Universitat Autònoma de Barcelona

Development and application of integrative tools for the functional and structural analyses of genomes

Andreu Paytuví Gallart

ADVERTIMENT. L'accés als continguts d'aquesta tesi queda condicionat a l'acceptació de les condicions d'ús establertes per la següent llicència Creative Commons: http://cat.creativecommons.org/?page_id=184

ADVERTENCIA. El acceso a los contenidos de esta tesis queda condicionado a la aceptación de las condiciones de uso establecidas por la siguiente licencia Creative Commons: http://es.creativecommons.org/blog/licencias/

WARNING. The access to the contents of this doctoral thesis it is limited to the acceptance of the use conditions set by the following Creative Commons license: https://creativecommons.org/licenses/?lang=en

DOCTORAL THESIS

Development and application of integrative tools for the functional and structural analyses of genomes

by
Andreu Paytuví Gallart

Supervisors:

Dra. Aurora Ruiz-Herrera Moreno

Dr. Riccardo Aiese Cigliano

Departament de Biologia Cel·lular, Fisiologia i Immunologia, Facultat de Biociències, Universitat Autònoma de Barcelona

&

Sequentia Biotech SL

Development and application of integrative tools for the functional and structural analyses of genomes

Thesis submitted for the degree of Doctor of Philosophy in Cell Biology by the Universitat Autònoma de Barcelona

Supervisor	Supervisor	PhD candidate
Aurora Ruiz-Herrera Moreno	Riccardo Aiese Cigliano	Andreu Paytuví Gallart

Bellaterra, 2019

This thesis was supported by grants from:

- The Industrial Doctorates Plan from the Agència de Gestió d'Ajuts Universitaris i de Recerca (AGUAR) (Generalitat de Catalunya) (2015 DI 003) to Sequentia Biotech SL and Universitat Autònoma de Barcelona (UAB).
- Spanish Ministry of Economy and Competitiveness (CGL2014-54317 and CGL2017-83802-P) to A. Ruiz-Herrera.

Table of contents

Abstract

Since the development of the Sanger **sequencing** in 1977, technological advances have revolutionized the -omics field. Large-scale sequencing projects have resulted in the generation of an enormous amount of data that have motivated the development of **bioinformatics** tools for its integration, organization and interpretation. Due to the fact that the amount of sequencing data produced worldwide doubles every 7 months, there is the need to improve data accessibility, processing and interpretation. In this sense, the main aim of this work is to develop bioinformatics tools for the analysis of the functional and structural characteristics of genomes. On the one hand, storage capacity and accessibility of -omics data has become a challenge, not only for raw data but also for post-processing results. And this is the case for **transcriptomics**, one of the most funded -omics. In order to overcome current limitations on the existing databases for plant lncRNAs, we developed **Green Non-Coding (GreeNC),** one of the most comprehensive online databases in the field that included 39 plant species and 6 algae, representing more than 200,000 lncRNAs. On the other hand, the availability of user-friendly tools to ensure feasible large-scale data analysis and management would help to democratize bioinformatics. Several software have recently emerged to allow the analysis of RNA-seq data in an accessible way. However, none of them provides an end-to-end solution. In this context, we took advantage of **cloud computing** to develop a cloud-based easy-to-use platform called **Artificial Intelligence RNA-seq (AIR)**. AIR is the first end-to-end solution for the analysis of RNA-seq data that is not limited to model species and does not require previous bioinformatics skills. Once developed, we validated AIR taking advantage of RNA-seq samples derived from mouse spermatogenic germ cells produced in our research group. We observed an increase in the prevalence of non-coding genes during **spermatogenesis** and detected silencing of the X chromosome. We also identified differentially expressed genes that were consistent with the sequential development of spermatogenesis. Precisely, it is known that the genome undergoes large **three-dimensional (3D)** conformational changes during spermatogenesis. To characterize such 3D re-organization, we made use of AIR and additional tools for Hi-C data analysis to generate an integrative atlas of the chromatin interactions and functional genomic characteristics of the mouse male germ line. Our results revealed previously undescribed patterns: (i) the sub-chromosomal organization scale is lost during prophase I, (ii) the sub-megabase organization scale becomes diffuse along spermatogenesis especially in sperm, (iii) specific events such as the telomere *bouquet* and the X chromosome inactivation were observed, and (iv) cell-specific open conformations correlated with the expression of genes with relevant functional roles. Overall, we have developed new bioinformatics solutions to enhance

accessibility, processing and interpretation of -omics data that permitted the analysis of functional and structural features of genomes.

Key words: sequencing, bioinformatics, lncRNAs, cloud computing, next-generation bioinformatics, RNA-seq, Hi-C, 3D genome organization, spermatogenesis, meiosis.

List of figures

List of tables

List of boxes

Acronyms and abbreviations

3C - Chromosome Conformation Capture
4C - Chromosome Conformation Capture on-Chip
5C - Chromosome Conformation Capture Carbon Copy
3D - three-dimensional
AIR - Artificial Intelligence RNA-seq
API - Application Programming Interface
AWS - Amazon Web Services
BAC - Bacterial Artificial Chromosome
BLAST - Basic Local Alignment Search Tool
Bp - Base pairs
BWA - Burrows-Wheeler Aligner
cDNA - complementary DNA
ChIP-seq - Chromatin Immunoprecipitation with sequencing
ChIA-PET - Chromatin Interaction Analysis with Paired-End Tag sequencing
CNV - Copy Number Variation
CPM - Counts Per Million
CSV - Comma Separated Values
CT - Chromosome Territories
CTCF - CCCTC-Binding Factor
DaaS - Data as a Service
DEG - Differentially Expressed Genes
DGE - Differential Gene Expression
DNA - Deoxyribonucleic Acid
DSB - Double Strand Breaks
EBI - European Bioinformatics Institute
EMBL - European Molecular Biology Laboratory
ENA - European Nucleotide Archive
ENCODE - Encyclopedia of DNA Elements
EST - Expressed Sequence Tags
FACS - Fluorescence Activated Cell Sorting
FASTA - FAST All
FISH - Fluorescence In Situ Hybridization
FPKM - Fragments Per Kilobase of transcript per Million mapped reads
FPR - False Positive Rate
FTP - File Transfer Protocol
GATK - Genomic Analysis Toolkit
GFF - General Feature Format
GO - Gene Ontology
GOEA - Gene Ontology Enrichment Analysis
GreeNC - Green Non-Coding database
GTF - Gene Transfer Format
GUI - Graphical User Interface
HGP - Human Genome Project
HTTP - Hypertext Transfer Protocol
HTTPS - Hypertext Transfer Protocol Secure
IaaS - Infrastructure as a Service
ICE - Iterative Correction and Eigenvector decomposition
lncRNA - long non-coding RNA
INDEL - Insertion/Deletion

JGI - Joint Genome Institute
Kbp - Kilo base-pairs
Mbp - Mega base-pairs
mRNA - messenger RNA
MSCI - Meiotic Sex Chromosome Inactivation
NAD - Nucleolus Associated Domains
NAR - Nucleic Acids Research
NAT - Natural Antisense Transcript
NCBI - National Centre for Biotechnology Information
NGS - Next Generation Sequencing
NHGRI - National Human Genome Research Institute
NIH - National Institutes of Health
ORF - Open Reading Frame
PaaS - Platform as a Service
PCA - Principal Component Analysis
PCR - Polymerase Chain Reaction
PGC - Primordial Germ Cells
piRNA - piwi-interacting RNA
PMSC - Post-Meiotic Sex Chromatin
RDBMS - Relational Database Management System
RE - Restriction Enzyme
REST - Representational State Transfer
RNA - Ribonucleic Acid
RNA-seq - RNA sequencing
RNAi - RNA interference
rRNA - ribosomal RNA
SaaS - Software as a Service
SDs - Segmental Duplications
siRNA - small interfering RNA
SMRT - Single Molecule Real-Time
SNP - Single Nucleotide Polymorphism
snoRNA - small nucleolar RNAs
snRNA - small nuclear RNA
SQL - Structured Query Language
SRA - Short Read Archive
SV - Structural Variant
TAD - Topologically Associating Domain
tRNA - transfer RNA
TSS - Transcriptional Start Site
UCSC - University of California Santa Cruz
VCS - Version Control Systems

Chapter 1: Introduction

1.1 Genomics

1.1.1 A general framework

Molecular biology is a branch of biological sciences that studies the structure, function and relationships between deoxyribonucleic acids (DNA), ribonucleic acids (RNA), and proteins. It reveals the essential principles underlying the transmission and expression of genetic information. One of the most important discoveries in molecular biology was the identification of the DNA structure by James Watson and Francis Crick in 1953 (Watson and Crick, 1953). Since then, the technological advances in the field, especially in the recent years, have made molecular biology the basis for the development of genomics (Powell, *et al*. 2007). Although the term "genome" was used for the first time back in 1920 to describe "the haploid chromosome set" (review in Goldman and Landweber, 2016) it was not until the description of the first DNA sequencing method (see section 1.1.2) that the term **genomics** was coined. This term contains the suffix -omics, which refers to a field of study in biology that involves large-scale information (Yadav, 2007). Thus, the field that studies the genetic information of an organism is called **genomics**, and it focuses on the structure, function, evolution, and editing of genomes. New techniques launched in the last decades have deepened our understanding of how genomes are organized and regulated. This has resulted in a large number of research areas that, together with genomics, provide a holistic perspective of the biology of the cell. These include: (i) **transcriptomics**, (ii) **epigenomics**, (iii) **proteomics**, (iv) **metabolomics** and (v) **microbiomics**.

In the context of the interrelationships between the above-mentioned technical advances, this thesis is focused on the processing, manipulation, and interpretation of genomics and transcriptomics data in their crosstalk to unveil the functional and structural organization of genomes. This includes the development and application of bioinformatics tools and databases necessary for data processing and sharing. In this sense, recent advances in high-throughput sequencing technologies have been essential in the field.

1.1.2 DNA sequencing: from hundreds to hundreds of billion bases

Determining the linear order of nucleotides is key to understand the information encoded in the molecule. Thus, obtaining the DNA primary sequence is the first step when investigating genomes, and this became true in 1977 thanks to the development of the so-called **Sanger sequencing** (a pioneer method to sequence DNA developed by Frederick Sanger and colleagues, Sanger, *et al.* 1977). Since it represented a methodological leap in the field, Sanger sequencing

was rapidly adopted for the scientific community. In fact, just one year after its initial release, several eukaryotes and prokaryotes genes together with the genomes of small bacteriophages were sequenced. This included the bacteriophage ϕX174, with 5,386 nucleotides long (Sanger, *et al*. 1977), the 16S ribosomal RNA gene from *E. coli* (Brosius, *et al*. 1978), the chicken ovalbumin (McReynolds, *et al*. 1978), the bacteriophage fd DNA (Beck, *et al*. 1978), or the bacteriophage G4 DNA (Godson, *et al*. 1978). The obtention of these initial sequences were tedious and time-consuming since it required manual reading of gels, thus preventing the approach for large genomes due to the big effort required. Nevertheless, these limitations were soon overcome by the development of fluorescence-based Sanger sequencing in 1986, becoming a semi-automatic method (Smith, *et al*. 1986). This facilitated the sequencing of the first living organism (*Haemophilus influenza*, the bacterium that causes influenza) (Fleischmann, *et al*. 1995) in 1995, followed by other model organisms such as yeast (*Saccharomyces cerevisiae*, Goffeau, *et al*. 1996) and the nematode *Caenorhabditis elegans* (*C. elegans* Sequencing Consortium, 1999), among others.

Thanks to the methodological advances in DNA sequencing, an international consortium was created in 1990 with the main aim to sequence the whole human genome (3×10^9 base pairs); this project was called the Human Genome Project (HGP) and represented the largest and the most expensive biological project in history. The project required the development of a significant number and range of methodological tools. Between 1990 and 1995, the consortium was especially focused on creating both genetic and physical maps (Guyer and Collins, 1995). Sequencing efforts were mainly done in the second part of the 90s using the hierarchical shotgun strategy, based on separately sequencing a series of overlapping fragments called Bacterial Artificial Chromosome (BAC) (Shendure, *et al.* 2017). As a result, the first draft of the human genome was released in 2001 with an estimated cost of $300 million (International Human Genome Sequencing Consortium, 2001; NHGRI, 2001). Shortly after, the mouse genome followed (Mouse Genome Sequencing Consortium, 2002).

Due to the extremely high costs associated to sequencing, soon became evident that the development of new and more affordable sequencing methods was much in need. Different companies played an important role to tackle this demand in the mid 2000's with the development of next-generation sequencing methods (NGS) (also known as "second-generation" sequencing). That was the case, for instance, of 454 Life Sciences (acquired by Roche Diagnostics) and Solexa (acquired by Illumina Inc). 454 Life Sciences launched the first commercially available NGS instrument in 2005 being able to obtain 25 million bases in few hours. Its sequencing technology was based on the emulsion PCR (Polymerase Chain Reaction)

and the pyrosequencing technology (Margulies, *et al*. 2005). Solexa Inc., on the other hand, launched its own NGS instrument in 2006 based on the reversible dye-terminators technology on a flow cell surface that allowed the sequencing of 1000 million bases (1 Gbp) (Bentley, *et al.* 2008). Both NGS methodological approaches broke Moore's Law (Moore, 1965), which states that the capacity of data processing doubles every 18 months, having as a consequence a remarkable decrease of the sequencing costs. Along with the new developments, Life Technologies (acquired by Thermo Fisher) launched in 2010 Ion Torrent (table 1), based on detecting the pH change when the nucleotide is incorporated to the template (Rothberg *et al.* 2011).

It comes as no surprise that the arrival of the NGS technologies promoted the release of new and more ambitious genomic projects. The high throughput capacity of NGS allowed the obtention of sequences covering a particular locus many times (read depth/coverage), thus having the possibility to call genomic variants such as Single Nucleotide Polymorphisms (SNPs), small insertions and deletions (INDELs), and Structural Variants (SVs), including Copy Number Variants (CNVs) and Segmental Duplications (SDs) (Xi, *et al*. 2010; Nielsen, *et al*. 2011). In this context, the 1000 Genomes Project, created in 2008, represented an advance in the field (The 1000 Genomes Project Consortium, 2010). The initial goal of this international consortium was the development of a catalogue of human variation from different ethnic groups around the world. The last phase of the 1000 Genomes Project (phase 3), which included 2,504 individuals from 26 different populations, identified 84.4 million of variants (The 1000 Genomes Project Consortium, 2015).

Meanwhile Roche Diagnostics discontinued the 454 platform in 2016, Illumina technology improved significantly in terms of higher sequencing throughput or longer reads, thus being able to re-sequence a human genome for less than $1,000 with the NovaSeq system in 2017 (Illumina, 2017) (table 1). Nevertheless, despite their immeasurable applications, NGS technologies present some drawbacks such as short read length and lack of portability (table 1). For instance, using only NGS-derived reads for the obtention of the primary sequence of a genome (genome assembly) may lead to incomplete, fragmented genomes due to their short length (Lee, *et al*. 2016).

In this context, the third-generation sequencing technologies arise as an alternative to produce highly accurate *de novo* assemblies and highly contiguous genome reconstructions. The first instrument developed under the umbrella of the third-generation sequencing technologies was commercially introduced in 2010 by Pacific Biosciences (PacBio; recently acquired by Illumina).

PacBio instruments are based on the SMRT (Single Molecule Real Time) technology, which allows real-time sequencing from uninterrupted template-directed synthesis using fluorescently-labelled nucleotides (Eid, *et al*. 2009). It is able to yield reads of several kilobases (up to 60 Kbp), overcoming the NGS read length of few hundreds (Eid, *et al*. 2009) (table 1).

Table 1. Overview of the most used next-generation and third-generation sequencing technologies. Table adapted from Besser, *et al*. (2017).

NGS / 3rd generation technologies	Throughput (Gb)	Read length (bp)	Strength	Weakness
Illumina MiSeq	0.3-15	2 x 300	- Read length (+) - Scalability	- Run length
Illumina NovaSeq	2000-6000	2 x 150	- Read accuracy - Throughput (++) - Low cost per sample	- High initial investment - Run length - Read length
Ion Torrent S5	0.6-15	Up to 400	- Read length (+) - Speed - Scalability	- Homopolymers*
Ion Torrent Proton	10-15	Up to 200	- Speed - Throughput (+)	- Homopolymers*
PacBio Sequel	5-10	Up to 60.000	- Read length (++) - Speed	- High error rate (+)
Oxford Nanopore MinION	0.1-1	Up to 100.000	- Read length (++) - Portability**	- High error rate (++) - Run length - Throughput

* Homopolymer: high error rate (insertions/deletions) on homopolymeric regions.
** Portability: very easy to be transported anywhere.

The also third-generation sequencing device MinION, released in 2015 by Oxford Nanopore technologies, also yields reads of several kilobases (up to 100 Kbp). Its size is as small as a pen drive, thus having high portability as it is wearable everywhere. Besides, it performs real-time sequencing, successfully used for Ebola surveillance (Quick, *et al*. 2016). However, although the third-generation technologies are very useful for genome assemblies due to their read length, they generate reads with higher error rates than the NGS technologies (Weirather, *et al*. 2017) (table 1).

As described above, both next-generation and third-generation technologies have advantages and drawbacks. However, one does not exclude the other, and the combination of both technologies increases the accuracy of some analyses, especially in genome assemblies (Koren, *et al*. 2012). In fact, the trend of using both technologies is increasing and the primary sequences of several genomes have been already obtained in this way, for example in gorilla or maize (Gordon, *et al*. 2016; Yinping, *et al*. 2017).

1.1.3 RNA sequencing: from its origins to expression atlas

The central dogma of molecular biology considers that DNA is transcribed into messenger RNA (mRNA) that, in turn, is codified to protein with the help of transfer RNA (tRNA) and ribosomal RNA (rRNA) (Crick, 1958). Like DNA, RNA is composed of nucleotides but, instead of thymine, this is replaced by uracil. The whole amount of RNA sequences in a cell is called the transcriptome and its study is crucial to obtain a better understanding of cell function, including the annotation of genes, their levels of expression, or the discovery of splicing variants and other post-transcriptional modifications.

As soon as the Sanger sequencing technology became available for DNA sequencing (Sanger, *et al.* 1977), it was subsequently applied for obtaining RNA sequences, with the first application in producing a rabbit muscle cDNA library (cloned cDNA fragments inserted into a collection of host cells) in 1983 (Scott, *et al*. 1983). These sequences are called "Expressed Sequence Tags" (ESTs) and consist of short sub-sequences of cDNA (Adams, *et al*. 1991). EST sequencing allowed the study of novel sequences and the discovery of gene structures along genome sequences. Gene annotation of the first draft of the human genome was performed using EST evidence, estimating between 30,000 and 35,000 protein-coding genes (representing 1.5% of the genome) (International Human Genome Sequencing Consortium, 2001). Despite its initial applications, EST sequencing presented several drawbacks, such as low throughput and qualitative results (not quantitative) due to the normalization that is normally applied during the process of cDNA library construction (Saccone and Pesole, 2003; Wang, *et al*. 2009). This, together with the high cost associated with this type of sequencing, motivated the use of DNA microarrays as a more affordable and quantitative alternative for the transcriptome study.

DNA microarrays are high-throughput technology based on the hybridization of fluorescently labelled cell-derived cDNA on high-density oligonucleotide microarrays. This technology was first described by Stephen Fodor and colleagues in 1991, who founded the microarray-specialized company Affymetrix one year later (Fodor, *et al*. 1991). The first use of DNA microarrays for gene expression analysis was carried out on *Arabidopsis thaliana* (Schena, *et al*. 1995) and soon after other model species followed, such as yeast (Shalon, *et al*. 1996). The potential of this technology was rapidly adopted in biomedicine as microarray-derived gene expression could be used to compare and classify different kinds of cancer (DeRisi, *et al*. 1996; Alon, *et al*. 1999; Golub, *et al*. 1999). The applications of DNA microarrays were diverse and include (i) gene expression analysis, (ii) transcription factor binding analysis, and (iii) genotyping (Bumgarner, 2013). But, since the oligonucleotide design for DNA microarrays relies on the

existing genomic sequences, this methodology was limited to model organisms with well-known transcriptomes. Moreover, reducing cross-hybridization was a challenge, since microarrays have high levels of background noise (Wang, *et al*. 2009).

The appearance of the NGS technologies in mid 2000s boosted the transcriptomics research along with DNA sequencing. In 2008, NGS technologies were also applied on the human transcriptome by means of RNA-sequencing (RNA-seq) (Morin, *et al*. 2008). This technique vanquished the limitations of EST sequencing and microarrays mentioned above in several ways. RNA-seq (i) is a high-throughput technique, (ii) does not rely on the existing genomic sequences, thus being able to predict gene structures (gene annotation) and to be applied on non-model organisms, and (iii) provides a wider dynamic range for gene quantification than microarrays (Wang, *et al*. 2009). These advances expedited the progress of ambitious projects, such as the Encyclopaedia of DNA Elements (ENCODE). ENCODE was launched by the US National Human Genome Research Institute (NHGRI) in September 2003 with the aim to uncover the role of the non-coding regions of the human genome. The appearance of NGS-derived techniques such as RNA-seq, among others, was used by ENCODE to boost the obtention of a catalogue of functional elements that includes, for instance, Transcription Start Sites (TSS), promoters, enhancers, nucleosome locations, or methylation sites. At its production phase, ENCODE concluded that 80.4% of the human genome participate in at least one biochemical RNA or chromatin associated event, challenging the initial concept that non-coding regions of the genome were "junk" DNA (The ENCODE Project Consortium, 2012). Consistent with the identification of functional elements on non-coding regions, pseudogenes, which were also described as "junk" DNA because of their coding capacity loss, were also revealed as true functional elements (Li, *et al*. 2013). In general, there is evidence that about 75% of human genome is transcribed (at least in some cell lines), being the coding regions a little minority (Djebali, *et al*. 2012).

Due to the advance in high-throughput sequencing, it is clear today that the central dogma of molecular biology needs to be revisited. In addition to tRNAs and rRNAs, new non-coding RNAs have been described and classified into small and long non-coding RNA (Barbosa Dogini, *et al*. 2014). On the one hand, small non-coding RNAs can be divided into different types: small interfering RNAs (siRNAs), small nucleolar RNAs (snoRNAs), small nuclear RNAs (snRNAs) and piwi-interacting RNAs (piRNAs) (Barbosa Dogini, *et al*. 2014). They are involved in a wide range of functions, such as splicing, gene regulation, transposon activity and RNA-editing (Barbosa Dogini, *et al*. 2014). On the other hand, long non-coding RNAs (lncRNA) are defined as non-translated molecules longer than 200 nucleotides with small or non-existing Open Reading

Frame (ORF). LncRNA are also involved in relevant cell pathways, including in chromatin remodelling, transcriptional control and post-transcriptional processing (Mercer, *et al*. 2009; Barbosa Dogini, *et al*. 2014; Dykes and Emanueli, 2017; Chen, *et al*. 2018).

In the human genome, the vast majority of lncRNAs are spliced with two to four exons, although lncRNAs from one to more than 10 exons have been also described. Besides, exons and introns of lncRNAs are slightly longer than those from coding genes (Derrien, *et al*. 2012). In terms of genomic location, lncRNAs can be either intergenic (they do not overlap any other gene) or can be overlapping an adjacent gene. This overlap can be either intronic or exonic. In humans, for example, the majority of lncRNAs are intergenic (Derrien, *et al*. 2012). Exon-overlapping transcripts at the opposite strand (antisense) are called Cis-Natural Antisense Transcripts (NATs) (Osato, *et al*. 2007). Also, their degree of conservation is lower compared to mRNAs, probably because secondary and tertiary RNA structure appears to be responsible of their wide variety of functions and regulatory roles (Mercer, *et al*. 2009; Johnsson, *et al*. 2014).

Overall, the study of the transcriptome has evolved dramatically since early 1980s due to the development of new technologies that deepened the knowledge of genes in terms of sequence, function, and expression. In this sense, on the one hand, single-cell RNA-seq technique is being increasingly adopted as it allows evaluating changes in individual cells (Tang, *et al*. 2009). On the other hand, initiatives such as the Expression Atlas, launched by the European Bioinformatics Institute (EBI), provides information about gene expression in different tissues and species (https://www.ebi.ac.uk/gxa) (Papatheodorou, *et al*. 2017). Currently, the Expression Atlas contains more than 110,000 assays from both DNA microarrays and RNA-seq across tens of species.

1.1.4 Towards the structural analyses of genomes

Genomes are not a matter of linear DNA sequences as they are highly organized and regulated inside the cell nucleus. The size and complexity of genomes require that their structural organization has direct implications in their function, as spatial gene positioning is intimately related with transcription, DNA replication and repair, as well as genome reshuffling (Farré, *et al*. 2015).

The first layer of organization of DNA is its own chemical modifications, specifically methylations on cytosines. DNA methylation is involved in regulating transcription, genomic imprinting, X chromosome inactivation, and inactivation of transposable elements (Jin, *et al*. 2011). But, beyond DNA methylation, DNA wraps around an octamer of histones (two each of H2A, H2B, H3, and H4) forming a nucleosome, which is the repeating unit of chromatin in eukaryotes

found approximately every 200 base pairs (Kornberg, 1974). Nucleosomes play an important role in regulating gene expression, as first suggested by Allfrey and Mirsky in 1964. Histones are susceptible to have chemical modifications that will either recruit non-histone proteins or loosen chromatin by disrupting nucleosome-nucleosome interactions, thus modulating DNA accessibility and transcription (Kouzarides, 2007). The field that studies these histone modifications as well as DNA methylation is called the **epigenome**, which is formally described as the study of changes in gene function that do not involve changes in the DNA sequence (Dupont, *et al*. 2009).

The most common histone modifications include acetylation, methylation, phosphorylation and ubiquitylation. The general picture is that certain histone modifications, such as H3K9ac, H3K27ac, H3K36me3, or H3K4me3, are known as markers for "open" chromatin conformations and, therefore, regions with active transcription (Barski, *et al*. 2007; Araki, *et al*. 2009). Other modifications, such as H3K27me3, or H3K9me3, are related to "closed" chromatin conformations and, therefore, regions where gene transcription is inactive (Barski, *et al*. 2007; Araki, *et al*. 2009). These signatures can be identified by NGS-derived Chromatin Immunoprecipitation coupled with sequencing (ChIP-seq) (Johnson, *et al*. 2007), an approach based on treating cells with formaldehyde to cross-link proteins to DNA, followed by cell disruption and sonication. DNA bound to the protein of interest is co-precipitated using a specific antibody. Subsequently, cross-links are reversed, and the purified DNA undergo sequencing (Landt, *et al*. 2012). Thus, ChIP-seq is indeed a powerful method for identifying genome-wide DNA binding sites for transcription factors and other proteins.

But gene expression is not only regulated by chemical modifications of DNA or histones on promoters or along gene bodies. Enhancers, silencers and insulators have been described as gene expression regulators acting in a spatial manner (Maston, *et al*. 2006). In fact, distal regulatory elements are connected to promoters and/or enhancers in a complex tri-dimensional (3D) network. Thus, the communication between widely spaced genomic elements is facilitated by the spatial organization of chromosomes that bring genes and their regulatory elements in close spatial proximity. The spatial organization of chromatin, also defined as the **nucleome** (Pennisi, 2015), has been initially studied by microscopy (Cremer and Cremer, 2001) and more recently by Chromosome Conformation Capture (3C)-based techniques (Dekker, *et al.*, 2002; Simonis, *et al*. 2006; Dostie, *et al*. 2006; Lieberman-Aiden, *et al*. 2009). The 3C technology was initially developed to reveal the spatial disposition of DNA between two chosen loci (one-versus-one method) (Dekker, *et al.*, 2002). In brief, this approach is based on treating cells with formaldehyde to cross-link nearby DNA segments, which in turn are cleaved with Restriction

Enzymes (REs) followed by a ligation step. The election of the RE determines the theoretically maximum resolution; for instance, 4-bp cutters (e.g. MboI) generate shorter fragments than 6-bp cutters (e.g. HindIII) (Lajoie, *et al*. 2015; Sati and Cavalli, 2017). Ligated fragments, which are detected by PCR, are chimeric: the ends of the fragments come from distinct interacting loci. Thus, if two distal sites on the DNA sequence form more ligation junction with each other than with other sequences, this indicates interaction *in vivo*.

Conformation capture techniques increased their throughput in the recent years thanks to improvements in the experimental procedures and NGS technologies (Schmitt, *et al*. 2016). These included: (i) Chromosome Conformation Capture on-Chip (4C) (Simons et al. 2006), (ii) Chromosome Conformation Capture Carbon Copy (5C) (Dostie, *et al*. 2006), (iii) ChIA-PET (a combination of ChIP and 3C) (Fullwood, *et al*. 2009), and (iv) Hi-C (Lieberman-Aiden et al. 2009) (figure 1).

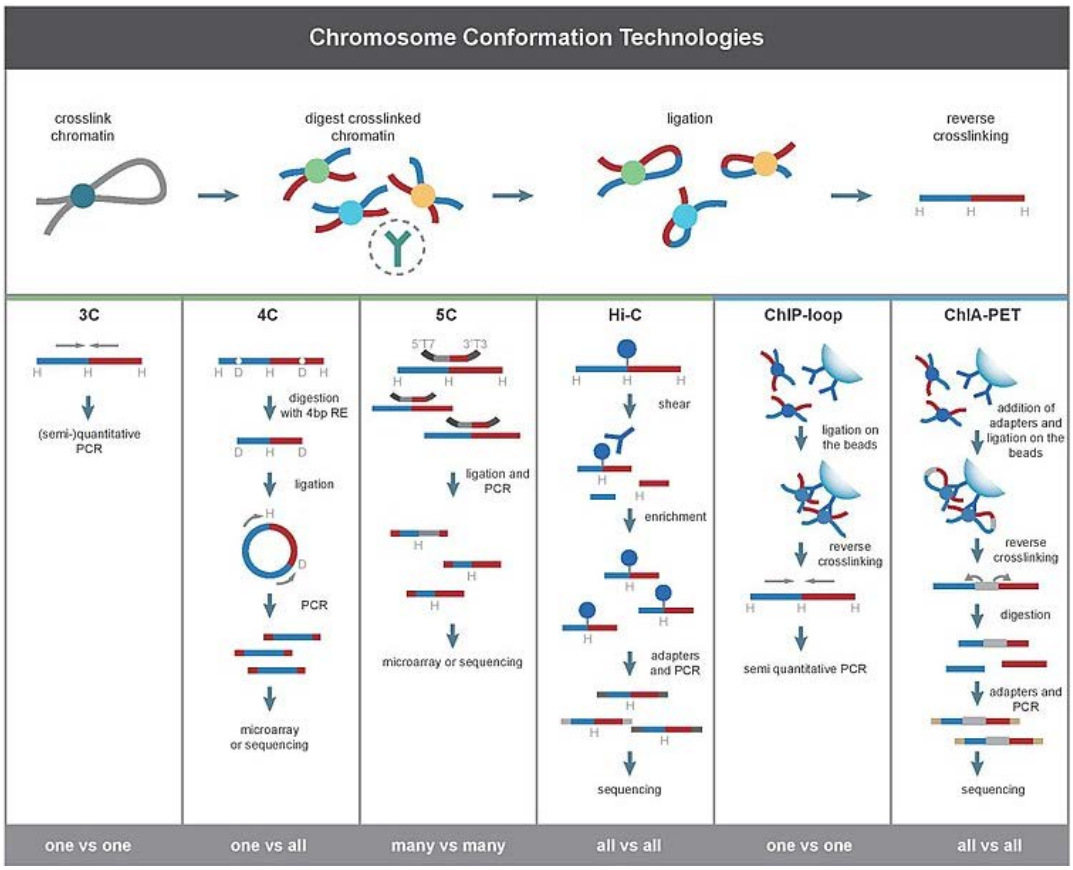


Figure 1. Comparison of different chromosome conformation capture methods. The first, horizontal panel shows the cross-linking, cleavage and ligation steps. Vertical panels show the specific steps for each technique. The vertical panels indicate the steps that are specific to separate methods. Figure extracted from de Wit and de Laat, (2012).

The 4C approach identifies the interaction of one locus with other loci (one-versus-all approach) (Simonis, *et al*. 2006), 5C could identifies the interaction of few loci with other few loci (many-

versus-many) (Dostie, *et al*. 2006). In both cases, 4C and 5C techniques, the approach initiates from a 3C library of fragments obtained after formaldehyde fixation, cleavage and ligation steps (figure 1).

A major breakthrough in the field was the development of the Hi-C technique (first applied on the human genome, Lieberman-Aiden, *et al*. 2009), which allowed the identification of any genome-wide contact between a pair of loci in a population of cells (all-versus-all approach) (figure 1). While 3C libraries usually contain both chimeric and unligated fragments, Hi-C libraries are enriched in chimeric fragments as it incorporates an additional step between cleavage and ligation: introduction of biotinylated nucleotides at ligation junctions which enables the specific purification of these junctions. By this way, biotinylated ligated fragments are pulled down and subsequently subject to NGS sequencing (Belton, *et al*. 2012). The cleavage and ligation steps take place after cell lysis. As a way to increase resolution and efficiency, the *in situ* Hi-C was recently described (Rao, *et al*. 2014). The main difference with previous approaches is that cleavage and ligation steps in intact nuclei before cell lysis. It has the following advantages relative to standard Hi-C: (i) reduction of spurious contacts due to random ligation, (ii) faster protocol, and (iii) higher resolution (Rao, *et al*. 2014).

The genome-wide applications of 3C techniques have served to provide insights into the spatial organization of chromatin, not only in human and mouse (Lieberman-Aiden, *et al*. 2009; Dixon, *et al*. 2012), but also in distant related species such as fruit fly (Sexton, *et al*. 2012), bacteria *Caulobacter crescentus* (Le, *et al*. 2013), fission yeast *Schizosaccharomyces pombe* (Mizuguchi, *et al*. 2014), nematode *C. elegans* (Crane, *et al*. 2015), or rice (Liu, *et al*. 2017). In this context, it becomes clear that functional activity of the genome can be determined by the association preferences of loci within each chromosome (Lieberman-Aiden, *et al*. 2009; Zhang, *et al*. 2012; Nagano, *et al*. 2013). Long before the development of conformation capture techniques, in the early 1900s, it was already known that chromosomes occupy specific locations in the nucleus called chromosome territories (CTs) (Cremer and Cremer, 2010). By using Fluorescence In Situ Hybridization (FISH) is was observed that gene-rich chromosomes tend to be located at the centre of the nucleus while gene-poor chromosomes tend to be at the nuclear periphery (Boyle, *et al*. 2001). The application of Hi-C in the human genome not only confirmed the existence of CTs, as it was seen gene-rich chromosomes preferentially interact, but also allowed the identification of new features of chromatin organization (Lieberman-Aiden, *et al*. 2009). In fact, this approach has revealed that genomes are compartmentalized into different levels of organization that include: (i) chromosomal territories, (ii) “open” (termed “A”)/”closed” (termed “B”) compartments inside chromosomal territories, (iii) Topologically Associated Domains

(TADs), and (iv) looping interactions (Dekker, *et al*. 2013; Phillips-Cremins, 2014; Rao, *et al*. 2014) (figure 2).

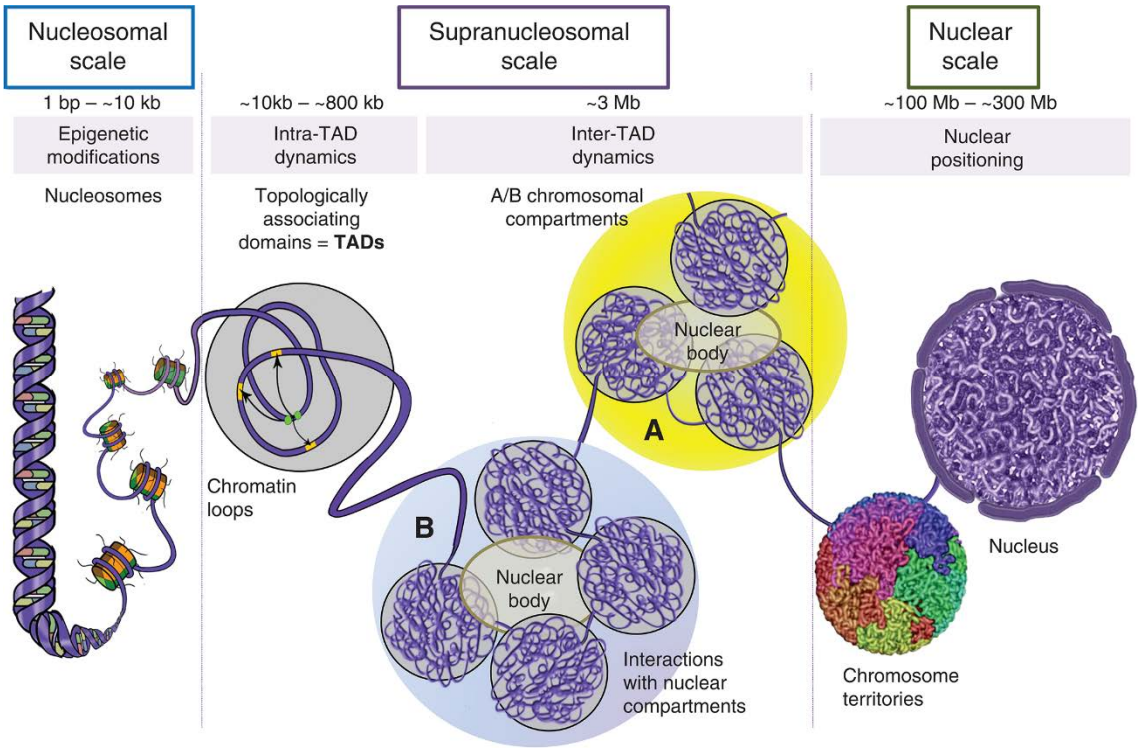


Figure 2. Overview of the genome organization from the nucleosomal scale to the nuclear scale. DNA wraps around octamers of histones forming nucleosomes (nucleosomal scale). Chromatin adopts a spatial conformation forming TADs, which are included into "open" A compartments or "closed" B compartments (supranucleosomal scale). Compartments are fractions of chromosomes, which are organized within the nucleus in chromosome territories (nuclear scale). Figure extracted from Golloshi, *et al*. (2017).

Compartments define the organization of the genome at the sub-chromosomal scale, thus they are the largest-scale position-specific interaction pattern detected by Hi-C (Lieberman-Aiden, *et al*. 2009). Genomic regions labelled as compartments "A" are gene-rich, actively transcribed, with higher chromatin accessibility and enriched with histone modifications related with "open" chromatin conformations; in contrast, compartments "B" are more densely packed genomic regions containing lower gene densities (Lieberman-Aiden, *et al*. 2009; Rao, *et al*. 2014). This compartmentalization is highly dynamic across cell types (Dixon, *et al*. 2015; Schmitt, *et al*. 2016). In 2014, Rao, *et al*. obtained for the human genome a very high resolution (1 Kbp) using by means of the *in situ* Hi-C. Considering this high resolution Hi-C experiment and other ChIP-seq marks, they revealed that compartment A might be split into 2 sub-compartments and compartment B split into 4 sub-compartments. On the one hand, the difference with subcompartments A1 and A2 is that subcompartment A1 finishes replicating earlier than A2. In addition, subcompartment A2 has lower GC content, longer genes, and is more strongly associated with H3K9me3 than A1. On the other hand, subcompartment B1 is associated with

facultative heterochromatin (↑H3K27me3); subcompartment B2 is enriched at both the nuclear lamina and the Nucleolus Associated Domains (NADs) and includes a great fraction of pericentromeric heterochromatin; subcompartment B3 is also enriched at the nuclear lamina but strongly depleted at NADs; finally, subcompartment B4 is highly enriched with genes from the KRAB-ZNF superfamily, the regions of which show a particular chromatin pattern with the presence of H3K36me3, H3K9me3 and H4K20me3 (Rao, *et al*. 2014).

At the sub-megabase scale, the Hi-C technique also revealed the existence of smaller organization domains within compartments, the so-called Topologically Associating Domains (TADs) (Sexton, *et al*. 2012; Dixon, *et al*. 2012). TADs are small domains that can be several kilobases (Kbp) in size, often containing distinct genes and multiple enhancers (Sexton, *et al*. 2012; Dekker and Heard, 2015). Thus, TADs can be defined as contiguous regions in which loci tend to interact more often with each other than with loci outside the TAD (Lajoie, *et al*. 2015). These small domains have been identified in a wide range of species from humans and mice (Dixon, *et al*. 2012) to Drosophila (Sexton, *et al*. 2012). Interestingly, TADs were not described in the *Arabidopsis thaliana* genome (Wang, *et al*. 2015). It has been reported that TAD boundaries in cultured human and mouse somatic cell lines are enriched with the transcription factor CTCF (also known as 11-zinc finger protein or CCCTC-binding factor, which has been detected in the majority of boundaries), cohesins (specifically RAD21 and SMC3, which have also been detected in the majority of boundaries), active transcription marks such as H3K4me3 and H3K36me3, nascent transcripts, and housekeeping genes, suggesting that they might contribute to TAD formation (Dixon, *et al*. 2012; Rao, *et al*. 2014; Sanborn, *et al*. 2015).

Studying the genome organization is not only important to understand gene regulation in somatic cells, but also to understand how chromosomes are transmitted during the formation of germ cells.

1.2 Bioinformatics

Methodological advances in DNA and RNA sequencing need to be coupled with the development of efficient analytical tools. Thus, bioinformatics can be defined as the use of computational tools to answer biological questions and to manage biological data. It represents an essential discipline that enables data analysis and identification of patterns in biological systems. Today, bioinformatics is routinely present in large-scale genetic studies (Russell, *et al*. 2018).

1.2.1 An historical view of the DNA sequencing data analysis

The birth of the term bioinformatics (generally synonymous with the term "computational biology") as currently known is tightly linked to the appearance of the Sanger sequencing technology (Hagen, 2000). In fact, shortly after the sequencing of the bacteriophage G4 genome (Godson, *et al*. 1978), computational algorithms were developed to compare this genome with the bacteriophage ϕX174, sequenced one year before (Sanger, *et al*. 1977). Later on, in 1979, different computational algorithms were developed mainly motivated by the need to establish homologies between genomes as they were sequenced and been available (Staden, *et al*. 1979).

As soon as the Sanger sequencing technology was adopted as an essential tool for biological research, the number of DNA fragments sequenced highly increased. This generated a need to develop public databases in order to decrease the cost of data acquisition, to speed up the dissemination of sequencing data, and to prevent the generation of duplicated data (Dayhoff, *et al*. 1981). In 1982, two currently used reference databases were released: (i) the Nucleotide Sequence Database (https://ebi.ac.uk/ena) supported by the EMBL (European Molecular Biology Laboratory) (Burks, *et al*. 1985) and (ii) the GenBank database (https://www.ncbi.nlm.nih.gov/genbank/) supported by the NIH (National Institute of Health, USA) (Hamm and Cameron, 1986). Currently, the Nucleotide Sequence Database from EMBL European Nucleotide Archive (ENA). GenBank and ENA collaborate between them by exchanging data on a daily basis. By January 21th of 2019, ENA included more than 2.207 million of sequences (ENA, 2019) (figure 3). In terms of complete genomes, GenBank currently contains more than 40 thousand genomes: 7,406 eukaryotic, 183,800 prokaryotic, 21,892 viral, 15,085 from plasmids, and 12,148 from organelles (GenBank, 2019).

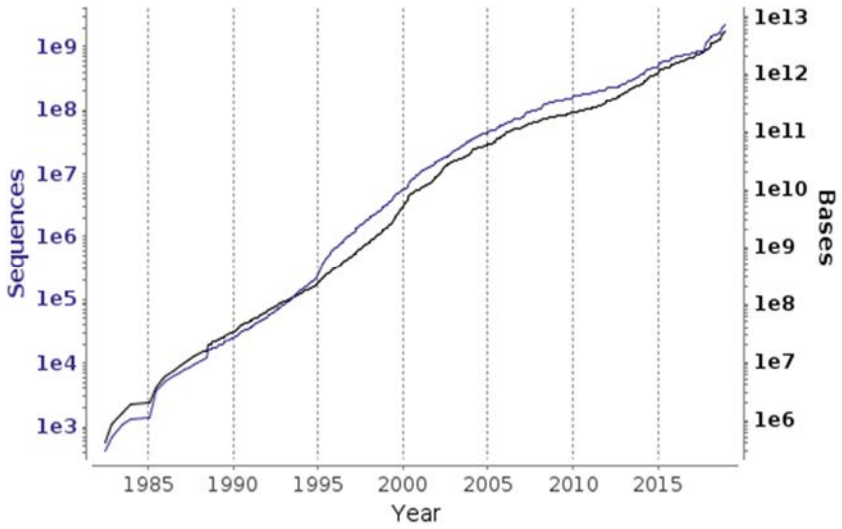


Figure 3. Growth of submitted sequences and base pairs to ENA between 1982 and 2018. Both vertical axes are in logarithmic scale. Figure extracted from the European Nucleotide Archive (2019).

Along with the increasing amount of sequences being generated and stored, there is the need to improve the performance of algorithms used for nucleotide search. In this sense, FASTA (FAST-All) was the first program (released in 1988) that was able to compare both protein and DNA sequences (Pearson and Lipman, 1988). In fact, the term FASTA is since then used to describe the file format of the same name, which is today the standard format to store sequences (box 1). Two years later, in 1990, the Basic Local Alignment Search Tool (BLAST) was developed, being orders of magnitude faster than FASTA (Altschul, *et al*. 1990). BLAST enables searches through the large number of DNA sequences that currently exist in the databases.

Box 1. **FASTA format.** It consists of a header starting with ">" that describes the sequence that will be found in the following lines below.

```
>NC_012920.1 Homo sapiens mitochondrion, complete genome
GATCACAGGTCTATCACCCTATTAACCACTCACGGGAGCTCTCCATGCATTTGGTATTTTCGTCTGGGGG
GTATGCACGCGATAGCATTGCGAGACGCTGGAGCCGGAGCACCCTATGTCGCAGTATCTGTCTTTGATTC
CTGCCTCATCCTATTATTTATCGCACCTACGTTCAATATTACAGGCGAACATACTTACTAAAGTGTGTTA
ATTAATTAATGCTTGTAGGACATAATAATAACAATTGAATGTCTGCACAGCCACTTTCCACACAGACATC
ATAACAAAAAATTTCCACCAAACCCCCCCTCCCCCGCTTCTGGCCACAGCACTTAAACACATCTCTGCCA
```

In the early 2000s, HGP resulted in an explosion in the number of raw sequences available from sequencing efforts. However, chromosomes needed to be assembled from the raw sequences. This purpose involves a process in which raw sequences are compared against themselves, thus aligning and merging fragments from a longer DNA sequence in order to reconstruct the original sequence. In this way, sequences grow longer as they are merged, ordered and oriented in order to assemble chromosomes. This process is known as genome assembly. In the HGP, it was performed using Phrap (Ewing, *et al*. 1998) and GigAssembler (Kent and Haussler, 2001).

Along with the release of the first draft of the human genome (see section 1.1.2), several relevant –and still active– bioinformatics projects were launched. That was the case of Bioconductor (https://www.bioconductor.org) (Gentleman, *et al*. 2004), Ensembl (https://www.ensembl.org) (Hubbard, *et al*. 2002), and the UCSC Genome browser (https://genome.ucsc.edu) (Kent, *et al*. 2002). In the context of the sequencing of the human and mouse genomes, databases such as Ensembl and the UCSC Genome Browser were created as online repositories in order to integrate and visually display the available genomic information, such as genome sequence, gene prediction, expression data, cross-species homologies or genetic variants, among others. Complementary, Bioconductor was developed not to be used as a repository of genomic data but to compile different methods and bioinformatics approaches to analyse genomic data for researchers with some bioinformatics background under the R programming language environment. In order to reach a more diverse audience, Galaxy (https://usegalaxy.org/) was further released in 2005 as an online platform

with the aim to allow researchers with no bioinformatics background to exploit the growing information available in public databases such as GenBank, ENA, Ensembl, the UCSC Genome Browser, among others (Giardine, *et al*. 2005).

The generation of millions of short reads by the massive use of NGS technologies motivated NCBI to develop in 2007 the Sequence Read Archive (SRA) in order to store and share all sequencing data available and made public (Wheeler, *et al*. 2007). However, the existing bioinformatics algorithms at that time were not prepared to handle millions of short sequences. In this context, several software were developed to address this problem, such as Velvet for genome assembly (Zerbino and Birney, 2008), Bowtie (Langmead, *et al*. 2009) and the Burrows-Wheeler Alignment tool (BWA) (Li and Durbin, 2009) for read mapping from genomic data, TopHat (Trapnell, *et al*. 2009) for read mapping from transcriptomic data, SAMtools (Li, *et al*. 2009) for manipulating mapping results, and the Genome Analysis Toolkit (GATK) (McKenna, *et al*. 2010) for Single Nucleotide Polymorphism (SNP) calling.

With the arrival of the third-generation sequencing technologies (see section 1.1.2), not only the spectacular increase in the read length but also the high sequencing error rate highlighted the need of new software adapted for these new technologies. That was the case of Falcon (Chin, *et al*. 2016) or Canu (Koren, *et al.* 2017), which were developed for genome assembly. Other existing tools, instead, were updated to be able work efficiently with long reads, such as BWA (from version 0.7.11 onwards) and the genome assembler Spades (from version 3.0 onwards) (Bankevich, *et al.* 2012). More recently, a new long-read mapper has been recently published, Minimap2 (Li, 2018), which claims to be much faster and more accurate than BWA. Therefore, big efforts have been made in bioinformatics in the last decade to improve the handling of the increasing amount of genomics data.

1.2.2 The bioinformatics bottleneck

Since the initial release of GenBank in 1982, the number of sequences stored has been growing exponentially, reaching more than 210 million sequences in December 2018 (GenBank, 2018) (figure 3). Likewise, the SRA database currently consist of more than 9 petabytes of raw sequencing data (SRA, 2019). As discussed in previous sections, this growth rate has been motivated by an improvement in the technology along with a continuous decrease of its costs. And not only specialized sequencing centres have access to high-throughput sequencing equipment; there are also several benchtop sequencers available in the market for smaller research institutions or individual laboratories. Illumina, for instance, has achieved an extremely high sequencing throughput, with longer reads and lower error rates than previous efforts (Liu,

et al. 2012; Schirmer, *et al*. 2016). The same company estimated in 2014 that the amount of sequencing data produced worldwide would double every 12 months. But neither the Moore's law nor the Illumina estimate was right: the trend reveals the amount of data doubles every 7 months approximately (Stephens, *et al*. 2015).

The production of large amounts of raw sequencing data has revealed the existence of some limitations in the scope of data accessibility, processing and interpretation, thus creating a bottleneck in bioinformatics. These limitations include three main aspects: (i) the development of new and more affordable alternatives to manage the current (and future) sequencing data flood, (ii) the acquisition of basic informatics and bioinformatics skills that would allow researchers to make use of their own data, thus ensuring feasible large-scale data analysis and management, and (iii) reproducibility.

1.2.2.1 Managing data acquisition, processing and sharing

Storage capacity and accessibility has become a challenge due to the amount of sequencing data being produced. In this sense, cloud computing systems might handle this problem (see section 1.2.3). In parallel, the amount of post-processing results generated also increases, thus becoming a challenge as they need to be organized and stored in an organized manner to allow their retrieval and to facilitate their accessibility to the research community (Schulz, *et al*. 2016). In fact, from early 2000s to 2016, in the Nucleic Acids Research journal, there has been a linear trend towards database development with a proliferation rate of 100 new databases a year (Imker, 2018) (figure 4).

In addition, available databases are introducing new and more affordable alternatives to manage big amounts of data. An intuitive Graphical User Interface (GUI) is utterly important for researchers interested in doing queries via web browser. However, at the time to retrieve big amounts of data in an easy and affordable way, an Application Programming Interface (API) is required (Helmy *et al*. 2016). Briefly, an API provides a way for sharing data in a structured way between software. For instance, a researcher working on the human gene BRCA2 can access to its annotated variants with few clicks through the Ensembl main page. In case this researcher wants to perform the same query several times, as the repository gets frequently updated, a bioinformatician could automate the process thus saving time. For this purpose, an API is needed; the bioinformatician would write a small program to connect to the Ensembl API (https://rest.ensembl.org/) to retrieve the list of annotated variants.

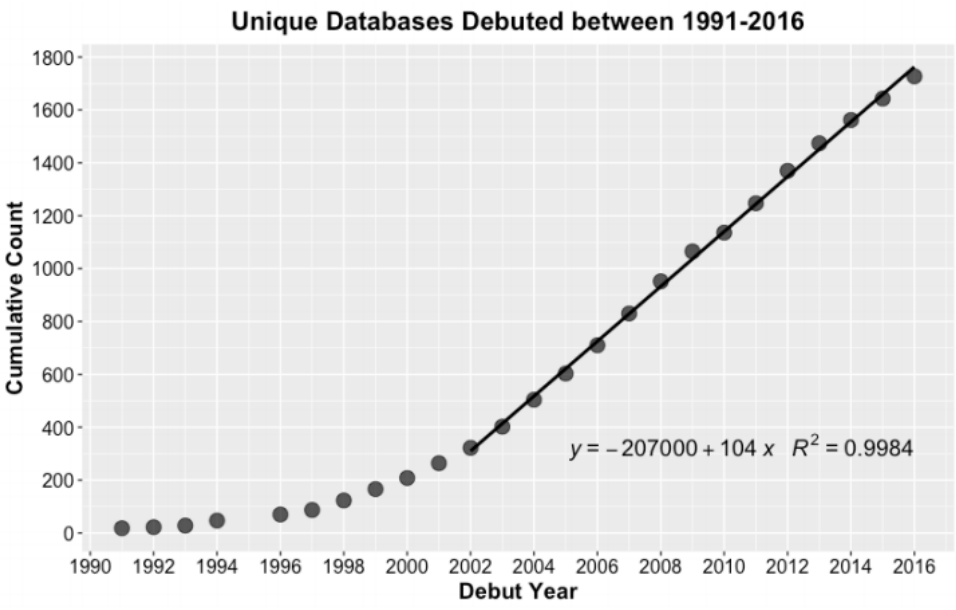


Figure 4. Cumulative count of new databases published in the Nucleic Acids Research (NAR) journal. Figure extracted from Imker, *et al*. (2018).

1.2.2.2 Developing the "know-how" in bioinformatics

The amount of sequencing data being produced is also requiring bioinformatics skills among the research community to be able to execute different software and informatics algorithms. These skills are mainly related to computer science and include: (i) being competent with the UNIX shell environment; (ii) previous knowledge of general programming languages such as Perl or Python; (iii) programming skills using languages specifically designed for queries, such as the Structured Query Language (SQL), for Relational Database Management Systems (RDBMS); (iv) knowledge of Version Control Systems (VCS) like Git to track changes done from version to version in a software and to share its code with the scientific community through platforms such as GitHub (https://github.com); (v) skills using available bioinformatics software in order to approach a biological problem with the best tool (Dudley and Butte, 2009).

There are different initiatives that have been tackling the need to acquire bioinformatics skills by promoting its learning. In this sense, the Rosalind project was born in 2012 as a platform for learning programming and bioinformatics through problem solving (http://rosalind.info). In addition, the European Bioinformatics Institute (EBI) is also well aware of this need: in 2017, EBI added 11 new courses and 38 webinars in its online training platform relative to 2016, besides of participating into another 341 training events (Cook, *et al*. 2019). In Europe, Barcelona is one of the most consolidated bioinformatics hubs (Biocat, 2017). In this region, over the last few years, 5 master's degrees and 1 bachelor's degree have been carried forward in Barcelona and surroundings (Bioinformatics Barcelona, 2018).

1.2.2.3 Reproducibility

Reproducibility (the replication of an analysis or experiment independently of the human resources and the geographical location) is another challenge in bioinformatics. Bioinformatics steps behind a publication are usually obscure as custom scripts and pipelines are not always shared. Besides, it has to be taken into account that software versions and manual data manipulation steps might affect the reproducibility of a bioinformatics analysis. In this sense, Sandve, *et al*. (2013) proposed 10 simple rules to increase reproducibility in bioinformatics analyses (table 2).

Table 2. **Rules for reproducible computational research** (Sandve, *et al*. 2013).

Rule	Description
1	For every result, keep track of how it was produced
2	Avoid manual data manipulation steps
3	Archive the exact versions of all external programs used
4	Version control all custom scripts
5	Record all intermediate results, when possible in standardized formats
6	For analyses that include randomness, note underlying random seeds
7	Always store raw data behind plots
8	Generate hierarchical analysis output, allowing layers of increasing detail to be inspected
9	Connect textual statements to underlying results
10	Provide public access to scripts, runs, and results

Precisely, following some of these rules would address the reproducibility problems mentioned so far: avoiding manual data manipulation steps to also avoid human errors (rule 2) or keeping track of software versions as well as the version of custom scripts and pipelines (rules 3 and 4).

Operating systems additionally affect the reproducibility of the experiments. As an example, Di Tommaso and collaborators carried out different bioinformatics analyses (gene annotation, transcript quantification and differential gene expression) using the same software version on Linux and Macintosh computers obtaining different results (Di Tommaso, *et al*. 2015; Di Tommaso, *et al*. 2017). Full reproducibility (same results in Linux and Macintosh) was finally obtained using Docker containers (Merkel, 2014), which work as lightweight operating systems with the needed software already installed inside.

In this context, the Bioinformatics and Genomics unit of the University of Torino identified very well this need to boost reproducibility in bioinformatics. They launched the initiative "Reproducible Bioinformatics" and provides several Docker images with ready-to-use pipelines following the rules proposed by Sandve, *et al*. (http://www.reproducible-bioinformatics.org/) (Kulkarni, *et al*. 2018).

1.2.3 Next generation bioinformatics

Traditionally, the most commonly used workflow in bioinformatics involves two main steps: (i) downloading the data obtained produced, and (ii) subsequently analyse the data in a powerful workstation or in an in-house server. However, this workflow assumes the availability of computational resources and bioinformatics skills for every research group or institution. How these steps are executed is relevant due to they can compromise the reproducibility of the analysis (see previous section 1.2.2). Thus, the traditional bioinformatics workflow might be replaced in the near future by the **next generation bioinformatics**, which could be defined as the field that brings the last Information Technology (IT) developments to bioinformatics with the aim to provide: (i) real-time data analysis solutions, (ii) fast access to the data, and (iii) efficient and interactive data visualization through web interface (de Brevern, *et al*. 2015).

Limitations in managing data acquisition and processing (see section 1.2.2) can be overcome by **cloud computing** systems (often referred as "the cloud"). Cloud-computing is defined as the technology that allows running application software and storing related data in central computer systems accessed through the Internet (Carr, 2009). Since the first description of its possible applications in the field of genomics, cloud computing systems represent a reliable alternative to traditional bioinformatics for several reasons (Dudley, *et al*. 2010; Stein, 2010). The main advantage that cloud computing can offer is scalability, as different instances (virtual servers) with customized resources (e.g. number of processors or memory available) can be opened to complete the analyses within a reasonable timeframe. Also, there is no need to perform initial capital investments, as cloud computing costs are variable by means of a fixed per-hour price (pay-per-use).

Cloud-based bioinformatics solutions are classified into four different categories (Dai, *et al.* 2012): (i) Data as a Service (DaaS), (ii) Software as a Service (SaaS), (iii) Platform as a Service (PaaS), and (iv) Infrastructure as a Service (IaaS). DaaS provides data on-demand through Internet, such as AWS Public Dataset (https://aws.amazon.com/es/opendata/public-datasets/) and Google Genomics Public Data (https://cloud.google.com/genomics/docs/public-datasets/). The difference between these sites and other databases such as GenBank is that AWS Public Dataset and Google Genomics Public Data are centralized sources of data, including archives of GenBank, Ensembl databases, data from the 1000 Genomes Project, among other repositories, that allow the access of the data in a standardized way. SaaS provides access to software, typically via web browser, for data analysis in the cloud (Rhyman, 2015). MG-RAST, for instance, is a SaaS for analysing metagenomics data (Meyer, *et al*. 2017). PaaS provides a platform where

bioinformaticians can deploy their own software and tools. Although Galaxy (Giardine, *et al*. 2005) or DNAnexus (https://dnanexus.com) are considered PaaS, they are also considered SaaS since any user can access to the deployed software for data analysis. Finally, IaaS provides virtualized resources such as virtual machines that contain ready-to-use software in it and can be executed on either cloud computing providers (e.g. Amazon EC2, Google Cloud or Microsoft Azure) or local workstation. CloudBioLinux is an example of IaaS (Krampis, *et al*. 2012).

The different "as a Service" categories described above bring the access to software and hardware more accessible to the scientific community. In this sense, cloud-based next generation bioinformatics is a promising alternative in terms of democratization of bioinformatics by facilitating the access to sequencing data analysis (Krampis and Wultsch, 2015). Specifically, SaaS platforms can solve the need to acquire powerful computers for data analysis, as computing is performed in the cloud, and partially solve the lack of bioinformatics skills due to the analysis is configured via web browser in a much nicer way than performing it via command line interface. However, current available SaaS platforms (e.g. DNAnexus or Galaxy) still require basic skills in bioinformatics at the very beginning of the process (defining specific software parameters) and at the very end (manipulating and transforming results for data interpretation and integration).

1.3 References

Adams, M. D., et al. "Complementary DNA Sequencing: Expressed Sequence Tags and Human Genome Project." *Science* , vol. 252, no. 5013, 1991, pp. 1651–56, doi:10.1126/SCIENCE.2047873.

Allfrey, V. G., Mirsky, A. E. "Structural Modifications of Histones and Their Possible Role in the Regulation of RNA Synthesis." *Science*, vol. 144, no. 3618, 1964, p. 559, doi:10.1126/science.144.3618.559.

Alon, U., et al. "Broad Patterns of Gene Expression Revealed by Clustering Analysis of Tumor and Normal Colon Tissues Probed by Oligonucleotide Arrays." *Proceedings of the National Academy of Sciences of the United States of America*, vol. 96, no. 12, 1999, pp. 6745–50, http://www.ncbi.nlm.nih.gov/pubmed/10359783.

Altschul, S. F., et al. "Basic Local Alignment Search Tool." *Journal of Molecular Biology*, vol. 215, no. 3, 1990, pp. 403–10, doi:10.1016/S0022-2836(05)80360-2.

Araki, Y., et al. "Genome-Wide Analysis of Histone Methylation Reveals Chromatin State-Based Regulation of Gene Transcription and Function of Memory CD8+ T Cells." *Immunity*, vol. 30, no. 6, 2009, pp. 912–25, doi:10.1016/j.immuni.2009.05.006.

Bankevich, A., et al. "SPAdes: A New Genome Assembly Algorithm and Its Applications to Single-Cell Sequencing." *Journal of Computational Biology*, vol. 19, no. 5, 2012, pp. 455–77, doi:10.1089/cmb.2012.0021.

Beck, E., et al. "Nucleotide Sequence of Bacteriophage Fd DNA." *Nucleic Acids Research*, vol. 5, no. 12, 1978, pp. 4495–503, http://www.ncbi.nlm.nih.gov/pubmed/745987.

Belton, J. M., et al. “Hi-C: A Comprehensive Technique to Capture the Conformation of Genomes.” *Methods*, vol. 58, no. 3, 2012, pp. 268–76, doi:10.1016/j.ymeth.2012.05.001.

Bentley, D. R., et al. “Accurate Whole Human Genome Sequencing Using Reversible Terminator Chemistry.” *Nature*, vol. 456, no. 7218, 2008, pp. 53–59, doi:10.1038/nature07517.

Berger, S. L. “The Complex Language of Chromatin Regulation during Transcription.” *Nature*, vol. 447, no. 7143, 2007, pp. 407–12, doi:10.1038/nature05915.

Besser, J., et al. “Next-Generation Sequencing Technologies and Their Application to the Study and Control of Bacterial Infections.” *Clinical Microbiology and Infection : The Official Publication of the European Society of Clinical Microbiology and Infectious Diseases*, vol. 24, no. 4, 2018, pp. 335–41, doi:10.1016/j.cmi.2017.10.013.

