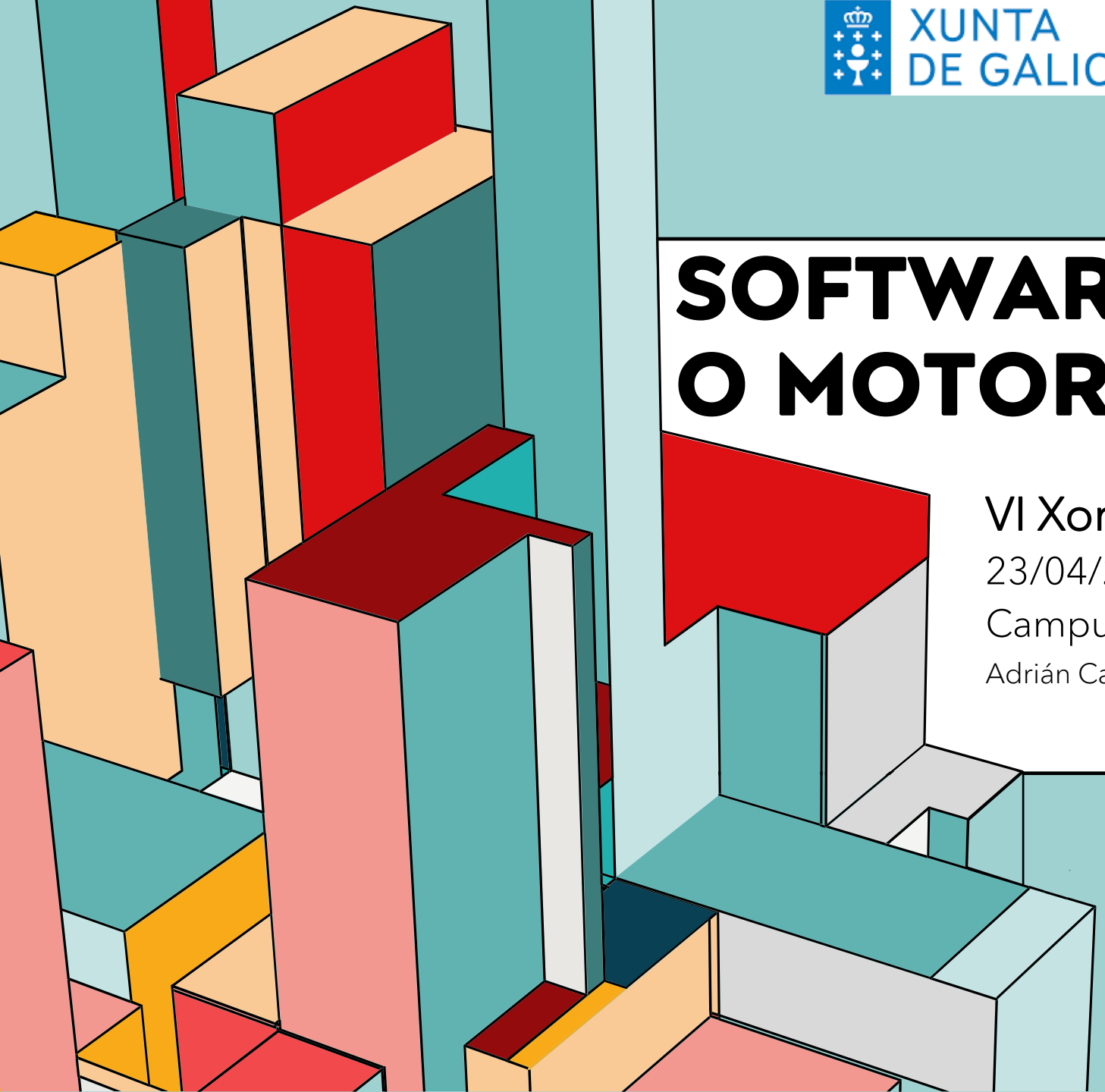


XUNTA DE GALICIA | CONSELLERÍA DE CULTURA, EDUCACIÓN, FORMACIÓN PROFESIONAL E UNIVERSIDADES

Xacobeo 21·22

SOFTWARE LIBRE: O MOTOR DA XENÓMICA

VI Xornada Software Libre Científico

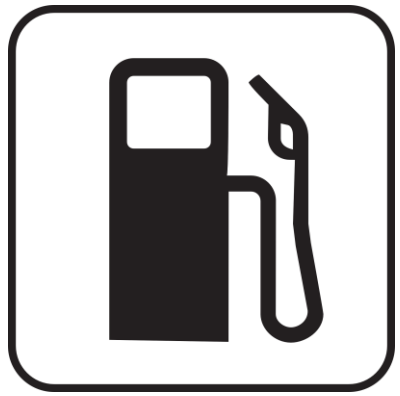
23/04/2024

Campus Terra (USC; Lugo)

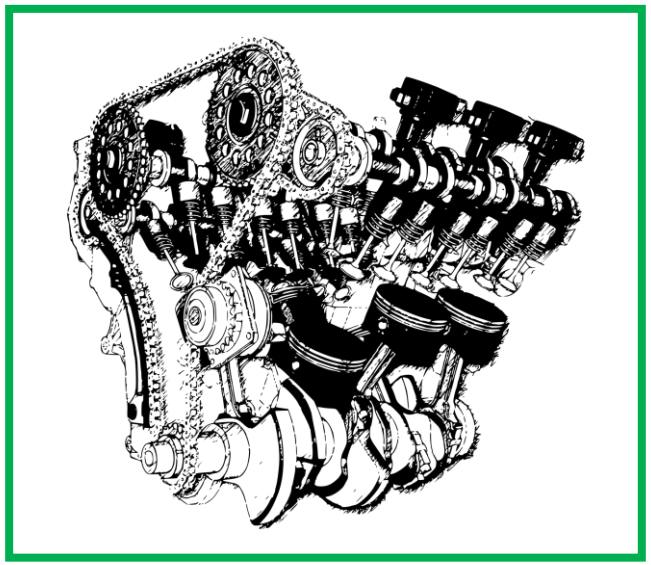
Adrián Casanova Chiclana

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

CAMPUS TERRA



DATOS BRUTOS
ADN



XENÓMICA

BIOINFORMÁTICA

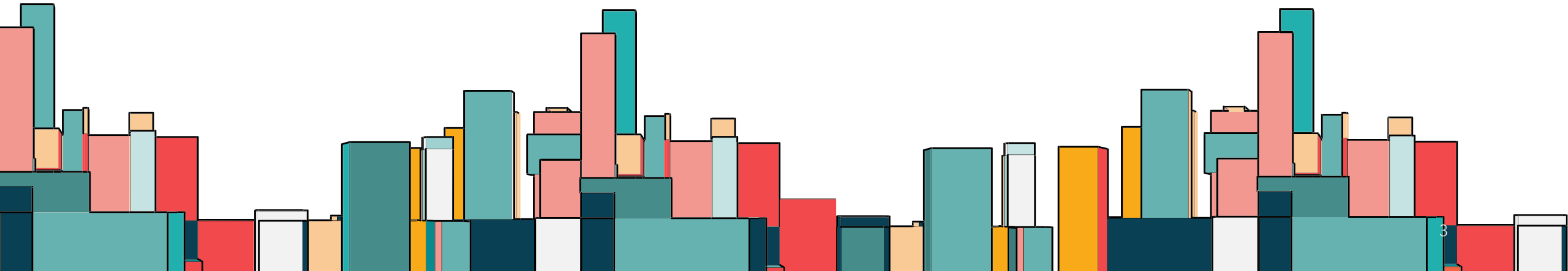
ÍNDICE

Conceptos
Para non perdernos

Contexto
Big (Bang) Data!

Un paseo xenómico
Abrochade os cintos!

Algúns resultados xenómicos
Con softwaRe libRe e gRatuíto



CONCEPTOS

Para non perdernos

QUE SON OS DATOS -ÓMICOS?

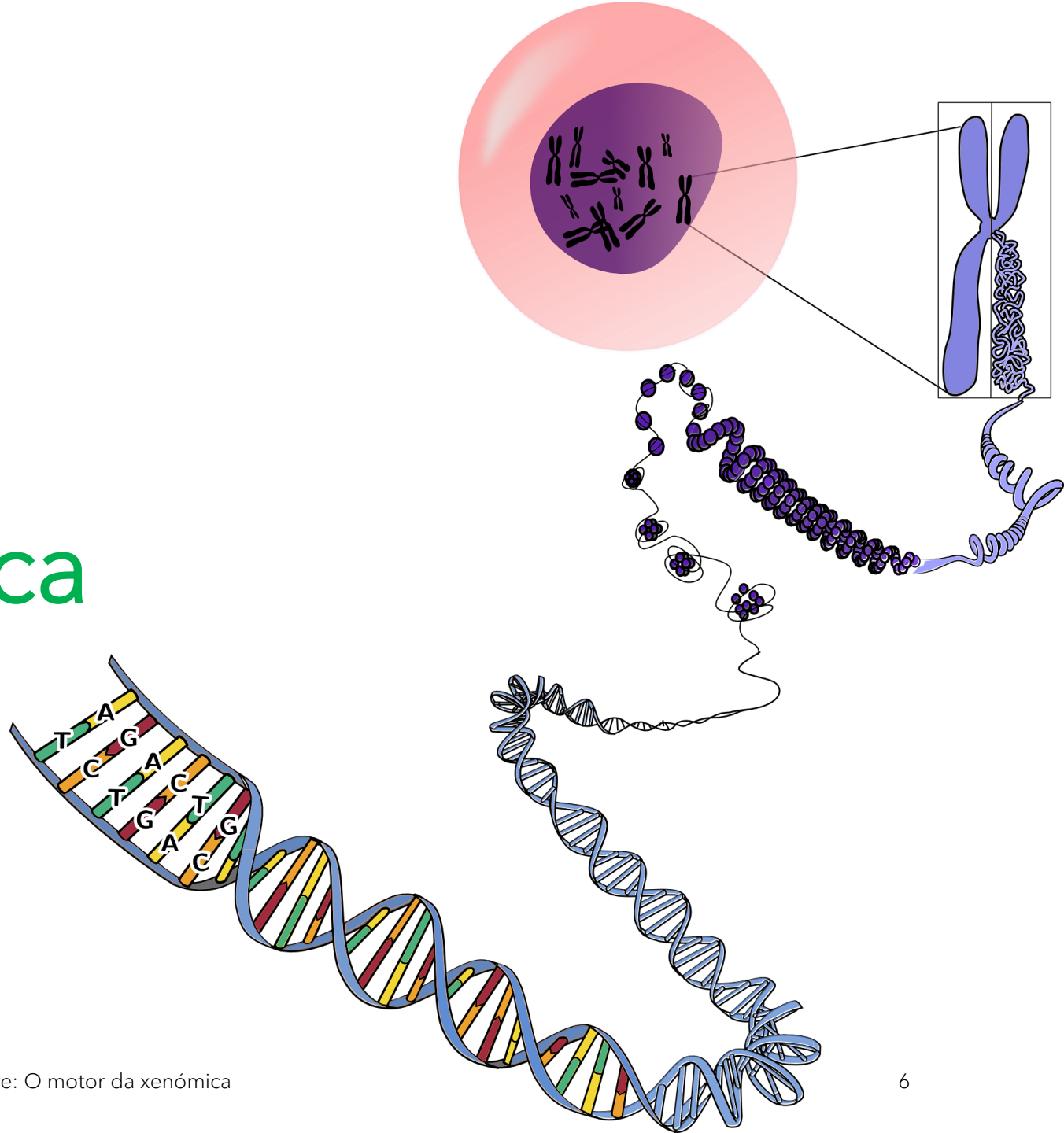
Este sufixo é empregado para abranguer a totalidade de entidades biolóxicas, como pode ser:

- Xenoma → Xenómica
- Transcriptoma → Transcriptómica
- Epixenoma → Epixenómica

...

