinformática.
- 2013-2017 Doctor en Sistemas Complejos.
- 2017-2021 Postdoctorado en algoritmos y cáncer (Francia).
- 2022 - Profesor Asistente UOH.
  - Di Genoma Lab
    - Combinamos el desarrollo de nuevos algoritmos, análisis de genomas y tecnologías ómicas de última generación para estudiar sistemas biológicos complejos.

# Sequencing technologies

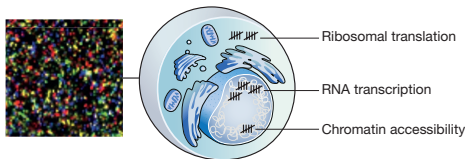
## De novo genome assembly



## Genome resequencing



## Sequencer as counting devices



Shendure, Jay, et al. DNA sequencing at 40: past, present and future. *Nature* 550.7676 (2017):

Sequence ID: CP030240.1 Length: 4582842 Number of Matches: 1

▼ Next Match ▲ Previous Match

Query	1	CACCAGCGTTAACAGCAATAACAGCAGGACGATAATCAGCAGGTGATTACGTGCCAGTCC	60
Sbjct	168128	CACCAGCGTTAACAGCAATAACAGCAGGACGATAATCAGCAGGTGATTACGTGCCAGTCC	168069

Sbict 168068 GCGGTAATGGTCATAAATTCGCTAAGAAAAATGTTGAAGGGCGGCATCCCTGCCAGCGC 168009

Seq1 16800 CAGCGCAC GCCGCCAC CCGC CGCGGTAA G CATGATT TGGG TCCACAGAC 167949

Query 181 GACG<sup>+</sup> GAGATCGCGCGT<sup>+</sup>GCCGTACT<sup>+</sup>T TATGCATGACAT<sup>+</sup>TGCCGGAACCGCAGAA CAGCAG 240

Aviary 167848 GACGTCAGATGGCGCTGGCTTCTGAGAGGTAAATCCGGAACCGCAGAACAGCAG 167889

Query 241 CGCTTTTGCCAGACTGTGGTTTAAGATGTGCAGCAGCGCGGCAAAAATTCCCAGCGGCC 300

Human genome (30X) = \$1500

Query 301 G 301

Human genome (30X, BGI) = ██████████ ~\$5

Escherichia coli strain ER1709 chromosome, complete genome

Sequence ID: CP030240.1 Length: 4582842 Number of Matches: 1

▼ Next Match ▲ Previous Match

Query	94	CGCCACGGCT---ACACGTCGGTAATGCACGGTTCGCC-ACCAGACATATGGCCAGAGC	148
Shict	129130	CGCCACGGCTGCACACACGTCGGTAATGCACGGTTCGCCACCGGAC--ATGGCCAGAGC	129187