Boyle, S., et al. “The Spatial Organization of Human Chromosomes within the Nuclei of Normal and Emerin-Mutant Cells.” *Human Molecular Genetics*, vol. 10, no. 3, 2001, pp. 211–19, http://www.ncbi.nlm.nih.gov/pubmed/11159939.

Brosius, J., et al. “Complete Nucleotide Sequence of a 16S Ribosomal RNA Gene from Escherichia Coli.” *Proceedings of the National Academy of Sciences of the United States of America*, vol. 75, no. 10, 1978, pp. 4801–05, http://www.ncbi.nlm.nih.gov/pubmed/368799.

Bumgarner, R. “Overview of DNA Microarrays: Types, Applications, and Their Future.” *Current Protocols in Molecular Biology*, vol. Chapter 22, 2013, p. Unit 22.1., doi:10.1002/0471142727.mb2201s101.

Burks, C., et al. “The GenBank Nucleic Acid Sequence Database.” *Computer Applications in the Biosciences : CABIOS*, vol. 1, no. 4, 1985, pp. 225–33, http://www.ncbi.nlm.nih.gov/pubmed/3880345.

C. elegans Sequencing Consortium. “Genome Sequence of the Nematode C. Elegans: A Platform for Investigating Biology.” *Science*, vol. 282, no. 5396, 1998, pp. 2012–18, http://www.ncbi.nlm.nih.gov/pubmed/9851916.

Carr, N. "Cloud computing" *Encyclopaedia Britannica*, 2009, https://www.britannica.com/technology/cloud-computing

Chen, W. et al. “Comprehensive analysis of coding-lncRNA gene co-expression network uncovers conserved functional lncRNAs in zebrafish” *BMC Genomics* vol. 19,Suppl 2 112, 2018, doi:10.1186/s12864-018-4458-7

Chin, C. S., et al. “Phased Diploid Genome Assembly with Single-Molecule Real-Time Sequencing.” *Nature Methods*, vol. 13, no. 12, 2016, pp. 1050–54, doi:10.1038/nmeth.4035.

Consortium, Mouse Genome Sequencing. “Initial Sequencing and Comparative Analysis of the Mouse Genome.” *Nature*, vol. 420, no. 6915, 2002, pp. 520–62, doi:10.1038/nature01262.

Cook, C. E., et al. “The European Bioinformatics Institute in 2018: Tools, Infrastructure and Training.” *Nucleic Acids Research*, vol. 47, no. D1, 2019, pp. D15–22, doi:10.1093/nar/gky1124.

Crane, E., et al. “Condensin-Driven Remodelling of X Chromosome Topology during Dosage Compensation.” *Nature*, vol. 523, no. 7559, 2015, pp. 240–44, doi:10.1038/nature14450.

Cremer, T., Cremer, M. “Chromosome Territories.” *Cold Spring Harbor Perspectives in Biology*, vol. 2, no. 3, 2010, p. a003889, doi:10.1101/cshperspect.a003889.

Dai, L., et al. “Bioinformatics Clouds for Big Data Manipulation.” *Biology Direct*, vol. 7, 2012, p. 43; discussion 43, doi:10.1186/1745-6150-7-43.

Dayhoff, M. O., et al. “Nucleic Acid Sequence Database.” *DNA*, vol. 1, no. 1, 1981, pp. 51–58, doi:10.1089/dna.1.1981.1.51.

de Brevern, A. G., et al. “Trends in IT Innovation to Build a Next Generation Bioinformatics Solution to Manage and Analyse Biological Big Data Produced by NGS Technologies.” *BioMed Research International*, vol. 2015, 2015, p. 904541, doi:10.1155/2015/904541.

de Wit, E., de Laat, W. “A Decade of 3C Technologies: Insights into Nuclear Organization.” *Genes & Development*, vol. 26, no. 1, 2012, pp. 11–24, doi:10.1101/gad.179804.111.

Dekker, J. “Two Ways to Fold the Genome during the Cell Cycle: Insights Obtained with Chromosome Conformation Capture.” *Epigenetics & Chromatin*, vol. 7, no. 1, 2014, p. 25, doi:10.1186/1756-8935-7-25.

Dekker, J., Heard, E. “Structural and functional diversity of Topologically Associating Domains” *FEBS letters,* vol. 589, no. 20 Pt A, 2015, pp. 2877-1884, doi:10.1016/j.febslet.2015.08.044.

Dekker, J, et al. “Exploring the Three-Dimensional Organization of Genomes: Interpreting Chromatin Interaction Data.” *Nature Reviews Genetics*, vol. 14, no. 6, 2013, pp. 390–403, doi:10.1038/nrg3454.

Dekker, J., et al. “Capturing Chromosome Conformation.” *Science*, vol. 295, no. 5558, 2002, pp. 1306–11, doi:10.1126/science.1067799.

DeRisi, J., et al. “Use of a CDNA Microarray to Analyse Gene Expression Patterns in Human Cancer.” *Nature Genetics*, vol. 14, no. 4, 1996, pp. 457–60, doi:10.1038/ng1296-457.

Derrien, T., et al. “The GENCODE v7 Catalog of Human Long Noncoding RNAs: Analysis of Their Gene Structure, Evolution, and Expression.” *Genome Research*, vol. 22, no. 9, 2012, pp. 1775–89, doi:10.1101/gr.132159.111.

Di Tommaso, P., et al. “Nextflow Enables Reproducible Computational Workflows.” *Nature Biotechnology*, vol. 35, no. 4, 2017, pp. 316–19, doi:10.1038/nbt.3820.

Di Tommaso, P., et al. “The Impact of Docker Containers on the Performance of Genomic Pipelines.” *PeerJ*, vol. 3, 2015, p. e1273, doi:10.7717/peerj.1273.

Dixon, J. R., et al. “Chromatin Architecture Reorganization during Stem Cell Differentiation.” *Nature*, vol. 518, no. 7539, 2015, pp. 331–36, doi:10.1038/nature14222.

Dixon, J. R., et al. “Topological Domains in Mammalian Genomes Identified by Analysis of Chromatin Interactions.” *Nature*, vol. 485, no. 7398, 2012, pp. 376–80, doi:10.1038/nature11082.

Djebali, S., et al. “Landscape of Transcription in Human Cells.” *Nature*, vol. 489, no. 7414, 2012, pp. 101–08, doi:10.1038/nature11233.

Dogini, D. B., et al. “The New World of RNAs.” *Genetics and Molecular Biology*, vol. 37, no. 1 Suppl, Sociedade Brasileira de Genética, 2014, pp. 285–93, http://www.ncbi.nlm.nih.gov/pubmed/24764762.

Dostie, J., et al. “Chromosome Conformation Capture Carbon Copy (5C): A Massively Parallel Solution for Mapping Interactions between Genomic Elements.” *Genome Research*, vol. 16, no. 10, 2006, pp. 1299–309, doi:10.1101/gr.5571506.

Dudley, J. T., Butte, A. J. “A Quick Guide for Developing Effective Bioinformatics Programming Skills.” *PLoS Computational Biology*, vol. 5, no. 12, 2009, p. e1000589, doi:10.1371/journal.pcbi.1000589.

Dudley, J. T., et al. “Translational Bioinformatics in the Cloud: An Affordable Alternative.” *Genome Medicine*, vol. 2, no. 8, 2010, p. 51, doi:10.1186/gm172.

Dupont, C., et al. “Epigenetics: Definition, Mechanisms and Clinical Perspective.” *Seminars in Reproductive Medicine*, vol. 27, no. 5, 2009, pp. 351–57, doi:10.1055/s-0029-1237423.

Durbin, R. M., et al. “A Map of Human Genome Variation from Population-Scale Sequencing.” *Nature*, vol. 467, no. 7319, 2010, pp. 1061–73, doi:10.1038/nature09534.

Dykes, I. M., Costanza, E. “Transcriptional and Post-transcriptional Gene Regulation by Long Non-coding RNA” *Genomics, proteomics & bioinformatics,* vol. 15, no. 3, 2017, pp. 177-186, doi:10.1016/j.gpb.2016.12.005.

Eid, J., et al. “Real-Time DNA Sequencing from Single Polymerase Molecules.” *Science*, vol. 323, no. 5910, 2009, pp. 133–38, doi:10.1126/science.1162986.

ENA. "Statistics" *European Bioinformatics Institute (EMBL-EBI)*, 2019, https://www.ebi.ac.uk/ena/about/statistics

Ewing, B., Green, P. “Base-Calling of Automated Sequencer Traces Using Phred. II. Error Probabilities.” *Genome Research*, vol. 8, no. 3, 1998, pp. 186–94, http://www.ncbi.nlm.nih.gov/pubmed/9521922.

Ewing, B., et al. “Base-Calling of Automated Sequencer Traces Using Phred. I. Accuracy Assessment.” *Genome Research*, vol. 8, no. 3, 1998, pp. 175–85, http://www.ncbi.nlm.nih.gov/pubmed/9521921.

Fleischmann, R. D., et al. “Whole-Genome Random Sequencing and Assembly of Haemophilus Influenzae Rd.” *Science*, vol. 269, no. 5223, 1995, pp. 496–512, http://www.ncbi.nlm.nih.gov/pubmed/7542800.

Fodor, S. P., et al. “Light-Directed, Spatially Addressable Parallel Chemical Synthesis.” *Science*, vol. 251, no. 4995, 1991, pp. 767–73, http://www.ncbi.nlm.nih.gov/pubmed/1990438.

Fullwood, M. J., et al. “An Oestrogen-Receptor-α-Bound Human Chromatin Interactome.” *Nature*, vol. 462, no. 7269, 2009, pp. 58–64, doi:10.1038/nature08497.

GenBank. "GenBank and WGS Statistics" *National Center for Biotechnology Information (NCBI)*, 2018, https://www.ncbi.nlm.nih.gov/genbank/statistics/

GenBank. "Genome List - Genome" *National Center for Biotechnology Information (NCBI)*, 2019, https://www.ncbi.nlm.nih.gov/genome/browse#!/overview/

Gentleman, R. C., et al. “Bioconductor: Open Software Development for Computational Biology and Bioinformatics.” *Genome Biology*, vol. 5, no. 10, 2004, p. R80, doi:10.1186/gb-2004-5-10-r80.

Giardine, B., et al. “Galaxy: A Platform for Interactive Large-Scale Genome Analysis.” *Genome Research*, vol. 15, no. 10, 2005, pp. 1451–55, doi:10.1101/gr.4086505.

Gibbs, R. A., et al. “A Global Reference for Human Genetic Variation.” *Nature*, vol. 526, no. 7571, Oct. 2015, pp. 68–74, doi:10.1038/nature15393.

Godson, G. N., et al. “Nucleotide Sequence of Bacteriophage G4 DNA.” *Nature*, vol. 276, no. 5685, 1978, pp. 236–47, doi:10.1038/276236a0.

Goffeau, A., et al. “Life with 6000 Genes.” *Science*, vol. 274, no. 5287, 1996, pp. 546, 563–67, http://www.ncbi.nlm.nih.gov/pubmed/8849441.

Golloshi, R., et al. “Genome Organization during the Cell Cycle: Unity in Division.” *Wiley Interdisciplinary Reviews: Systems Biology and Medicine*, vol. 9, no. 5, 2017, p. e1389, doi:10.1002/wsbm.1389.

Golub, T. R., et al. “Molecular Classification of Cancer: Class Discovery and Class Prediction by Gene Expression Monitoring.” *Science*, vol. 286, no. 5439, 1999, pp. 531–37, doi:10.1126/SCIENCE.286.5439.531.

Gordon, D., et al. “Long-Read Sequence Assembly of the Gorilla Genome.” *Science*, vol. 352, no. 6281, 2016, p. aae0344, doi:10.1126/science.aae0344.

Guyer, M. S., Collins, F. S. “How Is the Human Genome Project Doing, and What Have We Learned so Far?” *Proceedings of the National Academy of Sciences of the United States of America*, vol. 92, no. 24, 1995, pp. 10841–48, http://www.ncbi.nlm.nih.gov/pubmed/7479895.

Hagen, J. B. “The Origins of Bioinformatics.” *Nature Reviews Genetics*, vol. 1, no. 3, 2000, pp. 231–36, doi:10.1038/35042090.

Hamm, G. H., Cameron, G. N. “The EMBL Data Library.” *Nucleic Acids Research*, vol. 14, no. 1, 1986, pp. 5–9, http://www.ncbi.nlm.nih.gov/pubmed/3945550.

Hasin, Y., et al. “Multi-Omics Approaches to Disease.” *Genome Biology*, vol. 18, no. 1, 2017, p. 83, doi:10.1186/s13059-017-1215-1.

Helmy, M., et al. “Ten Simple Rules for Developing Public Biological Databases.” *PLOS Computational Biology*, vol. 12, no. 11, 2016, p. e1005128, doi:10.1371/journal.pcbi.1005128.

Hubbard, T., et al. “The Ensembl Genome Database Project.” *Nucleic Acids Research*, vol. 30, no. 1, 2002, pp. 38–41, doi:10.1093/nar/30.1.38.

Illumina. "Press Release: Illumina Introduces the NovaSeq Series—a New Architecture Designed to Usher in the $100 Genome" *Illumina*, 2017, https://www.illumina.com/company/news-center/press-releases/press-release-details.html?newsid=2236383

Imker, H. J. “25 Years of Molecular Biology Databases: A Study of Proliferation, Impact, and Maintenance.” *Frontiers in Research Metrics and Analytics*, vol. 3, 2018, p. 18, doi:10.3389/frma.2018.00018.

Jiao, Y., et al. “Improved Maize Reference Genome with Single-Molecule Technologies.” *Nature*, vol. 546, no. 7659, 2017, p. 524, doi:10.1038/nature22971.

Jin, B., et al. “DNA Methylation: Superior or Subordinate in the Epigenetic Hierarchy?” *Genes & Cancer*, vol. 2, no. 6, 2011, pp. 607–17, doi:10.1177/1947601910393957.

Johnson, D. S., et al. “Genome-Wide Mapping of in Vivo Protein-DNA Interactions.” *Science*, vol. 316, no. 5830, 2007, pp. 1497–502, doi:10.1126/science.1141319.

Kent, W. J., Haussler, D. “Assembly of the Working Draft of the Human Genome with GigAssembler.” *Genome Research*, vol. 11, no. 9, 2001, pp. 1541–48, doi:10.1101/gr.183201.

Kent, W. J., et al. “The Human Genome Browser at UCSC.” *Genome Research*, vol. 12, no. 6, 2002, pp. 996–1006, doi:10.1101/gr.229102.

Koren, S., et al. “Hybrid Error Correction and de Novo Assembly of Single-Molecule Sequencing Reads.” *Nature Biotechnology*, vol. 30, no. 7, 2012, pp. 693–700, doi:10.1038/nbt.2280.

Kornberg, R. D. “Chromatin Structure: A Repeating Unit of Histones and DNA.” *Science*, vol. 184, no. 4139, 1974, pp. 868–71, http://www.ncbi.nlm.nih.gov/pubmed/4825889.

Kouzarides, T. “Chromatin Modifications and Their Function.” *Cell*, vol. 128, no. 4, 2007, pp. 693–705, doi:10.1016/j.cell.2007.02.005.

Krampis, K., et al. “Cloud BioLinux: Pre-Configured and on-Demand Bioinformatics Computing for the Genomics Community.” *BMC Bioinformatics*, vol. 13, no. 1, 2012, p. 42, doi:10.1186/1471-2105-13-42.

Krampis, K., Wultsch, C. “A Review of Cloud Computing Bioinformatics Solutions for Next-Gen Sequencing Data Analysis and Research.” *Methods in Next Generation Sequencing*, vol. 2, no. 1, 2015, doi:10.1515/mngs-2015-0003.

Kulkarni, N., et al. “Reproducible Bioinformatics Project: A Community for Reproducible Bioinformatics Analysis Pipelines.” *BMC Bioinformatics*, vol. 19, no. S10, 2018, p. 349, doi:10.1186/s12859-018-2296-x.

Lajoie, B. R., et al. “The Hitchhiker’s Guide to Hi-C Analysis: Practical Guidelines.” *Methods*, vol. 72, 2015, pp. 65–75, doi:10.1016/j.ymeth.2014.10.031.

Lander, E. S., et al. “Initial Sequencing and Analysis of the Human Genome.” *Nature*, vol. 409, no. 6822, 2001, pp. 860–921, doi:10.1038/35057062.

Landt, S. G., et al. “ChIP-Seq Guidelines and Practices of the ENCODE and ModENCODE Consortia.” *Genome Research*, vol. 22, no. 9, 2012, pp. 1813–31, doi:10.1101/gr.136184.111.

Langmead, B., et al. “Ultrafast and Memory-Efficient Alignment of Short DNA Sequences to the Human Genome.” *Genome Biology*, vol. 10, no. 3, 2009, p. R25, doi:10.1186/gb-2009-10-3-r25.

Le, T. B. K., et al. “High-Resolution Mapping of the Spatial Organization of a Bacterial Chromosome.” *Science*, vol. 342, no. 6159, 2013, pp. 731–34, doi:10.1126/science.1242059.

Lee, H., et al. “Third-Generation Sequencing and the Future of Genomics.” *BioRxiv*, 2016, p. 048603, doi:10.1101/048603.

Li, H., Durbin, R. “Fast and Accurate Short Read Alignment with Burrows-Wheeler Transform.” *Bioinformatics*, vol. 25, no. 14, 2009, pp. 1754–60, doi:10.1093/bioinformatics/btp324.

Li, H. “Minimap2: Pairwise Alignment for Nucleotide Sequences.” *Bioinformatics*, vol. 34, no. 18, 2018, pp. 3094–100, doi:10.1093/bioinformatics/bty191.

Li, H., et al. “The Sequence Alignment/Map Format and SAMtools.” *Bioinformatics*, vol. 25, no. 16, 2009, pp. 2078–79, doi:10.1093/bioinformatics/btp352.

Lieberman-Aiden, E., et al. “Comprehensive Mapping of Long-Range Interactions Reveals Folding Principles of the Human Genome.” *Science*, vol. 326, no. 5950, 2009, pp. 289–93, doi:10.1126/science.1181369.

Liu, C., et al. “Prominent Topologically Associated Domains Differentiate Global Chromatin Packing in Rice from Arabidopsis.” *Nature Plants*, vol. 3, no. 9, 2017, pp. 742–48, doi:10.1038/s41477-017-0005-9.

Liu, L., et al. “Comparison of Next-Generation Sequencing Systems.” *Journal of Biomedicine and Biotechnology*, vol. 2012, 2012, pp. 1–11, doi:10.1155/2012/251364.

Margulies, M., et al. “Genome Sequencing in Microfabricated High-Density Picolitre Reactors.” *Nature*, vol. 437, no. 7057, 2005, pp. 376–80, doi:10.1038/nature03959.

Maston, G. A., et al. “Transcriptional Regulatory Elements in the Human Genome.” *Annual Review of Genomics and Human Genetics*, vol. 7, no. 1, 2006, pp. 29–59, doi:10.1146/annurev.genom.7.080505.115623.

McKenna, A., et al. “The Genome Analysis Toolkit: A MapReduce Framework for Analyzing next-Generation DNA Sequencing Data.” *Genome Research*, vol. 20, no. 9, 2010, pp. 1297–303, doi:10.1101/gr.107524.110.

McReynolds, L., et al. “Sequence of Chicken Ovalbumin MRNA.” *Nature*, vol. 273, no. 5665, 1978, pp. 723–28, doi:10.1038/273723a0.

Mercer, T. R., et al. “Long Non-Coding RNAs: Insights into Functions.” *Nature Reviews Genetics*, vol. 10, no. 3, 2009, pp. 155–59, doi:10.1038/nrg2521.

Merkel, D. “Docker: Lightweight Linux Containers for Consistent Development and Deployment.” *Linux Journal*, vol. 2014, no. 239, 2014, https://dl.acm.org/citation.cfm?id=2600241.

Meyer, F., et al. “MG-RAST Version 4—lessons Learned from a Decade of Low-Budget Ultra-High-Throughput Metagenome Analysis.” *Briefings in Bioinformatics*, 2017, doi:10.1093/bib/bbx105.

NHGRI. "2001 Release: First Analysis of Human Genome" *National Human Genome Research Institute (NHGRI)*, 2001, https://www.genome.gov/10002192/2001-release-first-analysis-of-human-genome/

Mizuguchi, T., et al. “Cohesin-Dependent Globules and Heterochromatin Shape 3D Genome Architecture in S. Pombe.” *Nature*, vol. 516, no. 7531, 2014, pp. 432–35, doi:10.1038/nature13833.

Moore, G. E. “Cramming More Components onto Integrated Circuits.” *Electronics*, vol. 38, no. 8, 1965, pp. 114.

Morin, R. D., et al. “Profiling the HeLa S3 Transcriptome Using Randomly Primed CDNA and Massively Parallel Short-Read Sequencing.” *BioTechniques*, vol. 45, no. 1, 2008, pp. 81–94, doi:10.2144/000112900.

Nagano, T., et al. “Single-Cell Hi-C Reveals Cell-to-Cell Variability in Chromosome Structure.” *Nature*, vol. 502, no. 7469, 2013, pp. 59–64, doi:10.1038/nature12593.

Nielsen, R., et al. “Genotype and SNP Calling from Next-Generation Sequencing Data.” *Nature Reviews. Genetics*, vol. 12, no. 6, 2011, pp. 443–51, doi:10.1038/nrg2986.

Osato, N., et al. “Transcriptional Interferences in Cis Natural Antisense Transcripts of Humans and Mice.” *Genetics*, vol. 176, no. 2, 2007, pp. 1299–306, doi:10.1534/genetics.106.069484.

Papatheodorou, I., et al. “Expression Atlas: Gene and Protein Expression across Multiple Studies and Organisms.” *Nucleic Acids Research*, vol. 46, no. D1, 2018, pp. D246–51, doi:10.1093/nar/gkx1158.

Pennisi, E. “Genomics. Inching toward the 3D Genome.” *Science*, vol. 347, no. 6217, 2015, p. 10, doi:10.1126/science.347.6217.10.

Phillips-Cremins, J. E., et al. “Architectural Protein Subclasses Shape 3D Organization of Genomes during Lineage Commitment.” *Cell*, vol. 153, no. 6, 2013, pp. 1281–95, doi:10.1016/j.cell.2013.04.053.

Powell, A., et al. “Disciplinary Baptisms: A Comparison of the Naming Stories of Genetics, Molecular Biology, Genomics, and Systems Biology.” *History and Philosophy of the Life Sciences*, vol. 29, 2007, pp. 5–32, doi:10.2307/23334194.

Putney, S. D., et al. “A New Troponin T and CDNA Clones for 13 Different Muscle Proteins, Found by Shotgun Sequencing.” *Nature*, vol. 302, no. 5910, 1983, pp. 718–21, doi:10.1038/302718a0.

Quick, J., et al. “Real-Time, Portable Genome Sequencing for Ebola Surveillance.” *Nature*, vol. 530, no. 7589, 2016, pp. 228–32, doi:10.1038/nature16996.

Rao, S. S. P., et al. “A 3D Map of the Human Genome at Kilobase Resolution Reveals Principles of Chromatin Looping.” *Cell*, vol. 159, no. 7, 2014, pp. 1665–80, doi:10.1016/j.cell.2014.11.021.

Rhyman, M. “Five Reasons Why Switching To SaaS Will Be The Best Investment You Make This Year.” *Forbes*, Forbes Magazine, 2017, www.forbes.com/sites/forbestechcouncil/2017/05/15/five-reasons-why-switching-to-saas-will-be-the-best-investment-you-make-this-year/#1852178636dc.

Rothberg, J. M., et al. “An Integrated Semiconductor Device Enabling Non-Optical Genome Sequencing.” *Nature*, vol. 475, no. 7356, 2011, pp. 348–52, doi:10.1038/nature10242.

Russell, P. H., et al. “A Large-Scale Analysis of Bioinformatics Code on GitHub.” *PloS One*, vol. 13, no. 10, 2018, p. e0205898, doi:10.1371/journal.pone.0205898.

Ryba, T., et al. “Evolutionarily Conserved Replication Timing Profiles Predict Long-Range Chromatin Interactions and Distinguish Closely Related Cell Types.” *Genome Research*, vol. 20, no. 6, 2010, pp. 761–70, doi:10.1101/gr.099655.109.

Saccone, C., Pesole G. *Handbook of Comparative Genomics: Principles and Methodology*. Wiley-Liss, 2003.

Sanborn, A. L., et al. “Chromatin Extrusion Explains Key Features of Loop and Domain Formation in Wild-Type and Engineered Genomes.” *Proceedings of the National Academy of Sciences of the United States of America*, vol. 112, no. 47, 2015, pp. E6456-65, doi:10.1073/pnas.1518552112.

Sanger, F., et al. “Nucleotide Sequence of Bacteriophage ΦX174 DNA.” *Nature*, vol. 265, no. 5596, 1977, pp. 687–95, doi:10.1038/265687a0.

Sanger, F., et al. “DNA Sequencing with Chain-Terminating Inhibitors.” *Proceedings of the National Academy of Sciences of the United States of America*, vol. 74, no. 12, 1977, pp. 5463–67, http://www.ncbi.nlm.nih.gov/pubmed/271968.

Sati, S., Cavalli, G. “Chromosome Conformation Capture Technologies and Their Impact in Understanding Genome Function.” *Chromosoma*, vol. 126, no. 1, 2017, pp. 33–44, doi:10.1007/s00412-016-0593-6.

Schena, M., et al. “Quantitative Monitoring of Gene Expression Patterns with a Complementary DNA Microarray.” *Science*, vol. 270, no. 5235, 1995, pp. 467–70, http://www.ncbi.nlm.nih.gov/pubmed/7569999.

Schena, M., et al. “Parallel Human Genome Analysis: Microarray-Based Expression Monitoring of 1000 Genes.” *Proceedings of the National Academy of Sciences of the United States of America*, vol. 93, no. 20, 1996, pp. 10614–19, http://www.ncbi.nlm.nih.gov/pubmed/8855227.

Schirmer, M., et al. “Illumina Error Profiles: Resolving Fine-Scale Variation in Metagenomic Sequencing Data.” *BMC Bioinformatics*, vol. 17, 2016, p. 125, doi:10.1186/s12859-016-0976-y.

Schmitt, A. D., et al. “Genome-Wide Mapping and Analysis of Chromosome Architecture.” *Nature Reviews Molecular Cell Biology*, vol. 17, no. 12, 2016, pp. 743–55, doi:10.1038/nrm.2016.104.

Schmitt, A. D., et al. “A Compendium of Chromatin Contact Maps Reveals Spatially Active Regions in the Human Genome.” *Cell Reports*, vol. 17, no. 8, 2016, pp. 2042–59, doi:10.1016/j.celrep.2016.10.061.

Schulz, W. L., et al. “Evaluation of Relational and NoSQL Database Architectures to Manage Genomic Annotations.” *Journal of Biomedical Informatics*, vol. 64, 2016, pp. 288–95, doi:10.1016/J.JBI.2016.10.015.

Sexton, T., et al. “Three-Dimensional Folding and Functional Organization Principles of the Drosophila Genome.” *Cell*, vol. 148, no. 3, 2012, pp. 458–72, doi:10.1016/j.cell.2012.01.010.

Shalon, D., et al. “A DNA Microarray System for Analyzing Complex DNA Samples Using Two-Color Fluorescent Probe Hybridization.” *Genome Research*, vol. 6, no. 7, 1996, pp. 639–45, http://www.ncbi.nlm.nih.gov/pubmed/8796352.

Shendure, J., et al. “DNA Sequencing at 40: Past, Present and Future.” *Nature*, vol. 550, no. 7676, 2017, pp. 345–53, doi:10.1038/nature24286.

Simonis, M., et al. “Nuclear Organization of Active and Inactive Chromatin Domains Uncovered by Chromosome Conformation Capture–on-Chip (4C).” *Nature Genetics*, vol. 38, no. 11, 2006, pp. 1348–54, doi:10.1038/ng1896.

Smith, L. M., et al. “Fluorescence Detection in Automated DNA Sequence Analysis.” *Nature*, vol. 321, no. 6071, 1986, pp. 674–79, doi:10.1038/321674a0.

SRA. "Documentation - SRA database growth" *National Center for Biotechnology Information (NCBI)*, 2019, https://www.ncbi.nlm.nih.gov/sra/docs/sragrowth/

Stein, L. D. “The Case for Cloud Computing in Genome Informatics.” *Genome Biology*, vol. 11, no. 5, 2010, p. 207, doi:10.1186/gb-2010-11-5-207.

Stephens, Z. D., et al. “Big Data: Astronomical or Genomical?” *PLoS Biology*, vol. 13, no. 7, 2015, p. e1002195, doi:10.1371/journal.pbio.1002195.

Tang, F., et al. “MRNA-Seq Whole-Transcriptome Analysis of a Single Cell.” *Nature Methods*, vol. 6, no. 5, 2009, pp. 377–82, doi:10.1038/nmeth.1315.

Trapnell, C., et al. “TopHat: Discovering Splice Junctions with RNA-Seq.” *Bioinformatics*, vol. 25, no. 9, 2009, pp. 1105–11, doi:10.1093/bioinformatics/btp120.

Venter, J. C., et al. “The Sequence of the Human Genome.” *Science*, vol. 291, no. 5507, 2001, pp. 1304–51, doi:10.1126/science.1058040.

Wang, C., et al. “Genome-Wide Analysis of Local Chromatin Packing in Arabidopsis Thaliana.” *Genome Research*, vol. 25, no. 2, 2015, pp. 246–56, doi:10.1101/gr.170332.113.

Wang, Z., et al. “RNA-Seq: A Revolutionary Tool for Transcriptomics.” *Nature Reviews. Genetics*, vol. 10, no. 1, Jan. 2009, pp. 57–63, doi:10.1038/nrg2484.

Watson, J. D., Crick, F. H. C. “Molecular Structure of Nucleic Acids: A Structure for Deoxyribose Nucleic Acid.” *Nature*, vol. 171, no. 4356, 1953, pp. 737–38, doi:10.1038/171737a0.

Weirather, J. L., et al. “Comprehensive Comparison of Pacific Biosciences and Oxford Nanopore Technologies and Their Applications to Transcriptome Analysis.” *F1000Research*, vol. 6, 2017, p. 100, doi:10.12688/f1000research.10571.2.

Welford, S. M., et al. “Detection of Differentially Expressed Genes in Primary Tumor Tissues Using Representational Differences Analysis Coupled to Microarray Hybridization.” *Nucleic Acids Research*, vol. 26, no. 12, 1998, pp. 3059–65, http://www.ncbi.nlm.nih.gov/pubmed/9611255.

Wheeler, D. L., et al. “Database Resources of the National Center for Biotechnology Information.” *Nucleic Acids Research*, vol. 36, no. Database issue, 2008, pp. D13-21, doi:10.1093/nar/gkm1000.

Xi, R., et al. “Detecting Structural Variations in the Human Genome Using next Generation Sequencing.” *Briefings in Functional Genomics*, vol. 9, no. 5–6, 2010, pp. 405–15, doi:10.1093/bfgp/elq025.

Yadav, S. P. “The Wholeness in Suffix -Omics, -Omes, and the Word Om.” *Journal of Biomolecular Techniques : JBT*, vol. 18, no. 5, 2007, p. 277, http://www.ncbi.nlm.nih.gov/pubmed/18166670.

Zerbino, D. R., Birney, E. “Velvet: Algorithms for de Novo Short Read Assembly Using de Bruijn Graphs.” *Genome Research*, vol. 18, no. 5, 2008, pp. 821–29, doi:10.1101/gr.074492.107.

Zhang, Y., et al. “Spatial Organization of the Mouse Genome and Its Role in Recurrent Chromosomal Translocations.” *Cell*, vol. 148, no. 5, 2012, pp. 908–21, doi:10.1016/j.cell.2012.02.002.

Chapter 2: Objectives

Recent advances from large-scale sequencing projects have resulted in the generation of an increasing amount of sequence data available for the research community. This revolution has motivated the development of efficient ways to integrate, organize and interpret -omics data. That will, in the long term, allow us to understand the essential principles underlying the transmission and expression of genetic information. Is in this context where we formulate the present work, which took place within the framework of a collaborative project between Aurora Ruiz-Herrera's research group from Universitat Autònoma de Barcelona and Sequentia Biotech S.L. under the Industrial Doctorates Programme from the Generalitat de Catalunya.

The **main aim** of this work is to develop and integrate next generation bioinformatics tools for the analysis of the functional and structural characteristics of genomes. The **specific objectives** are the following:

1. **Develop a public online database for lncRNAs**

   Storage capacity and accessibility of genomics and transcriptomics data has become a challenge in the recent years. In order to overcome current limitations on the existing databases, here we develop a comprehensive online database for plant lncRNAs.

2. **Development and application of a "Software as a Service" (SaaS) platform for the high-throughput analysis of RNA-seq data**

   Since one of the majors constrains in bioinformatics is the availability of user-friendly tools to ensure feasible large-scale data analysis and management here we develop and validate a new SaaS platform called AIR (Artificial Intelligence RNA-seq).

3. **Analyse the structural organization of the mouse genome germ line derived from Hi-C data**

   We made use of AIR and additional bioinformatics tools to generate an integrative atlas of the chromatin interactions and functional genomic characteristics of the mouse male germ line.

Chapter 3: Development of a database for lncRNAs in plant genomes

3.1 Introduction

The recent realization that genomes are made of large portions of non-coding challenged the central dogma of molecular biology (The ENCODE Project Consortium, 2012). This pivotal discovery boosted the launch of large genome-sequencing projects, such as the GENCODE project (part of the ENCODE project), which aimed to characterize and annotate all gene features in the human and mouse genome (Harrow, *et al*. 2012). Since a big portion of the human transcriptome has no apparent coding capacity, the study of non-coding RNA, especially long non-coding RNA (lncRNA), has become an emerging field (Carninci, *et al*. 2005; Mattick and Makunin, 2006; Mattick, 2009; Djebali, et al. 2012). In human, lncRNAs have been extensively studied to such an extent that, in 2012, a catalogue of 14.880 lncRNAs was released (Derrien, *et al*. 2012). Currently, the last human GENCODE annotation (version 29) contains 29.566 lncRNAs.

Although studies on lncRNA are by far more advanced in human and mice, recent studies in plants are highlighting the importance of lncRNA on relevant cell functions, such as transcriptional regulation and control of gene expression (Franco-Zorrilla, *et al*. 2007; Swiezewski, *et al*. 2009; Heo and Sung, 2011; Ding, *et al*. 2012; Shin and Chekanova, 2014; Gai, *et al*. 2018; Liu, *et al*. 2018). It is widely known that plant evolution is characterized by a large number of Whole Genome Duplications (WGD) and gene duplications, which are the origin of many pseudogenes (Clark and Donoghue, 2018). Besides, a large proportion of plant genomes are enriched in retrotransposons (Zou, *et al*. 2009). Therefore, since it is though that lncRNAs are derived from pseudogenes and retrotransposons (Milligan and Lipovich, 2014; Ganesh and Svoboda, 2016), this class of non-coding RNA might have a considerable importance in plants.

So far, only few lncRNAs have been functionally characterized in plants, highlighting the potential interest of lncRNAs in plant biology and in regulating important agronomic traits. This is the case of *IPS1*, *COLDAIR*, *COOLAIR*, *LDMAR* or *Enod40*, among others, which are implicated in a diversity of essential functions. *IPS1*, for instance, is a lncRNA expressed in *Arabidopsis thaliana* upon phosphate starvation and it is thought to counteract the activity of miR399 on *PHO2*, which in turn regulates the expression of phosphate transporter genes (Franco-Zorrilla, *et al*. 2007). *COLDAIR*, on the other hand it has been shown that may play a role in recruiting the histone methylase *PRC2* to interact with the PRC2 complex, so maintaining a stable silenced state of *FLC* to repress flowering during vernalization (Heo and Sung, 2011). *COOLAIR* and *ASL* are other *A. thaliana* lncRNAs that regulate the FLC expression (Swiezewski, *et al*. 2009; Shin

and Chekanova, 2014). In rice (*Oryza sativa*), the lncRNA *LDMAR* has been found to control photo-sensitive male sterility by regulating DNA methylation levels in the promoter region of *LDMAR* (Ding, *et al*. 2012). In *Medicago truncatula*, the lncRNA *Enod40* has been shown to participate in establishing symbiotic interactions with soil-bacteria by affecting nodule formation (Campalans, *et al*. 2004). Most recently, in mulberry (*Morus multicaulis*), the lncRNA *MuLnc1* has been associated with both biotic and abiotic stress through RNA interference (RNAi) (Gai, *et al*. 2018). Finally, the lncRNA *TWISTED LEAF* in rice is thought to affect the leaf morphological development by cis-regulating the gene *OsMYB60* via natural antisense mechanism (Liu, *et al*. 2018).

To gain further insights on the role of lncRNAs in plant biology, their comprehensive annotation is critical. Several genome-wide studies have been performed in several plant species (Boerner and McGinnis, 2012; Lin, *et al*. 2014; Shuai, *et al*. 2014; Flórez-Zapata, *et al*. 2016; Joshi, *et al*. 2016; Jain, *et al*. 2017), so there is a need to store and share this information, for instance, by means of online databases. Although many plant lncRNA databases currently exist (Chen, *et al*. 2012; Jin, *et al*. 2013; Xie, *et al*. 2014; Yi, *et al*. 2015; Quek, *et al*. 2015; Xuan, *et al*. 2015; Szczessniak, *et al*. 2016), all of them present important drawbacks that need to be addressed. These include lack of APIs and maintenance, higher GUI friendliness desirable, few species or lncRNAs available (table 3).

Table 3. Overview of the most relevant databases for plant lncRNAs.

Name	Publication details	Species	Number of lncRNAs	Observations
PlantNATsDB	Chen, *et al*. (2012)	70 plant species	> 2,000,000 (NATs*)	Lack of API; NATs != lncRNAs; GUI friendliness (++)
PLncDB	Jin, *et al*. (2013)	*A. thaliana*	> 13,000	Down (11/2018)
NONCODE v5	Xie, *et al*. (2014)	*A. thaliana*	> 3,000	Lack of API; not plant specific; GUI friendliness (++)
PNRD	Yi, *et al*. (2015)	4 plant species	> 5,000	Lack of API; GUI friendliness (+)
lncRNAdb v2.0	Quek, *et al*. (2015)	8 plant species	> 10	API available; functional lncRNAs; not plant specific; GUI friendliness (++)
PLNlncRbase	Xuan, *et al*. (2015)	45 plant species	> 1,000	Down (11/2018)
CANTATAdb v1.0	Szczessniak, *et al*. (2016)	10 species	> 45,000	Lack of API; GUI friendliness (+++)
CANTATAdb v2.0	None	39 species	> 239,000	Lack of API; GUI friendliness (+++)

** NATs: Natural Antisense Transcripts.*

With the aim to overcome current limitations on existing databases, the main objective of this work was to develop a well-suited online database of *in silico* predicted plant lncRNAs from a

wide range of plant species. This new database is called Green Non-Coding (GreeNC) database and represents one of the most comprehensive databases of lncRNAs available for the scientific community, thus becoming a meeting point for the plant lncRNA research.

3.2 Methods

3.2.1 Data source

A total of 45 publicly available transcriptomes were downloaded in FASTA format from Phytozome (Goodstein, *et al.* 2012) and consisted of 39 plant species and 6 algae species (Tuskan, *et al.* 2006; Merchant, *et al.* 2007; Ouyang, *et al.* 2007; Palenik, *et al.* 2007; Jaillon, *et al.* 2007; Rensing, *et al.* 2008; Ming, *et al.* 2008; Paterson, *et al.* 2009; Worden, *et al.* 2009; Schnable, *et al.* 2009; Huang, *et al.* 2009; Chan, *et al.* 2010; Vogel, *et al.* 2010; Schmutz, *et al.* 2010; Prochnik, *et al.* 2010; Velasco, *et al.* 2010; Hu, *et al.* 2011; Shulaev, *et al.* 2011; Xu, *et al.* 2011; Young, *et al.* 2011; Banks, *et al.* 2011; Paterson, *et al.* 2012; Blanc, *et al.* 2012; Lamesch, *et al.* 2012; Bennetzen, *et al.* 2012; Sato, *et al.* 2012; Prochnik, *et al.* 2012; Wang, *et al.* 2012; DePamphilis, *et al.* 2013; Verde, *et al.* 2013; Droc, *et al.* 2013; Zimmer, *et al.* 2013; Slotte, *et al.* 2013; Motamayor, *et al.* 2013; Yang, *et al.* 2013; Hellsten, *et al.* 2013; Wu, *et al.* 2014; Schmutz, *et al.* 2014; Wang, *et al.* 2014; International Wheat Genome Sequencing Consortium, 2014; Bartholomé, *et al.* 2015).

3.2.2 Identification of lncRNA

Two scripts were written to identify lncRNAs (see section 3.2.2.1). Then, lncRNAs were divided into high- and low- confidence groups (see section 3.2.2.2).

3.2.2.1 Pipeline design

The first script followed the approach developed at the McGinnis lab to identify lncRNAs in transcriptomes and is based on identifying the coding potential of each transcript and on similarity with known proteins (Boerner and McGinnis, 2012). The script retains transcripts longer than 200 nucleotides and with an ORF shorter than 120 aa by using Ugene (version 1.13) (http://ugene.unipro.ru/). Sequences were then BLASTxed (BLAST version 2.2.28+) (ftp://ftp.ncbi.nlm.nih.gov/blast/executables/LATEST) against SwissProt (2013/11) (UniProt Consortium, 2015). CPC (version 0.9-r2) (http://cpc.cbi.pku.edu.cn/) (Kong, *et al.* 2007) was also used, with the FrameFinder parameter "-r" set to "True" or "False" and the BLASTX parameter "-S" set to "3" or "1", depending on the group of transcripts being analysed.

The second script was written to discriminate other non-coding transcripts from lncRNAs and to identify possible miRNA precursors. Transcripts were analysed by cmscan (Infernal version

1.1rc4) against the RFAM database (release 11) (Nawrocki, *et al*. 2015). In addition, BLASTn (version 2.2.28+) was used against a database of mature plant miRNAs from miRBase (release 20) (Griffiths-Jones, 2010) and the putative miRNA coordinates were validated by MIReNA (version v2.0) (http://www.lcqb.upmc.fr/mirena/). Finally, MIReNA was called again, using the parameters "–valid", "–x", "–mfei -0.69", "–amfe –32", "–ratiomin 0.83", and "–ratiomax 1.17".

RepeatMasker (version open-4.0.5) (http://www.repeatmasker.org/) was used for repetitive element identification with the parameters: "-species Viridiplantae", "-no_is", "-gff", and "-nolow". The search engine used was RMBLAST (version 2.2.23+) against the RepBase database (released 31 January 2014) (Bao, *et al*. 2015).

3.2.2.2 lncRNA classification

The final set of lncRNAs was divided into two different groups: high-confidence and low-confidence. High-confidence lncRNAs included the transcripts without hits against SwissProt, described as non-coding by CPC, and considered non-precursors of miRNA. Otherwise, low-confidence lncRNAs included transcripts with hits against SwissProt, described as coding by CPC, or considered precursors of miRNA. Transcripts having predicted repetitive regions by RepeatMasker were also classified as low-confidence in order to exclude putative transposons.

3.2.3 Benchmark

The first script developed for the annotation of lncRNAs was tested with 480 lncRNAs and 1,268 coding genes annotated in *Arabidopsis thaliana* (TAIR10 genome version) resulting in a sensitivity of 92% and a specificity of 94.95%. The second script was tested with the 480 lncRNAs from *A. thaliana*, resulting in a sensitivity of 93% and a specificity of 97.6%.

3.2.4 Database structure

Data was imported into a MySQL (version 5.5) based relational database stored in an Ubuntu server (version 14.04). This database was then integrated into a MediaWiki by mapping relational data fields against wiki-predefined templates via Semantic MediaWiki. The use of templates facilitates printing information and styling it for different page types (e.g. genes and species). The template approach exposes the fields to the query system of Semantic MediaWiki, enhancing the search possibilities of the site. All transcript sequences were kept in a FASTA file with the same IDs as in the MySQL, and then formatted using makeblastdb (BLAST version 2.2.28+). In this way, sequences can be retrieved using their ID with blastdbcmd (BLAST version 2.2.28+) and, at the same time, other BLAST programs can be run against the resulting BLAST

database. Taking advantage of this, an Express NodeJS API web service was created to expose sequence retrieval, lncRNA information and BLAST searches.

3.3 Results

GreeNC includes approximately 175,000 gene pages with information on more than 200,000 transcripts (63% classified as high-confidence lncRNA) from 39 plant species and 6 algae. All information can be accessed through a graphical interface using any web browser or can be programmatically accessed via RESTful API.

3.3.1 Graphical interface

3.3.1.1 Main page

GreeNC is available at the following link: http://greenc.sciencedesigners.com. This address brings the user to the main page of the database (figure 5).

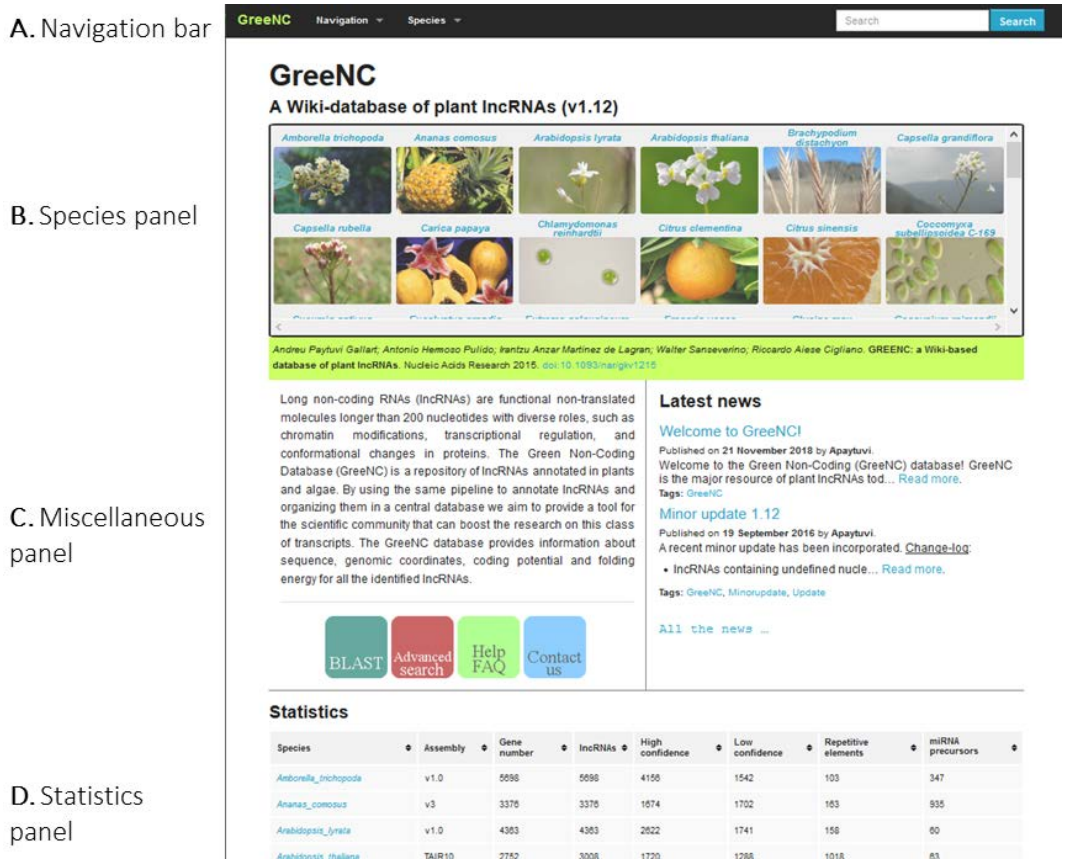


Figure 5. Screenshot of the main page in GreeNC. Four sections are highlighted: navigation bar (A), species panel (B), miscellaneous panel (C) and statistics panel (D).

The main page is divided into 4 different sections:

- Navigation bar. There is a black bar at the top of the web page with 2 drop-down menus and a search box. The first drop-down menu is called Navigation and allows the access to other sections of the database. The second drop-down menu lists every available species in GreeNC. Finally, the search box allows the search of any species or gene (figure 5A).
- Species panel. This panel contains a picture for each available species in order to access to any of them in a fast and visual way (figure 5B).
- Miscellaneous panel. This panel stores a general description of the database. Below this description there are 4 buttons that allow fast access to other sections of the database, such as BLAST, advanced search, Frequently Asked Questions (FAQ) or a page to contact the authors. To the right part of this panel there is a list of news created by the maintainers (figure 5C).
- Statistics panel. This panel shows a table with statistics about the genome assembly version, the number of genes and of lncRNAs per species (figure 5D).

3.3.1.2 Species page

From the navigation bar (second drop-down menu) or from the species panel it is possible to access to a species page. GreeNC is hierarchically organized into species pages, and under the species page anyone can access to gene pages from the corresponding species. The species page contains two different sections:

- Species title and description. This section contains the name of the species with its associated picture, synonyms of the species name, information about the used genome version and links that points to the corresponding NCBI Taxonomy page and to the FASTA file of the lncRNAs for that species (figure 6A).
- Gene list. This section contains a table showing genes that transcribe lncRNAs. This table also contains the chromosome, start and end positions of the gene and the number of lncRNAs it transcribes (figure 6B).

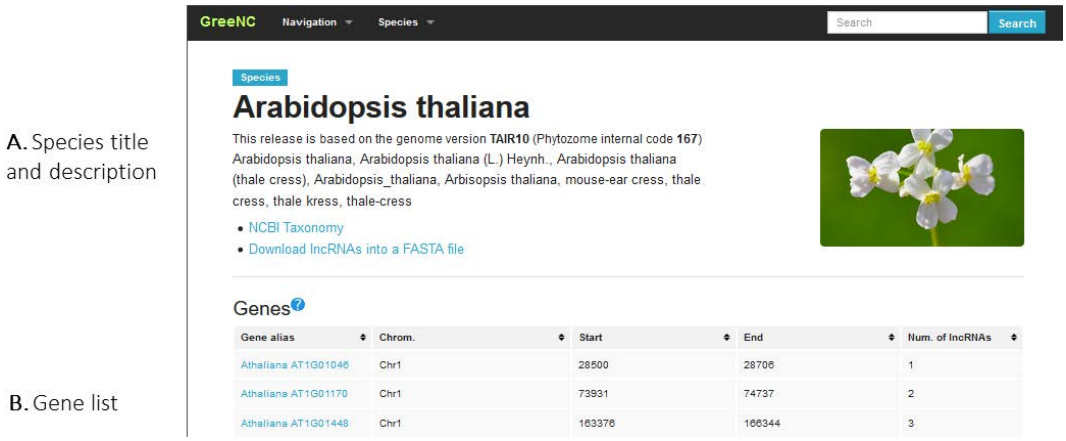


Figure 6. Screenshot of the *Arabidopsis thaliana* page in GreeNC. Two sections are highlighted: species title and description (A) and gene list (B).

3.3.1.3 Gene page

The gene page contains four sections (figure 7):

- **Gene information**. A table shows the gene name and alias, its coordinates, the database source and assembly, the species the gene comes from and whether this gene also transcribes coding transcripts or not (figure 7A).
- **Transcript features**. A table shows all lncRNAs the gene transcribes. Each row is relative to a lncRNA and displays its confidence (high or low), whether the lncRNA might be a miRNA precursor or not, its length, its sequence, and other features such as its ORF, coding potential, folding energies or GC content (figure 7B).
- **Matches to external databases**. This table displays whether the lncRNAs have matches to miRBase, Rfam, Swissprot, RepBase or NONCODE. It includes the database version, the hit name linked to the corresponding web page in the reference database, and the e-value (figure 7C).
- **Gene model**. This section shows a picture of the gene model: the axis that shows the coordinates of the gene, the gene feature and all transcripts being associated with it. Coding transcripts are drawn in magenta while high-confidence lncRNAs are drawn in green and low-confidence lncRNAs are drawn in cyan (figure 7D).

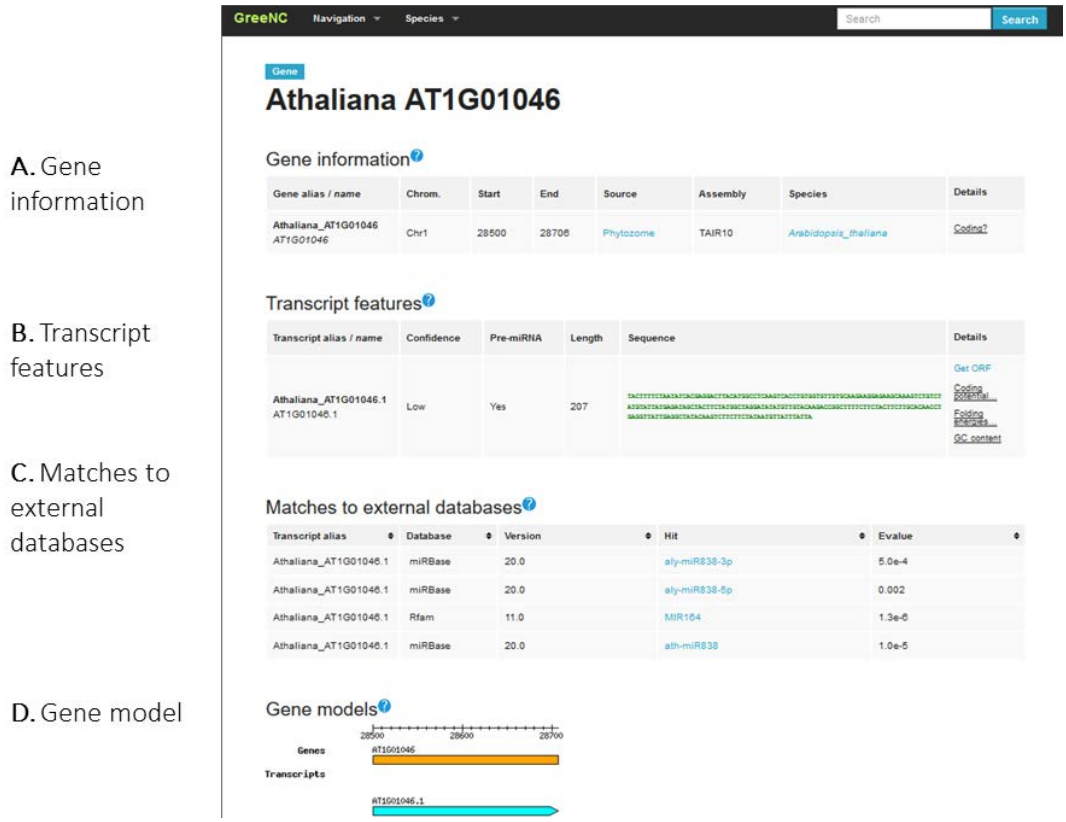


Figure 7. Screenshot of the gene page of an *A. thaliana*'s gene in GreeNC. Four sections are highlighted: gene information (A), transcript features (B), matches to external databases (C) and gene model (D).

3.3.1.4 Advanced search page

The *Advanced search* section can be accessed via the *Miscellaneous panel* from the main page of from the first drop-down menu in the navigation bar. There are three different query options to choose in this page:

- **Query by gene information**. This query option allows the user to filter the lncRNAs by species, coordinates and whether the genes transcribing lncRNAs also transcribe coding transcripts. The data can be also downloaded in a Comma Separated Values (CSV) table by selecting csv in the output option (figure 8).
- **Query by transcript feature**. This query option allows the user to filter the lncRNAs by species, confidence, length, coding potential, folding energies (AMFE and MFEI) or GC content. The data can be also downloaded in a CSV table by selecting csv in the output option.
- **Query by transcript matches to external databases**. This query option allows the user to filter the lncRNAs by species, hit to database, hit name and e-value. The data can be also downloaded in a CSV table by selecting csv in the output option.

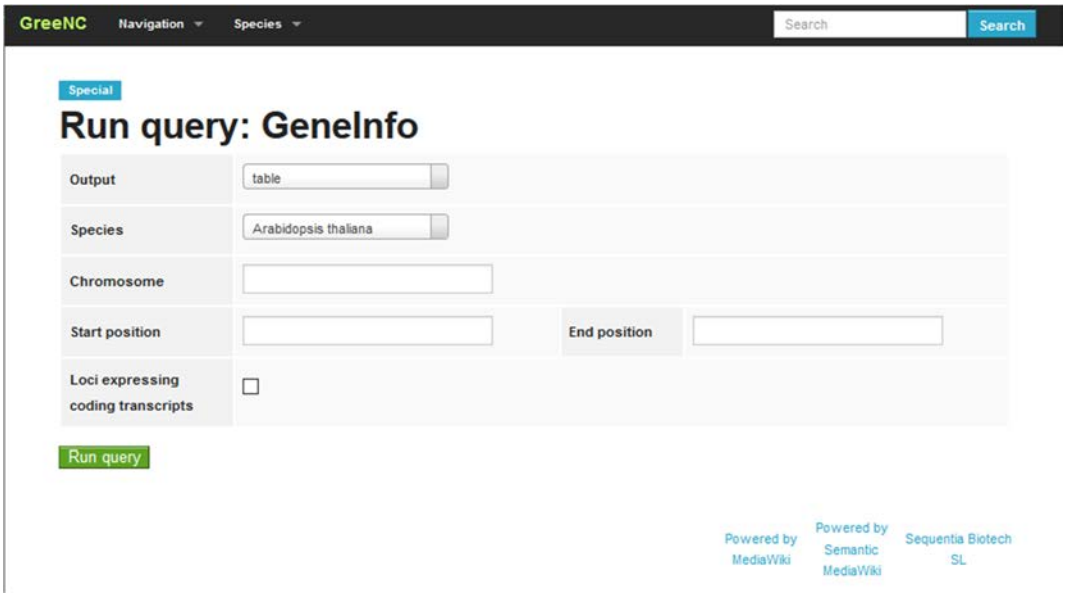


Figure 8. Screenshot of the advanced search page for query by gene information in GreeNC. It represents a form where the user can fill with search parameters.

3.3.1.5 Miscellaneous page

The user might also be interested in searching lncRNAs based on the similarity of some sequences. For this reason, we also made available a BLAST page where the user is able to BLASTn some sequence against the whole GreeNC database. The BLAST section can be accessed via the *Miscellaneous panel* from the main page of from the first drop-down menu in the navigation bar.

3.3.2 Programmatic access

The graphical interface is always a nice way to see the results. However, it fails when there is a need to retrieve the information through programming scripts. In order to overcome this limitation, GreeNC incorporates a RESTful (Representational State Transfer) API that provides the information via HTTP (Hypertext Transfer Protocol) GET requests.

It can be accessed under /api/ (http://greenc.sciencedesigners.com/api/) location. It contains 4 different resources: i) available databases (section 3.3.2.1, ii) available species (section 3.3.2.2), iii) transcripts information (section 3.3.2.3), and BLAST queries (section 3.3.2.4).

3.3.2.1 Available databases

The “db” function shows the list of the available BLAST databases in GreeNC. Currently, we only have a unique database containing all lncRNAs and its name is *greenc* (box 2).

Box 2. Bash command example to retrieve the list of databases available.

```
$ curl http://greenc.sciencedesigners.com/api/db/
{"nucl":{"greenc":{"path":"/home/ubuntu/db/lncrna/lncRNAall.fa"}}}
```

Once the user has chosen the database, the sequences of specific lncRNAs can be retrieved by adding the database name after /db/ followed by /entry/, the transcript alias, and /fasta (box 3).

Box 3. Bash command example to retrieve the sequence of a specific lncRNA in JSON format.

```
$ curl
http://greenc.sciencedesigners.com/api/db/greenc/entry/Athaliana_AT1G01170.1/fasta
{"def":">lcl|Athaliana_AT1G01170.1:1-
515","seq":"ACGACCGTCTTCCACCGTTGAATTCTTCTGGAACTGGAGTCCACTGTTTAAGCTTCACTGTCTCTGAATCGGC
AAAGCTT\nTAGAAGAAAATGGCATCAGGAGGTAAAGCCAAGTACATAATCGGTGCTCTCATCGGTTCTTTCGGAATCTCATACA
TCTT\nCGACAAAGTTATCTCTGATAATAAGATCTTTGGAGGGACTACTCCAGGAACTGTCTCTAACAAAGAATGGTGGGCAGCA
A\nCGGATGAGAAATTCCAAGCATGGCCAAGAACCGCTGGTCCTCCCGTTGTTATGAATCCCATTAGCCGTCAGAATTTCATC\n
GTCAAGACTCGTCCGGAATGAGAAAATAATAAGTTCAATGCTTTGATTTTCAGAATAAGATGAACGATGACGATGTTTTC\nTAA
ATCCGAGCTTGTACTAAATAACAATACATTACAACACGGTTTGCGGAACTACTCCACAGTCTATCTTCTGTTAAAAA\nACTCAA
ACAAGCTATTGCAAAAAGCCCTTACGAGA"}
```

The output is in JSON format and the sequence contains breaklines every 80 bases. FASTA format can be also retrieved by adding /2 at the end of the URL (box 4).

Box 4. Bash command example to retrieve the sequence of a specific lncRNA in FASTA format.

```
$ curl
http://greenc.sciencedesigners.com/api/db/greenc/entry/Athaliana_AT1G01170.1/fasta/2>
lcl|Athaliana_AT1G01170.1:1-515
ACGACCGTCTTCCACCGTTGAATTCTTCTGGAACTGGAGTCCACTGTTTAAGCTTCACTGTCTCTGAATCGGCAAAGCTT
TAGAAGAAAATGGCATCAGGAGGTAAAGCCAAGTACATAATCGGTGCTCTCATCGGTTCTTTCGGAATCTCATACATCTT
CGACAAAGTTATCTCTGATAATAAGATCTTTGGAGGGACTACTCCAGGAACTGTCTCTAACAAAGAATGGTGGGCAGCAA
CGGATGAGAAATTCCAAGCATGGCCAAGAACCGCTGGTCCTCCCGTTGTTATGAATCCCATTAGCCGTCAGAATTTCATC
GTCAAGACTCGTCCGGAATGAGAAAATAATAAGTTCAATGCTTTGATTTTCAGAATAAGATGAACGATGACGATGTTTTC
TAAATCCGAGCTTGTACTAAATAACAATACATTACAACACGGTTTGCGGAACTACTCCACAGTCTATCTTCTGTTAAAAA
ACTCAAACAAGCTATTGCAAAAAGCCCTTACGAGA
```

3.3.2.2 Available species

Another function of the API provides all species that GreeNC stores. The function is called "species" and yields a list or array (box 5).

Box 5. Bash command example to retrieve the list of available species.

```
$ curl http://greenc.sciencedesigners.com/api/species/
["Amborella_trichopoda","Ananas_comosus","Arabidopsis_lyrata","Arabidopsis_thaliana",
"Brachypodium_distachyon","Capsella_grandiflora","Capsella_rubella","Carica_papaya","
Chlamydomonas_reinhardtii","Citrus_clementina","Citrus_sinensis","Coccomyxa_subellips
oidea_C-
169","Cucumis_sativus","Eucalyptus_grandis","Eutrema_salsugineum","Fragaria_vesca","G
lycine_max","Gossypium_raimondii","Linum_usitatissimum","Malus_domestica","Manihot_es
culenta","Medicago_truncatula","Micromonas_pusilla_CCMP1545","Micromonas_pusilla_RCC2
99","Mimulus_guttatus","Musa_acuminata","Oryza_sativa_Japonica_Group","Ostreococcus_l
```

```
ucimarinus","Phaseolus_vulgaris","Physcomitrella_patens","Populus_trichocarpa","Prunus_persica","Ricinus_communis","Selaginella_moellendorffii","Setaria_italica","Solanum_lycopersicum","Solanum_tuberosum","Sorghum_bicolor","Spirodela_polyrhiza","Theobroma_cacao","Triticum_aestivum","Vitis_vinifera","Volvox_carteri","Zea_mays","Zostera_marina"]
```


3.3.2.3 Transcript information

The function "transcript" shows the transcript information for a lncRNA in JSON format. It is necessary to specify the transcript alias at the end or the URL (box 6). If information about more than one transcript needs to be retrieved, the transcript aliases need to be concatenated separated by "+" and placed at the end of the URL (for instance Athaliana_AT1G01170.1+Athaliana_AT1G01471.1).