ADN: A LINGUAXE DA VIDA

Xenoma → Xenómica



RESTRICTION SITE-ASSOCIATED DNA SEQUENCING (RAD-SEQ)

SL3-1 Xenoma completo (~3 Gb)

Subconxunto xenómico (<1% xenoma)

…AAAGCTGGCATCGATTGGATTGCGACCTCTTATCTGTGCACGTGCCTGCCACGAAGTTATAGGGCTAGTTACTGATGA…

AlfI enzima de restricción
(tesoira molecular)

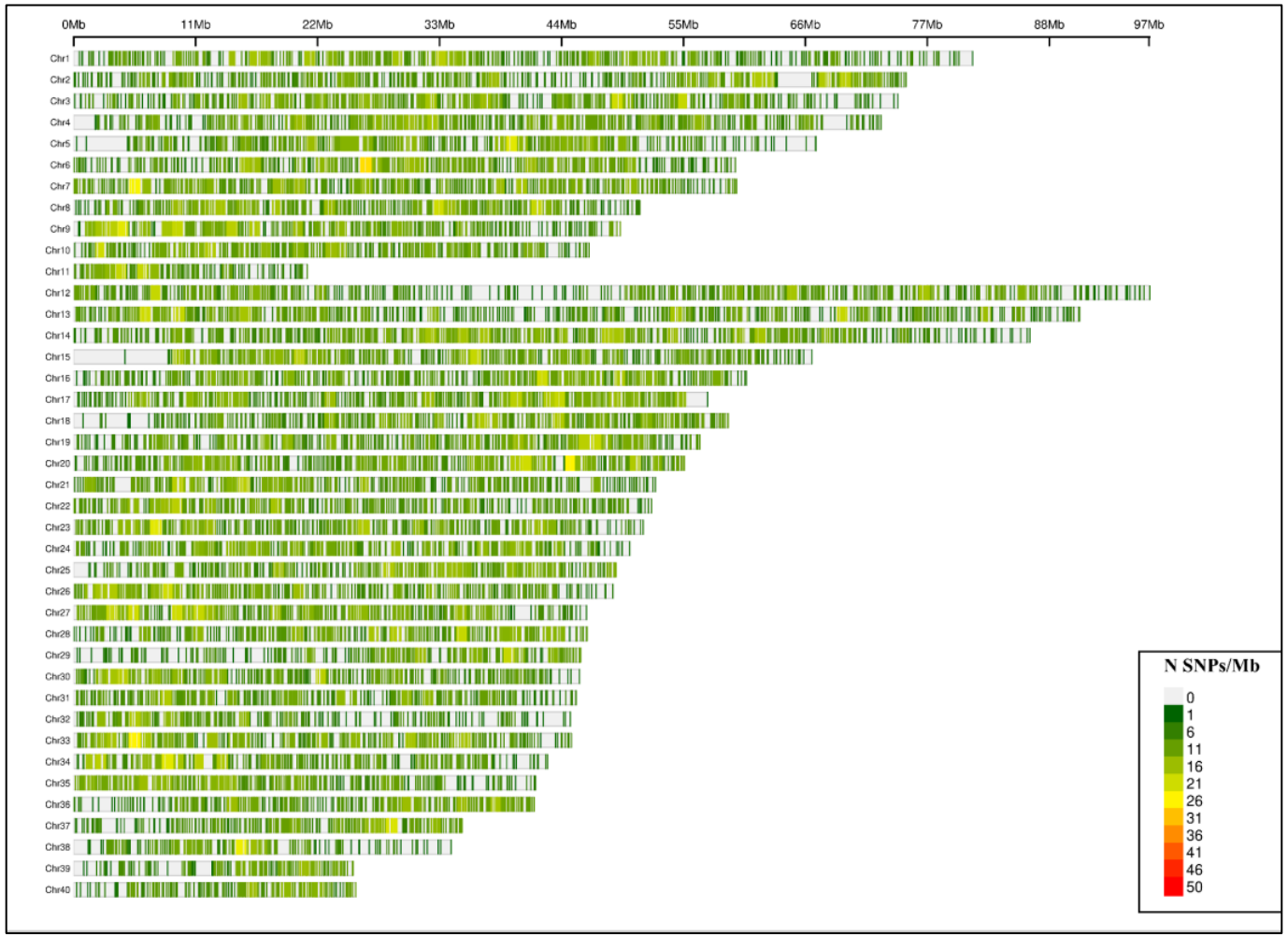
SNP 102456_27
C/G

CCTCTTATCTGTGCACGTGCCTGCCACGAAGTTATA
CCTCTTATCTGTGCACGTGCCTGCCACGAAGTTATA
CCTCTTATCTGTGCACGTGCCTGCCACGAAGTTATA
CCTCTTATCTGTGCACGTGCCTGCCACGAAGTTATA
CCTCTTATCTGTGCACGTGCCTGCCACGAAGTTATA
CCTCTTATCTGTGCACGTGCCTGCCAGGAAGTTATA
CCTCTTATCTGTGCACGTGCCTGCCAGGAAGTTATA

C
5x

G
2x

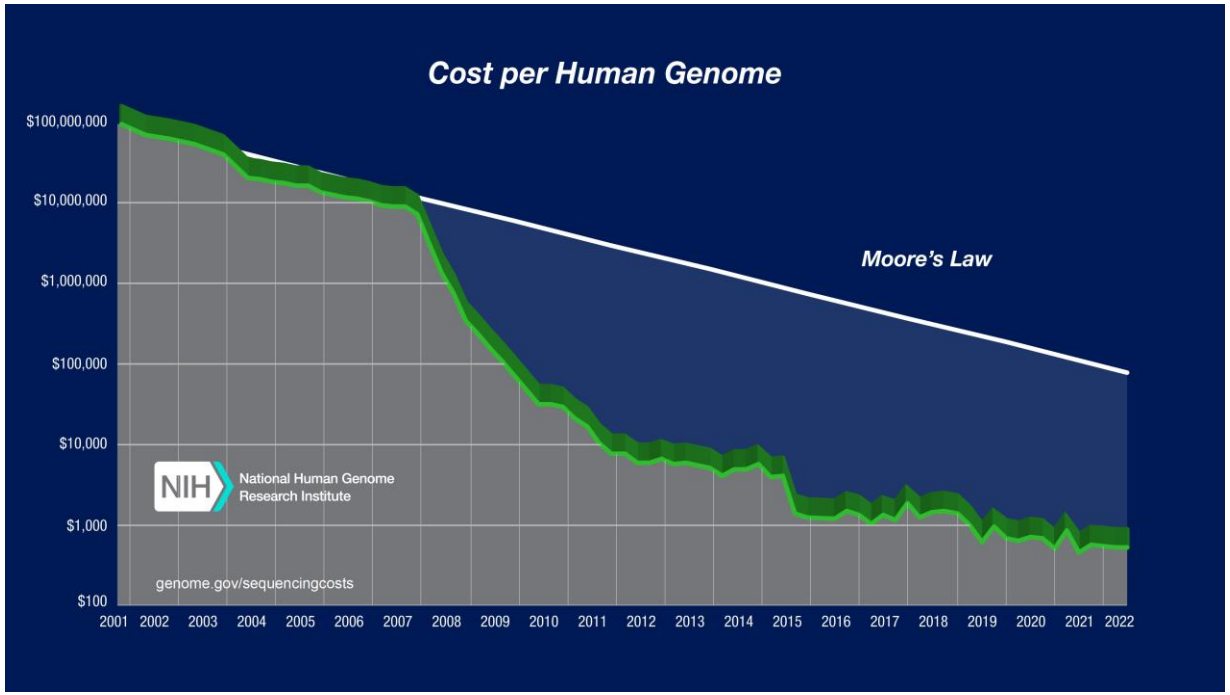
PANEL DE SNPs

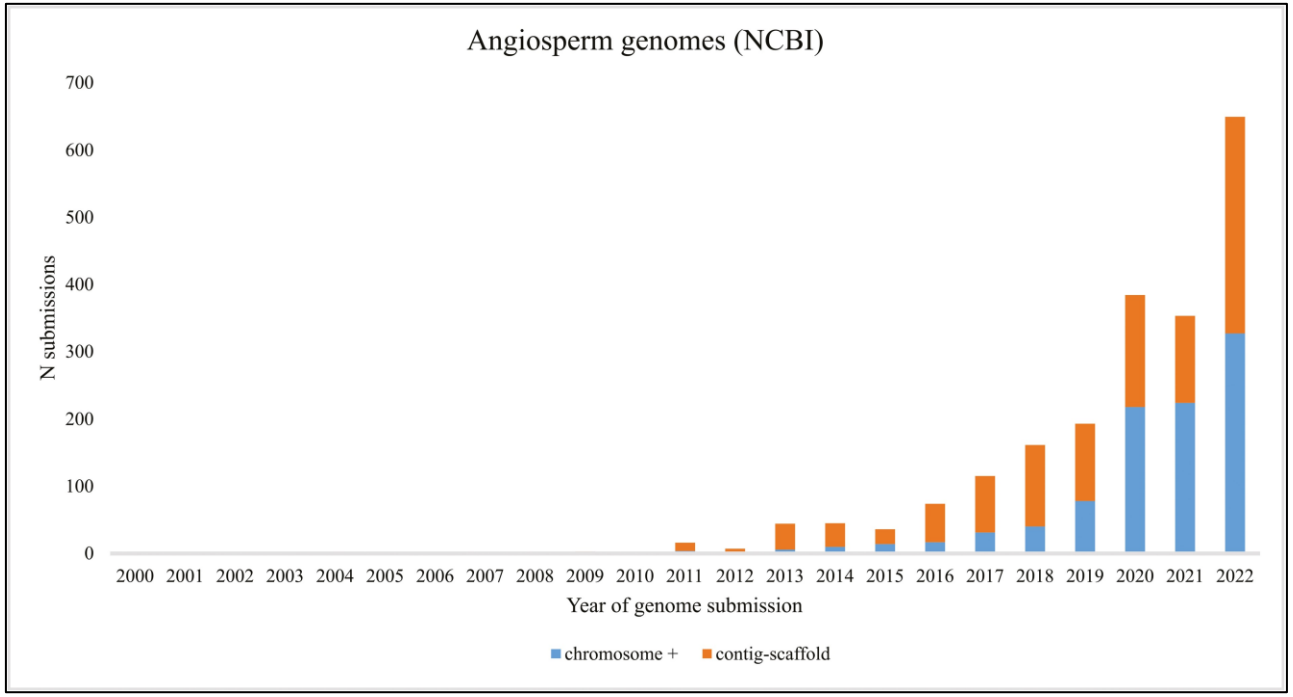


CONTEXTO

Big (Bang) Data!

