TALCA UNIVERSIDAD CHILE
Doctorado en Ciencias
mención Modelado de Sistemas Químicos y Biológicos
Facultad de Ingeniería

Genes Genomas y Proteomas

# COMPLETE SEQUENCING OF APE GENOMES

Gabriela Urra Gajardo

TALCA UNIVERSIDAD CHILE

Doctorado en Ciencias
mención Modelado de Sistemas Químicos y Biológicos
Facultad de Ingeniería

Article

# Complete sequencing of ape genomes

https://doi.org/10.1038/s41586-025-08816-3

Received: 31 July 2024

Accepted: 19 February 2025

Published online: 9 April 2025

Open access

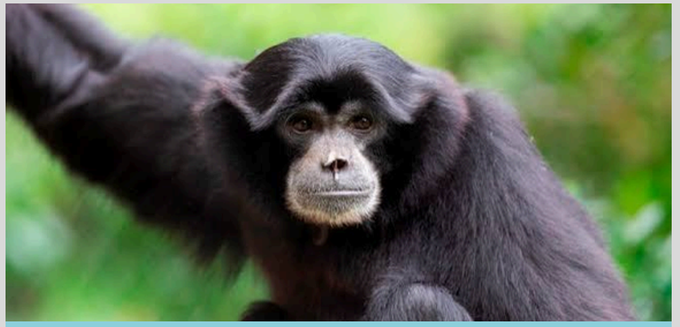
Check for updates

DongAhn Yoo[1], Arang Rhie[2], Prajna Hebbar[3], Francesca Antonacci[4], Glennis A. Logsdon[1,5], Steven J. Solar[2], Dmitry Antipov[2], Brandon D. Pickett[2], Yana Safonova[6], Francesco Montinaro[4,7], Yanting Luo[8], Joanna Malukiewicz[9,10], Jessica M. Storer[11], Jiadong Lin[1], Abigail N. Sequeira[12], Riley J. Mangan[13,14,15], Glenn Hickey[3], Graciela Monfort Anez[16], Parithi Balachandran[17], Anton Bankevich[6], Christine R. Beck[11,17,18], Arjun Biddanda[19], Matthew Borchers[16], Gerard G. Bouffard[20], Emry Brannan[21], Shelise Y. Brooks[20], Lucia Carbone[22,23], Laura Carrel[24], Agnes P. Chan[25], Juyun Crawford[20], Mark Diekhans[3], Eric Engelbrecht[26], Cedric Feschotte[27], Giulio Formenti[28], Gage H. Garcia[1], Luciana de Gennaro[4], David Gilbert[29], Richard E. Green[30], Andrea Guarracino[31], Ishaan Gupta[32], Diana Haddad[33], Junmin Han[34], Robert S. Harris[12], Gabrielle A. Hartley[11], William T. Harvey[1], Michael Hiller[35,36,37], Kendra Hoekzema[1], Marlys L. Houck[38], Hyeonsoo Jeong[1], Kaivan Kamali[12], Manolis Kellis[13,14], Bryce Kille[39], Chul Lee[40], Youngho Lee[41], William Lees[26,42], Alexandra P. Lewis[1], Qiuhui Li[43], Mark Loftus[44,45], Yong Hwee Eddie Loh[46], Hailey Loucks[3], Jian Ma[47], Yafei Mao[34,48,49], Juan F. I. Martinez[6], Patrick Masterson[33], Rajiv C. McCoy[19], Barbara McGrath[12], Sean McKinney[16], Britta S. Meyer[9], Karen H. Miga[3], Saswat K. Mohanty[12], Katherine M. Munson[1], Karol Pal[12], Matt Pennell[50], Pavel A. Pevzner[32], David Porubsky[1], Tamara Potapova[16], Francisca R. Ringeling[51], Joana L. Rocha[52], Oliver A. Ryder[38], Samuel Sacco[30], Swati Saha[26], Takayo Sasaki[29], Michael C. Schatz[43], Nicholas J. Schork[25], Cole Shanks[3], Linnéa Smeds[12], Dongmin R. Son[53], Cynthia Steiner[38], Alexander P. Sweeten[2], Michael G. Tassia[19], Françoise Thibaud-Nissen[33], Edmundo Torres-González[12], Mihir Trivedi[1], Wenjie Wei[54,55], Julie Wertz[1], Muyu Yang[47], Panpan Zhang[27], Shilong Zhang[34], Yang Zhang[47], Zhenmiao Zhang[32], Sarah A. Zhao[13], Yixin Zhu[50], Erich D. Jarvis[40,56], Jennifer L. Gerton[16], Iker Rivas-González[57], Benedict Paten[3], Zachary A. Szpiech[12], Christian D. Huber[12], Tobias L. Lenz[9], Miriam K. Konkel[44,45], Soojin V. Yi[53,58], Stefan Canzar[51], Corey T. Watson[26], Peter H. Sudmant[52,59], Erin Molloy[60], Erik Garrison[31], Craig B. Lowe[8], Mario Ventura[4], Rachel J. O'Neill[11,18,21], Sergey Koren[2], Kateryna D. Makova[12]✉, Adam M. Phillippy[2]✉ & Evan E. Eichler[1,56]✉

# BACKGROUND CONTEXT

This research focuses on the genomes of humans' closest living relatives, with the aim of **achieving their complete structural resolution** through the use of long-read sequencing technologies.

Previous sequencing technologies were **unable** to accurately assemble the **highly repetitive and structurally complex regions** of great ape genomes, resulting in incomplete genome assemblies and limiting comparative genomic analyses and our understanding of genome evolution.



*Symphalangus syndactylus*

*Pongo pygmaeus*

*Pongo abelii*

*Pan troglodytes*

*Pan paniscus*

*Gorilla gorilla*

# CURRENT KNOWLEDGE

1. Over the past two decades, great ape genome assemblies have progressively improved, culminating in the **first complete telomere-to-telomere** (T2T) human genome and complete ape sex chromosome assemblies.

2. Despite these advances, current great ape reference genomes remained incomplete, **lacking accurate resolution of highly dynamic genomic regions** and fully phased diploid assemblies.

3. The study proposes that combining long-read sequencing technologies (HiFi and ONT) with advanced assembly algorithms **enables the complete resolution** of great ape genomes, revealing structural and evolutionary variation previously missed by human reference-based approaches.

4. This aims to **generate haplotype-resolved diploid reference genomes** for six ape species and provide a comprehensive resource for future evolutionary studies.

Phase 1: Sample Selection

# METHODOLOGICAL WORKFLOW

High-quality DNA was obtained from fibroblast and lymphoblastoid cell lines of six ape species. Male individuals were preferentially selected to enable Y chromosome assembly, and parent–offspring trios were used to improve haplotype phasing.

**Table 1 | Summary of ape genome assemblies**

| Sample information | | | | Assembly stats (v.2.0) Haplotype 1 (haplotype 2) or maternal (paternal) | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| Common name (abbreviation) | Scientific name | Tissue | Sex | Accession | Total no. of bases (Gb) | Contig N50 (Mb) | No. of T2T contigs | No. of non-rDNA gaps or missing telomeres | QV |
| Chimpanzee (PTR) | *Pan troglodytes* | Lymphoblastoid | Male | PRJNA916736 | 3.14 (3.03) | 146.29 (140.84) | 19 (17) | 0 (2) | 66.0 |
| Bonobo (PPA)[a] | *Pan paniscus* | Fibroblast or lymphoblastoid | Male | PRJNA942951 | 3.21 (3.07) | 147.03 (147.48) | 20 (19) | 0 (0) | 62.7 |
| Gorilla (GGO)[a] | *Gorilla gorilla* | Fibroblast | Male | PRJNA942267 | 3.55 (3.35) | 151.43 (150.80) | 19 (22) | 4 (0) | 61.7 |
| Bornean orangutan (PPY) | *Pongo pygmaeus* | Fibroblast | Male | PRJNA916742 | 3.16 (3.05) | 140.59 (137.91) | 15 (13) | 1 (2) | 65.8 |
| Sumatran orangutan (PAB) | *Pongo abelii* | Fibroblast | Male | PRJNA916743 | 3.17 (3.08) | 146.20 (140.60) | 14 (13) | 2 (3) | 63.3 |
| Siamang (SSY) | *Symphalangus syndactylus* | Lymphoblastoid | Male | PRJNA916729 | 3.24 (3.12) | 146.71 (144.67) | 24 (20) | 0 (5) | 66.4 |
| Average | | | | – | 3.24 (3.12) | 146.38 (143.72) | 18.5 (17.3) | 1.2 (2) | 64.3 |

[a]Sample with parental sequence data. Contig N50 is the shortest contiguous assembled fragment length that contains 50% of the genome (values approximate the length of chromosomes). QV represents the accuracy score from the Illumina–HiFi hybrid-based approach. Numbers in parentheses denote the values for haplotype 2 or the paternal haplotype.

# METHODOLOGICAL WORKFLOW

Three complementary datasets were generated: **PacBio HiFi, Oxford Nanopore** (ONT) and **Hi-C or parent–offspring trio** data. By integrating these complementary technologies, the sequencing strategy maximized assembly accuracy, continuity, and haplotype resolution, providing the **foundation for complete, haplotype-resolved diploid genome assemblies**.

## PacBio HiFi

Generated highly accurate long reads (≈90× coverage), providing the backbone of the genome assembly.

## Oxford Nanopore (ONT)

Produced ultra-long reads (≈136× coverage, including >100 kb reads) capable of spanning highly repetitive regions such as centromeres and segmental duplications.

## Hi-C / parent–offspring trio

Used for chromosome-scale haplotype phasing. These approaches distinguished maternal from paternal haplotypes (or H1/H2 in PTR, PPY, PAB, SSY), enabling the reconstruction of complete diploid genomes.

# METHODOLOGICAL WORKFLOW

Compares syntenic regions of **human chromosomes 7 and 16** across ape genomes with Verkko and SVbyEye.

**Figure 1** validates the quality of the new assemblies, revealing previously unresolved chromosomal complexity and improving our understanding of primate genome evolution.

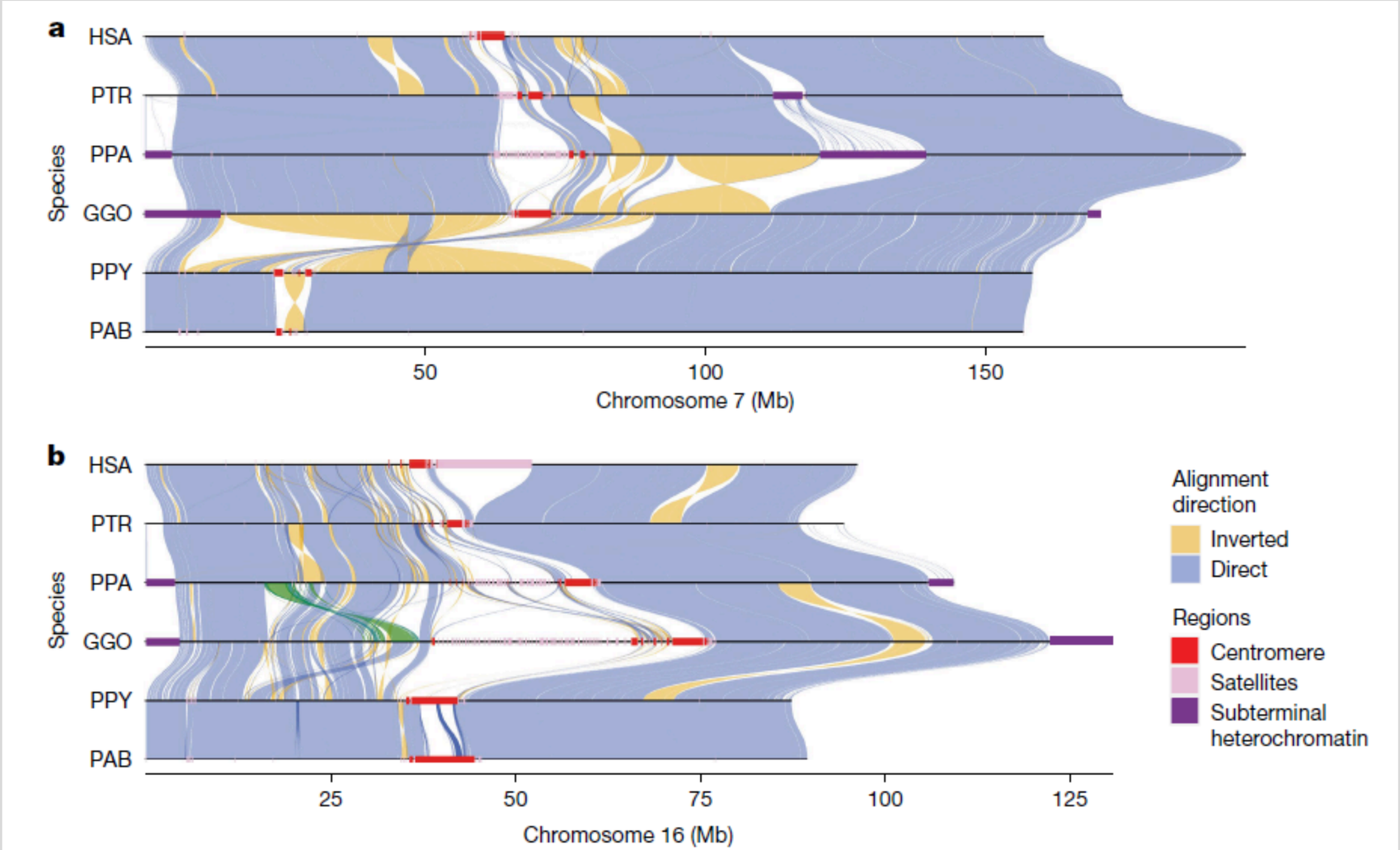



**Fig. 1 | Chromosomal-level assembly of complete genomes for great apes.**

# METHODOLOGICAL WORKFLOW

**Importance of comparing HSA7 and HSA16:**

- Demonstrates the ability of T2T assemblies to accurately resolve ape chromosome architecture.
- Reveals conserved syntenic regions and structural changes such as inversions and repetitive regions.

