

# Procesamiento Masivo de Datos

---

Alex Di Genova

August 22, 2022

Universidad de O'higgins

Bienvenida curso de PMD

Planificación curso PMD

# **Bienvenida curso de PMD**

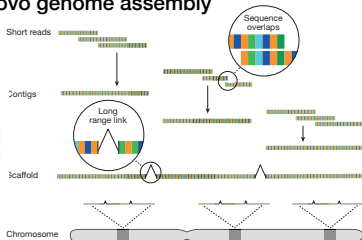
---

Alex Di Genova

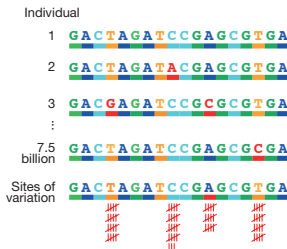
- 2003–2008 Ingeniero en Bioinformática.
- 2013-2017 Doctor en Sistemas Complejos.
- 2017-2021 Postdoctorado en algoritmos y cáncer (Francia).
- 2022 - Profesor Asistente UOH.
  - Di Genoma Lab
    - Combinamos el desarrollo de nuevos algoritmos, análisis de genomas y tecnologías ómicas de última generación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

# Sequencing technologies

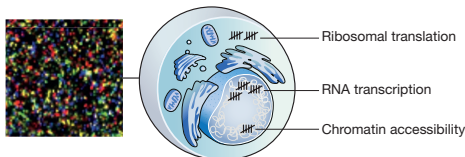
## De novo genome assembly



## 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
Subject	168128	CACCAGCGTTAACAGCAATAACAGCAGGACGATAATCAGCAGGTGATTACGTGCCAGTCC	168069

Sbict 168068 GCGGTAATGGTCATAAATTCGCTAAGAAAAATGTTGAAGGGCGGCATCCCTGCCAGCGC 168009

dict 16100 CAGCGCAC GCCGCCA C C C GCGGTA T G CATGATT G S TCCACAGAC 167949

167848 GACGGTCGACATGGGCGCCGCGCTACTTTCAGCTGTAACTCCGGAACCGCAGAACAGCAG 167889

**Human genome (30X) – \$1500**

Human genome (30X, BGI) = ~~1000~~ \$5

[View all posts by](#) [Bryan Smith](#)

Short

Range 1: 129130 to 13

1661 bits(899)

Query 94 CGC

Sbjct 129130 CGC

Query 149 G--A

Sbjct 129188 GTC

Query 199 TGC

Sbjct 129248 TGC

Query 257 GT-0

Sbjct 129306 GTC

Page 309 TTT

Sbjct 129364 TTT

Query 368 C2-7

Sbjct 129420 CAT

Journal Pre-proof

Sbjct 129478 TCA

Humanity 475 AC 10

Long

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

▼ Next Match ▲ Previous Match

Query 94 CGCCACGGCT----ACACGTCGGTAATGCACGGTTCGCC--ACCAGACATATGGCCAGAGC 148

[illegible]

Genbank 100 TCCGAAAACGGTCGTCGGC TCACCAACAGATCTGCTTCAGCGGACGGCGTTCGAGT-ACG 256

Shint 120249  120205

**Query** 257 GCG CGGAGCG TGGGTCTAC GGCAT CATTAAGGTCGTTGGGTCCTTACGCCAATATGCG 309

Shint 120206 GTCCGCAAGTCCGACGCGACGCGCGACGCGCA-AAGCGTC-AAGCGCTGCAAGCGCTAAATG 120262

200     TTTTACGGGCTTCGACATTCGAGCGATGCGTTCGCGCAACGATGCGCGCTC-ACCAATG 267

Read length = 15kb (Max 2Mb)