As tecnoloxías de secuenciación de alto rendemento (*High Throughput Sequencing*; HTS) supuxeron un gran avance para os estudos xenéticos debido á **forte redución dos custos de secuenciación**, implicando a democratización das aproximacións xenómicas.





Wetterstrand KA. DNA Sequencing Costs: Data | NHGRI. 2022 (8 xaneiro 2024): https://www.genome.gov/about-genomics/fact-sheets/DNA-Sequencing-Costs-Data.

Casanova, A. Digest: Focus on chromosomes: how to understand angiosperm radiation. Evolution 77(6), 1491–1492 (2023). https://doi.org/10.1093/evolut/qpad053

Incremento exponencial dos datos xerados e dispoñibles (en aberto)

Genome assembly ASM3086718v1 reference

Download datasets curl FTP

Actions

Submitted GenBank assembly	GCA_030867185.1
Taxon	Tagetes erecta (African marigold)
Isolate	WSJ
WGS project	JAUHHV01
Assembly type	haploid
Submitter	Agricultural Genomics Institute at Shenzhen,Chinese Academy of Agricultural Sciences
Date	Aug 24, 2023

View the legacy Assembly page

View annotated genes

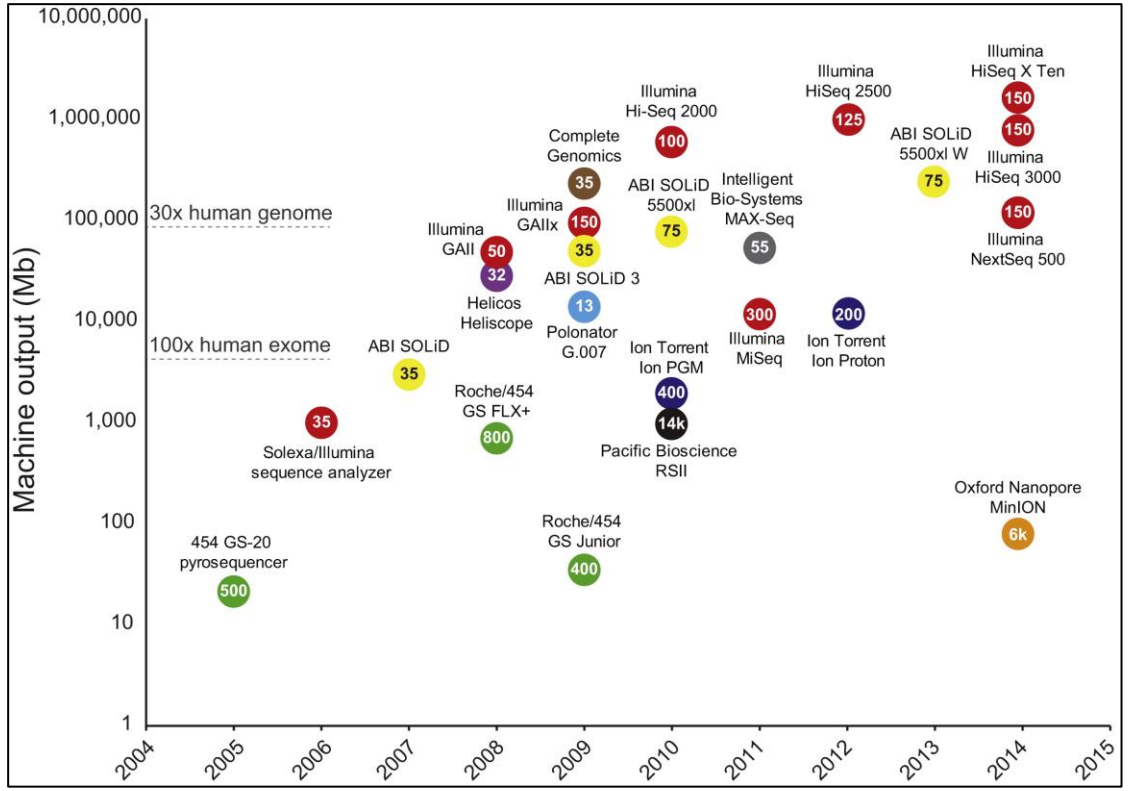
BLAST the reference genome

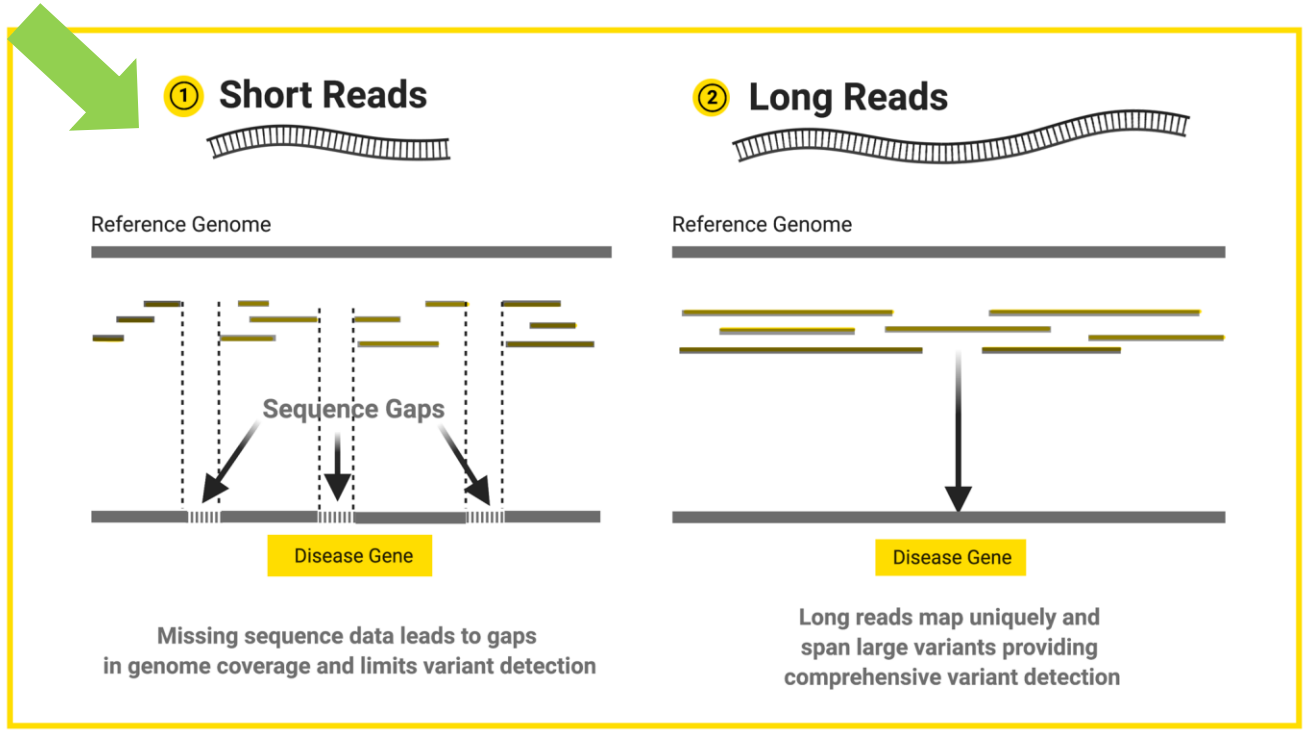
https://www.ncbi.nlm.nih.gov/datasets/genome/GCA_030867185.1/

Kai Yan, Joseph Wong | cc-by-nc-sa. Encyclopedia Of Life (EOL): https://eol.org/

Nos últimos anos desenvolvéronse múltiples tecnoloxías de secuenciación

Diferentes tecnoloxías → Diferentes outputs → Diferentes pipelines bioinformáticas

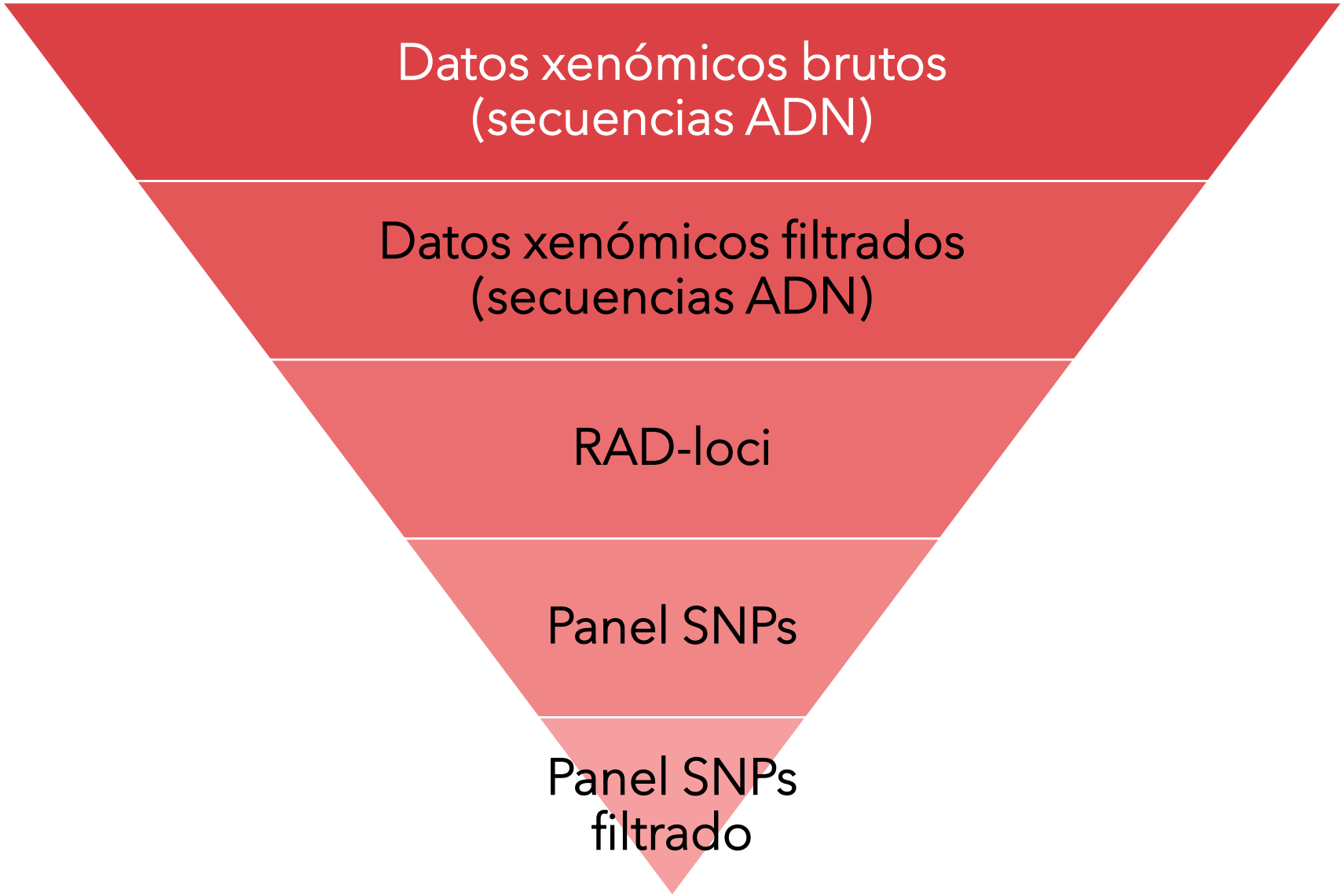




Reuter, J. A., Spacek, D. V., & Snyder, M. P. High-throughput sequencing technologies. Molecular cell 58(4), 586–597 (2015). https://doi.org/10.1016/j.molcel.2015.05.004