- Identifies a gorilla-specific 4.8 Mb inverted transposition in HSA16, previously misclassified as a simple inversion.
- Enables analysis of centromeres, satellite expansions, and heterochromatin regions.
- **Allows direct comparison of ape genomes without relying solely on the human genome as a reference.**

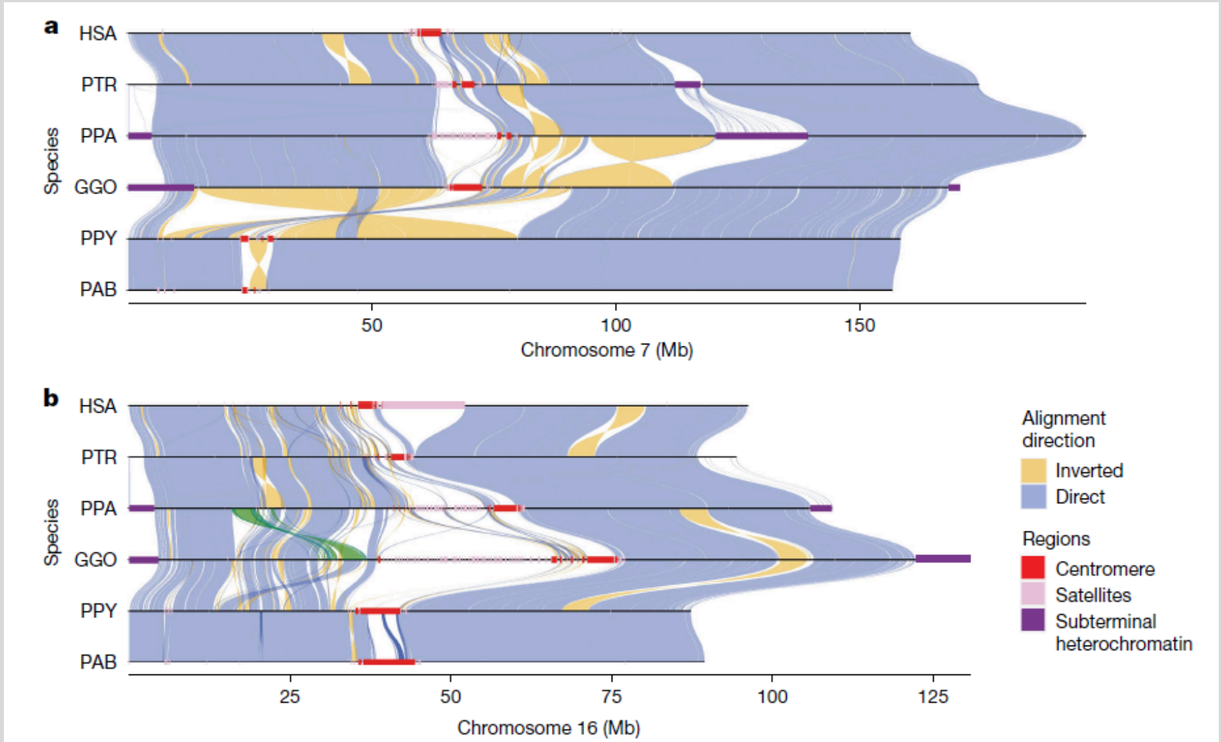



**Fig. 1 | Chromosomal-level assembly of complete genomes for great apes.**

# METHODOLOGICAL WORKFLOW

- The final genome assembly was subjected to comprehensive quality assessment using Merqury to evaluate nucleotide accuracy, completeness, and potential assembly errors, through the **Quality Value (QV) metric**.
- The resulting assembly achieved a QV above 60, corresponding to **fewer than one error per million bases**, supporting the high accuracy of the reconstruction.
- Additionally, 74% of chromosomes were assembled in a gapless telomere-to-telomere (T2T) state, demonstrating the **high continuity and completeness** achieved by the assembly strategy.

**80.8%** of chromosomes have at least one perfect end-to-end copy

**Table 1 | Summary of ape genome assemblies**

| Sample information | | | | Assembly stats (v.2.0) Haplotype 1 (haplotype 2) or maternal (paternal) | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| Common name (abbreviation) | Scientific name | Tissue | Sex | Accession | Total no. of bases (Gb) | Contig N50 (Mb) | No. of T2T contigs | No. of non-rDNA gaps or missing telomeres | QV |
| Chimpanzee (PTR) | *Pan troglodytes* | Lymphoblastoid | Male | PRJNA916736 | 3.14 (3.03) | 146.29 (140.84) | 19 (17) | 0 (2) | 66.0 |
| Bonobo (PPA)[a] | *Pan paniscus* | Fibroblast or lymphoblastoid | Male | PRJNA942951 | 3.21 (3.07) | 147.03 (147.48) | 20 (19) | 0 (0) | 62.7 |
| Gorilla (GGO)[a] | *Gorilla gorilla* | Fibroblast | Male | PRJNA942267 | 3.55 (3.35) | 151.43 (150.80) | 19 (22) | 4 (0) | 61.7 |
| Bornean orangutan (PPY) | *Pongo pygmaeus* | Fibroblast | Male | PRJNA916742 | 3.16 (3.05) | 140.59 (137.91) | 15 (13) | 1 (2) | 65.8 |
| Sumatran orangutan (PAB) | *Pongo abelii* | Fibroblast | Male | PRJNA916743 | 3.17 (3.08) | 146.20 (140.60) | 14 (13) | 2 (3) | 63.3 |
| Siamang (SSY) | *Symphalangus syndactylus* | Lymphoblastoid | Male | PRJNA916729 | 3.24 (3.12) | 146.71 (144.67) | 24 (20) | 0 (5) | 66.4 |
| Average | | | | – | 3.24 (3.12) | 146.38 (143.72) | 18.5 (17.3) | 1.2 (2) | 64.3 |

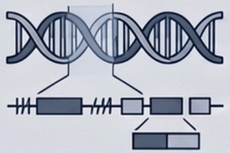
[a]Sample with parental sequence data. Contig N50 is the shortest contiguous assembled fragment length that contains 50% of the genome (values approximate the length of chromosomes). QV represents the accuracy score from the Illumina–HiFi hybrid-based approach. Numbers in parentheses denote the values for haplotype2 or the paternal haplotype.

# METHODOLOGICAL WORKFLOW

## FASE 1: ENSAMBLAJE Y VALIDACIÓN DE MODELOS

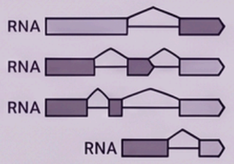
### Genoma Ensamblado (T2T)

Generación de ensamblajes completos y continuos (gapless) mediante herramientas como Verkko y curación manual.



### Predicción Génica Inicial

Identificación de regiones génicas probables basadas en marcas de anotación y modelos *ab initio*.

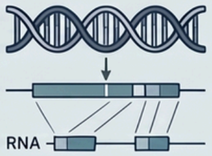



### Secuenciación Iso-Seq de Longitud Completa

Captura de transcritos de RNA completos para identificar isoformas y estructuras de exones sin ambigüedad.

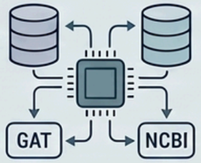
Identified potential coding and non-coding genes. These provided direct evidence of expressed transcripts.

## FASE 2: REFINAMIENTO Y ANOTACIÓN SISTÉMICA



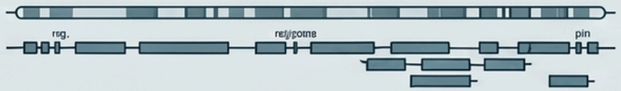

### Validación y Refinamiento de Modelos

Comparación de alineamientos genoma-transcrito para corregir y validar la estructura de los modelos génicos.




### Pipeline de Anotación CAT + NCBI

Uso del Comparative Annotation Toolkit (CAT) y NCBI para integrar anotaciones de referencia y bases de datos.

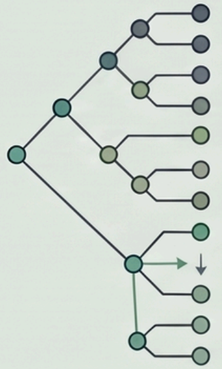


### Anotación Genómica Final

Generación de un mapa genómico definitivo con genes curados y adoptado como referencia en RefSeq.

The refined models were integrated with comparative annotation.

## FASE 3: ANÁLISIS DE ORTOLOGÍA Y EVOLUCIÓN



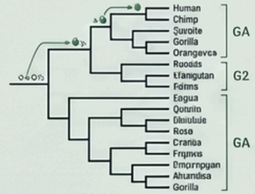
### Análisis de Ortología con TOGA

Inferencia de relaciones ortólogas para detectar la pérdida o ganancia de genes en linajes específicos.

Gene conservatión, gaim expancia o genes en linajes específicos.

Gene genica', gain y pordis.

Inferencia de replaciones ortólogas para detectar la pérdida o ganancia de genes en linajes específicos.



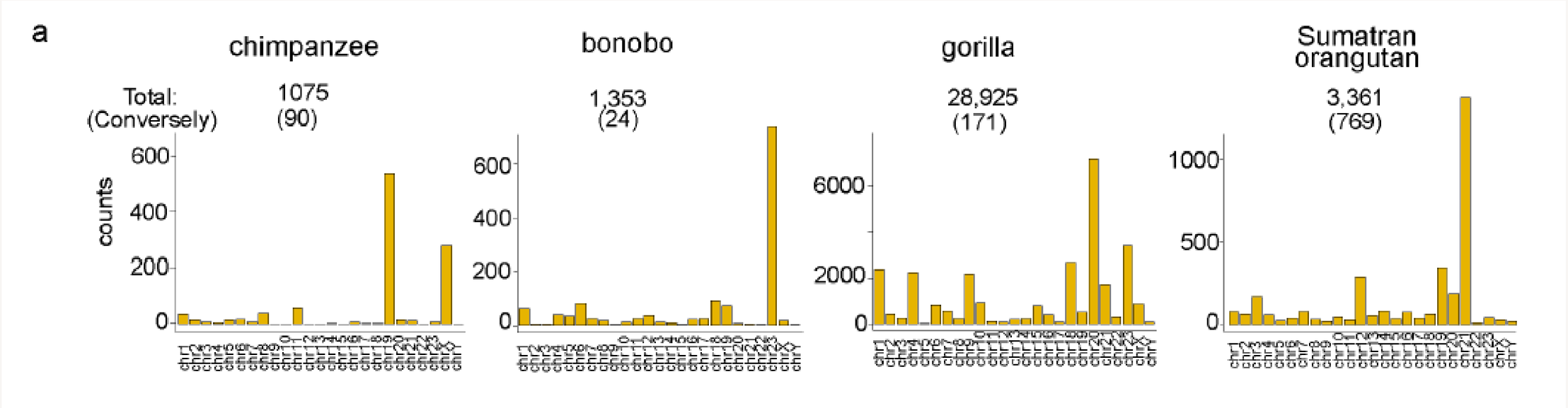

### Análisis Genético Evolutivo

Evaluación de la expansión de familias génicas y selección positiva mediante árboles filogenéticos.

Identification of gene gains, losses, and lineage-specific expansions.

