Mapping and Sequencing:

Track Name	Meaning
Base Position	Shows the base pairs (A/T/G/C) in a region.
CenSat Annotation	Centromeric Satellite annotations (used in CHM13 genome).
CHM13 unique	Sequences unique to the CHM13 reference genome (vs GRCh38).
rDNA models	Ribosomal DNA repeat regions.
Assembly	Information on the genome build (e.g. CHM13, GRCh38).
GC Percent	Shows GC content across the genome. (Guanine and Cytosine) GC Content (%)=Total number of basesNumber of G's + Number of C's ×100
Human liftOver	Mapping of regions between genome assemblies (e.g. GRCh38 ⇄ CHM13).
Mappability	Indicates how uniquely regions can be aligned by sequencing reads.
Microsatellites / Problematic Regions / Restr Enzymes / Short Match	Special sequence features like repeats, or restriction enzyme cut sites.

CHM13→ The first gapless human genome sequence developed by telomere to telomere sequence.

CHM13 is a **reference genome**, like a master copy of the entire human genome. When you do genome analysis (e.g., variant calling, alignment, GWAS), your sample's DNA sequences are **compared against this reference**. If a region in your sample is **not present in CHM13**, it's considered novel. If a region **only exists in CHM13** and is missing in your sample, it might be **deleted** or just **not sequenced**.

Human Liftover→ **Liftover** is like **translating locations** from one version of the human genome to another.

Think of it like this:

Imagine you have a map of a city from 2009 and a new map from 2022.

- **The locations (addresses) may have shifted because new roads were added, old ones were removed, etc.**
- **You want to find the same house on the new map, even if its coordinates changed.**

Here, the references for the genome changed with time from GRCh38 to CHM13, so to convert data from the former to latter, you use liftover.

Gene and Gene Predictions

Track Name	Meaning
CAT/Liftoff Genes	Genes transferred (“lifted”) from GRCh38 to CHM13 using prediction tools.
NCBI RefSeq	Curated gene annotations from NCBI (trusted reference genes).
CRISPR Targets	Regions that are potential targets for CRISPR gene editing.

Phenotype and Literature

Track Name	Meaning
ClinVar Variants	Shows **clinically significant mutations**, like those causing diseases.
GWAS Variants	Variants found in **genome-wide association studies (GWAS)** that correlate with traits or diseases. You’ve set this one to **"full"** — so you’ll see detailed variant info.

mRNA and EST

Track Name	Meaning
CHM13 PROseq / RNA-Seq	Gene expression data from **PRO-sequencing or RNA-seq** using CHM13.
RefSeq mRNAs	Messenger RNAs (from RefSeq) — shows where genes are expressed as RNA.

Expression and Regulation

Track Name	Meaning
CpG Islands	Regions rich in CG sequences — usually near **gene promoters**.
T2T ENCODE	Regulatory features (enhancers, promoters, etc.) from **ENCODE** project mapped to CHM13.

Comparative Genomics

Track Name	Meaning
Cactus Alignment	Multi-genome alignment to compare CHM13 with other species.
Primate Chain/Net	Alignments between human and **primate genomes** (e.g., chimp, gorilla).

Liftoff gene→ a gene that is used to map gene annotation based a reference gene.

Chr 1

We are talking about the isoforms, full mRNA transcript made of exons after alternative splicing, where the untranslatable introns are removed. Different isoforms of the same gene’s mRNA can be

produced using different combinations of exons, different start and end points, and even different proteins produced.

Alternative splicing→ process where one gene can produce multiple mRNA transcripts by joining multiple exons in different ways.

When the spliceosomes at times skip some exons or even include a bit of introns into the transcript creating different transcripts.

The following color key is used:

- Blue: protein coding→ Exons
- Green: non-coding→ Introns
- Pink: pseudogenes

What are Microsatellites?

Microsatellites (also called **Short Tandem Repeats** or **STRs**) are **short sequences of DNA**, usually **2–6 base pairs long**, that are **repeated in a row**.

Type	Repeat Motif	Example Sequence	Common Features
GT	GT/TG	GTGTGTGTGT	Most common dinucleotide microsatellite, can form Z-DNA→ seen more in non-coding regions
GC	GC/CG	GCGCGCGCGC	GC-rich, stable, often in promoters, may form hairpins→can cause polymerase slippage during the translation.
AT	AT/TA	ATATATATAT	Less stable, flexible DNA, common in non-coding DNA
GA/CT			Stabler than AT, but less than GC; **Polymorphic**: These repeats vary between individuals → used in **population genetics** and **linkage studies; No major human diseases are directly caused by TC/GA repeats (unlike CAG or CGG), but their instability makes them interesting in genome evolution and marker studies.**

 Colored bars in each primate row indicate regions that **align to the human (CHM13) genome**.