Sharman, S. (2021, February 3). *Piecing together the genome: The long and short of it all*. HudsonAlpha Institute for Biotechnology. https://www.hudsonalpha.org/piecing-together-the-genome-the-long-and-short-of-it-all/

Datos xenómicos



300 GB

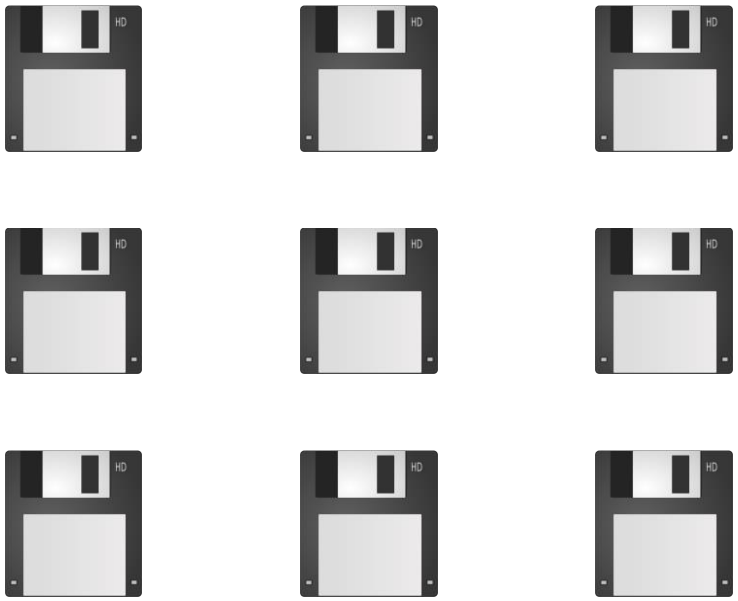
250 GB

50 GB

1.5 GB

0.5 GB

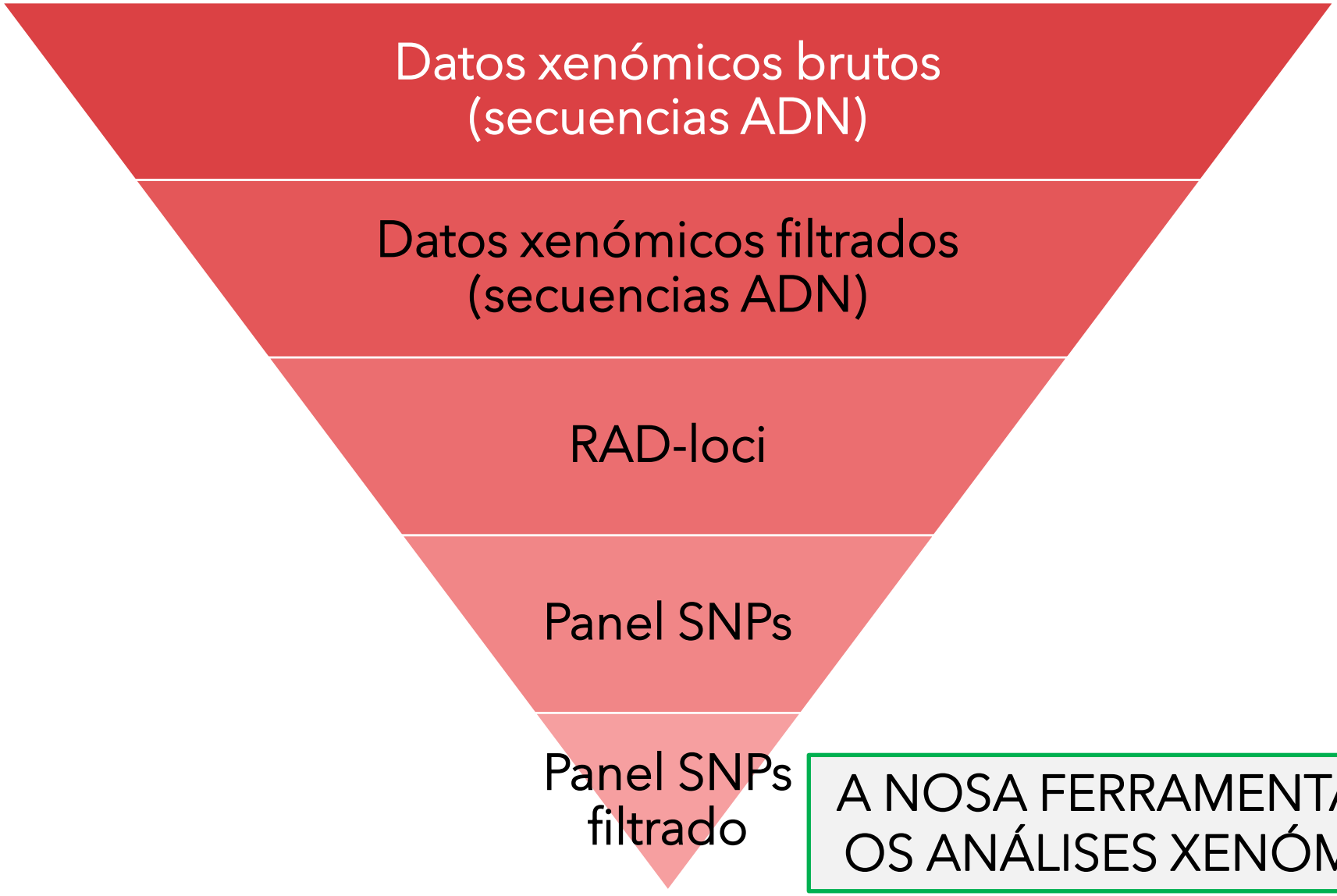
- ✓ Hardware
- ✓ Soporte técnico/mantemento
- ✓ Software



30th Anniversary CESGA
GALICIA SUPERCOMPUTING CENTER

UN PASEO XENÓMICO

Abrochade os cintos!



300 GB

250 GB

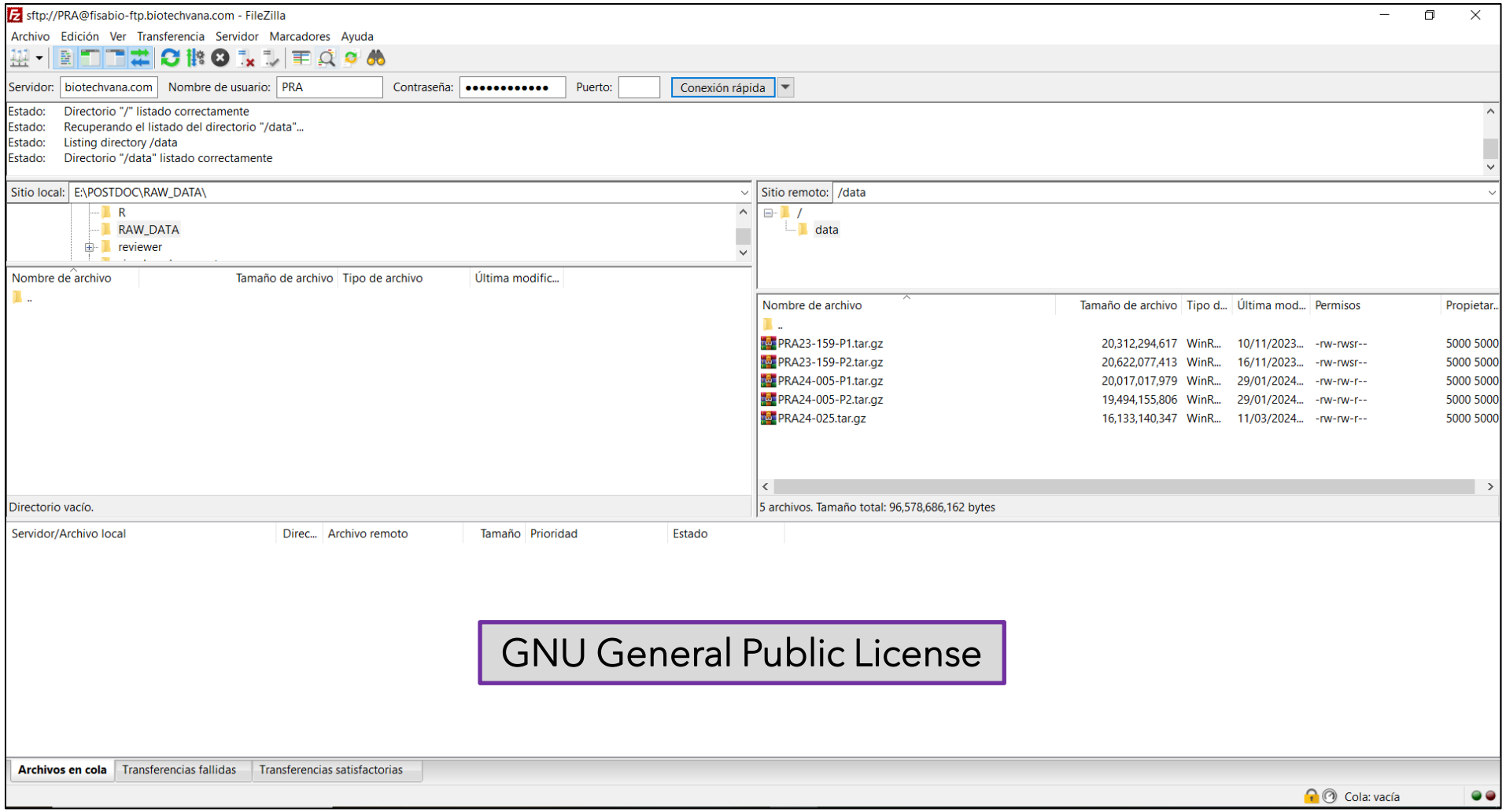
50 GB

1.5 GB

0.5 GB

A NOSA FERRAMENTA PARA OS ANÁLISES XENÓMICOS!