# METHODOLOGICAL WORKFLOW

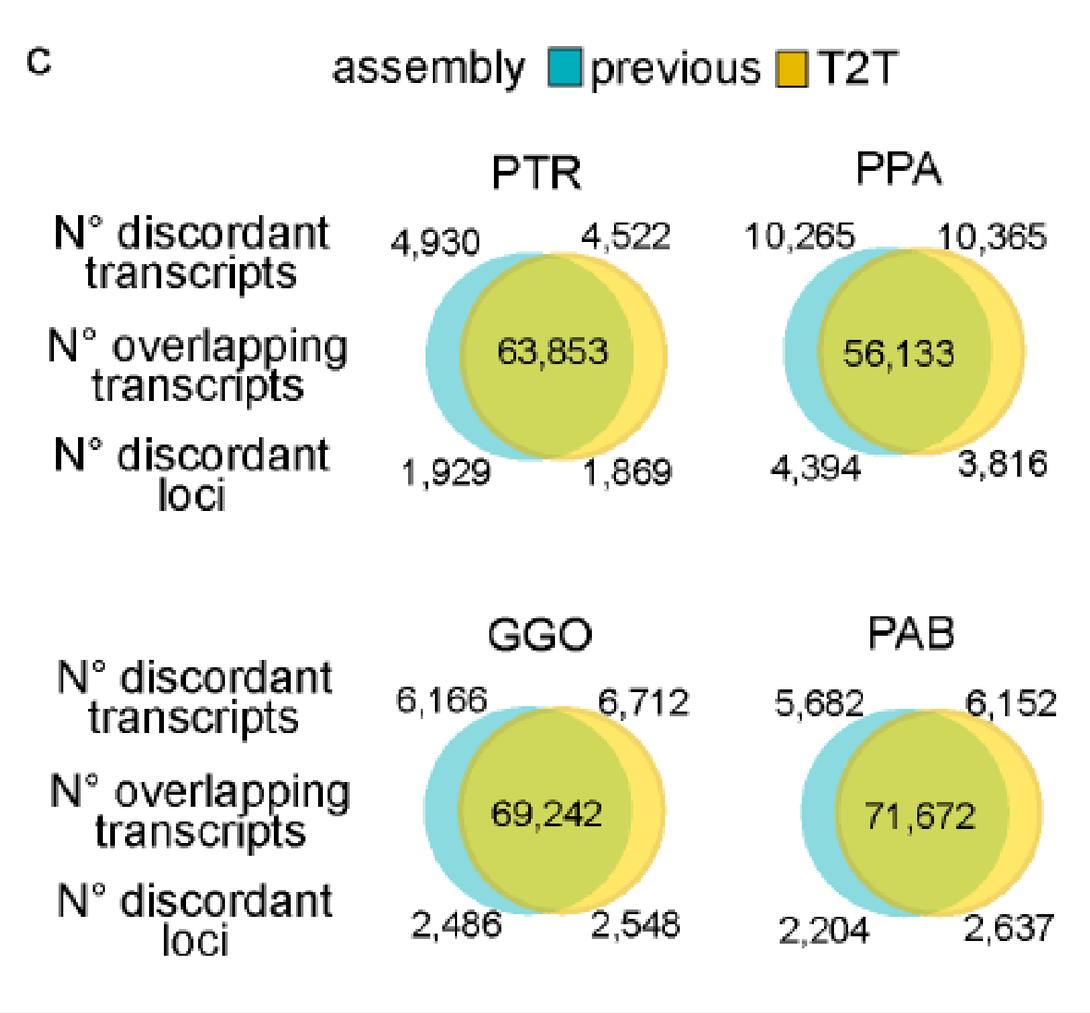
The importance of this workflow is demonstrated in **Supplementary Figure VIII.29.**




**Panel A shows that T2T assemblies recover transcript reads that failed to map to previous genome assemblies, increasing transcriptome completeness.**
Numbers above each plot show reads gained with the T2T assembly; numbers in parentheses show the few reads lost compared with previous assemblies.

# METHODOLOGICAL WORKFLOW

The importance of this workflow is demonstrated in **Supplementary Figure VIII.29.**



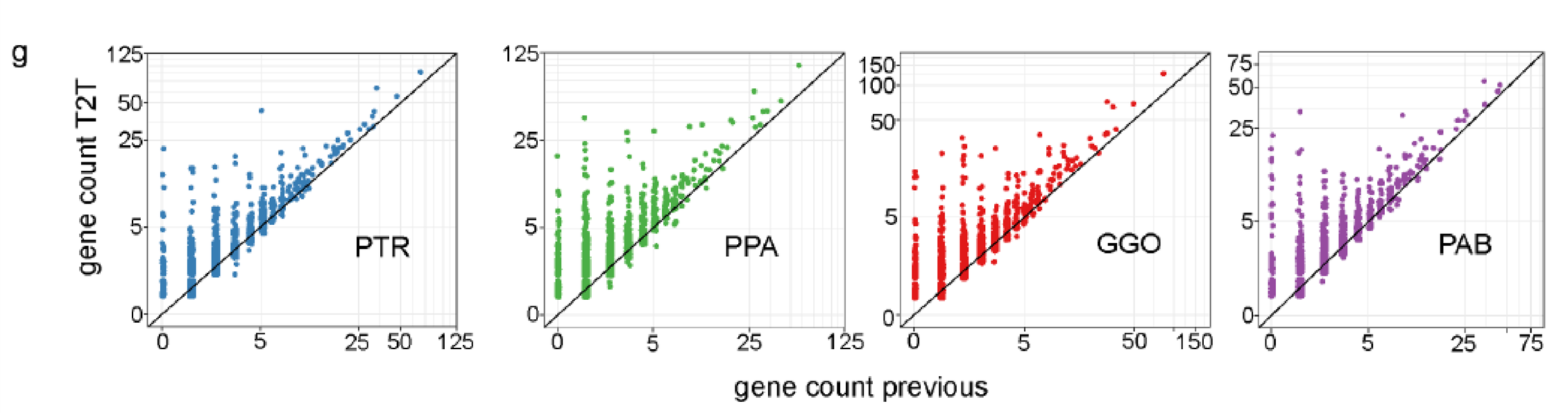

**Panel C compares transcript predictions from Iso-Seq data mapped to T2T (yellow) and previous (blue) genome assemblies across four great ape species, highlighting the improvements provided by T2T assemblies.**

- N° overlapping transcripts: how many transcripts are exactly the same in both versions (same string of introns).
- N° of discordant transcripts: how many transcripts are unique to each assembly.
- N° discordant loci: genomic regions where gene models differ between previous and T2T assemblies.

Chimpanzee (**PTR**), Bonobo (**PPA**), Gorilla (**GGO**) and Sumatran Orangutan (**PAB**)

# METHODOLOGICAL WORKFLOW

The importance of this workflow is demonstrated in **Supplementary Figure VIII.29.**




**Panel G compares gene copy numbers between previous and T2T assemblies.** Each point represents a specific gene family, and points on the diagonal indicate identical copy numbers in both assamblies. X-axis: copy numbers in previous assemblies. Y-axis: copy numbers in T2T assemblies. Many gene families lie above the diagonal, indicating that T2T assemblies identify more gene copies than previous assemblies.

Chimpanzee (**PTR**), Bonobo (**PPA**), Gorilla (**GGO**) and Sumatran Orangutan (**PAB**)

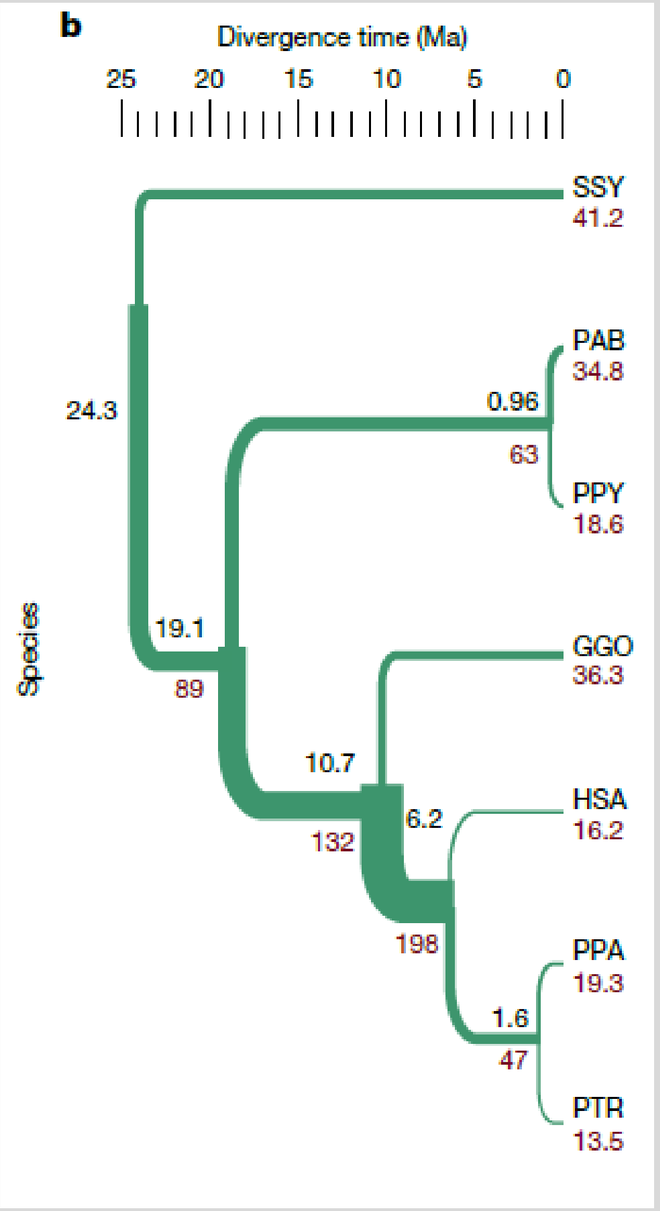
# METHODOLOGICAL WORKFLOW

They built global alignments to mitigate human reference bias.

**Progressive Cactus** provided reliable whole-genome alignments across species. These aligned regions were subsequently analysed with TRAILS to estimate speciation times, incomplete lineage sorting and ancestral effective population sizes, contributing to the phylogenetic reconstruction shown in **Figure 2, panel B**.

- The **black numbers** represent the estimated separation times between lineages, expressed in millions of years.
- The **red numbers** for **ancestral nodes** represent the ancestral effective population size, estimated using TRAILS considering the incomplete lineage classification.
- The **red numbers** for **terminal lineages** represent the most recent effective population size, calculated using MSMC2.

**This global approach was complemented by a pangenomic graph analysis...**



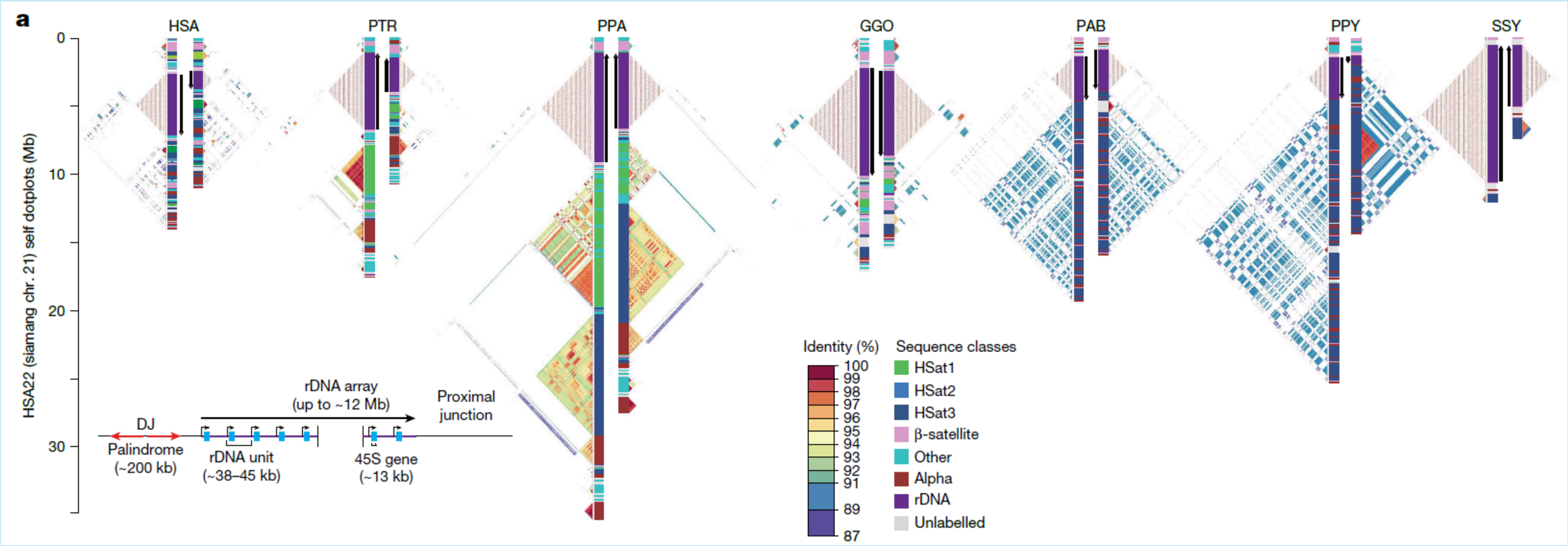
# METHODOLOGICAL WORKFLOW

This stage corresponds to a **diverse complementary biological building blocks**: sucha as centromeres, acrocentric chromosomes, and methylation. All of them take advantage of the fact that T2T assemblies allow the study of repetitive regions that were previously incomplete or absent.

- **Acrocentric chromosomes**: chromosomes whose centromere is near one of the ends, generating a long arm and a small short arm, generally rich in repetitions.
- **NOR regions**: nucleolus organizer regions that contain multiple copies of ribosomal DNA genes, necessary to produce the RNAs that form ribosomes.
- **Centromere**: region of the chromosome where the kinetochore forms and microtubules attach during cell division, allowing the correct separation of chromosomes.
- **Satellite DNA**: short sequences repeated many times in tandem. They are mainly concentrated in centromeric, pericentromeric, interstitial and subterminal heterochromatic regions.
- **Subterminal heterochromatin**: Compact, repetitive DNA located near telomeres. It typically has low transcriptional activity and participates in the organization and stability of chromosome ends.
- **Epigenetics**: the study of modifications that regulate gene activity without altering the DNA sequence, such as cytosine methylation.
- **Segmental duplications**: relatively large fragments of the genome that appear repeated in two or more regions and maintain a high sequence similarity.