Box 6. Bash command example to retrieve information about one or several lncRNAs.


```
$ curl http://greenc.sciencedesigners.com/api/transcript/Athaliana_AT1G01170.1
[{"Athaliana_AT1G01170.1": {
    "transcript_name":"AT1G01170.1",
    "features": {
        "length":515,
        "cpc_type":"noncoding",
        "cpc_potential":-0.756,
        "mfei":-21.573,
        "amfe":-0.519,
        "gc_content":41.553
    },
    "swissprot":{},
    "rfam":{},
    "repbase":{},
    "confidence":"High",
    "gene_alias":"Athaliana_AT1G01170",
    "gene_name":"AT1G01170",
    "coord": {
        "chromosome":"Chr1",
        "start":73931,
        "end":74737,
        "strand":"-",
        "species":"Athaliana"
    }
}}]
```


3.3.2.4 BLAST queries

It is also possible to perform BLAST searched via API instead of using the GUI. While the Bash commands seen so far were doing GET requests to the server, the "blast" function of the API needs a POST request. The variables that need to be defined are "seq" (required), which needs to store the query sequence, and the "e-value" (optional). The output is in JSON format containing the different alignments, positions and scores (box 7).

Box 7. Bash command example to perform a BLAST search given a query sequence.

```
$ curl -d
"evalue=0.001&seq=TACTTTTCTAATATCACGAGGACTTACATGGCCTCAAGTCACCTGTGGTGTTGTGCAAGAAGGAGAA
GCAAAGTCTGTCTATGTATTATGAGATAGCTACTTCTATGGCTAGGATATAGTTGTACAAGACCGGCTTTTCTTCTACTTCTTGC
ACAACCTGAGGTTATTGAGGCTATACAAGTCTTCTTCTATAATGTTATTTATTA" -X POST
http://greenc.sciencedesigners.com/api/blast
{"_id":"",
    "type":"blast",
    "ref":"",
    "db":"blastdb",
    "program":"blastn",
    "seqtype":"prot",
    "maxiters":"1",
    "username":"Anonymous",
    "date":"",
    "params": {
        "expect":0.001,
        "gap_open":5,
        "gap_extend":2,
        "filter":"L;m;"
    },
    "results": [object]
}
```

3.4 Discussion

Over the last few years, it has emerged the idea that lncRNAs might play an important role in transcriptional regulation and control of gene expression rather than being just transcriptional noise (Carninci, *et al*. 2005; Mattick and Makunin, 2006; Mattick, 2009; Djebali, et al. 2012; Derrien, *et al*. 2012). In order to boost the lncRNA research in plants, we have developed the Green Non-Coding (GreeNC) database, an online database of plant lncRNAs with a user-friendly and intuitive access for researchers and API for bioinformaticians and informatics pipelines. It is based on a MediaWiki running on the Amazon Web Services with the Semantic MediaWiki extension incorporated.

Among the most relevant databases for plant lncRNA currently available, GreeNC database is the most comprehensive in terms of the number of species, as it contains 39 plant species - including important species for agriculture such as tomato, orange, cucumber, apple, wheat, maize, or potato- and 6 algae (a total of 45 species). Although PlantNATsDB contains 70 species, this database is not specific for lncRNAs, but specific for NATs. Some lncRNAs might be NATs, but these terms are not interchangeable (St Laurent, *et al*. 2015). In terms of the number of lncRNAs stored, CANTATAdb v2.0 currently stores few thousands more lncRNAs than GreeNC. Nevertheless, CANTATAdb v1.0 got updated in 2018 to the 2.0 version. At the beginning of 2016, when GreeNC was officially released, CANTATAdb only contained 45.000 lncRNAs and 10

species. Therefore, GreeNC also represented the biggest database in terms of number of lncRNAs from early 2016 to the beginning of 2018.

Since its initial release in 2016, GreeNC has been visited more than 12,000 times, including the view of more than 57,000 pages according to Google Analytics. In addition, GreeNC has proven itself as a useful repository for lncRNA plant research as it was used by several genome-wide identification studies. These include studies where several hundreds of unannotated lncRNAs have been identified (Kwenda, *et al*. 2016; Li, *et al*. 2017), while others have used GreeNC to characterize a lncRNA related to the flowering of *Paspalum notatum* (Ochogavía *et al*. 2017). In addition, GreeNC has also been used for training machine-learning algorithms in a successful way (da Costa Negri, *et al*. 2018).

Moreover, the big amount of information stored in GreeNC has been extracted from high sensitivity and specificity pipelines and it is available via an easy-to-use and friendly GUI as well as via API. The database not only follows the key accessibility suggestions of Helmy *et al*. 2016 for database creation; the access to the database's knowledge is enhanced due to the Semantic MediaWiki extension. This extension allows the creation of query pages such as the advanced search page described in section 3.3.1.4.

Summarizing, the Green Non-Coding database can be considered a valid reference database for plant lncRNA research. Since genomic projects are making available new expression details across different tissues and novel lncRNAs are continuously reported in the literature, future updates of GreeNC will include this information to complete its resources for lncRNA research in plants.

3.5 References

Banks, J. A., et al. "The Selaginella Genome Identifies Genetic Changes Associated with the Evolution of Vascular Plants." *Science*, vol. 332, no. 6032, 2011, pp. 960–63, doi:10.1126/science.1203810.

Bao, W., et al. "Repbase Update, a Database of Repetitive Elements in Eukaryotic Genomes." *Mobile DNA*, vol. 6, no. 1, 2015, p. 11, doi:10.1186/s13100-015-0041-9.

Bartholomé, J., et al. "High-Resolution Genetic Maps of Eucalyptus Improve Eucalyptus Grandis Genome Assembly." *New Phytologist*, vol. 206, no. 4, 2015, pp. 1283–96, doi:10.1111/nph.13150.

Bennetzen, J. L., et al. "Reference Genome Sequence of the Model Plant Setaria." *Nature Biotechnology*, vol. 30, no. 6, 2012, pp. 555–61, doi:10.1038/nbt.2196.

Blanc, G., et al. "The Genome of the Polar Eukaryotic Microalga Coccomyxa Subellipsoidea Reveals Traits of Cold Adaptation." *Genome Biology*, vol. 13, no. 5, 2012, doi:10.1186/gb-2012-13-5-r39.

Boerner, S., McGinnis, K. M. "Computational Identification and Functional Predictions of Long Noncoding RNA in Zea Mays." *PLoS ONE*, vol. 7, no. 8, 2012, p. e43047, doi:10.1371/journal.pone.0043047.

Campalans, A., et al. “Enod40, a Short Open Reading Frame-Containing MRNA, Induces Cytoplasmic Localization of a Nuclear RNA Binding Protein in Medicago Truncatula.” *The Plant Cell Online*, vol. 16, no. 4, 2004, pp. 1047–59, doi:10.1105/tpc.019406.

Carninci, P., et al. “The Transcriptional Landscape of the Mammalian Genome.” *Science*, vol. 309, no. 5740, 2005, pp. 1559–63, doi:10.1126/science.1112014.

Chan, A. P., et al. “Draft Genome Sequence of the Oilseed Species Ricinus Communis.” *Nature Biotechnology*, vol. 28, no. 9, 2010, pp. 951–56, doi:10.1038/nbt.1674.

Clark, J. W., Donoghue, P. C. J. “Whole-Genome Duplication and Plant Macroevolution.” *Trends in Plant Science*, vol. 23, no. 10, 2018, pp. 933–45, doi:10.1016/j.tplants.2018.07.006.

Consortium, The ENCODE Project. “An Integrated Encyclopedia of DNA Elements in the Human Genome.” *Nature*, vol. 489, no. 7414, 2012, pp. 57–74, doi:10.1038/nature11247.

Consortium, The Tomato Genome. “The Tomato Genome Sequence Provides Insights into Fleshy Fruit Evolution.” *Nature*, vol. 485, no. 7400, 2012, pp. 635–41, doi:10.1038/nature11119.

Derrien, T., et al. “The GENCODE v7 Catalog of Human Long Noncoding RNAs: Analysis of Their Gene Structure, Evolution, and Expression.” *Genome Research*, vol. 22, no. 9, 2012, pp. 1775–89, doi:10.1101/gr.132159.111.

Ding, J., et al. “A Long Noncoding RNA Regulates Photoperiod-Sensitive Male Sterility, an Essential Component of Hybrid Rice.” *Proceedings of the National Academy of Sciences*, vol. 109, no. 7, 2012, pp. 2654–59, doi:10.1073/pnas.1121374109.

Djebali, S., et al. “Landscape of Transcription in Human Cells.” *Nature*, vol. 489, no. 7414, 2012, pp. 101–08, doi:10.1038/nature11233.

Droc, G., et al. “The Banana Genome Hub.” *Database*, vol. 2013, 2013, doi:10.1093/database/bat035.

Flórez-Zapata, N. M. V., et al. “Long Non-Coding RNAs Are Major Contributors to Transcriptome Changes in Sunflower Meiocytes with Different Recombination Rates.” *BMC Genomics*, vol. 17, no. 1, 2016, p. 490, doi:10.1186/s12864-016-2776-1.

Franco-Zorrilla, J. M., et al. “Target Mimicry Provides a New Mechanism for Regulation of MicroRNA Activity.” *Nature Genetics*, vol. 39, no. 8, 2007, pp. 1033–37, doi:10.1038/ng2079.

Gai, Y. P., et al. “A Novel LncRNA, MuLnc1, Associated With Environmental Stress in Mulberry (Morus Multicaulis).” *Frontiers in Plant Science*, vol. 9, 2018, p. 669, doi:10.3389/fpls.2018.00669.

Ganesh, S., Svoboda, P. “Retrotransposon-Associated Long Non-Coding RNAs in Mice and Men.” *Pflügers Archiv - European Journal of Physiology*, vol. 468, no. 6, 2016, pp. 1049–60, doi:10.1007/s00424-016-1818-5.

Goodstein, D. M., et al. “Phytozome: A Comparative Platform for Green Plant Genomics.” *Nucleic Acids Research*, vol. 40, no. D1, 2012, pp. D1178–86, doi:10.1093/nar/gkr944.

Griffiths-Jones, S. “MiRBase: MicroRNA Sequences and Annotation.” *Current Protocols in Bioinformatics*, vol. Chapter 12, 2010, p. Unit 12.9.1-10, doi:10.1002/0471250953.bi1209s29.

Harrow, J., et al. “GENCODE: The Reference Human Genome Annotation for The ENCODE Project.” *Genome Research*, vol. 22, no. 9, 2012, pp. 1760–74, doi:10.1101/gr.135350.111.

Hellsten, U., et al. “Fine-Scale Variation in Meiotic Recombination in Mimulus Inferred from Population Shotgun Sequencing.” *Proceedings of the National Academy of Sciences*, vol. 110, no. 48, 2013, pp. 19478–82, doi:10.1073/pnas.1319032110.

Helmy, M., et al. “Ten Simple Rules for Developing Public Biological Databases.” *PLOS Computational Biology*, vol. 12, no. 11, 2016, p. e1005128, doi:10.1371/journal.pcbi.1005128.

Heo, J. B., Sung, S. “Vernalization-Mediated Epigenetic Silencing by a Long Intronic Noncoding RNA.” *Science*, vol. 331, no. 6013, 2011, pp. 76–79, doi:10.1126/science.1197349.

Hu, T. T., et al. “The Arabidopsis Lyrata Genome Sequence and the Basis of Rapid Genome Size Change.” *Nature Genetics*, vol. 43, no. 5, 2011, pp. 476–81, doi:10.1038/ng.807.

Huang, S., et al. “The Genome of the Cucumber, Cucumis Sativus L.” *Nature Genetics*, vol. 41, no. 12, 2009, pp. 1275–81, doi:10.1038/ng.475.

Jaillon, O., et al. “The Grapevine Genome Sequence Suggests Ancestral Hexaploidization in Major Angiosperm Phyla.” *Nature*, vol. 449, no. 7161, 2007, pp. 463–67, doi:10.1038/nature06148.

Jain, P., et al. “Identification of Long Non-Coding RNA in Rice Lines Resistant to Rice Blast Pathogen Maganaporthe Oryzae.” *Bioinformation*, vol. 13, no. 8, 2017, pp. 249–55, doi:10.6026/97320630013249.

Johnsson, P., et al. “Evolutionary Conservation of Long Non-Coding RNAs; Sequence, Structure, Function.” *Biochimica et Biophysica Acta*, vol. 1840, no. 3, 2014, pp. 1063–71, doi:10.1016/j.bbagen.2013.10.035.

Joshi, R. K., et al. “Genome Wide Identification and Functional Prediction of Long Non-Coding RNAs Responsive to Sclerotinia Sclerotiorum Infection in Brassica Napus.” *PLOS ONE*, vol. 11, no. 7, 2016, p. e0158784, doi:10.1371/journal.pone.0158784.

Kong, L., et al. “CPC: Assess the Protein-Coding Potential of Transcripts Using Sequence Features and Support Vector Machine.” *Nucleic Acids Research*, vol. 35, no. Web Server issue, 2007, pp. W345-9, doi:10.1093/nar/gkm391.

Kwenda, S., et al. “Genome-Wide Identification of Potato Long Intergenic Noncoding RNAs Responsive to Pectobacterium Carotovorum Subspecies Brasiliense Infection.” *BMC Genomics*, vol. 17, no. 1, 2016, p. 614, doi:10.1186/s12864-016-2967-9.

Lamesch, P., et al. “The Arabidopsis Information Resource (TAIR): Improved Gene Annotation and New Tools.” *Nucleic Acids Research*, vol. 40, no. D1, 2012, pp. D1202–10, doi:10.1093/nar/gkr1090.

Li, L., et al. “Genome-Wide Discovery and Characterization of Maize Long Non-Coding RNAs.” *Genome Biology*, vol. 15, no. 2, 2014, p. R40, doi:10.1186/gb-2014-15-2-r40.

Li, S., et al. “Genome-Wide Identification and Functional Prediction of Cold and/or Drought-Responsive LncRNAs in Cassava.” *Scientific Reports*, vol. 7, no. 1, 2017, p. 45981, doi:10.1038/srep45981.

Li, W., et al. “Pseudogenes: Pseudo or Real Functional Elements?” *Journal of Genetics and Genomics*, vol. 40, no. 4, 2013, pp. 171–77, doi:10.1016/j.jgg.2013.03.003.

Liu, X., et al. “A Novel Antisense Long Noncoding RNA, *TWISTED LEAF* , Maintains Leaf Blade Flattening by Regulating Its Associated Sense R2R3-MYB Gene in Rice.” *New Phytologist*, vol. 218, no. 2, 2018, pp. 774–88, doi:10.1111/nph.15023.

Mattick, J. S. “The Genetic Signatures of Noncoding RNAs.” *PLoS Genetics*, vol. 5, no. 4, 2009, p. e1000459, doi:10.1371/journal.pgen.1000459.

Mattick, J. S., Makunin, I. V. “Non-Coding RNA.” *Human Molecular Genetics*, vol. 15, no. suppl_1, 2006, pp. R17–29, doi:10.1093/hmg/ddl046.

Mayer, K. F. X., et al. “A Chromosome-Based Draft Sequence of the Hexaploid Bread Wheat (Triticum Aestivum) Genome.” *Science*, vol. 345, no. 6194, 2014, pp. 1251788–1251788, doi:10.1126/science.1251788.

Merchant, S. S., et al. “The Chlamydomonas Genome Reveals the Evolution of Key Animal and Plant Functions.” *Science*, vol. 318, no. 5848, 2007, pp. 245–50, doi:10.1126/science.1143609.

Milligan, M. J., Lipovich, L. “Pseudogene-Derived LncRNAs: Emerging Regulators of Gene Expression.” *Frontiers in Genetics*, vol. 5, 2014, p. 476, doi:10.3389/fgene.2014.00476.

Ming, R., et al. “The Draft Genome of the Transgenic Tropical Fruit Tree Papaya (Carica Papaya Linnaeus).” *Nature*, vol. 452, no. 7190, Apr. 2008, pp. 991–96, doi:10.1038/nature06856.

Motamayor, Juan C., et al. “The Genome Sequence of the Most Widely Cultivated Cacao Type and Its Use to Identify Candidate Genes Regulating Pod Color.” *Genome Biology*, vol. 14, no. 6, 2013, p. r53, doi:10.1186/gb-2013-14-6-r53.

Nawrocki, E. P., et al. “Rfam 12.0: Updates to the RNA Families Database.” *Nucleic Acids Research*, vol. 43, no. D1, 2015, pp. D130–37, doi:10.1093/nar/gku1063.

Negri, T. D. C., et al. “Pattern Recognition Analysis on Long Noncoding RNAs: A Tool for Prediction in Plants.” *Briefings in Bioinformatics*, 2018, doi:10.1093/bib/bby034.

Ochogavía, A., et al. “Structure, Target-Specificity and Expression of PN_LNC_N13, a Long Non-Coding RNA Differentially Expressed in Apomictic and Sexual Paspalum Notatum.” *Plant Molecular Biology*, vol. 96, no. 1–2, 2018, pp. 53–67, doi:10.1007/s11103-017-0679-4.

Ohno, S. “So Much ‘Junk’ DNA in Our Genome.” *Brookhaven Symposia in Biology*, vol. 23, 1972, pp. 366–70, http://www.ncbi.nlm.nih.gov/pubmed/5065367.

Ouyang, S., et al. “The TIGR Rice Genome Annotation Resource: Improvements and New Features.” *Nucleic Acids Research*, vol. 35, no. Database issue, 2007, pp. D883-7, doi:10.1093/nar/gkl976.

Palenik, B., et al. “The Tiny Eukaryote Ostreococcus Provides Genomic Insights into the Paradox of Plankton Speciation.” *Proceedings of the National Academy of Sciences*, vol. 104, no. 18, 2007, pp. 7705–10, doi:10.1073/pnas.0611046104.

Paterson, A. H., et al. “The Sorghum Bicolor Genome and the Diversification of Grasses.” *Nature*, vol. 457, no. 7229, 2009, pp. 551–56, doi:10.1038/nature07723.

Paterson, A. H., et al. “Repeated Polyploidization of Gossypium Genomes and the Evolution of Spinnable Cotton Fibres.” *Nature*, vol. 492, no. 7429, 2012, pp. 423–27, doi:10.1038/nature11798.

Prochnik, S. E., et al. “Genomic Analysis of Organismal Complexity in the Multicellular Green Alga Volvox Carteri.” *Science*, vol. 329, no. 5988, 2010, pp. 223–26, doi:10.1126/science.1188800.

Prochnik, S., et al. “The Cassava Genome: Current Progress, Future Directions.” *Tropical Plant Biology*, vol. 5, no. 1, 2012, pp. 88–94, doi:10.1007/s12042-011-9088-z.

Rensing, S. A., et al. “The Physcomitrella Genome Reveals Evolutionary Insights into the Conquest of Land by Plants.” *Science*, vol. 319, no. 5859, 2008, pp. 64–69, doi:10.1126/science.1150646.

Schmutz, J., et al. “Genome Sequence of the Palaeopolyploid Soybean.” *Nature*, vol. 463, no. 7278, 2010, pp. 178–83, doi:10.1038/nature08670.

Schmutz, J., et al. “A Reference Genome for Common Bean and Genome-Wide Analysis of Dual Domestications.” *Nature Genetics*, vol. 46, no. 7, 2014, pp. 707–13, doi:10.1038/ng.3008.

Schnable, P. S., et al. “The B73 Maize Genome: Complexity, Diversity, and Dynamics.” *Science*, vol. 326, no. 5956, 2009, pp. 1112–15, doi:10.1126/science.1178534.

Shin, J. H., Chekanova, J. A. “Arabidopsis RRP6L1 and RRP6L2 Function in FLOWERING LOCUS C Silencing via Regulation of Antisense RNA Synthesis.” *PLoS Genetics*, vol. 10, no. 9, 2014, p. e1004612, doi:10.1371/journal.pgen.1004612.

Shuai, P., et al. “Genome-Wide Identification and Functional Prediction of Novel and Drought-Responsive LincRNAs in Populus Trichocarpa.” *Journal of Experimental Botany*, vol. 65, no. 17, 2014, pp. 4975–83, doi:10.1093/jxb/eru256.

Shulaev, V., et al. “The Genome of Woodland Strawberry (Fragaria Vesca).” *Nature Genetics*, vol. 43, no. 2, 2011, pp. 109–16, doi:10.1038/ng.740.

Slotte, T., et al. “The Capsella Rubella Genome and the Genomic Consequences of Rapid Mating System Evolution.” *Nature Genetics*, vol. 45, no. 7, 2013, pp. 831–35, doi:10.1038/ng.2669.

St Laurent, G., et al. “The Landscape of Long Noncoding RNA Classification.” *Trends in Genetics : TIG*, vol. 31, no. 5, 2015, pp. 239–51, doi:10.1016/j.tig.2015.03.007.

Swiezewski, S., et al. “Cold-Induced Silencing by Long Antisense Transcripts of an Arabidopsis Polycomb Target.” *Nature*, vol. 462, no. 7274, 2009, pp. 799–802, doi:10.1038/nature08618.

Tuskan, G. A., et al. “The Genome of Black Cottonwood, Populus Trichocarpa (Torr. & Gray).” *Science*, vol. 313, no. 5793, 2006, pp. 1596–604, doi:10.1126/science.1128691.

UniProt Consortium. “UniProt: A Hub for Protein Information.” *Nucleic Acids Research*, vol. 43, no. D1, 2015, pp. D204–12, doi:10.1093/nar/gku989.

Velasco, R., et al. “The Genome of the Domesticated Apple (Malus × Domestica Borkh.).” *Nature Genetics*, vol. 42, no. 10, 2010, pp. 833–39, doi:10.1038/ng.654.

Verde, I., et al. “The High-Quality Draft Genome of Peach (Prunus Persica) Identifies Unique Patterns of Genetic Diversity, Domestication and Genome Evolution.” *Nature Genetics*, vol. 45, no. 5, 2013, pp. 487–94, doi:10.1038/ng.2586.

Vogel, J. P., et al. “Genome Sequencing and Analysis of the Model Grass Brachypodium Distachyon.” *Nature*, vol. 463, no. 7282, 2010, pp. 763–68, doi:10.1038/nature08747.

Wang, W., et al. “The Spirodela Polyrhiza Genome Reveals Insights into Its Neotenous Reduction Fast Growth and Aquatic Lifestyle.” *Nature Communications*, vol. 5, no. 1, 2014, p. 3311, doi:10.1038/ncomms4311.

Wang, Z., et al. “The Genome of Flax (*Linum Usitatissimum*) Assembled *de Novo* from Short Shotgun Sequence Reads.” *The Plant Journal*, vol. 72, no. 3, 2012, pp. 461–73, doi:10.1111/j.1365-313X.2012.05093.x.

Worden, A. Z., et al. “Green Evolution and Dynamic Adaptations Revealed by Genomes of the Marine Picoeukaryotes Micromonas.” *Science*, vol. 324, no. 5924, 2009, pp. 268–72, doi:10.1126/science.1167222.

Wu, G. A., et al. “Sequencing of Diverse Mandarin, Pummelo and Orange Genomes Reveals Complex History of Admixture during Citrus Domestication.” *Nature Biotechnology*, vol. 32, no. 7, 2014, pp. 656–62, doi:10.1038/nbt.2906.

Xu, X., et al. “Genome Sequence and Analysis of the Tuber Crop Potato.” *Nature*, vol. 475, no. 7355, 2011, pp. 189–95, doi:10.1038/nature10158.

Yang, R., et al. “The Reference Genome of the Halophytic Plant Eutrema Salsugineum.” *Frontiers in Plant Science*, vol. 4, 2013, p. 46, doi:10.3389/fpls.2013.00046.

Young, N. D., et al. “The Medicago Genome Provides Insight into the Evolution of Rhizobial Symbioses.” *Nature*, vol. 480, no. 7378, 2011, pp. 520–24, doi:10.1038/nature10625.

Zimmer, A. D., et al. “Reannotation and Extended Community Resources for the Genome of the Non-Seed Plant Physcomitrella Patens Provide Insights into the Evolution of Plant Gene Structures and Functions.” *BMC Genomics*, vol. 14, no. 1, 2013, p. 498, doi:10.1186/1471-2164-14-498.

Zou, J., et al. “Retrotransposons - a Major Driving Force in Plant Genome Evolution and a Useful Tool for Genome Analysis.” *Journal of Crop Science and Biotechnology*, vol. 12, no. 1, 2009, pp. 1–8, doi:10.1007/s12892-009-0070-3.

Chapter 4: Development and application of a "Software as a Service" platform for the high-throughput analysis of RNA-seq data

4.1 Introduction

Transcriptomics is, at present, the most funded -omics field after genomics (Ulrich, 2016). This field is mainly based on RNA-seq technology, which generated, during 2017 alone, more than 200,000 samples that were submitted to the Sequence Read Archive (SRA) (items counted from SRA advanced search). This number continuously increase since, in only one year, the number of submitted samples doubled (more than 400,000 samples in 2018). Fuelled by technological advancements and the lowering of sequencing costs (see section 1.1.2), RNA-seq will become even more common in large-scale -omics studies, paving the way for the development of technologies such as single cell RNA-seq, which provides more resolution (Tang, *et al*. 2009; Saliba, *et al*. 2014).

Data derived from RNA-seq has been traditionally analysed by means of command-line interface that requires a deep knowledge of bioinformatics skills. Depending on the aim of the experiment, this typically includes the analysis of expression profiling (genes expressed in a given sample), Differential Gene Expression (DGE) analysis (genes significantly differentially expressed between samples) and functional characterization of expressed genes such as Gene Ontology Enrichment Analysis (GOEA) (gene functions significantly enriched in a subset of genes relative to the full set of genes from a given genome).

In this context, any software aiming to reach a wide range of users should include the following characteristics: it should (i) be easily accessible and cross-platform compatible, (ii) be reproducible with the use of Docker (Merkel, 2014), (iii) not require previous informatics or bioinformatics skills, thus being an end-to-end solution, (iv) include a wide range of non-model species, and (v) not be restricted by computational resources. Focusing on RNA-seq data analysis, several software with Graphical User Interface (GUI) -many of them cloud-based and commercially available- have emerged recently to allow the analysis of RNA-seq data, providing more accessibility to the research community (Kearse, *et al*. 2012; Illumina, 2014; Malhotra, *et al*. 2017). In this context, cloud-based systems for RNA-seq data analysis are gaining importance due to their cross-platform compatibility and their low computational requirements as the analyses are performed in remote machines. However, most of the software currently available

fail to fulfil some of the mentioned requirements (table 4). For instance, there is a limitation in the number of genomes available, truncating the possibility of working on the many non-model sequenced genomes. In addition, software currently available also require previous knowledge of bioinformatics such as understanding basic concepts (e.g. trimming or mapping) and being aware of additional tools that will be used in the analyses and their parameters that sometimes need to be tuned.

Table 4. Overview of the most relevant software available for DGE analyses. Bold words show the most optimal implementation of the feature displayed in the first column.

Software characteristics	Seven Bridges	DNAnexus	Genestack	Illumina BaseSpace	Galaxy	Geneious
Cloud-based	**Yes**	**Yes**	**Yes**	**Yes**	**Yes**	No
Dockers	**Yes**	**Yes**	**Yes**	**Yes**	**Yes**	No
Genomes available*	Limited	Limited	Limited	Limited	Limited	Limited
Computational resources	**Low**	**Low**	**Low**	**Low**	**Low**	High
Speed of analysis	**Fast**	**Fast**	**Fast**	**Fast**	N/A**	N/A**
Previous bioinformatics knowledge	Yes	Yes	Yes	Yes	Yes	Yes
Previous informatics knowledge	**No**	**No**	**No**	**No**	Yes	**No**
End to end solution	No	No	No	No	No	No

** Unlimited: all genomes sequenced and available. Limited: only few genomes from model organisms or need to be provided by the user. ** Not applicable: it highly depends on the computational resources available.*

With the aim to overcome the above-mentioned limitations, the main objective of this study is two-fold: (i) develop a new SaaS platform called Artificial Intelligence RNA-seq (AIR) for a user-friendly and effective way to analyse RNA-seq data, and (ii) validate its performance and usability taking advantage of RNA-seq data from mouse germ cells generated in our laboratory.

4.2 Methods

4.2.1 Reference genomes retrieval

AIR retrieves genomes and their associated gene annotations from three repositories: (i) Ensembl (https://ensembl.org/) and Ensembl Genomes (http://ensemblgenomes.org/), (ii) National Centre for Biotechnology Information (NCBI) RefSeq (https://ncbi.nlm.nih.gov/refseq/), and (iii) Joint Genome Institute (JGI) (https://jgi.doe.gov/). Three in-house Bash scripts were written for downloading genomes from each repository; they mainly use open-source software such as wget and curl. The access to Ensembl and NCBI is via File Transfer Protocol (FTP) without user credentials while the JGI site needs user credentials via Hypertext Transfer Protocol Secure (HTTPS), storing the session information in cookies for further downloading.

Additional information is required in order to obtain functional information of genes in the cases of Ensembl and NCBI RefSeq repositories. The Gene Ontology Consortium (http://geneontology.org/) provides the file "goa_uniprot_all.gaf.gz" with a relationship between UniProt IDs and Gene Ontology (GO) terms. Finally, in the case of NCBI, it also needs the file "gene_refseq_uniprotkb_collab.gz", which contains the association between RefSeq accession names and UniProt IDs.

4.2.2 Integrity and quality check of FASTQ data

AIR allows the upload of Illumina sequencing data from a wide range of files with the ".fq.gz", ".fastq.gz", ".txt.gz", ".fq.bz2", ".fastq.bz2", ".txt.bz2", ".fq", ".fastq", and ".txt" extensions. However, within the file, the only allowed format is FASTQ. In-house scripts written in Python and AWK validates the format of the uploaded files. Valid files undergo a quality check and trimming step using BBDuk (Bushnell, 2014) (minimum length of 35 bp and a minimum Phred-quality score of 25) and FastQC (Andrews, 2014).

4.2.3 Mapping and gene expression quantification

After quality check, trimmed reads are mapped with the splicing-aware mapper STAR (Dobin, *et al*. 2013) against the reference genome. The alignment mode is "end to end". The maximum number of mismatches allowed, defined by the "--outFilterMismatchNmax" argument, depends on whether the data come from the same genotype relative to the reference genome selected, from a different genotype, or from a different species. The value of this argument is 3, 5, or 10 whether the data is defined as "same genotype", "different genotype", or "different species", respectively. While chimeric alignments are not allowed, protrude alignments are permitted with the argument "--alignEndsProtrude 100 ConcordantPair". The gene expression quantification is performed with featureCounts (Liao, *et al*. 2014) with the following parameters: "-C", "-Q 30", and "-p" (in case of paired-end datasets).

4.2.4 Statistical analysis

Lowly expressed genes are filtered using HTSFilter (Rau, *et al*. 2013) with the following parameters: "s.len=50" and "s.max=200". The filtered set of genes is subsequently normalized with the function "normalizeData" from the same package using the Trimmed Mean of M-values (TMM) method. Afterwards, the filtered set of genes is given to four different statistical methods for the identification of Differentially Expressed Genes (DEG): DESeq2 (Love, *et al*. 2014), edgeR (Robinson, *et al*. 2010), EBSeq (Leng, *et al*. 2013) and NOISeq (Tarazona, *et al*. 2015). While DESeq2 and edgeR analyses are performed with default parameters, EBseq needs

an undetermined number of iterations in order to estimate parameters of the distribution it is based on. The correct number of iterations is reached when parameters converges the expected probability. We define convergence when the differences of parameters of the iteration i and the iteration i-1 is below 0.01. EBSeq starts with 5 iterations with a step of 5 more iterations if convergence is not reached. Finally, NOISeq is executed with noise correction using the TMM normalization approach, 100 permutations, no count filter applied, and 10 K-means clusters. Since NOISeq does not yield q-values but a probability of differential expression equivalent to 1 - q-values, its q-values are calculated as 1 - probability of differential expression.

Gene Ontology Enrichment Analysis (GOEA) is performed by means of a hypergeometric test for each GO term, thus identifying significant enriched GO terms relative to the expected genome background (Tian, *et al*. 2017). P-values are corrected with the Benjamini-Hochberg procedure to reduce false positives. We considered as significantly enriched the terms with q-values equal or smaller than 0.05. Results are shown in an interactive Voronoi treemap (https://carrotsearch.com/foamtree/).

Then, several plots, such as Volcano plots, MA-plot, Principal Component Analysis (PCA), or heatmaps are generated using the R libraries reshape2, ggplot and ggrepel. The R stats package is used for hierarchical clustering using the Euclidean distance.

4.3 Results

The Artificial Intelligence RNA-seq (AIR) is a rapid, easy-to-use SaaS platform for the analysis of RNA-seq data. It is made of three main sections (figure 9): (i) the **bioinformatics core** with all scripts and programs needed to perform the bioinformatics analysis, (ii) a **cloud-based architecture** to store and perform analyses online by using cloud computing, and (iii) a front-end with a **graphical user interface** accessible online. AIR is already setup to work on more than 150,000 genomes and it is available at https://transcriptomics.sequentiabiotech.com.

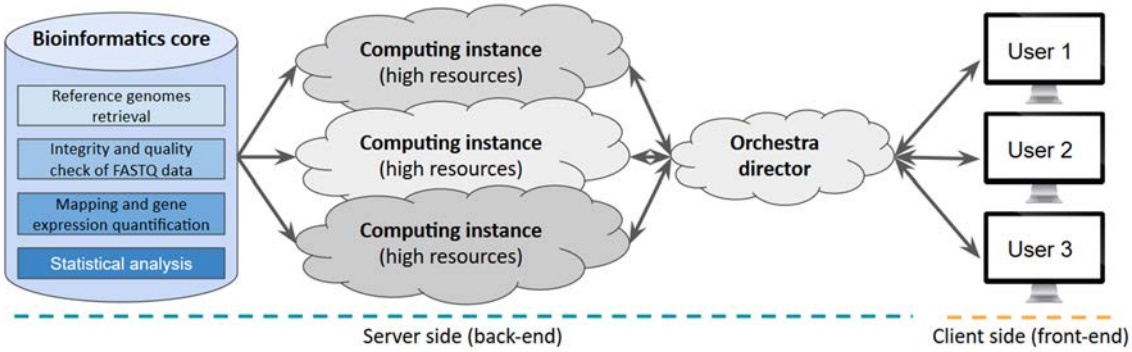


Figure 9. Overview of the three main sections of AIR. Users access to the graphical user interface through the web browser (client-side/front-end) to start an analysis. This user request is received by the cloud-based architecture (grey colour tones), specifically received by a remote machine called "orchestra director", which opens a computing

instance with high resources for the analysis. Computing instances are connected to the bioinformatics core (blue colour tones), which is a disk containing all pipelines for data processing and analysis.

4.3.1 Bioinformatics core

4.3.1.1 Reference genomes retrieval

Three different in-house Bash scripts have been written to connect to the Ensembl, to the National Centre for Biotechnology Information (NCBI) RefSeq and to the Joint Genome Institute (JGI) sites in order to retrieve the available genomes.

In particular, the Ensembl- and the NCBI-genomes retrieval scripts connect via File Transfer Protocol (FTP). They download the genome in FASTA format, the gene annotation in General Transfer Format (GTF) or in General Feature Format (GFF), and the UniProt annotation. UniProt IDs are then intersected with the Gene Ontology Consortium information, thus obtaining a final relationship between gene names and associated GO terms. Finally, the JGI-genome retrieval script connects to the site through its API after providing user credentials, as it needs an account to log in. From its repository, the genome in FASTA file, the gene annotation in GFF, and three additional text files containing gene definitions, gene names and the corresponding GO terms are downloaded. As a last step, each script indices the downloaded genome in FASTA format with STAR (Dobin, *et al.* 2013), as this RNA-seq aligner is the one used in a further step to map the samples provided by the user on the genome.

4.3.1.2 Integrity and quality check of FASTQ data

In house Python/AWK scripts were designed to read the full file and to check its integrity. These scripts check whether: (i) headers start with "@", (ii) sequences and quality scores have the same length, and (iii) the third line of each read starts with "+". In addition, paired-end files are detected and paired using the orientation identified from the header of the files, the read number and the read names. Reversed paired-end files without a detected mate or bad-formatted FASTQ files do not pass the check and cannot be used in further steps.

For the quality check or trimming step, a Python module was written to remove the bad quality portions of the reads in three steps. First, the module calls FastQC (Andrews, 2014) to gather quality metrics before trimming. Second, the actual trimming is performed using BBDuk (Bushnell, 2014). The third step involves another FastQC call on the final trimmed files. Quality metrics are stored into a JSON file, which is subsequently provided to the AIR front-end for visualization.

4.3.1.3 Mapping and gene expression quantification

The high-quality reads resulting from the quality check and trimming processes are mapped against the reference genome with STAR (Dobin, *et al*. 2013) using the end-to-end alignment mode. After the mapping step, the gene expression quantification is performed with featureCounts (Liao, *et al*. 2014), considering only the uniquely mapped reads and properly-paired reads in case of samples sequenced with the paired-end strategy. The quality of the alignments is assessed using Qualimap (Okonechnikov, *et al*. 2016).

4.3.1.4 Statistical analysis

From the featureCounts output, the statistical analysis starts by filtering lowly expressed genes using HTSFilter (Rau, *et al*. 2013). A Principal Component Analysis (PCA) is subsequently performed using the filtered set of genes and normalizing them with the Trimmed Mean of M-values (TMM) method. The filtered set of genes from the HTSFilter is given to four different statistical methods for the identification of Differentially Expressed Genes (DEG): DESeq2 (Love, *et al*. 2014), edgeR (Robinson, *et al*. 2010), EBSeq (Leng, *et al*. 2013) and NOISeq (Tarazona, *et al*. 2015) (see section 4.2.1.4). Additional plots such as the Volcano plot or the MA-plot are generated (figure 10A). A heatmap of expression patterns as Z-scaled FPKM (Fragments Per Kilobase of transcript per Million mapped reads) values from the DEG is also generated; the order of the genes is additionally established after hierarchical clustering using the Euclidean distance (figure 10B).

Significant genes are reported if their corrected q-values are equal or smaller than 0.05. The selected genes undergo a GOEA performed by the hypergeometric test approach with false discovery rate adjustment (Tian, *et al*. 2017). The GOEA results are shown in a Voronoi treemap (https://carrotsearch.com/foamtree/) (figure 10C). GO terms are represented as boxes. On the one hand, the size of the box depends on the significance of the enrichment of the GO term. Thus, the size of the box depends on the q-value of the GO term: the smaller it is, the bigger the box. On the other hand, the colour of the box depends on the enrichment score. The higher the enrichment, the reddish the colour.

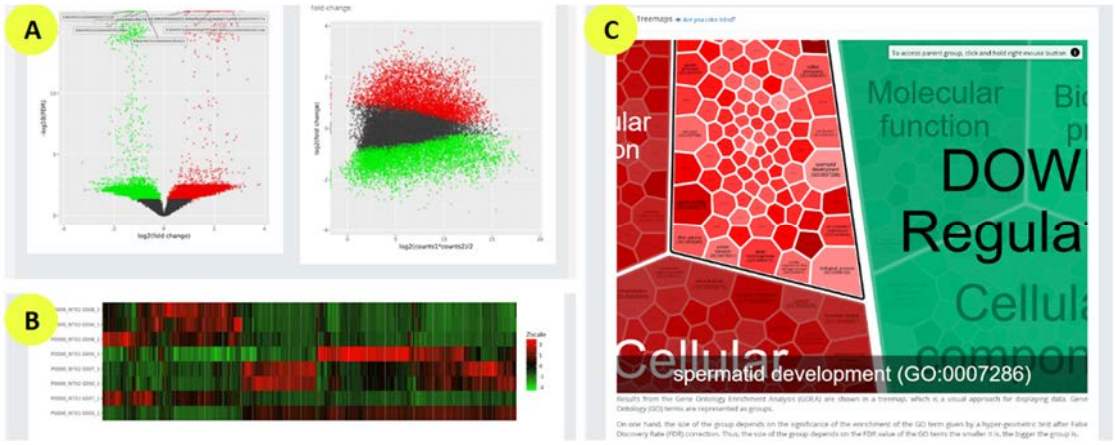


Figure 10. Overview of the different graphical outputs generated by the statistical analysis. (A) Volcano and MA plots; green dots are down-regulated genes while read dots are up-regulated ones. (B) Heatmap of expression patterns from the DEG. (C) Enriched GO terms from the GOEA in a Voronoi treemap.

4.3.2 Graphical user interface

The AIR platform is accessible from the main page https://transcriptomics.cloud or from https://transcriptomics.sequentiabiotech.com. Once logged in, the website is composed of different pages from which the user can find his/her own information, upload and validate sequencing data, create a new analysis, or viewing the final results. The frontend is coded in HTML/CSS and JavaScript.

4.3.2.1 User page

At the subdirectory /home (https://transcriptomics.sequentiabiotech.com/home), it brings the user to the user page of the platform, which can be divided into 3 different parts:

- **Navigation bar**. It is the left-side menu of the page and is shown in all pages of AIR to ease the navigation. The main pages of AIR are accessible from this menu (figure 11A).
- **User information**. This panel shows how many projects have been completed, how many samples the user has uploaded, how many gigabytes of data the user has uploaded, or how many euros are available in the user's wallet to be spent (figure 11B).
- **Shortcut buttons**. Two big buttons are shown at the right-side of the web page: "Samples" addresses you to the page to upload your samples, while "Analysis" addresses you to the page to define an analysis with already uploaded samples (figure 11C).

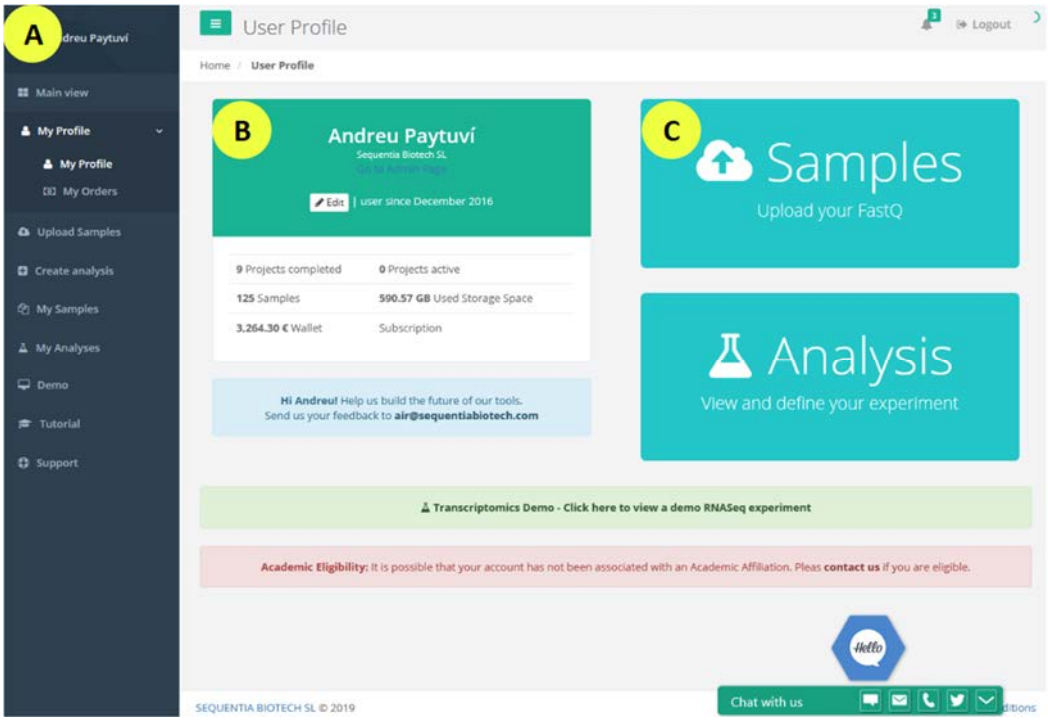


Figure 11. Screenshot of the user page in AIR. (A) Navigation bar from which the main pages of AIR are accessible. (B) User information showing how many projects have been completed, how many samples the user has uploaded, how many gigabytes of data the user has uploaded, or how many euros are available in the user's wallet. (C) Shortcut buttons to upload your samples ("Samples") or to define an analysis ("Analysis").

4.3.2.2 Sample upload page

AIR accepts raw Illumina sequencing data as input. The process of uploading the samples and making the accessible for further analyses goes through three steps:

1. **Upload files tab**. From this page, the user can upload samples stored on the computer or upload samples stored in Google Drive (figure 12).
2. **Validate samples tab**. Once the samples have been fully uploaded, they undergo a validation and integrity check: the main purpose of this step is to verify that the uploaded files contain sequencing data in FASTQ format and checks whether the FASTQ format is well formatted or truncated. It is important to validate file pairs (forward and reverse files) from paired-end sequencing data together, as this step also identifies pairs of samples in case paired-end sequencing data is uploaded.
3. **Resource payment page**. Successfully validated samples are eligible for analysis. Beforehand, these samples need to be selected in the resource payment page.

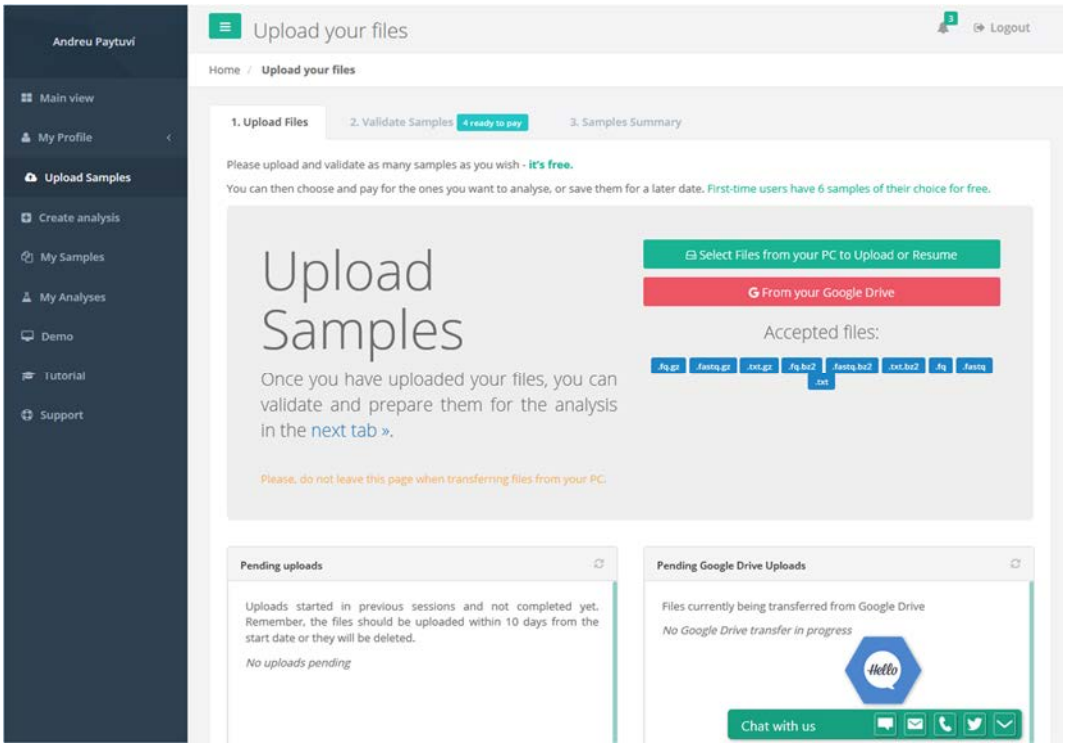


Figure 12. **Screenshot of the sample upload page in AIR.** From this page, a user can upload samples for further analysis.

4.3.2.3 Create analysis page

Once the samples are uploaded, the creation of an analysis also goes through 3 steps:

1. Analysis name and sequencing strategy. The user can specify the type of sequencing (either single or paired-end) and label the analysis with a custom name (figure 13).
2. Definition of groups. The user provides information about their experimental design with an easy drag-and-drop function. The minimum number of experimental groups is 2, and the name of each group is customizable.
3. Genome selection. A search box is available, displaying the best-matching genomes according to what has been typed. Currently, more than 150.000 genomes are already available. In addition, the user can indicate whether the samples have the same genotype of the reference genome, or if they are expected to be similar but not identical, or if they are related species.

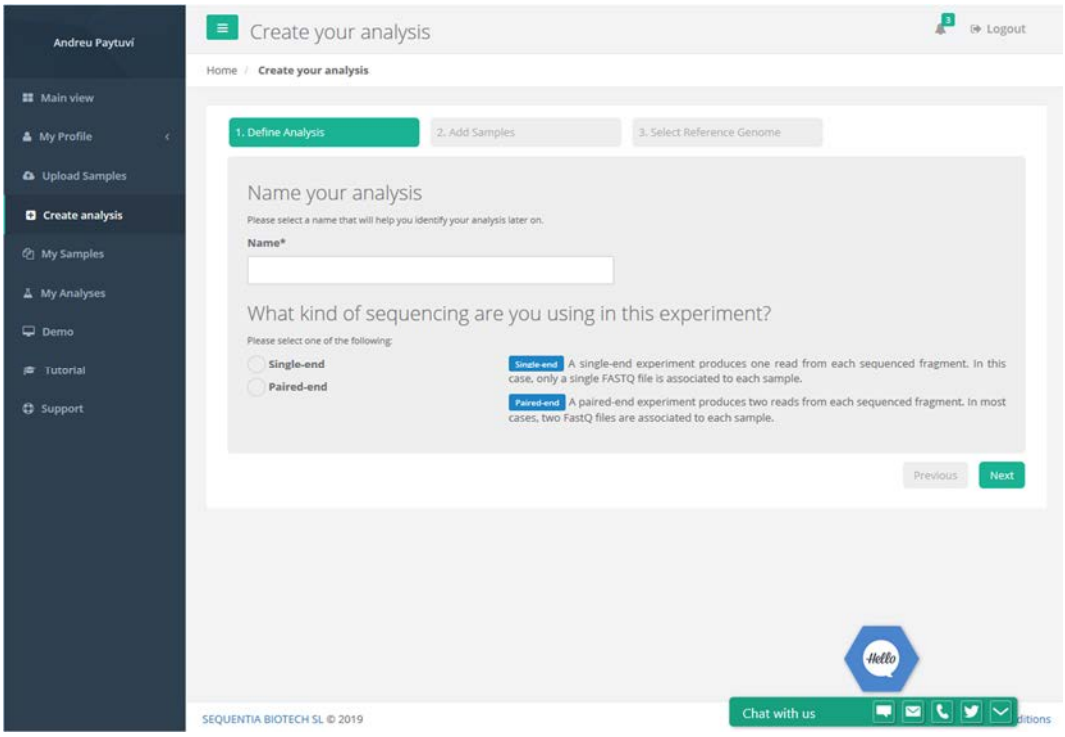


Figure 13. Screenshot of the analysis creation page in AIR. This page is the first step to define an analysis. In this page, the user must specify an analysis name and select the sequencing strategy used for the uploaded data (either single-end or paired-end).

4.3.2.4 Analysis page

When an analysis is started, it automatically appears in the section "My analyses" from the left-side menu. The analysis main page, which is accessible by clicking on the analysis name from "My analyses", contains 2 different parts:

- **Status of the analysis**. The analysis is divided into different parts: (i) trimming and quality check of the data, (ii) mapping data against the reference genome, and (iii) the statistical analysis. In order to follow the status of each analysis and to access to the generated results for each part, a workflow is shown in the middle of the page. In it, it is possible to see the degree of completion for each part as well as to access to the results once the corresponding part is done (figure 14A).
- **Downloadables**. The raw FASTQ files, the trimmed FASTQ files, the alignment files, and raw and normalized expression tables can be downloaded from the box at the right part of the page (figure 14B).

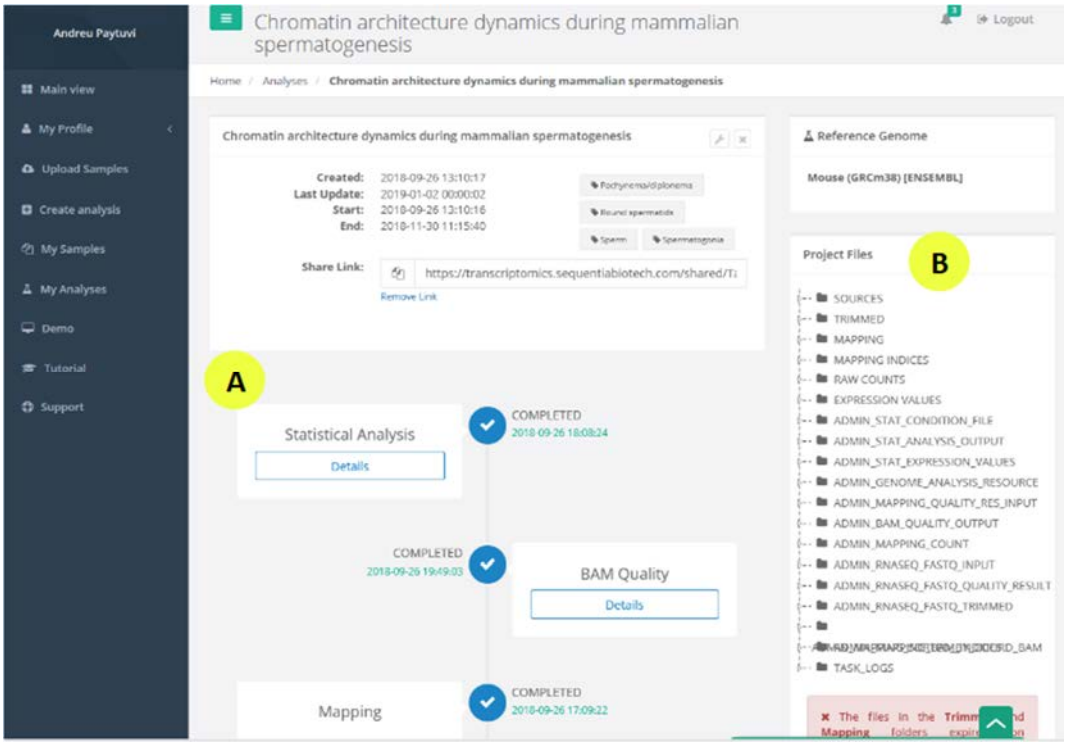


Figure 14. Screenshot of the analysis page in AIR. (A) Status of the analysis consisting of a workflow that allows the user to follow the status of the analysis as well as to access to the different results of the analysis. (B) Downloadables. Different generated output from the different analysis steps is available to be downloaded.

4.3.2.5 Trimming and mapping results

Results of the trimming step are shown after clicking the corresponding button in the workflow shown in the middle of the analysis page. The page is divided into 2 sections: (i) for the untrimmed data, and (ii) for the trimmed data. The information provided is summarized in 5 parts:

- Basic statistics. It shows a table displaying how many reads per file, its GC average content, the sequence length range, and the Illumina platform. The number of reads in the trimmed samples will always be less due to the fact that the bad-quality portions of the reads are removed and some reads are as short that they are removed as well.
- GC content. It shows a plot with the GC content distribution for each file.
- Per Base Sequence Quality. It shows the Phred quality scores qualities along reads; lines show the quality median while shadows show the lower and upper quartile. For Illumina data, the quality at the end of the reads tends to be lower. The per base quality in the trimmed samples should be very high.
- Sequence length Distribution. It shows a histogram with the distribution of read lengths. Along the trimming process, several reads have different unequal lengths as the bad quality portions of the reads are removed.

- **Adapter Content**. It shows the percentage of adapter along the read length.

Results of the mapping step are shown after clicking the button "BAM quality" in the workflow shown in the middle of the analysis page (figure 14A). The page shows a table with general statistics with the absolute number and percentage of reads that mapped to the genome 0 times (unmapped), 1 time (uniquely mapped), or more than 1 times (multi-mapped). In addition, it also shows how many reads and percentage of them mapping on genes (figure 15A). By clicking to the sample name, placed in the first column of the table, different plots and detailed statistics are shown (figure 15B).

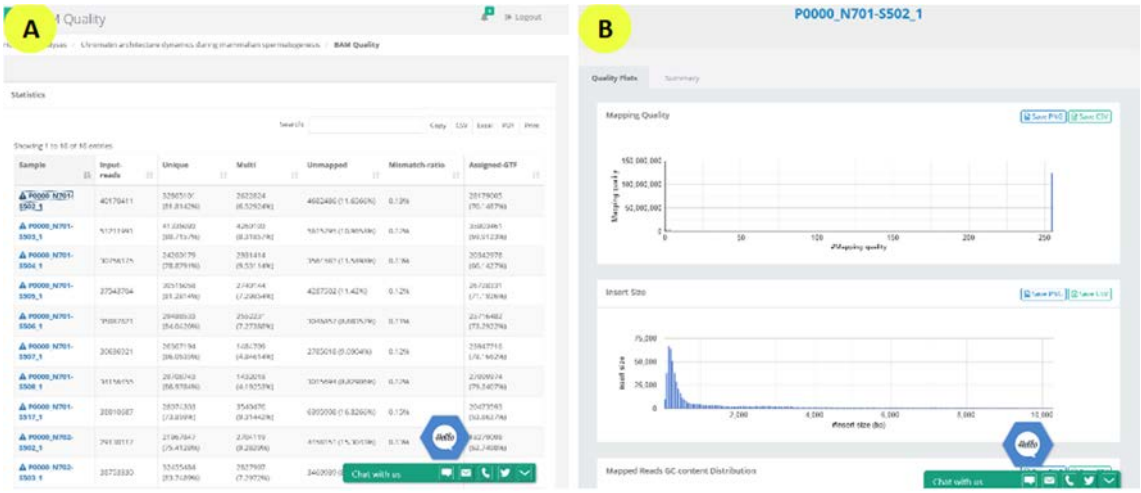


Figure 15. Screenshot of the mapping results in AIR. (A) Table with general mapping statistics. (B) Plots displaying different mapping information for a specific sample, after clicking its name from the table.

4.3.2.6 Statistical analysis results

The statistics page is made out of multiple rows, using each one the same reference for the different pairwise comparisons. The pairwise comparisons are represented in boxes, which contain four buttons, one for each statistical approach used (figure 16).

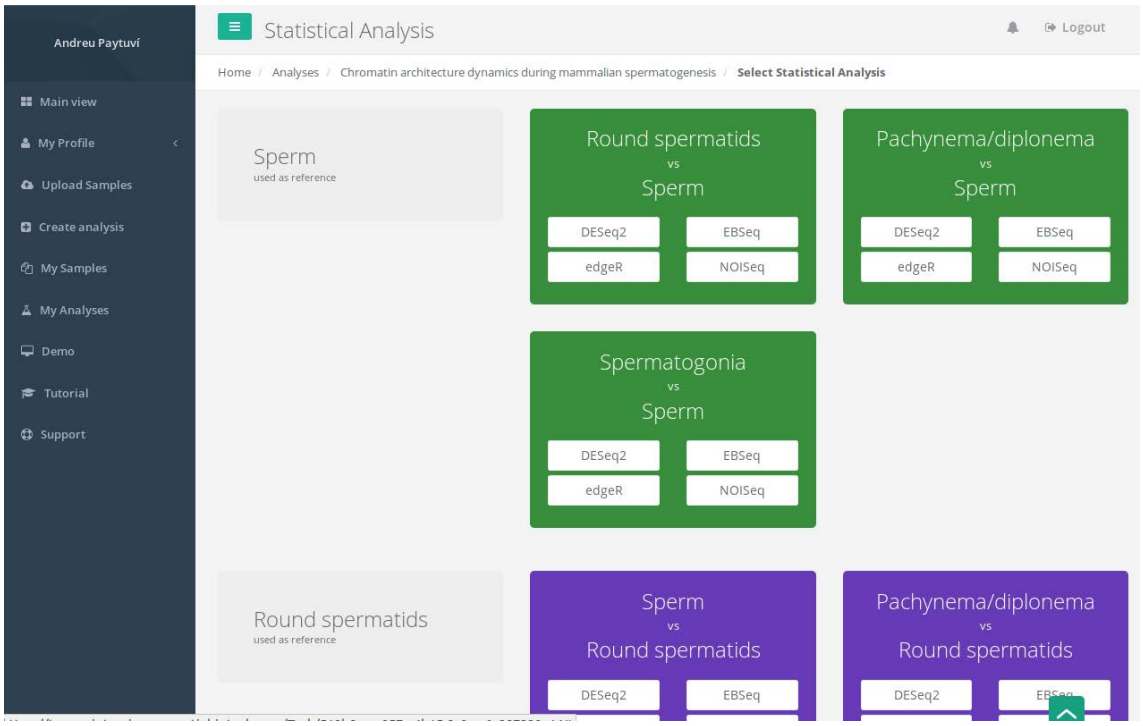


Figure 16. Screenshot of the statistical page in AIR. This page shows a row with green boxes, displaying pair-wise comparisons with sperm as **reference, and a row with purple boxes, displaying pair-wise comparisons with round spermatids as reference.**

AIR uses four R packages with different statistical approaches:

- **DESeq2** (Love, *et al*. 2014). It provides methods to test for differential expression by use of negative binomial generalized linear models. DESeq2 is one of the leading methods for a DGE analysis with 7073 citations (November 2018). When the variability across the replicates is high, it can only identify few differentially expressed genes.
- **edgeR** (Robinson, *et al*. 2010). It is the most cited package to perform a DGE analysis with 9226 citations (November 2018). Its statistical approach fits the data to a negative binomial model and uses an empirical Bayes estimation and exact tests to call DEG. When the variability across the replicates is high, it can only identify few differentially expressed genes.
- **NOISeq** (Tarazona, *et al*. 2015). It is a non-parametric approach suitable for experiments with high variability among replicates under the same condition. For these cases, NOISeq is able to identify more differentially expressed genes compared to edgeR and DESeq2, but its False Positive Rate (FPR) is higher (Schurch, *et al*. 2016).
- **EBseq** (Leng, *et al*. 2013). It applies empirical Bayes methods and a negative binomial distribution for DGE. Although it is particularly suited to perform time course analyses, it generally shows higher false positive rates with respect to edgeR and DESeq2 (Ching, *et al*. 2014; Schurch, *et al*. 2016).

At the bottom of the page, each statistical approach is accompanied by an explanation on the methodology and recommendations on their best use. When accessing to a specific statistical approach, the page shows: (i) a PCA showing the clustering of the samples, (ii) general charts for the interpretation of the experiment such as the Volcano and the MA plots (figure 10A), and (iii) several tabs in which the DEG and the GOEA results are shown.

The PCA is interactive, allowing direct manipulation of samples. If a given sample lies far apart from its group replicates (e.g. an outlier), the sample can be clicked in order to remove it from the analysis. Then, DEGs are shown in tables (including gene IDs, gene names, gene descriptions, and statistical values such as p-values, q-values, and fold changes) and their expression patterns (Z-scaled FPKM values) are presented in a heatmap. In this heatmap, changes in expression levels are displayed in different colours: from green (less expressed) to red (more expressed) relative to the corresponding reference (figure 10B). The order of the genes is established after hierarchical clustering using the Euclidean distance.

Moreover, the GOEA of the significant genes are shown in an interactive Voronoi treemap split by the three main categories in the gene ontology (Biological Process, BP; Cellular Component, CC; Molecular Function, MF) and divided by up-regulated and down-regulated genes (figure 10C). All data, including raw and normalized expression values and tables with the DGE and GOEA results are downloadable in spreadsheet file formats (e.g. CSV or xlsx).

4.3.3 Validation of AIR using RNA-seq data from mouse germ cells

Once the AIR pipeline was developed, we validated its suitability taking advantage of RNA-seq data produced in our lab (Vara and Paytuví-Gallart, *et al*. submitted). This included data obtained from four populations of highly enriched mouse germ cells isolated by Fluorescence-Activated Cell Sorter (FACS): spermatogonia, primary spermatocytes (pachynema/diplonema), round spermatids and sperm (figure 17).

