Reference Genome Assembly of *Marmota marmota* Reveals Molecular Bases of Hibernation and Life at High Altitude

***Author*:** Sofia Natale

***Master's Degree Course*:** Bioinformatics, University of Bologna, Italy

***Final submission*:** August 2025

Abstract

High-altitude environments impose selective pressures such as chronic hypoxia, low temperatures and seasonal food scarcity, while hibernation provides a complex strategy for energy conservation. The Alpine marmot (*Marmota marmota*) represents an ideal model to investigate the genomic and regulatory basis of these combined adaptations. A chromosome-scale reference genome is produced through a hybrid assembly of PacBio HiFi long reads polished with Illumina short reads and scaffolded with Hi-C. Functional genomics is addressed by integrating RNA-seq, ATAC-seq, and H3K27ac ChIP-seq in key tissues (liver, brown adipose tissue, heart, hypothalamus) from hibernating and active individuals. Whole-genome bisulfite sequencing (WGBS) provides epigenetic maps of methylation dynamics. To capture natural variation, Pool-seq of 10 populations across the altitudinal gradient (900–2,800 m) of the Stelvio National Park (Italy) enables the identification of loci under selection linked to altitude and hibernation. Integrative analysis of transcriptional, chromatin and methylation profiles with population signals highlights candidate genes and regulatory elements involved in thermogenesis, lipid metabolism, hypoxia response and circannual rhythms. This study establishes M. marmota as a genomic reference for mammalian environmental adaptation and provides insights into conserved metabolic pathways relevant to human physiology.

Key words: Alpine marmot, reference genome, multi-omics integration, population genomics, environmental adaptation

1. Introduction

1.1 Biological and ecological features of the Alpine Marmot

The Alpine marmot (*Marmota marmota*) is a rodent species endemic to the European Alps, where it occupies open grasslands and rocky slopes between subalpine and alpine zones. It lives in highly social colonies and builds extensive underground burrow systems that offer protection from predators and create stable microclimates [1] . The species exhibits an annual cycle characterized by extreme seasonality, with a short summer devoted to reproduction, foraging and fat accumulation, followed by a prolonged winter hibernation during which metabolic activity is drastically reduced [2].

This combination of ecological specialization and extreme physiological transitions makes *M. marmota* an ideal system to investigate how vertebrate genomes regulate energy balance, stress resistance and survival strategies in environments characterized by fluctuation and scarcity.

1.2 Environmental adaptation across alpine landscapes

Alpine marmots are distributed along a wide altitudinal gradient, from subalpine meadows around 900 m to alpine ridges above 2,800 m, where environmental pressures intensify with elevation. At lower sites, colonies experience milder winters but must contend with stronger predation and interspecific competition, while at higher elevations they face chronic hypoxia, prolonged snow cover and a sharply shortened growing season.

This ecological variability generates contrasting selective regimes that act on physiology and genome regulation across populations [3].

Adaptation to these conditions operates at multiple levels. Ecologically, marmots synchronize their life cycle with the short alpine summer, concentrating reproduction and fat accumulation within a narrow window of resource availability. Physiologically, survival during hibernation requires profound metabolic suppression: cardiac and respiratory activity drop drastically, body temperature drops to only a few degrees above the surrounding environment and lipid reserves provide the sole energy source. At the molecular scale, hypoxia-inducible signalling, mitochondrial adaptations and epigenetic reprogramming sustain these transitions, ensuring reversible shifts between active and inactive states [4].

These multi-layered adjustments demonstrate how the species has evolved to withstand the dual challenges of high altitude and prolonged hibernation, offering a living model of resilience under environmental extremes.

1.3 Genomic variation along the altitudinal gradient

The distribution of Alpine marmot colonies across the Stelvio National Park provides a unique opportunity to investigate how genomic diversity is shaped by contrasting environments. Studying allele frequency shifts and regulatory variants along this gradient offers insights into the genetic basis of local adaptation, complementing ecological and physiological observations [5].

A chromosome-scale reference genome is essential to interpret these patterns, as it enables the mapping of population-level variation onto functional elements. In addition to advancing evolutionary understanding, this perspective is crucial for conservation, since the amount of genetic diversity will influence the species' ability to respond to ongoing climate change [6].

Figure 1: ***from Stelvio National Park*** (https://www.stelviopark.it/). The Alpine marmot (*Marmota marmota*), emblematic species of the European Alps and model for studying adaptation to high altitude and hibernation.