# METHODOLOGICAL WORKFLOW

In **Figure 5, panel A**, the organization of the NOR+ short arms equivalent to human chromosome 22 is compared in different species, as well as the siamang chromosome 21. Although these regions share basic components, **their length, composition, and orientation vary considerably between species** and between the two haplotypes of an individual.




 Human (**HSA**), Chimpanzee (**PTR**), Bonobo (**PPA**), Gorilla (**GGO**), Sumatran Orangutan (**PAB**), Bornean Orangutan (**PPY**) and Siamang (**SSY**)

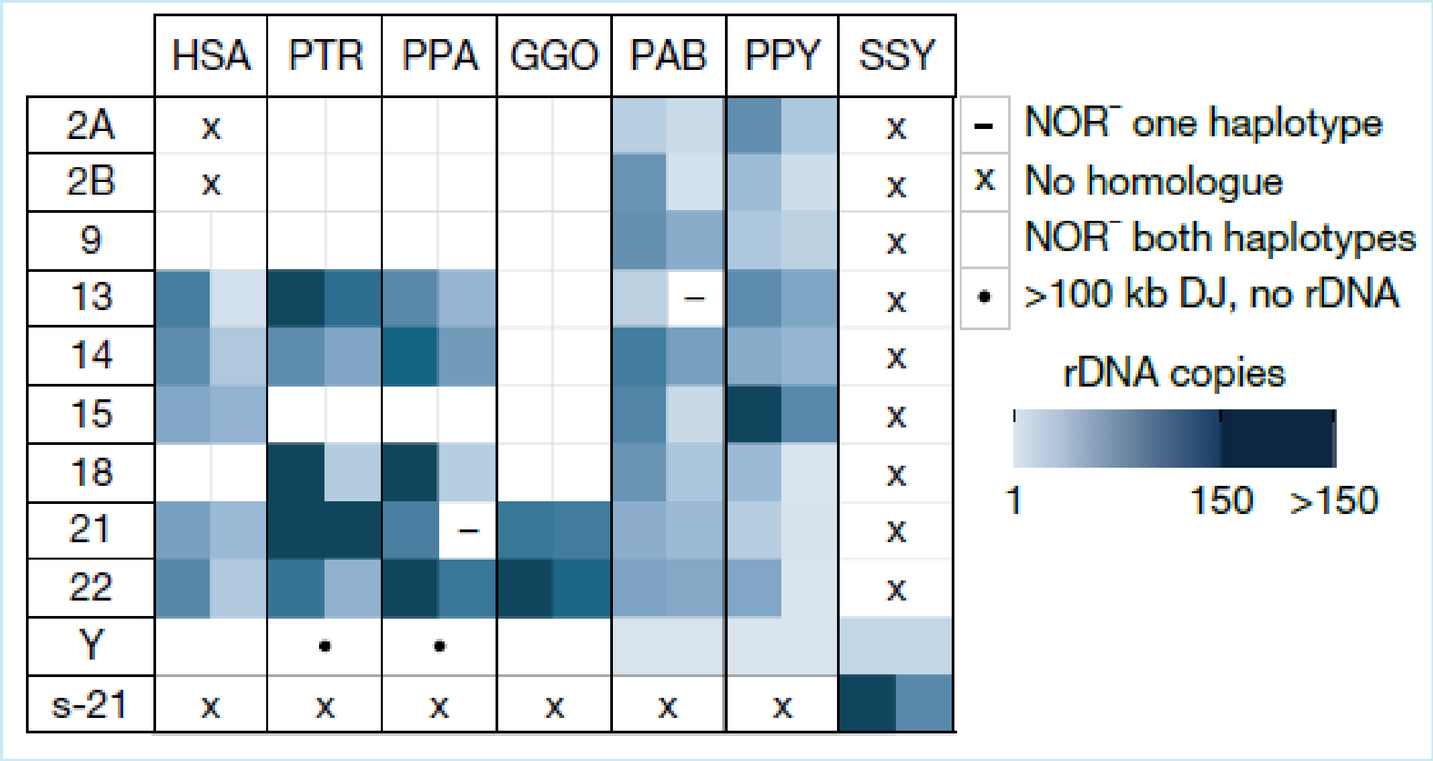
# METHODOLOGICAL WORKFLOW

In **Figure 5, panel B**, shows which chromosomes contain the NORs and how many rDNA units each haplotype has. The location of the NORs is not constant across species.

Each cell is divided into two parts, one for each haplotype. The rows correspond to chromosomes defined according to their homology with human chromosomes.The **intensity of the blue indicates the estimated number of rDNA units**:

- Light blue: few copies
- Dark blue: many copies

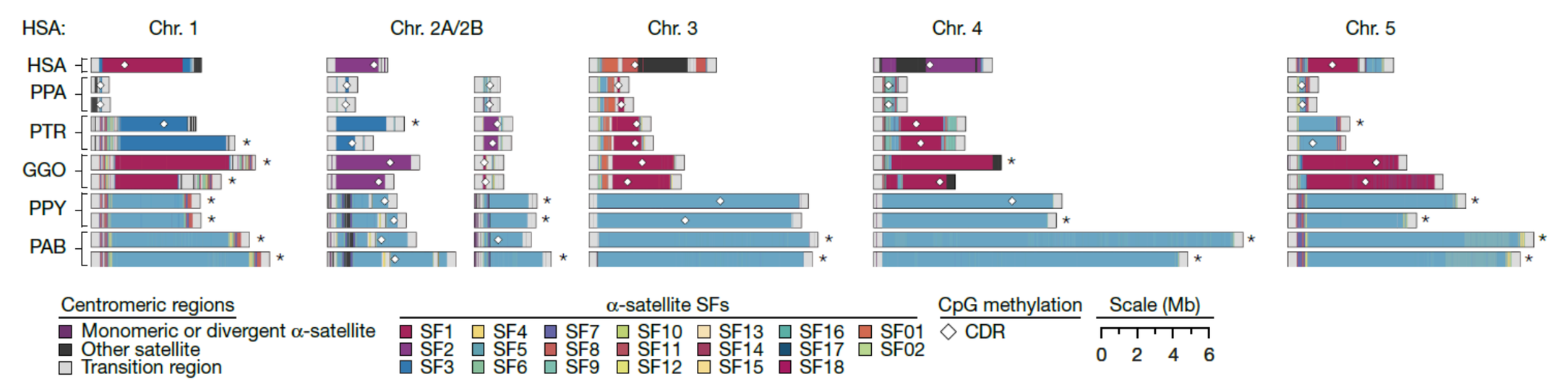
A **white cell** indicates that both haplotypes lack a NOR.
The **dash** indicates loss of the NOR in one of the haplotypes.
The **x** indicates that there is no direct homologous chromosome.
The **dot** indicates that there is an extensive DJ region, but without an rDNA array.




 Human (**HSA**), Chimpanzee (**PTR**), Bonobo (**PPA**), Gorilla (**GGO**), Sumatran Orangutan (**PAB**), Bornean Orangutan (**PPY**) and Siamang (**SSY**)

# METHODOLOGICAL WORKFLOW

In **Figure 6, panel A**, compares the size and composition of alpha-satellite HOR arrays across homologous chromosomes. Centromeres vary markedly among species and haplotypes, revealing rapid evolution of centromeric satellite DNA despite conservation of centromere function.



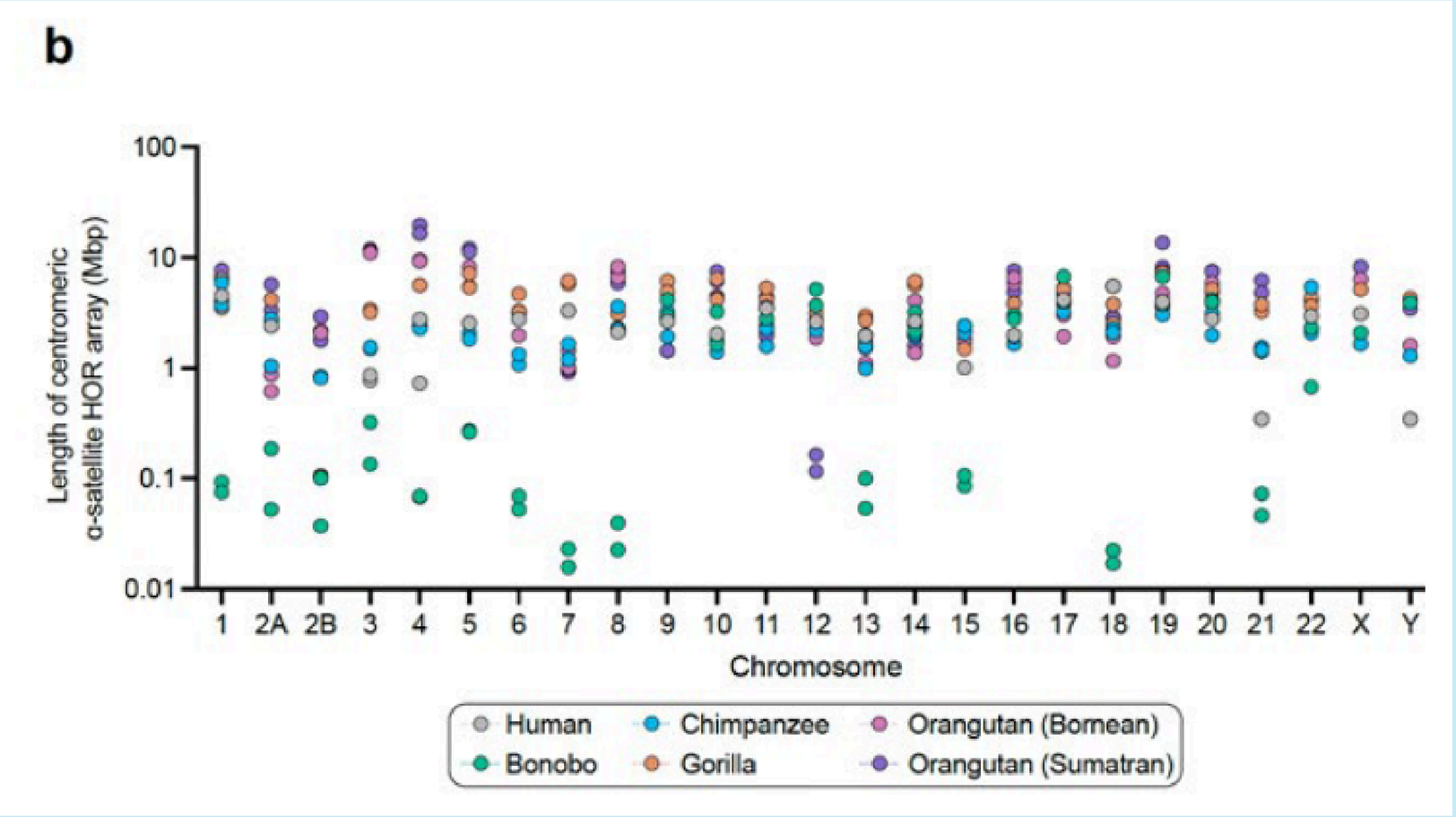

Each bar shows the organization of centromeric repetitive DNA. The colored blocks represent alpha-satellite arrays organized into HORs. Each color identifies a different suprachromosomal family (SF).
**White diamond**: Indicates a region with lower CpG methylation.
**Asterisk**: Indicates that a problem or uncertainty was detected in the assembly of that centromere.

 Human (**HSA**), Chimpanzee (**PTR**), Bonobo (**PPA**), Gorilla (**GGO**), Sumatran Orangutan (**PAB**) and Bornean Orangutan (**PPY**)

# METHODOLOGICAL WORKFLOW

In **Supplementary Figure XIX.67, panel B**, compares the length of centromeric alpha-satellite HOR arrays across homologous chromosomes.