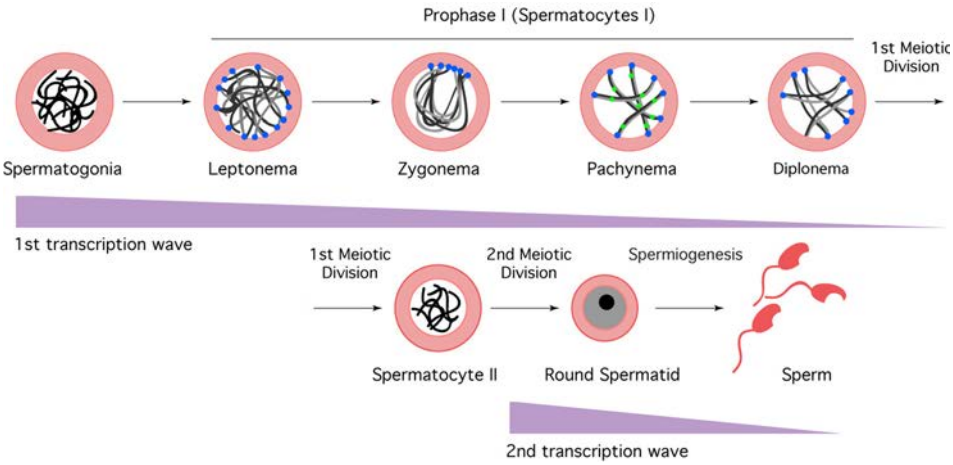


Figure 17. Overview of the spermatogenesis process. The two extensive waves of transcription that take place during this process can be seen below cell types (adapted from Reig-Viader *et al*. 2016 and de Mateo & Sassone-Corsi, 2014).

Pools containing between 20,000 and 40,000 cells were obtained by FACS-sorting from adult mice (C57BL/6J strain). Four independent biological replicates were included in the analysis. Briefly, full-length low-input RNA sequencing libraries were prepared by members of our research group using the Smart-seq2 protocol (Picelli, *et al*. 2014). Sequencing of Nextera® libraries was carried out on a HSeq2500 (Illumina) to obtain more than 30 million pair-end reads per sample. A total of sixteen replicates were uploaded in AIR and validated before conducting further analysis. All replicates together summed to 45 gigabytes and 1,228 million of reads. The analysis took five hours and 48 minutes to complete.

4.3.3.1 Quality metrics

Before the mapping step, sequence portions of the reads with low Phred quality scores were removed in order to increase the overall quality of the data. On average, samples did not lose more than 10% of the read pairs (table 5).

Table 5. General statistics before and after quality check and trimming step. It includes the number of reads and sequence length distributions of each biological replicate. Abbreviations: Sg – Spermatogonia; P/D – Pachynema/Diplonema; RS – Round Spermatids; R1 – Replicate 1, R2 – Replicate 2.

Biological replicate	File ID	Read pairs before trimming	Read length before trimming	Read pairs after trimming	Read length distribution after trimming	Percentage of survival
P/D R1	P0000_N701-S502_1	44072247	76	40170411	35-76	91.14
P/D R1	P0000_N701-S502_2	44072247	76	40170411	35-76	91.14
RS R1	P0000_N701-S503_1	56528181	76	51211991	35-76	90.59
RS R1	P0000_N701-S503_2	56528181	76	51211991	35-76	90.59
Sperm R1	P0000_N701-	36247245	76	30756175	35-76	84.85

	S504_1					
Sperm R1	P0000_N701-S504_2	36247245	76	30756175	35-76	84.85
Sg R1	P0000_N701-S505_1	41736421	76	37543704	35-76	89.95
Sg R1	P0000_N701-S505_2	41736421	76	37543704	35-76	89.95
P/D R2	P0000_N701-S506_1	37627009	76	35087621	35-76	93.25
P/D R2	P0000_N701-S506_2	37627009	76	35087621	35-76	93.25
RS R2	P0000_N701-S507_1	33223896	76	30636921	35-76	92.21
RS R2	P0000_N701-S507_2	33223896	76	30636921	35-76	92.21
Sperm R2	P0000_N701-S508_1	37132726	76	34156455	35-76	91.98
Sperm R2	P0000_N701-S508_2	37132726	76	34156455	35-76	91.98
Sg R2	P0000_N701-S517_1	40742610	76	38010687	35-76	93.29
Sg R2	P0000_N701-S517_2	40742610	76	38010687	35-76	93.29
P/D R3	P0000_N702-S502_1	31553176	76	29130117	35-76	92.32
P/D R3	P0000_N702-S502_2	31553176	76	29130117	35-76	92.32
RS R3	P0000_N702-S503_1	43455067	76	38753330	35-76	89.18
RS R3	P0000_N702-S503_2	43455067	76	38753330	35-76	89,18
Sperm R3	P0000_N702-S504_1	36639224	76	33023282	35-76	90,13
Sperm R3	P0000_N702-S504_2	36639224	76	33023282	35-76	90,13
Sg R3	P0000_N702-S505_1	38771263	76	34953928	35-76	90,15
Sg R3	P0000_N702-S505_2	38771263	76	34953928	35-76	90,15
P/D R4	P0000_N702-S506_1	37276664	76	34320275	35-76	92,06
P/D R4	P0000_N702-S506_2	37276664	76	34320275	35-76	92,06
RS R4	P0000_N702-S507_1	30567301	76	27825912	35-76	91,03
RS R4	P0000_N702-S507_2	30567301	76	27825912	35-76	91,03
Sperm R4	P0000_N702-S508_1	33184700	76	30854197	35-76	92,97
Sperm R4	P0000_N702-S508_2	33184700	76	30854197	35-76	92,97
Sg R4	P0000_N702-S517_1	35598432	76	32262491	35-76	90,62
Sg R4	P0000_N702-	35598432	76	32262491	35-76	90,62

	S517_2					

The read pairs were then mapped against the mouse genome GRCm38 (retrieved from Ensembl release 89). On average, 80% of the read pairs mapped in just one locus on the genome (uniquely mapped) and over 70% of read pairs were assigned to genes (table 6).

Table 6. Mapping efficiency statistics. It includes the number and percentage of reads that mapped once on the genome (uniquely mapped), reads that mapped multiple times on the genome (multi-mapped), reads that did not map (unmapped), and reads that mapped once on genes (uniquely mapped on genes). Abbreviations: Sg – Spermatogonia; P/D – Pachynema/Diplonema; RS – Round Spermatids; R1 – Replicate 1, R2 – Replicate 2.

Biological replicate	Sample ID	Number of read pairs	Uniquely mapped read pairs (%)	Multi-mapped read pairs (%)	Unmapped reads (%)	Uniquely mapped read pairs on genes (%)
P/D R1	P0000_N701-S502	40170411	32865101 (81.81%)	2622824 (6.52%)	4682486 (11.65%)	28179005 (70.14%)
RS R1	P0000_N701-S503	51211991	41336093 (80.71%)	4260103 (8.31%)	5615795 (10.96%)	35803461 (69.91%)
Sperm R1	P0000_N701-S504	30756175	24260179 (78.87%)	2931414 (9.53%)	3564582 (11.58%)	20342978 (66.14%)
Sg R1	P0000_N701-S505	37543704	30516058 (81.28%)	2740144 (7.29%)	4287502 (11.42%)	26728331 (71.19%)
P/D R2	P0000_N701-S506	35087621	29488533 (84.04%)	2552231 (7.27%)	3046857 (8.68%)	25716482 (73.29%)
RS R2	P0000_N701-S507	30636921	26367194 (86.06%)	1484709 (4.84%)	2785018 (9.09%)	23947716 (78.16%)
Sperm R2	P0000_N701-S508	34156455	29708743 (86.97%)	1432018 (4.19%)	3015694 (8.82%)	27099974 (79.34%)
Sg R2	P0000_N701-S517	38010687	28074303 (73.85%)	3540476 (9.31%)	6395908 (16.82%)	20473593 (53.86%)
P/D R3	P0000_N702-S502	29130117	21967847 (75.41%)	2704119 (9.28%)	4458151 (15.30%)	18279098 (62.74%)
RS R3	P0000_N702-S503	38753330	32455484 (83.74%)	2827907 (7.29%)	3469939 (8.95%)	29049945 (74.96%)
Sperm R3	P0000_N702-S504	33023282	28523977 (86.37%)	1404504 (4.25%)	3094801 (9.37%)	25838380 (78.24%)
Sg R3	P0000_N702-S505	34953928	27053571 (77.39%)	3075743 (8.79%)	4824614 (13.80%)	22602831 (64.66%)
P/D R4	P0000_N702-S506	34320275	26319970 (76.68%)	3217903 (9.37%)	4782402 (13.93%)	22198845 (64.68%)
RS R4	P0000_N702-S507	27825912	22246393 (79.94%)	2035509 (7.31%)	3544010 (12.73%)	19287752 (69.31%)
Sperm R4	P0000_N702-S508	30854197	26588200 (86.17%)	1306126 (4.23%)	2959871 (9.59%)	23990832 (77.75%)
Sg R4	P0000_N702-S517	32262491	23839185 (73.89%)	4079114 (12.64%)	4344192 (13.46%)	19311631 (59.85%)

4.3.3.2 Transcriptional profile of germ cells

Once the quality check was finished (trimming and mapping efficiency statistics), we analyzed differences in terms of gene expression among cell types. These differences can be browsed in the "Statistical analysis" page (figure 16). Since the analysis was defined with four conditions (four cell types) and comparisons are pairwise, the following comparisons were carried out (the reference is interchangeable):

- Spermatogonia *versus* pachynema/diplonema
- Spermatogonia *versus* round spermatids
- Spermatogonia *versus* sperm
- Pachynema/diplonema *versus* round spermatids
- Pachynema/diplonema *versus* sperm
- Round spermatids *versus* sperm

The PCA shown in each of the comparisons revealed that some replicates did not cluster with the other replicates from the same condition (outliers), thus suggesting high variability within conditions (figure 18 - Standard PCA). The PCA was then redone after applying on the data the NOISeq batch effect correction (figure 18 - NOISeq correction).

The high variability within conditions would explain the fact that almost no DEGs are found by edgeR and DESeq2. Nevertheless, NOISeq corrected the data in a way that replicates clustered with other replicates from the same condition more consistently. In this sense, since NOISeq is suitable for differential gene expression analysis from samples with high variability, it identified several thousands of DEG in each comparison (table 7).

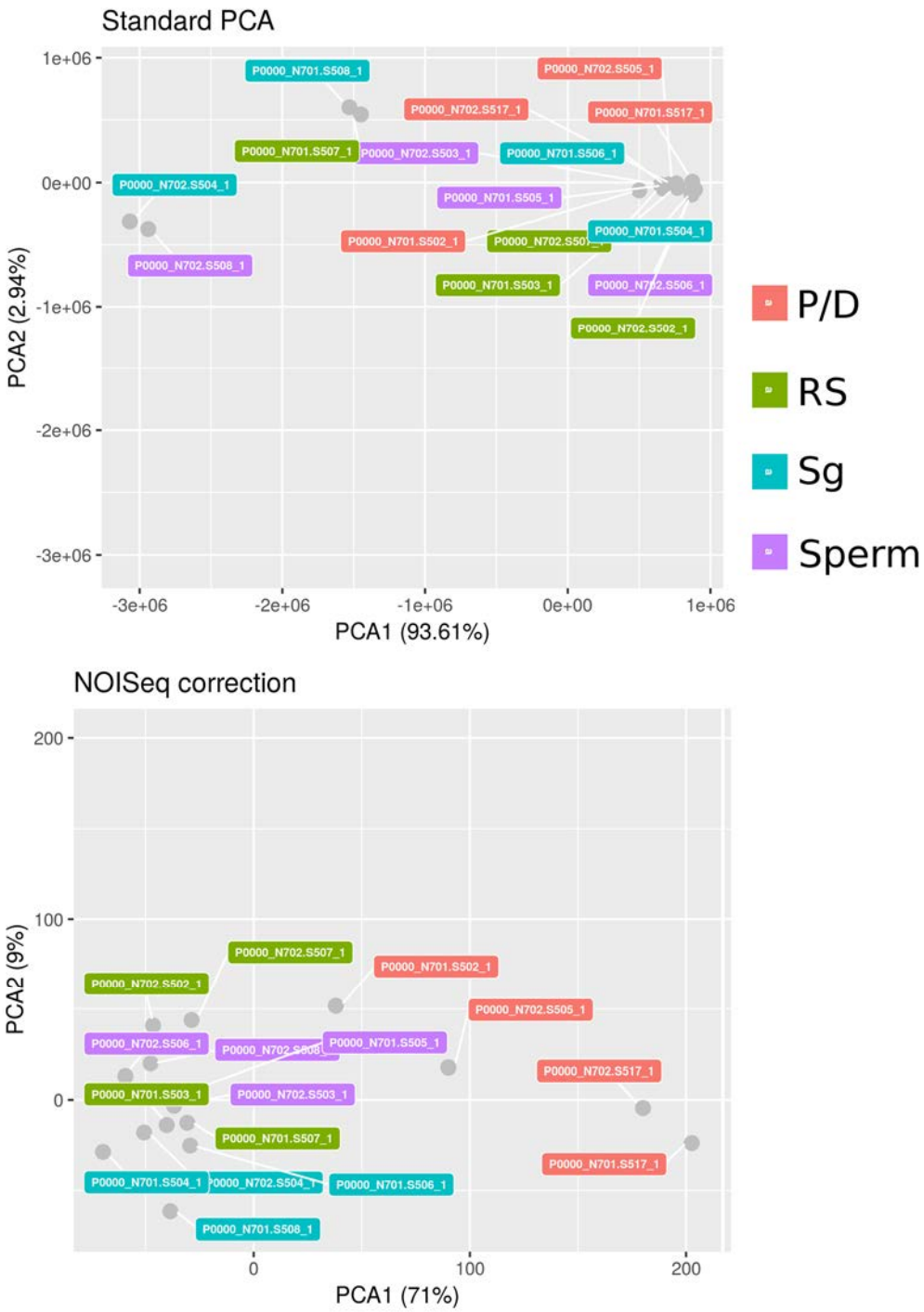


Figure 18. Principal Component Analyses (PCA) showing sample clustering. It shows the clustering of spermatogonia (Sg), pachynema/diplonema (P/D), round spermatids (RS) and sperm replicates before and after NOISeq batch effect correction.

Table 7. Number of DEGs for each comparison and statistical approach. A DEG was considered to have a q-value below 0.05. Abbreviations: Sg – Spermatogonia; P/D – Pachynema/Diplonema; RS – Round Spermatids.

Comparison	DESeq2	edgeR	NOISeq (q-value 0.05)	NOISeq (q-value 0.01)	EBSeq
Sg *versus* P/D	2,208	1.889	19,115	11,081	1,648
Sg *versus* RS	3,701	4.234	14,041	3,826	2,269
Sg *versus* Sperm	5,950	6.223	22,157	16,471	4,217
P/D *versus* RS	0	0	9,798	2,347	2
P/D *versus* Sperm	1,130	60	13,525	7,278	455
RS *versus* Sperm	69	85	11,023	3,760	1,064

AIR has the possibility to exclude outlier samples from the analysis as long as there are at least two replicates per experimental condition. Considering the pairwise comparison spermatogonia *versus* pachynema/diplonema, removing the outlier sample "P0000_N701-S505" belonging to spermatogonia, the number of DEG in DESeq2 increases up to 4,999 (+126%). All genes identified as DEG by the DESeq2 method were also identified by NOISeq. However, the number of DEG in NOISeq (q-value 0.05) was very high (table 7). Thus, in order to reduce the number of false positives, we filtered the results using a q-value of 0.01. With this threshold, the number of DEG in NOISeq was reduced to 11,081 (-57%) maintaining 95% of genes identified as DEG by DESeq2 without the outlier. Since changes in terms of gene expression are expected during gametogenesis, subsequent analyses were performed using the results obtained by the NOISeq method (q-value < 0.01).

Expression values and DEG tables were downloaded from AIR for further analyses. In order to obtain representative expression values for each gene and cell type, expression values of replicates from the same cell type and gene were averaged. Genome-wide, we detected that the number of expressing genes considering a Count Per Million (CPM) higher than 1 is reduced along spermatogenesis with 19,145, 15,480, 14,706 and 13,646 expressed genes in spermatogonia, pachynema/diplonema, round spermatids and sperm, respectively (table 8). Moreover, the number of expressed genes in the chromosome X is remarkably reduced from spermatogonia to pachynema/diplonema, as the number of expressed genes (CPM > 1) in the chromosome X of spermatogonia is 788 and in pachynema/diplonema 487 (-38.19%) (table 8).

Table 8. Number of expressing genes considering all chromosomes, autosomal chromosomes, or chromosome X. Reduction percentages relative to the cell type located at the left side column are shown within parenthesis. Abbreviations: Sg – Spermatogonia; P/D – Pachynema/Diplonema; RS – Round Spermatids.

	Sg	P/D	RS	Sperm
All chromosomes	19,145	15,480 (-19.14%)	14,706 (-4.99%)	13,646 (-7.37%)
Autosomes	18,355	14,992 (-18.32%)	14,225 (-5.11%)	13,263 (-6.76%)
X chromosome	788	487 (-38.19%)	481 (-1.23%)	359 (-25.36%)

In all pairwise comparisons analysed, the vast majority of DEGs detected were protein-coding (supplementary tables 1-2). However, the prevalence of protein-coding genes was continuously reduced along spermatogenesis. The net balance in spermatogonia *versus* pachynema/diplonema, round spermatids *versus* sperm, and spermatogonia *versus* sperm suggests a decrease of protein-coding genes and an increase of non-coding genes such as lncRNA genes, antisense RNAs (asRNA) genes and pseudogenes (figure 19). The net balance in pachynema/diplonema *versus* round spermatids is negative in all gene biotypes, indicating a higher transcriptional activity in round spermatids than in pachynema/diplonema primary spermatocytes. The overall balance in spermatogenesis, considering spermatogonia (cell type at the beginning of spermatogenesis) and sperm (cell type at the end of spermatogenesis), genes more expressed in spermatogonia are 80.34% protein-coding and 13.17% non-coding while genes more expressed in sperm are 56.99% protein-coding and 36.74% non-coding.

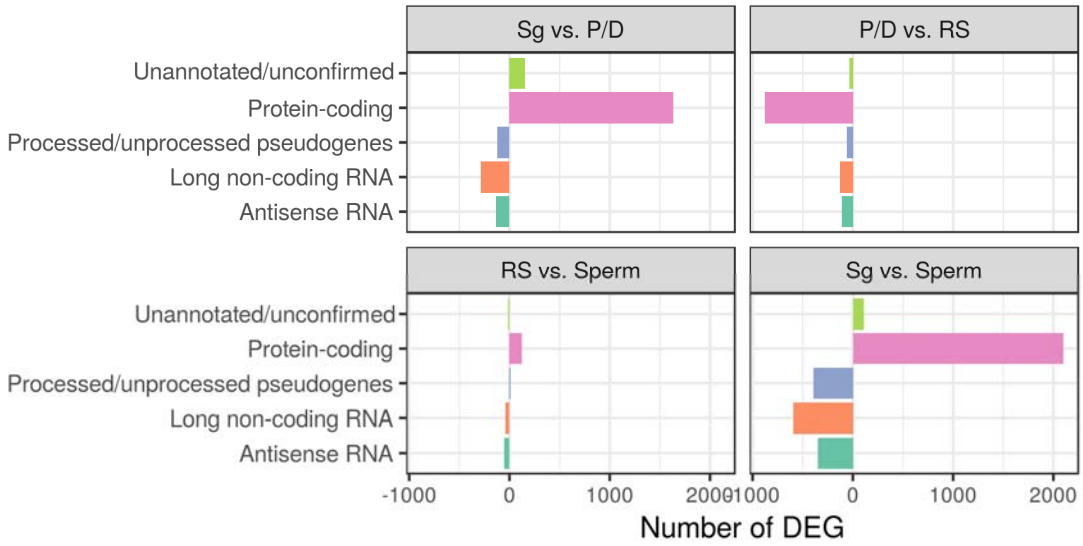


Figure 19. Balance of DEGs separated by biotype for each pair-wise comparison. It includes unannotated and unconfirmed genes, known protein-coding genes, pseudogenes, long noncoding RNA (lncRNA), and antisense RNA.

Several of the DEGs detected in our study were identified as relevant in spermatogenesis. For instance, the up-regulated genes in spermatogonia relative to pachytene/diplotene *Dmc1* and *Tex15* (fold change of 2.26 and 1.04, respectively) are essential for the Double Strand Breaks (DSBs) repair produced in leptotene/zygotene stages of primary spermatocytes (Yang, *et al.* 2008). The lncRNA *Xist* was also found among the up-regulated genes of spermatogonia relative to pachytene/diplotene (fold change of 1.64). Genes encoding synaptonemal complex proteins *Sycp1*, *Sycp2* and *Sycp3* were found as down-regulated in spermatogonia relative to pachynema/diplonema (fold-change of -0.47, -1.45 and -1.68, respectively). Also, acrosomal-associated genes such as *Spaca3*, *Spaca7* and *Spaca9* start being increasingly expressed in

round spermatids relative to pachynema/diplonema (fold changes of 0.79, 0.62, 1.48, respectively). More genes related to the *Spaca* family, which stands for Sperm Acrosome Associated, were found more expressed in round spermatids relative to sperm: *Spaca1*, *Spaca3*, *Spaca4*, *Spaca5*, *Spaca6* and *Spaca7* (fold changes of 2.39, 1.97, 1.98, 2.25, 1.15 and 0.78, respectively). In addition, the gene *Zpbp* (Zona Pellucida Binding Protein) is also more expressed in round spermatids relative to sperm (fold change of 1.22). In contrast, protamines *Prm1*, *Prm2* and *Prm3* are more expressed in sperm than in round spermatids (fold change of 1.54, 1.61 and 1.14, respectively).

4.4 Discussion

4.4.1 Development and applicability of the AIR platform

The number of RNA-seq samples submitted to SRA increased from more than 200,000 to more than 400,000 in the last two years, suggesting (i) an upward trend in the usage of the RNA-seq technology and (ii) an increasing accumulation of raw RNA-seq sequencing data that might create a bottleneck for bioinformatics analyses (see section 1.2.2). This has motivated the development of different solutions to analyse RNA-seq data in the recent years (Kearse, *et al*. 2012; Illumina, 2014; Malhotra, *et al*. 2017). Most of them are cloud-based systems, so hardware limitation is not a problem for the final user due to the computational power and storage are provided by Google Cloud, Amazon, Microsoft Azure, among others. However, current solutions for RNA-seq analyses have some limitations: (i) they require previous bioinformatics knowledge in order to select tools and tool-specific parameters, and (ii) they are limited to model species (table 9). In this sense, none of them is an end-to-end solution.

Like most of the currently available solutions for RNA-seq analysis, the Artificial Intelligence RNA-seq (AIR) presented here is a cloud-based platform, thus being accessible and cross-platform. In this way, the analyses are fast due to the fact that a computing instance with customized resources is exclusively opened for the user who, in turn, only requires a computer with enough computational resources to open a web browser (table 9). AIR also makes use of Docker containers, enhancing the reproducibility of bioinformatics analyses. However, AIR has been designed to be an end-to-end solution. On the one hand, users are not required to have previous informatics or bioinformatics knowledge due to the fact that analyses can be performed with few clicks on a web-based Graphical User Interface (GUI). On the other hand, there is no limitation in terms of choosing the genome as all sequenced organisms archived in Ensembl, NCBI and JGI are available in AIR (table 9). Considering that an end-to-end solution includes all stages of a process, we could say that AIR covers all stages of an RNA-seq analysis

from the input to the final result. No stage remains blocked due to lack of bioinformatics knowledge or working on non-model species. In this sense, AIR outperforms current RNA-seq solutions for non-bioinformaticians and scientist working on non-model species.

In fact, since its initial release in 2017, 754 users have created an account in AIR and 155 analyses have been performed on 46 different genomes. For instance, AIR has been already used to study obesity in humans (Gerlini, *et al*. 2018) and the expression changes after activating the transcription factor PPARγ also in human (Kim, *et al*. 2019). In this way, AIR is postulated as a highly valuable next-generation bioinformatics tool for RNA-seq data analysis, reaching in this case consistent results within short (a matter of few hours) timeframe.

Table 9. AIR features compared with the most relevant software available for DGE analyses. Bold words show the most optimal implementation of the feature displayed in the first column.

Software characteristics	AIR	Seven Bridges	DNAnexus	Genestack	BaseSpace	Galaxy	Geneious
Cloud-based	**Yes**	**Yes**	**Yes**	**Yes**	**Yes**	**Yes**	No
Dockers	**Yes**	**Yes**	**Yes**	**Yes**	**Yes**	**Yes**	No
Genomes available*	**Unlimited**	Limited	Limited	Limited	Limited	Limited	Limited
Computational resources	**Low**	**Low**	**Low**	**Low**	**Low**	**Low**	High
Speed of analysis	**Fast**	**Fast**	**Fast**	**Fast**	**Fast**	N/A**	N/A**
Previous bioinformatics knowledge	**No**	Yes	Yes	Yes	Yes	Yes	Yes
Previous informatics knowledge	**No**	**No**	**No**	**No**	**No**	Yes	**No**
End to end solution	**Yes**	No	No	No	No	No	No

** Unlimited: all genomes sequenced and available at the NCBI, Ensembl and JGI sites (currently > 150.000). Limited: only few genomes from model organisms or need to be provided by the user.*

*** Not applicable: it highly depends on the computational resources available.*

4.4.2 Transcriptional profiling and differential expression analysis of germ cells

In terms of applicability, AIR performance was validated taking advantage of RNA-seq data from mouse germ cells produced in our laboratory, including a total of 16 samples derived from four cell types involved in mouse spermatogenesis. It only took 5 hours and 48 minutes to process 45 gigabytes of data, including trimming, mapping and statistical analysis with four statistical algorithms (DESeq2, edgeR, NOISeq and EBseq) and all possible pair-wise comparisons. The dynamic tables and plots generated by AIR allow fast data mining, thus minimizing human resources, boosting research and accelerating results.

The fact that AIR includes four different statistical approaches allowed us to choose the most appropriate for the nature of the data. Being a dynamic biological process, germ cells presented high variability within conditions in the form of a poor clustering in the PCA (figure 18). This

might be indicative of having an unknown source of technical noise in the experiment (Tarazona, *et al*. 2013). Under these conditions, DESeq2, edgeR and EBseq performed poorly thus reporting a very low number of DEGs. In contract, NOISeq was able to correct the data for possible biases identifying this way a considerable number of DEGs.

Spermatogenesis is a highly regulated process at both the transcriptional and post-transcriptional levels (Bettegowda and Wilkinson, 2010; Hammoud, *et al*. 2014). Bivalent promoters (promoters with H3K4me3 and H3K27me3 histone modifications) are frequently observed in germ cells as part of the transcriptional regulation (Hammoud, *et al*. 2014). At the post-transcriptional levels, small non-coding RNA has been extensively studied in spermatogenesis, especially piRNA, which is required for transposon silencing (Fu and Wang, 2014). Other small non-coding RNAs, such as miRNA and siRNA, are involved in the regulation of gene expression (Yadav and Kotaja, 2014; Hilz, *et al*. 2016).

Our RNA-seq results revealed that the tendency along spermatogenesis is to reduce the expression of protein-coding genes in favour of genes that transcribe non-coding transcripts such as lncRNA, asRNA and pseudogenes. lncRNAs have been described as being involved in chromatin remodelling, transcriptional control and post-transcriptional processing (Mercer, *et al*. 2009; Barbosa Dogini, *et al*. 2014). Although the role of lncRNAs in spermatogenesis needs to be further investigated, several lncRNAs have been already identified as testis-specific (Hong, *et al*. 2018), involved in male germ cell development (Luk, *et al*. 2014) or involved in fertility (Wen, *et al*. 2016; Wichman, *et al*. 2017). Therefore, the progressive increase of lncRNAs along spermatogenesis might be indicative of potential functional roles. In addition, due to the fact that RNA was selected by the poly-A tail at the time to prepare the RNA-seq library, lncRNAs may even have even more relevance in spermatogenesis since they might have been underestimated due to the fact that most of them do not have the poly-A tail (Derrien, *et al*. 2012; Zhao, *et al*. 2018).

We also observed a decrease in the number of expressing genes during spermatogenesis, from 19,145 genes detected in spermatogonia to 13,646 in sperm. It has been generally accepted the existence of two waves of active transcription during spermatogenesis: the first one before meiosis and the second one before spermiogenesis starting in primary spermatocytes (Sassone-Corsi, 2002; de Mateo and Sassone-Corsi, 2014; da Cruz, *et al.* 2016). In this way, the first wave of active transcription in spermatogonia seems to promote the expression of a more variety of genes than the second one as the number of expressing genes is approximately 19,000 in spermatogonia and 15,500 in pachynema/diplonema. Interestingly, the reduction of the

expressing genes along spermatogenesis in chromosome X was sharper than in autosomes. In fact, the silencing of chromosome X has been described during meiosis I (Meiotic Sex Chromosome Inactivation, MSCI) (Turner, 2007; Yan and McCarrey, 2009). Remarkably, a substantial number of genes escaped global silencing in the X chromosome in primary spermatocytes (487 genes) and round spermatids (481 genes).

A high number of meiosis-related genes are differentially expressed during the different steps of spermatogenesis. Briefly, at leptonema, the pairing and the alignment of homologous chromosomes is promoted and it is maintained by the synaptonemal complex (Zickler and Kleckner, 1999). Precisely, the synaptonemal complex proteins *Sycp1*, *Sycp2* and *Sycp3* were found more expressed in pachynema/diplonema than in spermatogonia. Also, at this step, DSBs are initiated and are required for recombination (Keeney, *et al.* 1997; Martinez-Garay, *et al*. 2002). In these sense, important genes for DSBs such as *Dmc1* and *Tex15* were found more expressed in spermatogonia than in pachynema/diplonema. On the other hand, several Sperm Acrosome Associated (*Spaca*) genes are more expressed in round spermatids relative to the other steps of spermatogenesis. This gene family have roles in the binding of spermatozoa to the egg plasma membrane (Korfanty, *et al*. 2012). In addition, the gene *Zpbp* (Zona Pellucida Binding Protein), which is key for fertilization as participates in the interaction with the zona pellucida of the oocyte (Swegen, *et al*. 2018), is also more expressed in round spermatids relative to sperm. Therefore, round spermatids are already expressing genes that will be of essential importance in sperm. Finally, consistent with the replacement of histones by protamines during the spermiogenesis process (Balhorn, *et al*. 1984; Hud, *et al*. 1993; Johnson, *et al*. 2011), protamines *Prm1*, *Prm2* and *Prm3* were found more expressed in sperm than in round spermatids.

Overall, our germ cell transcriptome analysis highlighted several DEGs that were consistent with the sequential development of spermatogenesis and the specific events being carried out in each cell type (e.g. the MSCI, the assembly of the synaptonemal complex or the DSBs). Moreover, the transformation of round spermatids into spermatozoa was accompanied by the transcription of genes related to spermiogenesis and sperm function. This agrees with the second wave of active transcription described in round spermatids (da Cruz, *et al*. 2016).

4.5 References

Andrews, S. “FastQC.” *Babraham Bioinformatics*, 2010, doi:citeulike-article-id:11583827.

Bettegowda, A., Wilkinson, M. F. “Transcription and Post-Transcriptional Regulation of Spermatogenesis.” *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, vol. 365, no. 1546, 2010, pp. 1637–51, doi:10.1098/rstb.2009.0196.

Bushnell, B. “BBMap: A Fast, Accurate, Splice-Aware Aligner.” *Joint Genome Institute, Department of Energy*, 2014, doi:10.1186/1471-2105-13-238.

Ching, T., et al. “Power Analysis and Sample Size Estimation for RNA-Seq Differential Expression.” *RNA*, vol. 20, no. 11, 2014, pp. 1684–96, doi:10.1261/rna.046011.114.

Dobin, A., et al. “STAR: Ultrafast Universal RNA-Seq Aligner.” *Bioinformatics*, vol. 29, no. 1, 2013, pp. 15–21, doi:10.1093/bioinformatics/bts635.

Fu, Q., Wang, P. J. “Mammalian PiRNAs: Biogenesis, Function, and Mysteries.” *Spermatogenesis*, vol. 4, 2014, p. e27889, doi:10.4161/spmg.27889.

Gerlini, R., et al. “Glucose Tolerance and Insulin Sensitivity Define Adipocyte Transcriptional Programs in Human Obesity.” *Molecular Metabolism*, vol. 18, 2018, pp. 42–50, doi:10.1016/J.MOLMET.2018.09.004.

Hammoud, S. S., et al. “Chromatin and Transcription Transitions of Mammalian Adult Germline Stem Cells and Spermatogenesis.” *Cell Stem Cell*, vol. 15, no. 2, 2014, pp. 239–53, doi:10.1016/j.stem.2014.04.006.

Hilz, S., et al. “The Roles of MicroRNAs and SiRNAs in Mammalian Spermatogenesis.” *Development (Cambridge, England)*, vol. 143, no. 17, 2016, pp. 3061–73, doi:10.1242/dev.136721.

Hong, S. H., et al. “Profiling of Testis-Specific Long Noncoding RNAs in Mice.” *BMC Genomics*, vol. 19, no. 1, 2018, p. 539, doi:10.1186/s12864-018-4931-3.

Illumina. “Using Illumina BaseSpace Apps to Analyze RNA Sequencing Data” *Illumina*, 2016, https://www.illumina.com/content/dam/illumina-marketing/documents/products/technotes/technote-basespace-rna-seq.pdf.

Kearse, M., et al. “Geneious Basic: An Integrated and Extendable Desktop Software Platform for the Organization and Analysis of Sequence Data.” *Bioinformatics*, vol. 28, no. 12, 2012, pp. 1647–49, doi:10.1093/bioinformatics/bts199.

Kim, S. W., et al. “Transcriptome Analysis after PPARγ Activation in Human Meibomian Gland Epithelial Cells (HMGEC).” *The Ocular Surface*, 2019, doi:10.1016/J.JTOS.2019.02.003.

Korfanty, J., et al. “Identification of a New Mouse Sperm Acrosome-Associated Protein.” *Reproduction*, vol. 143, no. 6, 2012, pp. 749–57, doi:10.1530/REP-11-0270.

Leng, N., et al. “EBSeq: An Empirical Bayes Hierarchical Model for Inference in RNA-Seq Experiments.” *Bioinformatics*, vol. 29, no. 8, 2013, pp. 1035–43, doi:10.1093/bioinformatics/btt087.

Liao, Y., et al. “FeatureCounts: An Efficient General Purpose Program for Assigning Sequence Reads to Genomic Features.” *Bioinformatics*, vol. 30, no. 7, 2014, pp. 923–30, doi:10.1093/bioinformatics/btt656.

Love, M. I., et al. “Moderated Estimation of Fold Change and Dispersion for RNA-Seq Data with DESeq2.” *Genome Biology*, vol. 15, no. 12, 2014, p. 550, doi:10.1186/s13059-014-0550-8.

Luk, A. C. I., et al. “Long Noncoding RNAs in Spermatogenesis: Insights from Recent High-Throughput Transcriptome Studies.” *Reproduction*, vol. 147, no. 5, 2014, pp. R131–41, doi:10.1530/REP-13-0594.

Malhotra, R., et al. “Using the Seven Bridges Cancer Genomics Cloud to Access and Analyze Petabytes of Cancer Data.” *Current Protocols in Bioinformatics*, vol. 60, no. 1, 2017, p. 11.16.1-11.16.32, doi:10.1002/cpbi.39.

Martinez-Garay, I., et al. “A New Gene Family (FAM9) of Low-Copy Repeats in Xp22.3 Expressed Exclusively in Testis: Implications for Recombinations in This Region.” *Genomics*, vol. 80, no. 3, 2002, pp. 259–67, doi:10.1006/GENO.2002.6834.

Merkel, D. “Docker: Lightweight Linux Containers for Consistent Development and Deployment.” *Linux Journal*, vol. 2014, no. 239, 2014, https://dl.acm.org/citation.cfm?id=2600241.

Okonechnikov, K., et al. “Qualimap 2: Advanced Multi-Sample Quality Control for High-Throughput Sequencing Data.” *Bioinformatics*, vol. 32, no. 2, 2016, pp. 292–94, doi:10.1093/bioinformatics/btv566.

Picelli, S., et al. “Full-Length RNA-Seq from Single Cells Using Smart-Seq2.” *Nature Protocols*, vol. 9, no. 1, 2014, pp. 171–81, doi:10.1038/nprot.2014.006.

Rau, A., et al. “Data-Based Filtering for Replicated High-Throughput Transcriptome Sequencing Experiments.” *Bioinformatics*, vol. 29, no. 17, 2013, pp. 2146–52, doi:10.1093/bioinformatics/btt350.

Reig-Viader, R., et al. “Telomere Homeostasis in Mammalian Germ Cells: A Review.” *Chromosoma*, vol. 125, no. 2, 2016, pp. 337–51, doi:10.1007/s00412-015-0555-4.

Robinson, M. D., et al. “EdgeR: A Bioconductor Package for Differential Expression Analysis of Digital Gene Expression Data.” *Bioinformatics*, vol. 26, no. 1, 2010, pp. 139–40, doi:10.1093/bioinformatics/btp616.

Saliba, A. E., et al. “Single-Cell RNA-Seq: Advances and Future Challenges.” *Nucleic Acids Research*, vol. 42, no. 14, 2014, pp. 8845–60, doi:10.1093/nar/gku555.

Schurch, N. J., et al. “How Many Biological Replicates Are Needed in an RNA-Seq Experiment and Which Differential Expression Tool Should You Use?” *RNA*, vol. 22, no. 6, 2016, pp. 839–51, doi:10.1261/rna.053959.115.

Swegen, A., et al. “Unraveling Infertility: Deciphering the Molecular Basis of Idiopathic Infertility in a Thoroughbred Stallion.” *Journal of Equine Veterinary Science*, vol. 66, 2018, p. 90, doi:10.1016/J.JEVS.2018.05.056.

Tang, F., et al. “MRNA-Seq Whole-Transcriptome Analysis of a Single Cell.” *Nature Methods*, vol. 6, no. 5, 2009, pp. 377–82, doi:10.1038/nmeth.1315.

Tarazona, S., et al. “Data Quality Aware Analysis of Differential Expression in RNA-Seq with NOISeq R/Bioc Package.” *Nucleic Acids Research*, vol. 43, no. 21, 2015, p. e140, doi:10.1093/nar/gkv711.

Tian, T., et al. “AgriGO v2.0: A GO Analysis Toolkit for the Agricultural Community, 2017 Update.” *Nucleic Acids Research*, vol. 45, no. W1, 2017, pp. W122–29, doi:10.1093/nar/gkx382.

Turner, J. M. A. “Meiotic Sex Chromosome Inactivation.” *Development*, vol. 134, no. 10, 2007, pp. 1823–31, doi:10.1242/dev.000018.

Ulrich, T. “Opinionome: What Will Be the next Big –Ome?” *Broad Institute*, 2016, www.broadinstitute.org/blog/opinionome-what-will-be-next-big-–ome.

Wang, Z., et al. “RNA-Seq: A Revolutionary Tool for Transcriptomics.” *Nature Reviews Genetics*, vol. 10, no. 1, 2009, pp. 57–63, doi:10.1038/nrg2484.

Wen, K., et al. “Critical Roles of Long Noncoding RNAs in *Drosophila* Spermatogenesis.” *Genome Research*, vol. 26, no. 9, 2016, pp. 1233–44, doi:10.1101/gr.199547.115.

Wichman, L., et al. “Dynamic Expression of Long Noncoding RNAs Reveals Their Potential Roles in Spermatogenesis and Fertility†.” *Biology of Reproduction*, vol. 97, no. 2, 2017, pp. 313–23, doi:10.1093/biolre/iox084.

Yadav, R. P., Kotaja N. “Small RNAs in Spermatogenesis.” *Molecular and Cellular Endocrinology*, vol. 382, no. 1, 2014, pp. 498–508, doi:10.1016/J.MCE.2013.04.015.

Yan, W., McCarrey, J. R. “Sex Chromosome Inactivation in the Male.” *Epigenetics*, vol. 4, no. 7, 2009, pp. 452–56, http://www.ncbi.nlm.nih.gov/pubmed/19838052.

Yang, F., et al. “Mouse TEX15 Is Essential for DNA Double-Strand Break Repair and Chromosomal Synapsis during Male Meiosis.” *The Journal of Cell Biology*, vol. 180, no. 4, 2008, pp. 673–79, doi:10.1083/jcb.200709057.

Zhao, S., et al. “Evaluation of Two Main RNA-Seq Approaches for Gene Quantification in Clinical RNA Sequencing: PolyA+ Selection versus RRNA Depletion.” *Scientific Reports*, vol. 8, no. 1, 2018, p. 4781, doi:10.1038/s41598-018-23226-4.

Chapter 5: Analysis of the structural organization of the mouse genome during spermatogenesis

5.1 Introduction

Mammalian genomes are packaged into a specifically tailored chromatin structure that consists of several superimposed layers of organization. These include chemical modifications on DNA and histones (the **epigenome**) and the high-order chromatin organization (the **nucleome**). This organisation is achieved by chromatin folding into loops, Topologically Associating Domains (TADs), and compartments (A and B), which ultimately can influence the transcriptional activity of genomic regions (Lieberman-Aiden, *et al*. 2009; Dixon, *et al*. 2012; Rao, *et al*. 2014) (see section 1.1.4).

How these different levels of chromatin organization changes during the cell cycle has just begun to be elucidated (Dekker, *et al*. 2013). Studies in somatic cells have shown how the highly compartmentalized folding state of the genome in interphase is lost during mitosis (Naumova, *et al*. 2013; Gibcus, *et al*. 2018). But, despite the exciting recent advances in the field, the general picture of how mammalian genomes are packaged inside cells is incomplete. Even more fragmentary is our understanding of the heritability of genome organization. In this context, germ cells represent a unique cell model, where unipotent diploid cells undergo extensive cellular differentiation (meiosis) to form highly differentiated haploid cells that ultimately form a totipotent embryo after fertilization. In the case of mammalian males, germ cells are produced during **spermatogenesis** (figure 17). These sequential developmental stages involve dramatic and highly regulated chromosomal movements and chromatin remodelling, whose regulatory pathways are far from being understood.

Spermatogenesis begins with the cell differentiation of Primordial Germ Cells (PGCs) to spermatogonia (Reig-Viader, *et al*. 2016). This transition is composed of different mitotic cell divisions yielding a pool of uncommitted spermatogonia (spermatogonia A) that differentiate into committed spermatogonia (spermatogonia B) that enter meiosis (Grisworld, 2016). Thus, spermatogonia differentiate into primary spermatocytes, which undergo both first (resulting in secondary spermatocytes) and second (resulting in round spermatids) meiotic divisions. Finally, round spermatids undergo spermiogenesis, a differentiation phase that involves an intermediate step (elongated spermatids) to become male gametes (spermatozoa) ready for fecundation (figure 17).

In detail, primary spermatocytes pass though different stages along prophase of meiosis I (prophase I): leptonema, zygonema, pachynema and diplonema (reviewed in Reig-Viader, *et al*. 2016). It is in the primary spermatocytes at the leptotene stage (leptonema) where telomeres cluster at the nuclear envelope forming the so-called *bouquet* stage (Scherthan, *et al.* 1996; Reig-Viader, *et al.* 2016; Boateng, *et al*. 2013). This structure promotes the pairing and the alignment of homologous chromosomes, maintained by a protein structure called synaptonemal complex (Zickler and Kleckner, 1999). Also, at leptonema, DSBs required for recombination are initiated by the endonuclease protein SPO11 (Keeney, *et al.* 1997). DSBs are then repaired at the zygotene stage (zygonema), leading to synapses between homologous chromosomes. Subsequently, at pachytene stage (pachynema), the homologous chromosomes are fully synapsed, and recombination is resolved producing crossover (exchange of genetic material between two homologous chromosomes) and non-crossover events (repair of DSBs not resulting in exchange of genetic material between two homologous chromosomes) (Handel and Schimenti, 2010). At diplotene stage (diplonema), homologous chromosomes start to segregate by the disassembling the synaptonemal complex (Handel and Schimenti, 2010). After the first meiotic division, secondary spermatocytes, which are already haploid cells with two chromatids, are formed. They undergo the second meiotic division (meiosis II), resulting in round spermatids (haploid cells with one chromatid). Through spermiogenesis, round spermatids become first elongated spermatids and then motile sperm. This process includes changes in cell morphology and DNA packaging through the replacement of histones by protamines (Balhorn, *et al*. 1984; Hud, *et al*. 1993; Johnson, *et al*. 2011).

How the higher-order chromatin organisation is configured in mammalian pre-meiotic, meiotic and post-meiotic germ cells and how it is related to gene expression still remains largely unexplored. Recent studies in mouse (Jung, *et al*. 2017; Ke, *et al*. 2017; Wang, *et al*. 2019; Alavattam, *et al*. 2019; Patel, *et al*. 2019) and macaque (Wang, *et al*. 2019) suggested the existence of a remarkable reprogramming of chromatin architecture during mammalian spermatogenesis and early embryogenesis. However, how the different levels of chromatin organization are configured during all stages of spermatogenesis and how it is related with gene expression remain unknown. In this sense, the main aim of this work is to elucidate the organization and function of the three-dimensional (3D) genome during mouse spermatogenesis.

5.2 Material and methods

5.2.1 Material

We took advantage of *in situ* Hi-C data produced in our lab (Vara and Paytuví-Gallart, *et al*. submitted) for the study of the high-order chromatin structure. This included six different highly enriched germ cell populations from adult mice (C57BL/6J strain) isolated by FACS: spermatogonia (two replicates), primary spermatocytes at pachytene/diplotene (P/D) stages (three replicates) and at leptotene/zygotene (L/Z) stages (one replicate), secondary spermatocytes (two replicates), round spermatids (two replicates), and sperm (two replicates) (figure 17). In addition, Hi-C data from a population of fibroblasts (two replicates) was also produced as a somatic profile. Briefly, Hi-C libraries were prepared by members of our research group using the *in situ* Hi-C protocol (Rao, *et al*. 2014). Pools containing between 1.7 and 113 million cells were obtained by FACS. Sequencing of Hi-C libraries was carried out on a Hi-Seq 2500 v4 (Illumina) to obtain an average of 247 million pair-end reads per sample. All replicates together summed more than 200 gigabytes in size and 3,400 million reads (table 10).

5.2.2 Quality check of FASTQ data

Sequenced raw data underwent a quality check and trimming step using BBDuk (version 10/2015) (Bushnell, 2014). Setting a minimum read length of 35 bp and a minimum Phred quality score of 20, adapters and low-quality reads were removed while preserving their longest high-quality regions.

5.2.3 Hi-C data processing, binning and normalization

The workflow for Hi-C processing includes the following steps: (i) read mapping, (ii) fragment assignment, (iii) fragment filtering, (iv) binning, (v) bin level filtering, and (vi) balancing (figure 20A). It begins with paired-end sequencing data that is aligned against the reference genome as if it was single-end data, mapping each mate forward and reverse separately (figure 20A - read mapping). Hi-C fragments are chimeric; therefore, some reads might have covered the junction site, thus being also chimeric (the 5' and the 3' portions of a read coming from different loci). As chimeric reads will not align on the genome, reads are first shortened by truncating them before mapping. Multi-mapped reads are made longer by extending them by few nucleotides and mapped again along different iterations (iterative mapping). The workflow continues assigning each read to a restriction fragment, as they can be inferred from the genomic sequence (figure 20A - fragment assignment). However, during the Hi-C library preparation, some artefacts can appear: (i) self-ligated fragments ("self-circles"), (ii) un-ligated fragments ("dangling ends"), (iii) fragments not derived from restriction sites ("internal fragments"), and

(iv) re-ligated adjacent fragments ("contiguous sequences") (figure 20B). Reads derived from artefact fragments, from the PCR amplification step prior to sequencing ("PCR duplicates") and located too far from a restriction site thus being inconsistent with the library insert size are removed (figure 20A - fragment filter). Valid reads are usually binned into fixed genomic interval sizes (from 40 Kbp to 1Mbp) leading to the creation of a square matrix that stores all bin-bin interactions (figure 20A - binning). Since some genomic regions have low mappability or high repeat content, rows/columns from the square matrix lying on these regions should be removed as they are source of noise (figure 20A – bin filtering). Finally, an iterative correction (also called "balancing") is applied on the square matrix to correct any biases, such as the GC content, the mappability, or the number of fragments in each bin (Lajoie, *et al*. 2015; Ferhat and Noble, 2015).

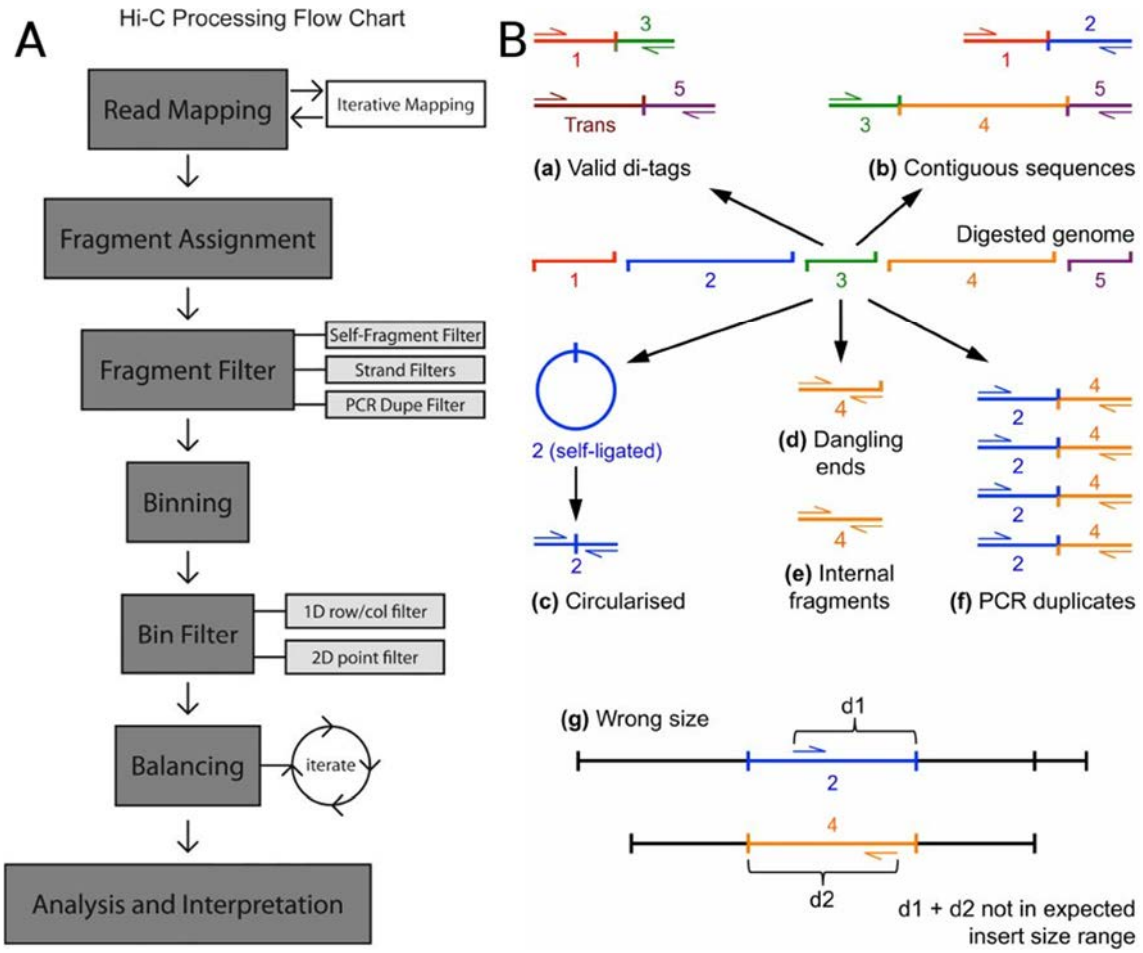


Figure 20. Overview of the Hi-C analysis workflow. (A) Workflow for Hi-C processing (adapted from Lajoie, *et al*. 2015). (B) Experimental artefacts that arise from the Hi-C library preparation (adapted from Wingett, *et al*. 2015).

After the quality check of our Hi-C data (section 5.2.2), reads were processed with TADbit (version 0.2.0.23) (Serra, *et al*. 2017), which makes use of the GEM (version 1.7.1) mapper (Marco-Sola, *et al*. 2012) to iteratively map them against the mouse genome (version mm10).

Reads were mapped from 15 bp towards using a step size of 5 bp. The filters used to remove possible artefacts were the following: “self-circle”, “dangling-end”, “error”, “extra dangling-end”, “too short”, “too large”, “duplicated”, and “random breaks”. The filter “error” removes reads coming from the same restriction fragment (like dangling-ends), but the forward and the reverse read map to the same strand. The filter “extra dangling-end” removes reads mapping on re-ligated adjacent fragments or contiguous sequences. The filters “too short” and “too large” removes those reads mapping on very short or very large restriction fragments. Finally, the filter “random breaks” remove those reads having a RE cutting site in a distance that is not consistent with the insert size distribution of the Hi-C library. The maximum molecule length parameter was set at 2 times the 99.9 percentile of the insert size distribution, returned by the “insert_size” function of TADbit. The maximum distance of a read to a cleavage site was set to the 99.9 percentile of the insert size distribution.

An in-house Python script was used for binning and data normalization. This script imported the “HiC_data” function of TADbit, read the map files generated after the artefacts filtering step, bin the reads into a square matrix of 50 Kbp, and stored the matrix into a file in NPZ format (raw matrix). Afterwards, HiCExplorer (version 1.8.1) (Ramírez, *et al*. 2018) was used to correct the raw matrix with the ICE (Iterative Correction and Eigenvector decomposition) approach the resulting matrix, setting a maximum number of iterations of 500.

5.2.4 Correlation coefficient analysis

In order to validate the reproducibility of the Hi-C replicates, pairwise comparisons between biological replicates were performed using HiCRep (version 1.4) (Yang, *et al*. 2017) under a smoothing parameter of 5 and a considered distance over 10 Mbp. Since HiCRep only handles intra-chromosome raw matrices, each pair-wise comparison yielded 20 correlation scores (19 autosomal chromosomes and the sex chromosome X). The correlation between two replicates was defined as the mean of the 20 correlation scores.

5.2.5 Inter-chromosome and intra-chromosome interaction ratio

ICE-normalised data stored in matrices were exported with HiCExplorer to the GInteractions format, which consists of 7 columns: chromosome, start and end from bin 1, chromosome, start and end from bin 2, and the amount of interaction. The GInteractions tables were imported in R for further quantification of intra-chromosome and intra-chromosome interactions and plotting.

5.2.6 Inter-subcentromeric interaction quantification

ICE-normalised matrices were scaled with a factor of 1,000,000/sum(matrix) and exported with HiCExplorer to GInteractions format. The GInteractions tables were imported in R for this inter-telomere interaction quantification. Since the telomeric and centromeric regions (annotated from the beginning of each chromosome to 2.9 Mbp according to the UCSC Table Browser) were masked due to the low-count filtering step prior to ICE normalization, we only considered inter-chromosome interactions between loci located within genomic positions 3 to 3.5 Mbp in each chromosome. Differences in the subcentromeric interaction frequencies between cell types were assessed with the Wilcoxon test.

5.2.7 Distance-dependent interaction frequency

The contact probability as a function of genomic distance, P(s), measures the probability of interaction between loci at a given distance. In this sense, ICE-corrected matrices were scaled with a factor of 1/sum(matrix). The resulting matrices were then input to "hicPlotDistVsCounts" from the HiCExplorer package in order to obtain the P(s).

5.2.8 Simulation of somatic contamination in sperm samples

In order to validate our enriched sperm population, we simulated six Hi-C sperm datasets of 100 million reads with different proportions, from 0 to 100% by steps of 20%, of fibroblast reads. Both sperm and fibroblasts reads were derived from our generated libraries. Previously published data on sperm (SRR3225862 and SRR3225863 accessions from Jung, *et al*. 2017) were also downloaded from NCBI SRA database. These datasets underwent a quality check, Hi-C data processing, binning and normalization steps (see section 5.2.2). The resulting raw Hi-C matrices were used for correlation coefficient analysis while the ICE-normalised matrices were used to calculate the averaged contact probability P(s) (see section 5.2.4).

5.2.9 A/B compartments and TADs calling

Raw matrices were used for the definition of A/B compartments. Columns with a low number of counts were filtered out using TADbit, setting the parameter "*min_count*" to 10. Since TADbit fits the column count distribution into a polynomial distribution, columns with a number of counts smaller than the first antimode of the distribution, which cannot be smaller than the *min_count* parameter, are filtered out. Then, genome-wide matrices were normalized by the expected interactions at a given distance and by visibility by means of one iteration of the ICE method. The correlation analysis was also performed with TADbit thus getting the first 5 eigenvectors. In-house scripts computed A/B compartments generally from the first eigenvector

(with the exception of P/D), using 0 as threshold to differentiate both compartments and the gene density to label them. In the case of P/D, all 5 eigenvectors were examined visually in order to select one of them for each chromosome. By convention, eigenvector values belonging to compartment A were forced to be positive values and eigenvector values belonging to compartment B were forced to be negative values.

TADs were identified using an in-house script that imported the "Chromosome" module of TADbit and using the raw and the ICE-normalised matrices as input, each chromosome separately. Filtered bins due to low count were considered in order to mask those regions at the time to call TADs.

TAD signal is referred to the insulation score, defined as the average of interactions in a sliding window diamond along the matrix diagonal (Lajoie, *et al*. 2015). In this context, TAD signals for each cell type were obtained by first normalizing the different matrices in terms of number of reads. Each matrix was then scaled with a factor of 100,000,000/sum(matrix) by means of a custom script. Afterwards, TAD signals were obtained from the output given by the "hicFindTADs" program from HiCExplorer.

5.2.10 Compartment switching

BED files with a resolution of 50 Kbp were available from the A/B compartments calling step (see section 5.2.7). Each genomic bin of 50 Kbp had its corresponding compartment attributed. Pairwise comparisons between cell types (genome-wide and per-chromosome) were performed; the ratio of compartment switching was calculated as the number of genomic bins with a compartment change (A>B or B>A) divided by the total number of bins. From these files, a matrix was created with 50 kbp-binned genomic coordinates as rows and cell types as columns, filled by the corresponding compartment labelling in each bin and cell type. Cell-specific A compartments were defined as those bins being compartment A in a cell type and compartment B in the remaining cell types.

5.2.11 Compartments and gene expression relationship

Compartments A and B of each cell type were intersected with BEDTools (version 2.26) against a BED file with the TSS of genes derived from the GRCm38 gene annotation from Ensembl (release 89). Genes in each compartment were grepped (Bash command) with the table of FPKM values downloaded from AIR (see chapter 4), generating the expression profiles represented as boxplots for each cell type and compartment. Statistical significance among pairwise comparisons was tested using the Wilcoxon test using a p-value threshold of 0.05.

5.2.12 HiCloud genome browser

HiCloud is a user-friendly tool developed within the framework of this thesis to visualize and integrate Hi-C data with other epigenetic features into a genome browser in order to browse the results all encapsulated into a unique single place. HiCloud uses technologies such as NodeJS in the back-end and HTML/CSS in the front-end and takes advantage of HiCExplorer to generate the graphical output (supplementary figure 30).

5.3 Results

5.3.1 Quality metrics and correlation coefficient analysis

After quality-trimming, we detected that on average, no more than 5% of reads were lost (table 10; supplementary table 3). Then, quality-trimmed reads were mapped, obtaining an average mapping efficiency (forward and reverse reads uniquely mapped) of 72.1%. These reads were subsequently filtered to remove Hi-C artefacts, such as duplicates (9.8%), reads mapping in very short restriction fragments ("too short") (5.8%), dangling-ends (3.8%) and extra dangling-ends (3.8%) the greatest sources or artefacts (table 10; supplementary table 3). Interestingly, the average of dangling-ends in all samples but round spermatids was 1.8% while in round spermatids alone was 15.8%. After the artefact removal step, an average of 265 million of paired-end reads was obtained per cell type (table 10, supplementary table 3).

Table 10. Hi-C quality metrics per cell type. It includes the number of reads before and after quality check and trimming, the number of reads that mapped once in the genome (uniquely mapped), the percentage of reads classified as artefacts (see section 5.2.3), and the final number of valid reads. Legend: Fib: fibroblast; Sg: spermatogonia; L/Z: leptonema/zygonema; P/D: pachynema/diplonema; SpII: secondary spermatocytes; RS: round spermatids.

INFO PER CELL TYPE	Fib	Sg	L/Z	P/D	SpII	RS	Sperm
Raw read pairs	500514408	1428752701	225472936	761934779	444722954	494535554	1128671379
Trimmed read pairs	486002613	1366310545	216917242	724352845	423752920	471961823	1087438630
Uniquely mapped read pairs	346917699	387340642	158331725	522268513	314131439	335780080	324425739
Self-circle (% relative uniquely mapped)	0.12	0.17	0.11	0.20	0.19	0.30	0.09
Dangling-end (% relative uniquely mapped)	1.72	0.33	0.06	3.25	2.07	15.84	0.26
Error (% relative	0.86	0.03	0.03	0.77	1.23	3.55	0.03

uniquely mapped)							
Extra dangling-end (% relative uniquely mapped)	3.03	3.58	4.95	4.24	2.68	6.40	4.15
Too short (% relative uniquely mapped)	5.56	6.97	6.55	5.85	4.90	6.43	5.50
Too large (% relative uniquely mapped)	0.01	0.00	0.00	0.00	0.00	0.01	0.00
Duplicated (% relative uniquely mapped)	4.74	8.34	3.98	8.94	12.65	16.03	5.87
Random breaks (% relative uniquely mapped)	0.38	0.14	0.01	0.60	0.34	2.75	0.16
Total valid read pairs	293977083	316529749	134826221	411123699	244223890	189129404	276055642
Total valid (% relative to Raw)	58.73	22.15	59.8	53.96	54.92	38.24	24.46
Total valid (% relative to Trimmed)	60.49	23.17	62.16	56.76	57.63	40.07	25.39
Total valid (% relative to Mapped uniquely)	84.74	81.72	85.15	78.72	77.75	56.33	85.09

As we analysed different replicates per cell type (excluding leptonema/zygonema), we calculated pairwise correlation scores among them with the aim to assess the reproducibility of the Hi-C data generated. Pairwise correlation coefficients showed high correlation scores (between 0.82 and 0.98), thus validating our results (figure 21). Among cell types, primary spermatocytes at pachynema/diplonema stage showed high correlation values (between 0.82 and 0.86) with primary spermatocytes at leptonema/zygonema stage, suggesting similarities in the higher-order chromatin structure during prophase I. The correlation between round spermatids and secondary spermatocytes was also high (>0.90).