Query 149 G--ATGGC-A-GCAGTCAAGGCT-A--C-ACGCGTC-GGCCAACGGTCAT-CCTGCCTGA 198

Shift 129188 GTCATGGCGATACCTTTAACGGTCAGGCTACGGGTCAGCCCGGGCGGTCATCCCTGCCTGA 129247

Query 199 TGC AAAAAGCTGTCTGCC-TCACGAACAGATGTCTTTCAGCCACGCGTTTGCACT-ACT 256

Sbict 129248 TGCAAAAAGCTGTCTGCCATCACGAACAGATG--TTTCAGCCCACGCGTTTGCGCTTGCT 129305

Query 257 GT-C-CTACT--TCTCTGAAG-CGGA---CATAAGCGTCTACCGGTGGAACGCTAAATGT 308

Sbjct 129306 GTCCGCAACTGCTCACCGAGCCCGGACCGCA-AAGCGTC-ACCGGTGGAACGCTAAATGT 129363

GenBank 309 TTTTACCGGTTGCAGATTCTGGGTATCGAATGCTCTGAAGTAAGGTCGTC-ACAATT 367

Sbjct 129364 TTTTACCGTTCAGATTCAAGG-TATCGACG-CTGAA-AGA-GCGCGTCTGCAATT 129419

Average error rate = 15%

Sbjct 129420 CATTTCGCATATAGCCCTCCGCTTCTCTCGCCCACTAC-TGGCGATCGCC-AGCGGGG 129477

Human genome (50X) = \$5000

Subject 129478 TCAT-AAACAAACTGTGCTGCGGT-AAATGCACCTGTAACGCCGGGAATTGTTTGC-GAA 129534

# Sequencing technologies

Long

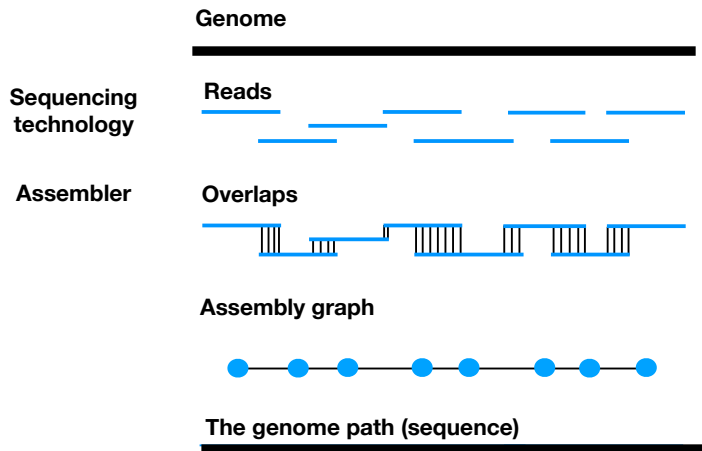
Read length = 15kb (Max 2Mb)

Query: 358 CG-TTGGCATAAGCCGCTGC-CT-GCTCTC-CTAAGAC

Human genome (30X-ONT) = \$5000

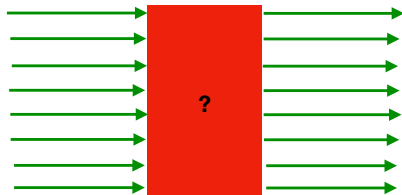
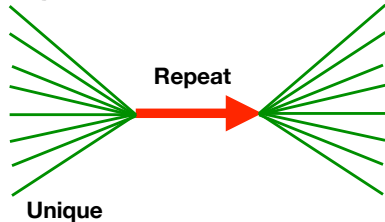
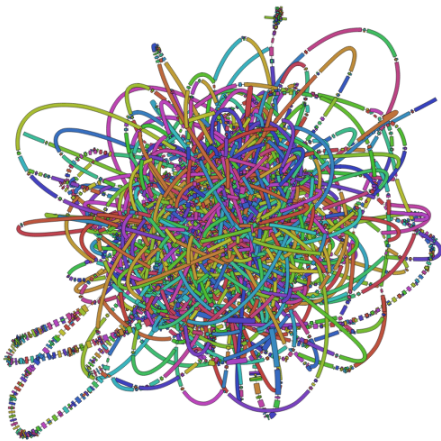
Human genome (30x PAC) — \$10000

# Genome assembly



40 years of genome assembly

# Genome assembly is complex



How do you get through?

How do we make “the” genome path?



# Hybrid assembly: How can we combine short and long reads?

Sequencing technology

Genome

Identical repeats



Short reads (< 300 bp, base error < 0.1%)



Long reads (>10 kb, base error <15%)



Wengan Assembler

(1) Assemble short-reads



(2) Scaffold using long reads



(3) Refine repetitive regions (polishing)