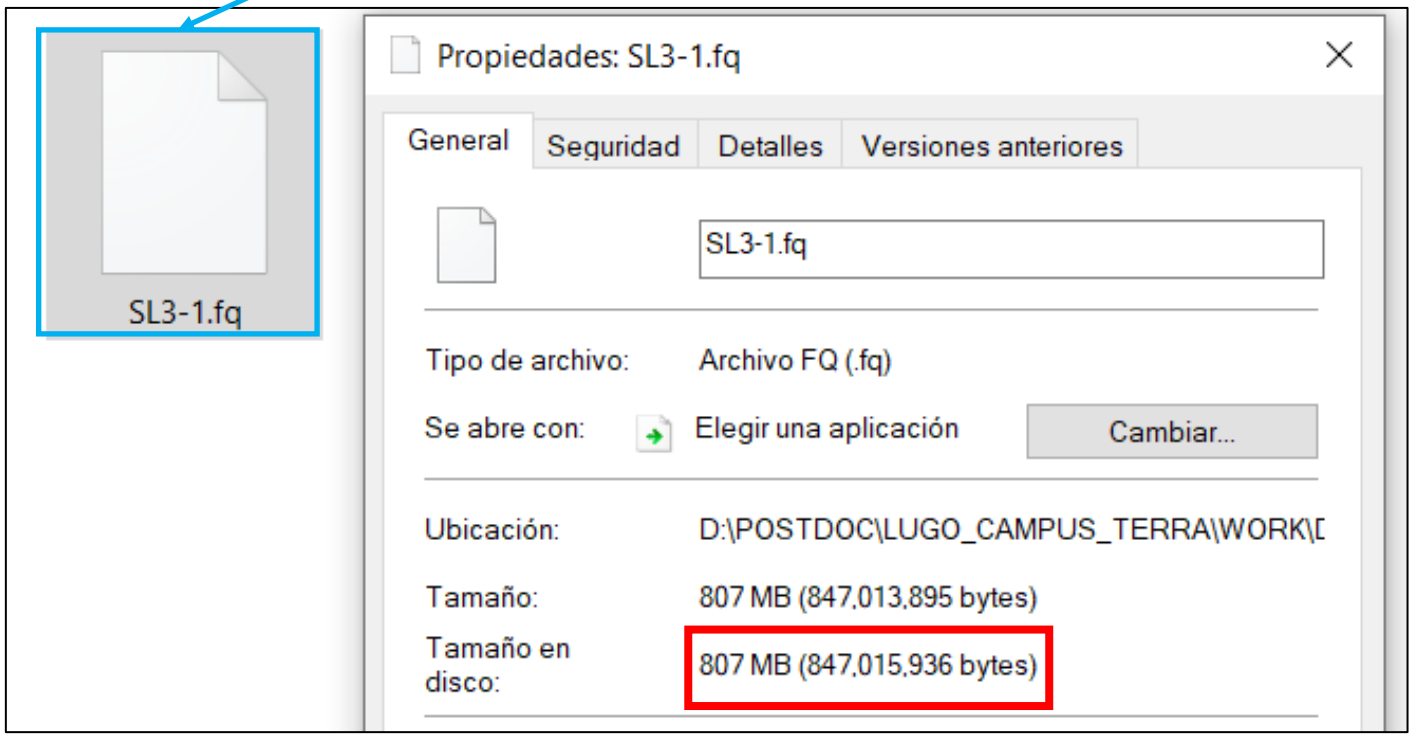
DESCARGA DOS DATOS BRUTOS



GNU General Public License

DATOS BRUTOS

UN ARQUIVO FASTQ = UNHA MOSTRA



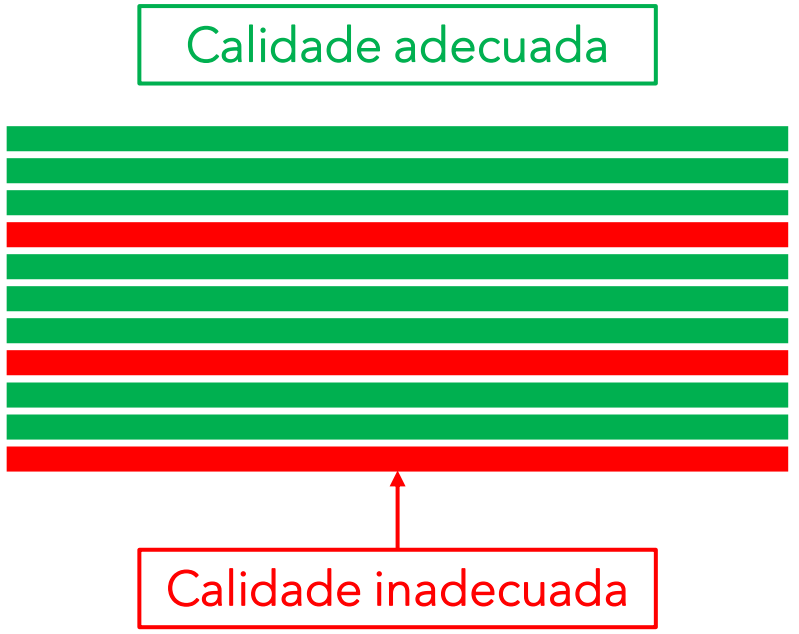
A xenómica é unha ciencia de Big Data!!

DATOS BRUTOS

```
@NB501698:26:H7FK7BGX3:1:11101:26297:1053 1:N:0:GATGAAT
GCTAGNTTGTTAGCACCGGCCTGCTAACTGTCTGAA
+
A6AAA#EEEEEEEE/EEEEEEAEEEEEEEEEEEEA/
@NB501698:26:H7FK7BGX3:1:11101:23614:1053 1:N:0:GATGAAT
GATGTNTCCAGGGCAATAAATTGCAATGTCCGTACA
+
AAAAA#EEEEEEEEAE<EEEEEEEEEEEEEEAEAEE
@NB501698:26:H7FK7BGX3:1:11101:17059:1054 1:N:0:GATGAAT
GGGGANGTACATGCAGTTTGCTGCTATGTCTGAGGG
+
AAAAA#EEEEEEEEEEEEEEEEEEEEEEEEEEEEE
@NB501698:26:H7FK7BGX3:1:11101:19501:1058 1:N:0:GATGAAT
TACGTNCATGCAGCATCTGCATGCCAACCATTTGAG
+
6AAAA#EEEEEEEE<EEEEEEEEEE6EEEEEEEEE
@NB501698:26:H7FK7BGX3:1:11101:23832:1064 1:N:0:GATGAAT
GTCTCTACGGCAGCAGAATTGTGCAGCCAGCAGCTG
+
AAAAAEEEEEEEEEEEEEEEEEEEEEEEEEEEEEE
@NB501698:26:H7FK7BGX3:1:11101:17066:1066 1:N:0:GATGAAT
AATGAAAACAACGCAGCCTTCTGCCTCTAGAGGACG
+
AAAAAEEEEEEEEE6EEEEAEEEEEAEEEEEEEAEE
@NB501698:26:H7FK7BGX3:1:11101:4774:1067 1:N:0:GATGAAT
TCCCTACGTAGTGCACTACGGTGCTTCTGACCAGGT
+
AAAAAEEEEEEEEEEEEEEEEEEEEEEAEAEEEEEE
@NB501698:26:H7FK7BGX3:1:11101:1726:1067 1:N:0:GATGAAT
AATGAAAGAAAAGCAGACAGATGCACTTCTAGCTCT
+
/AAAAEEEEEEEEEEEEEEEEE/EEEEEEAEAEE
@NB501698:26:H7FK7BGX3:1:11101:8689:1068 1:N:0:GATGAAT
```

O **FASTQ** é un formato de texto plano que serve para almacenar secuencias biolóxicas e a súa calidade de secuenciación.

Para cada lectura (*read*; fragmento de ADN secuenciado) corresponden CATRO liñas.



AVALIACIÓN CALIDADE DATOS BRUTOS

FastQC Report

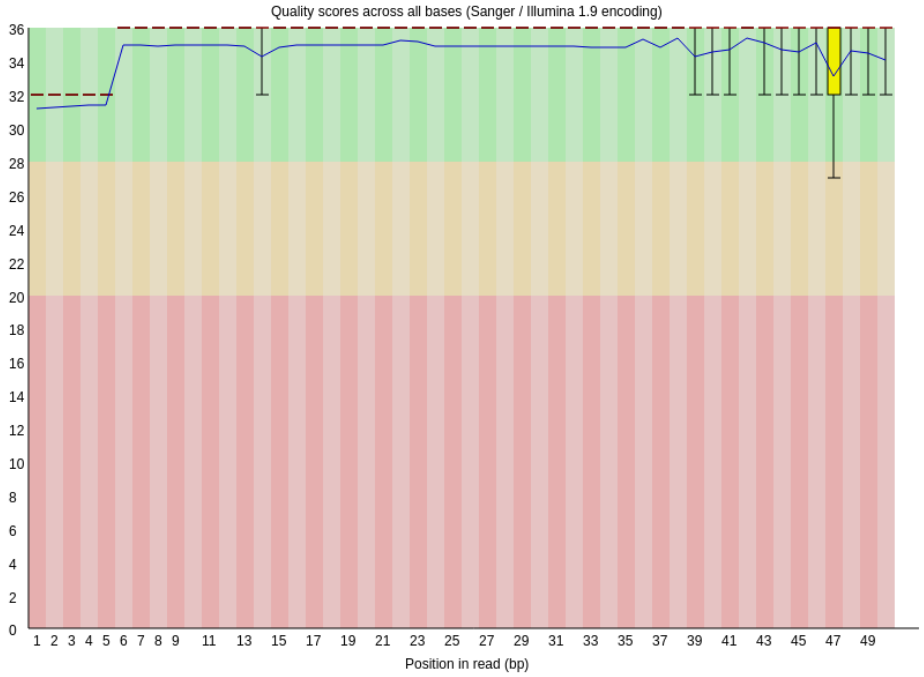
Summary

- Basic Statistics
- Per base sequence quality
- Per tile sequence quality
- Per sequence quality scores
- Per base sequence content
- Per sequence GC content
- Per base N content
- Sequence Length Distribution
- Sequence Duplication Levels
- Overrepresented sequences
- Adapter Content

Basic Statistics

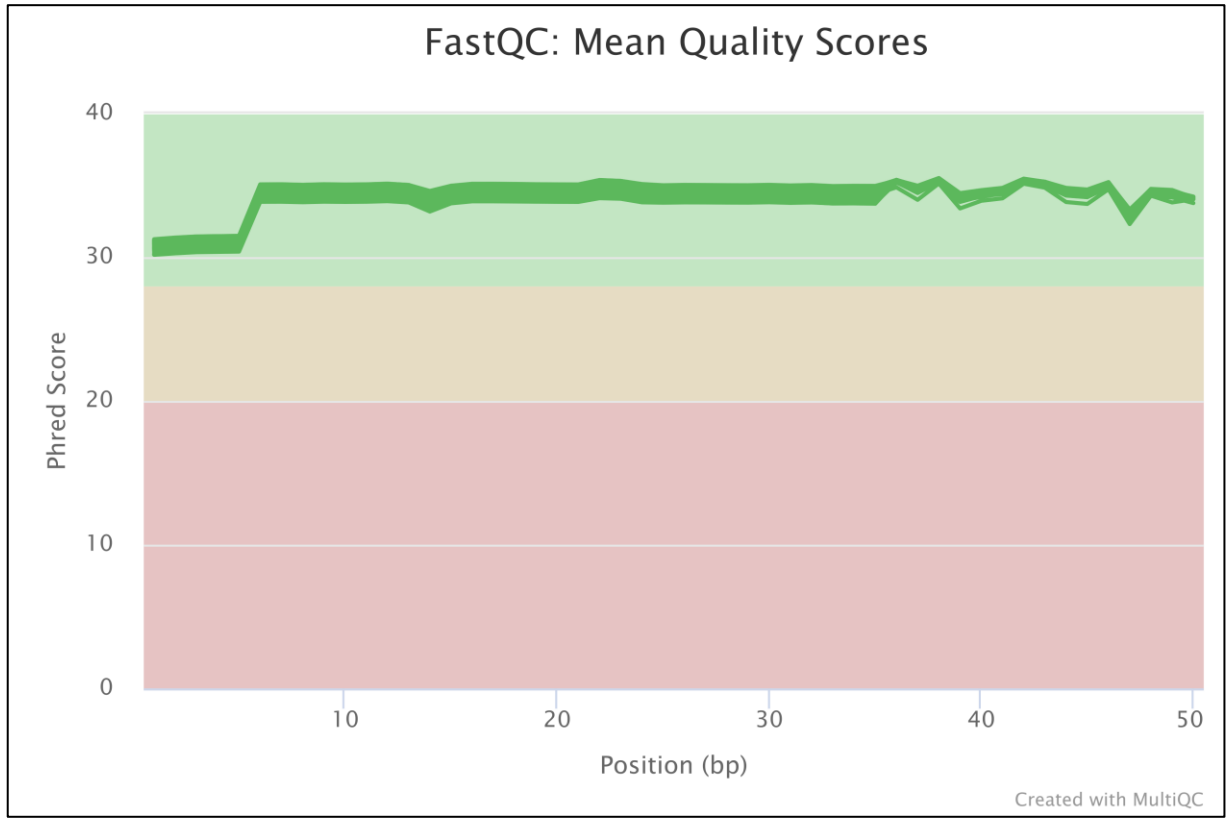
Measure	Value
Filename	DG_C1_01_S49_R1_001.fastq.gz
File type	Conventional base calls
Encoding	Sanger / Illumina 1.9
Total Sequences	3667274
Total Bases	140.4 Mbp
Sequences flagged as poor quality	0
Sequence length	35-50
%GC	49