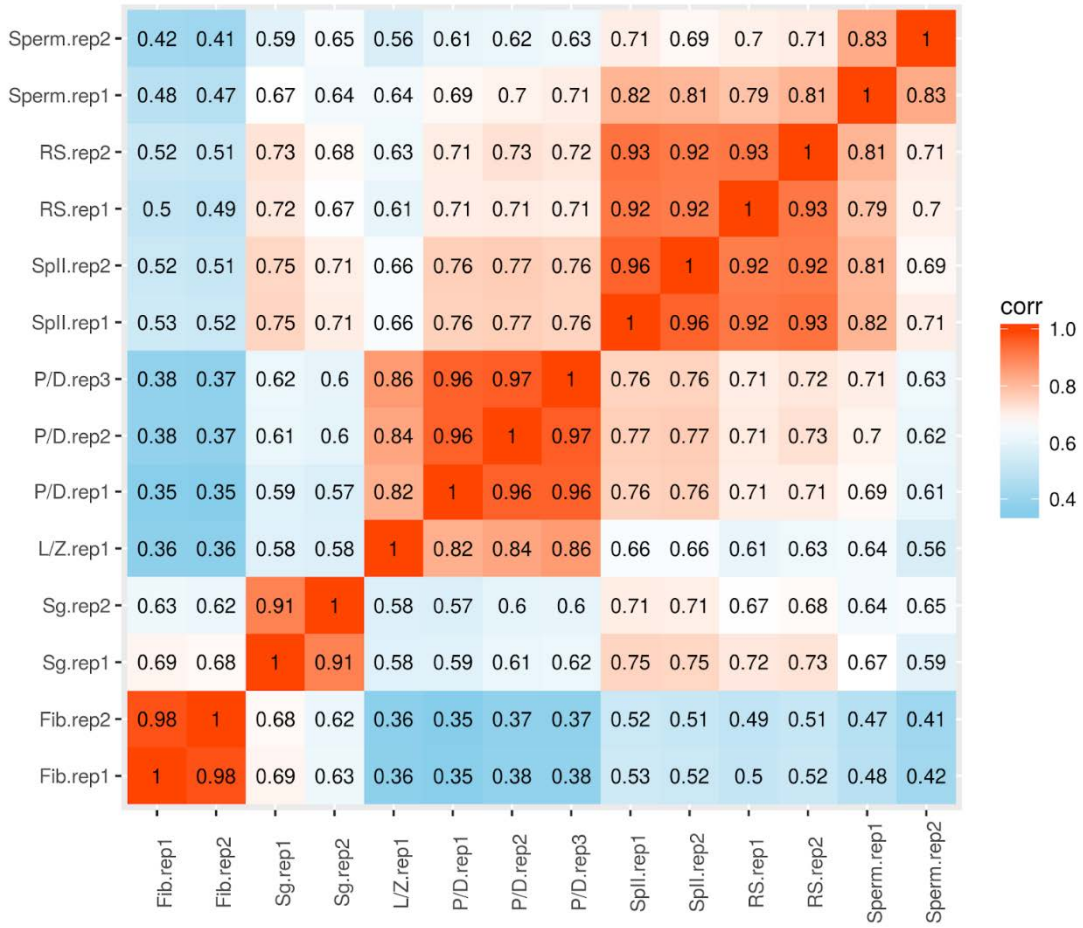


Figure 21. Heatmap with correlation values among replicates. It is based on the pairwise similarity score calculated using HiCRep. Legend: Fib: fibroblast; Sg: spermatogonia; L/Z: leptonema/zygonema; P/D: pachynema/diplonema; SpII: secondary spermatocytes; RS: round spermatids; rep1: replicate 1; rep2: replicate 2.

5.3.2 The higher-order chromatin structure along spermatogenesis

Once the reproducibility of the Hi-C data was confirmed, interaction matrices from replicates from the same cell type were merged thus obtaining a more representative interaction matrix for each cell type. From these matrices, genome-wide heatmaps were created at 500 Kbp resolution and per-chromosome heatmaps at 50 Kbp resolution (supplementary figures 1-8). Genome-wide heatmaps show the interaction pattern among different chromosomes while per-chromosome heatmaps show the interaction pattern within the same chromosome.

5.3.2.1 Inter-chromosome and intra-chromosome interaction ratio

Genome-wide Hi-C heatmaps reveal certain chromosome organisation patterns. That is, there is higher interaction frequency between pair of loci from the same chromosome than between pair of loci from different chromosomes (Lajoie, *et al.* 2015). Interactions within the same chromosome are called intra-chromosome interactions while interactions between

chromosomes are called inter-chromosome interactions. In this sense, we quantified the ratio of interchromosome and intrachromosome interactions for each chromosome and cell type to interrogate whether chromosomes have different interaction patterns.

The inter/intra-chromosome interaction ratio was below one in all chromosomes of all cell types with the exception of chromosome 19 in fibroblast and the vast majority of chromosomes in sperm, meaning that these chromosomes have more inter-chromosome interactions than intra-chromosome interactions. In addition, an upward tendency was shown in the autosomal chromosomes of fibroblasts, spermatogonia, round spermatids and sperm as the inter/intra-chromosome interaction ratio slightly increased with the chromosome number and, therefore, with the chromosome size (figure 22). In contrast, the ratio in chromosome X was remarkably reduced, in comparison with similar-size autosomal chromosomes, in round spermatids, sperm and, more weakly, in pachynema/diplonema. In all chromosomes in leptonema/zygonema and in all autosomes in pachynema/diplonema, the ratio of inter/intra-chromosome interactions remained stable across all chromosomes.

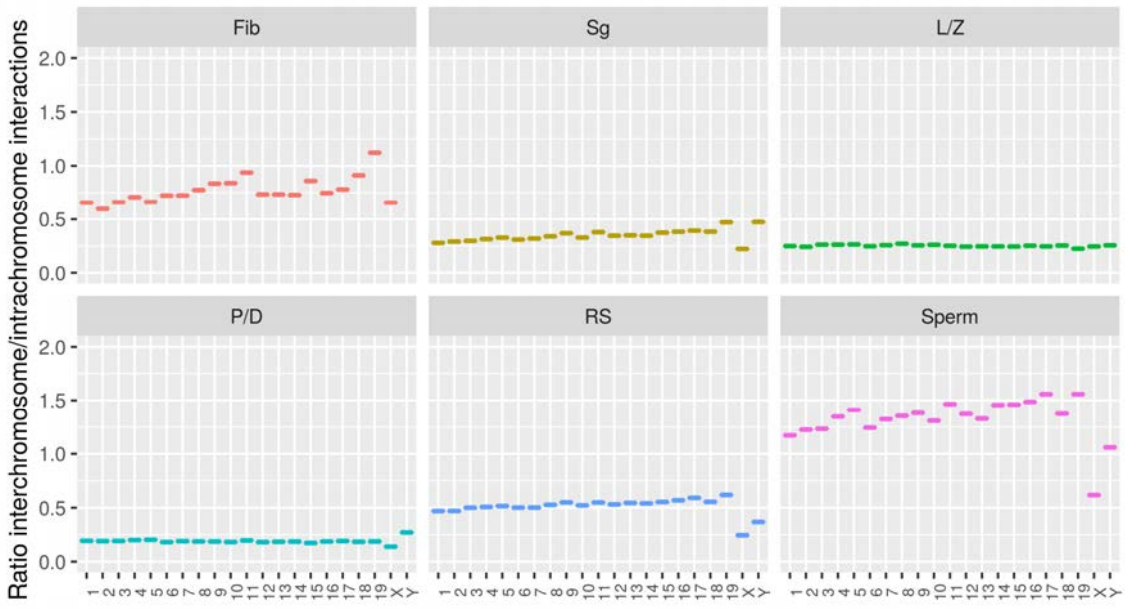


Figure 22. Inter-chromosome/intra-chromosome interaction ratio for each chromosome and cell type. Legend: Fib: fibroblast; Sg: spermatogonia; L/Z: leptonema/zygonema; P/D: pachynema/diplonema; RS: round spermatids.

5.3.2.2 Distance-dependent interaction frequency

The amount of interaction between a pair of loci located in the same chromosome depends on the genomic distance between them. Nearby loci are more likely to interact; thus, the amount of interaction decreases when the genomic distance between two loci increases (Lieberman-Aiden, *et al*. 2009; Lajoie, *et al*. 2015). However, the strength of this interaction decrease is

related with chromosome organization. Lieberman-Aiden and colleagues described a lineal decrease of interactions between distances from 500 Kbp to 7Mbp with a slope of -1.08 (Lieberman-Aiden, *et al*. 2009). This slope was compared with the slopes given by two different polymeric models: (i) the "fractal" globule, which represents a self-organized polymer with a long-lived and non-equilibrium conformation, and (ii) the "equilibrium" globule, which represents a polymer with a densely knotted conformation. Since the "fractal" and the "equilibrium" models showed slopes of -1 and -3/2, respectively, the "fractal" globule was considered to be a more suitable model for chromatin organization during interphase (Lieberman-Aiden, *et al*. 2009; Mirny, 2011).

The average intrachromosomal contact probability as a function of genomic distance, P(s), was investigated as a way to analyse whether the dynamics of the higher-order chromatin structure during spermatogenesis was also translated into differences in levels of chromosome organisation. Firstly, we computed the contact probability as a function of genomic distance P(s) (figure 23A). In general terms, all cell types appear to follow the same pattern at shorter distances (below 0.5 Mbp). In contrast, at medium distances (from 0.5 to 7 Mbp), the decay in the interaction of loci is slower in primary spermatocytes, followed by the group of cells composed of secondary spermatocytes, round spermatids and sperm. The group of cells composed of fibroblasts and spermatogonia has the fastest interaction decay at these distances. However, at long distances (above 7 Mbp), fibroblasts and spermatogonia show the slowest decay, suggesting that interaction of loci separated by long distances is maintained. Primary spermatocytes show an abrupt drop in the interaction of loci at long distances.

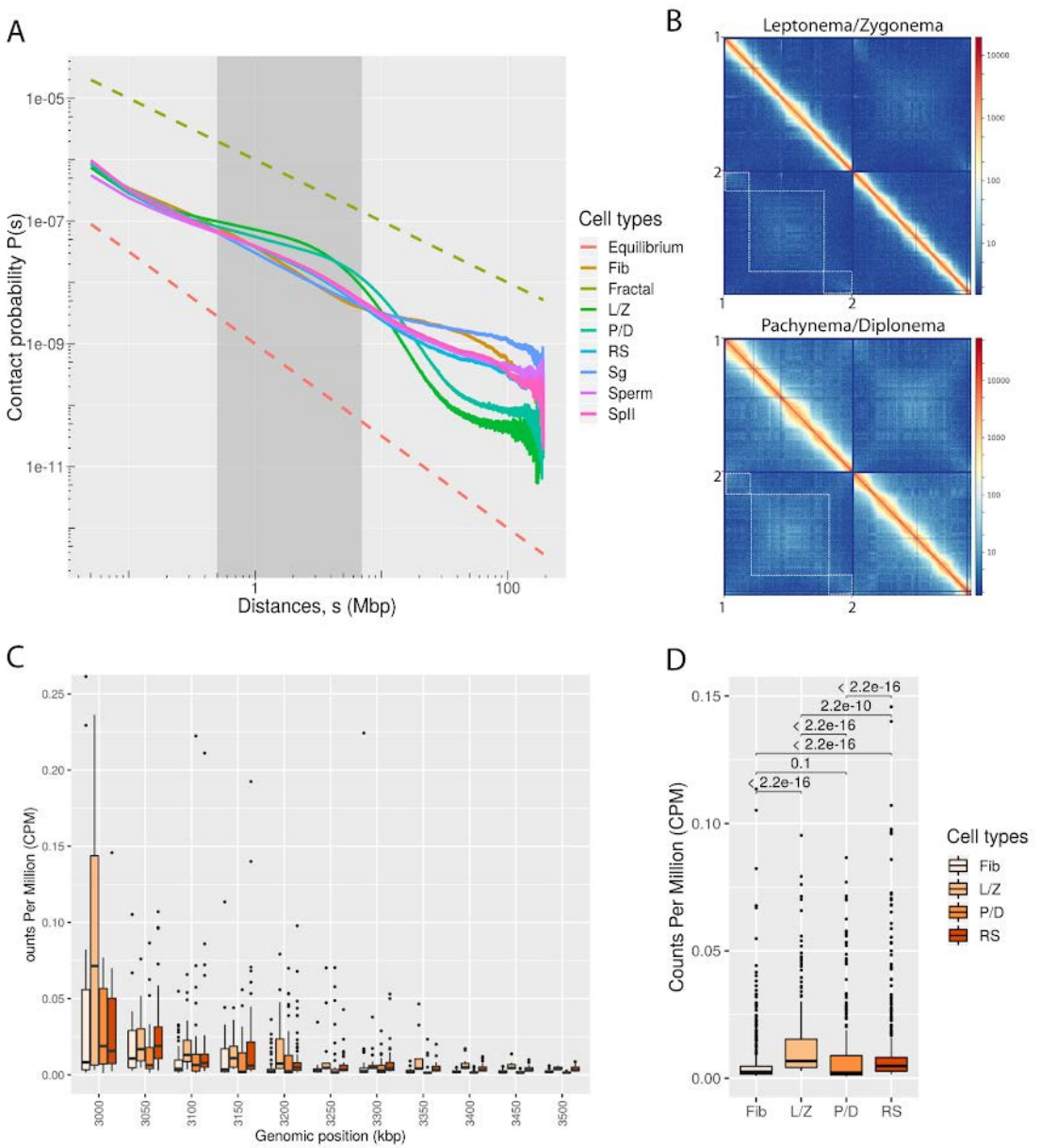


Figure 23. Intra and inter-chromosomal contacts. (A) Contact probability as a function of genomic distance P(s) for fibroblasts and the different germ cell types. Discontinuous straight lines correspond to the fractal (green) and equilibrium (red) models. Grey-shadowed area expands the genomic region from 0.5 to 7 Mbp. (B) Inter-chromosomal interactions between mouse chromosomes 1 and 2 in leptonema/zygonema and pachynema/diplonema stages. Dotted-white boxes show high contact regions. (C) Boxplots showing inter-chromosomal interactions (CPM) in subcentromeric regions (from 3000 to 3500 Kbp) among all cell types analysed. (D) Boxplots showing inter-chromosomal interactions (CPM) in subcentromeric regions (3000-3500 Kbp) among all cell types analysed. Inter-subcentromeric interactions in L/Z and RS were significantly higher relative to fibroblast (Wilcoxon test, p-value < 2.2e-16). Legend: Fib: fibroblast; Sg: spermatogonia; L/Z: leptonema/zygonema; P/D: pachynema/diplonema; RS: round spermatids.

Specifically, fibroblasts and spermatogonia, both being in interphase stage, shared a similar contact probability P(s) patterns with slopes from 0.5 to 7 Mbp of -1.20 (r^2 = 0.99) and -1.03 (r^2 = 0.99), respectively (figure 23A). However, at the genomic distance of 10 Mbp, fibroblasts presented a slighted change in the slope by lowering the amount of interactions at longer

distances whereas spermatogonia maintains the slope up to 100 Mbp distance. On the other hand, prophase I cells display two abrupt changes in slopes; the first one between 2.5 and 4.5 Mbp and the second one at 40 Mbp (figure 23A). The former slope change makes prophase I cells not to fit with the power-law decay described by Lieberman-Aiden, *et al*. (2009) between distances from 0.5 to 7 Mbp, so the slopes were calculated from 0.5 to 2.5 Mbp for L/Z (slope = -0.51, r^2 = 0.99) and from 0.5 - 4.5 Mbp for P/D (slope = -0.60, r^2 = 0.99). This pattern resembles what it has been previously reported for the mitotic chromosome (slope = 0.5; Naumova, *et al*. 2013). Specifically, the change in the slope observed in the meiotic chromosome between 2.5 - 4.5 Mbp preceded by a rapid fall-off resembles the drop of the P(s) curve recently described at 2 Mbp in prometaphase cells (Gibcus, *et al*. 2018). In case of secondary spermatocytes, round spermatids and sperm, they shared a similar contact probability P(s) patterns with slopes from 0.5 to 7 Mbp of -0.89 (r^2 = 0.98), -0.96 (r^2 = 0.98) and -0.92 (r^2 = 0.98), respectively (figure 23A).

Remarkably, the presence of enriched inter-chromosomal contacts between telomeres in both leptonema/zygonema and pachynema/diplonema stages, being more prominent in the former, was observed (figure 23B). In order to know whether this observation was statistically significant, we quantified the number of counts between bins with the same start coordinates at the subcentromeric regions from different chromosomes in fibroblast, leptonema/zygonema, pachynema/diplonema and round spermatids (figure 23C). Especially leptonema/zygonema but also round spermatids presented a significantly higher amount of interaction (Wilcoxon test, p-value < 0.05) when compared to fibroblast (figure 23D).

5.3.2.3 Genomic compartments

Compartments belong to the sub-chromosome organization scale and consists of alternated genomic regions of “open” and “closed” chromatin states termed “A” and “B”, respectively. On the one hand, compartments A are correlated with histone modifications that characterize accessible chromatin, such as H3K9ac, H3K27ac, H3K36me3, or H3K4me3 (Barski, *et al*. 2007; Araki, *et al*. 2009). On the other hand, compartments B are correlated with histone modifications associated to “closed” chromatin states, such as H3K27me3 or H3K9me3 (Barski, *et al*. 2007; Araki, *et al*. 2009). In this sense, the definition of compartments might provide insights into the chromosome organization changes and their functional roles during spermatogenesis.

Looking at the chromosome-specific interaction heatmaps, plaid patterns, which are indicative of the presence of A/B compartments, were well defined in the case of fibroblasts and spermatogonia (figure 24A-B; supplementary figures 2-3). On the contrary, such patterns were

mainly lost in primary spermatocytes, especially in leptonema/zygonema (figure 24C-D; supplementary figures 4-5). Precisely, leptonema/zygonema correspond to initial stages of meiotic prophase I, where homologous chromosome condensate, align, pair and synapse. In the case of secondary spermatocytes, round spermatids and sperm, they show a blurry-like plaid pattern suggesting an intermediate status between fibroblasts and leptonema/zygonema (figure 24E-F; supplementary figures 6-8).

Changes in chromosome conformation were also evident when analysing the dynamics of compartment profiles using the eigenvector decomposition (Imakaev, *et al*. 2012) (figure 25A-B). The eigenvector values showed blocks of contiguous positive and negative values in both fibroblasts and spermatogonia, showing the presence of compartments in these cell lines. Nevertheless, at the beginning of prophase I, the eigenvector values were close to 0 in leptonema/zygonema stage, consistent with an absence of compartments. Compartments appear again in the late prophase I, where the synaptonemal complex unassemble, as the eigenvector in pachynema/diplonema shows again blocks of contiguous positive and negative values (figure 25A). Nevertheless, unlike fibroblasts and spermatogonia, the first eigenvector did not represent compartments in the vast majority of chromosomes in pachynema/diplonema. In the case of round spermatids and sperm, the first eigenvector was consistent with the presence of compartments (figure 25A-B).

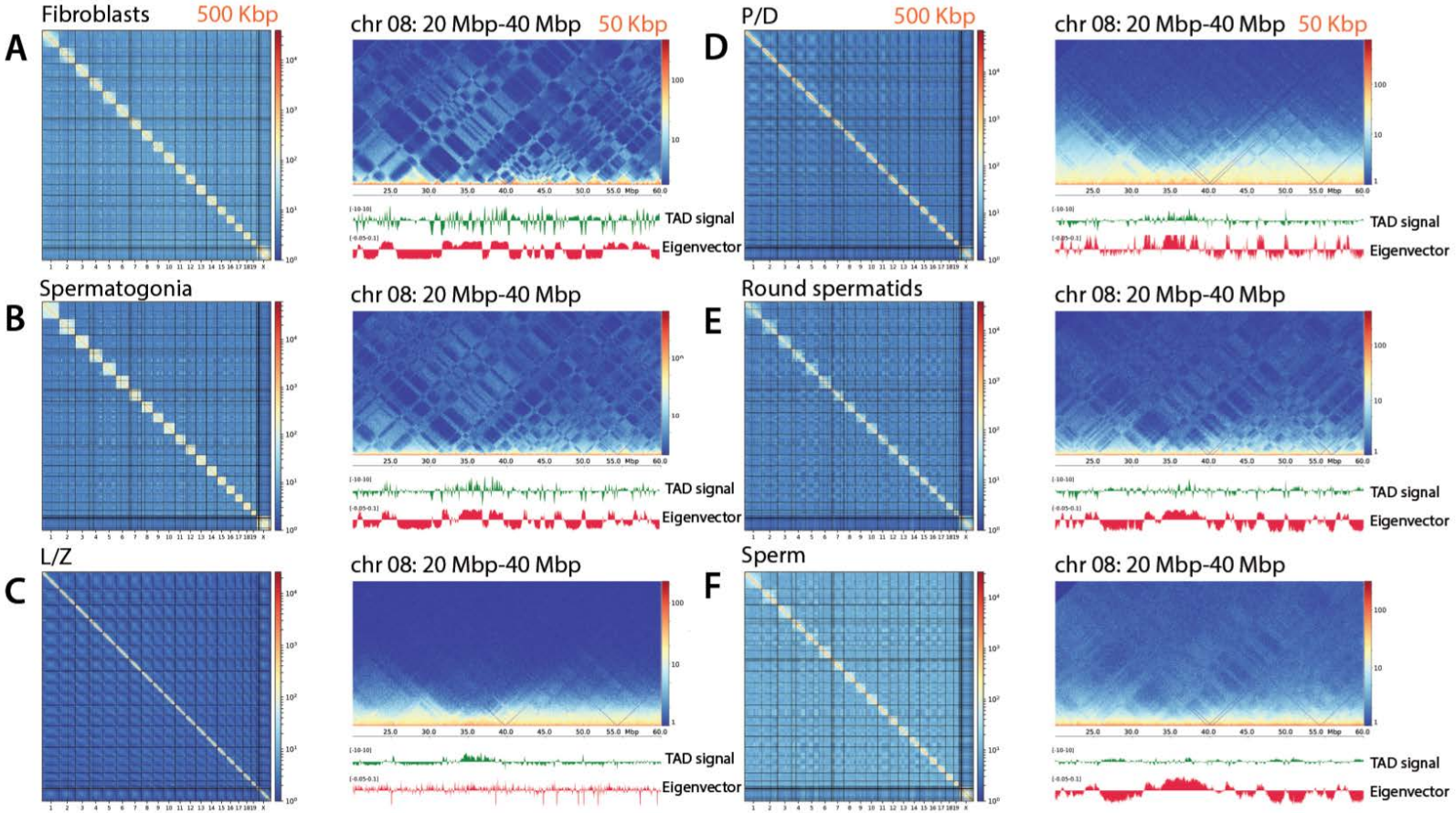


Figure 24. Chromosomal organization during in interphase, pre-meiotic, meiotic and post-meiotic cells. Genome-wide ICE-corrected heatmaps at 500 kbp and chromosome 8 region-specific ICE-corrected heatmaps at 50 kbp with TAD and compartment signal (eigenvector values) for (A) fibroblast, (B) spermatogonia, (C) leptonema/zygonema (L/Z), (D) pachynema/diplonema (P/D), (E) round spermatids, and (F) sperm.

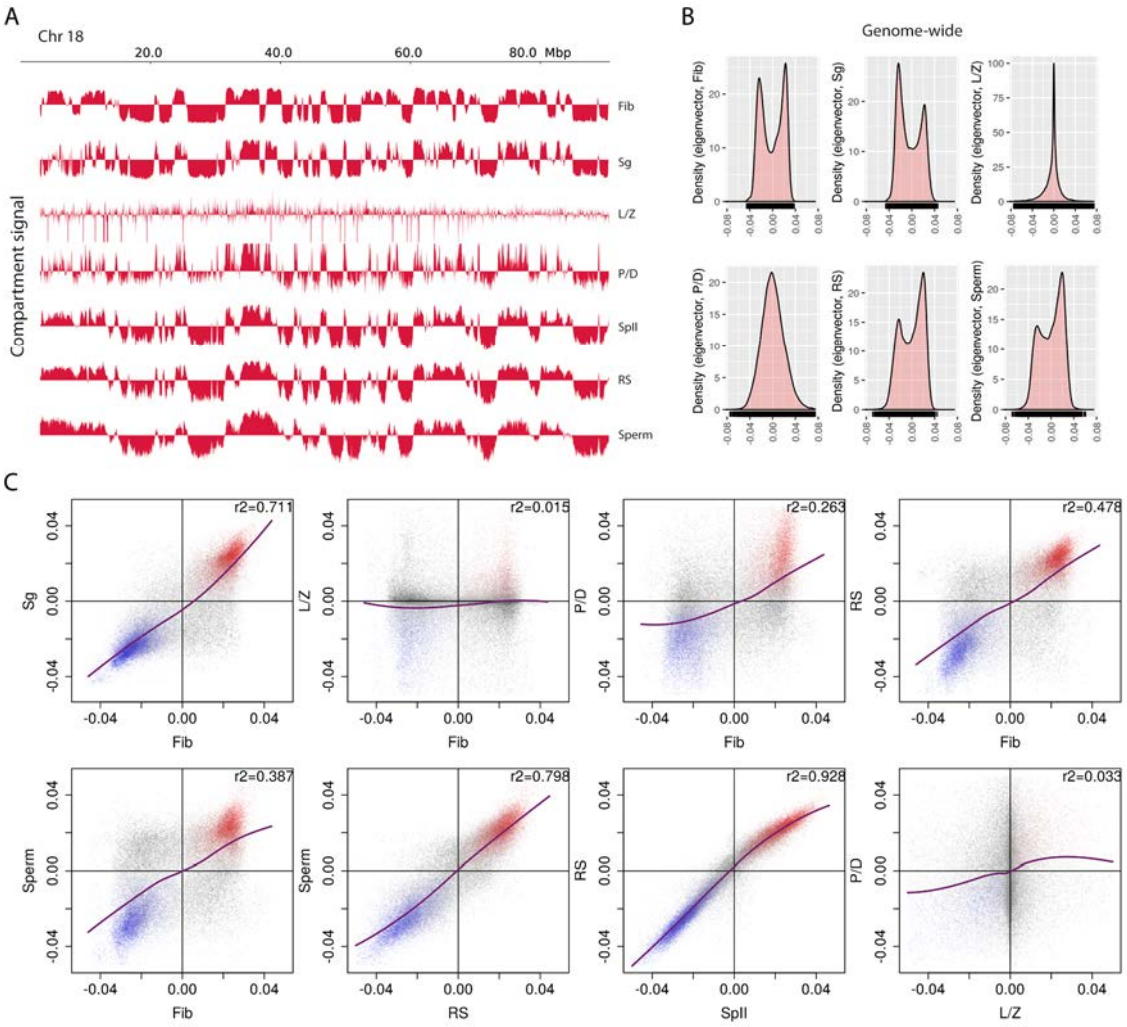


Figure 25. Sub-chromosome organization scale and eigenvector decomposition. (A) Representation of compartment signal (eigenvector values) across mouse chromosome 18 in all cell types. (B) Density plots representing eigenvector values for each cell type, considering all chromosomes but sex chromosomes. (C) Pair-wise representation of eigenvectors between cell types genome-wide. Each dot represents a 50Kbp bin in the genome. Bins representing A compartment conservation are depicted in red, whereas in blue are depicted bins with B compartment conservation. Bins with unclear signal or compartment switching are represented in grey. The purple line is a LOESS curve showing the tendency of the compartment switching. R-squared correlation values are represented for each pairwise comparison. Legend: Fib: fibroblast; Sg: spermatogonia; L/Z: leptonema/zygonema; P/D: pachynema/diplonema; SpII: secondary spermatocytes; RS: round spermatids.

Apart from leptonema/zygonema, the remaining cell types analysed showed compartmentalization. In this sense, regression analyses among the eigenvectors of cell types were carried out as a way to test the degree of compartment conservation during spermatogenesis (figure 25C). R-squared values from linear regressions were particularly high in the case of secondary spermatocytes *versus* round spermatids (r^2=0.93), denoting very high degree of compartment conservation. R-squared values were also high in round spermatids *versus* sperm (r^2=0.80). In the case of fibroblast, the R-squared value *versus* spermatogonia showed a remarkable compartment conservation (r^2=0.71). Nevertheless, as spermatogenesis progresses, the degree of compartment conservation in fibroblast *versus*

pachynema/diplonema, round spermatids, and sperm decreases considerably (r^2=0.26, r^2=0.48, and r^2=0.39, respectively). Finally, as expected due to the absence of compartments (figure 25A), the eigenvector of leptonema/zygonema has no correlation with either fibroblast or pachynema/diplonema (r^2=0).

5.3.2.4 Topologically associating domains

While compartments denote the chromosome organization at the megabase scale, the organization at the sub-megabase scale is revealed by the presence of TADs (Sexton, *et al*. 2012; Dixon, *et al*. 2012). TADs represent loop-like structures of several kilobases in size with elevated interaction frequencies between loci located within the same TAD, where promoter-enhancer contacts take place (Lajoie, *et al*. 2015). To further investigate the dynamics of the higher-order chromatin structure during spermatogenesis at the sub-megabase scale, we analysed TAD insulator scores (TAD signal), TADs size and the robustness of TAD boundaries.

We first quantified the TAD signal as insulation scores in all cell types (figure 26A). In fibroblasts and spermatogonia, the TAD signal was highly variable thus indicating the presence of these structures (figure 26A-B). In contrast, consistent with the A/B compartment patterns, we detected a substantial reduction in the variance of the TAD signal in leptonema/zygonema (figure 26B). The TAD signal slightly recovered in pachynema/diplonema, secondary spermatocytes and round spermatids, but was drastically reduced in sperm (variance of TAD signal close to 0). Hi-C heatmaps focused on specific genomic regions confirm these patterns (supplementary figures 9-15).

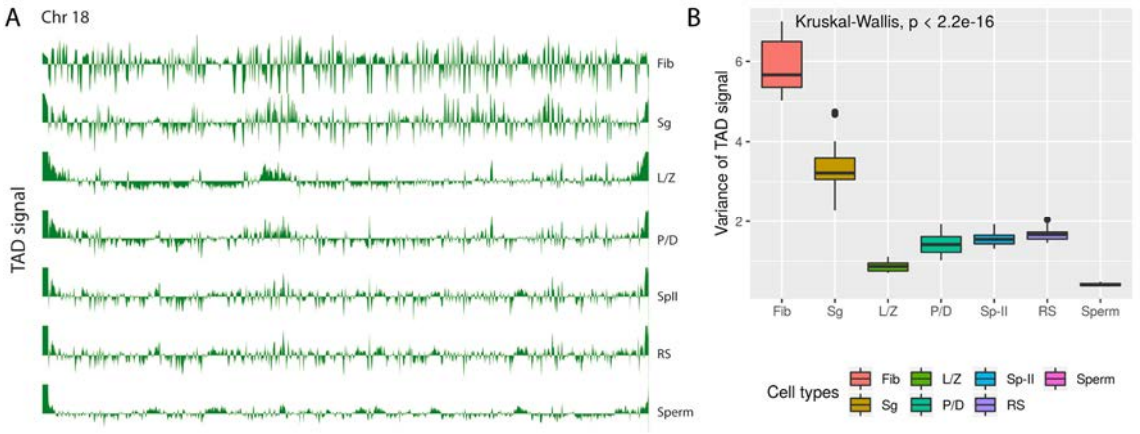


Figure 26. Sub-megabase organization scale and TAD signal. (A) Representation of the TAD signal (insulation score) across mouse chromosome 18 in all cell types. (B) Variance of the TAD insulation scores in all cell types, considering all chromosomes but sex chromosomes. The Kruskal-Wallis statistical test reveals TAD signal variances are different among the cell types analysed (p-value < 2.2e-16). Legend: Fib: fibroblast; Sg: spermatogonia; L/Z: leptonema/zygonema; P/D: pachynema/diplonema; SpII: secondary spermatocytes; RS: round spermatids.

The extreme reduction of the TAD signal observed in sperm cells differed from what has been previously described in the literature, in which the presence of TADs in sperm has been reported (Ke, *et al*. 2017; Jung, *et al*. 2017; Wang, *et al*. 2019). To further validate the pattern observed in our enriched sperm population, we simulated Hi-C datasets derived from sperm including different proportions (from 0 to 100%, by steps of 20%) of reads from fibroblast Hi-C libraries in order to test whether the patterns observed previously (Ke, *et al*. 2017; Jung, *et al*. 2017; Wang, *et al*. 2019) are due to the presence of somatic contamination. Our sample containing 100% sperm showed a very distinct pattern with those samples containing some percentage of fibroblast reads (figure 27). The sample SRP071784 (retrieved from Jung, *et al*. 2017) has the highest correlation (0.72) with the simulated sample with 20% sperm and 80% fibroblast or 40% sperm and 60% fibroblast (figure 27A). In addition, the contact probability as a function of genomic distance shows that the pattern of sample SRP071784 is much more similar to fibroblast than to sperm (figure 27B). Also, Hi-C heatmaps show that a minority of fibroblast reads in a sperm sample dramatically disrupts the interaction pattern (figure 27C). Therefore, these results suggest the sample SRP071784 is much more similar to our fibroblast data than to our sperm data.

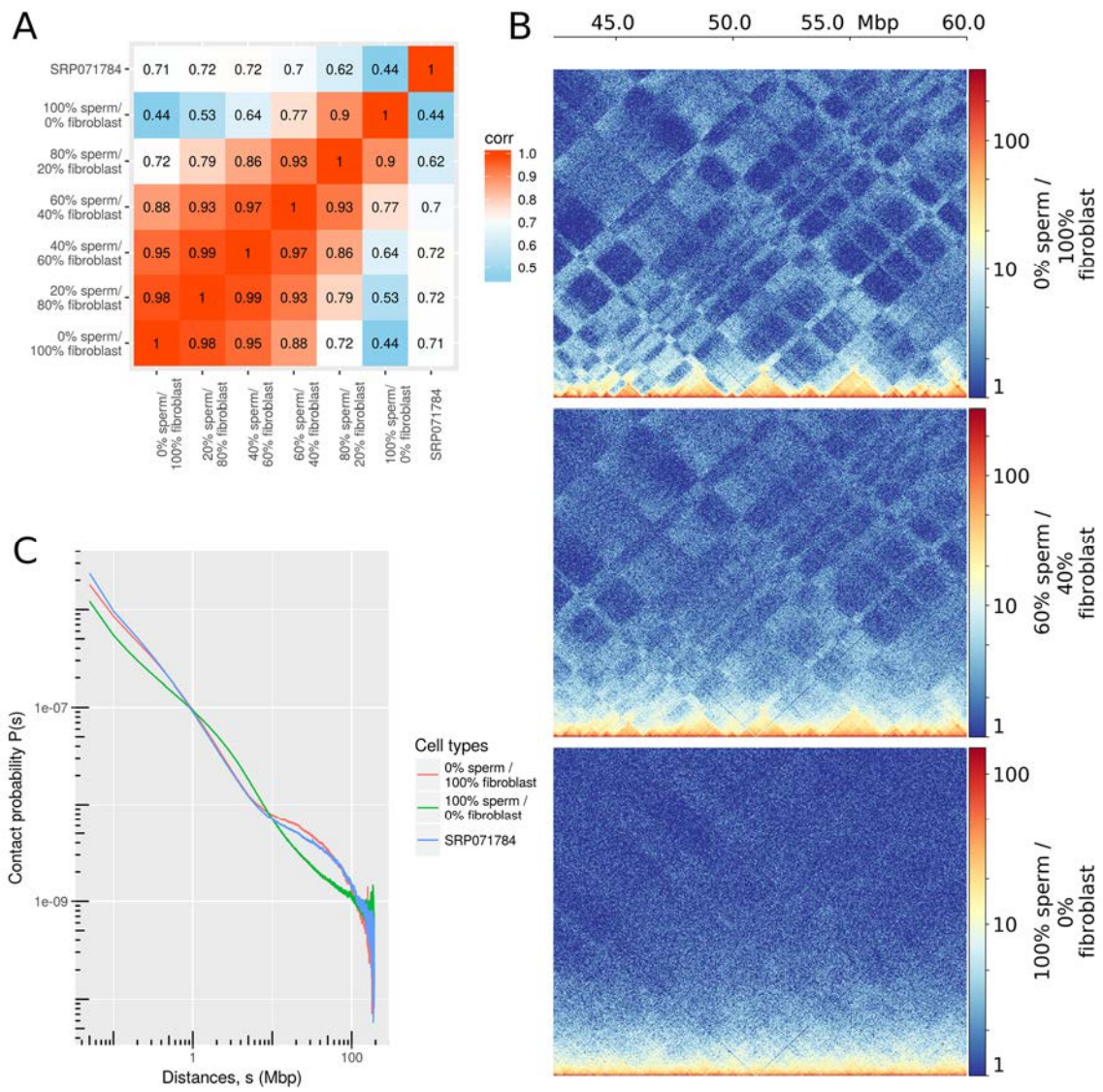


Figure 27. Simulations of samples with different fibroblast and sperm content. (A) Heatmap showing correlation values, based on the pairwise similarity score calculated using HiCRep, among samples with different fibroblast and sperm content as well as the merged sperm replicates from SRP071784. (B) Contact probability as a function of genomic distance for samples with different fibroblast and sperm content as well as the merged sperm replicates from SRP071784 (retrieved from Jung, *et al*. 2017). (C) Chromosome 8 region-specific ICE-corrected heatmaps at 50 kbp for samples with different fibroblast and sperm content as well as the merged sperm replicates from SRP071784.

In terms of the number of TADs, TADbit identified a total of 2,002 TADs with an average length of 1.3 Mbp in fibroblasts, a number slightly higher than in spermatogonia (834 TADs, mean size of 3.26Mbp) (figure 28; supplementary table 4). Although the total number of TADs was dramatically reduced in primary spermatocytes (305 TADs in L/Z and 294 TADs in P/D), the strength score of TAD boundaries was extremely high (74.25% and 79.59% of TADs with scores between 10 and 9 in L/Z and P/D, respectively, an indicator of a high confidence in the prediction (supplementary table 4). This pattern contrasted with secondary spermatocytes and

round spermatids; both cell types presented large numbers of small TADs but low boundaries scores (figure 28; supplementary table 4).

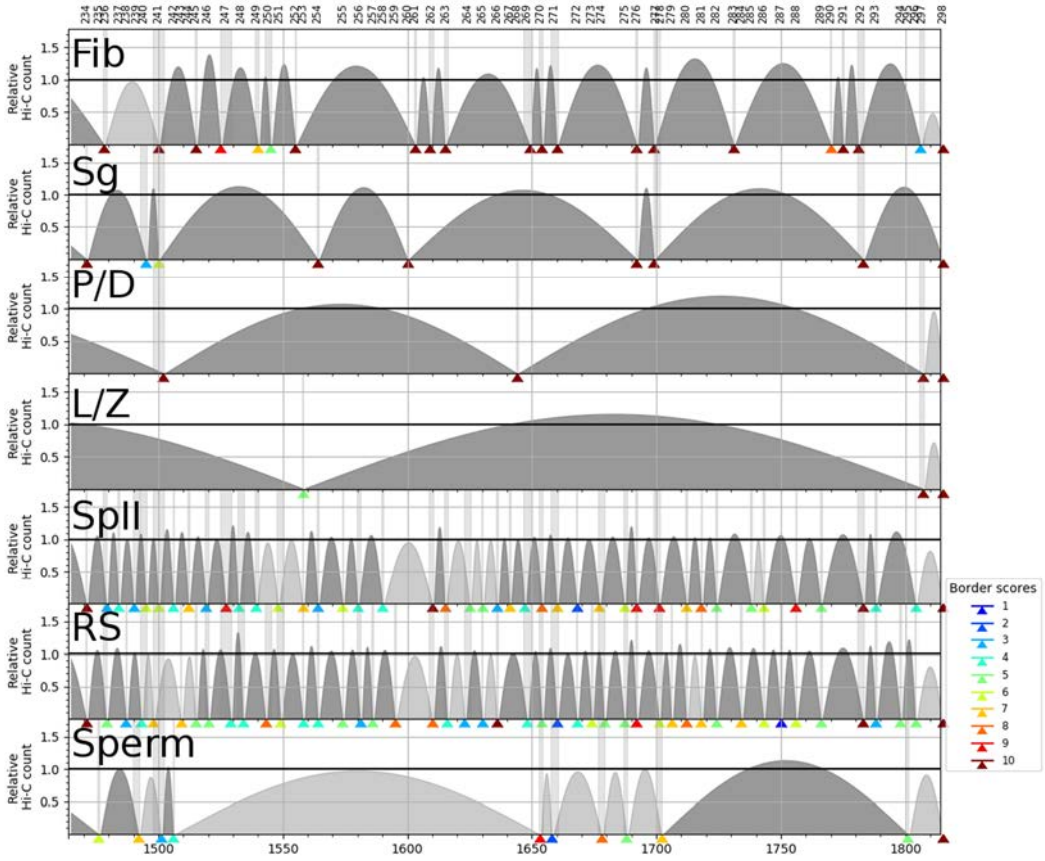


Figure 28. TAD border alignment in chromosome 18 (70-90 Mbp) between all cell types. Dark and grey arches represent TADs with higher and lower than expected intra-TAD interactions, respectively. TAD border robustness (from 1 to 10) is represented by a colour gradient.

5.3.4 Functional compartment switching during spermatogenesis

As detailed above, we detected changes in genome organization during spermatogenesis at different levels: (i) different inter-/intra-chromosome interaction patterns (see 5.3.2.1), (ii) distance-dependent interaction frequencies (see section 5.3.2.2), (iii) genomic compartments (see section 5.3.2.3), and (iv) TAD signal and number of TADs (see section 5.3.2.4). Also, the analysis of RNA-seq data showed that germ cells presented different transcriptional profiles (see chapter 4). In the light of these observations and considering that compartments A have been described to be associated with open chromatin state regions (Lieberman-Aiden, *et al*. 2009), we integrated gene expression with cell-specific A compartments as a way to investigate the relationship between 3D structure and function during spermatogenesis.

5.3.4.1 Insights on compartment switching and gene expression

We first quantified the percentage of genome labelled as compartment A in each cell type and then we analysed compartment switching during spermatogenesis (figure 29). The proportion of the mouse genome organised in compartments A was 45.68% in fibroblast, reduced in spermatogonia (39.36%) to raise again in round spermatids (46.94%) and sperm (48.64%). In terms of compartment switching, 13.9% of compartments change from A to B or from B to A between fibroblasts and spermatogonia, 22.8% between spermatogonia and pachynema/diplonema, 24% between pachynema/diplonema and secondary spermatocytes, 6.3% between secondary spermatocytes and round spermatids, and 11.28% between round spermatids and sperm.

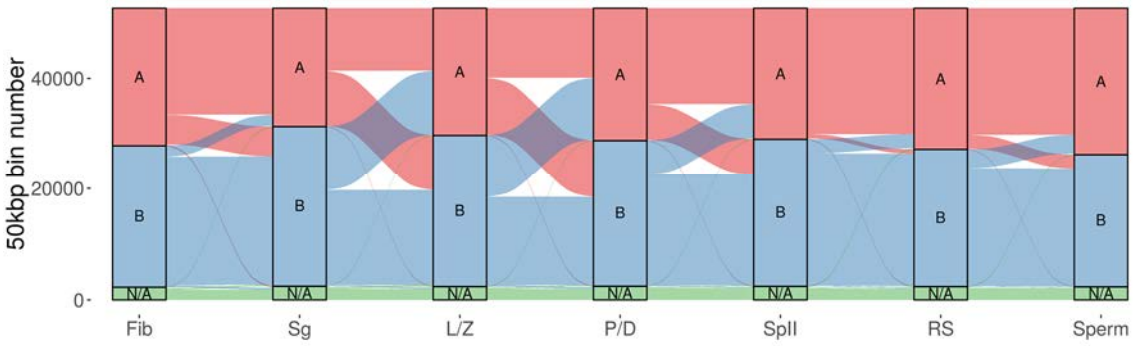


Figure 29. Alluvial plot showing the global dynamics of A/B compartment switch during spermatogenesis. Each line represents a 50 kbp bin in the genome. Legend: Fib: fibroblast; Sg: spermatogonia; L/Z: leptonema/zygonema; P/D: pachynema/diplonema; SpII: secondary spermatocytes; RS: round spermatids; N/A: not assigned.

RNA-seq data generated from spermatogonia, pachynema/diplonema, round spermatids and sperm (see chapter 4) was then integrated with compartments assignment. To this aim, raw expression values from AIR were downloaded and converted to Counts Per Million (CPM). Since the X chromosome behaved in a different way than autosomes in terms of inter-chromosome/intra-chromosome interaction ratio, we performed the following analyses considering autosomes and the X chromosome separately.

Consistent with the presence of a relationship between chromatin remodelling and active transcription, genes located in compartments A were significantly more expressed than those in B compartments across all cell types in autosomal chromosomes (Wilcoxon test, p-value < 2.2e-16) (figure 30). This pattern was also confirmed for chromosome X in spermatogonia (Wilcoxon test, p-value < 6.9e-16). Nevertheless, there was no significant differences in gene expression

between compartments A and B in chromosome X in pachynema/diplonema, round spermatids and sperm (Wilcoxon test, p-value > 0.05).

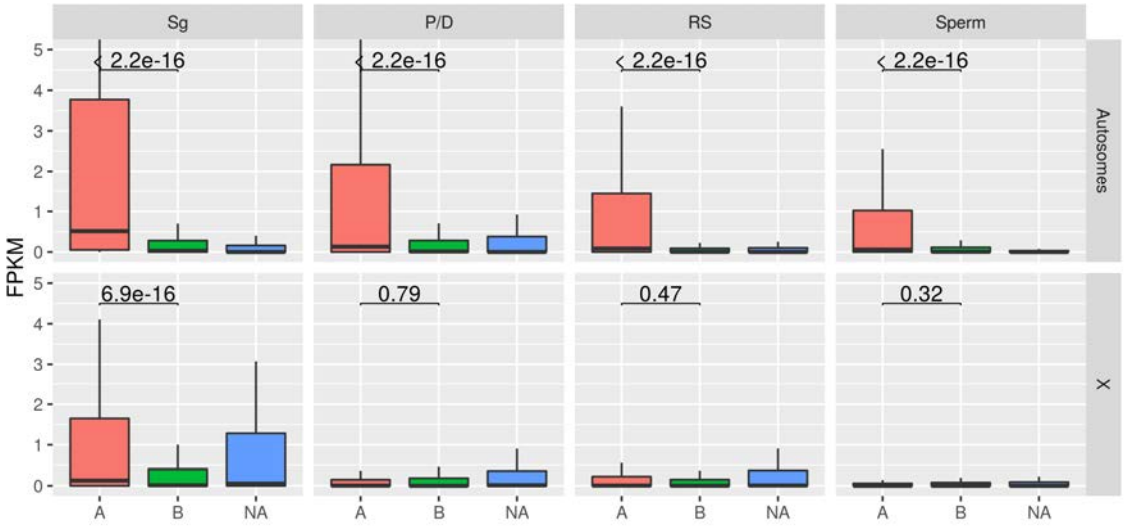


Figure 30. Box plots representing gene expression in autosomal chromosomes and chromosome X according to A/B compartment assignment. Differences in gene expression were assessed between compartments A and B with the Wilcoxon test. A compartments in autosomal chromosomes and the X chromosome in spermatogonia showed statistically significant differences (p-value < 6.9e-16). Legend: Sg: spermatogonia; P/D: pachynema/diplonema; RS: round spermatids; N/A: not assigned.

In searching for transcriptional signatures of compartment switching during spermatogenesis, we first identified cell-specific A-compartments (regions labelled as compartment A in a cell type and labelled as compartment B in the remaining cell types). The cell type with more extension of A-specific regions was pachynema/diplonema with 306 Mbp (8.7% of the genome). This was followed by round spermatids with 108.2 Mbp (3.1% of the genome) (table 11).

Table 11. Description of the A-specific regions in the four cell types analysed by RNA-seq. It includes genomic extension and the number of expressed genes. Legend: Sg: spermatogonia; P/D: pachynema/diplonema; RS: round spermatids; germline: includes A compartment regions in Sg, P/D, RS and sperm but classified as B compartment in fibroblast.

Cell type	Genomic extension (Mbp)	Number of expressing genes	Ratio of expressing protein-coding genes
Germline	21.75	160	0.75
Sg	9	43	0.74
Sg + P/D	5.45	14	0.92
P/D	306	330	0.55
P/D + RS	57.2	147	0.59
RS	108.2	196	0.59
RS + Sperm	37.95	200	0.52
Sperm	57.15	154	0.55

Subsequently, the number of expressing genes (CPM > 1) in A-specific regions was assessed (table 11). The number of genes involved in these regions ranges from few tens to few hundreds depending on the cell type. Interestingly, as observed in section 4.3.3.2, the ratio of

protein-coding genes is higher in A-specific regions where spermatogonia is involved (> 0.74) while in the other cell types the ratio is lower (< 0.60).

5.3.4.2 Functional signatures of compartment switching

The expressing genes involved in compartment A-specific regions were considered for GOEA in order to identify functional signatures during spermatogenesis. Several GO terms related to morphogenesis and cell differentiation were enriched with statistical significance in all cell types excluding sperm: "anatomical structure formation involved in morphogenesis, GO:0048646" and "anatomical structure morphogenesis, GO:0009653" (figure 31; supplementary table 5). In these GO terms, we found genes such as *Mpp5*. In addition, in all cell types, including sperm, the gene *Stag3*, involved in "cellular process involved in reproduction in multicellular organism, GO:0022412", specifically in "male meiosis sister chromatid cohesion, GO:0007065", was identified. The gene *Immp2l* was also found, involved in "response to stress, GO:0006950". Finally, several genes belonging to the enriched GO category, "catabolic process, GO:0009056", specifically "aggrephagy, GO:0035973" were described.

Specifically, in spermatogonia, we found the genes *Gm1993* and *Gm5169*, which are involved in the "synaptonemal complex, (GO:0000795)" under the GO term "cell cycle, GO:0007049". The gene *Klhl13* is also being involved in the mitotic cell division of the cell. On the other hand, the gene *Pot1A* was found under the GO term "regulation of molecular function, GO:0065009". *Pot1A* is part of the telomere shelterin complex. The gene *Diaph2* was also identified among the A-specific expressed genes in spermatogonia.

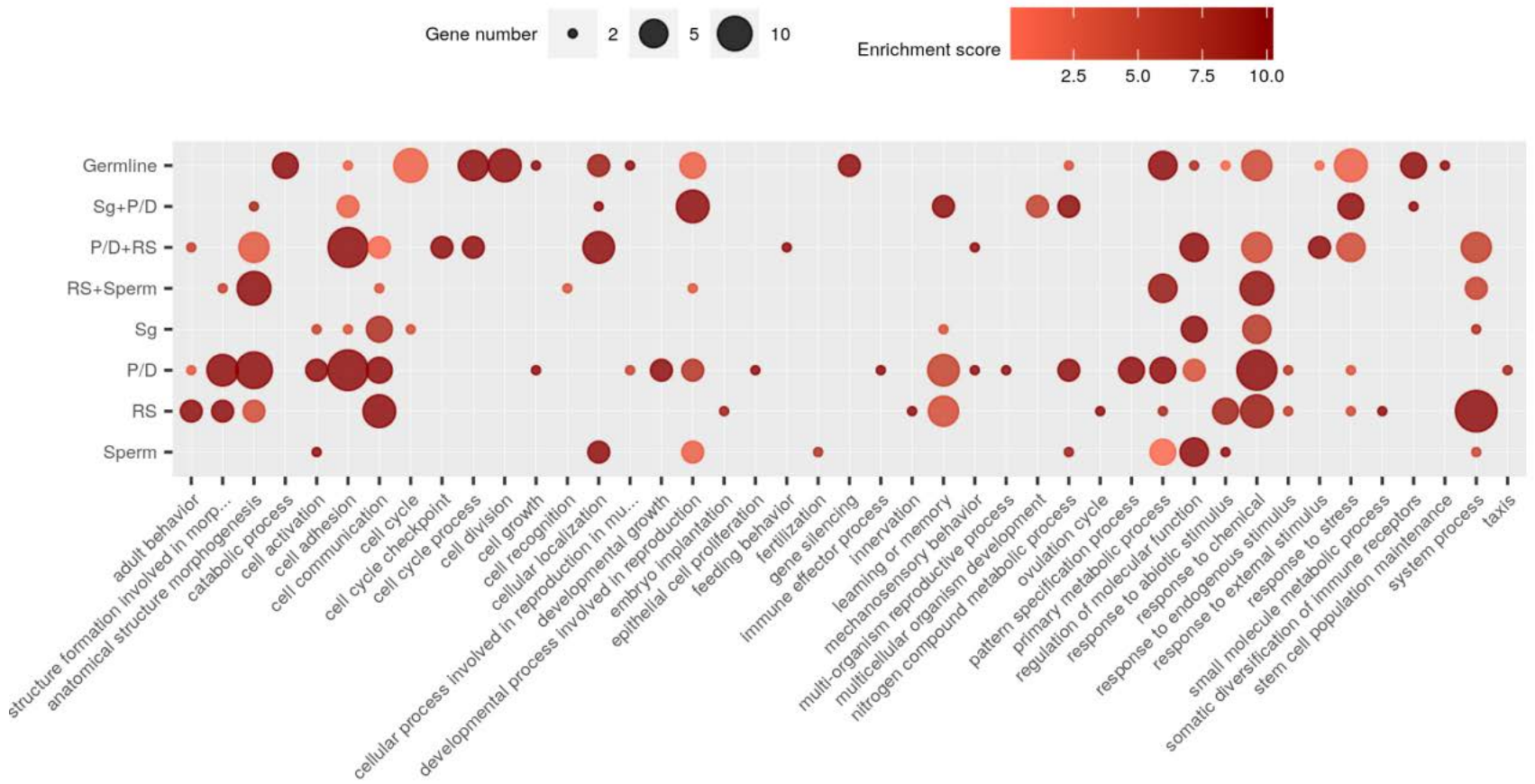


Figure 31. Bubble plot of the significant enriched GO terms from the GOEA analysis. GO terms are summarized up to the level 3 (GO terms with 2 parents). The size of the bubble is related with the number of genes having the corresponding GO term. The colour of the bubble depends on the enrichment score (enrichment scores higher than 10 are set to 10). For a proper visualization, GO terms with 1 gene were hidden from the figure. Detailed information is available at the supplementary table 5.

In primary spermatocytes (pachynema/diplonema) we found and enrichment of the GO term "cell communication, GO:0007154" (i.e. *Fgf10)*. Likewise, the GO term "cell adhesion, GO:0007155", specifically its progenitor "cell-matrix adhesion involved in ameboidal cell migration, GO:0003366", was very enriched in both pachynema/diplonema and round spermatids. The other two most representative GO terms were related with the cell cycle: "cell cycle process, GO:0022402" and "cell cycle checkpoint, GO:0000075". In this sense, genes such as *Mad2l1* and *Ppp2r1a* were spot out under these categories. Another relevant category was "response to chemical, GO:0042221", in which the gene *Abcg2* was identified.

The most representative GO term in round spermatids was "system process, GO:0003008". Inside this category, twenty-seven genes involved in "sensory perception of smell, GO:0007608" were identified. Specifically, 12 out of 24 genes are olfactory receptors localized in 3 different genomic clusters (chromosome 7:86.3-86.4Mb; chromosome 7:102.6-106.8Mb; chromosome 11:49.2-42.3Mb). *Chrna7* was also identified under the GO term "response to odorant, GO:1990834". Besides of it, other genes related with "response to chemical, GO:0042221", such as *Gabrb1* and *Cyp2r1*, were found in the A-specific regions of round spermatids.

In both round spermatids and sperm, the dopamine receptor *Drd2* was found under the GO term "response to chemical, GO:0042221". Also, in both round spermatids and sperm, several genes have been identified under the GO term "cellular response to caffeine (GO:0071313)". In sperm, genes *Plcz1* and *Smcp* were found with implications in "fertilization, GO:0009566". Finally, the acrosomal hyaluronoglucosaminidases *HYAL4* and *HYAL6* were identified under the GO term "primary metabolic process, GO:0044238".

5.4 Discussion

Spermatogenesis involves a continuous process of cell division and differentiation, ranging from undifferentiated diploid cells (PGS) to specialized haploid cells (spermatozoa) (Reig-Viader, *et al*. 2016). In this sense, our data reveals that the 3D genome organization of germ cells is highly dynamic and correlates with gene expression.

5.4.1 Dynamics of the higher-order chromatin organization during gametogenesis

While A/B compartments and TADs are present in spermatogonia, this higher order chromatin organisation is mainly lost during early prophase I in leptonema/zygonema. At this stage, meiotic chromosomes are organized into large DNA loops attached to a protein scaffold composed of specific meiotic cohesins (e.g. REC8 and RAD21L, Gutiérrez-Caballero, *et al.*, 2011; Llano, *et al*. 2012) and proteins of the synaptonemal complex (e.g. SYCP3) (Henderson and

Keeney, 2005). Thus, this particular chromosomal organization necessarily affects the way chromosomes are organized. In late prophase I, in pachynema/diplonema, the weak TAD signal and the low number of TADs suggest that TADs are still absent at this stage. However, compartments start to appear again, though they are not as clear as in spermatogonia. Precisely, it is at the diplotene stage where the synaptonemal complex is unassembled, thus lowering the rigidity of chromosomes when they are aligned side-by-side with their homologous. In fact, the analysis of contact probability as a function of genomic distance P(s) reveals pachynema/diplonema maintains more interactions between long-distance separated *loci* than leptonema/zygonema (figure 23A).

The P(s) analysis also suggests differences in meiotic chromosome folding when compared to what has been reported in metaphase chromosomes (Naumova, *et al*. 2013; Gibcus, *et al*. 2018). That is, prophase I cells display two changes in P(s), the first one between 2.5 and 4.5 Mbp and the second at 40 Mbp. Although the first slope fall-off observed in the meiotic chromosome between 2.5 - 4.5 Mbp resembles what has been described in prometaphase cells (Gibcus, *et al*. 2018), the second decrease in contact probability detected at longer distances (40 Mbp) was not present in the mitotic chromosome. This suggests that the chromatin is organised differently in mitotic and meiotic chromosomes. This particular chromosome organisation can be the result of the particular assembly of chromosomes during prophase I: (i) telomeres contact with the nuclear envelope (bouquet) (Scherthan, *et al*. 1996; Reig-Viader, *et al*. 2016) in leptonema/zygonema thus creating a loop-like structure that make distant regions hardly to interact, or (ii) chromatin is anchored as long DNA loops in the synaptonemal complex, thus preventing interactions below 40 Mbp to occur.

A striking pattern was observed during spermatogenesis when considering TAD signal. Its variance was reduced through the spermatogenesis process when comparing to fibroblasts and spermatogonia, being the TAD signal variance extremely low in sperm (figure 26). A complete absence of the TAD signal has been described in mitotic chromosomes (Naumova, *et al*. 2013; Gibcus, *et al*. 2018) and, in this regard, meiotic chromosomes (i.e., primary spermatocytes) mirror this pattern. However, in the case of sperm, we detected an extremely low TAD signal, contrary to what has been previously reported (Jung, *et al*. 2017; Ke, *et al*. 2017; Wang, *et al*. 2019). While we used FACS to obtain highly enriched populations of sperm separated from the rest of the germ cell populations, Jung, *et al*., Ke, *et al*. and Wang *et al*. obtained the swimming sperm from the supernatant after incubating dissected cauda epididymis. Our Hi-C simulations mixing sperm and fibroblast reads suggest that the higher order chromatin structures previously reported might be due to the presence of somatic contamination (figure 27). In this context, the

absence of TAD signal observed in sperm might be related to the histone replacement process that takes place during spermiogenesis. According to the literature, the vast majority of histones are replaced by protamines in sperm, folding DNA into toroidal subunits of approximately 50 kbp (Balhorn, *et al*. 1984; Hud, *et al*. 1993; Johnson, *et al*. 2011). Therefore, the highly compacted chromatin that characterises sperm is associated with the presence of A/B compartments at the Mbp scale but not with the formation of TAD structures at the finer scale (i.e. kbp).

The highly compacted sperm chromatin might also be related with the inter-chromosome/intra-chromosome interaction ratio. Most sperm chromosomes have more inter-chromosome interactions than intra-chromosome interactions, suggestive of highly compacted chromosome structure into a very compressive environment, thus favouring inter-chromosome interactions than intra-chromosome interactions. Additionally, in sperm, but also in fibroblast, spermatogonia and round spermatids, the inter-chromosome/intra-chromosome interaction upward rate tendency suggests the so-called chromosome territories. Gene-rich chromosomes, which in fact are the shortest chromosomes, tend to be located at the centre of the nucleus (Boyle, *et al*. 2001). In this sense, as observed in our data, the shortest chromosomes might interact more with the other chromosomes (higher inter-chromosome/intra-chromosome interaction ratio) than the longest ones.

Also related with the inter-/intra-chromosome interaction ratio, the X chromosome shows a distinct pattern in regards the autosomal chromosomes of pachynema/diplonema, round spermatids and sperm. This is consistent with the Meiotic Sex Chromosome Inactivation (MSCI) previously described during meiosis I, a silencing process that is maintained in post-meiotic cells (Namekawa, *et al*. 2006; Turner, 2007; Yan and McCarrey, 2009). The fact that the sex chromosomes are isolated from the autosomal ones (Yan and McCarrey, 2009) explains the reduction of the inter-/intra-chromosome interaction ratio in these cell types. Gene expression data also confirmed this pattern, as we identified a sharper reduction of the expressing genes in the X chromosome than in autosomal chromosomes between spermatogonia and meiotic/post-meiotic cells (see section 4.3.3.2).

Additional chromosomal features include the clustering of telomeres in the nuclear envelope, the so-called *bouquet* (Scherthan, *et al.* 1996; Reig-Viader, *et al.* 2016), which was identified in leptonema/zygonema by Hi-C maps. This cell type shows significantly higher interaction among subcentromeric regions than fibroblast (figure 23D). Round spermatids also show higher interaction among subcentromeric regions than fibroblast (figure 23D). At this stage of the

spermatogenesis, especially during spermiogenesis, chromosomes adopt a looped conformation because of centromere association (Haaf and Ward, 1995; Zalensky, *et al*. 1995; Meyer-Ficca, *et al*. 1998) that explains this higher interaction.

5.4.1 Functional insights of the higher-order chromatin organization during gametogenesis

The study of transcriptional signatures of compartment switching revealed different expressing genes in cell-specific A compartments that are related with spermatogenesis. This is the case of morphogenesis-related genes implicated in cell differentiation process along spermatogenesis. In this sense, the gene family *Mpp* (Membrane palmitoylated protein) (e.g. *Mpp5*) were found in centrosomes or in the mitotic spindle, thus involved in microtubule-related functions such as cytoskeleton rearrangements (Matsumoto-Taniura, *et al*. 1996). Continuous contact with different stimulus, being either chemical or hormonal, influence cell differentiation progress or sperm motility. For example, *Drd2* is a dopamine receptor the ligand of which, dopamine, has been found in sperm affecting its motility (Urra, *et al*. 2014). In addition, Saucedo, *et al*. 2018 reported fibroblast growth factors (*Fgfs*) such as *Fgf10* to influence sperm motility, having their receptors mainly in the acrosome of round spermatids and sperm.

Spermatogenesis is a very dynamic process that involves a series of cell divisions. In this sense, expressing genes related to cell cycle processes were identified in cell-specific A compartments. For instance, *Klhl13* is involved in the mitotic cell division of cells according to the Mouse Genome Database (MGD) (Smith, *et al*. 2018). It is in primary spermatocytes, specifically at the stage of leptotene, where the synaptonemal complex starts being assembled. Also, according to MGD (Smith, *et al*. 2018), the expressing genes *Gm1993* and *Gm5169*, found in spermatogonia-specific A compartments, are related with the synaptonemal complex. At further stages of primary spermatocytes, chromosome synapsis and meiotic recombination is carried out. The gene *Stag3* has been described as essential for DNA repair and synapsis between homologous chromosomes (Llano, *et al*. 2014). Also, while *Mad2l1* is related with the spindle checkpoint, *Deup1* has been described as essential for the centriole formation (Zhao, *et al*. 2013). As centrioles are needed for development of flagella in living organisms, *Deup1* plays a key role in spermiogenesis.