Genes 2019, 10, 1100

Average error rate = 15%

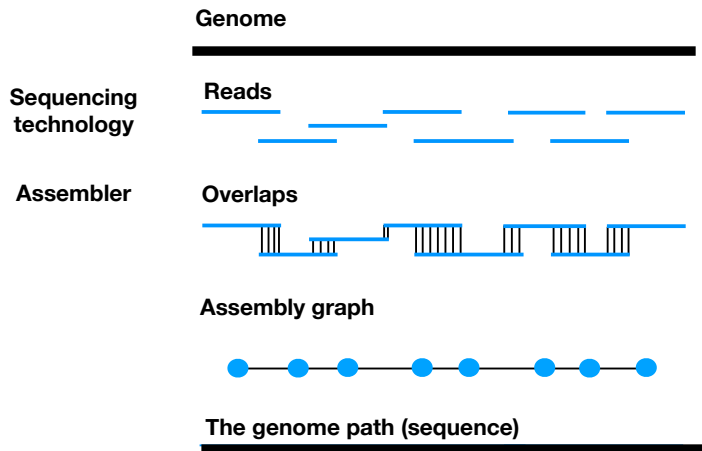
## Human genome (30X ONT) – \$5000

Shit 120470 TCGT- AAAGAAAGCTCTCGTCCCGT- AAATCCAGCTCTAAGCCCGCCCAATTCCTTCC- GAA 120524

# Human genome (30X PAC) – \$14999 \$1

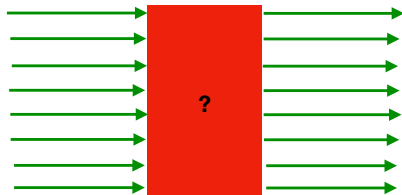
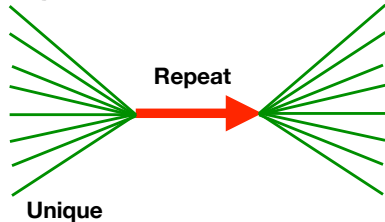
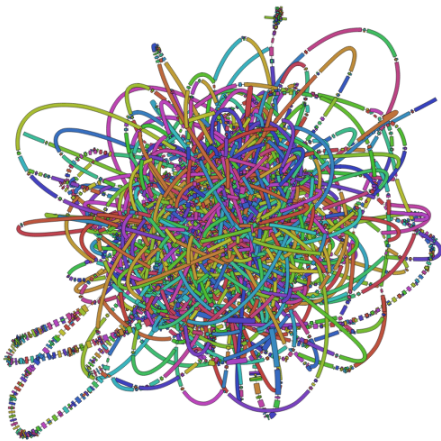
5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32 33 34 35 36 37 38 39 40 41 42 43 44 45 46 47 48 49 50 51 52 53 54 55 56 57 58 59 60 61 62 63 64 65 66 67 68 69 70 71 72 73 74 75 76 77 78 79 80 81 82 83 84 85 86 87 88 89 90 91 92 93 94 95 96 97 98 99 100 101 102 103 104 105 106 107 108 109 110 111 112 113 114 115 116 117 118 119 120 121 122 123 124 125 126 127 128 129 130 131 132 133 134 135 136 137 138 139 140 141 142 143 144 145 146 147 148 149 150 151 152 153 154 155 156 157 158 159 160 161 162 163 164 165 166 167 168 169 170 171 172 173 174 175 176 177 178 179 180 181 182 183 184 185 186 187 188 189 190 191 192 193 194 195 196 197 198 199 200 201 202 203 204 205 206 207 208 209 210 211 212 213 214 215 216 217 218 219 220 221 222 223 224 225 226 227 228 229 230 231 232 233 234 235 236 237 238 239 240 241 242 243 244 245 246 247 248 249 250 251 252 253 254 255 256 257 258 259 260 261 262 263 264 265 266 267 268 269 270 271 272 273 274 275 276 277 278 279 280 281 282 283 284 285 286 287 288 289 290 291 292 293 294 295 296 297 298 299 300 301 302 303 304 305 306 307 308 309 310 311 312 313 314 315 316 317 318 319 320 321 322 323 324 325 326 327 328 329 330 331 332 333 334 335 336 337 338 339 340 341 342 343 344 345 346 347 348 349 350 351 352 353 354 355 356 357 358 359 360 361 362 363 364 365 366 367 368 369 370 371 372 373 374 375 376 377 378 379 380 381 382 383 384 385 386 387 388 389 390 391 392 393 394 395 396 397 398 399 400 401 402 403 404 405 406 407 408 409 410 411 412 413 414 415 416 417 418 419 420 421 422 423 424 425 426 427 428 429 430 431 432 433 434 435 436 437 438 439 440 441 442 443 444 445 446 447 448 449 450 451 452 453 454 455 456 457 458 459 460 461 462 463 464 465 466 467 468 469 470 471 472 473 474 475 476 477 478 479 480 481 482 483 484 485 486 487 488 489 490 491 492 493 494 495 496 497 498 499 500 501 502 503 504 505 506 507 508 509 510 511 512 513 514 515 516 517 518 519 520 521 522 523 524 525 526 527 528 529 530 531 532 533 534 535 