The log-scale distribution reveals **extensive variation among species and haplotypes**, ranging from bonobo minicentromeres to the large arrays found in gorillas and orangutans. This demonstrates rapid expansion and contraction of centromeric satellite DNA **despite conservation of centromere function.**



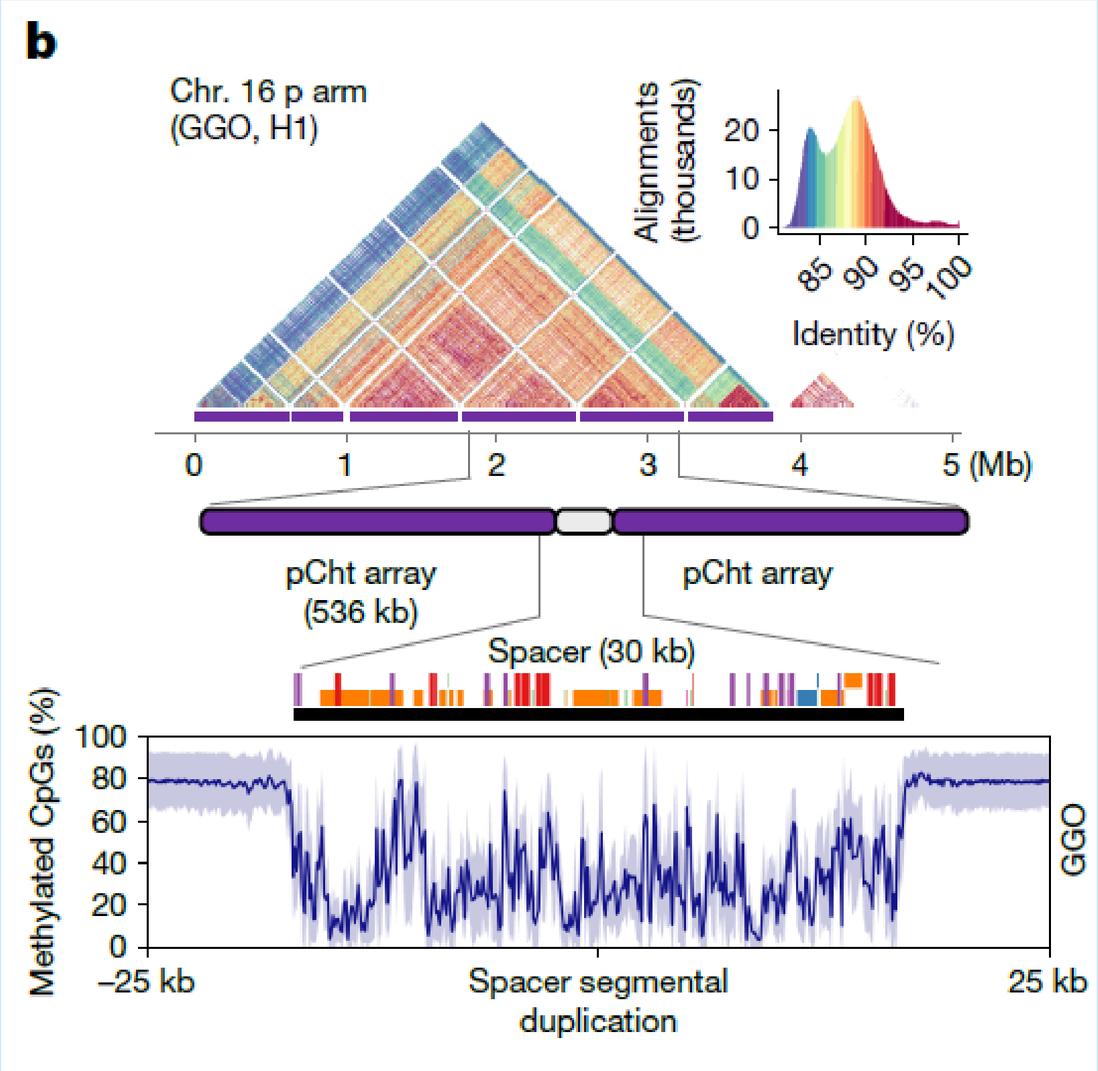

 Human (**HSA**), Chimpanzee (**PTR**), Bonobo (**PPA**), Gorilla (**GGO**), Sumatran Orangutan (**PAB**) and Bornean Orangutan (**PPY**)

# METHODOLOGICAL WORKFLOW

In **Figure 7, panel B**, analyzes a subterminal region of the p arm of gorilla chromosome 16.
The **subterminal heterochromatin** of the gorilla has a periodic organization consisting of large pCht satellite blocks interrupted by duplicated and hypomethylated spacers.

- The **X-axis** represents the distance from the center of the spacer sequence.
- The **Y-axis** indicates the percentage of methylated CpGs.
- The **blue line** shows the average methylation.
- The **light blue band** represents its variability.

A **decrease in methylation is observed at the spacer**, surrounded by the most methylated pCht arrays. This creates a hypomethylated zone within a highly repetitive heterochromatic region.




**pCht satellite:** A 32-bp AT-rich repetitive sequence that forms large subterminal heterochromatic arrays in African ape chromosomes.

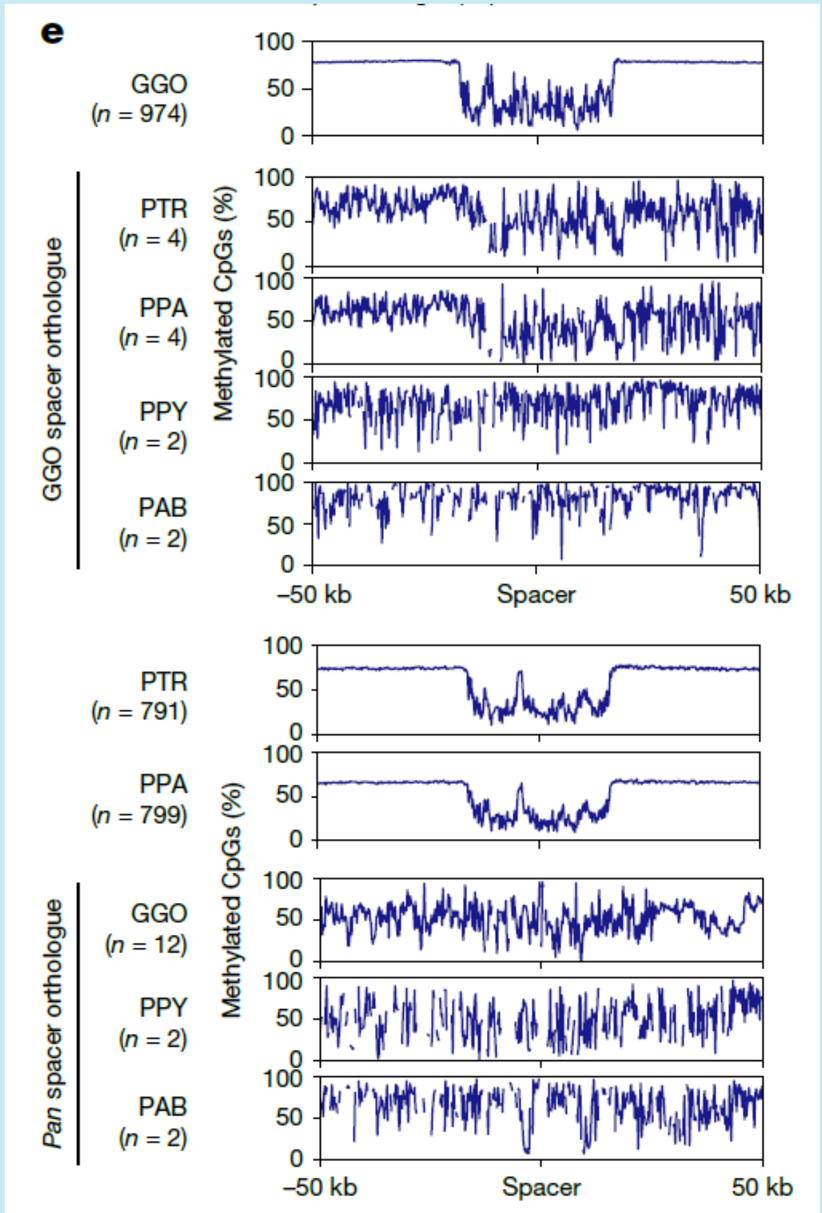
# METHODOLOGICAL WORKFLOW

In **Figure 7, panel E**, this panel seeks to determine whether hypomethylation depends solely on the spacer sequence or on its location within subterminal heterochromatin.

When the spacer is located within the subterminal heterochromatin, a clear drop in methylation is observed in its central region. In contrast, in the ancestral copy or in interstitial copies located outside the subterminal cap, this drop generally does not appear.
**Therefore, the same sequence can show different methylation states depending on its chromosomal context.**

Hypomethylation of the spacer is not determined solely by its sequence, but also by its association with the subterminal heterochromatin cap. This suggests an epigenetic effect associated with the subteminal context.

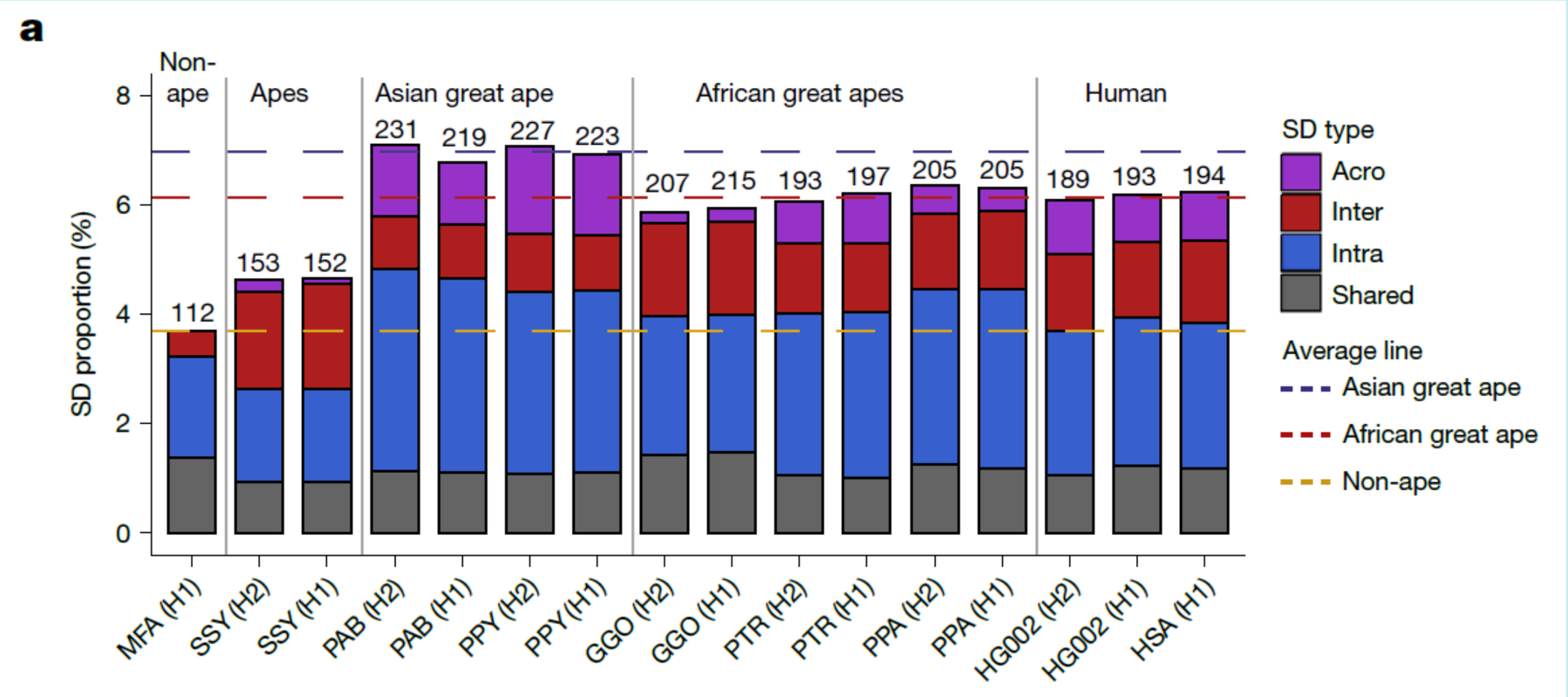
Human (**HSA**), Chimpanzee (**PTR**), Bonobo (**PPA**), Gorilla (**GGO**), Sumatran Orangutan (**PAB**) and Bornean Orangutan (**PPY**)



# METHODOLOGICAL WORKFLOW

In **Figure 8, panel A**, compare the distribution of segmental duplications across genomes. The y-axis and stacked bars show the **proportion of total SD content associated with acrocentric chromosomes**, located within the same chromosome, shared across categories, or distributed between different chromosomes.

The numbers above the bars indicate the **total amount of segmentally duplicated DNA** in megabases, rather than the number of duplication events.