In order to fertilize the oocyte, sperm will use chemotaxis for guidance. Several olfactory receptors, which have been previously described as important for the sperm guidance towards the oocyte (Flegel, *et al*. 2015), were identified in cell-specific A compartments and being expressed. Other membrane receptors were found: *Chrna7* has been described as important for

normal sperm motility (Bray, *et al.* 2005), while *Gabrb1* might have implications in the fertility of male (Jodar, *et al*. 2012). The gene *Cyp2r1* is involved in the vitamin D metabolism, which might also be involved in sperm metabolism and motility (Rehman, *et al*. 2018). *Plcz1* localizes in the acrosome with involved in fertilization, while *Smcp,* according to its Entrez Gene (Maglott, *et al*. 2005) entry, is thought to stabilize the mitochondrial sheath. At the end, fertilization is reached when the membrane of sperm fuses with the oocyte through the acrosome reaction, which releases hyaluronidases such as *Hyal4* and *Hyal6*. On the other hand, the capacity of sperm for membrane fusion is correlated with the cholesterol concentration of the surrounding environment (Cross, *et al*. 1998). The gene *Abcg2* has been identified as a mediator of cholesterol removal (Scharenberg, *et al*. 2009). Finally, additional genes have been associated with infertility. Mutations in *Immp2l* produce infertile mouse females or subfertile males (Lu, *et al*. 2008). *Diaph2* was found to be associated with ovarian failure, but it has been also associated with sperm morphology in bulls (Bione, *et al*. 1998; Fortes, *et al*. 2013). Also, *Ppp2r1a* together with *Stag3* have been described as essential for fertility in mice (Hu, *et al*. 2014).

Overall, our results reveal previously undescribed stages of compartmentalisation for meiotic chromosomes in both early (prophase I) and late stages of spermatogenesis (round spermatids and sperm) and provide evidence on the existence of a fine tuning between chromatin remodelling and gene expression in germ cells.

5.5 References

Alavattam, K. G., et al. “Attenuated Chromatin Compartmentalization in Meiosis and Its Maturation in Sperm Development.” *Nature Structural & Molecular Biology*, 2019, p. 1, doi:10.1038/s41594-019-0189-y.

Araki, Y., et al. “Genome-Wide Analysis of Histone Methylation Reveals Chromatin State-Based Regulation of Gene Transcription and Function of Memory CD8+ T Cells.” *Immunity*, vol. 30, no. 6, 2009, pp. 912–25, doi:10.1016/j.immuni.2009.05.006.

Ay, F., Noble, W. S. “Analysis Methods for Studying the 3D Architecture of the Genome.” *Genome Biology*, vol. 16, no. 1, 2015, p. 183, doi:10.1186/s13059-015-0745-7.

Balhorn, R., et al. “DNA Packaging in Mouse Spermatids. Synthesis of Protamine Variants and Four Transition Proteins.” *Experimental Cell Research*, vol. 150, no. 2, 1984, pp. 298–308, http://www.ncbi.nlm.nih.gov/pubmed/6692853.

Barski, A., et al. “High-Resolution Profiling of Histone Methylations in the Human Genome.” *Cell*, vol. 129, no. 4, 2007, pp. 823–37, doi:10.1016/j.cell.2007.05.009.

Bione, S., et al. “A Human Homologue of the Drosophila Melanogaster Diaphanous Gene Is Disrupted in a Patient with Premature Ovarian Failure: Evidence for Conserved Function in Oogenesis and Implications for Human Sterility.” *The American Journal of Human Genetics*, vol. 62, no. 3, 1998, pp. 533–41, doi:10.1086/301761.

Boateng, K. A., et al. “Homologous Pairing Preceding SPO11-Mediated Double-Strand Breaks in Mice.” *Developmental Cell*, vol. 24, no. 2, 2013, pp. 196–205, doi:10.1016/J.DEVCEL.2012.12.002.

Boyle, S., et al. “The Spatial Organization of Human Chromosomes within the Nuclei of Normal and Emerin-Mutant Cells.” *Human Molecular Genetics*, vol. 10, no. 3, 2001, pp. 211–19, doi:10.1093/hmg/10.3.211.

Bray, C., et al. “Mice Deficient in CHRNA7, a Subunit of the Nicotinic Acetylcholine Receptor, Produce Sperm with Impaired Motility1.” *Biology of Reproduction*, vol. 73, no. 4, 2005, pp. 807–14, doi:10.1095/biolreprod.105.042184.

Bushnell, B. “*BBMap: A Fast, Accurate, Splice-Aware Aligner*”. 2014, https://www.osti.gov/biblio/1241166-bbmap-fast-accurate-splice-aware-aligner.

Cross, N. L. “Role of Cholesterol in Sperm Capacitation1.” *Biology of Reproduction*, vol. 59, no. 1, 1998, pp. 7–11, doi:10.1095/biolreprod59.1.7.

Dixon, J. R., et al. “Topological Domains in Mammalian Genomes Identified by Analysis of Chromatin Interactions.” *Nature*, vol. 485, no. 7398, 2012, pp. 376–80, doi:10.1038/nature11082.

Flegel, C., et al. “Characterization of the Olfactory Receptors Expressed in Human Spermatozoa.” *Frontiers in Molecular Biosciences*, vol. 2, 2015, p. 73, doi:10.3389/fmolb.2015.00073.

Fortes, M. R. S., et al. “Genome-Wide Association Study for Inhibin, Luteinizing Hormone, Insulin-like Growth Factor 1, Testicular Size and Semen Traits in Bovine Species.” *Andrology*, vol. 1, no. 4, 2013, pp. 644–50, doi:10.1111/j.2047-2927.2013.00101.x.

Gibcus, J. H., et al. “A Pathway for Mitotic Chromosome Formation.” *Science*, vol. 359, no. 6376, 2018, p. eaao6135, doi:10.1126/science.aao6135.

Griswold, M. D. “Spermatogenesis: The Commitment to Meiosis.” *Physiological Reviews*, vol. 96, no. 1, 2016, pp. 1–17, doi:10.1152/physrev.00013.2015.

Gutiérrez-Caballero, C., et al. “Identification and Molecular Characterization of the Mammalian α-Kleisin RAD21L.” *Cell Cycle*, vol. 10, no. 9, 2011, pp. 1477–87, doi:10.4161/cc.10.9.15515.

Haaf, T, Ward, D. C., “Higher Order Nuclear Structure in Mammalian Sperm Revealed by in Situ Hybridization and Extended Chromatin Fibers.” *Experimental Cell Research*, vol. 219, no. 2, 1995, pp. 604–11, doi:10.1006/excr.1995.1270.

Handel, M. A., Schimenti, J. C. “Genetics of Mammalian Meiosis: Regulation, Dynamics and Impact on Fertility.” *Nature Reviews Genetics*, vol. 11, no. 2, 2010, pp. 124–36, doi:10.1038/nrg2723.

Henderson, K. A., Keeney, S. “Synaptonemal Complex Formation: Where Does It Start?” *BioEssays*, vol. 27, no. 10, 2005, pp. 995–98, doi:10.1002/bies.20310.

Hu, M. W., et al. “Scaffold Subunit Aalpha of PP2A Is Essential for Female Meiosis and Fertility in Mice1.” *Biology of Reproduction*, vol. 91, no. 1, 2014, p. 19, doi:10.1095/biolreprod.114.120220.

Hud, N. V., et al. “Identification of the Elemental Packing Unit of DNA in Mammalian Sperm Cells by Atomic Force Microscopy.” *Biochemical and Biophysical Research Communications*, vol. 193, no. 3, 1993, pp. 1347–54, doi:10.1006/bbrc.1993.1773.

Imakaev, M., et al. “Iterative Correction of Hi-C Data Reveals Hallmarks of Chromosome Organization.” *Nature Methods*, vol. 9, no. 10, 2012, pp. 999–1003, doi:10.1038/nmeth.2148.

Jodar, M., et al. “Differential RNAs in the Sperm Cells of Asthenozoospermic Patients.” *Human Reproduction*, vol. 27, no. 5, 2012, pp. 1431–38, doi:10.1093/humrep/des021.

Jodar, M., et al. “Absence of Sperm Rna Elements Correlates With Idiopathic Male Infertility.” *Science translational medicine*, vol. 7, no. 295, 2016, p. 295re6, doi: 10.1126/scitranslmed.aab1287.

Johnson, G. D., et al. “The Sperm Nucleus: Chromatin, RNA, and the Nuclear Matrix.” *Reproduction*, vol. 141, no. 1, 2011, pp. 21–36, doi:10.1530/REP-10-0322.

Ke, Y., et al. “3D Chromatin Structures of Mature Gametes and Structural Reprogramming during Mammalian Embryogenesis.” *Cell*, vol. 170, no. 2, 2017, p. 367–381.e20, doi: 10.1016/j.cell.2017.06.029.

Keeney, S., et al. “Meiosis-Specific DNA Double-Strand Breaks Are Catalyzed by Spo11, a Member of a Widely Conserved Protein Family.” *Cell*, vol. 88, no. 3, 1997, pp. 375–84, http://www.ncbi.nlm.nih.gov/pubmed/9039264.

Lajoie, B. R., et al. “The Hitchhiker’s Guide to Hi-C Analysis: Practical Guidelines.” *Methods*, vol. 72, 2015, pp. 65–75, doi:10.1016/j.ymeth.2014.10.031.

Lieberman-Aiden, E., et al. “Comprehensive Mapping of Long-Range Interactions Reveals Folding Principles of the Human Genome.” *Science*, vol. 326, no. 5950, 2009, pp. 289–93, doi:10.1126/science.1181369.

Llano, E., et al. “Meiotic Cohesin Complexes Are Essential for the Formation of the Axial Element in Mice.” *The Journal of Cell Biology*, vol. 197, no. 7, 2012, pp. 877–85, doi:10.1083/jcb.201201100.

Lu, B., et al. “A Mutation in the Inner Mitochondrial Membrane Peptidase 2-Like Gene (Immp2l) Affects Mitochondrial Function and Impairs Fertility in Mice1.” *Biology of Reproduction*, vol. 78, no. 4, 2008, pp. 601–10, doi:10.1095/biolreprod.107.065987.

Maglott, D., et al. “Entrez Gene: Gene-Centered Information at NCBI.” *Nucleic Acids Research*, vol. 33, no. Database issue, 2004, pp. D54–58, doi:10.1093/nar/gki031.

Marco-Sola, S., et al. “The GEM Mapper: Fast, Accurate and Versatile Alignment by Filtration.” *Nature Methods*, vol. 9, no. 12, 2012, pp. 1185–88, doi:10.1038/nmeth.2221.

Matsumoto-Taniura, N., et al. “Identification of Novel M Phase Phosphoproteins by Expression Cloning.” *Molecular Biology of the Cell*, vol. 7, no. 9, 1996, pp. 1455–69, doi:10.1091/mbc.7.9.1455.

Mirny, L. A. “The Fractal Globule as a Model of Chromatin Architecture in the Cell.” *Chromosome Research*, vol. 19, no. 1, 2011, pp. 37–51, doi:10.1007/s10577-010-9177-0.

Muller, H., et al. “Characterizing Meiotic Chromosomes’ Structure and Pairing Using a Designer Sequence Optimized for Hi-C.” *Molecular Systems Biology*, vol. 14, no. 7, 2018, p. e8293, doi:10.15252/msb.20188293.

Namekawa, S. H., et al. “Postmeiotic Sex Chromatin in the Male Germline of Mice.” *Current Biology*, vol. 16, no. 7, 2006, pp. 660–67, doi:10.1016/j.cub.2006.01.066.

Naumova, N., et al. “Organization of the Mitotic Chromosome.” *Science*, vol. 342, no. 6161, 2013, pp. 948–53, doi:10.1126/science.1236083.

Patel, L, et al. “Dynamic Reorganization of the Genome Shapes the Recombination Landscape in Meiotic Prophase.” *Nature Structural & Molecular Biology*, 2019, p. 1, doi:10.1038/s41594-019-0187-0.

Ramírez, F., et al. “High-Resolution TADs Reveal DNA Sequences Underlying Genome Organization in Flies.” *Nature Communications*, vol. 9, no. 1, 2018, p. 189, doi:10.1038/s41467-017-02525-w.

Rao, S. S. P., et al. “A 3D Map of the Human Genome at Kilobase Resolution Reveals Principles of Chromatin Looping.” *Cell*, vol. 159, no. 7, 2014, pp. 1665–80, doi:10.1016/j.cell.2014.11.021.

Rehman, R., et al. “Association Between Vitamin D, Reproductive Hormones and Sperm Parameters in Infertile Male Subjects.” *Frontiers in Endocrinology*, vol. 9, 2018, p. 607, doi:10.3389/fendo.2018.00607.

Reig-Viader, R., et al. “Telomere Homeostasis in Mammalian Germ Cells: A Review.” *Chromosoma*, vol. 125, no. 2, 2016, pp. 337–51, doi:10.1007/s00412-015-0555-4.

Ren, X., et al. “Is transcription in sperm stationary or dynamic?” *Journal of Reproduction and Development*, vol. 63, no. 5, 2017, pp: 439-443, doi: 10.1262/jrd.2016-093.

Saucedo, L., et al. “Deficiency of Fibroblast Growth Factor 2 (FGF-2) Leads to Abnormal Spermatogenesis and Altered Sperm Physiology.” *Journal of Cellular Physiology*, vol. 233, no. 12, 2018, pp. 9640–51, doi:10.1002/jcp.26876.

Saucedo, L., et al. “Involvement of Fibroblast Growth Factor 2 (FGF2) and Its Receptors in the Regulation of Mouse Sperm Physiology.” *Reproduction*, vol. 156, no. 2, 2018, pp. 163–72, doi:10.1530/REP-18-0133.

Scharenberg, C., et al. “ABCG2 Is Expressed in Late Spermatogenesis and Is Associated with the Acrosome.” *Biochemical and Biophysical Research Communications*, vol. 378, no. 2, 2009, pp. 302–07, doi:10.1016/j.bbrc.2008.11.058.

Scherthan, H., et al. “Centromere and Telomere Movements during Early Meiotic Prophase of Mouse and Man Are Associated with the Onset of Chromosome Pairing.” *The Journal of Cell Biology*, vol. 134, no. 5, 1996, pp. 1109–25, doi:10.1083/JCB.134.5.1109.

Serra, F., et al. “Automatic analysis and 3D-modelling of Hi-C data using TADbit reveals structural features of the fly chromatin colors”. *PLOS Computational Biology*, vol. 13, no. 7, 2017, pp: e1005665.

Sexton, T., et al. “Three-Dimensional Folding and Functional Organization Principles of the Drosophila Genome.” *Cell*, vol. 148, no. 3, 2012, pp. 458–72, doi:10.1016/j.cell.2012.01.010.

Strouboulis, J., et al. “Functional Compartmentalization of the Nucleus.” *Journal of Cell Science*, vol. 109 (Pt 8), no. 8, 1996, pp. 1991–2000, http://www.ncbi.nlm.nih.gov/pubmed/8856494.

Turner, J. M. A. “Meiotic Sex Chromosome Inactivation.” *Development*, vol. 134, no. 10, 2007, pp. 1823–31, doi:10.1242/dev.000018.

Urra, J. A., et al. “Presence and Function of Dopamine Transporter (DAT) in Stallion Sperm: Dopamine Modulates Sperm Motility and Acrosomal Integrity.” *PLoS ONE*, vol. 9, no. 11, 2014, p. e112834, doi:10.1371/journal.pone.0112834.

Wang, Y., et al. “Reprogramming of Meiotic Chromatin Architecture during Spermatogenesis.” *Molecular Cell*, vol. 73, no. 3, 2019, p. 547–561.e6, doi:10.1016/J.MOLCEL.2018.11.019.

Yan, W., McCarrey, J. R. “Sex Chromosome Inactivation in the Male.” *Epigenetics*, vol. 4, no. 7, 2009, pp. 452–56, http://www.ncbi.nlm.nih.gov/pubmed/19838052.

Yang, T., et al. “HiCRep: Assessing the Reproducibility of Hi-C Data Using a Stratum-Adjusted Correlation Coefficient.” *Genome Research*, vol. 27, no. 11, 2017, pp. 1939–49, doi:10.1101/gr.220640.117.

Zalensky, A. O., et al. “Well-Defined Genome Architecture in the Human Sperm Nucleus.” *Chromosoma*, vol. 103, no. 9, 1995, pp. 577–90, doi:10.1007/BF00357684.

Zhao, H., et al. “The Cep63 Paralogue Deup1 Enables Massive de Novo Centriole Biogenesis for Vertebrate Multiciliogenesis.” *Nature Cell Biology*, vol. 15, no. 12, 2013, pp. 1434–44, doi:10.1038/ncb2880.

Zickler, D., Kleckner, N. “Meiotic Chromosomes: Integrating Structure and Function.” *Annual Review of Genetics*, vol. 33, no. 1, 1999, pp. 603–754, doi:10.1146/annurev.genet.33.1.603.

Chapter 6: General discussion

Several technological advances have led in the last decades to a better understanding of how genomes are organized and regulated. Since the first sequencing method was released in 1977, new technologies such as next-generation and third-generation sequencing technologies boosted the development of different -omics fields (Margulies, *et al*. 2005; Bentley, *et al*. 2008; Eid, *et al*. 2009; Rothberg *et al.* 2011). This permitted, for instance, the study of genetic variants (e.g. the 1000 Genomes Project Consortium, 2015), gene expression and regulation (e.g. the ENCODE Project Consortium, 2012), or the most recent exploration of the tri-dimensional (3D) organization of genomes (e.g. Lieberman-Aiden, *et al*. 2009). The increased performance of new sequencing technologies came in parallel with the development of bioinformatics. As genomes were released together with its transcriptional and epigenetics profiles, different databases emerged to make -omics information accessible to the scientific community (Hubbard, *et al*. 2002; Kent, *et al*. 2002). The arrival of new bioinformatics tools soon followed to handle new sequencing reads yielded by high-throughput technologies. Notwithstanding recent advances in the field, the currently amount of -omics data requires further maturation and expansion of bioinformatics tools.

In this context, this thesis takes advantage of different technologies to develop and integrate next-generation bioinformatics tools to increase our understanding of genomes at both the **functional** and the **structural** levels.

6.1 Towards the development of online databases and cloud platforms for the analysis of transcriptomics data

The amount of sequencing data produced worldwide doubles every 7 months, thus growing at a faster rate than the expected by Illumina or by Moore's law (Stephens, *et al*. 2015) (figure 32). Moore's law (formulated by Gordon Moore, cofounder of Intel) states that the number of transistors in a chip (processing capacity) doubles every 18 months (Moore, 1965). Due to the fact that sequencing is growing at a higher rate than processing capacity, the flood of sequencing data that is expected to be generated in the next few years is likely to create a **bottleneck** in bioinformatics.

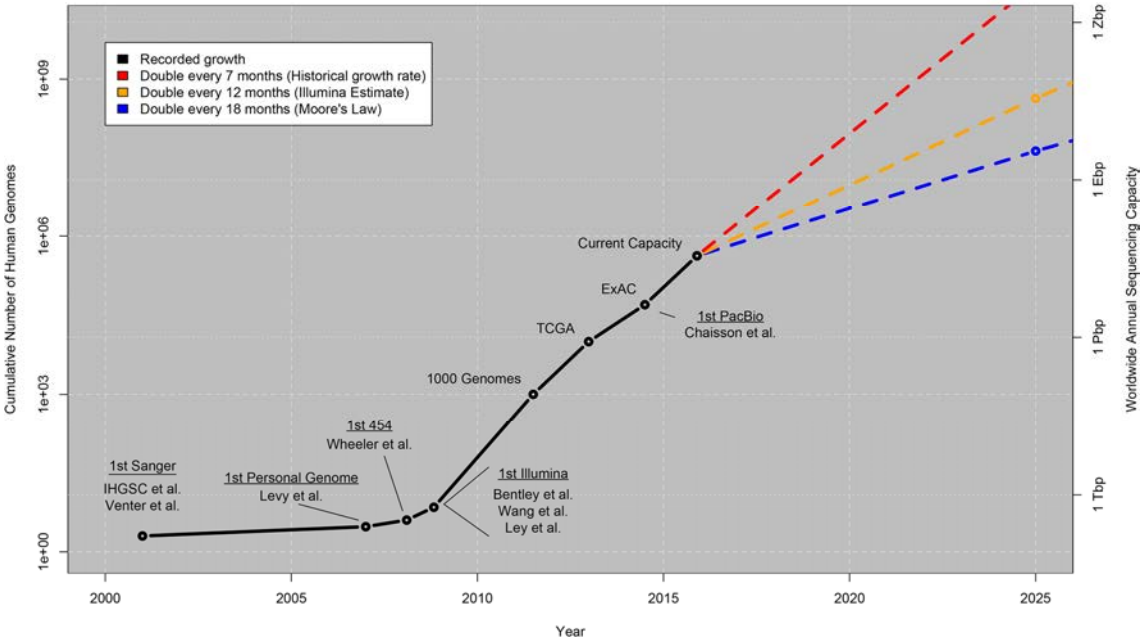


Figure 32. Growth of DNA sequencing. The plot shows the cumulative number of human genomes generated so far from early 2000s and the growth projection towards 2025. Figure extracted from Stephens, *et al*. (2015).

Under this scenario, data accessibility needs to be improved in order to power researchers with useful tools for allowing data mining and to facilitate reaching conclusions efficiently. In parallel, data processing should also be ameliorated, not in terms of hardware, which processing growth is limited by the Moore's law, but in terms of efficiency, automatization and democratization using IT solutions currently available.

With all these factors in mind, we have developed in this thesis two bioinformatics solutions for the transcriptomics field. In particular, the first tool (GreeNC) has been implemented to boost data accessibility of lncRNA annotation in plants and the second one (AIR) to democratize data processing of high-throughput RNA-seq data in any type of organism.

6.1.1 GreeNC: a comprehensive online database of plant lncRNAs

The analysis of sequencing data generates post-processing results, which are far more complex than raw data: it includes a wide variety of data types (e.g. integer and floating-point numbers, strings, Booleans, lists, key – value relationships) and variables (e.g. GC content, gene length, gene expression). In this sense, post-processing results need to be organized and stored in an organized manner to allow their retrieval and to facilitate their accessibility. There are two types of informatics databases to store data: SQL (e.g. MySQL) and NoSQL (e.g. MongoDB). Traditionally, SQL was preferred for database development; however, this paradigm is changing due to NoSQL databases, which are (i) more efficient in managing big amounts of data and (ii)

more flexible as there is no need to pre-define the number of data types and variables to be stored (schema-less approach) (Schulz, *et al*. 2016).

Since transcriptomics is one of the most funded -omics fields at the moment, large amounts of data are generated in a short period of time (Ulrich, 2016). Thus, efforts to enhance data accessibility are much in need. As an example, it has been suggested that most of the human transcriptome is non-coding, increasing the interest in this field (Carninci, *et al*. 2005; Mattick and Makunin, 2006; Mattick, 2009; Derrien, *et al*. 2012; Djebali, *et al*. 2012). In this context, the latest GENCODE annotation (version 29) currently contains more non-coding genes (e.g. lncRNA genes) than coding genes in both human and mouse genomes. However, although lncRNAs are known to be involved in important processes such as chromatin remodelling, transcriptional control and post-transcriptional processing (Mercer, *et al*. 2009; Barbosa Dogini, *et al*. 2014), their research is far more extended in human and mouse than in other organisms (e.g. plants).

Despite being a largely unexplored field, some lncRNAs have been already characterized in plants, most of them regulating important agronomic traits such as flowering, morphological development, or stress response (Franco-Zorrilla, *et al*. 2007; Swiezewski, *et al*. 2009; Heo and Sung, 2011; Ding, *et al*. 2012; Shin and Chekanova, 2014; Gai, *et al*. 2018; Liu, *et al*. 2018). Being plant transcriptomics an emerging field, a comprehensive annotation of lncRNAs is required to ease their functional characterization in plant biology. In this context, we developed GreeNC, a public online database of plant lncRNAs derived from genome-wide studies on 39 plant species and 6 algae. Before the release of GreeNC, the available databases for plant lncRNAs presented several drawbacks such as lack of APIs, low GUI friendliness and few species or lncRNAs available. This motivated the development of GreeNC, which in turn fulfils most of the rules suggested for database creation (Helmy *et al*. 2016) (table 2): (i) high data quality due to high accurate pipelines (rule 2); (ii) easy-to-use and friendly GUI besides of the availability of an API using modern technologies such as NodeJS (rules 3-5); and (iii) simple query options, including the possibility of batch downloads (rules 6 and 7).

Today, GreeNC is the most comprehensive database in terms of the number of species in comparison with currently available plant databases. GreeNC also represented the largest database in terms of number of lncRNAs from early 2016 to the beginning of 2018 until the release of CANTATAdb v2.0 (239,000 lncRNAs), which slightly surpassed GreeNC (203,000 lncRNAs). GreeNC has provided to the scientific community a valuable lncRNA resource for nearly three years since its launch in 2016. During this period, several studies have used GreeNC for lncRNA identification and characterization. For instance, novel lncRNAs were identified in

banana (*Musa acuminata*) (Li, *et al*. 2017) and potato (*Solanum tuberosum*) (Kwenda, *et al*. 2016). In addition, surveys in GreeNC allowed the identification of a candidate transcript involved in apomictic development as a potential lncRNA in *Paspalum notatum* (Ochogavía *et al*. 2017). Machine-learning algorithms were also trained with the lncRNA resource of GreeNC (da Costa Negri, *et al*. 2018). In this way, GreeNC has proven to be a useful repository to boost the research in the field.

From the technological point of view, GreeNC was developed using the relational database MySQL to store the data. This database was chosen since the number of fields to store for each lncRNA (e.g. species, gene, coordinates, coding potential or folding energies) was well defined. However, genomic projects are making available new expression data across different tissues and novel lncRNAs are continuously reported in the literature. In parallel, the availability of expression data would allow the construction of both coding and lncRNA gene co-expression network to infer putative functionalities (Chen, *et al*. 2018). The incorporation of this data might require a shift towards NoSQL databases due to (i) the amount of data stored will scale up and NoSQL databases offer better performance, (ii) variables such as gene expression might come from different sources, thus requiring an evolving data model that is provided by schema-less NoSQL approaches, and (iii) NoSQL databases are more suitable to store graphs (e.g. gene networks) (de Brevern, *et al*. 2015; Schulz, *et al*. 2016). Future updates of GreeNC will include expression data and gene networks to complete its resources for lncRNA research in plants.

6.1.2 AIR: the first end-to-end solution for high-throughput RNA-seq analysis

According to Cisco Global Cloud Index, cloud computing is growing so fast that by 2021 94% of the worldwide workload will be performed in the cloud, being SaaS the most used cloud service. The adoption of SaaS comes with a series of advantages: (i) the initial costs for its adoption are low (pay-per-use model) in comparison with an on-premise solution, (ii) support and training are usually available, (iii) since hardware is not on-premise, neither hardware maintenance nor dedicating a physical space to allocate servers are required, (iv) system upgrades are carried out by the provider, (v) deployment is immediate due to its cloud-based nature, and (vi) it is adaptive to tour workload needs (scalability) (Rhyman, 2017).

The accelerated growth of sequencing is produced not only by ultra-high-throughput sequencing equipment acquired in specialized sequencing centres (e.g. Illumina NovaSeq), but also by benchtop sequencers that permitted the democratization of sequencing as they target smaller research institutions or even individual laboratories (e.g. Illumina MiSeq or Ion Torrent S5) (table 1). It is at these smaller institutions where a bioinformatics facility might represent a

constraint, thus lacking expertise to perform genome-wide studies. In these cases, the use of SaaS focused on bioinformatics applications can be a solution. Following the democratization of sequencing, next-generation bioinformatics (SaaS platforms applied on bioinformatics) democratize bioinformatics, thus solving the need to acquire (i) powerful computers for data analysis and (ii) bioinformatics skills due to easy-to-use interfaces (de Brevern, *et al*. 2015).

Several next-generation bioinformatics solutions have recently emerged in the field of RNA-seq data analysis (Illumina, 2014; Malhotra, *et al*. 2017). However, they still require basic skills in bioinformatics at the very beginning of the process (defining specific software parameters) and at the very end (manipulating and transforming results for data interpretation and integration). In addition, non-model species are not available by default. In this way, current next-generation bioinformatics solutions for the analysis of RNA-seq data do not reach full democratization. In order to overcome these limitations, we developed AIR, a SaaS platform.

One of the advantages of AIR is that users are not required to have previous informatics or bioinformatics knowledge as they only need to upload their samples and to start an analysis with few clicks. In addition, it is not only limited to model species: all genomes available in Ensembl, NCBI and JGI can be used (more than 150,000 genomes available). All stages of the RNA-seq data analysis are handled by AIR, from the quality check and trimming to the statistical analysis with the creation of tables and plots and no further involvement of the user. In this sense, it represents the first end-to-end solution in the field. AIR, which has today more than 750 users and 1,000 RNA-seq samples analysed, has been already cited in two different publications that studied obesity and transcriptional changes after activating the transcription factor PPARγ in human (Gerlini, *et al*. 2018; Kim, *et al*. 2019). For instance, the use of AIR to study obesity in human ease the identification of potential targets for managing metabolic health (Gerlini, *et al*. 2018).

Technologies behind AIR, such as Docker, should be spread out in bioinformatics beyond cloud-based systems. As highlighted by Di Tommaso, *et al*. (2015 and 2017), reproducible results with the same data across computers and operating systems is only achievable with Dockers. In this sense, centralization of bioinformatics analyses in SaaS solutions would enhance reproducibility in the field. This goes in parallel with the application of the rules for reproducible computational research suggested by Sandve, *et al*. 2013 (table 2). Centralization of bioinformatics analyses in SaaS solutions would likely fulfil most of the rules suggested. In the case of AIR, for instance, it does not fulfil rules 8 (it is not applicable), 9 (it should be done by the final user) and 10 (results can be shared through a public URL, but scripts are not provided due to industrial interest).

Altogether, pipeline automatization for RNA-seq data included in a smart cloud-based system allowed the development of AIR, which is ready for reproducible bioinformatics research in transcriptomics bringing at the same time the possibility to perform RNA-seq data analyses to institutions without bioinformatics facilities or researchers with poor expertise in bioinformatics analyses.

6.2 Principles of chromosome assembly during spermatogenesis

In this work, we also elucidated the 3D organization of the mouse genome during spermatogenesis. Specifically, we studied the dynamics of the higher-order chromatin organization in: (i) pre-meiotic (spermatogonia), (ii) meiotic (primary spermatocytes at pachytene/diplotene and leptotene/zygotene stages and secondary spermatocytes) and (iii) post-meiotic cells (round spermatids and sperm). Our study permitted to unveil principles of chromosome assembly during the formation of germ cells, which was reflected at different levels of resolution: (i) intra-/inter-chromosomal interaction ratios, (ii) distance-dependent interaction frequencies, (iii) genomic compartments and (iv) topological domains. Moreover, we showed evidence of a delicate fine-tuning between chromatin remodelling and cell-specific gene expression.

6.2.1 Commitment to enter meiosis is accompanied by changes in chromosome occupancy

Mammalian spermatogenesis begins with the differentiation of PGCs to spermatogonia (reviewed in Reig-Viader, *et al*. 2016). The interaction patterns observed in both genome-wide and per-chromosome interaction heatmaps suggested that this transition is accompanied by an interphase-like genome organization. That is, plaid patterns were well defined, which were indicative of genome A/B compartmentalization (supplementary figures 1 and 3). This observation was also confirmed by analysing distance-dependent interaction frequencies (contact probability) (figure 23A). In this case, the decrease of interaction as a function of genomic distance in spermatogonia followed a similar pattern than fibroblasts. In contrast, we detected distinct patterns in the inter-chromosome/intra-chromosome interaction ratio when compared to somatic cells. In fibroblasts, small chromosomes (e.g. 18 or 19) showed higher intra-/inter-chromosomal interaction ratio than large ones (e.g. 1, 2 or X). This pattern is likely due to the presence of chromosomal territories, where each chromosome is physically separated and occupies a distinct volume within the nucleus (Cremer and Cremer 2010). Since it is known that small and gene-rich chromosomes tend to be located at the centre of the nucleus

(Boyle, *et al*. 2001), they are more likely to interact with each other. In contrast, inter-chromosomal interactions are minimized in spermatogonia (figure 22); that is, intra-/inter-chromosomal interaction ratio decreased by 2-fold approximately and differences among chromosomes were reduced. These results suggest that commitment to enter meiosis is accompanied by a drastic remodelling of chromosomal occupancy inside the nucleus, which appears to be already established in spermatogonia.

At the sub-megabase organization scale, TADs were detected in spermatogonia, although the TAD number was reduced when compared to fibroblasts (from 2,002 to 834 TADs) as well as TAD signal variance (figures 26 and 28). Altogether, our observations suggest that the 3D organization of the genome in spermatogonia is being remodelled prior to prophase I, not only in terms of chromosome localization inside the nucleus but also at the sub-megabase scale.

6.2.2 Compartmentalization is highly re-arranged during prophase I

As meiosis progresses (prophase I), we detected that the high-order chromatin organization reorganized at several levels. We have to take into account that, at this stage of meiosis (prophase I), homologous chromosomes condensate, align and pair (at leptonema), start to synapse (at zygonema) and recombine (at pachynema) (reviewed in Reig-Viader, *et al*. 2016). In the light of our results, these processes are affecting the way chromosomes are organised in prophase I. Overall, both genome-wide and per-chromosome interaction heatmaps suggested the existence of strong local and weak long-range interactions in primary spermatocytes (supplementary figures 1, 4 and 5). This was indicative of highly condensed chromosomes, as previously described for mitotic chromosomes (Naumova, *et al*. 2013).

At the largest chromosome scale, we detected that the inter-/intra-chromosome interaction ratio reached a minimum for all chromosomes in primary spermatocytes (ratio of 0.25) when compared to fibroblasts (ratio of 0.75 on average) and it was stable for all chromosomes (figure 22). These results suggest that chromosome territories are lost in prophase I. In fact, it is known that there is a clustering of telomeres called *bouquet* in primary spermatocytes at leptotene stage that promotes the pairing of homologous chromosomes (Scherthan, *et al.* 1996; Reig-Viader, *et al.* 2016). This telomeric attachment to the nuclear envelope is essential for successful synapsis between homologs (Boateng, *et al*. 2013). Remarkably, the *bouquet* structure was detected in our analysis (and recently validated by Alavattam, *et al*. 2019) by assessing the inter-chromosome subcentromeric interactions as there was statistically higher interaction in leptonema/zygonema relative to fibroblasts (figure 23D). We did not find

statistically significant differences between fibroblast and pachynema/diplonema, suggesting that the *bouquet* is no longer maintained at later stages of prophase I.

The analysis of contact probability in primary spermatocytes also confirmed that the overall reorganization of the genome in prophase I was translated into changes of the condensation status of meiotic chromosomes. Our analysis showed that primary spermatocytes presented the slowest decrease in contact probability when compared to the rest of cell types analysed, with slopes close to -0.5 between genomic distances of 0.5 Mbp and 2.5 (in leptonema/zygonema) or 4.5 Mbp (in pachynema/diplonema) followed by a rapid fall-off. This slope resembles what has been described for mitotic prophase and metaphase chromosomes (Naumova, *et al*. 2013; Gibcus, *et al*. 2018). These studies indicated that the mitotic prophase chromosomes present a slow decrease in contact probability at short distances (slope of -0.5) followed by a rapid fall-off at 3 Mbp (Gibcus, *et al*. 2018). This drop-off is increasingly delayed when cells progress towards metaphase, in which this rapid fall-off is at 10 Mbp (Naumova, *et al*. 2013; Gibcus, *et al*. 2018). Based on these observations, it has been proposed that DNA loops of 60-80 Kbp wrap around a scaffold of proteins in the mitotic pre-metaphase and metaphase chromosomes (Gibcus, *et al*. 2018). Precisely, the rapid fall-off in contact probability observed in pre-metaphase and metaphase chromosomes coincide with the total length of DNA that wraps around the scaffold per turn. In this way, the length increases towards metaphase, shortening metaphase chromosomes (Gibcus, *et al*. 2018). However, chromosomes are organized differently in meiosis: DNA loops are attached to the synaptonemal complex, which brings homologous chromosomes together (Henderson and Keeney, 2005). Consistent with our data, recent studies also described meiotic chromosomes retain a slope of 0.5 in distances up to 5 Mbp, with higher contact probability at longer distances in pachynema relative to leptonema (Wang, *et al*. 2019; Patel, *et al*. 2019). The fact that pachynema/diplonema shows higher contact probability at longer distances than leptonema/zygonema would be explained by an elongation of the DNA loops attached to the synaptonemal complex, thus promoting more interaction at longer loci. Precisely, it has been proposed a loop length of 0.8-1 Mbp in zygonema and 1.5-2 Mbp in pachynema (Patel, *et al*. 2019), mirroring previous cytological studies (Klecner, *et al*. 2003).

In terms of compartmentalization at the sub-chromosome scale, we did not observe A/B compartmentalization in leptonema/zygonema; however, it seems A/B compartments start to appear in pachynema/diplonema again, although this was not evident when analysing the first eigenvector. The unassembling of the synaptonemal complex at diplotene stage might be involved in the reappearance of compartmentalization since it would add rigidity to the chromosomes when homologous are aligned side-by-side. The fact that compartments were

not inferred from the first eigenvector suggests the presence of attenuated A/B compartmentalization, as recently described (Alavattam, *et al*. 2019). Other studies have also suggested A/B compartmentalization in pachynema (Patel, *et al*. 2019; Wang, *et al*. 2019); specifically, Wang, *et al*. (2019) did not predict compartmentalization from the full Hi-C matrix but from submatrices of 10 Mbp along the diagonal, thus predicting compartments using local interaction patterns. This approach was not performed in this work but could be useful to provide clearer eigenvector profiles (supplementary figure 12).

At the sub-megabase scale, we identified a sharp drop in the TAD signal variance, especially in leptonema/zygonema but also in pachynema/diplonema (figure 26). The total number of TADs identified in these cell types was around 300, giving TAD structures of 9 Mbp on average (supplementary table 4). In mammals, TADs range from tens of Kbp up to 2 Mbp with an average of 800 Kbp (Dekker and Heard, 2015). In this context, given the low TAD signal variance, the low number of TADs and their big size, we suggest an absence of this structure in primary spermatocytes (also suggested by Wang, *et al*. 2019 and Patel, *et al*. 2019). The particular organization of primary spermatocytes would prevent the formation of these sub-megabase structures. The length of the DNA loops attached to the synaptonemal complex, estimated between 0.8-1 Mbp in zygonema and between 1.5-2 Mbp in pachynema, would prevent TAD formation. The shortest loops observed in leptonema/zygonema would also explain the lower TAD signal variance observed in this cell type when compared to pachynema/diplonema.

6.2.3 Reprogramming of genome compartmentalization in post-meiotic cells

The end of meiosis results in the formation of haploid cells (round spermatids), which undergo spermiogenesis, a differentiation process that involves an intermediate step (elongated spermatids) to produce male gametes ready for fertilization (sperm) (reviewed in Handel and Schimenti, 2010). The sperm genome is highly compacted, and it is achieved by replacing histones by protamines during spermiogenesis (Balhorn, *et al*. 1984; Hud, *et al*. 1993; Johnson, *et al*. 2011). Protamines are arginine-rich proteins (positively charged) that changes the electrostatic environment of the DNA, thus changing its conformation into toroidal structures of 50 Kbp allowing high compaction (Johnson, *et al*. 2011).

The analysis of genome-wide and per-chromosome interaction heatmaps revealed that round spermatids showed plaid patterns that were blurrier relative to fibroblast and spermatogonia (supplementary figures 1 and 7). This was suggestive of more condensed genomes, as recently reported (Wang, *et al*. 2019; Alavattam, *et al*. 2019). This pattern of genome condensation was confirmed by the contact probability analysis: round spermatids presented an intermediate

state between prophase I (e.g. primary spermatocytes) and interphase-like (e.g. fibroblast and spermatogonia) cell types with less long-range interactions than fibroblast but more than primary spermatocytes (figure 23). The analysis of inter-/intra-chromosome interaction ratio also revealed a weak recovery from prophase I; that is, the interaction ratio was higher in round spermatids for all chromosomes, with exception of sex chromosomes (figure 22). However, like spermatogonia, evidence of chromosome territories was weaker than fibroblasts (figure 22). In fact, it is at round spermatids where centromeres cluster around the chromocenter, which is a visible aggregation of centromeric heterochromatin the centre of the nucleus (Haaf and Ward, 1995; Zalensky, *et al*. 1995; Meyer-Ficca, *et al*. 1998). This was statistically significantly confirmed by assessing the inter-subcentromeric interactions (figure 23D) (and validated by Alavattam, *et al*. 2019).

At the sub-chromosome organization scale, our analyses revealed A/B compartmentalization in round spermatids with a percentage of compartment A (46.94%) higher than in fibroblast (45.68%). However, at the sub-megabase scale, the TAD signal variance remained similar to pachynema/diplonema. In this way, TADs might be also absent in these cell types considering the low TAD signal variance and the fact that most of the predicted TAD boundaries had very low-quality scores (figure 26; supplementary table 4). Consistent with an unclear presence of TADs in this cell type, TADs in round spermatids were described as weak or being similar to pachynema (Wang, *et al*. 2019; Alavattam, *et al*. 2019).

Remarkably in sperm, the analysis of the inter-/intra-chromosomal interactions ratio suggested the presence of chromosomal territories. In fact, inter-chromosomal interactions were greater than intra-chromosomal interactions in all but the X chromosome, with the interactions inversely correlated with chromosomal size (figure 22). Since ratios were higher than in fibroblasts, the higher-order chromatin structure is likely densely packed in sperm, thus favouring inter-chromosome interactions although chromosome territories remain.

Mirroring the pattern observed in round spermatids, blurry plaid patterns were detected in sperm when analysing genome-wide and per-chromosome interaction heatmaps (supplementary figures 1 and 8) together with A/B compartmentalization (figure 25). In fact, sperm showed the higher percentage of compartment A (48.64%). The contact probability analysis also revealed that sperm presented an intermediate state of chromatin condensation between prophase I and interphase-like (figure 23). These results suggest that the sperm genome is more condensed than the interphase-like cell types. In fact, the reorganization of the sperm genome into toroidal structures due to the replacement of histones by protamines leads

to high compactness (Johnson, *et al*. 2011). These toroidal structures might extend compartments A and change the sub-megabase organization scale.

The TAD signal variance observed in sperm was extremely low in sperm, even lower than leptonema/zygonema. These findings contrast with the patterns reported by recent studies: (i) well-defined plaid-patterns, and (ii) well-defined TAD structures (Jung, *et al*. 2017; Ke, *et al*. 2017; Wang, *et al*. 2019). For example, Wang, *et al*. (2019) stated that 78% of TAD boundaries are shared between fibroblast and sperm. The patterns reported were obtained from the isolation of sperm from the supernatant after incubating dissected cauda epididymis. In contrast, we isolated sperm by FACS, ensuring low percentage of somatic contamination. In addition, we observed in our simulations using samples with mixed compositions of fibroblast and sperm that small fractions of somatic contamination might affect the plaid-pattern observed in the interaction heatmaps (figure 27). Altogether, correlation analyses, contact probability and interaction heatmap patterns from these simulations confirm our results in sperm.

6.2.4 Dynamics of the X chromosome architecture during spermatogenesis

In eutherian males, sex chromosomes (X and Y) synapse at the Pseudo-Autosomal Region (PAR), thus the X chromosome remains mostly asynapsed during prophase I (Yan and McCarrey, 2009). The asynapsed status of sex chromosomes has been suggested to trigger the inactivation of sex chromosomes (the so-called Meiotic Sex Chromosome Inactivation, MSCI) during the transition from zygonema to pachynema (Turner, 2007; Yan and McCarrey, 2009). In this way, chromosomes X and Y are condensed and physically separated from the autosomes in the periphery of the nucleus forming the sex body (Turner, 2007; Yan and McCarrey, 2009). The sex body is diluted in further stages of spermatogenesis (e.g. spermatids), but the sex chromosome inactivation is maintained and sex chromosomes appear as heterochromatic domains called Post-Meiotic Sex Chromatin (PMSC) (Namekawa, *et al*. 2006; Turner, 2007).

In this context, the dynamics of the X chromosome architecture during spermatogenesis was also assessed in our study. The inter-chromosome/intra-chromosome interaction ratio revealed that the X chromosome had a comparable ratio to similar-size chromosomes (e.g. 1 or 2) in fibroblasts and spermatogonia. The inter-/intra-chromosome interaction ratio remained stable in leptonema/zygonema with no visible differences between autosome and sexual chromosomes. This tendency remained the same in pachynema/diplonema for autosomal chromosomes, but the ratio of the X chromosome slightly decreased (figure 22). Since the separation of sex chromosomes in the sex body at the nucleus periphery would reduce the

probability of inter-interactions, the decrease of the inter-/intra-chromosome interaction ratio in the X chromosome might suggest the presence of MSCI (also observed in Wang, *et al*. 2019 and Alavattam, *et al*. 2019). In both round spermatids and sperm, the X chromosome also showed an inter-/intra-chromosome interaction ratio below chromosomes of similar size, but it was more evident than in pachynema/diplonema. In this sense, the specific chromosome organization in pachynema/diplonema might have masked MSCI due to the low inter-chromosome interactions shown in autosomes.

During MSCI, the X chromosome is not only isolated at the nucleus periphery, but it is also transcriptionally silenced (Turner, 2007). In this sense, the inactivation of the X chromosome can be also studied using RNA-seq data. In recent studies, a reduction in the transcriptional activity was also recently observed in both pachynema and round spermatids (Wang, *et al*. 2019). According to our data, we observed a sharp reduction in the number of expressing genes in chromosome X (-38.3%) relative to the autosomes (-18.2%) in pachynema/diplonema *versus* spermatogonia (table 8). Reduction in the transcriptional activity was also maintained in round spermatids and sperm, thus confirming the inactivation of the X chromosome in late meiotic and post-meiotic cells.

Consistent with the per-chromosome interaction heatmaps (supplementary figures 2-8), which showed no plaid patterns, A/B compartmentalization was not detected in the X chromosomes of leptonema/zygonema, pachynema/diplonema, round spermatids and sperm (supplementary figures 18-22). Accordantly, absence of compartmentalization in pachynema and round spermatids was recently described (Alavattam, *et al*. 2019). In contrast, plaid patterns and A/B compartmentalization were found in fibroblasts and spermatogonia (supplementary figures 16 and 17). At the sub-megabase organization scale, TADs could be inferred in both fibroblasts and, in a weaker way, in spermatogonia (supplementary figures 9 and 10), but were not present in the remaining cell types, which showed low TAD signal variance (supplementary figures 25-29) (confirmed in pachynema and round spermatids by Wang, *et al*. 2019). Altogether, the inter-/intra-chromosome interaction ratio, the RNA-seq data analysis and the higher-order chromatin organization of the X chromosome, MSCI was confirmed in late prophase I and PMSC in post-meiotic cell types.

6.3 Functional signatures of spermatogenesis

Beyond changes in the higher-order chromatin structure, spermatogenesis also relies on highly regulated gene expression mechanisms at both the transcriptional and post-transcriptional level (Bettegowda and Wilkinson, 2010; de Mateo and Sassone-Corsi, 2014; Hammoud, *et al*. 2014).

This is the case of small non-coding RNA, such as piRNAs, which play a relevant role in gametogenesis. In fact, disruption of the piRNA pathway has been related to meiotic arrest at the zygotene stage (Fu and Wang, 2014). During spermatogenesis, two transcription waves of piRNA have been suggested, one before meiosis I (pre-pachytene piRNA) and another one in the transition from pachynema to round spermatids (pachytene piRNA) (de Mateo and Sassone-Corsi, 2014; Fu and Wang, 2014). Likewise, two additional transcriptional waves of total mRNA are known to occur, the first one before meiosis I and the second one at primary spermatocytes (Sassone-Corsi, 2002; de Mateo and Sassone-Corsi, 2014; da Cruz, *et al.* 2016). In this context, we took advantage of the RNA-seq data generated in our laboratory to identify functional signatures of spermatogenesis. This also served us to stablish a link between the dramatic chromatin remodelling that takes places during spermatogenesis and the regulatory pathways involved in this process.

Overall, we detected that expression of protein-coding genes decreased in favour of non-coding genes (pseudogenes, lncRNA genes and asRNA genes) during spermatogenesis (figure 19). Since compartments A are correlated with open chromatin state regions (Lieberman-Aiden, *et al*. 2009), we analysed the biotypes of the expressing genes detected in cell-specific A compartment regions. Consistent with our transcriptome analysis, about half of the expressing genes in A-specific compartment regions in meiotic and post-meiotic germ cells are non-coding (table 11). We have to take into consideration that RNAs were selected by their poly-A tail when preparing the RNA-seq library. Since most lncRNAs do not have poly-A tail, lncRNAs might have been underestimated in our experiment (Derrien, *et al*. 2012; Zhao, *et al*. 2018). Since this class of non-coding RNA might be involved in chromatin remodelling and transcriptional regulation (e.g. mediating histone modifications), post-transcriptional processing (e.g. regulating splicing or being a source of miRNA), or chromatin looping (e.g. mediating proximity between enhancer and promoter) (Mercer, *et al*. 2009; Barbosa Dogini, *et al*. 2014; Dykes and Emanueli, 2017), the potential roles of lncRNA in spermatogenesis are promising.

Focusing on specific functions of protein coding genes, several spermatogenesis-related genes were identified in spermatogonia. That was the case, for instance, of predicted genes *Gm1993* and *Gm5169* which contain *Sycp3*-like domains. *Sycp3* (Synaptonemal Complex Protein 3) is a component of the synaptonemal complex established during meiosis. In this sense, the presence of genes with *Sycp3*-like domains in spermatogonia suggests that genes implicated in meiotic processes are already transcribed before entering meiosis (Martinez-Garay, *et al*. 2002). This was also the case of genes involved in repair of the DSBs (e.g. *Dmc1* and *Tex15*), or *Pot1A* as part of the shelterin complex at the *bouquet* stage at leptonema (Wang, *et al*. 2018). These

are examples of genes significantly more expressed in spermatogonia than in pachynema/diplonema.

Involved in chromatin cohesion, *Stag3* was found significantly more expressed in pachynema/diplonema than in spermatogonia, although it was already expressed in spermatogonia. In fact, *Stag3* was found in an A-specific region common in all germ cells. Additional genes involved in the formation of the synaptonemal complex, such as *Sycp* genes (e.g. *Sycp1*, *Sycp2* or *Sycp3*) were also significantly more expressed in pachynema/diplonema than spermatogonia or round spermatids. Precisely, the synaptonemal complex is still assembled in pachynema, thus it is consistent with the active expression of the *Sycp* family at this stage. In addition, in A-specific regions, an *Hyal5* (hyaluronidase) and *Fgf10* were found and being significantly more expressed in pachynema/diplonema than in spermatogonia. On the one hand, hyaluronidases are involved in the acrosome reaction, thus important for fertilization. On the other hand, receptors for *Fgf10* are in the acrosome and stimulate sperm motility (Saucedo, *et al*. 2018).

Several important genes for fertilization were found significantly more expressed in round spermatids than in other stages of spermatogenesis. That was the case, for example, of genes involved in the formation of the sperm acrosome (*Spaca*), the zona pellucida binding protein (*Zpbp*), *Abcg2* or *Plcz1* (Korfanty, *et al*. 2012; Swegen, *et al*. 2018; Cross, *et al*. 1998; Scharenberg, *et al*. 2009). In the case of hyaluronidases, *Hyal5*, which was already found as being significantly more expressed in pachynema/diplonema than spermatogonia, significantly increased its expression in round spermatids. Other hyaluronidases, such as *Hyal4* and *Hyal6*, were identified in A-specific regions of round spermatids and sperm. Genes related with sperm motility were also identified: *Drd2* (dopamine receptor D2) and *Deup1* were found as up-regulated in round spermatids relative to pachynema/diplonema, besides of being located in an A-specific compartment region shared between round spermatids and sperm. While dopamine affects sperm motility (Urra, *et al*. 2014), *Deup1* is essential for centriole formation and development of flagella (Zhao, *et al*. 2013). Protamines (*Prm1*, *Prm2* and *Prm3*) were also found significantly more expressed in round spermatids than pachynema/diplonema, consistent with the replacement of histones by protamines during spermiogenesis (Johnson, *et al*. 2011).

Previous works have reported the presence of genes related to meiotic recombination and chromosome segregation were remarkable in leptonema/zygonema while the presence of genes related to sperm motility and “sperm-egg recognition” were remarkable from pachynema to round spermatids (da Cruz, *et al*. 2016). Consistent with these findings, we identified 18

genes with “sperm motility” ontology annotation and 10 genes with “sperm-egg recognition” ontology annotation being significantly more expressed in pachynema/diplonema than spermatogonia. In addition, we identified 46 genes with “meiotic cell cycle” also being more expressed in pachynema/diplonema than spermatogonia. In this way, we found pachynema/diplonema to be involved in early expression of important genes for spermiogenesis but also expressing genes for regulating the progression of meiosis I. At this stage, though, recombination is resolved. Protein complexes in charge of this task should have been previously transcribed and coded at early stages. In this sense, we identified 4 genes with “DNA recombination” ontology annotation (*Msh2*, *Tex11*, *Prdm9*, *Swap70*) being significantly more expressed in spermatogonia than in pachynema/diplonema.

In the case of sperm, but also in the other cell types, the expression of compartments A were significantly higher than the expression of compartments B (figure 30). In addition, the percentage of compartment A that covers the genome gradually increased during spermatogenesis, reaching its maximum in sperm (figure 29). These results are in agreement with the recent idea that sperm are not inactive cells (Jodar, *et al*. 2016; Jung *et al*. 2017). In fact, it has been shown that a large number of sperm promoters are in an active epigenetic state, suggesting that this genetic information can influence embryo development upon fertilization (Jung, *et al.* 2017). Nevertheless, like the number of the expressing genes, gene expression progressively decreased during spermatogenesis (figure 30). Although RNA is found in sperm, it should be further studied whether it is a result of active transcription or it is the remnants from previous stages (Ren, *et al*. 2017). We cannot either confirm nor reject the hypothesis of active transcription in sperm since we did not use the so-called spike-in controls in our experimental design. In this way, we could only evaluate the relative changes in gene expression without the possibility to assess the absolute changes.

Overall, our work provides a comprehensive overview of the functional regulation of the high-order chromatin organization in all stages of meiosis and along spermiogenesis. We unravelled previously undescribed stages of genome compartmentalization and 3D organization together with the in-depth profiling of functional cell-specific signatures identified from gene expression data and compartment switching dynamics. In this way, we provide evidence on the existence of a fine-tuning between chromatin remodelling and gene expression in germ cells.

Chapter 7: Conclusions

The conclusions of this work are the following:

1. GreeNC represents one of the most comprehensive plant lncRNA databases including 45 species and more than 200,000 lncRNAs.

2. GreeNC overcomes limitations identified in previous databases as it (i) includes high data quality, (ii) represents an easy-to-use and friendly GUI, (iii) availability of an API, (iv) contains simple query options, and (v) offers the possibility of batch downloads.

3. AIR is the first end-to-end solution for the analysis of high-throughput RNA-seq data that does not require previous bioinformatics skills and its not limited to model species (more than 150,000 genomes available).

4. AIR permitted the analysis of 45 gigabytes in less than 6 hours of RNA-seq data derived from mouse germ cells, resulting in the identification several DEGs related with meiosis and spermiogenesis consistent with the sequential development of spermatogenesis.

5. The analysis of the dynamics of the higher-order chromatin organization of the mouse genome during spermatogenesis reveals principles of chromosome assembly at different levels of resolution: (i) intra-/inter-chromosomal interaction ratio, (ii) distance-dependent interaction frequencies, (iii) genomic compartments and (iv) topologically associating domains.

6. Chromosome territories are partially or totally lost in pre-meiotic, meiotic or post-meiotic cells with the exception of sperm. This is consistent with the chromosome organization events that takes place during meiosis I (*bouquet,* pairing of homologous chromosomes and recombination) or round spermatids (centromere clustering at the chromocenters).

7. Three chromosome condensation patterns are observed during spermatogenesis: (i) interphase-like (e.g. fibroblasts and spermatogonia), (ii) mitotic-like (e.g. primary spermatocytes), and (iii) an intermediate state (e.g. round spermatids and sperm).

8. The genome organization at the sub-chromosome scale (e.g. A/B compartments) is maintained in all cell types but mainly lost in primary spermatocytes.

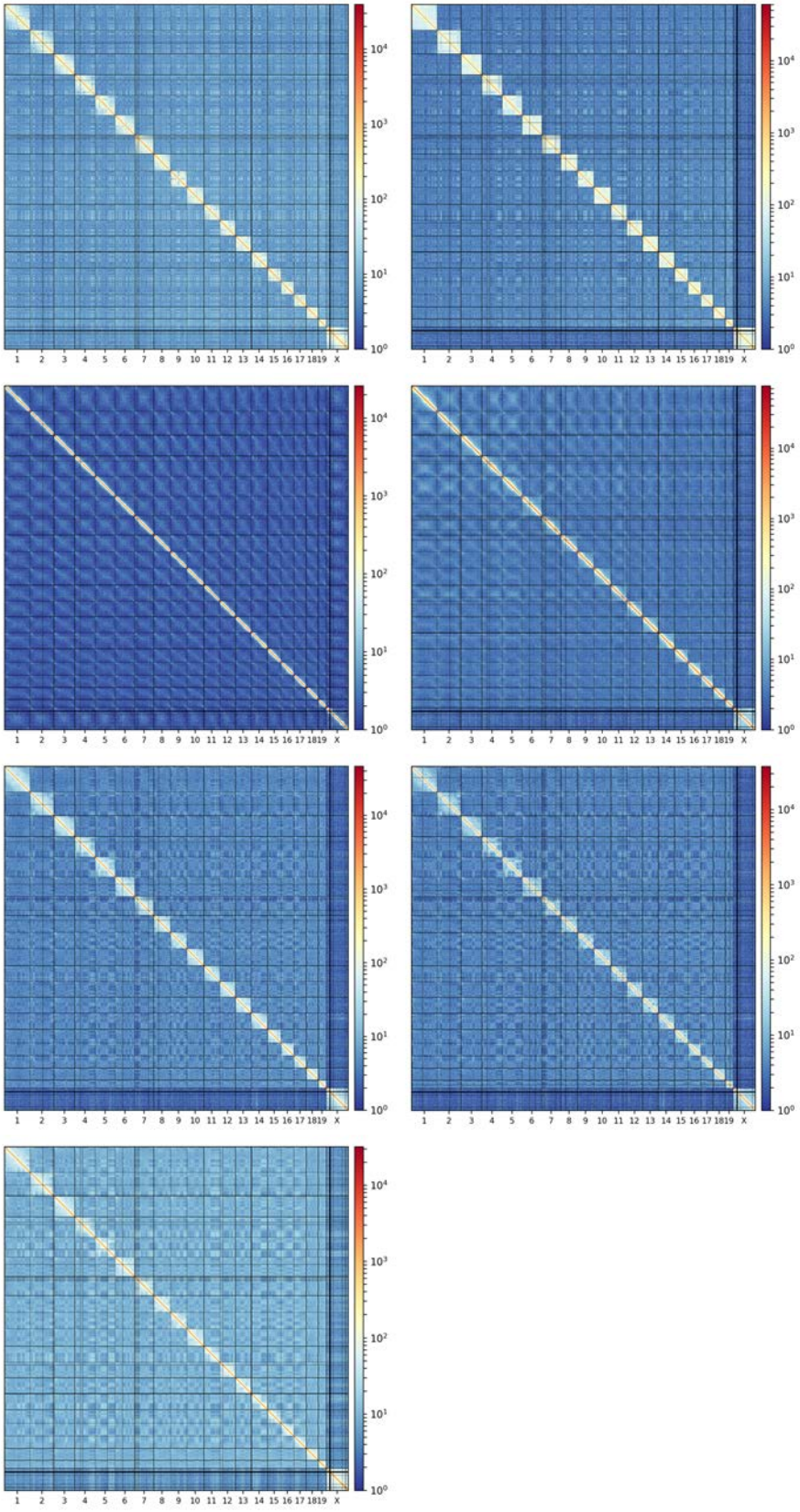
9. Several cell-specific A compartments regions expressing genes related with meiosis and spermiogenesis functions are present in the genome of germ cells. This pattern is consistent with the existence of a fine-tuning between chromatin remodelling and gene expression.

10. At the sub-megabase organization scale, TADs are detected in spermatogonia, but their presence remains poorly defined or inexistent in primary spermatocytes and sperm.

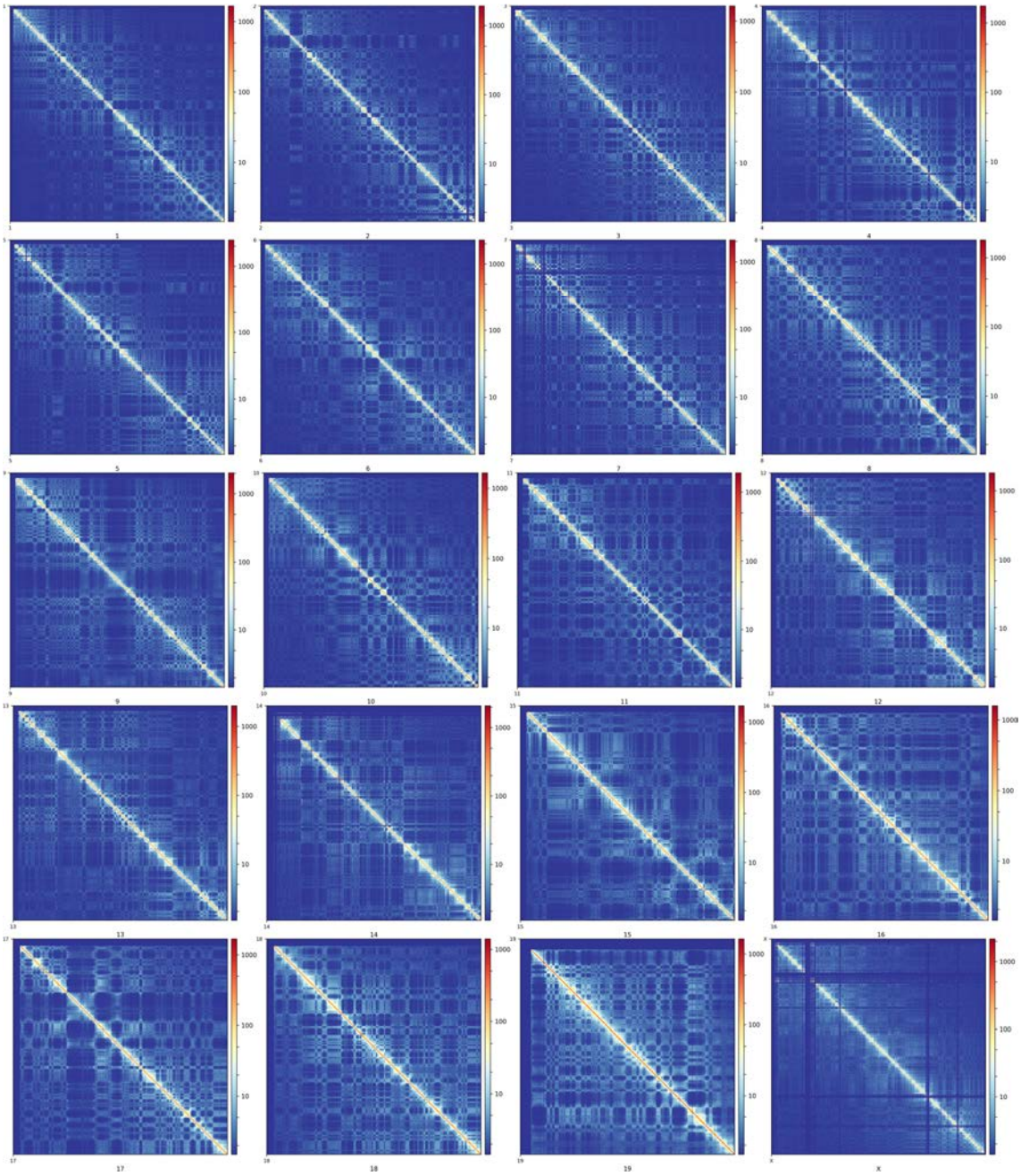
11. The X chromosome suffers global chromatin remodelling and silencing during prophase I; neither A/B compartmentalization nor TADs are detected in primary spermatocytes. This observation is consistent with the meiotic sex chromosome inactivation (MSCI).

12. The formation of post-meiotic sex chromatin (PMSC) correlates with chromatin remodelling that results in low inter-/intra-chromosomal interactions and an absence of neither A/B compartmentalization in the X chromosome of round spermatids.