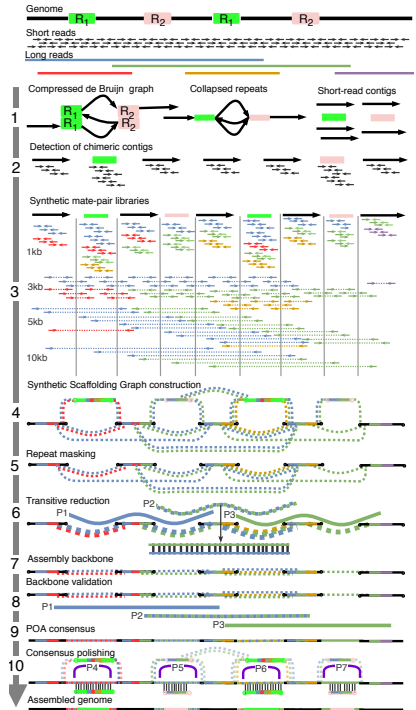


The resulting assembly is both *contiguous and accurate*



# Wengan: a new assembly paradigm

- **Full** hybrid assembler.
- Avoids entirely all-vs-all read comparisons (**fast**).
- A new assembly graph (*GoogleMaps*).
- 1.5 years of development.
- ~20k lines of code (C++, PERL)
- <https://github.com/adigenova/wengan>
- **Di Genova, A.** (2018). Fast-SG: an alignment-free algorithm for hybrid assembly. *GigaScience*, 7(5).
- **Di Genova, A.** (2021). Wengan: Efficient and high-quality hybrid de novo assembly of human genomes. *Nature Biotechnology*.



# Wengan



## ARTICLES

<https://doi.org/10.1038/s41587-020-00747-w>

nature  
biotechnology



## OPEN

## Efficient hybrid de novo assembly of human genomes with WENGAN

Alex Di Genova<sup>1,2</sup>✉, Elena Buena-Atienza<sup>3,4</sup>, Stephan Ossowski<sup>3,4</sup> and Marie-France Sagot<sup>1,2</sup>✉

Generating accurate genome assemblies of large, repeat-rich human genomes has proved difficult using only long, error-prone reads, and most human genomes assembled from long reads add accurate short reads to polish the consensus sequence. Here we report an algorithm for hybrid assembly, WENGAN, that provides very high quality at low computational cost. We demonstrate de novo assembly of four human genomes using a combination of sequencing data generated on ONT PromethION, PacBio Sequel, Illumina and MGI technology. WENGAN implements efficient algorithms to improve assembly contiguity as well as consensus quality. The resulting genome assemblies have high contiguity (contig NG50: 17.24–80.64 Mb), few assembly errors (contig NGA50: 11.8–59.59 Mb), good consensus quality (QV: 27.84–42.88) and high gene completeness (BUSCO complete: 94.6–95.2%), while consuming low computational resources (CPU hours: 187–1,200). In particular, the WENGAN assembly of the haploid CHM13 sample achieved a contig NG50 of 80.64 Mb (NGA50: 59.59 Mb), which surpasses the contiguity of the current human reference genome (GRCh38 contig NG50: 57.88 Mb).

**WENGAN** is a Mapudungun word.

**WENGAN** means "Making the path".

# Planificación curso PMD

---

- 4 Unidades (14 semanas)
  - Distribución y Paralelismo (4 semanas)
  - Modelamiento de Procesamiento Distribuido (4 semanas)
  - Modelos de Almacenamiento Escalable (3 semanas)
  - Bases de datos Distribuidas (3 semanas)

- Controles (75%, potenciales fechas de controles):
  - Control 1: Semana del 2 Octubre.
  - Control 2: Semana del 30 Octubre.
  - Control 3: Semana del 27 Noviembre.

- Controles (75%, potenciales fechas de controles):
  - Control 1: Semana del 2 Octubre.
  - Control 2: Semana del 30 Octubre.
  - Control 3: Semana del 27 Noviembre.
- Tareas (25%, potenciales fechas de entrega):
  - Tarea 1: Semana del 25 Septiembre.
  - Tarea 2: Semana del 20 Noviembre.

- Controles (75%, potenciales fechas de controles):
  - Control 1: Semana del 2 Octubre.
  - Control 2: Semana del 30 Octubre.
  - Control 3: Semana del 27 Noviembre.
- Tareas (25%, potenciales fechas de entrega):
  - Tarea 1: Semana del 25 Septiembre.
  - Tarea 2: Semana del 20 Noviembre.
- Examen (40%, potencial fecha del examen):
  - Examen : Semana del 11 Diciembre.



## Condiciones y Políticas de Evaluación

- El promedio de actividades complementarias se considerará como un cuarto control (control IV) y tendrá una ponderación de 25%. El promedio de controles I,II, III, IV con sus respectivas ponderaciones corresponderán a la nota de cátedra del curso.
- Examen: Los alumnos que cuenten con una nota de cátedra menor a 5. Deberán rendir el examen.
- Nota final: Promedio ponderado del examen (40%) y promedio de los controles (60%).
- Los alumnos eximidos recibirán como nota de examen el promedio de las notas de los controles de cátedra. Si lo desean, podrán rendir el examen, en cuyo caso se considerará la nota obtenida sólo si ésta es superior al promedio de las notas de los controles.