536 537 538 539 540 541 542 543 544 545 546 547 548 549 550 551 552 553 554 555 556 557 558 559 560 561 562 563 564 565 566 567 568 569 570 571 572 573 574 575 576 577 578 579 580 581 582 583 584 585 586 587 588 589 590 591 592 593 594 595 596 597 598 599 600 601 602 603 604 605 606 607 608 609 610 611 612 613 614 615 616 617 618 619 620 621 622 623 624 625 626 627 628 629 630 631 632 633 634 635 636 637 638 639 640 641 642 643 644 645 646 647 648 649 650 651 652 653 654 655 656 657 658 659 660 661 662 663 664 665 666 667 668 669 670 671 672 673 674 675 676 677 678 679 680 681 682 683 684 685 686 687 688 689 690 691 692 693 694 695 696 697 698 699 700 701 702 703 704 705 706 707 708 709 710 711 712 713 714 715 716 717 718 719 720 721 722 723 724 725 726 727 728 729 730 731 732 733 734 735 736 737 738 739 740 741 742 743 744 745 746 747 748 749 750 751 752 753 754 755 756 757 758 759 760 761 762 763 764 765 766 767 768 769 770 771 772 773 774 775 776 777 778 779 780 781 782 783 784 785 786 787 788 789 790 791 792 793 794 795 796 797 798 799 800 801 802 803 804 805 806 807 808 809 810 811 812 813 814 815 816 817 818 819 820 821 822 823 824 825 826 827 828 829 830 831 832 833 834 835 836 837 838 839 840 841 842 843 844 845 846 847 848 849 850 851 852 853 854 855 856 857 858 859 860 861 862 863 864 865 866 867 868 869 870 871 872 873 874 875 876 877 878 879 880 881 882 883 884 885 886 887 888 889 890 891 892 893 894 895 896 897 898 899 900 901 902 903 904 905 906 907 908 909 910 911 912 913 914 915 916 917 918 919 920 921 922 923 924 925 926 927 928 929 930 931 932 933 934 935 936 937 938 939 940 941 942 943 944 945 946 947 948 949 950 951 952 953 954 955 956 957 958 959 960 961 962 963 964 965 966 967 968 969 970 971 972 973 974 975 976 977 978 979 980 981 982 983 984 985 986 987 988 989 990 991 992 993 994 995 996 997 998 999 1000 1001 1002 1003 1004 1005 1006 1007 1008 1009 1010 1011 1012 1013 1014 1015 1016 1017 1018 1019 1020 1021 1022 1023 1024 1025 1026 1027 1028 1029 1030 1031 1032 1033 1034 1035 1036 1037 1038 1039 1040 1041 1042

# 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)

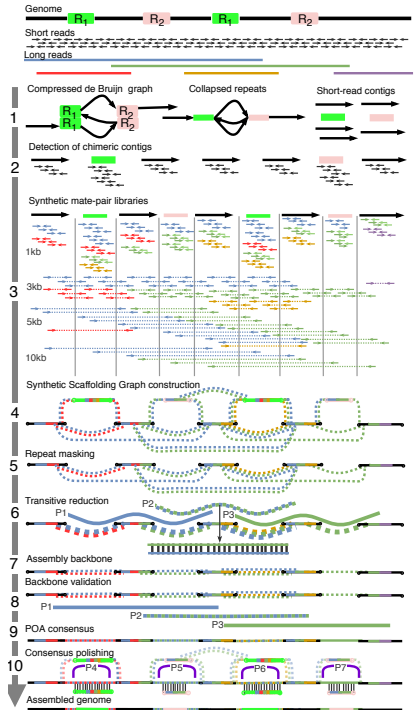


The resulting assembly is both *contiguous and accurate*



# 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 (70%):
  - Control 1: Semana del 3 Octubre.
  - Control 2: Semana del 28 Noviembre.