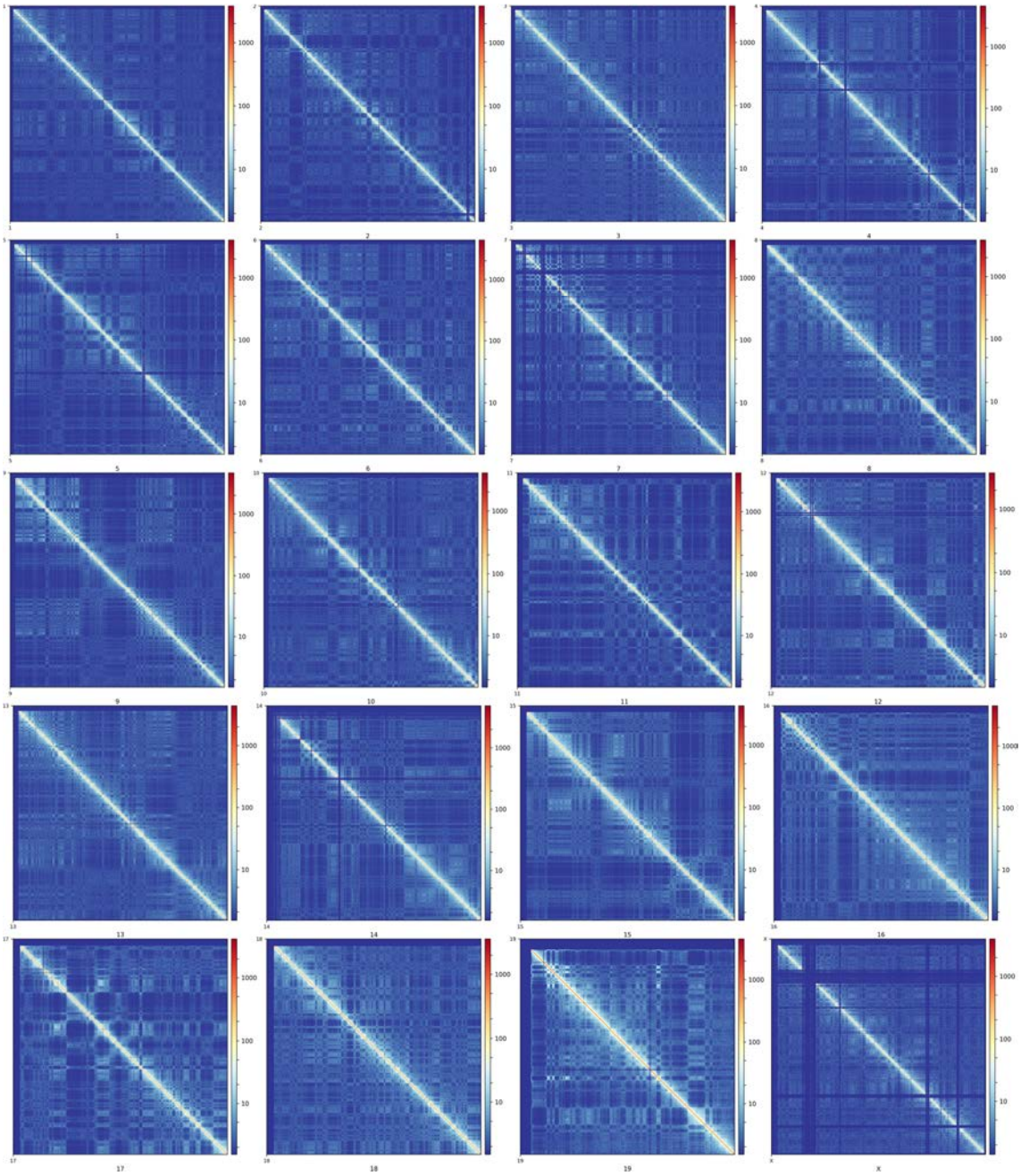
Supplementary figures



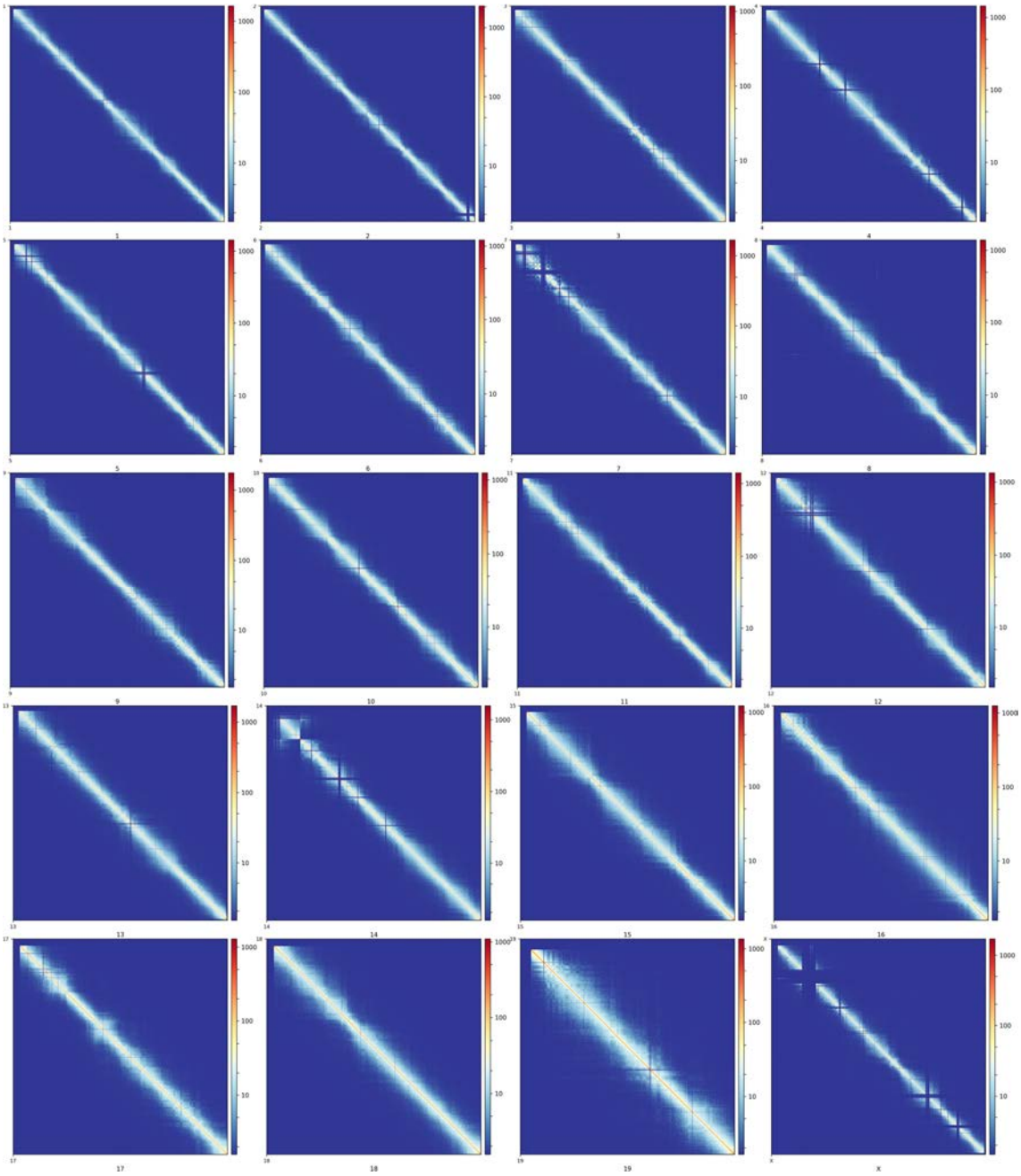
Supplementary figure 1. Genome-wide ICE-corrected interaction heatmaps. First row: fibroblast (left) and spermatogonia (right). Second row: leptonema/zygonema (left) and pachynema/diplonema (right). Third row: secondary spermatocytes (left) and round spermatids (right). Forth row: sperm.



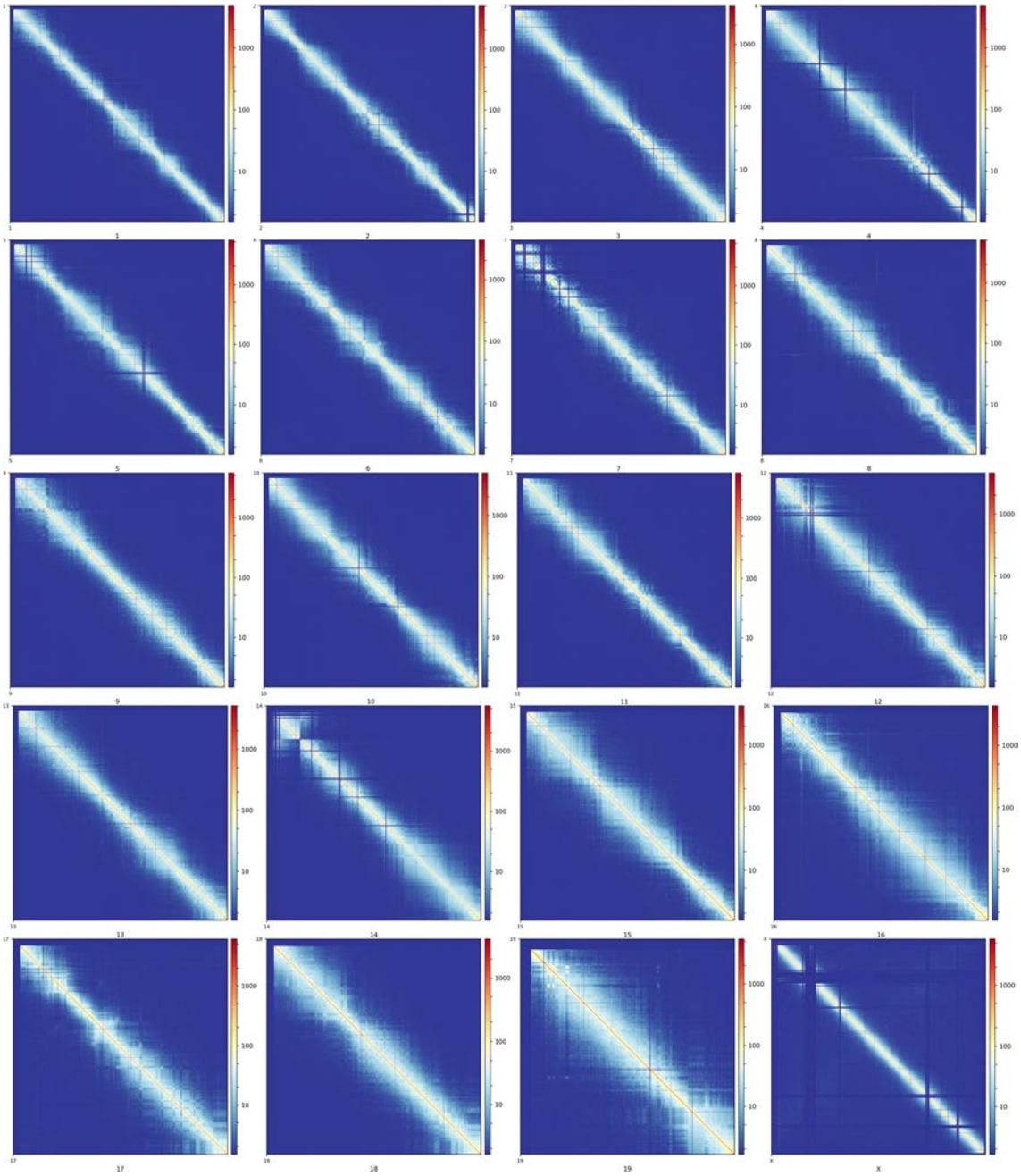
Supplementary figure 2. Per-chromosome ICE-corrected interaction heatmaps in fibroblast. First row: chromosomes 1-4. Second row: chromosomes 5-8. Third row: chromosomes 9-12. Forth row: chromosomes 13-16. Fifth row: chromosomes 17-19 and X.



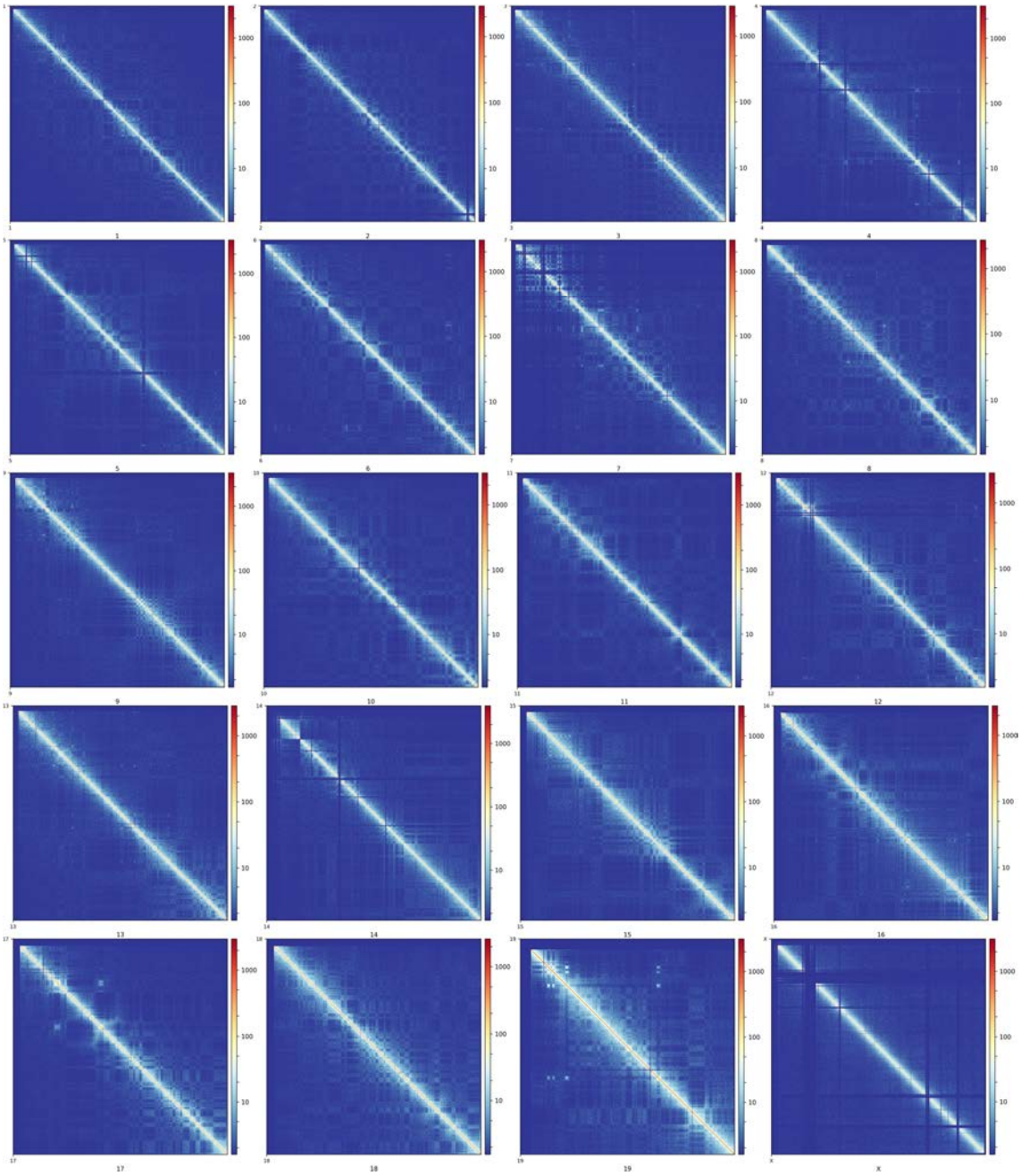
Supplementary figure 3. Per-chromosome ICE-corrected interaction heatmaps in spermatogonia. First row: chromosomes 1-4. Second row: chromosomes 5-8. Third row: chromosomes 9-12. Forth row: chromosomes 13-16. Fifth row: chromosomes 17-19 and X.



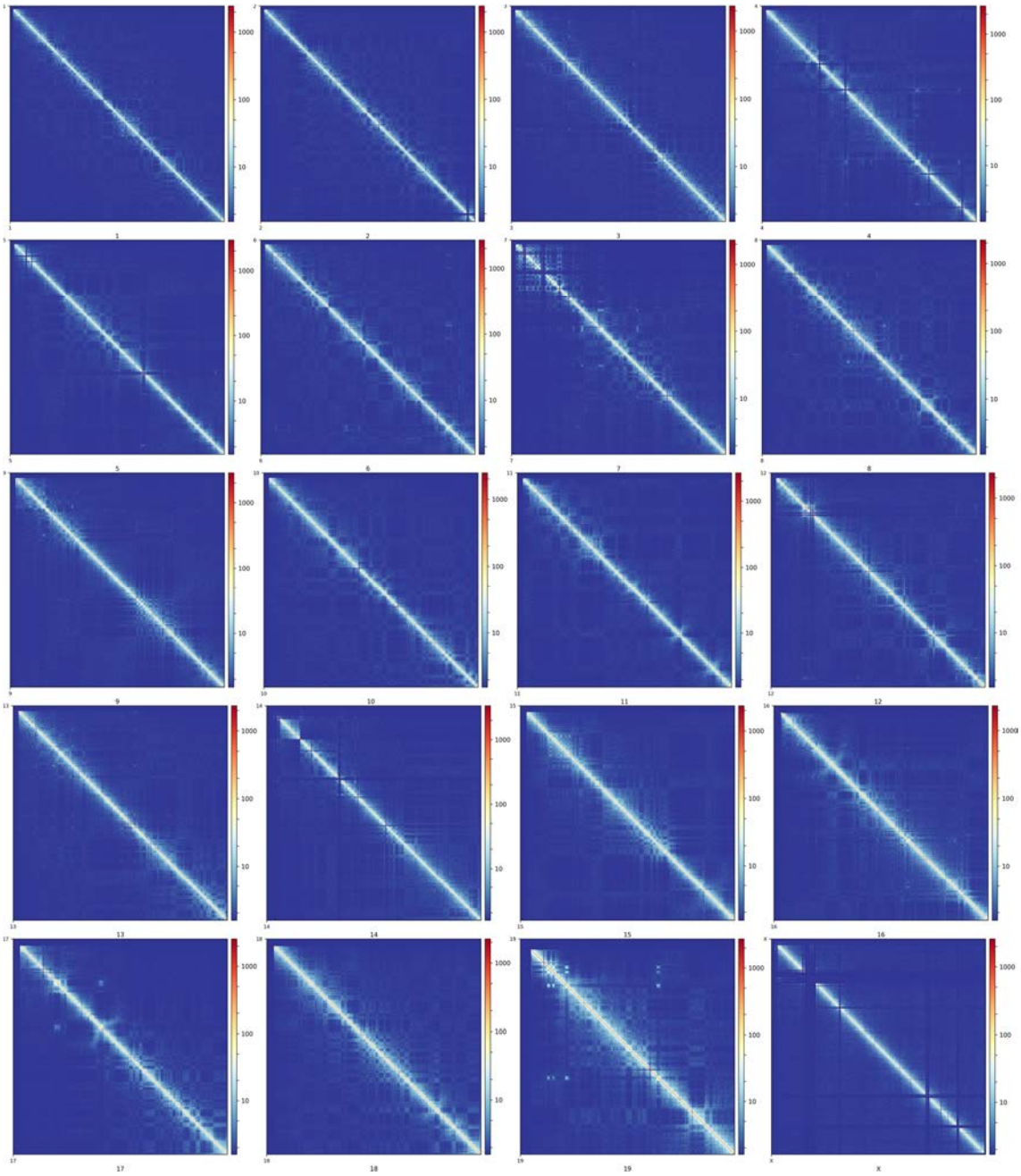
Supplementary figure 4. Per-chromosome ICE-corrected interaction heatmaps in leptonema/zygonema. First row: chromosomes 1-4. Second row: chromosomes 5-8. Third row: chromosomes 9-12. Forth row: chromosomes 13-16. Fifth row: chromosomes 17-19 and X.



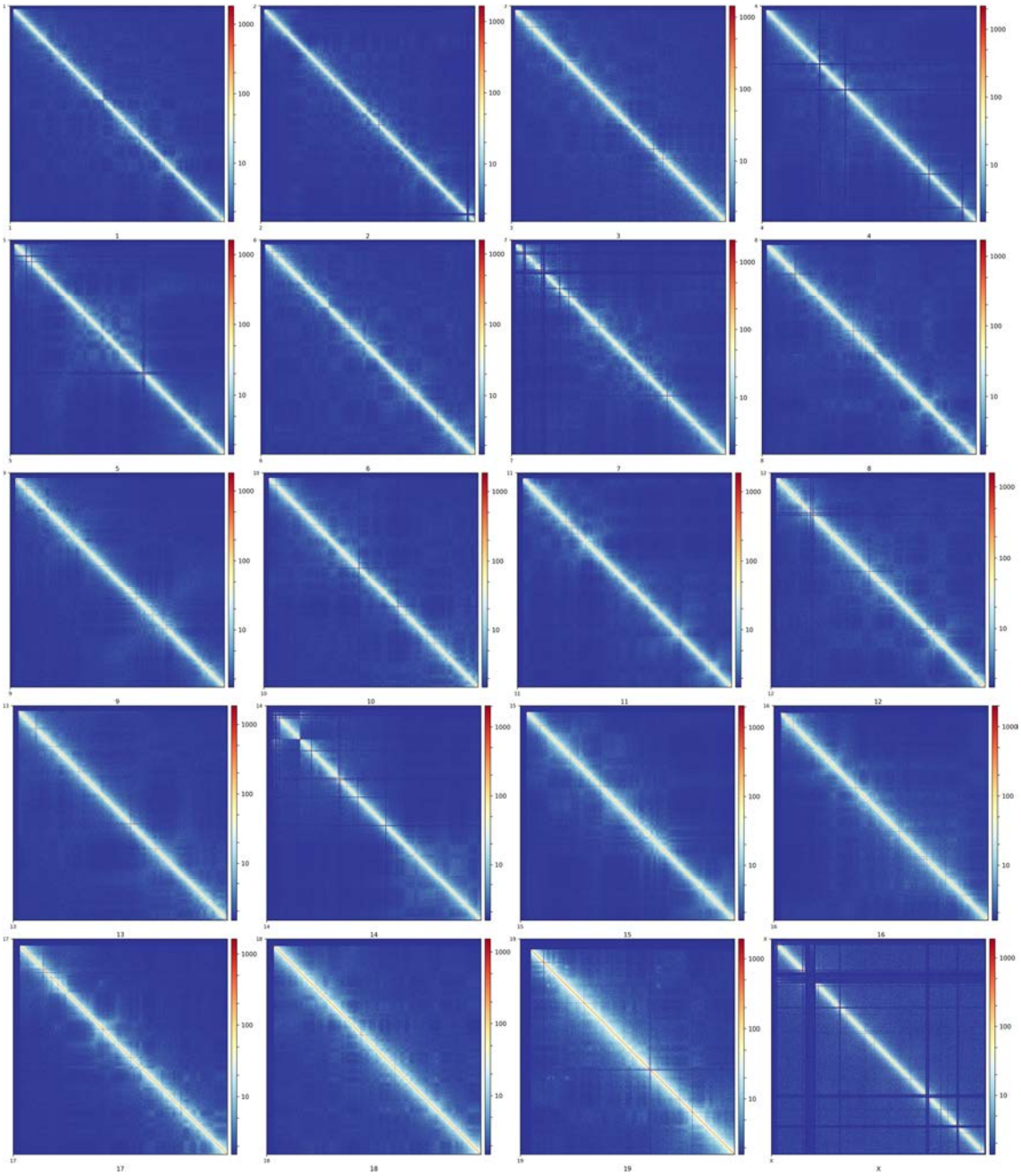
Supplementary figure 5. Per-chromosome ICE-corrected interaction heatmaps in pachynema/zygonema. First row: chromosomes 1-4. Second row: chromosomes 5-8. Third row: chromosomes 9-12. Forth row: chromosomes 13-16. Fifth row: chromosomes 17-19 and X.



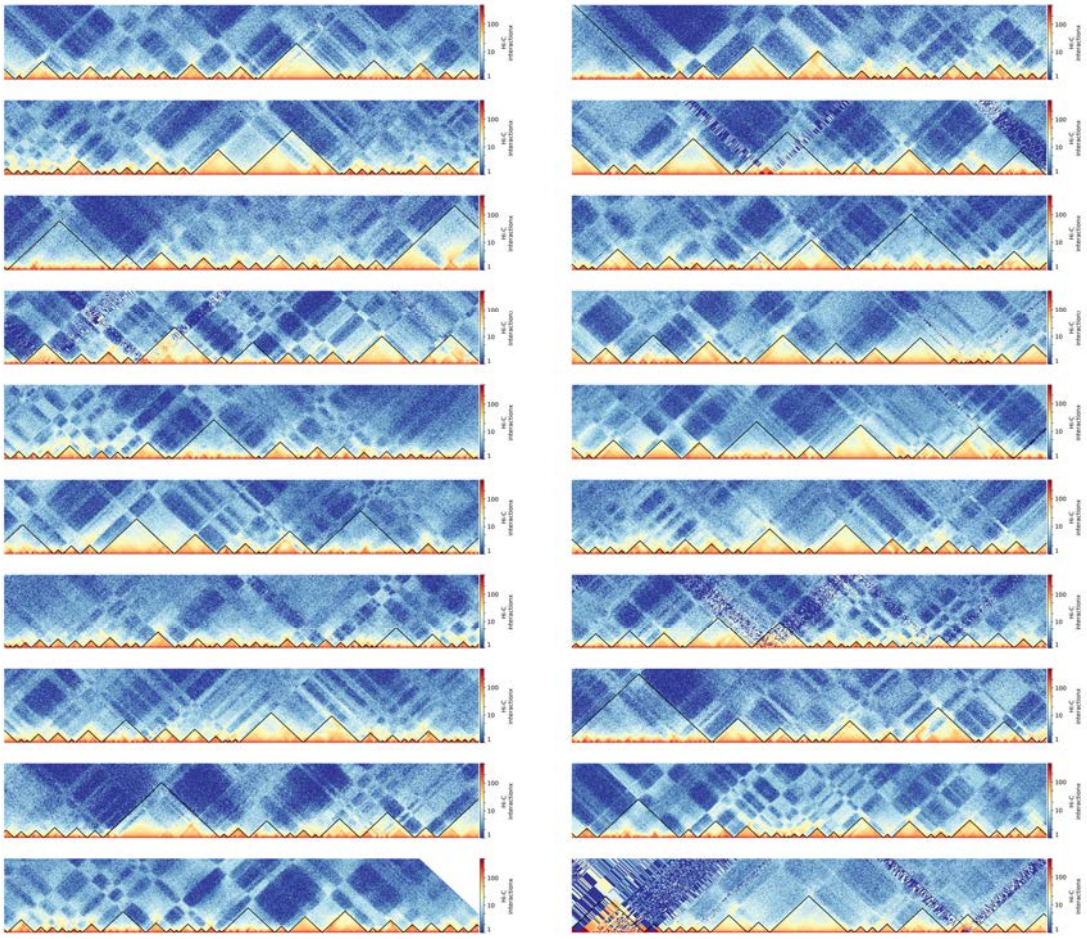
Supplementary figure 6. Per-chromosome ICE-corrected interaction heatmaps in secondary spermatocytes. First row: chromosomes 1-4. Second row: chromosomes 5-8. Third row: chromosomes 9-12. Forth row: chromosomes 13-16. Fifth row: chromosomes 17-19 and X.



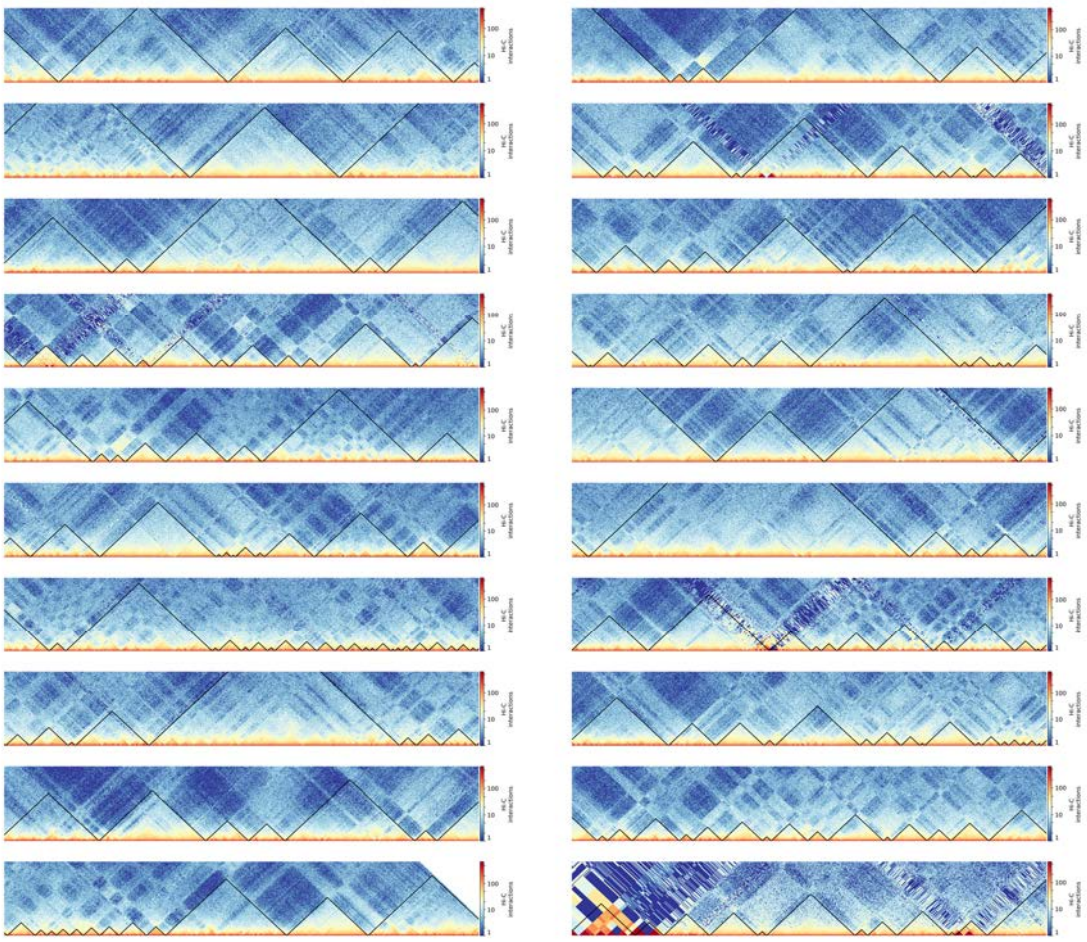
Supplementary figure 7. Per-chromosome ICE-corrected interaction heatmaps in round spermatids. First row: chromosomes 1-4. Second row: chromosomes 5-8. Third row: chromosomes 9-12. Forth row: chromosomes 13-16. Fifth row: chromosomes 17-19 and X.



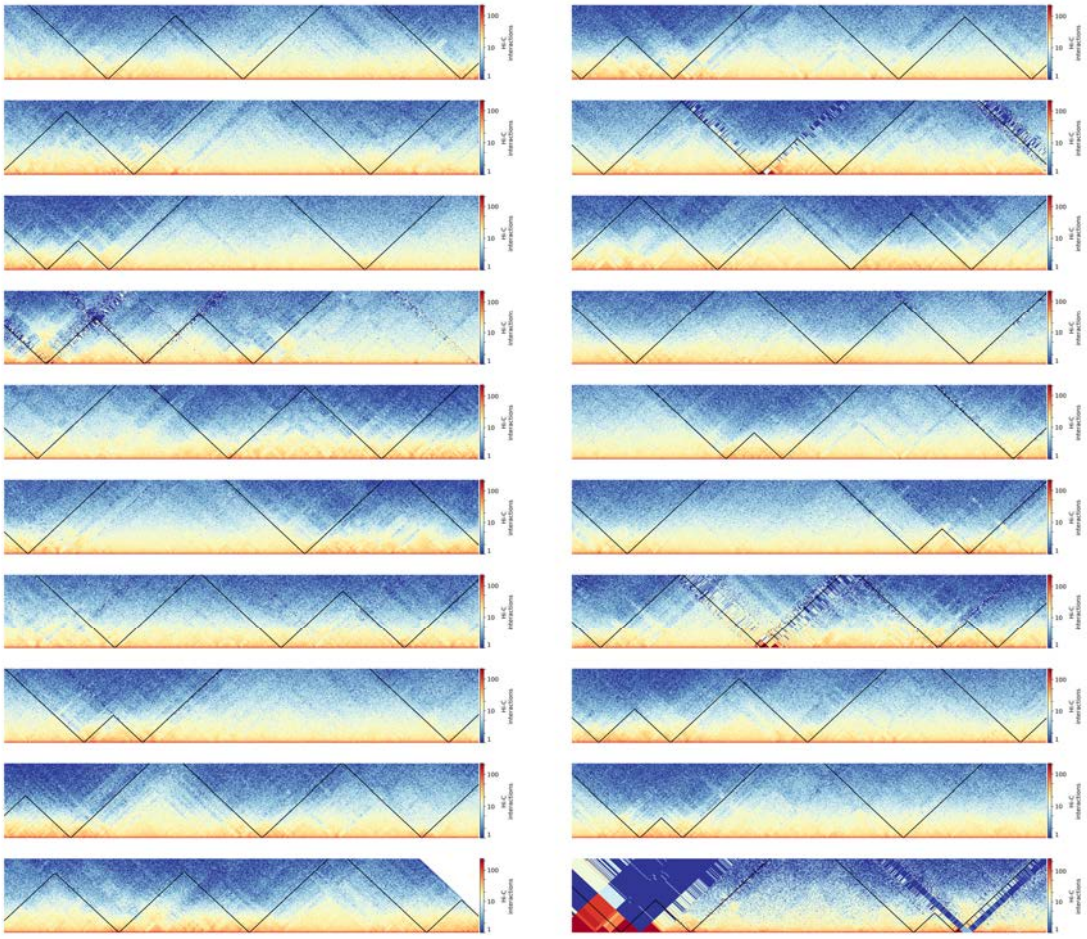
Supplementary figure 8. Per-chromosome ICE-corrected interaction heatmaps in sperm. First row: chromosomes 1-4. Second row: chromosomes 5-8. Third row: chromosomes 9-12. Forth row: chromosomes 13-16. Fifth row: chromosomes 17-19 and X.



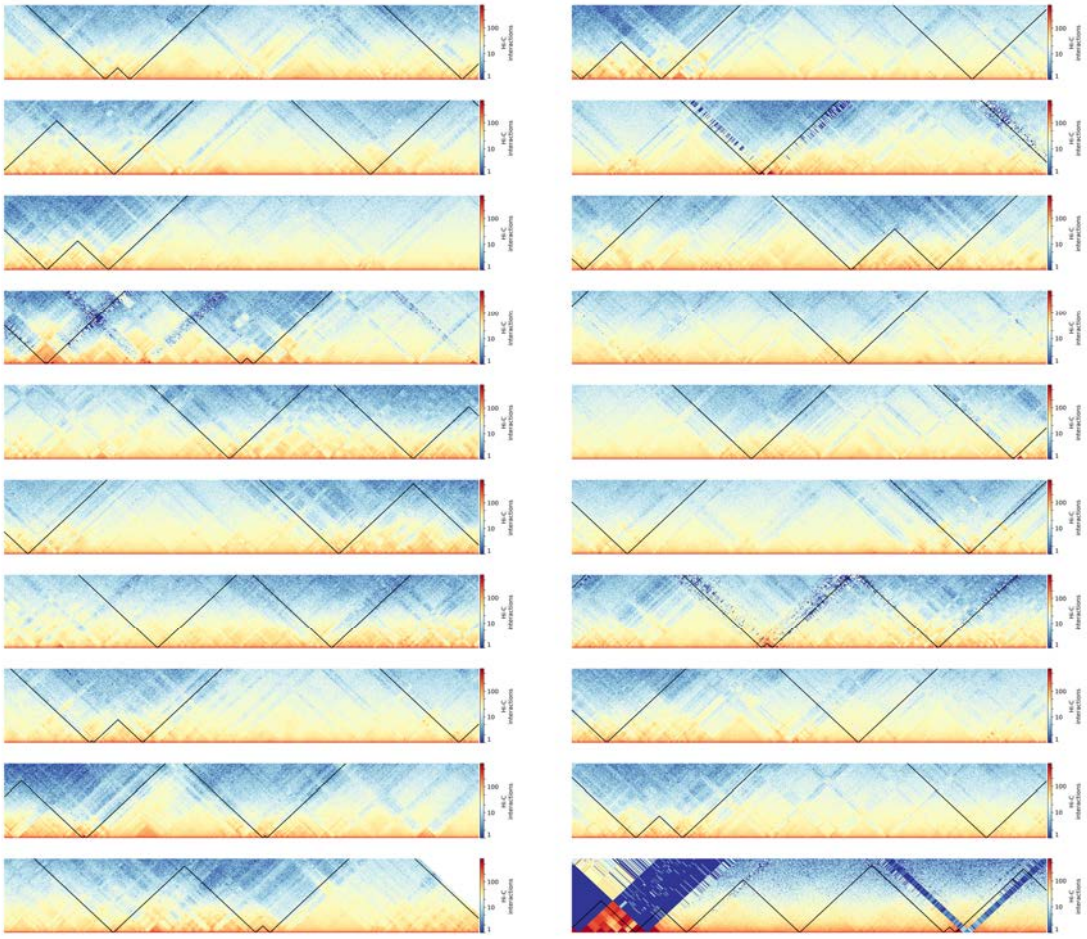
Supplementary figure 9. Focused, per-chromosome ICE-corrected interaction heatmaps in fibroblast. These heatmaps show the region between 50-80 Mbp for each chromosome. First row: chromosomes 1 (left) and 2 (right). Second row: chromosomes 3 (left) and 4 (right). Third row: chromosomes 5 (left) and 6 (right). Forth row: chromosomes 7 (left) and 8 (right). Fifth row: chromosomes 9 (left) and 10 (right). Sixth row: chromosomes 11 (left) and 12 (right). Seventh row: chromosomes 13 (left) and 14 (right). Eighth row: chromosomes 15 (left) and 16 (right). Ninth row: chromosomes 17 (left) and 18 (right). Tenth row: chromosomes 19 (left) and X (right).



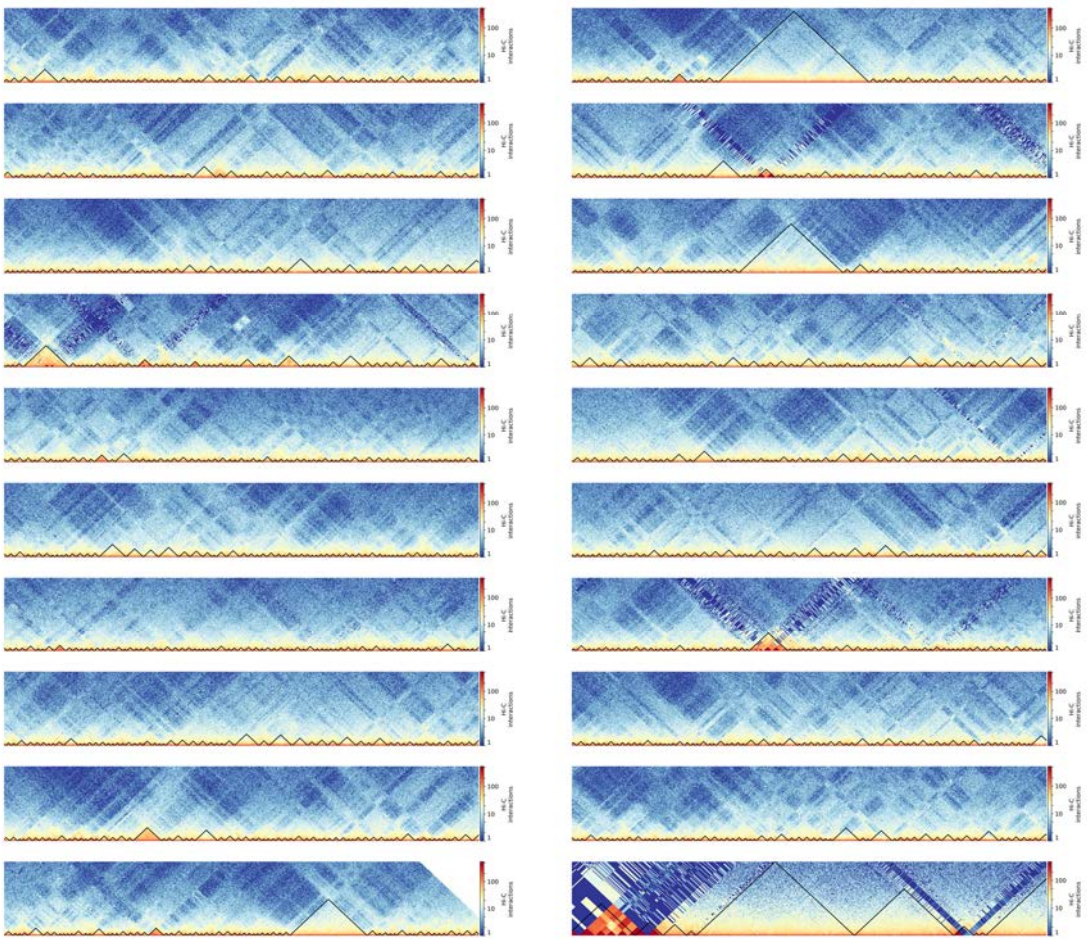
Supplementary figure 10. Focused, per-chromosome ICE-corrected interaction heatmaps in spermatogonia. These heatmaps show the region between 50-80 Mbp for each chromosome. First row: chromosomes 1 (left) and 2 (right). Second row: chromosomes 3 (left) and 4 (right). Third row: chromosomes 5 (left) and 6 (right). Forth row: chromosomes 7 (left) and 8 (right). Fifth row: chromosomes 9 (left) and 10 (right). Sixth row: chromosomes 11 (left) and 12 (right). Seventh row: chromosomes 13 (left) and 14 (right). Eighth row: chromosomes 15 (left) and 16 (right). Ninth row: chromosomes 17 (left) and 18 (right). Tenth row: chromosomes 19 (left) and X (right).



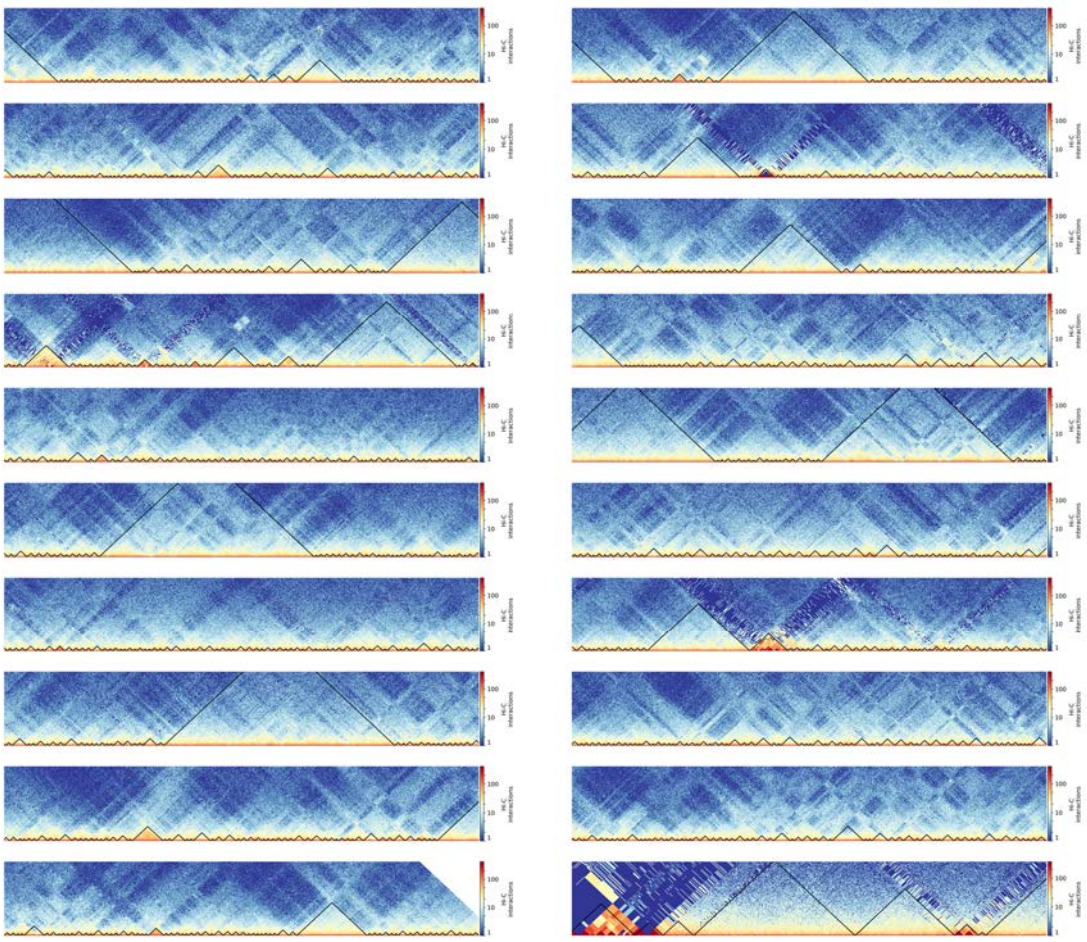
Supplementary figure 11. Focused, per-chromosome ICE-corrected interaction heatmaps in leptonema/zygonema. These heatmaps show the region between 50-80 Mbp for each chromosome. First row: chromosomes 1 (left) and 2 (right). Second row: chromosomes 3 (left) and 4 (right). Third row: chromosomes 5 (left) and 6 (right). Forth row: chromosomes 7 (left) and 8 (right). Fifth row: chromosomes 9 (left) and 10 (right). Sixth row: chromosomes 11 (left) and 12 (right). Seventh row: chromosomes 13 (left) and 14 (right). Eighth row: chromosomes 15 (left) and 16 (right). Ninth row: chromosomes 17 (left) and 18 (right). Tenth row: chromosomes 19 (left) and X (right).



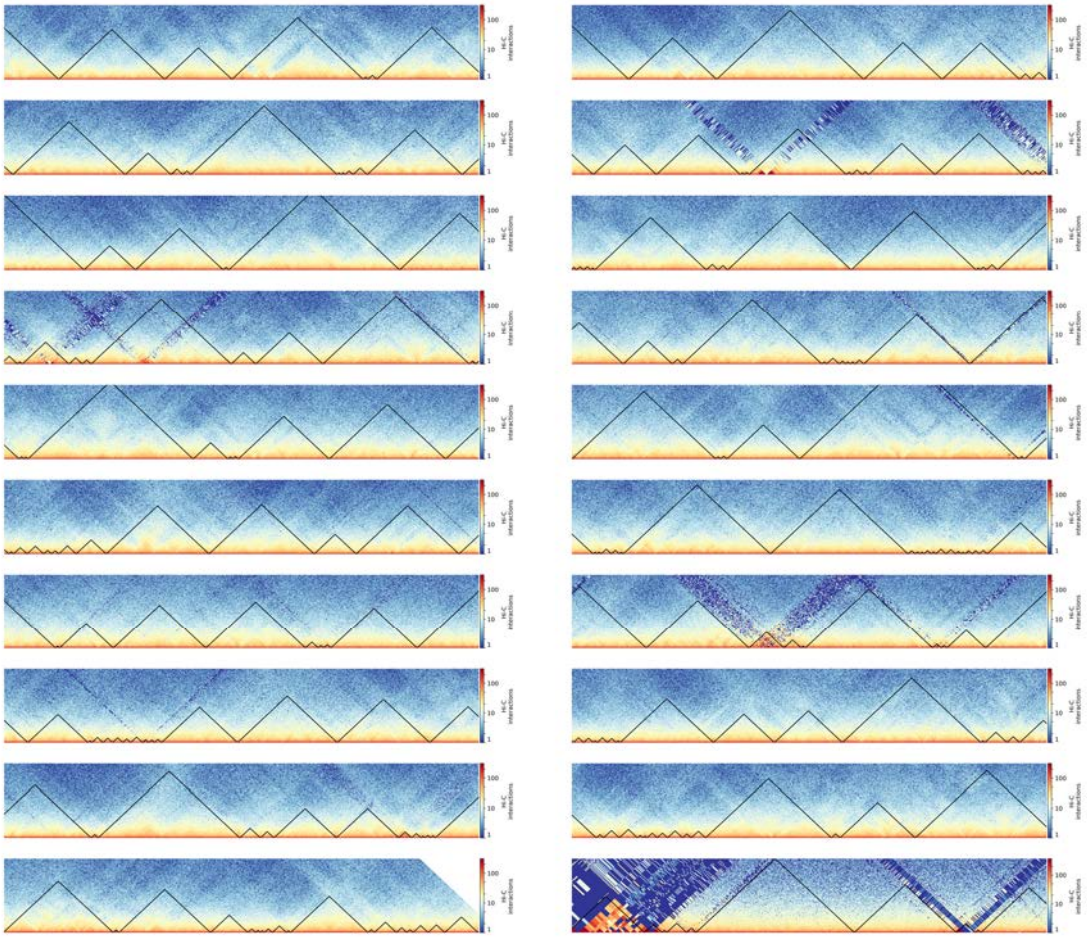
Supplementary figure 12. Focused, per-chromosome ICE-corrected interaction heatmaps in pachynema/diplonema. These heatmaps show the region between 50-80 Mbp for each chromosome. First row: chromosomes 1 (left) and 2 (right). Second row: chromosomes 3 (left) and 4 (right). Third row: chromosomes 5 (left) and 6 (right). Forth row: chromosomes 7 (left) and 8 (right). Fifth row: chromosomes 9 (left) and 10 (right). Sixth row: chromosomes 11 (left) and 12 (right). Seventh row: chromosomes 13 (left) and 14 (right). Eighth row: chromosomes 15 (left) and 16 (right). Ninth row: chromosomes 17 (left) and 18 (right). Tenth row: chromosomes 19 (left) and X (right).



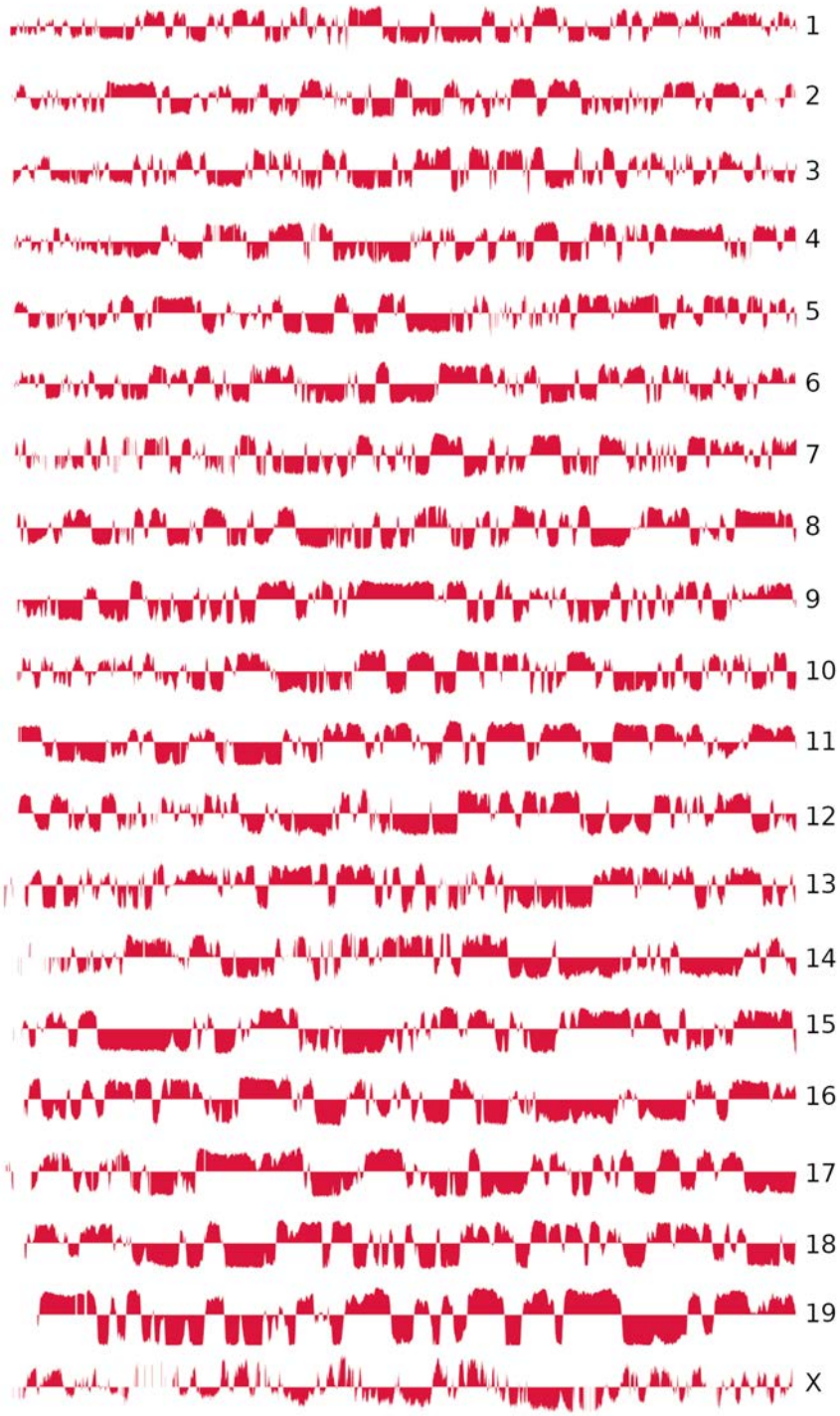
Supplementary figure 13. Focused, per-chromosome ICE-corrected interaction heatmaps in secondary spermatocytes. These heatmaps show the region between 50-80 Mbp for each chromosome. First row: chromosomes 1 (left) and 2 (right). Second row: chromosomes 3 (left) and 4 (right). Third row: chromosomes 5 (left) and 6 (right). Forth row: chromosomes 7 (left) and 8 (right). Fifth row: chromosomes 9 (left) and 10 (right). Sixth row: chromosomes 11 (left) and 12 (right). Seventh row: chromosomes 13 (left) and 14 (right). Eighth row: chromosomes 15 (left) and 16 (right). Ninth row: chromosomes 17 (left) and 18 (right). Tenth row: chromosomes 19 (left) and X (right).



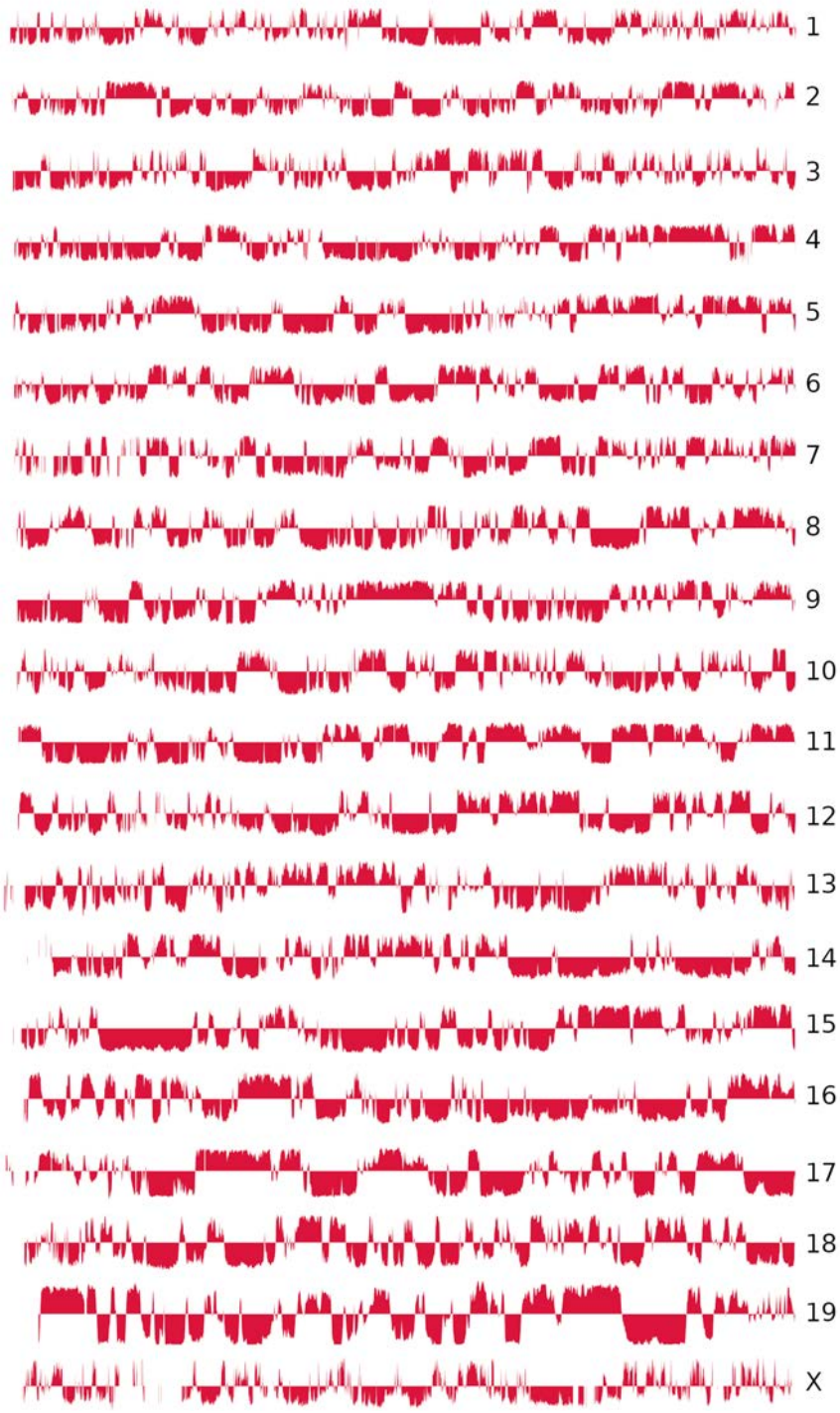
Supplementary figure 14. Focused, per-chromosome ICE-corrected interaction heatmaps in round spermatids. These heatmaps show the region between 50-80 Mbp for each chromosome. First row: chromosomes 1 (left) and 2 (right). Second row: chromosomes 3 (left) and 4 (right). Third row: chromosomes 5 (left) and 6 (right). Forth row: chromosomes 7 (left) and 8 (right). Fifth row: chromosomes 9 (left) and 10 (right). Sixth row: chromosomes 11 (left) and 12 (right). Seventh row: chromosomes 13 (left) and 14 (right). Eighth row: chromosomes 15 (left) and 16 (right). Ninth row: chromosomes 17 (left) and 18 (right). Tenth row: chromosomes 19 (left) and X (right).



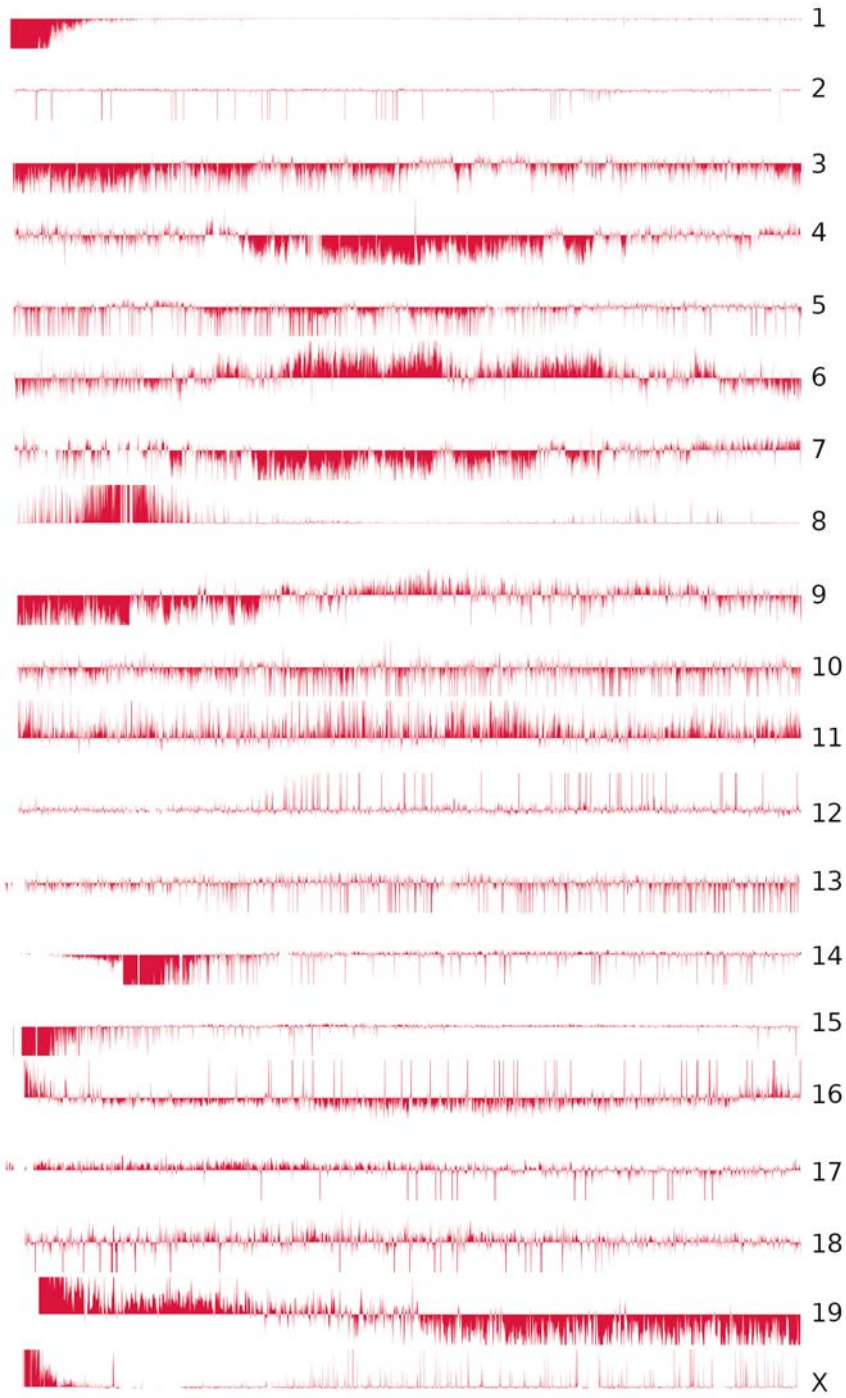
Supplementary figure 15. Focused, per-chromosome ICE-corrected interaction heatmaps in sperm. These heatmaps show the region between 50-80 Mbp for each chromosome. First row: chromosomes 1 (left) and 2 (right). Second row: chromosomes 3 (left) and 4 (right). Third row: chromosomes 5 (left) and 6 (right). Forth row: chromosomes 7 (left) and 8 (right). Fifth row: chromosomes 9 (left) and 10 (right). Sixth row: chromosomes 11 (left) and 12 (right). Seventh row: chromosomes 13 (left) and 14 (right). Eighth row: chromosomes 15 (left) and 16 (right). Ninth row: chromosomes 17 (left) and 18 (right). Tenth row: chromosomes 19 (left) and X (right).