1.4 Aim of the project

Despite the ecological importance of the Alpine marmot, no high-quality reference genome is currently available. This gap limits the ability to investigate how mammals combine high-altitude survival with prolonged hibernation. The aim of this project is to generate a chromosome-scale reference genome of *M. marmota* through a hybrid approach integrating PacBio HiFi long reads, Illumina short reads and Hi-C scaffolding. This resource will serve as the foundation for accurate gene and regulatory annotation and for the integration of transcriptomic, epigenomic and population genomic data.

By anchoring these multi-omics layers, the project will clarify the molecular basis of environmental adaptation in marmots and establish the species as a reference model. Beyond its intrinsic value, this genomic framework will provide insights into conserved mechanisms of stress resistance and metabolic flexibility with implications for both biomedical research and biodiversity conservation.

2. Material and Methods

2.1 Visual workflow

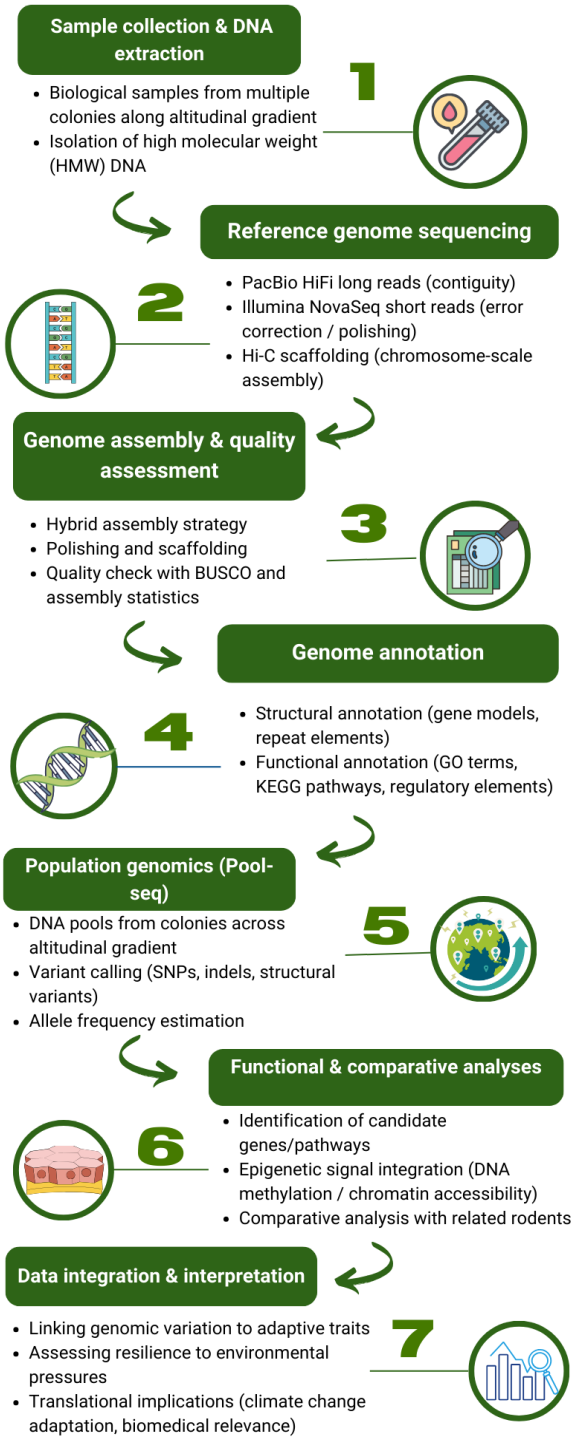

Figure 2: Overview of the experimental and analytical pipeline, summarizing the main steps from genome sequencing to data integration and interpretation.

2.2 Sample collection

Sampling was conducted within the Stelvio National Park (Italian Alps), where Alpine marmot (*Marmota marmota*) colonies are distributed along a wide altitudinal gradient (1,500–3,000 m a.s.l.). This area was chosen as it provides representative environmental conditions for both low and high-elevation populations.

Two seasonal campaigns were organized to capture the two major physiological states of the species. During the active phase (June–September), individuals were trapped while foraging above ground, whereas during the hibernation phase (December–February) they were accessed directly inside their burrow systems under veterinary supervision. This dual design ensured the collection of tissues from both active and hibernating marmots.

