Supplementary Materials

Genomic approximations for the study and conservation of mammals

Aproximaciones genómicas para el estudio y la conservación de mamíferos

Gabriela Castellanos-Morales[1]*, Jorge Ortega[2], Anahí Canedo-Texón[1], Carlos A. Barrera[2], Jesús Antonio Rocamontes-Morales[1] and Katia Hernández-Bolaños[1]

S1. List of abbreviations used in the present article.

bp: Base pair (1 bp)

CR: Control Region of mitogenome

ddRADseq: Double digest restriction associated DNA sequencing

DEG: differentially expressed genes

DNA: Deoxyribonucleic acid

eDNA: Environmental DNA

F_{ST}: Fixation Index

Kb: Kilobasepair (1000 pb)

lcWCS: Low-coverage whole genome sequencing

mitogenome: Mitochondrial genome

mRNA: Messenger RNA

mtDNA: Mitochondrial DNA

NGS: Next Generation Sequencing

NUMTs: Nuclear-mitochondrial segments

PCR: polymerase chain reaction

PCGs: Protein-coding genes

PAVs: Presence/absence variations

RAD-seq: Restriction site-associated DNA sequencing

RNA: Ribonucleic Acid

RNA-Seq: RNA Sequencing

ROH: Runs of homozygosity

rRNA: Ribosomal RNA

RRS: Reduced Representation Sequencing

SNPs: Single nucleotide polymorphisms

SVs: Structural variants

tRNA: Transfer RNA

WGS: Whole-Genome Sequencing

S2. List of concepts frequently used in genomic studies.

Beta diversity: In ecology, this refers to changes in the species composition between two regions. It identifies species turnover along environmental gradients or between sites.

Coverage: The proportion of the genome that has been sequenced at a certain depth. Gives an idea of how much of the total genome length has been effectively sequenced. Typically designed in percentage. For example, 95% of coverage means that 95% of the genome has been sequenced at least once.

Depth: Also known as sequencing depth or depth of coverage. Refers to the number of times each nucleotide has been sequenced. Usually it is designated as 10x, 30x, etc; it means that each nucleotide is sequenced 30 times, for example. Higher depth of coverage increases sequencing accuracy and ensures a better detection of genetic variation.

DNA Sequencing: Refers to laboratory techniques for determining the sequence of nucleotides of a DNA template. It is used to describe and understand the function of genes and genomes of living beings including prokaryotic and eukaryotic.

Double digest restriction associated DNA sequencing (ddRADseq): A reduced representation technique that use two different restriction enzymes at the library preparation step.

Epigenomics: Studies the modifications that DNA can reversibly undergo. The heritability mechanisms that regulate gene expression and, thus, the ability of organisms to respond to environmental changes.

Genomics: Discipline to study the genetic information from an organism, comprises the characterization of the structure, function, location, regulation process, and the sequences of DNA of each of the genes in a biological system.

Genotyping-by-sequencing (GBS): Encompass all reduced representation techniques used to identify genetic variants such as SNPs, small insertions, deletions, and microsatellites, using high-throughput technologies. These methods allow obtaining millions of markers and probes, and detecting sample outliers across samples in the absence of a reference genome.

Low-coverage whole genome sequencing (lcWCS): Approach to generate genome-wide high-throughput sequencing data at very low depth. A powerful and cost-effective

approach for population genomic studies. For this strategy it is preferred to have a reference genome.

Next Generation Sequencing (NGS): Innovative technology refers such high-throughput, that allows parallel sequencing of millions to billions of DNA/RNA fragments in large scale, with the aim of performing any genomic studies without previously known genetic information.

Nuclear-mitochondrial segments (NUMTs): mitochondrial DNA fragments copied into nuclear DNA by horizontal transfer in eukaryotes. Plays an important role in genomic variability and genome evolution. Occur through non-homologous end joining at double-stranded breaks in the nuclear genome.

Metagenomics. Discipline that focus on the genetic characterization (identity and potential function) of microbial communities contained in a biological sample. Metagenomics have allowed the discovery of novel branches in the tree of life.

Omics sciences: Characterized by employing high-performance technologies to generate massive data (also called big data) in a single experiment performed from a single sample; allow an integrative and detailed screening of the functioning of biological systems and the influence of the surrounding environment.

Pangenomics: Comprehensive approach to understanding and elucidating genetic diversity and capturing several categories of genomic variation using a great collection of genomes, rather than a single reference, to reduce bias and capture intra species diversity.

Phylosymbiosis: Refers to the differential levels of concordance between the gut microbiota and host phylogenetic history observed in some mammalian orders (Moeller and Sanders 2020).

Polymerase chain reaction (PCR): One of the most important molecular biology techniques. It is used to amplify specific DNA sequences or regions from any biological sample (blood, tissues, and environmental matrix, etc); its principle is based on the use of DNA polymerase which performs the replication of template DNA, thus, generating billions of copies.

Population genomics: Large-scale study of genetic variation across individuals from different populations; across the integration of high-throughput sequencing into genetics, evolution and ecology to understand evolutionary processes.

Presence/absence variations (PAVs): Structural variants in which a genomic segment containing one or more genes that contribute to ecological adaptation is present in some individuals but absent in others.

Reduced Representation Sequencing (RRS): High-throughput sequencing approach using restriction enzymes to recognize and cleave specific sites in genomic DNA, and generate and sequence fragments of different sizes. This technique significantly reduces genome complexity (1 to 10% of the genome) and thus results in cost-effective alternatives. Typically, no reference genome is required but demands higher depth. This group of techniques includes GBS/RAD for genome sequencing.

Restriction Associated DNA Sequencing (RADseq): A specific reduced representation sequencing techniqus used to detect random genetic markers such SNP

and to genotype species. Based in the use of one restriction enzyme, without previous genomic information. Usually sequenced at high depth, across the entire genome.

RNA Sequencing (RNA-Seq): Protocol implemented by *Next Generation Sequencing* technologies to perform transcriptomic studies. Consist in converting RNA to complementary DNA (cDNA) to sequence and quantify mRNA molecules with the aim of quantifying gene expression profiles associated with different environmental conditions, tissues, cell types, or dynamics over time.

Runs of homozygosity (ROH): Continuous identical segments (haplotype) which are inherited from each parent, manifest high homozygosity rate; are not randomly distributed, appear in low recombination areas through genome. Commonly useful to study inbreeding and also reflecting past demographic history.

Single nucleotide polymorphisms (SNPs): A type of genetic variation, which is the result of possible errors (mutations), changing in a single base pair. SNPs discovery is use in genomics studies to assess levels of genetic variation, genetic differentiation, inbreeding and selection.

Structural variants (SVs): Genomic rearrangements, segments of DNA more than 50 bp in length such as insertions, deletions, duplications, inversions, and translocations. The largest source of functional interindividual genomic variation, that may have evolutionary impacts on species, phenotype differences and population fitness.

Transcriptomics: Science implicated in the study of the second genetic molecular level, means the RNA transcribed from the genome, including the associated sequence, structure, function, regulation, translation, expression, and degradation process. Captures the gene profile associated with specific moments in the biological samples, it is an instant photograph of gene expression.

Whole-genome sequencing (WGS): A methodological framework to get a high-resolution view of the entire DNA sequence from any individual/organism genome which produces a large amount of data. This approach is available to capture coding and no coding information, detect variations and structural genomic rearrangements, etc.