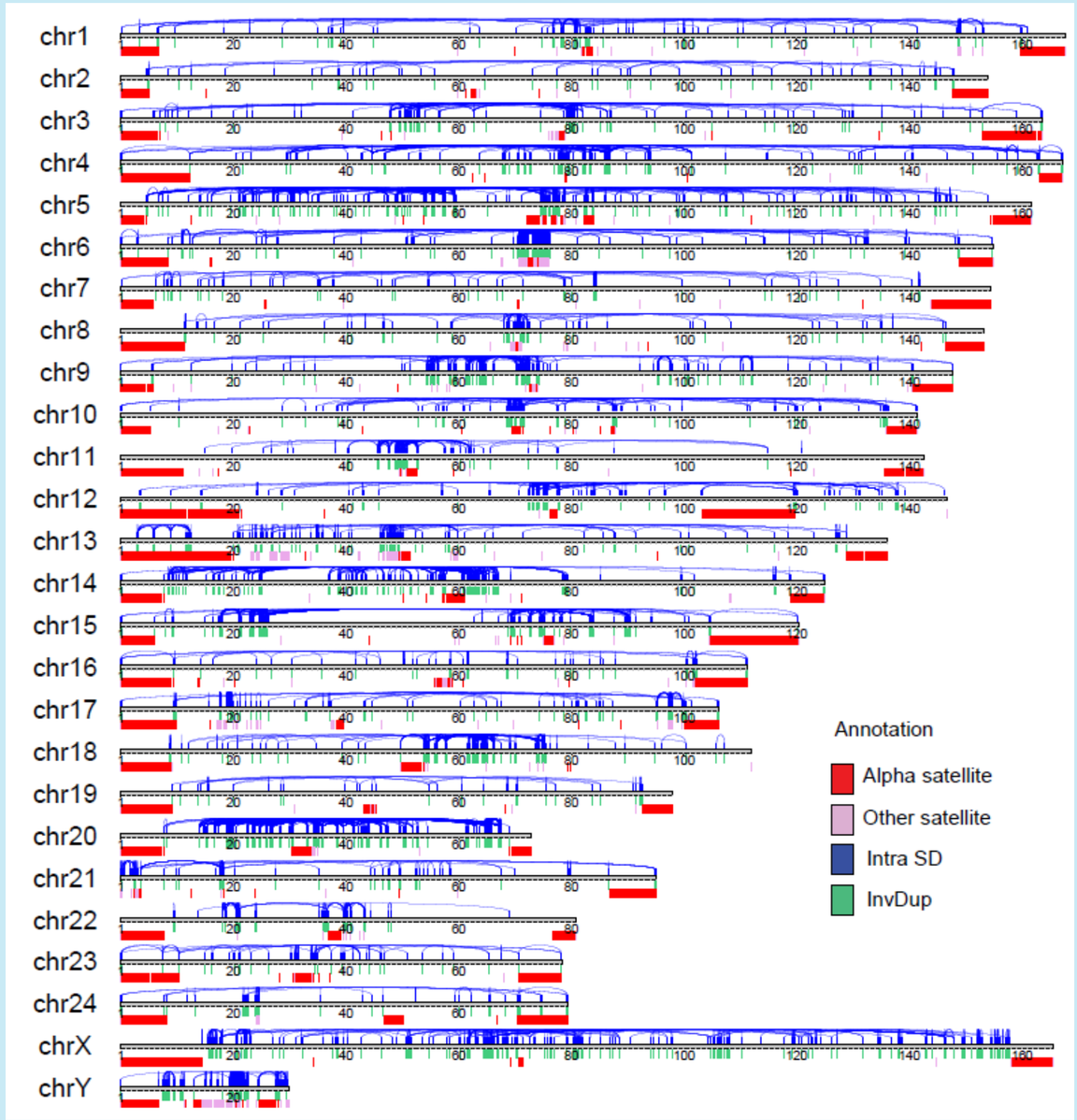

Human (**HSA**), Chimpanzee (**PTR**), Bonobo (**PPA**), Gorilla (**GGO**), Sumatran Orangutan (**PAB**), Bornean Orangutan (**PPY**), Siamang (**SSY**) and Long-tailed Macaque (**MFA**)

# METHODOLOGICAL WORKFLOW

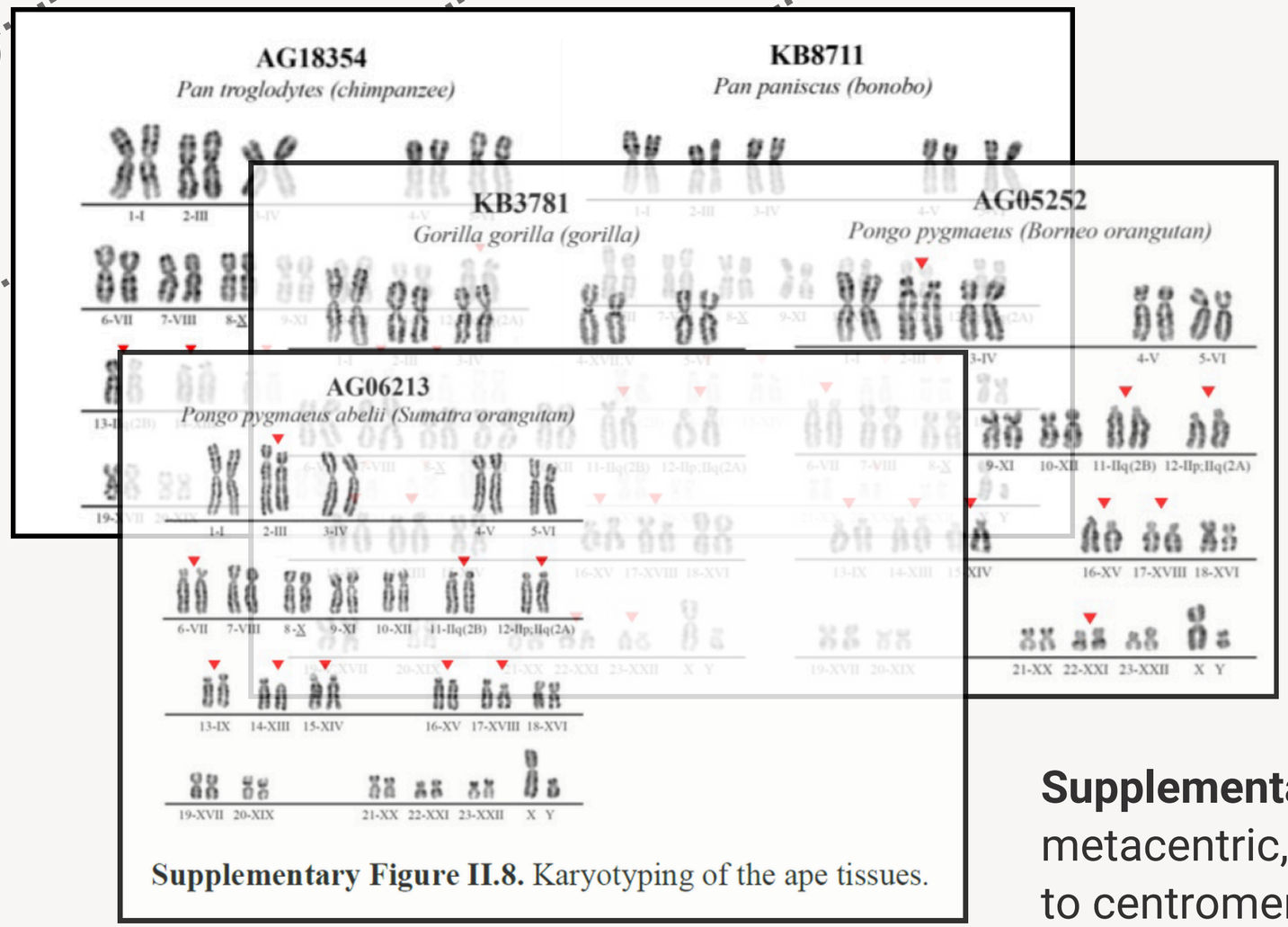
In **Extended Data Fig. 1**, it presents a complete chromosome map of the **siamang genome**, showing where the major repetitive regions and segmental duplications are located.

The most striking feature is the **abundance of large alpha-satellite blocks at the ends of numerous chromosomes**. This indicates that the siamang possesses an exceptionally high amount (~10%) of subterminal heterochromatin.

**Unlike the great apes**, the siamang exhibits a distinctive genomic polarization: the ends of 96 of the 100 chromosomes are dominated by **vast regions of tandemly repetitive DNA**, **flanked by segmental duplications** that mediate the structural complexity of the arms.



# ANALYSIS RESULTS




**Supplementary Figure II.8.** Karyotyping of the ape tissues.

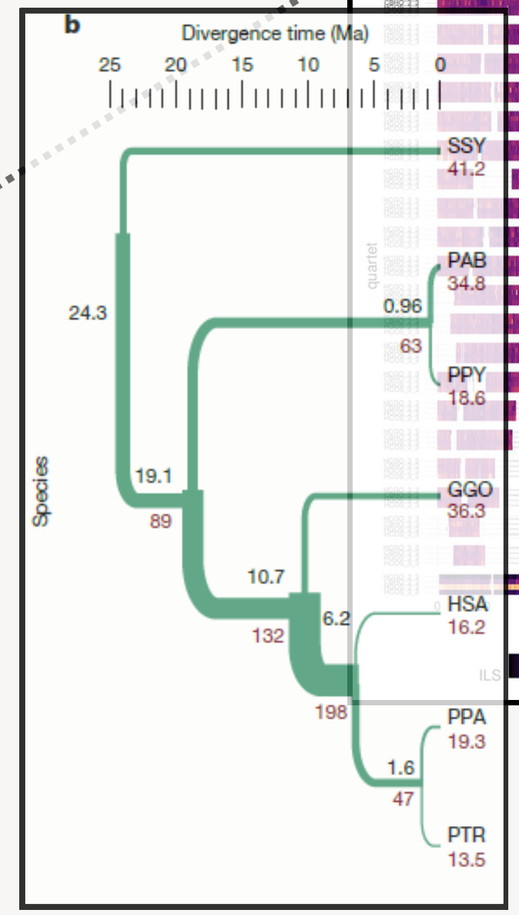
## R1: Construction of Reference T2T Diploid Assemblies

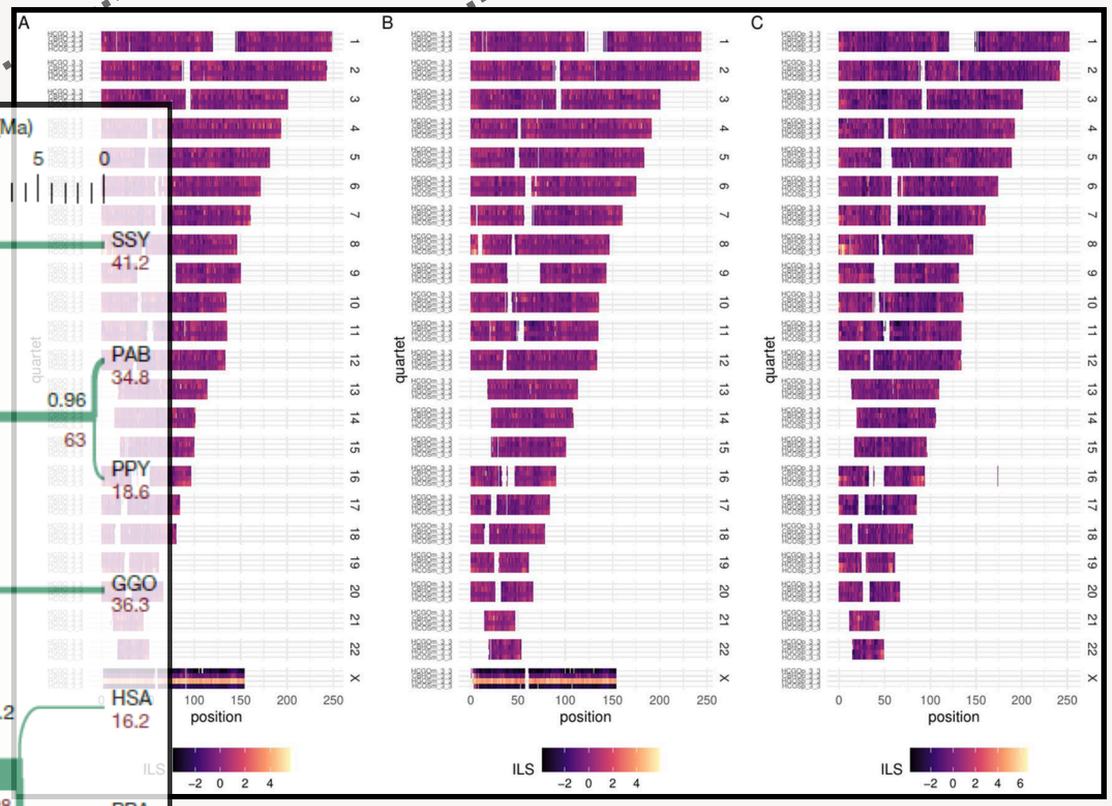
The study generated complete, phased diploid reference genomes for six ape species with **quality comparable to the human T2T-CHM13 assembly**. Validation with Merqury k-mer analyses supported their high accuracy, providing an unprecedented resource for studying previously inaccessible regions such as centromeres, satellite DNA and segmental duplications. However, the **interiors of some rDNA arrays and portions of the largest centromeres remained unresolved**.

**Supplementary Figure II.8** classifies chromosomes as metacentric, submetacentric, or acrocentric according to centromere position and chromosome-arm lengths.