Captured animals were sedated to allow the collection of biopsies from target tissues (liver, brown adipose tissue, heart, hypothalamus) as well as whole blood samples for high-molecular-weight DNA extraction. Individuals were continuously monitored and released at the capture site following complete recovery. All material was preserved in liquid nitrogen or transported in dry shippers to maintain the cold chain, and subsequently stored at –80 °C until processing.

All procedures were conducted under permits issued by the Stelvio National Park authority and approved by national ethical committees (IACUC), ensuring compliance with Italian regulations on wildlife research and animal welfare.

In total, approximately 25–30 individual marmots were sampled for reference genome sequencing and functional multi-omics assays, while ~800 individuals grouped into 80 pools were collected for population genomics (Pool-seq). This strategy provides both high-resolution molecular data at the individual level and broad coverage of population-level variation across the altitudinal gradient.

2.3 Reference genome sequencing

The construction of a chromosome-scale reference genome of *Marmota marmota* was the central methodological step of this project, as it provides the indispensable scaffold for all functional genomics and population analyses. To ensure high-quality input material, high molecular weight (HMW) genomic DNA was extracted from a single reference individual sampled in the Stelvio National Park. This individual was selected to minimize heterozygosity and maximize assembly accuracy. DNA integrity was verified by pulsed-field gel electrophoresis, confirming the presence of fragments exceeding 50 kb, a prerequisite for long-read sequencing.

A hybrid sequencing strategy was implemented to combine the complementary advantages of three platforms:

PacBio HiFi long reads [7] – Four SMRT cells were sequenced, yielding >30× genome coverage with average read lengths of 15–20 kb and single-molecule accuracy above 99.9%. These reads constituted the primary substrate for de novo assembly, providing the long-range continuity required to resolve complex repetitive regions, segmental duplications and structural variants typical of mammalian genomes.

Illumina NovaSeq short reads [8] – 1–2 lanes of 150 bp paired-end reads (~30× coverage) were generated. Despite their limited read length, Illumina reads offer an error rate of <0.1%, making them ideal for polishing. They were aligned back to the PacBio draft assembly to correct small insertion–deletion errors and nucleotide substitutions, thereby ensuring single-base accuracy across the genome.

Hi-C chromatin conformation sequencing [9] – Two Hi-C libraries were constructed and sequenced on the NovaSeq platform, generating genome-wide chromatin contact maps. Hi-C captures physical proximity between genomic regions, enabling the anchoring and orientation of contigs into chromosome-scale scaffolds. This step not only increased contiguity (target scaffold N50 > 50 Mb) but also ensured biologically accurate chromosomal organization.

The assembly pipeline followed a stepwise process:

1. De novo assembly of HiFi reads using a long-read assembler (e.g., Hifiasm).
2. Polishing with Illumina short reads using tools such as Pilon.
3. Scaffolding with Hi-C data through 3D contact map integration (e.g., Juicer/3D-DNA).

2.4 Genome assembly and annotation

The chromosome-scale assembly obtained through the hybrid sequencing strategy was subjected to rigorous quality assessment and subsequently annotated to provide a comprehensive genomic resource for *Marmota marmota*.

Genome assembly quality assessment – Assembly statistics were first computed to evaluate contiguity and completeness. Scaffold N50 was used as the main metric of long-range continuity. Genome completeness was quantified using BUSCO (Benchmarking Universal Single-Copy Orthologs) [10] with the Mammalia dataset, providing an estimate of the proportion of conserved single-copy orthologs successfully recovered. In addition, raw Illumina reads were mapped back to the assembled genome to calculate the overall mapping rate, serving as an independent measure of assembly accuracy and structural integrity.

Genome annotation – Structural annotation of the genome was performed using a combined approach. Gene prediction was carried out through ab initio models (e.g. AUGUSTUS), trained on mammalian genomic features, and

refined by comparative alignment with closely related rodent species such as *Marmota monax* and *Mus musculus* [11]. This allowed accurate identification of exon–intron boundaries, alternative transcripts, and untranslated regions (UTRs). Functional annotation of the predicted gene set was performed by mapping protein sequences against reference databases (UniProtKB/SwissProt, Pfam, KEGG), enabling the assignment of Gene Ontology (GO) terms and pathway classifications. Repetitive elements and transposable elements were annotated using RepeatMasker [23], providing insights into genome architecture and evolutionary dynamics.

2.5 Population genomics

Population-level variation in *Marmota marmota* was investigated through pooled sequencing (Pool-seq), an approach that enables cost-effective allele frequency estimation by sequencing DNA pools rather than individual genomes [12]. This design provides sufficient statistical power for detecting selective signatures while ensuring broad representation of population diversity.