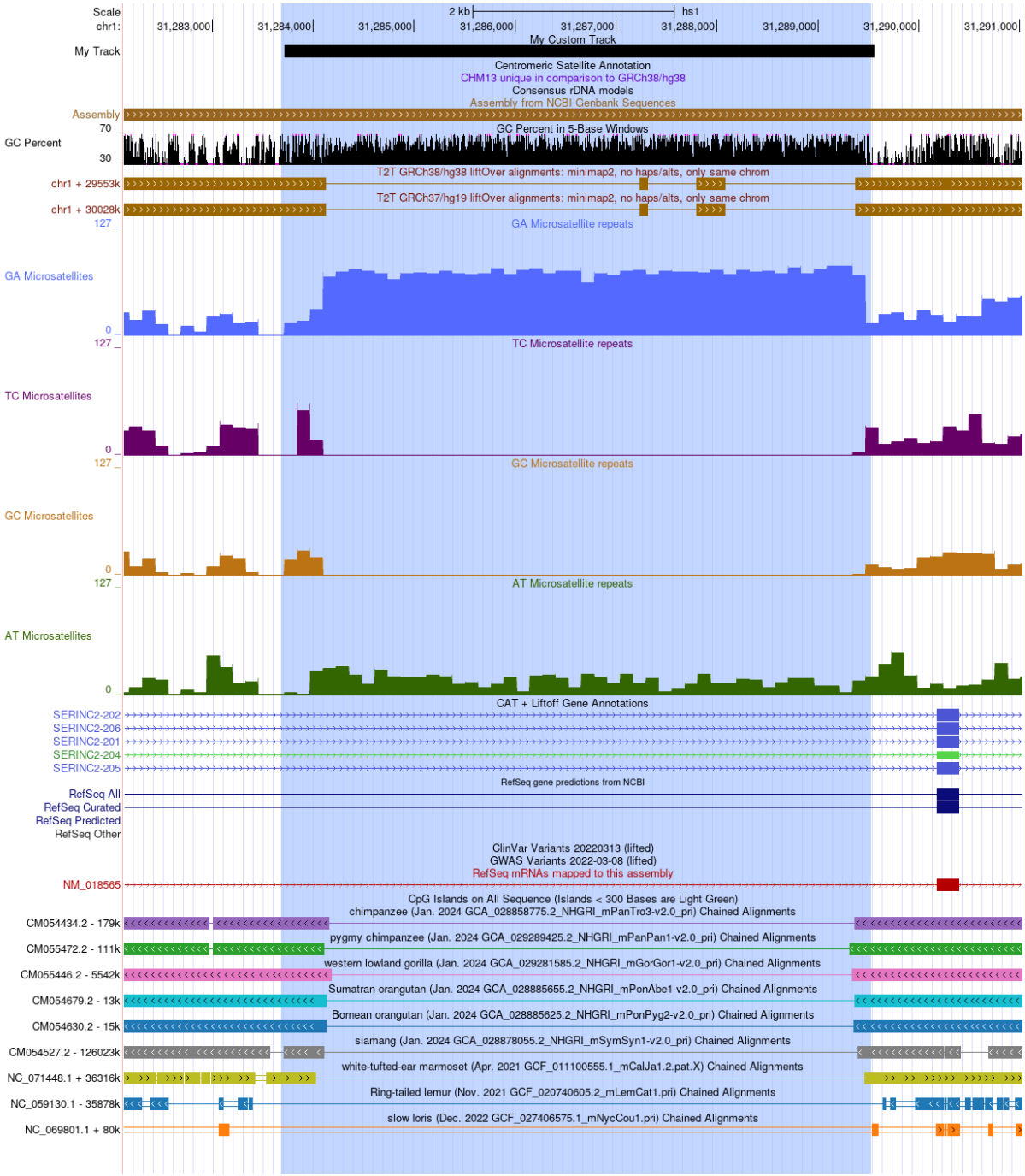
 If a primate has **no alignment block** in a region → it likely **lacks that sequence** or it's too diverged.

If a primate **lacks alignment** in a GA/AT-rich region:

- That region is possibly **species-specific** or **repetitive**, and difficult to map.
- Could indicate **insertions/deletions (indels)** or **tandem repeat expansions** that are not conserved
- • Check whether the **exons of SERINC2 isoforms** (e.g., SERINC2-204) overlap with aligned regions in primates.
- • Exons **shared across multiple primates** = **highly conserved**, likely **functionally important**.
- • Exons only in humans (and not aligned in other primates) = **possible human-specific splicing** or annotation issue.

If only humans (CHM13) have GA-rich sequences in a region **lacking in other primates**, it's likely:

- A **tandem repeat expansion** in humans
- Or an **assembly difference** (T2T captured more of the genome than GRCh38)



1. Based on the highlighted part, it is clear that the region with high microsatellites belong to non-coding strands and exactly the region where they are less, you could see alignment with primate samples too.
2. It's clear that the GA satellite region in human genome is not aligning with the primates, indicating that it is species specific due to some mutation that evolved or tandem repeats.
3. Exons on the human cluster reference do align with the primates, indicating they have common coding region, but have microsatellite which are absent in primates.
4. No ClinVariant and Gwas variant located—not a big danger in disease situation
5. By Human liftover, I was able to assemble the genes common from GRCh38 to CHM13.
6. Also, note that GC percent is just how much G and C present, not as tandem repeats.

chr2

1. Even though the GC percent is high due to CpG Islands, there are no GC tandem repeats as microsatellites.
2. There are no microsatellites found in the region highlighted, which also aligns with the primates genomes.
3. The Transcript basically carried introns and no exons located in the region considered.
4. ClinVariant and GWAS is lifted or coordinates are arranged and linked.