- Controles (70%):
  - Control 1: Semana del 3 Octubre.
  - Control 2: Semana del 28 Noviembre.
- Tareas (30%):
  - Tarea 1: Semana del 26 Septiembre.
  - Tarea 2: Semana del 7 Noviembre.
  - Tarea 3: Semana del 28 Noviembre

## Condiciones y Políticas de Evaluación

- El promedio de actividades complementarias se considerará como un tercer control (control III) y tendrá una ponderación de 30%. El promedio de controles I,II y III con sus respectivas ponderaciones corresponderán a la nota final del curso. El curso será aprobado con una nota promedio igual o superior a 4,0.



## Condiciones y Políticas de Evaluación

- El promedio de actividades complementarias se considerará como un tercer control (control III) y tendrá una ponderación de 30%. El promedio de controles I,II y III con sus respectivas ponderaciones corresponderán a la nota final del curso. El curso será aprobado con una nota promedio igual o superior a 4,0.
- Estudiantes que se ausenten a un control tendrán la oportunidad de recuperarlo durante el periodo correspondiente al final del semestre. El control recuperativo es de carácter **acumulativo**, por lo tanto, contendrá contenido de las cuatro unidades del curso. Adicionalmente, alumnos que quieran remplazar una calificación en un control o actividades complementarias, también podrán rendir el control recuperativo.

## Condiciones y Políticas de Evaluación

- El promedio de actividades complementarias se considerará como un tercer control (control III) y tendrá una ponderación de 30%. El promedio de controles I,II y III con sus respectivas ponderaciones corresponderán a la nota final del curso. El curso será aprobado con una nota promedio igual o superior a 4,0.
- Estudiantes que se ausenten a un control tendrán la oportunidad de recuperarlo durante el periodo correspondiente al final del semestre. El control recuperativo es de carácter **acumulativo**, por lo tanto, contendrá contenido de las cuatro unidades del curso. Adicionalmente, alumnos que quieran remplazar una calificación en un control o actividades complementarias, también podrán rendir el control recuperativo.
- 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>
  - Clases
  - Código

# Materiales

- Repositorio GitHub – <https://github.com/adigenova/uohpmd>
  - Clases
  - Código
- Ucampus – <https://ucampus.uoh.cl/uoh/2022/2/COM400>
  - Comunicación (Consultas, noticias, evaluaciones)
  - Planificación

- Repositorio GitHub –  
<https://github.com/adigenova/uohpmd>
  - Clases
  - Código
- Ucampus – <https://ucampus.uoh.cl/uoh/2022/2/COM400>
  - 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

- 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 máquinas (MPI, RMI)

## Resultados de Aprendizaje

- Conocer los principios fundamentales de diseño de sistemas distribuidos y paralelos
- Implementar comunicación entre máquinas (MPI, RMI)
- Utilizar Nextflow/Hadoop para distribuir tareas computacionales básicas.



## Resultados de Aprendizaje

- Conocer los principios fundamentales de diseño de sistemas distribuidos y paralelos
- Implementar comunicación entre máquinas (MPI, RMI)
- Utilizar Nextflow/Hadoop para distribuir tareas computacionales básicas.
- Conocer la taxonomía de modelos de datos NoSQL, sus lenguajes de consulta y construir una base de datos NO-SQL.

## Resultados de Aprendizaje

- Conocer los principios fundamentales de diseño de sistemas distribuidos y paralelos
- Implementar comunicación entre máquinas (MPI, RMI)
- Utilizar Nextflow/Hadoop para distribuir tareas computacionales básicas.
- Conocer la taxonomía de modelos de datos NoSQL, sus lenguajes de consulta y construir una base de datos NO-SQL.
- Construir una base de datos distribuida.

Consultas o comentarios?  
Muchas Gracias.