## Condiciones y Políticas de Evaluación

- Estudiantes que se ausenten a un control tendrán la oportunidad de recuperarlo con el examen. La nota del examen reemplazará la nota más baja de los controles de la asignatura, solo en caso de ser la nota de examen superior.
- Un/a estudiante que cometa plagio obtendrá un 1,0 en la evaluación y el caso será informado a Escuela de Ingeniería.

- Repositorio GitHub –  
<https://github.com/adigenova/uohpmd>
  - Slides Clases (Lunes y Miercoles : 12:00 - 13:30)
  - Código
- Ayudante : Iván Bozo Catalán (Jueves 16:15 - 17: 45)

# Materiales

- Repositorio GitHub –  
<https://github.com/adigenova/uohpmd>
  - Slides Clases (Lunes y Miercoles : 12:00 - 13:30)
  - Código
- Ayudante : Iván Bozo Catalán (Jueves 16:15 - 17: 45)
- Ucampus –  
<https://ucampus.uoh.cl/uoh/2023/2/COM4002>
  - Comunicación (Consultas, noticias, evaluaciones)
  - Planificación

- Repositorio GitHub –  
<https://github.com/adigenova/uohpmd>
  - Slides Clases (Lunes y Miercoles : 12:00 - 13:30)
  - Código
- Ayudante : Iván Bozo Catalán (Jueves 16:15 - 17: 45)
- Ucampus –  
<https://ucampus.uoh.cl/uoh/2023/2/COM4002>
  - Comunicación (Consultas, noticias, evaluaciones)
  - Planificación
- Bibliografía
  - S. Tanenbaum, M. Van Steen. Distributed Systems: Principles and Paradigms (2nd Edition). Prentice Hall, 2006
  - P. J. Sadalage, M. Fowler. NoSQL Distilled: A Brief Guide to the Emerging World of Polyglot Persistence. Addison-Wesley Professional, 2012

## Resultados de Aprendizaje

- Conocer los principios fundamentales de diseño de sistemas distribuidos y paralelos

## Resultados de Aprendizaje

- Conocer los principios fundamentales de diseño de sistemas distribuidos y paralelos
- Implementar comunicación entre procesos de una máquina (OpenMP, pthread)

## Resultados de Aprendizaje

- Conocer los principios fundamentales de diseño de sistemas distribuidos y paralelos
- Implementar comunicación entre procesos de una máquina (OpenMP, pthread)
- Implementar comunicación entre máquinas (MPI)



## Resultados de Aprendizaje

- Conocer los principios fundamentales de diseño de sistemas distribuidos y paralelos
- Implementar comunicación entre procesos de una máquina (OpenMP, pthread)
- Implementar comunicación entre máquinas (MPI)
- Utilizar Nextflow/Hadoop para distribuir tareas computacionales en un clúster.

## Resultados de Aprendizaje

- Conocer los principios fundamentales de diseño de sistemas distribuidos y paralelos
- Implementar comunicación entre procesos de una máquina (OpenMP, pthread)
- Implementar comunicación entre máquinas (MPI)
- Utilizar Nextflow/Hadoop para distribuir tareas computacionales en un clúster.
- Conocer la taxonomía de modelos de datos NoSQL, sus lenguajes de consulta y manipulación de datos con modelos NO-SQL.

## Resultados de Aprendizaje

- Conocer los principios fundamentales de diseño de sistemas distribuidos y paralelos
- Implementar comunicación entre procesos de una máquina (OpenMP, pthread)
- Implementar comunicación entre máquinas (MPI)
- Utilizar Nextflow/Hadoop para distribuir tareas computacionales en un clúster.
- Conocer la taxonomía de modelos de datos NoSQL, sus lenguajes de consulta y manipulación de datos con modelos NO-SQL.
- Construir y manipular una base de datos distribuida.

Consultas o comentarios?  
Muchas Gracias.