A total of 80 DNA pools were generated, representing 10 populations sampled across the altitudinal gradient of the Stelvio National Park. Populations were stratified into lowland groups (1,200–1,800 m a.s.l.) and highland groups (2,200–2,800 m a.s.l.), reflecting contrasting ecological conditions. Within each population, individuals were subdivided into four technical pools of 10 specimens, thereby reducing stochastic noise introduced by pooling and increasing the accuracy of allele frequency estimation. This framework allowed for robust comparisons between low- and high-altitude habitats, as well as between individuals sampled in active and hibernating physiological states.

Sequencing libraries were prepared from each pool and run across three lanes of the Illumina NovaSeq 6000 platform, generating 150 bp paired-end reads with a target depth of 50–80× per pool. Raw reads underwent quality assessment with FastQC and MultiQC, followed by adapter removal and base-quality trimming with fastp [13]. Filtered reads were then aligned to the Alpine marmot reference genome using BWA-MEM [14], achieving accurate mapping across both unique and repetitive regions.

Variant discovery was carried out using GATK HaplotypeCaller, following best practices for pooled samples by setting ploidy according to the number of individuals per pool [15]. Stringent filtering criteria were applied to retain only high-confidence SNPs and indels. Allele frequencies were then directly derived from the read count information encoded in the VCF (AD, DP fields) and normalized with BCFtools, which enabled accurate estimation of population-level allele frequencies while accounting for sequencing depth variation. This approach, integrated with window-based averaging and site-level filtering, provided robust allele frequency estimation.

Population genetic parameters were computed to investigate patterns of variation across altitude and physiological state.

Analyses included:

- Genetic diversity: nucleotide diversity (π) and Watterson's θ.
- Population differentiation: pairwise F_ST between lowland and highland groups.
- Neutrality tests: Tajima's D to detect departures from neutral expectations.

2.6 Functional and comparative analysis

Functional and comparative analyses were implemented to characterize the molecular basis of hibernation and altitude adaptation in *Marmota marmota* by integrating transcriptomic, epigenomic and regulatory datasets within a comparative evolutionary framework.

Transcriptome profiling (RNA-seq) [16]- RNA was extracted from liver, brown adipose tissue (BAT), heart, and hypothalamus collected from both active (summer) and hibernating (winter) individuals. Stranded poly(A)-enriched libraries were constructed and sequenced on the Illumina NovaSeq 6000 platform, generating 150 bp paired-end reads with an average depth of 40–60 million reads per sample. After quality control with FastQC and MultiQC, adapters and low-quality bases were removed with fastp. Reads were aligned to the reference genome using STAR with a two-pass mapping strategy to maximize splice junction detection. Transcript quantification was performed with Salmon and differential expression analyses were conducted with DESeq2 [24]. Functional enrichment of differentially expressed genes was assessed using GO and KEGG databases.

Chromatin accessibility and enhancer landscapes (ATAC-seq and H3K27ac ChIP-seq) - Chromatin accessibility and regulatory landscapes were investigated using ATAC-seq [17] and ChIP-seq. ATAC-seq was applied to nuclei isolated from liver, BAT and hypothalamus. H3K27ac ChIP-seq was performed on liver and hypothalamus using validated antibodies, providing information on active enhancers and promoters. Sequencing reads from both assays were aligned to the Alpine marmot reference genome with BWA-MEM and genomic regions of enrichment were compared using bedtools [18] to identify overlaps between ATAC and H3K27ac peaks, which were then used to define high-confidence active enhancers and to perform differential analyses of regulatory elements across physiological states.

DNA methylation profiling (WGBS) [19] - Whole-genome bisulfite sequencing was carried out on liver and hypothalamus DNA from active and hibernating marmots, targeting a coverage of 20–30× per sample. Reads were processed with Bismark [25] for alignment and methylation calling. DMRs were identified with methylKit [26] (≥20% difference, FDR ≤ 0.05) and cross-referenced with promoters, enhancers and gene bodies for integration with RNA-seq and ATAC/ChIP-seq data.

Orthologous gene families were identified across *Marmota marmota* and related rodents (*M. monax, Sciurus vulgaris, Mus musculus*) using OrthoFinder [20]. Functional annotation was performed with eggNOG-mapper [21], while synteny analysis was conducted using MCScanX [22]. Comparative analyses highlighted conserved versus lineage-specific expansions of gene families associated with hibernation and environmental adaptation.