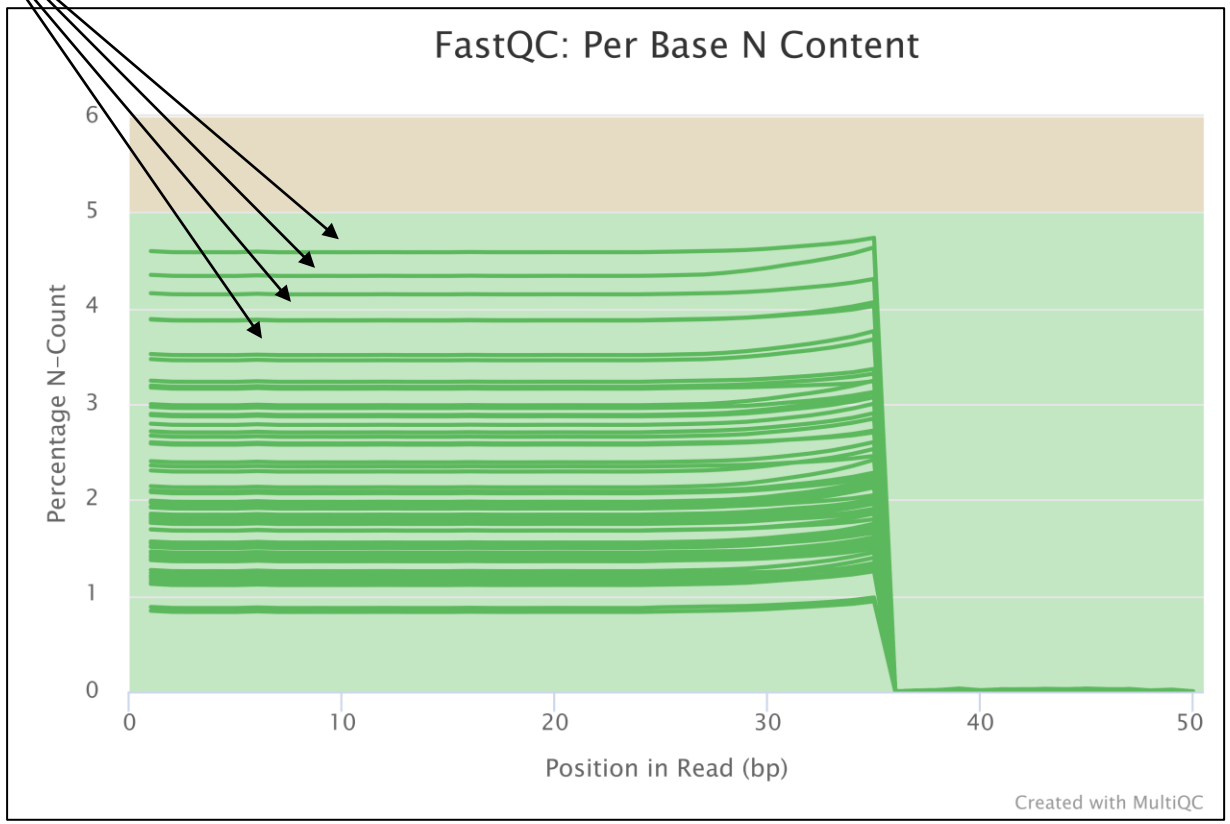
Per base sequence quality



AVALIACIÓN CALIDADE DATOS BRUTOS

Diferentes mostras





Trimmomatic: A flexible read trimming tool for Illumina NGS data

Citations

Bolger, A. M., Lohse, M., & Usadel, B. (2014). Trimmomatic: A flexible trimmer for Illumina Sequence Data. *Bioinformatics*, btu170.

FILTRAXE POR CALIDADE DAS SECUENCIAS (*READS*) ADN

Trimmomatic / LICENSE

BjoernUsadel Create LICENSE — 6b5cf32 · 3 months ago — History

Code | Blame | 677 lines (556 loc) · 34.6 KB — Raw

```
The license does NOT apply to the sequence of the Illumina sequences contained in the software.
THE ILLUMINA sequences (adapters) etc REMAIN COPYRIGHTED and owned by Illumina and are used in trimmomatic by permission.
The software and sourecode are licensed by the ....
                    GNU GENERAL PUBLIC LICENSE
                       Version 3, 29 June 2007

 Copyright (C) 2007 Free Software Foundation, Inc. <https://fsf.org/>
 Everyone is permitted to copy and distribute verbatim copies
 of this license document, but changing it is not allowed.

                            Preamble

  The GNU General Public License is a free, copyleft license for
software and other kinds of works.

  The licenses for most software and other practical works are designed
to take away your freedom to share and change the works.  By contrast,
```

https://github.com/usadellab/Trimmomatic/blob/main/LICENSE

FILTRAXE POR CALIDADE DAS SECUENCIAS (*READS*) ADN

fastp

A tool designed to provide fast all-in-one preprocessing for FastQ files. This tool is developed in C++ with multithreading supported to afford high performance.

Citation: Shifu Chen. 2023. Ultrafast one-pass FASTQ data preprocessing, quality control, and deduplication using fastp. iMeta 2: e107. https://doi.org/10.1002/imt2.107

MIT license

fastp

A tool designed to provide fast all-in-one preprocessing for FastQ files. This tool is developed in C++ with multithreading supported to afford high performance.

Citation: Shifu Chen. 2023. Ultrafast one-pass FASTQ data preprocessing, quality control, and deduplication using fastp. iMeta 2: e107. https://doi.org/10.1002/imt2.107

FILTRAXE POR CALIDADE DAS SECUENCIAS (*READS*) ADN

RECOMENDACIÓNS BIOINFORMÁTICAS

LER OS MANUAIS E PUBLICACIÓNS!

LER AS LICENCIAS!

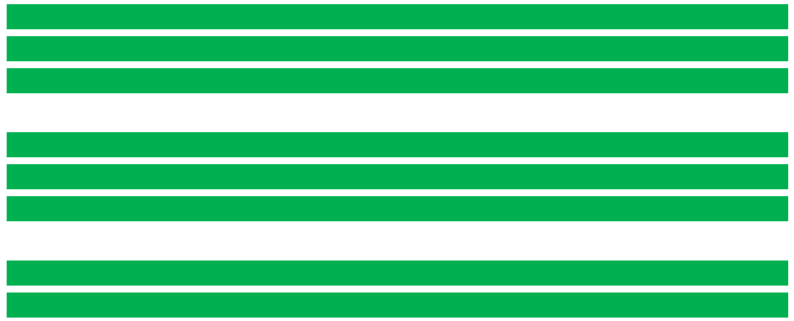
...

```
MIT License

Copyright (c) 2016 OpenGene - Open Source Genetics Toolbox

Permission is hereby granted, free of charge, to any person obtaining a copy
of this software and associated documentation files (the "Software"), to deal
in the Software without restriction, including without limitation the rights
to use, copy, modify, merge, publish, distribute, sublicense, and/or sell
copies of the Software, and to permit persons to whom the Software is
furnished to do so, subject to the following conditions:

The above copyright notice and this permission notice shall be included in all
copies or substantial portions of the Software.

THE SOFTWARE IS PROVIDED "AS IS", WITHOUT WARRANTY OF ANY KIND, EXPRESS OR
IMPLIED, INCLUDING BUT NOT LIMITED TO THE WARRANTIES OF MERCHANTABILITY,
FITNESS FOR A PARTICULAR PURPOSE AND NONINFRINGEMENT. IN NO EVENT SHALL THE
AUTHORS OR COPYRIGHT HOLDERS BE LIABLE FOR ANY CLAIM, DAMAGES OR OTHER
LIABILITY, WHETHER IN AN ACTION OF CONTRACT, TORT OR OTHERWISE, ARISING FROM,
OUT OF OR IN CONNECTION WITH THE SOFTWARE OR THE USE OR OTHER DEALINGS IN THE
SOFTWARE.
```

DATOS FILTRADOS

```
@NB501698:26:H7FK7BGX3:1:11101:26297:1053 1:N:0:GATGAAT
GCTAGNTTGTTAGCACCGGCCTGCTAACTGTCTGAA
+
A6AAA#EEEEEEEE/EEEEEEAEEEEEEEEEEEEA/
@NB501698:26:H7FK7BGX3:1:11101:23614:1053 1:N:0:GATGAAT
GATGTNTCCAGGGCAATAAATTGCAATGTCCGTACA
+
AAAAA#EEEEEEEEAE<EEEEEEEEEEEEEEAEAEE
@NB501698:26:H7FK7BGX3:1:11101:17059:1054 1:N:0:GATGAAT
GGGGANGTACATGCAGTTTGCTGCTATGTCTGAGGG
+
AAAAA#EEEEEEEEEEEEEEEEEEEEEEEEEEEEE
@NB501698:26:H7FK7BGX3:1:11101:19501:1058 1:N:0:GATGAAT
TACGTNCATGCAGCATCTGCATGCCAACCATTTGAG
+
6AAAA#EEEEEEEE<EEEEEEEEEE6EEEEEEEEE
@NB501698:26:H7FK7BGX3:1:11101:23832:1064 1:N:0:GATGAAT
GTCTCTACGGCAGCAGAATTGTGCAGCCAGCAGCTG
+
AAAAAEEEEEEEEEEEEEEEEEEEEEEEEEEEEEE
@NB501698:26:H7FK7BGX3:1:11101:17066:1066 1:N:0:GATGAAT
AATGAAAACAACGCAGCCTTCTGCCTCTAGAGGACG
+
AAAAAEEEEEEEEE6EEEEAEEEEEAEEEEEEEAEE
@NB501698:26:H7FK7BGX3:1:11101:4774:1067 1:N:0:GATGAAT
TCCCTACGTAGTGCACTACGGTGCTTCTGACCAGGT
+
AAAAAEEEEEEEEEEEEEEEEEEEEEEAEAEEEEEE
@NB501698:26:H7FK7BGX3:1:11101:1726:1067 1:N:0:GATGAAT
AATGAAAGAAAAGCAGACAGATGCACTTCTAGCTCT
+
/AAAAEEEEEEEEEEEEEEEEE/EEEEEEAEAEE
@NB501698:26:H7FK7BGX3:1:11101:8689:1068 1:N:0:GATGAAT
```