# ANALYSIS RESULTS

## R2: Evolutionary Refinement and Incomplete Lineage Sorting (ILS)



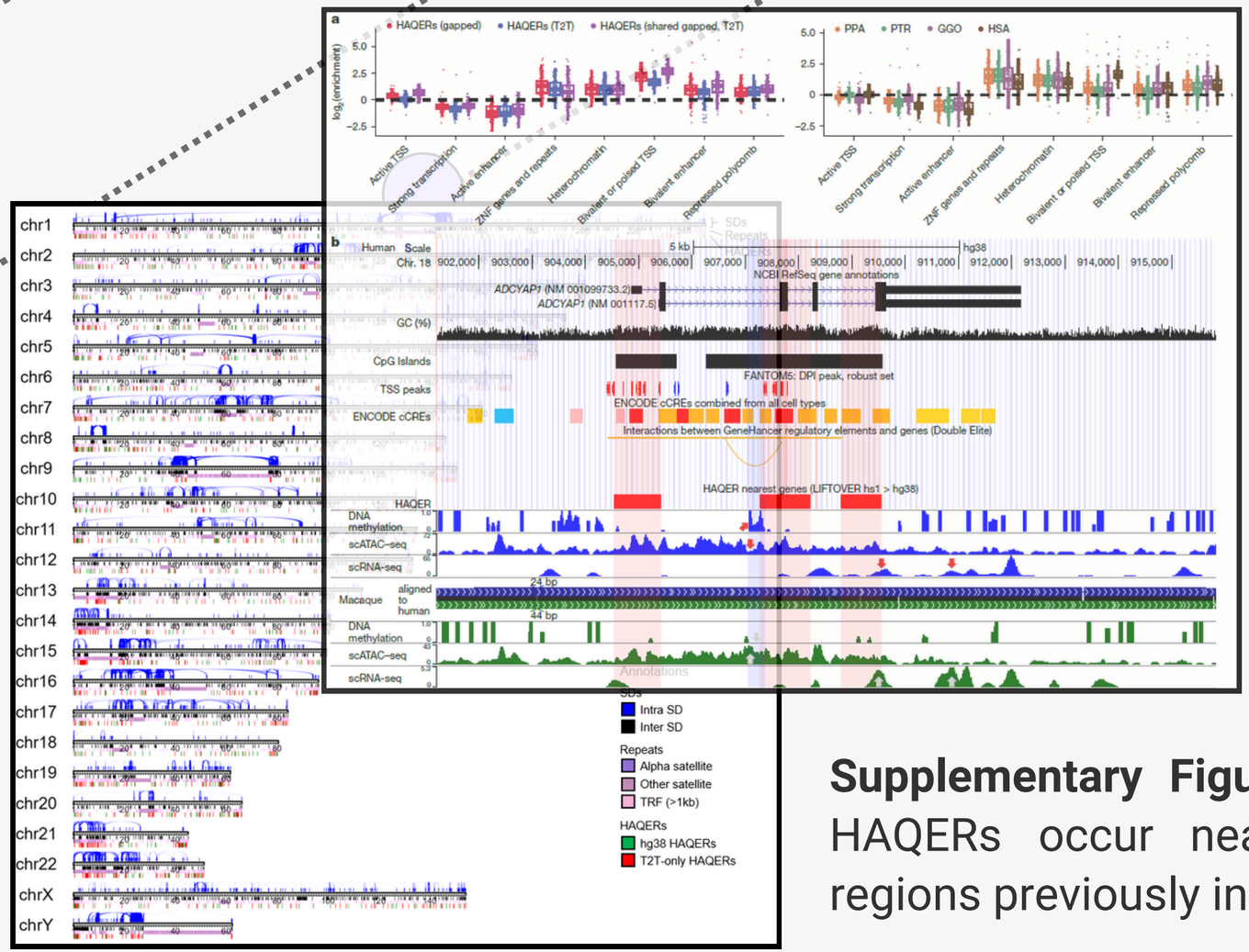




Using T2T whole-genome alignments, TRAILS and MSMC2 refined ape speciation times and effective population sizes, as summarized in **Figure 2b**. The analysis estimated the human–chimpanzee split at **5.5–6.3 million years ago** and detected incomplete lineage sorting **across 39.5% of autosomes** and **24% of the X chromosome**, showing that many genomic regions follow evolutionary histories different from the species tree.

**Supplementary Figure VI.16** maps this mosaic pattern across chromosomes using alternative haplotypes, with highly consistent results between analyses ($R^2$ = 0.88–0.89).

# ANALYSIS RESULTS

## R3: Rapidly Evolving Regions (AQERs) and Genetic Regulation

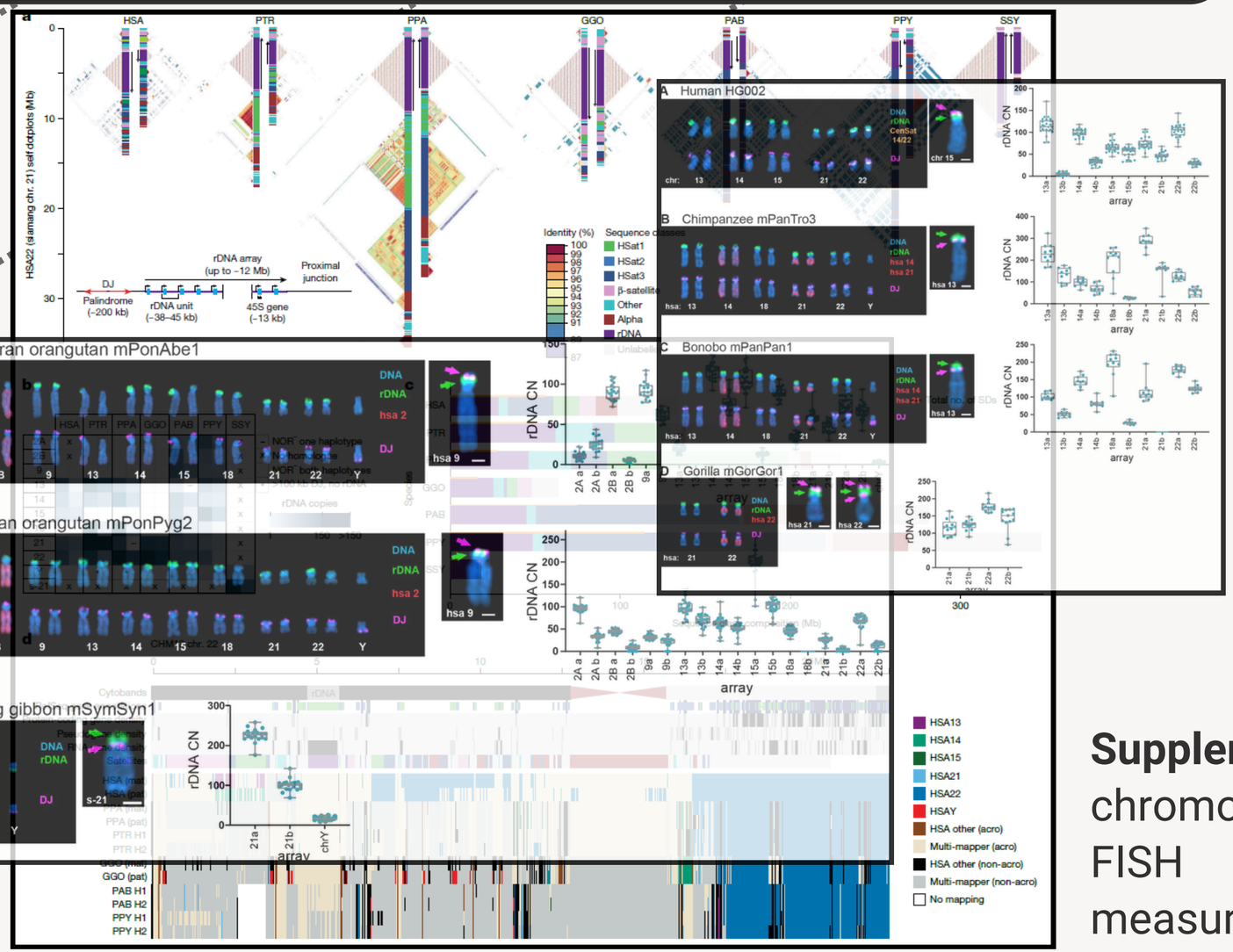



The T2T genomes identified **13,128 ape quickly evolved regions (AQERs)** across four lineages, including 3,268 human-specific HAQERs, more than twice the 1,581 detected using previous gapped assemblies. **Figure 4a** shows that HAQERs are strongly enriched in bivalent regulatory elements, particularly poised promoters, while **Figure 4b** highlights HAQERs within ADCYAP1, a gene associated with vocal-learning-related neural pathways and human-specific epigenetic patterns.

**Supplementary Figure XIII.49** shows that many newly detected HAQERs occur near segmental duplications and satellite-rich regions previously inaccessible to analysis.

# ANALYSIS RESULTS

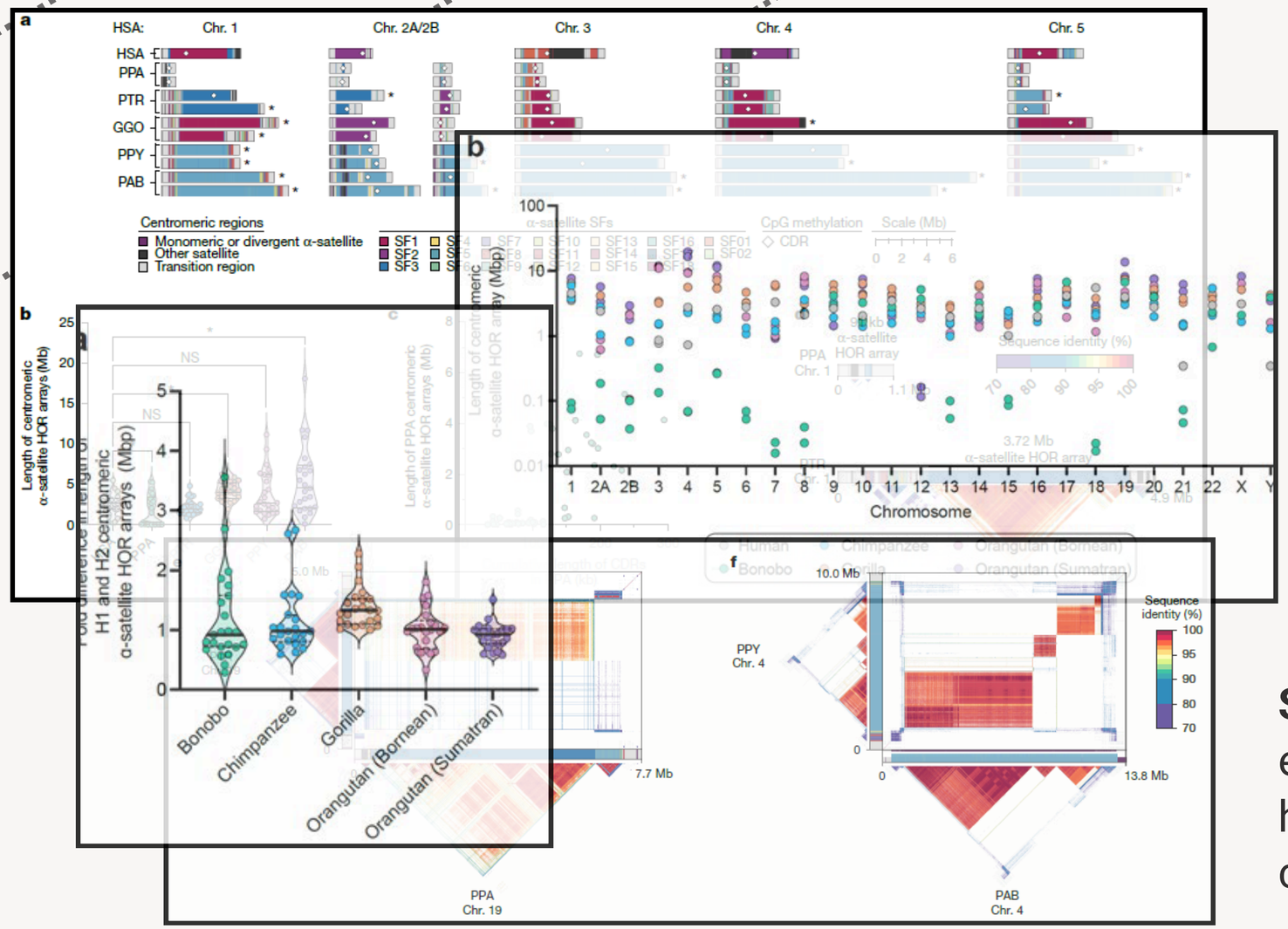
## R4: Architecture of Acrocentric Chromosomes and rDNA




T2T assemblies resolved the repetitive short arms of ape acrocentric chromosomes, revealing major variation in NOR location and rDNA copy number. **Figure 5** shows that individual arrays contain **1–287 rDNA units**, while total diploid copy number ranges from **343 in siamang to 1,142 in chimpanzee**. Despite extensive structural variation, the rRNA-coding regions remain highly conserved, with 18S and 5.8S sequences exceeding 99% identity.

**Supplementary Figure XVIII.59** validates the chromosome-specific distribution of NORs using FISH and independent copy-number measurements.