2.7 Data interpretation

All datasets were interpreted within a unified framework, consistently anchored to the chromosome-scale reference genome as a common coordinate system. The analytical strategy was guided by three complementary principles:

1. Integration across data types - Genomic variation, expression dynamics and regulatory profiles were considered jointly, with signals evaluated in the same genomic regions to highlight concordant patterns.
2. Context-based contrasts – Molecular differences were assessed by comparing populations from distinct environments (lowland vs. highland) and individuals in different physiological states (active vs. hibernating), ensuring that variation was interpreted in its ecological and biological context.
3. Strength of evidence – Candidate loci were prioritized when supported by multiple, independent layers of data and further explored through functional enrichment and network analyses to evaluate their role in adaptation.

This approach ensured that interpretations were both biologically meaningful and robust, relying on convergent signals rather than isolated observations.

3. Estimated cost

The total planned budget for the project was **300,000 €.** According to the detailed allocation reported below, the effective expenditure amounts to **294,000 €**, reflecting the ambition and complexity of assembling a chromosome-scale reference genome and integrating it with downstream functional and population analyses. The budget covers laboratory and field activities, personnel, sequencing and computational resources, as well as logistics and dissemination, thereby ensuring both the generation of high-quality data and its comprehensive analysis within the project timeframe.

Budget breakdown:

Sample collection & DNA extraction

- ***Fieldwork & logistics: 40,000 €***
  2 seasonal campaigns in the Alps, trapping, consumables, travel, cold shipments.
- ***Laboratory consumables: 8,000 €***
  HMW extraction kits, QC reagents, plastics, buffers.
- ***Shipping & cold chain: 8,000 €***
  Dry shipper, liquid nitrogen/dry ice, sample transport.
- ***Permits & ethics: 5,000 €***
  Authorizations, IACUC

Reference genome sequencing

- ***Reference genome (Illumina + PacBio HiFi + Hi-C): 27,400 €***
  4 SMRT cells HiFi, HMW libraries, 1–2 NovaSeq PE150 lanes for polishing, 2 Hi-C libraries, QC and assembly.
- ***Hi-C extra: 2,700 €***
  1 additional library per alternative state.

Genome assembly & quality assesment

- ***HPC & storage: 15,000 €***
  2–3 years, 2–4 TB active + backup, GPU spot.

Population genomics

- ***Pool-seq: 20,000 €***
  80 pools (10 sites × 2 elevations × 4 technical pools), libraries + 3 NovaSeq PE150 lanes, QC.

Functional & comparative analyses

- ***RNA-seq: 12,400 €***
  6 tissues × 2 states × 4 replicates = 48 libraries, prep + ~1.5 NovaSeq lanes.
- ***ATAC-seq: 5,000 €***
  3 tissues × 2 states × 4 replicates = 24 libraries, prep + 1 NovaSeq lane.
- ***ChIP-seq H3K27ac: 8,000 €***
  2 key tissues × 3 replicates = 12 libraries + 1 NovaSeq lane.
- ***WGBS (Methyl-seq): 15,000 €***
  2 key tissues × 2 states × 3 replicates = 12 samples, libraries + 1–2 NovaSeq PE150 lanes.

Personnel & Dissemination

- *Laboratory technician (18 months):* ***37,500 €***
- *Bioinformatics postdoc (24 months):* ***70,000 €***
- *Dissemination / Open Access:* ***8,000 €***

Total effective cost: 294,000 €

A graphical summary of the main budget allocations is reported in Fig. 3, highlighting the relative contribution of fieldwork, sequencing, computational resources, personnel and dissemination:

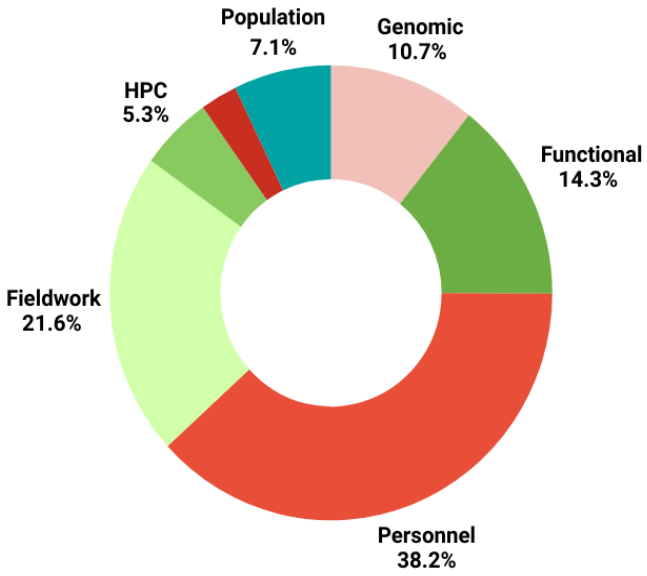

Figure 3: Percentage distribution of the project budget. The largest share of the budget is dedicated to personnel (38.2%), followed by fieldwork (21.6%) and functional sequencing (14.3%). Genomic and population sequencing account for 10.7% and 7.1% respectively, while HPC (5.3%) and dissemination (2.8%) represent smaller fractions of the total costs.

4.Results

The following results are anticipated from the integration of genome assembly, population genomics and functional multi-omics in *Marmota marmota*.

The hybrid sequencing strategy should yield a chromosome-scale reference genome with high contiguity and completeness. Assembly metrics are expected to reach scaffold N50 > 50 Mb, BUSCO completeness >95% and Illumina mapping rates above 95%. Annotation is anticipated to recover ~20,000 protein-coding genes together with non-coding transcripts and regulatory elements, while RepeatMasker is expected to confirm a substantial fraction of transposable elements, in line with other rodent genomes.

Population genomics (80 pools, sequenced at 50–80× coverage per pool) is expected to reveal reduced nucleotide diversity (π, θ) in high-altitude colonies compared to lowland ones, alongside elevated F_ST and extreme Tajima's D values in candidate regions. These regions are likely to cluster around genes linked to oxygen sensing, energy metabolism and physiological adaptation to hypoxia.

Functional multi-omics analyses are anticipated to reveal complementary signatures. RNA-seq should identify several hundred differentially expressed genes between active and hibernating states, enriched in mitochondrial metabolism, lipid catabolism, thermogenesis and circannual rhythm pathways. ATAC-seq and H3K27ac ChIP-seq are expected to identify thousands of accessible chromatin regions and enhancers, with differential activity across physiological states. Motif enrichment analyses should highlight transcription factor families including HIF, PPAR and circadian regulators. WGBS is expected to detect hundreds to thousands of differentially methylated regions (≥20% difference, FDR ≤ 0.05), particularly overlapping promoters and enhancers, consistent with epigenetic regulation of metabolic and hypoxia-related genes.

Comparative analyses with related rodents are expected to identify both shared orthogroups and lineage-specific gene family expansions. Selection scans (dN/dS) are anticipated to reveal signatures of adaptive evolution in marmot-specific genes associated with energy storage, mitochondrial function and hypoxia signalling.

Overall, the study is expected to deliver a high-quality genomic resource and identify the key molecular mechanisms that enable hibernation and survival at high altitude.

5. Discussion

5.1 Key strengths of the project

This project stands out for its ability to integrate ecological context and molecular resolution in addressing two interconnected adaptive strategies: hibernation and high-altitude survival. The dual sampling framework, capturing both seasonal states and altitudinal contrasts, provides a biologically meaningful basis for interpreting genomic and regulatory variation. The use of a hybrid sequencing strategy (PacBio HiFi, Illumina, and Hi-C) delivers a highly accurate reference genome, providing a solid foundation for subsequent analyses. Multi-omics profiling across transcriptomic, chromatin and methylation layers strengthens the identification of adaptive loci, as candidate genes and regulatory elements are validated through convergent evidence rather than single assays. Comparative approaches add an evolutionary dimension, distinguishing conserved mechanisms from marmot-specific adaptations. Overall, the project design balances methodological innovation with ecological realism, increasing both the robustness of findings and their interpretability.

5.2 Limitations of the study

Despite its innovative scope, the project faces several limitations. Hybrid genome assemblies, even when highly contiguous, may fail to fully resolve repetitive or structurally complex regions, potentially leaving gaps in functional annotation. Predictions based on sequence homology or domain similarity can misassign gene functions, underscoring the need for experimental validation. Functional assays such as RNA-seq, ATAC-seq and WGBS capture regulatory dynamics but do not directly demonstrate protein activity or physiological impact, leaving an interpretive gap between molecular signatures and organismal traits.

Population genomic scans, while powerful for detecting selection, may be confounded by demographic history or unequal sampling effort. Finally, the geographical focus on a single park limits the representation of the species' broader range, restricting conclusions about adaptation across the Alps as a whole. These caveats highlight the importance of cautious interpretation and complementary follow-up studies.