Calidade adecuada



Calidade inadecuada

ALIÑAMENTO DAS SECUENCIAS (SE HAI XENOMA DE REFERENCIA DISPOÑIBLE)

Burrows-Wheeler Aligner

Introduction

BWA is a software package for mapping low-divergent sequences against a large reference genome, such as the human genome. It consists of three algorithms: BWA-backtrack, BWA-SW and BWA-MEM. The first algorithm is designed for Illumina sequence reads up to 100bp, while the rest two for longer sequences ranged from 70bp to 1Mbp. BWA-MEM and BWA-SW share similar features such as long-read support and split alignment, but BWA-MEM, which is the latest, is generally recommended for high-quality queries as it is faster and more accurate. BWA-MEM also has better performance than BWA-backtrack for 70-100bp Illumina reads.

https://bio-bwa.sourceforge.net/

GPL-3.0 license

Bowtie

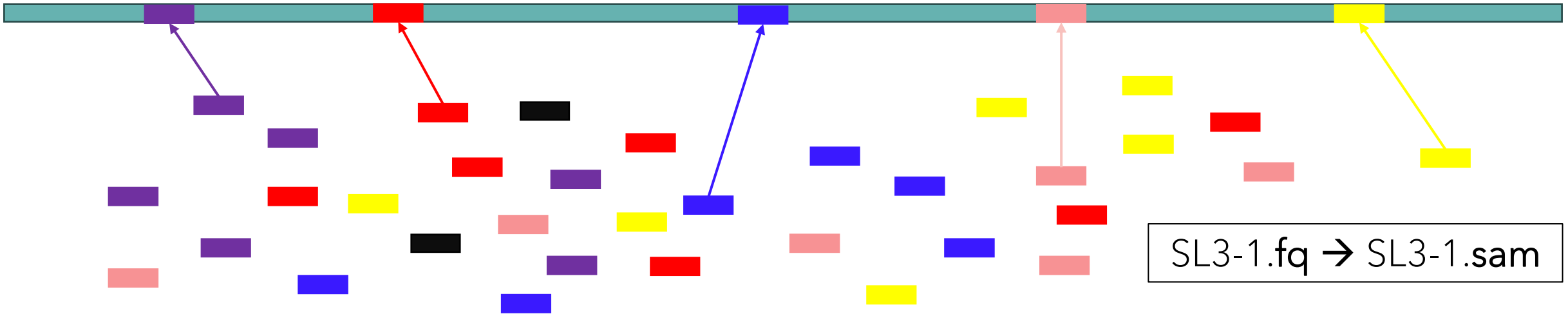
An ultrafast memory-efficient short read aligner

Bowtie is an ultrafast, memory-efficient short read aligner. It aligns short DNA sequences (reads) to the human genome at a rate of over 25 million 35-bp reads per hour. Bowtie indexes the genome with a Burrows-Wheeler index to keep its memory footprint small: typically about 2.2 GB for the human genome (2.9 GB for paired-end).

https://bowtie-bio.sourceforge.net/index.shtml

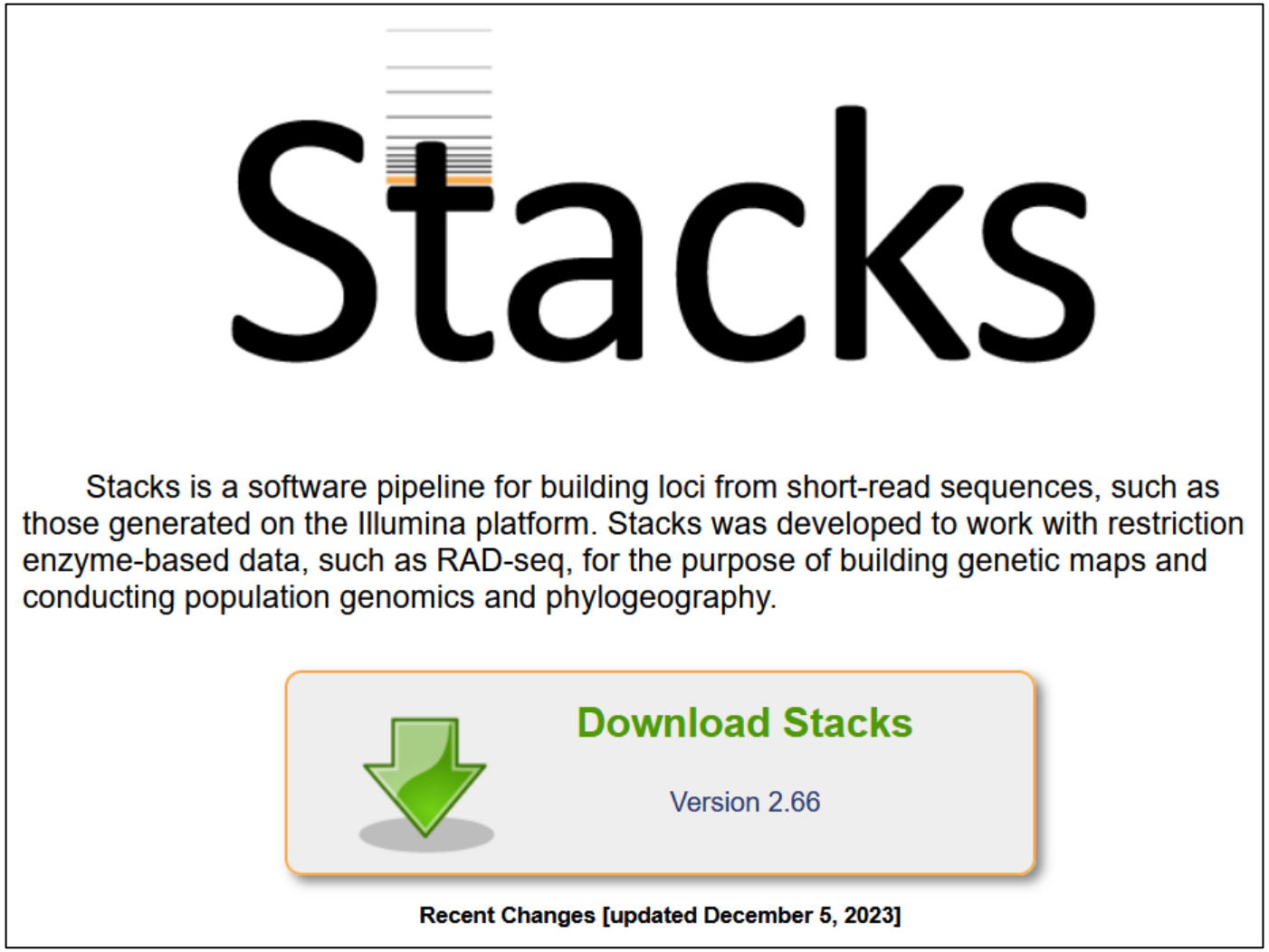
Artistic licence 2.0

Xenoma de referencia



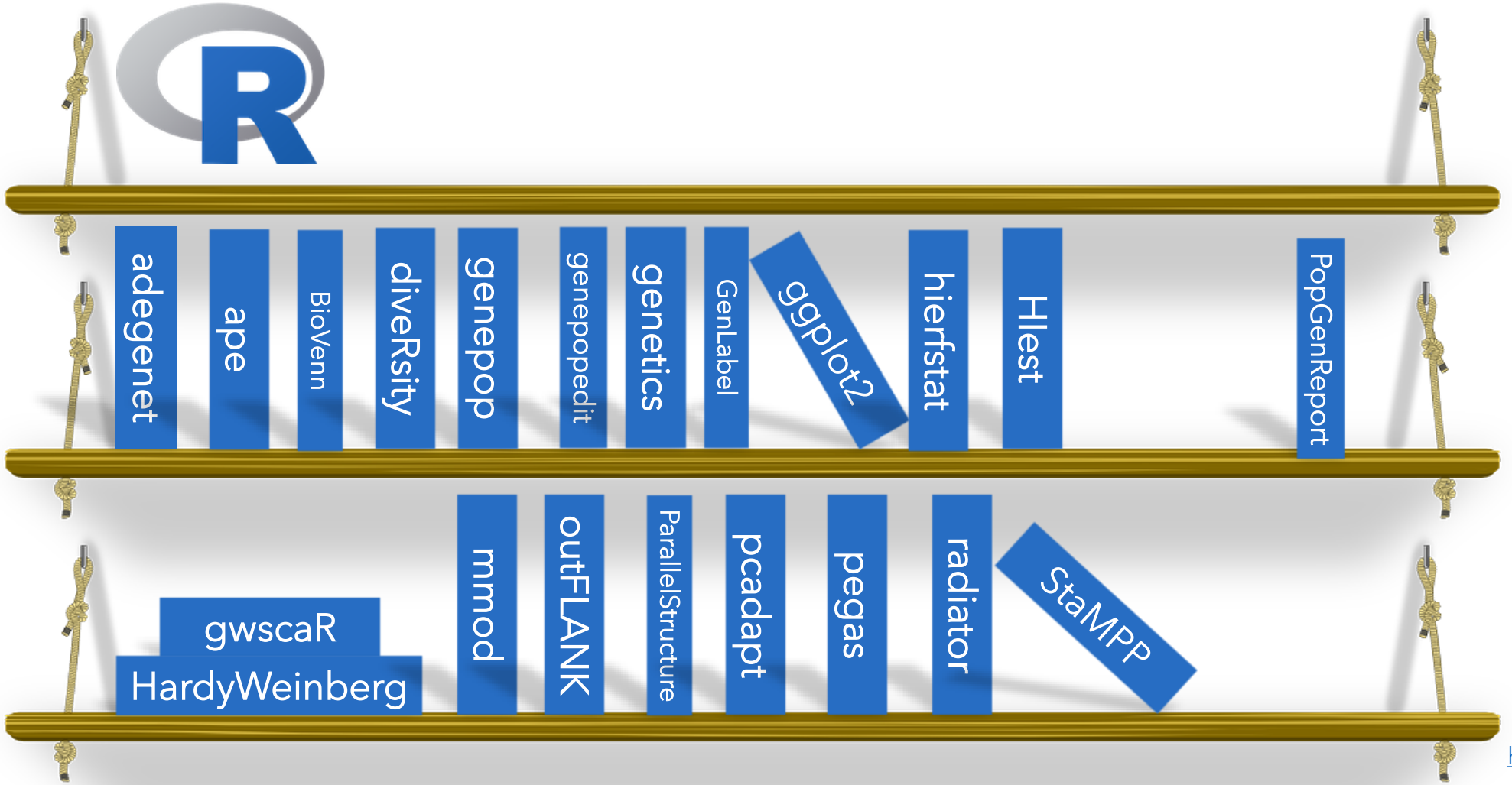
SL3-1.fq → SL3-1.sam

OBTENCIÓN RAD-LOCI E XENOTIPADO DE SNPS



https://catchenlab.life.illinois.edu/stacks/

FILTRADO DE SNPs | PANEL DEFINITIVO DE SNPs | ANÁLISES XENÓMICAS



VCFtools

A set of tools written in Perl and C++ for working with VCF files.