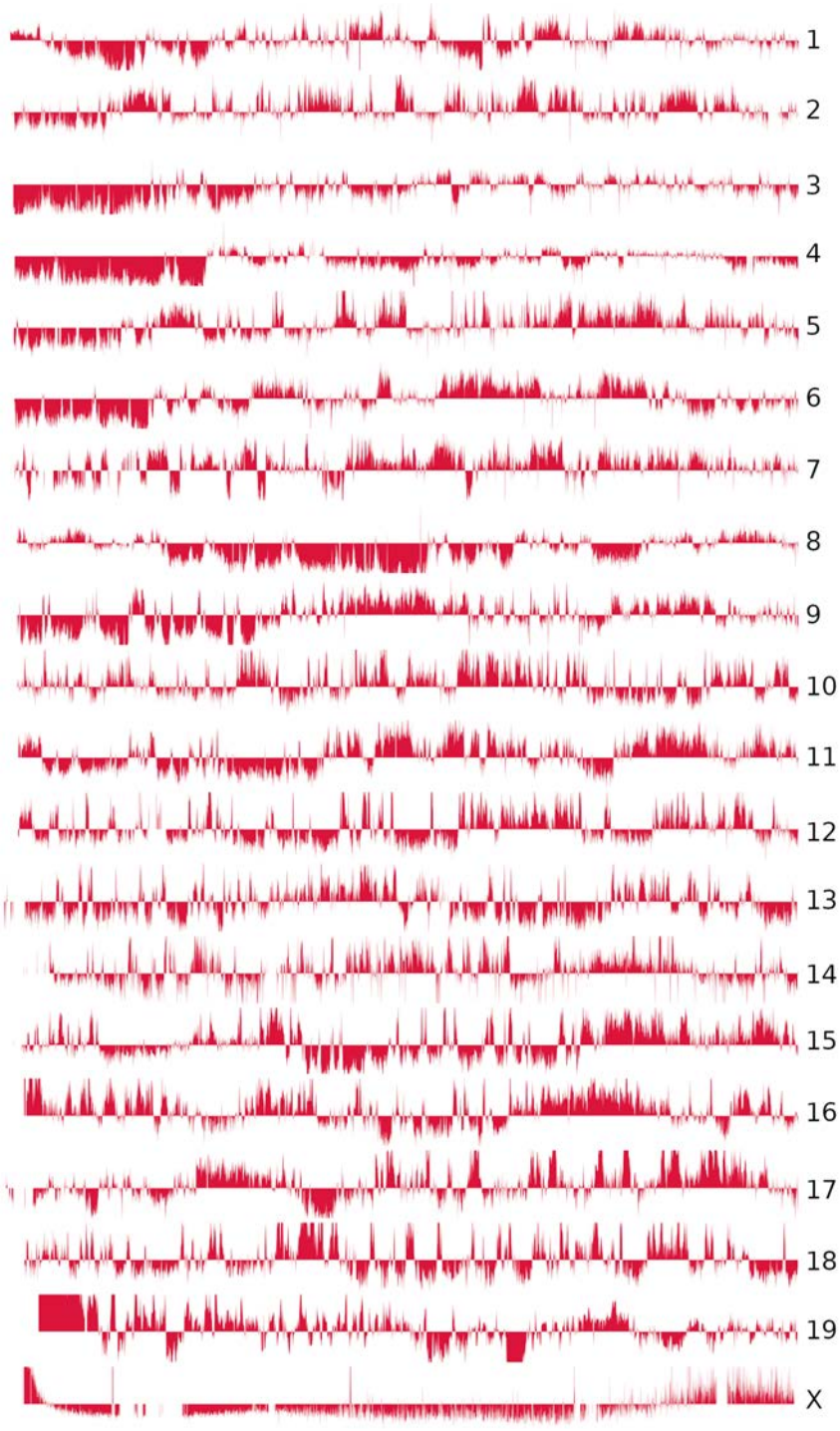
Supplementary figure 16. Per-chromosome eigenvector in fibroblast. Positive values represent A compartments while negative values represent B compartments.



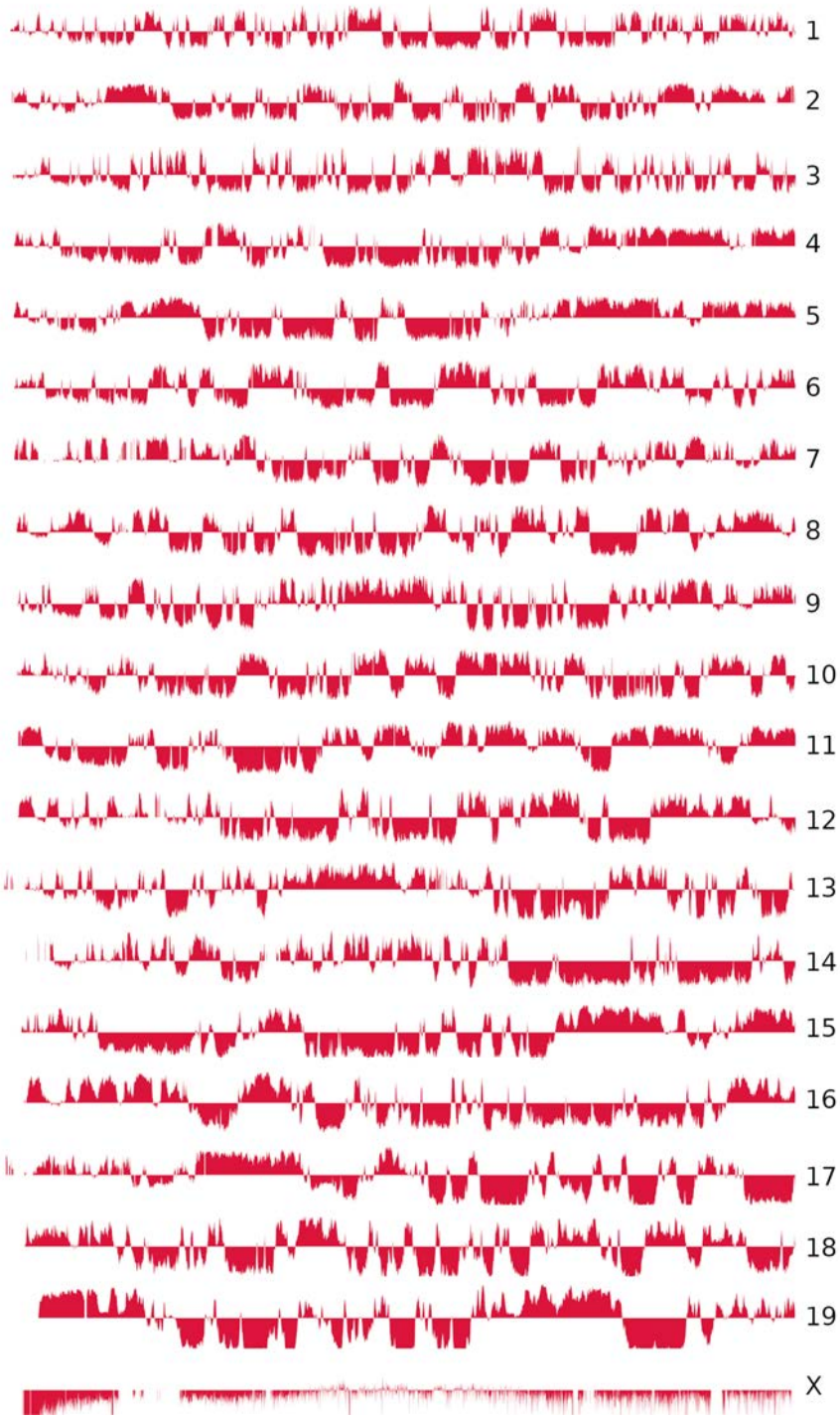
Supplementary figure 17. **Per-chromosome eigenvector in spermatogonia.** Positive values represent A compartments while negative values represent B compartments.



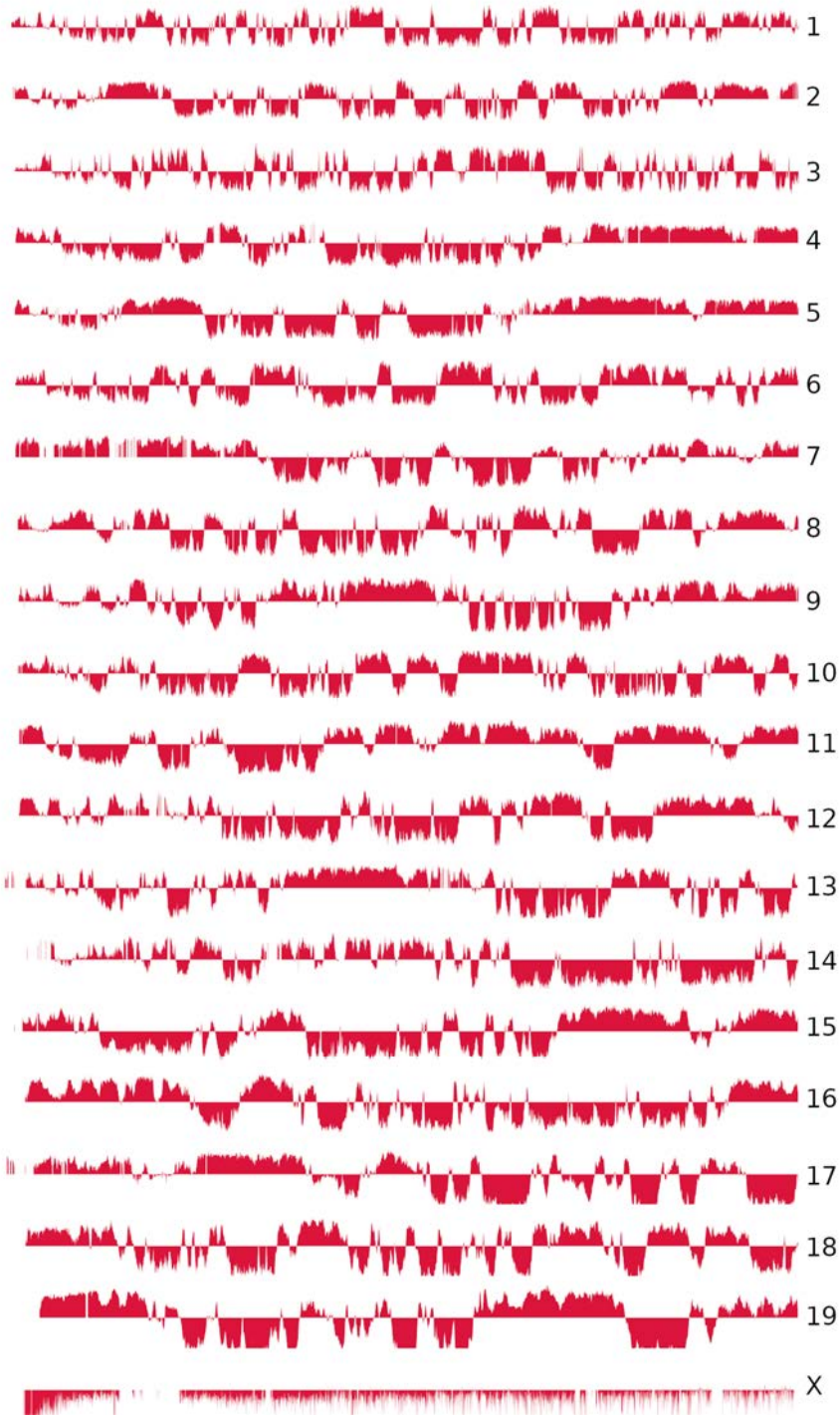
Supplementary figure 18. Per-chromosome eigenvector in leptonema/zygonema. Positive values represent A compartments while negative values represent B compartments.



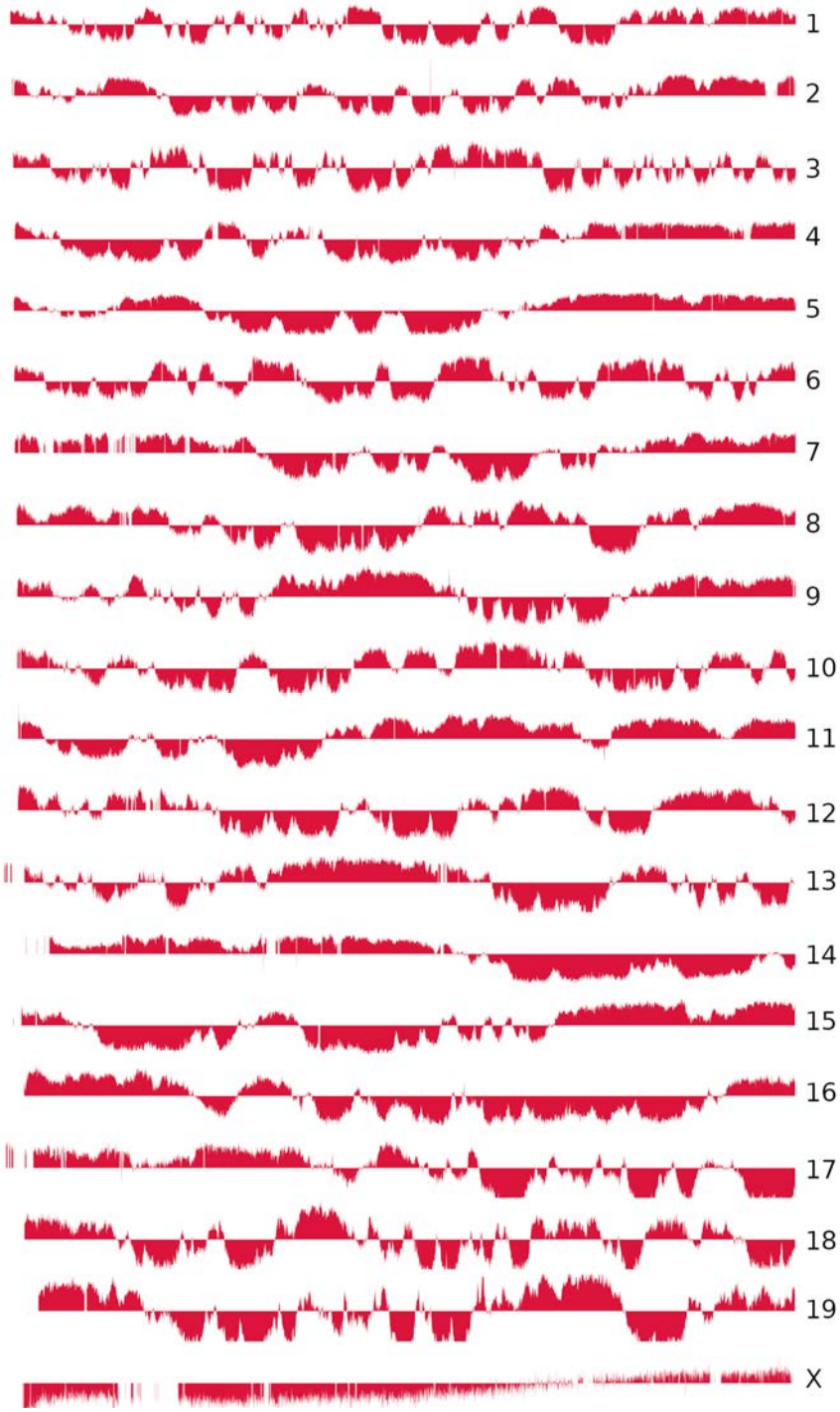
Supplementary figure 19. Per-chromosome eigenvector in pachynema/diplonema. Positive values represent A compartments while negative values represent B compartments.



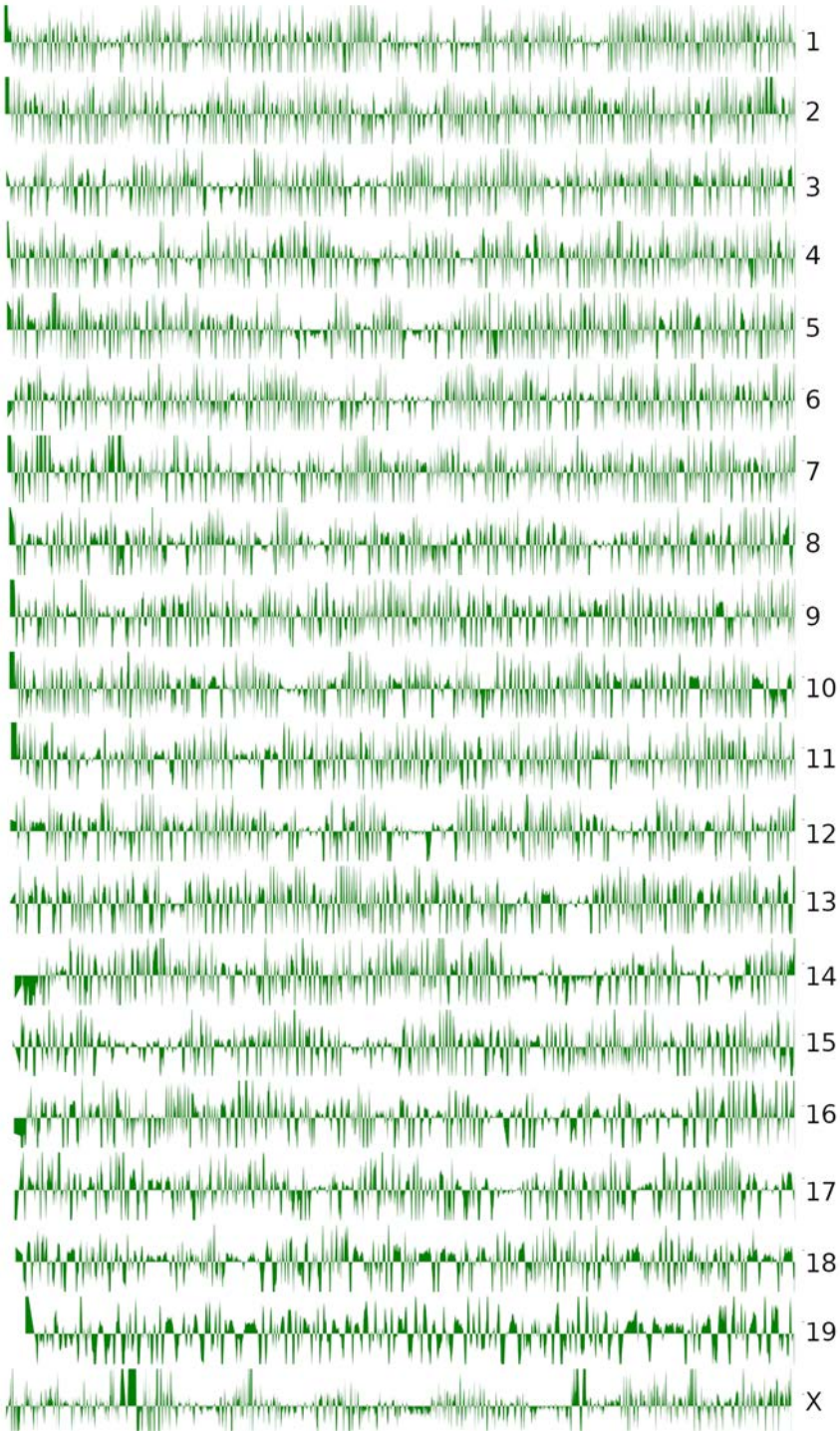
Supplementary figure 20. Per-chromosome eigenvector in secondary spermatocytes. Positive values represent A compartments while negative values represent B compartments.



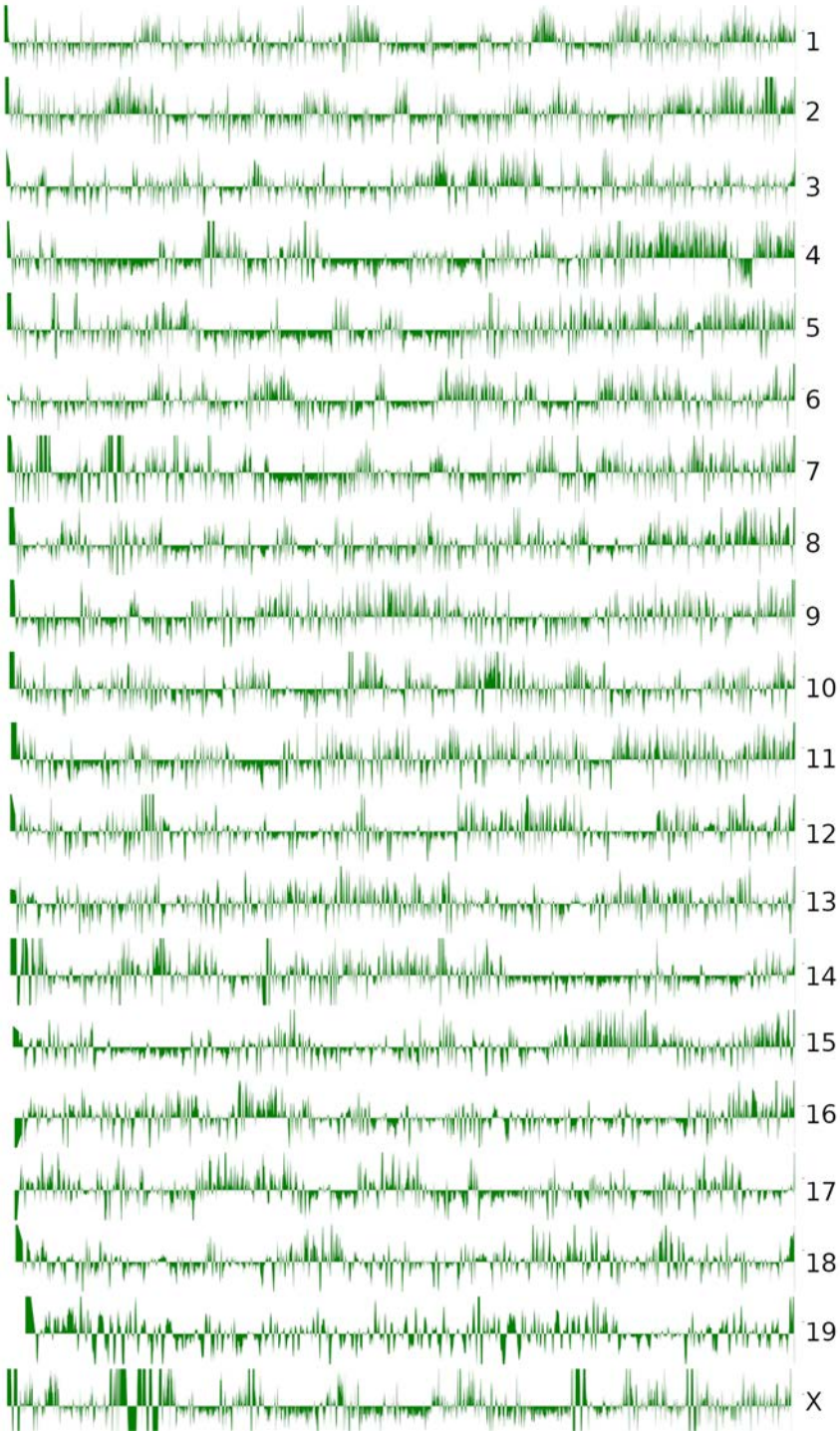
Supplementary figure 21. Per-chromosome eigenvector in round spermatids. Positive values represent A compartments while negative values represent B compartments.



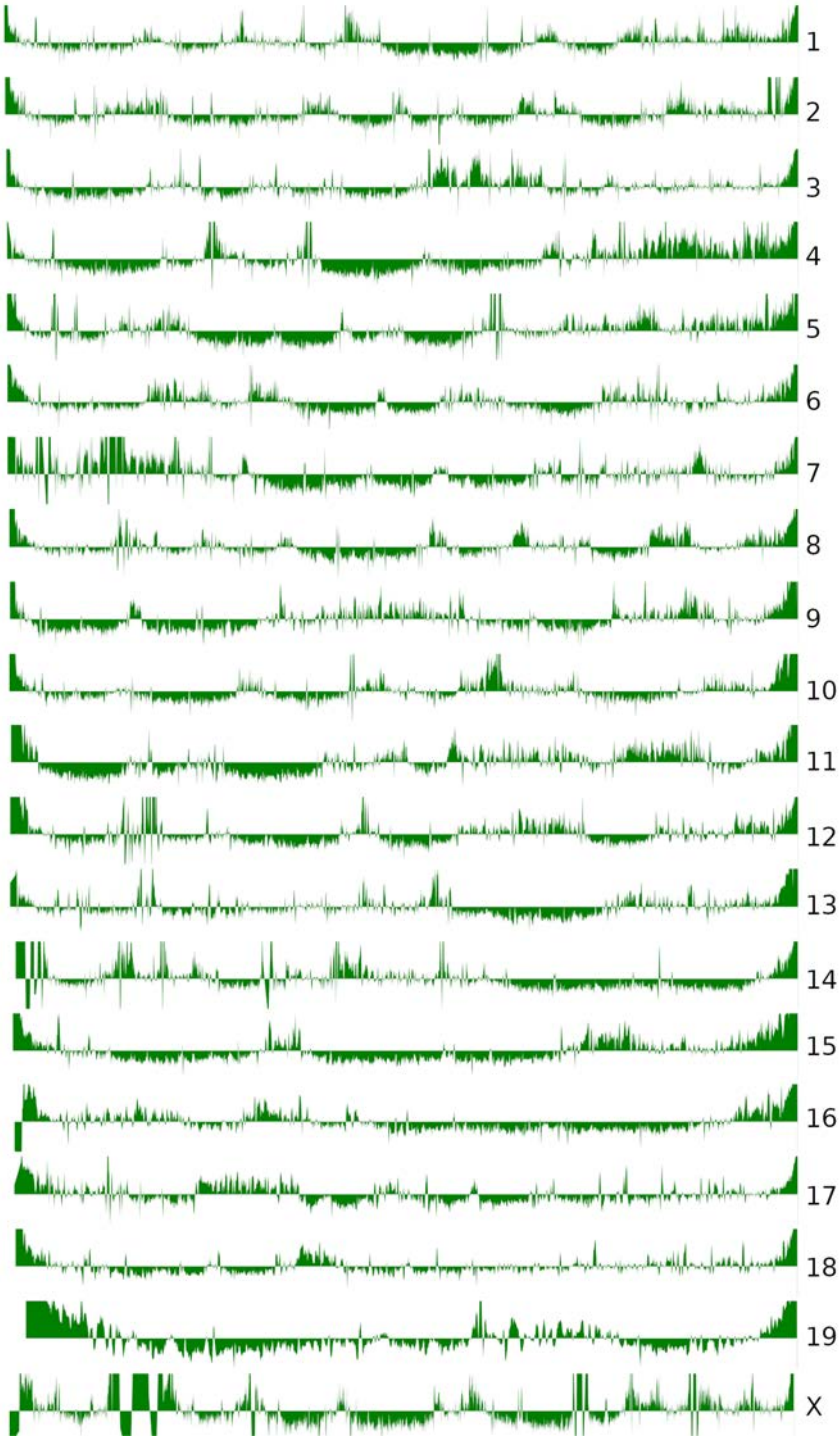
Supplementary figure 22. Per-chromosome eigenvector in sperm. Positive values represent A compartments while negative values represent B compartments.



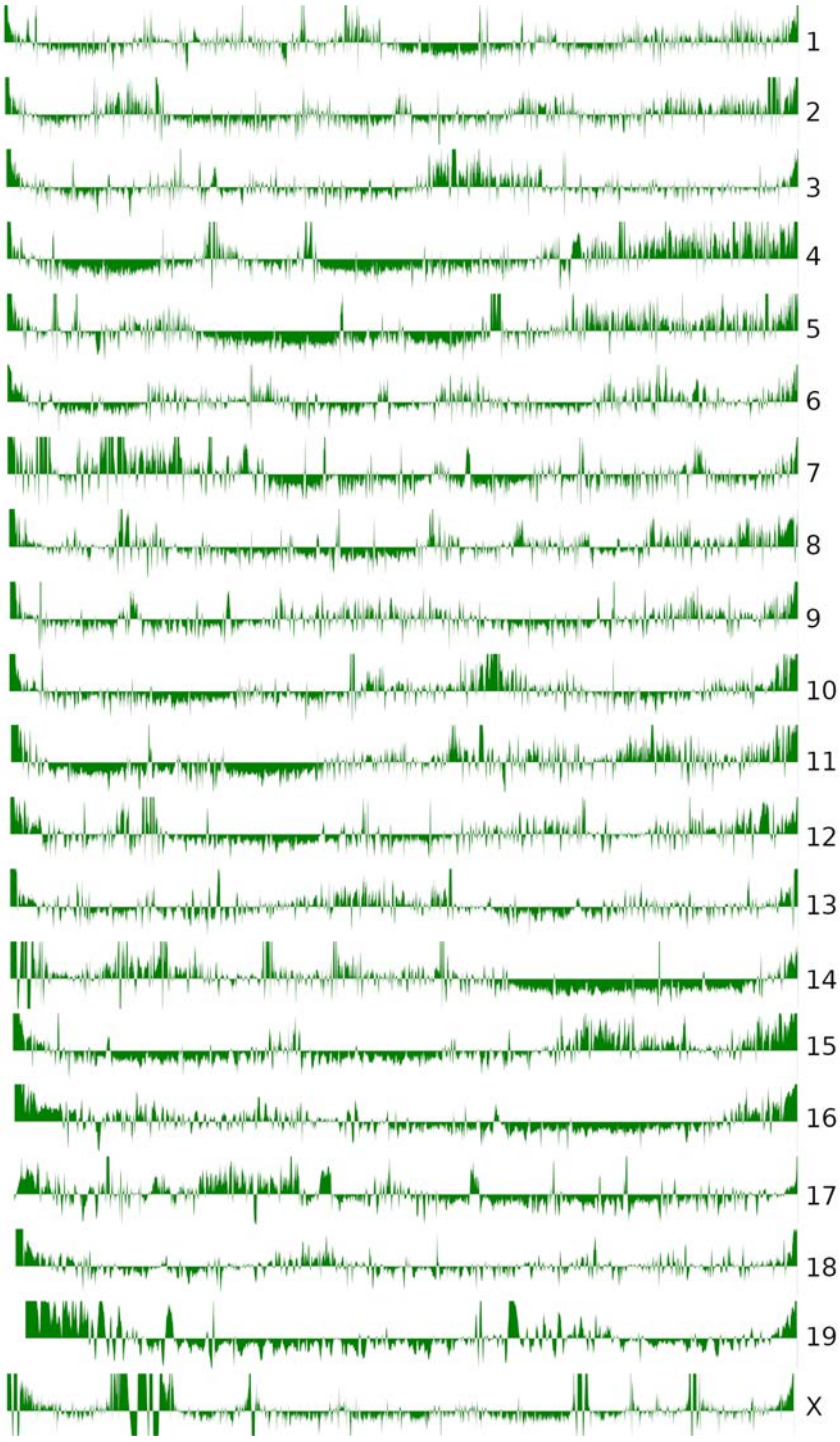
Supplementary figure 23. Per-chromosome TAD signal (insulator score) in fibroblast.



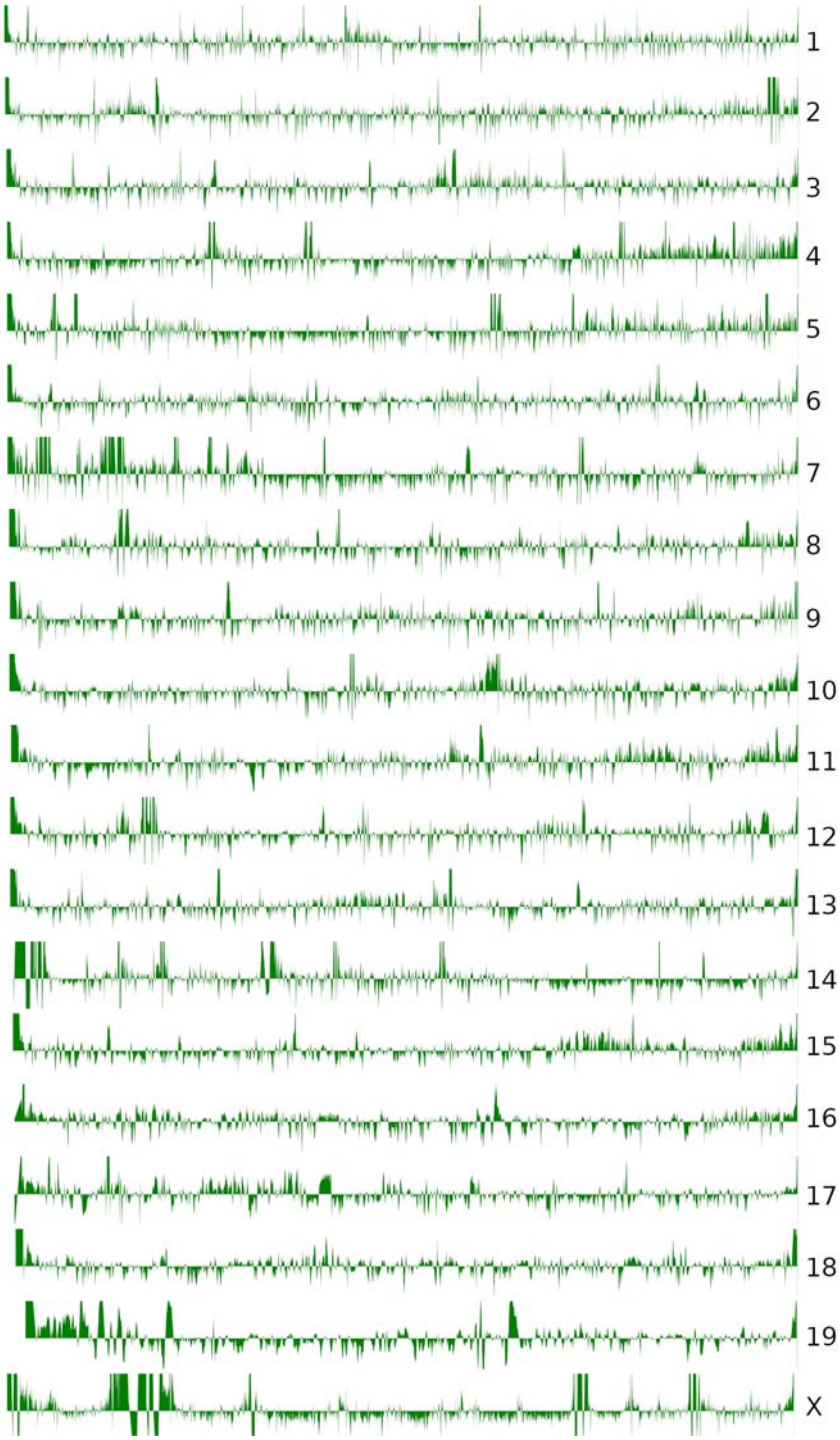
Supplementary figure 24. Per-chromosome TAD signal (insulator score) in spermatogonia.



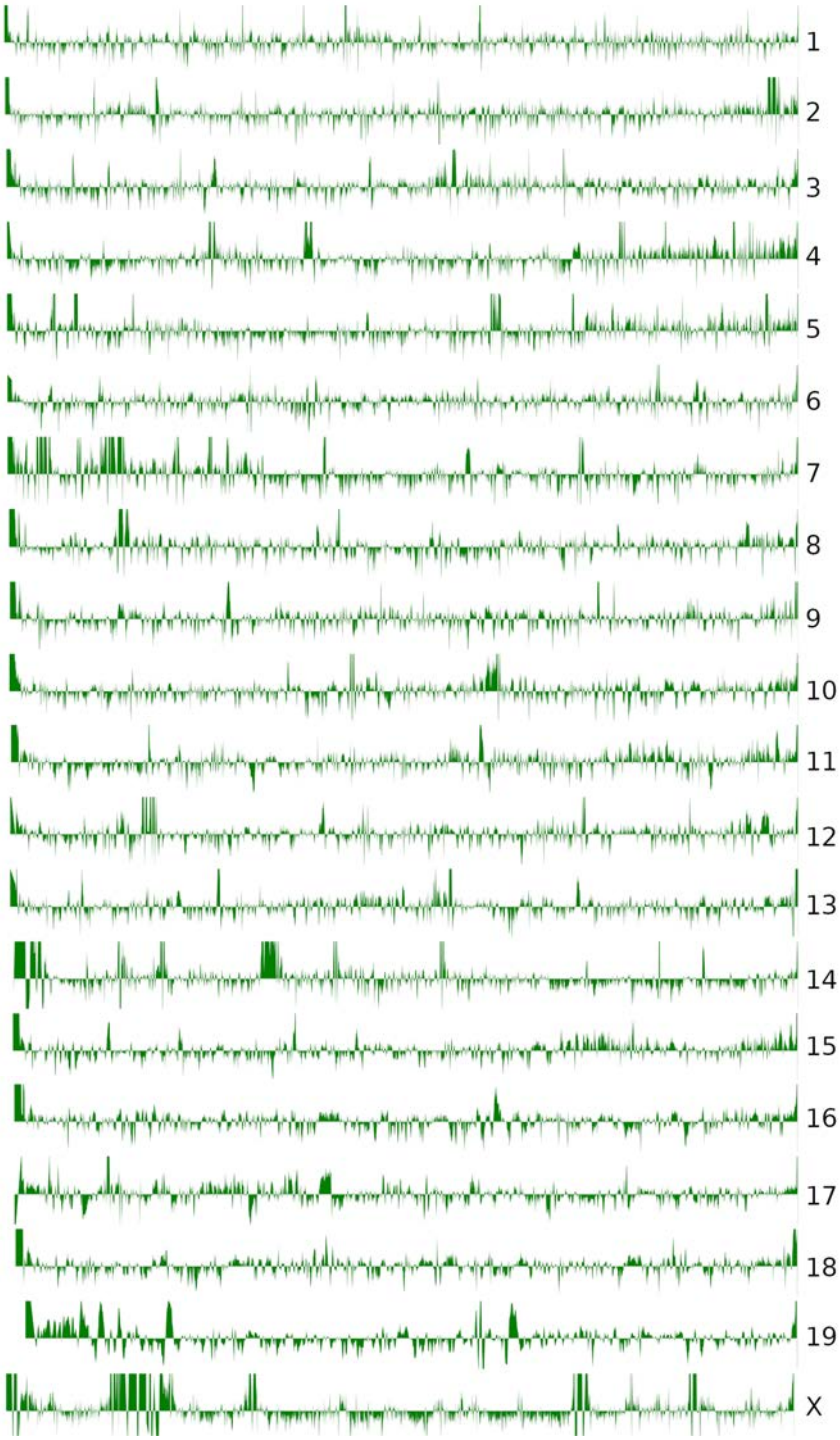
Supplementary figure 25. Per-chromosome TAD signal (insulator score) in leptonema/zygonema.



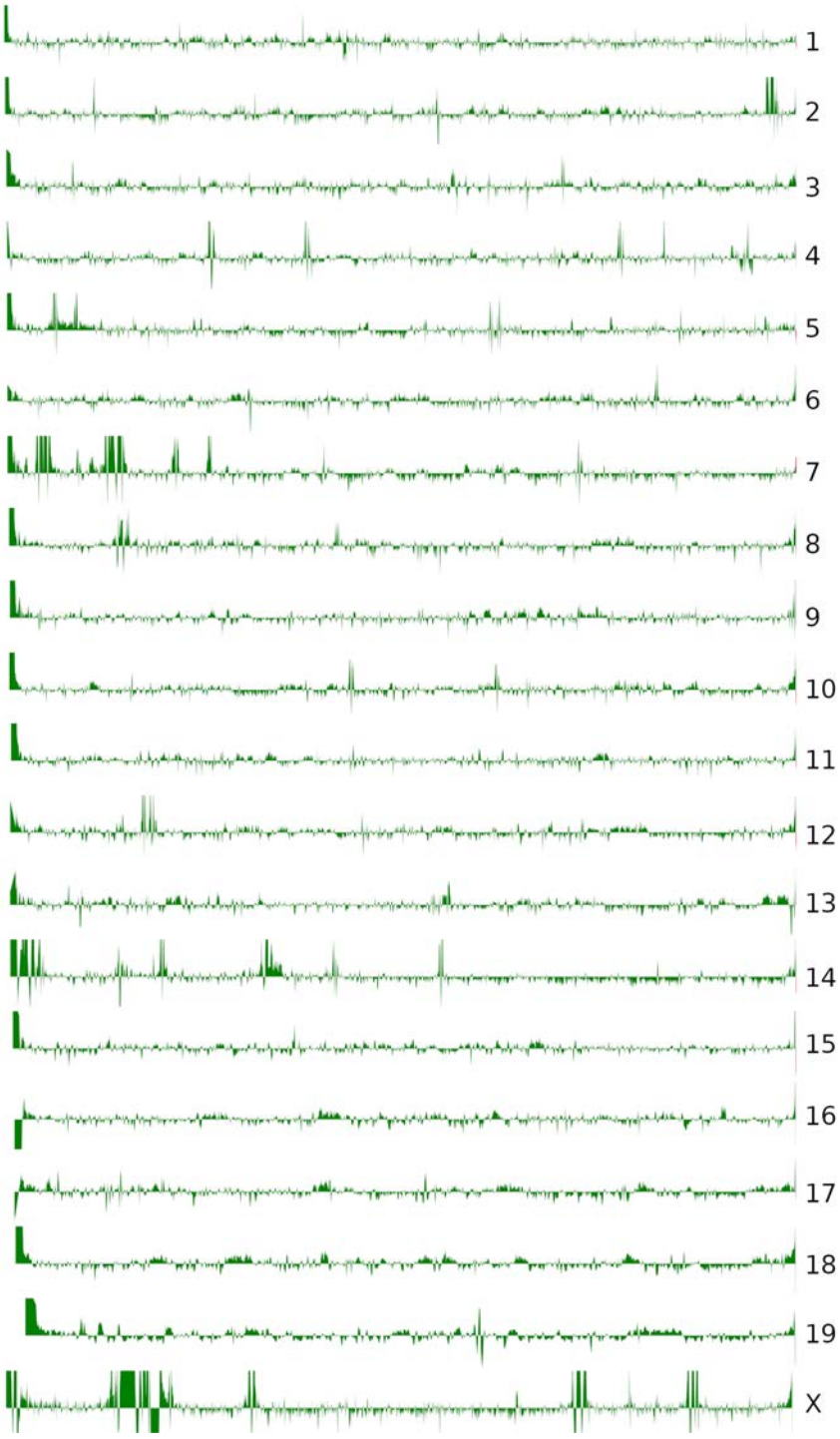
Supplementary figure 26. Per-chromosome TAD signal (insulator score) in pachynema/diplonema.



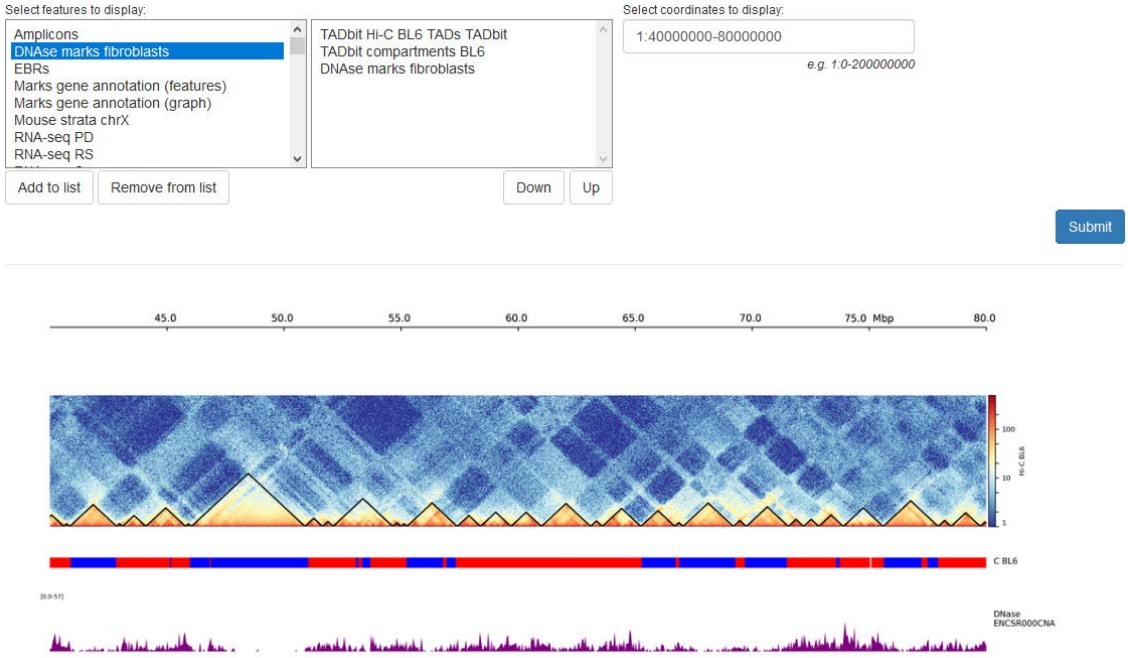
Supplementary figure 27. Per-chromosome TAD signal (insulator score) in secondary spermatocytes.



Supplementary figure 28. Per-chromosome TAD signal (insulator score) in round spermatids.



Supplementary figure 29. Per-chromosome TAD signal (insulator score) in sperm.



Supplementary figure 30. Screenshot of HiCloud. The top part of the figure shows the control panel to select the data or the coordinates to display. The bottom part of the figure shows the graphical output generated.

Supplementary tables

Supplementary table 1. Number of DEGs according to biotype. Legend: Sg: spermatogonia; P/D: pachynema/diplonema; RS: round spermatids.

Comparison	Type	DEG number	asRNA	lncRNA	Pseudogenes	Protein coding genes	smRNA	Unannotated
Sg *versus* P/D	Up-regulated	5091	121	209	151	4251	18	341
	Down-regulated	3842	253	495	273	2619	20	182
Sg *versus* RS	Up-regulated	2574	26	58	27	2368	84	11
	Down-regulated	1105	49	76	32	924	22	2
Sg *versus* sperm	Up-regulated	6956	185	392	339	5592	23	425
	Down-regulated	6124	532	986	732	3490	47	337
P/D *versus* RS	Up-regulated	350	8	25	3	304	0	10
	Down-regulated	1609	118	153	62	1185	0	91
P/D *versus* sperm	Up-regulated	2619	104	217	148	2026	4	120
	Down-regulated	3206	224	399	266	2155	7	155
RS *versus* sperm	Up-regulated	1636	59	127	134	1275	0	41
	Down-regulated	1626	115	167	122	1153	0	69

Supplementary table 2. **Percentage of DEGs according to biotype.** Legend: Sg: spermatogonia; P/D: pachynema/diplonema; RS: round spermatids.

Comparison	Type	DEG number	asRNA	lncRNA	Pseudogenes	Protein coding genes	smRNA	Unannotated
Sg *versus* P/D	Up-regulated	5091	2.38	4.11	2.97	83.5	0.35	6.7
	Down-regulated	3842	6.59	12.88	7.11	68.17	0.52	4.74
Sg *versus* RS	Up-regulated	2574	1.01	2.25	1.05	92	3.26	0.43
	Down-regulated	1105	4.43	6.88	2.9	83.62	1.99	0.18
Sg *versus* sperm	Up-regulated	6956	2.66	5.64	4.87	80.39	0.33	6.11
	Down-regulated	6124	8.69	16.1	11.95	56.99	0.77	5.5
P/D *versus* RS	Up-regulated	350	2.29	7.14	0.86	86.86	0	2.86
	Down-regulated	1609	7.33	9.51	3.85	73.65	0	5.66
P/D *versus* sperm	Up-regulated	2619	3.97	8.29	5.65	77.36	0.15	4.58
	Down-regulated	3206	6.99	12.45	8.3	67.22	0.22	4.83
RS *versus* sperm	Up-regulated	1636	3.61	7.76	8.19	77.93	0	2.51
	Down-regulated	1626	7.07	10.27	7.5	70.91	0	4.24

Supplementary table 3. **Hi-C quality metrics per replicate.** It includes the number of reads before and after quality check and trimming, the number of reads that mapped once in the genome (uniquely mapped), the percentage of reads classified as artefacts (see section 5.2.3), and the final number of valid reads. Legend: Fib: fibroblast; Sg: spermatogonia; L/Z: leptonema/zygonema; P/D: pachynema/diplonema; SpII: secondary spermatocytes; RS: round spermatids; rep1: replicate 1; rep2: replicate 2; rep3: replicate 3.

Reads information	Fib (rep1)	Fib (rep2)	Sg (rep1)	Sg (rep2)	L/Z (rep1)	P/D (rep1)	P/D (rep2)	P/D (rep 3)	SpII (rep1)	SpII (rep2)	RS (rep1)	RS (rep2)	Sperm (rep1)	Sperm (rep2)
# raw	248031330	252483078	278816401	287484075	225472936	233132942	250584996	278216841	215981536	228741418	239271647	255263907	248883003	219947094
# trimmed	240260284	245742329	267840097	274617612	216917242	216040083	243182843	265129919	209031852	214721068	223813558	248148265	240177238	211815348
# uniquely mapped	174084630	172833069	193623984	193716658	158331725	156290679	173770045	192207789	158968833	155162606	155273080	180507000	172429305	151996434
% self-circle (relative uniquely mapped)	0.13	0.12	0.17	0.17	0.11	0.25	0.24	0.12	0.15	0.24	0.27	0.34	0.09	0.08
% dangling-end (relative uniquely mapped)	0.32	3.13	0.12	0.55	0.06	5.36	4.33	0.06	1.10	3.04	12.77	18.90	0.15	0.37
% error (relative uniquely mapped)	0.19	1.53	0.04	0.03	0.03	1.70	0.59	0.03	1.18	1.28	5.59	1.52	0.03	0.03
extra dangling-end (relative uniquely mapped)	2.94	3.12	4.70	2.46	4.95	3.36	4.63	4.72	2.22	3.14	4.60	8.20	3.57	4.74
% too short (relative uniquely mapped)	4.84	6.28	7.34	6.60	6.55	5.82	5.75	5.98	3.62	6.18	6.79	6.07	5.61	5.38
% too large (relative uniquely mapped)	0.01	0,01	0.00	0.00	0.00	0.01	0.01	0.00	0.01	0.00	0.01	0.02	0.01	0.00
% duplicated (relative uniquely mapped)	3.34	6.15	7.62	9.05	3.98	18.91	4.05	3.85	11.11	14.18	22.16	9.91	3.87	7.87
% random breaks (relative uniquely mapped)	0.09	0.67	0.03	0.25	0.01	1.00	0.80	0.01	0.09	0.58	1.92	3.58	0.06	0.26
# total valid	154701718	139275365	157495526	159034223	134826221	105426311	140428838	165268550	129615416	114608474	84132360	104997044	150537722	125517920
% total valid reads	62.37	55.16	56.49	55.32	59.80	45.22	56.04	59.40	60.01	50.10	35.16	41.13	60.49	57.07

(relative to raw)														
% total valid (relative to trimmed)	64.39	56.68	58.80	57.91	62.16	48.80	57.75	62.33	62.01	53.38	37.59	42.31	62.68	59.26
% total valid (relative to uniquely mapped)	88.87	80.58	81.34	82.10	85.15	67.46	80.81	85.98	81.54	73.86	54.18	58.17	87.30	82.58

Supplementary table 4. **Compartment and TAD statistics**. It includes the mean size of compartments and TADs, the number of TADs, and the percentage of robustness of TAD boundaries. Legend: Fib: fibroblast; Sg: spermatogonia; L/Z: leptonema/zygonema; P/D: pachynema/diplonema; SpII: secondary spermatocytes; RS: round spermatids.

Cell type	Compartment mean size (bp)	TAD number	TAD mean size (bp)	% Robust TADs (score 10 and 9)
Fib	1000367	2002	1361638	72.32
Sg	755820	834	3268585	71.70
P/D	282056	294	9272109	79.59
L/Z	150888	305	8937705	74.24
SpII	760457	5004	544764	12.23
RS	869856	4649	586363	8.49
Sperm	933584	1042	2616123	14.87

Supplementary table 5. Significant enriched GO terms from the GOEA analysis. GO terms were summarized up to the level 3 (GO terms with 2 parents). This table provides the summarized GO terms and the name of the expressing genes from A-specific compartment regions for each cell type.

GO terms	Germline	Sg+P/D	P/D+RS	RS+Sperm	Sg	P/D	RS	Sperm
GO:0000075 cell cycle checkpoint	Nsun2		Mad2l1; Spdl1					
GO:0000920 cell separation after cytokinesis	Chmp2a							
GO:0001775 cell activation			Anxa3		Rap2b; P2ry12	Dock2		Lcp2; Plcz1
GO:0001816 cytokine production					Maf		Adamts3	
GO:0001909 leukocyte mediated cytotoxicity							Ctsc	
GO:0002200 somatic diversification of immune receptors	Lig4	Exo1						
GO:0002252 immune effector process			C8b			Cfh		
GO:0002253 activation of immune response						Cd36		
GO:0002532 production of molecular mediator involved in inflammatory response						Chia1		
GO:0003006 developmental process involved in reproduction	Taf4b; Dazl; Aspm	Dmrt1; Dmrt3	Dach1	Spef2; Cyp19a1	Diaph2	Fgf10		Dpy19l2; Spag16; Capza3
GO:0003008 system process	Immp2l		Ttn; Tmc1; Scn1a; Grm7	Drd2	Fli1; Gucy1a1	Ppargc1a	Chrna7; Olfr654; Olfr554; Olfr1392; Grm5; Olfr703; Trdn; Mkks; Olfr275; Gabrb3; Prkg1; Olfr697; Gabra5; Dlg2; Gja1; Olfr301; Olfr308; Olfr701; Grm1; Olfr653; Calca; Olfr1393; Olfr303; Eya4	Trpm8
GO:0006457 protein folding	Emc6						Mkks	

GO:0006807 nitrogen compound metabolic process	Immp1l; Immp2l	Kmo; Spcs3				Pcsk2	Trhde	Dmd; Lgsn
GO:0006950 response to stress	Paxip1; Rad21; Psmd14; Smarcad1; Lig4; Lig1; Immp2l	Reln; Stxbp4; Exo1; Kmo	Pnpt1; Vrk2; Dpp4; Grm7			Cd36; Oxr1	Gja1	Dmd
GO:0006955 immune response		Exo1						Sftpd
GO:0007017 microtubule-based process			Ppp2r1a	Deup1				
GO:0007049 cell cycle	Nsun2; Poc5; Bora; Rad21; Syce1; Pik3c3; Aspm; Stag3; Lig4	Exo1			Gm1993; Gm5169			
GO:0007154 cell communication			Syt10; Nrxn1; Grm7	Drd2	Grm8; Sv2c; Rasd2	Fgf10; Cnr1; Gabra1	Fgf12; Dlg2; Sv2b; Gabra5; Gja1; Olfr275	
GO:0007155 cell adhesion	Cdh13; Pcdhga12	Fat1; Reln	Cdh2; Nrxn1; Chl1; Dpp4; Cdh18; Pcdh9; Ctnna2; Cadm2; Cntnap5b; Cntn6		P2ry12; Rap2b	Plcb1; Cdh12; Cdh9; Klra8; Gm28710; Cntn3; Pcdh17; Igfbp7; Cntnap2; Flrt2; Cd36; Dsg1b; Dsg1a; Cyfip2; Itga4	Adgrl3	Pdlim1
GO:0007163 establishment or maintenance of cell polarity	Mpp5	Fat1		Lin7a		Dock2	Gja1	
GO:0007272 ensheathment of neurons	Qk			Aspa				
GO:0007275 multicellular organism development		Dmrt1; Dmrt3; Reln						
GO:0007389 pattern specification process		Reln	Ttc21b	Sox17		Fgf10; Cobl		
GO:0007566 embryo implantation							Calca; A1cf	
GO:0007568 aging			Prmt6					
GO:0007585 respiratory gaseous exchange			Dach1					Sftpd
GO:0007611 learning or memory		Reln	Grm7		Fgf13	Pak7; Cnr1; Plcb1; Adgrb3; Amph;	Grm5; Chrna7; Gabra5; Atp8a1	

						Cntnap2		
GO:0007625 grooming behavior	Ctns			Drd2				
GO:0007626 locomotory behavior		Dmrt3	Dpp4				Alk	
GO:0007631 feeding behavior			Grm7; Dach1					
GO:0007638 mechanosensory behavior			Etv1; Nrxn1	Drd2		Cntnap2; Foxp2		
GO:0008037 cell recognition	Spaca3		Cadm2	Glipr1l1; Prss37		Cntnap2		
GO:0009056 catabolic process	Pik3c3; Wdfy3; Tbc1d5			Oxct1				
GO:0009058 biosynthetic process		Kmo	Hmgcll1			Dio2		
GO:0009566 fertilization								Plcz1; Smcp
GO:0009605 response to external stimulus	Lypd8; Spaca3		Tmc1; Csmd1; Ttn	Drd2	Gucy1a1	Foxp2	Alk	Dmd
GO:0009628 response to abiotic stimulus	Lig4; Paxip1			Rp1			Gja1; Calca; Grm1	Trpm8
GO:0009653 anatomical structure morphogenesis	Mpp5	Dmrt1; Reln	Meis1; Cdh2; Ctnna2; Ttn	Rp1; Sox17; Wdpcp; Drd2; Spef2	Tenm3	Fgf10; Bmp5; Epha7; Hpgd; Aldh1a1	Zfpm2	
GO:0009719 response to endogenous stimulus	Qdpr		Pnpt1			Igfbp7; Ppargc1a	Gja1; Mkks	Pdgfd
GO:0014854 response to inactivity				Drd2				Dmd
GO:0014874 response to stimulus involved in regulation of muscle adaptation						Ppargc1a		
GO:0016049 cell growth	Mex3c; Tmem108					Cobl; Cyfip2		
GO:0016458 gene silencing	Sox6; Cnot6l; Lin28b							
GO:0019725 cellular homeostasis							Trim32	Dmd
GO:0019748 secondary metabolic process						Ddc		
GO:0019827 stem cell population maintenance	Aspm; Mcph1		Cdh2					
GO:0019835 cytolysis			C8b			Reg3g		
GO:0021700 developmental maturation			Nrxn1			Adgrb3	Gja1	Snx19
GO:0022402 cell cycle process	Chmp2a; Smarcad1;		Ppp2r1a; Pnpt1	Deup1	Klhl13			Magi2

	Aspm; Mcph1; Rad21; Haspin							
GO:0022404 molting cycle process	Nsun2							
GO:0022406 membrane docking			Nrxn1	Exoc5				
GO:0022412 cellular process involved in reproduction	Stag3		Ppp2r1a			Hyal5; Spaca6		
GO:0022602 ovulation cycle process							Chrna7	
GO:0030029 actin filament-based process					Elmo1			
GO:0030534 adult behavior			Nrxn1; Grm7	Drd2		Pcdh17; Cntnap2	Olfr275; Chrna7	
GO:0031294 lymphocyte costimulation			Dpp4					
GO:0031987 locomotion involved in locomotory behavior							Adgrl3	
GO:0032940 secretion by cell			Abca12			Fgf10		Cadps
GO:0033058 directional locomotion			Ttn					
GO:0033687 osteoblast proliferation			Fignl1					
GO:0034381 plasma lipoprotein particle clearance						Cd36		
GO:0035265 organ growth			Ttn			Fgf10		
GO:0035637 multicellular organismal signaling		Dmrt3	Grm7					
GO:0035640 exploration behavior			Chl1					
GO:0036093 germ cell proliferation		Dmrt1						
GO:0042221 response to chemical	Pik3c3; Lig4; Prkcb; Tmem108; Qdpr		Pde1c; Ttn; Abcg2; Anxa3	Cftr; Drd2; Oxct1; Grin2b; Rnls	P2ry12; Adamts7; Fgf13	Ppargc1a; Gabra1; F830016B08Rik; Cd36; Ddc; Gm4841; Ifi202b	Gabrb3; Grm5; Cyp2r1; Ryr3; Gabrb1; Chrna7	Rfc3
GO:0042330 taxis						Epha7; Flrt2		
GO:0042698 ovulation cycle							Gabrb1; Chrna7	
GO:0044085 cellular component biogenesis	Serinc1							
GO:0044236 multicellular organism metabolic process						Ppargc1a		

GO:0044238 primary metabolic process	Rpe; Lrat; Lipe			Cyp19a1; St8sia3		Aldh1a1; Chia1; Ppargc1a; Aoah	Ndst3; Cyp2r1	Pdha2; Hyal6; Hyal4; Chil5
GO:0044281 small molecule metabolic process	Ppip5k2					Ttpa	Gpd2; Cyp2r1	
GO:0044419 interspecies interaction between organisms	Chmp2a			Vta1		Reg3g		
GO:0044703 multi-organism reproductive process						Hyal5; Hpgd		Smcp
GO:0045058 T cell selection						Dock2		
GO:0048532 anatomical structure arrangement								Dmd
GO:0048589 developmental growth						Fgf10		
GO:0048609 multicellular organismal reproductive process	Immp2l		Abcg2		Diaph2	Hpgd		
GO:0048646 anatomical structure formation involved in morphogenesis		Reln		Sox17		Bmp5; Fgf10	Zfpm2; Kif16b	
GO:0048771 tissue remodeling							Gja1	
GO:0048871 multicellular organismal homeostasis	Mex3c			Drd2			Arrdc3	
GO:0050673 epithelial cell proliferation	Cdh13					Fgf10		
GO:0050900 leukocyte migration						Itga4		
GO:0051235 maintenance of location	Taf3							Pex5l
GO:0051301 cell division	Nsun2; Bora; Rad21; Syce1; Pik3c3; Aspm; Lig4							
GO:0051606 detection of stimulus			Scn1a					
GO:0051641 cellular localization	Aspm; Cdh13; Cep112	Reln	Nrxn1; Cenpq; Cdh2			Itga4	Dlg2	Magi2; Syne1; Dmd
GO:0051703 intraspecies interaction between organisms					Grid1			
GO:0060384 innervation							Gabrb3; Gabra5	
GO:0060466 activation of meiosis involved in egg activation						Plcb1		
GO:0060710 chorio-allantoic fusion						Bmp5		

GO:0061744 motor behavior			Dpp4					
GO:0065009 regulation of molecular function	Serinc1	Reln	Nrxn1; Grm7; Crbn; Pot1b		Pot1a; Fgf13	Eno1b; Itga4; Cd36		Hjurp; Sftpd; Rfc3; Mtrr; Dync1i1
GO:0071625 vocalization behavior			Nrxn1			Foxp2		

List of publications

Paytuví Gallart, A., et al. "GREENC: a Wiki-based database of plant lncRNAs" *Nucleic acids research* vol. 44, no. D1, 2015, pp: D1161-6.

Capilla, L., Sánchez-Guillén, R. A., Farré, M., Paytuví-Gallart, A., Malinverni, R., Ventura, J., Larkin, D. M., Ruiz-Herrera, A. "Mammalian Comparative Genomics Reveals Genetic and Epigenetic Features Associated with Genome Reshuffling in Rodentia" *Genome biology and evolution*, vol. 8, no. 12, 2016, pp: 3703-3717.

Ruggieri, V., et al. "Exploiting the great potential of Sequence Capture data by a new tool, SUPER-CAP" *DNA Research*, vol. 24, no. 1, 2016, pp: 81-91.

Osuna-Cruz, CM and Paytuvi-Gallart, A., et al. "PRGdb 3.0: a comprehensive platform for prediction and analysis of plant disease resistance genes" *Nucleic acids research*, vol. 46, no. D1, 2017, pp: D1197-D1201.

Paytuvi-Gallart, A., Sanseverino, W., Aiese Cigliano, R. "A walkthrough to the use of GreeNC, the plant lncRNA database". *Plant Long Non-Coding RNAs*, Julia Chekanova and Hsiao-Lin Wang (Eds). In press.

Vara, C., Paytuví-Gallart, A., et al. "Three-dimensional genomic structure and cohesin occupancy shape transcriptional activity during spermatogenesis". Manuscript submitted for publication.

Acknowledgements

First, I want to thank my supervisor Aurora Ruiz-Herrera. It has been an honour to be her PhD student. I appreciate all her contributions of time and ideas to make my PhD productive and stimulating. I would also like to thank Covadonga Vara, who has been my PhD partner from the beginning of the project generating part of the data shown in this work.

I also want to acknowledge Riccardo Aiese Cigliano, my supervisor at Sequentia Biotech, and Walter Sanseverino for their contributions and support as well as for allowing me to carry out the PhD in the framework of an Industrial Doctorate. I do not want to forget the other members of Sequentia Biotech, thank you for collaboration.

It has been a stressful time and my beloved life partner Esther Llobera knows it very well. Thank you, Esther, for your patience and support along these years.