5.3 Applications and future perspectives

The implications of this study extend beyond its immediate evolutionary focus. At the biomedical level, dissecting the molecular pathways of hypoxia tolerance, metabolic suppression and circadian regulation may inform research on ischemia resistance, metabolic disorders, obesity and therapeutic hypothermia. At the ecological scale, the identification of genomic markers associated with resilience or vulnerability provides valuable tools for conservation planning, particularly in the context of climate change, which threatens to disrupt hibernation cycles and shorten alpine winters. Methodologically, the pipeline developed here offers a transferable framework for investigating other non-model species under extreme environmental pressures. Future directions include expanding sampling to additional alpine regions, incorporating proteomic or metabolomic data to link regulation with phenotype and experimentally validating key genes and regulatory elements. Such developments will consolidate the Alpine marmot not only as a genomic reference for environmental adaptation but also as a model with direct relevance to human health and biodiversity conservation.

6. Conclusion

This project establishes the Alpine marmot as a reference model for understanding how mammals endure prolonged hibernation and survive under chronic hypoxia. The integrative strategy adopted here does more than generate a genome: it provides a framework to explore how physiology, environment, and evolution interact at multiple molecular levels. The implications reach well beyond marmots, offering new perspectives for biomedical research on metabolic suppression and oxygen deprivation, while also delivering practical tools for biodiversity conservation in a rapidly changing Alpine ecosystem.

7. References

1. Gossmann, T. I., & Ralser, M. (2020). Marmota marmota. *Trends in genetics : TIG*, *36*(5), 383–384. https://doi.org/10.1016/j.tig.2020.01.006
2. Ortmann, S., & Heldmaier, G. (2000). Regulation of body temperature and energy requirements of hibernating alpine marmots (Marmota marmota). *American journal of physiology. Regulatory, integrative and comparative physiology*, *278*(3), R698–R704. https://doi.org/10.1152/ajpregu.2000.278.3.R698
3. Basak, N., & Thangaraj, K. (2021). High-altitude adaptation: Role of genetic and epigenetic factors. *Journal of biosciences*, *46*, 107.
4. Storz, J. F., & Scott, G. R. (2021). Phenotypic plasticity, genetic assimilation, and genetic compensation in hypoxia adaptation of high-altitude vertebrates. *Comparative biochemistry and physiology. Part A, Molecular & integrative physiology*, *253*, 110865. https://doi.org/10.1016/j.cbpa.2020.110865
5. Tigano, A., & Friesen, V. L. (2016). Genomics of local adaptation with gene flow. Molecular ecology, 25(10), 2144–2164. https://doi.org/10.1111/mec.13606
6. Theodoridis, S., Fordham, D. A., Brown, S. C., Li, S., Rahbek, C., & Nogues-Bravo, D. (2020). Evolutionary history and past climate change shape the distribution of genetic diversity in terrestrial mammals. *Nature communications*, *11*(1), 2557. https://doi.org/10.1038/s41467-020-16449-5
7. Rhoads, A., & Au, K. F. (2015). PacBio Sequencing and Its Applications. *Genomics, proteomics & bioinformatics*, *13*(5), 278–289. https://doi.org/10.1016/j.gpb.2015.08.002
8. Modi, A., Vai, S., Caramelli, D., & Lari, M. (2021). The Illumina Sequencing Protocol and the NovaSeq 6000 System. *Methods in molecular biology (Clifton, N.J.)*, *2242*, 15–42. https://doi.org/10.1007/978-1-0716-1099-2_2
9. Lafontaine, D. L., Yang, L., Dekker, J., & Gibcus, J. H. (2021). Hi-C 3.0: Improved Protocol for Genome-Wide Chromosome Conformation Capture. *Current protocols*, *1*(7), e198. https://doi.org/10.1002/cpz1.198
10. Seppey, M., Manni, M., & Zdobnov, E. M. (2019). BUSCO: Assessing Genome Assembly and Annotation Completeness. Methods in molecular biology (Clifton, N.J.), 1962, 227–245. https://doi.org/10.1007/978-1-4939-9173-0_14
11. Stanke, M., Keller, O., Gunduz, I., Hayes, A., Waack, S., & Morgenstern, B. (2006). AUGUSTUS: ab initio prediction of alternative transcripts. Nucleic acids research, 34(Web Server issue), W435–W439. https://doi.org/10.1093/nar/gkl200
12. Hivert, V., Leblois, R., Petit, E. J., Gautier, M., & Vitalis, R. (2018). Measuring Genetic Differentiation from Pool-seq