https://vcftools.github.io/

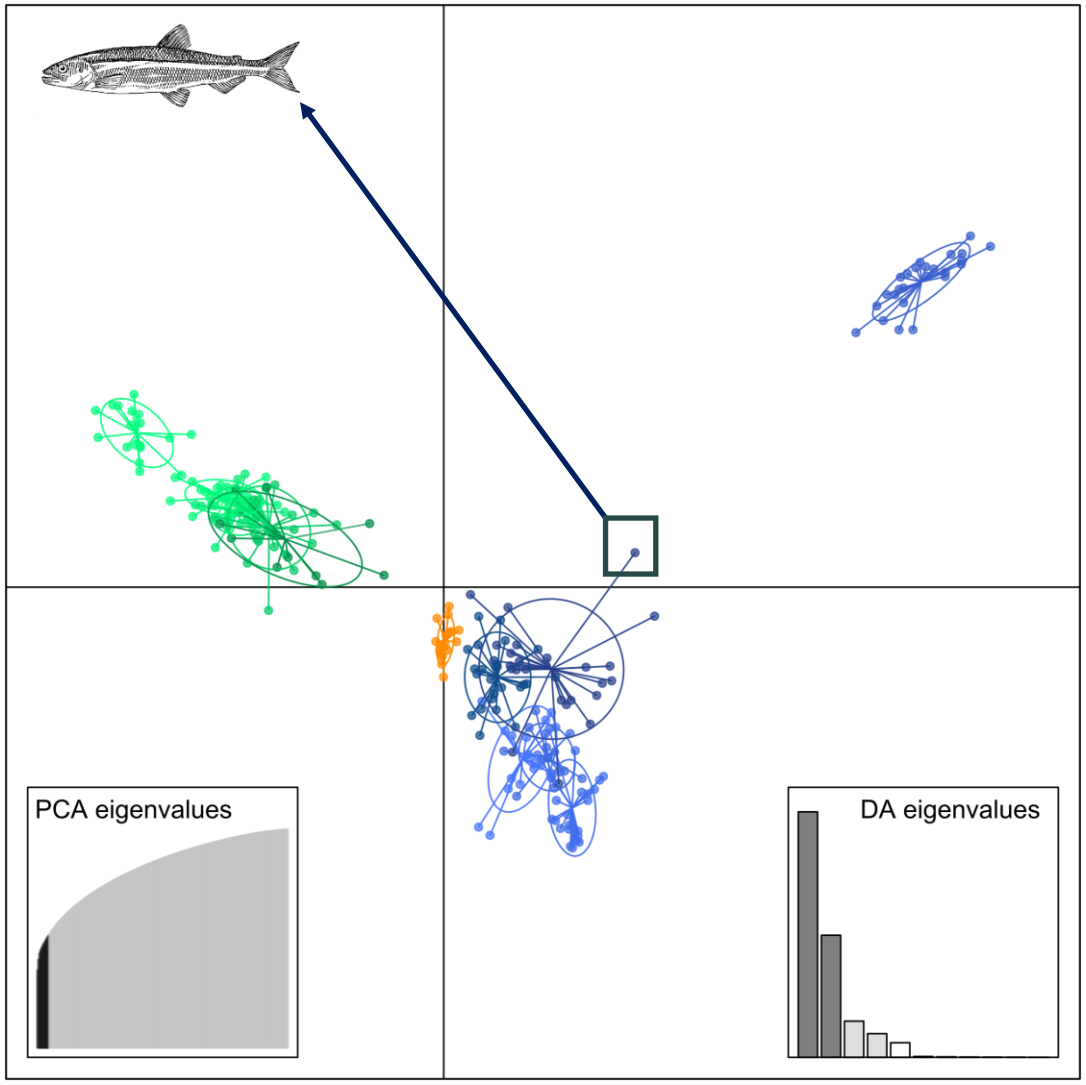
README GPL-3.0 license GPL-3.0 license

BayeScan

https://github.com/mfoll/BayeScan

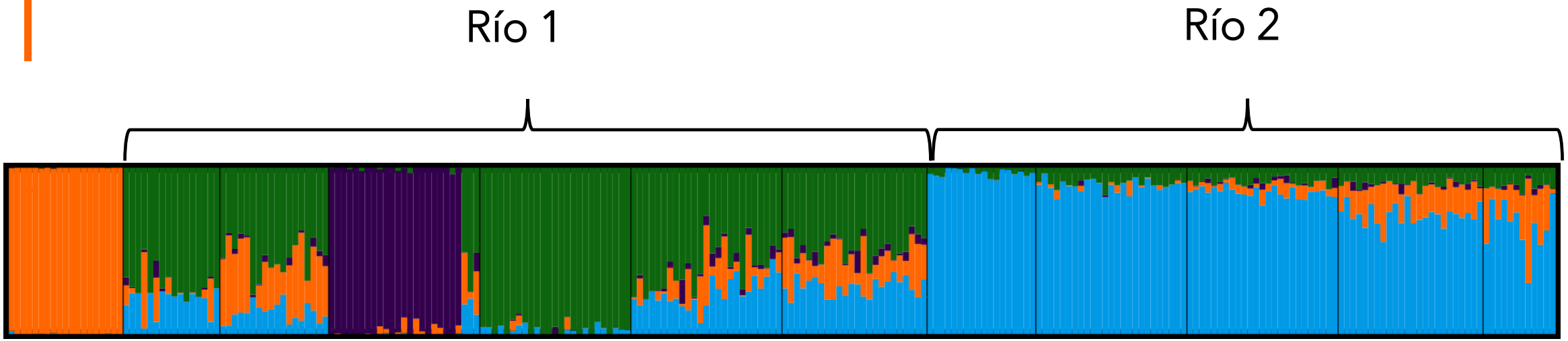
ALGÚNS RESULTADOS XENÓMICOS

Con softwaRe libRe e gratuíto

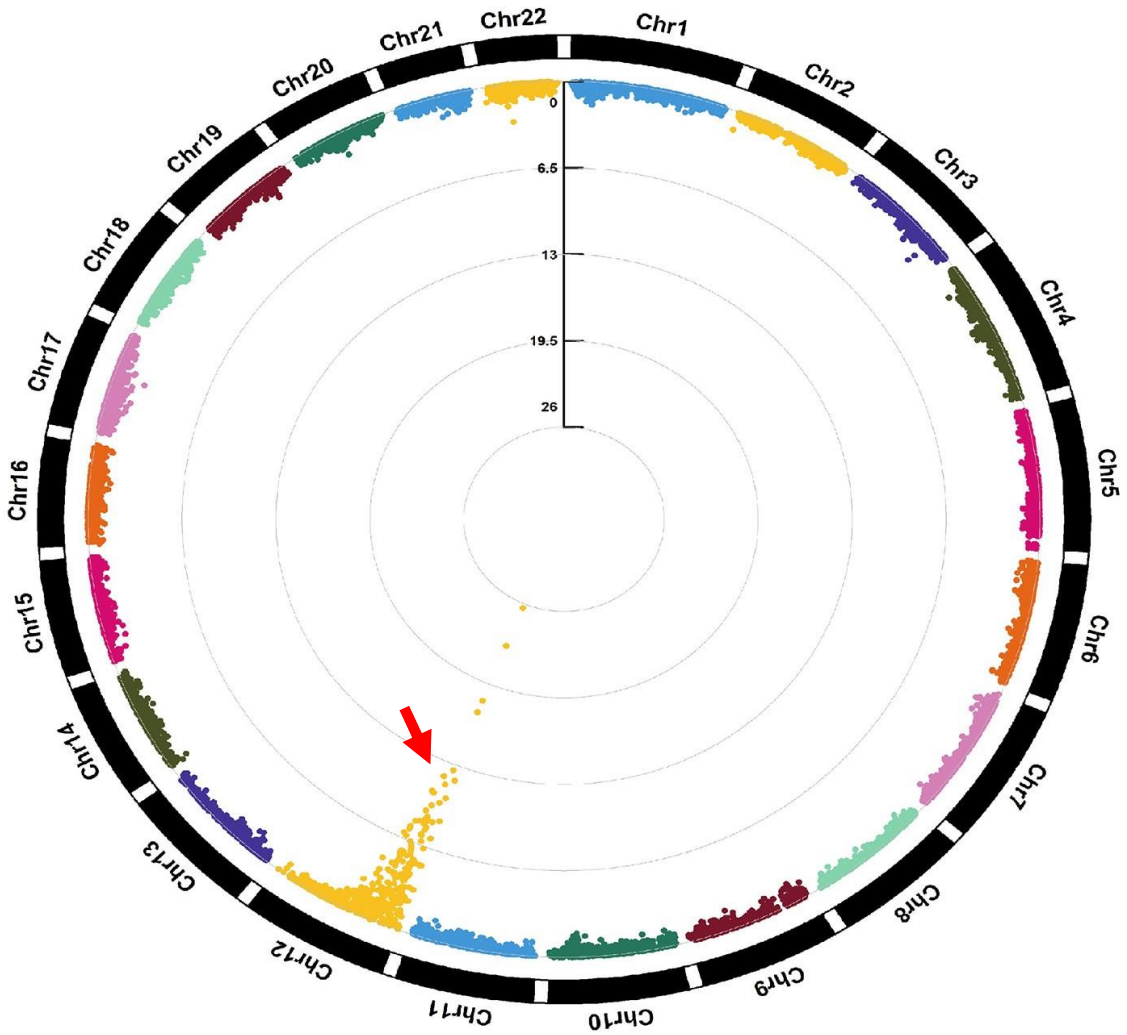


- É fundamental coñecer a estrutura das poboacións naturais para xestionar a súa conservación.
- Troita común (*Salmo trutta*)
- Paquete de R: {**adegenet**}

PISCIFACTORÍA



- É fundamental coñecer a estrutura das poboacións naturais para xestionar a súa conservación.
- Troita común (*Salmo trutta*)
- Paquete de R: {**ParallelStructure**} | Análisis *multi-core* no CESGA



- SNPs significativamente asociados ao sexo do rodaballo. Sistema ZZ/ZW
- Xene candidato determinación sexual: *sox2*
- Rodaballo (*Scophthalmus maximus*)
- Paquete de R: {**GenABEL**}

Martínez, P., Robledo, D., Taboada, X., Blanco, A., Moser, M., Maroso, F., Hermida, M., Gómez-Tato, A., Álvarez-Blázquez, B., Cabaleiro, S., Piferrer, F., Bouza, C., Lien, S., & Viñas, A. M. A genome-wide association study, supported by a new chromosome-level genome assembly, suggests *sox2* as a main driver of the undifferentiated ZZ/ZW sex determination of turbot (*Scophthalmus maximus*). *Genomics*, 113(4), 1705–1718 (2021). https://doi.org/https://doi.org/10.1016/j.ygeno.2021.04.007

O melloR traballo, en equipo!
Con softwaRe libre e gratuíto