# ANALYSIS RESULTS

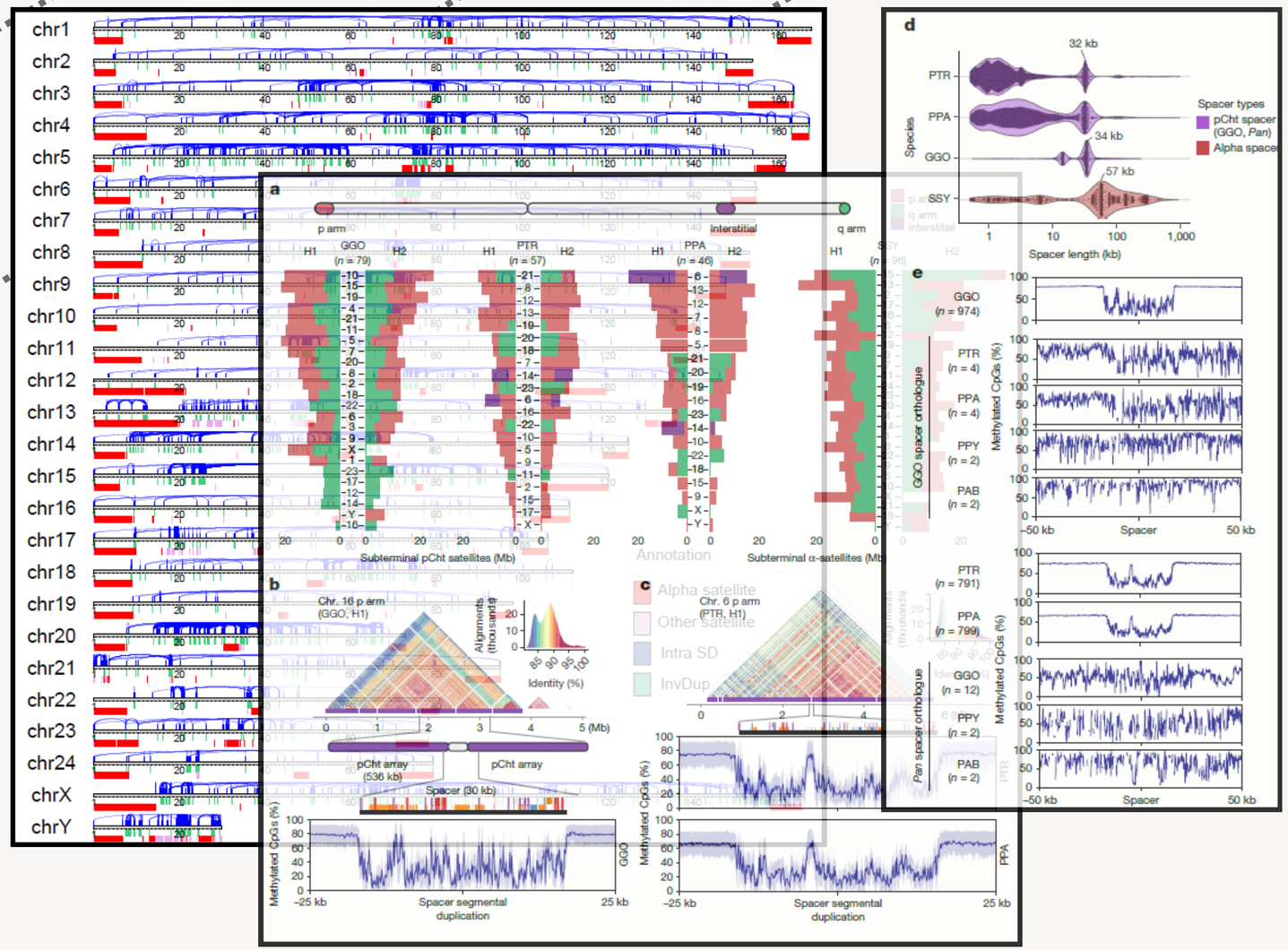



## R5: Dynamics of Centromeres and Satellites HOR

T2T assemblies **resolved 227 of 230 centromeres** across five non-human ape genomes and revealed rapid lineage-specific expansion and contraction of α-satellite higher-order repeat (HOR) arrays. **Figure 6** shows their variation in size, sequence composition and suprachromosomal families, as well as hypomethylated centromere dip regions (CDRs) consistent with kinetochore location.

**Supplementary Figure XIX.67** confirms extensive differences between haplotypes and homologous chromosomes across species.

# ANALYSIS RESULTS

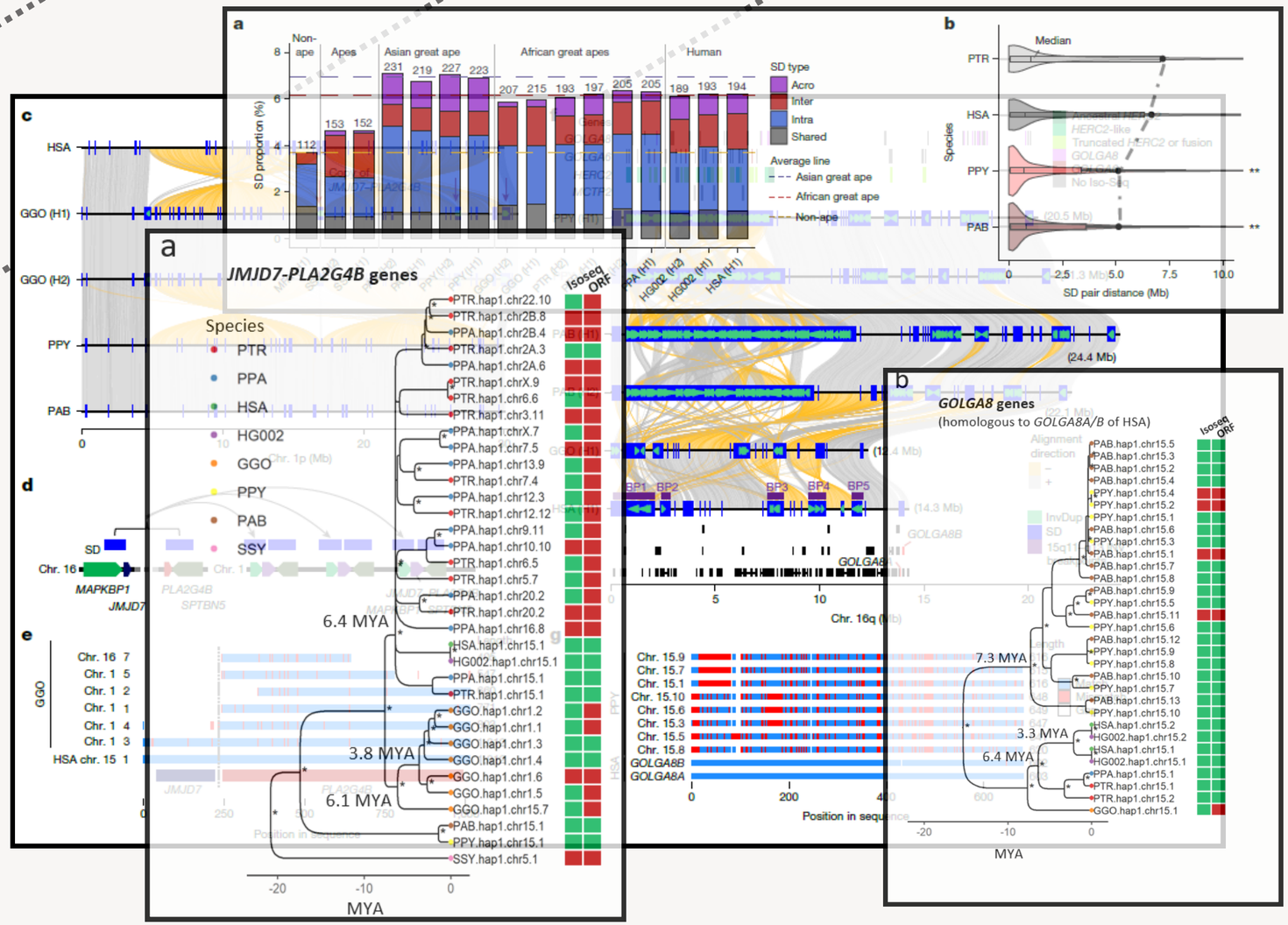



## R6: Subterminal Heterochromatin and Chromosomal "Caps"

T2T assemblies resolved massive subterminal heterochromatic caps in chimpanzees, bonobos, gorillas and siamangs. **Figure 7** shows that these regions can extend up to **26 Mb** and consist of highly methylated satellite arrays interrupted by duplicated, hypomethylated spacers detected using ONT methylation profiles.

**Extended Data Figure 1** highlights the extreme siamang expansion, where subterminal satellites comprise 642 Mb, or 10.1% of the genome, and occur at 96 of 100 chromosome ends.

# ANALYSIS RESULTS




## R7: Segmental Duplications (SD) and Gene Emergence

T2T assemblies resolved an average of 208 Mb of segmental duplications per great-ape haplotype, reaching 225.3 Mb in orangutans, and revealed lineage-specific gene expansions associated with chromosomal rearrangements. **Figure 8** reveals lineage-specific duplication expansions: a **30-Mb double inversion in gorilla** chromosome 1 and a **6.8–10.8-Mb orangutan expansion**.

**Supplementary Figure XXI.78** supports the phylogenetic relationships, transcription and coding potential of these expanded genes.

# CONCLUSIONS

T2T technology eliminates gaps in the primate reference frame, demonstrating that "dark matter" (centromeres, satellites, SDs) is vital to understanding species biology

The discovery that 39.5% of the genome has ILS confirms that human genetic inheritance is a mosaic of shared ancestry, more complex than previous linear models indicated.

The discovery of functional 15 kb centromeres in bonobos revolutionizes our understanding of chromosome architecture, proving that function does not require massive redundancy.

Rapid changes in non-coding regions and segmental duplications are the main drivers of genetic innovation, facilitating unique human traits such as neuronal specialization for language.

# CRITICAL EVALUATION

## STRENGTHS

This is the first time that T2T (Telomere-to-Telomere) quality has been achieved in six ape species simultaneously, eliminating the bias of the human reference.

The combination of genomics (sequence), transcriptomics (Iso-Seq) and epigenomics (methylation) provides robust functional validation to computational assemblies.

The commitment to open science by publishing all resources in a UCSC hub and in public repositories is exemplary.

# CRITICAL EVALUATION

## WEAKNESSES AND LIMITATIONS

The study is based on only two haplotypes per individual per species. This limits the ability to distinguish whether a structural variant is a fixed characteristic of the species or an individual polymorphism (especially in SDs).

Despite the overall success, the centromeres of the Sumatran orangutan (only 27% complete) remain a challenge due to their extreme complexity.

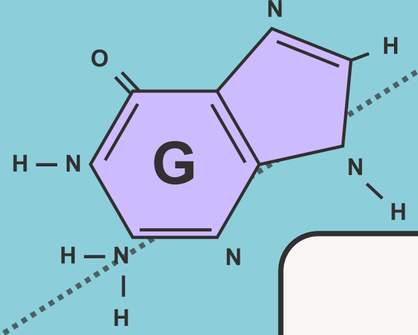
Although the mapping bias was eliminated, annotation bias towards humans persists, since the functions of ape genes are inferred from what we know about our own species.

# CRITICAL EVALUATION

## SUGGESTIONS FOR IMPROVEMENT

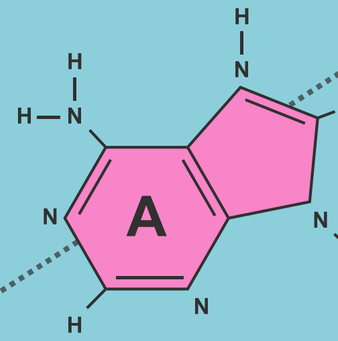
To validate that the reported gene expansions (as in the siamang or gorilla) are defining traits of the species, it would be necessary to sequence at least 5-10 additional individuals per lineage using short reads mapped to these new T2T references.

Perform gene editing experiments (CRISPR) on monkey cell lines to test whether variants in bivalent promoters of HAQERs actually change spatiotemporal gene expression.




The authors were only able to resolve certain genomic regions by telomere-to-telomere assembly. What does this mean for interpreting evolutionary conclusions drawn from previous reference genomes with gaps? How does it change our understanding of genome evolution?

*The transition from gapped genomic drafts to complete telomere-to-telomere assemblies represents a paradigm shift in evolutionary biology.*



## MITIGATION OF REFERENCE BIAS AND CORRECTION OF DIVERGENCE

T2T assemblies revealed that ape genome divergence had been underestimated, with 12.5–27.3% of genomic sequence failing to align consistently in simple one-to-one comparisons.

## RESOLUTION OF GENOMIC "DARK MATTER"

Previous reference genomes systematically excluded the most dynamic and repetitive regions. T2T assemblies uncovered 770–1,482 previously unknown ape gene models, 68.6% of which are associated with lineage-specific segmental duplications.

## DISCOVERY OF FUNCTIONAL PLASTICITY

Bonobos possess extremely small centromeres, some as short as 15 kb and up to 450-fold smaller than typical arrays, showing that centromere function can be maintained across a much broader range of structures than previously recognized.

TALCA UNIVERSIDAD CHILE
Doctorado en Ciencias
mención Modelado de Sistemas Químicos y Biológicos
Facultad de Ingeniería

Genes Genomas y Proteomas

# COMPLETE SEQUENCING OF APE GENOMES

Gabriela Urra Gajardo