Data. *Genetics*, *210*(1), 315–330. https://doi.org/10.1534/genetics.118.300900
13. Leggett, R. M., Ramirez-Gonzalez, R. H., Clavijo, B. J., Waite, D., & Davey, R. P. (2013). Sequencing quality assessment tools to enable data-driven informatics for high throughput genomics. *Frontiers in genetics*, *4*, 288. https://doi.org/10.3389/fgene.2013.00288
14. Heng Li, Richard Durbin, Fast and accurate short read alignment with Burrows–Wheeler transform, *Bioinformatics*, Volume 25, Issue 14, July 2009, Pages 1754–1760, https://doi.org/10.1093/bioinformatics/btp324
15. Danecek, P., Auton, A., Abecasis, G., Albers, C. A., Banks, E., DePristo, M. A., Handsaker, R. E., Lunter, G., Marth, G. T., Sherry, S. T., McVean, G., Durbin, R., & 1000 Genomes Project Analysis Group (2011). The variant call format and VCFtools. *Bioinformatics (Oxford, England)*, *27*(15), 2156–2158. https://doi.org/10.1093/bioinformatics/btr330
16. Li, J., Varghese, R. S., & Ressom, H. W. (2024). RNA-Seq Data Analysis. *Methods in molecular biology (Clifton, N.J.)*, *2822*, 263–290. https://doi.org/10.1007/978-1-0716-3918-4_18
17. Grandi, F. C., Modi, H., Kampman, L., & Corces, M. R. (2022). Chromatin accessibility profiling by ATAC-seq. *Nature protocols*, *17*(6), 1518–1552. https://doi.org/10.1038/s41596-022-00692-9
18. Quinlan, A. R., & Hall, I. M. (2010). BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics (Oxford, England)*, *26*(6), 841–842. https://doi.org/10.1093/bioinformatics/btq033
19. Liu, X., Pang, Y., Shan, J., Wang, Y., Zheng, Y., Xue, Y., Zhou, X., Wang, W., Sun, Y., Yan, X., Shi, J., Wang, X., Gu, H., & Zhang, F. (2024). Beyond the base pairs: comparative genome-wide DNA methylation profiling across sequencing technologies. *Briefings in bioinformatics*, *25*(5), bbae440. https://doi.org/10.1093/bib/bbae440
20. Emms, D. M., & Kelly, S. (2019). OrthoFinder: phylogenetic orthology inference for comparative genomics. Genome biology, 20(1), 238. https://doi.org/10.1186/s13059-019-1832-y
21. Huerta-Cepas, J., Forslund, K., Coelho, L. P., Szklarczyk, D., Jensen, L. J., von Mering, C., & Bork, P. (2017). Fast Genome-Wide Functional Annotation through Orthology Assignment by eggNOG-Mapper. *Molecular biology and evolution*, *34*(8), 2115–2122. https://doi.org/10.1093/molbev/msx148
22. Wang, Y., Tang, H., Debarry, J. D., Tan, X., Li, J., Wang, X., Lee, T. H., Jin, H., Marler, B., Guo, H., Kissinger, J. C., & Paterson, A. H. (2012). MCScanX: a toolkit for detection and evolutionary analysis of gene synteny and collinearity. *Nucleic acids research*, *40*(7), e49. https://doi.org/10.1093/nar/gkr1293
23. Tarailo-Graovac, M., & Chen, N. (2009). Using RepeatMasker to identify repetitive elements in genomic sequences. *Current protocols in bioinformatics*, *Chapter 4*, 4.10.1–4.10.14. https://doi.org/10.1002/0471250953.bi0410s25
24. Love, M. I., Huber, W., & Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome biology*, *15*(12), 550. https://doi.org/10.1186/s13059-014-0550-8
25. Krueger, F., & Andrews, S. R. (2011). Bismark: a flexible aligner and methylation caller for Bisulfite-Seq applications. *Bioinformatics (Oxford, England)*, *27*(11), 1571–1572. https://doi.org/10.1093/bioinformatics/btr167
26. Akalin, A., Kormaksson, M., Li, S., Garrett-Bakelman, F. E., Figueroa, M. E., Melnick, A., & Mason, C. E. (2012). methylKit: a comprehensive R package for the analysis of genome-wide DNA methylation profiles. *Genome biology*, *13*(10), R87. https://doi.org/10.1186/gb-2012-13-10-r87