UNIVERSITÀ DEGLI STUDI DI MILANO

DIPARTIMENTO DI SCIENZE E POLITICHE AMBIENTALI

DOTTORATO IN SCIENCE AMBIENTALI XXXI CICLO

THIRD-GENERATION SEQUENCING AND ASSEMBLY OF THE BARN SWALLOW GENOME AND A STUDY ON THE EVOLUTION OF THE HUNTINGTIN GENE

GIULIO PAOLO FORMENTI
matricola R11215

TUTOR: PROF. NICOLA SAINO
CO-TUTOR: PROF. ELENA CATTANEO

COORDINATORE DI DOTTORATO: PROF. NICOLA SAINO

A.A. 2017/2018

PREFACE **4**

SECTION A

INTRODUCTION **10**

Long Walk to Genomics: A Brief History of DNA Sequencing and Genome Assembly **11**

1. Sequencing of nucleic acids in the XX century 11
2. The age of the Human Genome Project 15
3. NextGen: DNA sequencing in the third millennium 19
4. Genome sequencing in the Third Generation Sequencing era 23

The Barn Swallow **29**

1. General description 29
2. Genetic studies in the barn swallow 33
3. A genome for the barn swallow 34

METHODS **37**

RESULTS **45**

DISCUSSION **55**

Appendix 1 62
Appendix 2 63
References 80

SECTION B

INTRODUCTION **101**

1. Huntington's Disease 102
2. Huntington's Disease gene evolution 106

AIMS ..... Errore. Il segnalibro non è definito.

RESULTS **114**

1. The kick-off: database search 115
2. Sampling 121
3. Development of protocols for Htt exon 1 amplification and sequencing 126
4. Sequencing results 137
5. A focus on primates 142

Appendix 1 165
Appendix 2 169
Appendix 3 172
Appendix 4 174
References 177

PREFACE

The present section summarizes my achievements during the three years of graduate studies in the framework of the Ph.D. programme in Environmental Sciences at the University of Milan.

I devoted the first two years to a project aimed at studying the evolutionary origins of Huntington's Disease-causing mutation. This project was part of an on-going effort from the Laboratory of Stem Cell Biology and Pharmacology of Neurodegenerative Diseases directed by Prof. Elena Cattaneo at the University of Milan. The project and its results are outlined in the section B of this document. A manuscript is in preparation reporting part of the data from this work together with other data from the laboratory.

I devoted my last year of gratudate studies to a new project related to the barn swallow, the organism of choice of my Ph.D. thesis supervisor Prof. Nicola Saino. The aim of the project was to produce a state-of-the-art genome assembly for the European barn swallow (*Hirundo rustica rustica*). This work is outlined in part A of this document. The sequencing data and assembly results are reported in a publication entitled *"SMRT long-read sequencing and Direct Label and Stain optical maps allow the generation of a high-quality genome assembly for the European barn swallow (*Hirundo rustica rustica*)"* that has already undergone the first round of review (minor revisions requested) in the peer-reviewed journal Gigascience (IF 7.5, 2016). The manuscript is available as preprint on BioRxiv (https://doi.org/10.1101/374512) and as **Appendix 2** of section A. Moreover, this work was presented at an international meeting on Third-Generation Sequencing and Genomics that my supervisor and I have organized at the University of Milan. Invited speakers included Deborah Moine (Senior Scientist, Pacific Biosciences), Sandra Bauer (Senior Scientist, Bionano Genomics) and Erich Jarvis (G10K Coordinator, The Rockefeller University).

During my Ph.D. I also contributed to conception and writing of the following proposals:

	Amount	**Status**
Templeton Foundation, Prof. Nicola Saino et al. (2018)	$ 220,000	Under review
ERC Advanced, Prof. Nicola Saino (2018)	€ 2.9 M	Under review
PRIN, Prof. Nicola Saino et al. (2018)	€ 800,000	Under review
PRIN, Prof. Antonio Torroni, Prof. Luca Gianfranceschi (2018)	€ 744,000	Under review
ERC Advanced, Prof. Elena Cattaneo (2016)	€ 2 M	Awarded

Over the three years, and particularly in the last year, I have been able to attend several international scientific conferences and meetings. Specifically, in November 2017 I attended the PacBio User Group Meeting in Barcellona; in February 2018 I attended the PacBio Day at the Functional Genomics Center of the ETH of Zurich; in April 2018 I attended the 10X Genomics UGM in Uppsala; in June 2018 I presented in a talk our preliminary results at the Bionano UGM in Evry

(Paris) [1]; the same month I also presented a poster at the SMRT Conference in Leiden [2]; finally, in September 2018 I presented a poster outlining the final results of the barn swallow genome assembly during the G10K meeting at the Rockefeller University in New York [3]. I have also been invited to present our results in a talk at the next PacBio UGM that will take place in November 2018 in Lisbon.

Throughout the Ph.D., I have contributed to scientific dissemination. In particular, I have collaborated to the weekly newspaper "Pagina99" with several feature articles on stem cell research, genome sequencing and genome editing.

As Ph.D. students representative in the Academic Senate (elected in January 2016) I have carried out several initiatives aimed at improving the quality of Ph.D. courses. In 2015, I publicly advocated for a Ph.D. salary increase at the University of Milan. The initiative led to a 20% salary increase for all Ph.D. students at the University of Milan in 2016 (~3.5 millions euro investment in three years). In 2016, I advocated for the creation of a Ph.D. Student Council at the University of Milan. In 2017, the Academic senate instituted it and in 2018 it was institutionalized in the University Statute. In 2017, together with colleagues from other universities, I publicly advocated for a Ph.D. salary increase at the national level. The initiative led to a 12,5% salary increase for all Italian Ph.D. students in 2018 (60 millions euro investment in three years). Between 2017-2018, I proposed the creation of an Italian Ph.D. students association. This is now the second largest association of its kind in Italy and has already promoted several initiatives since its foundation, including the establishment of a Ph.D. fellowship entitled to Giulio Regeni in several Italian universities. The association is currently promoting new national regulations for Ph.D. programs and postdoctoral opportunities.

References

[1] Formenti G., Bonisoli-Alquati A., Chiara M., Gianfranceschi L., Horner D., Poveda L. and Saino N. Bionano Direct Label and Stain optical mapping for reference-level scaffolding of SMRT long read-based contigs in the barn swallow (Hirundo rustica). Oral presentation given at the Bionano UGM, Genoscope, Evry, France, June 2018.

[2] Formenti G.*, Chiara M.*, Poveda L., Bonisoli-Alquati A., Gianfranceschi L., Horner D.S. and Saino N. A novel barn swallow draft genome assembly based on SMRT long-reads and DLS optical mapping. Poster presented at the SMRTLeiden Meeting, Leiden University Medical Center, Leiden, June 2018.

[3] Formenti G.*, Chiara M.*, Poveda L., Francoijs KJ., Bonisoli-Alquati A., Canova L., Gianfranceschi L., Horner D., Saino N. SMRT long-read sequencing and Direct Label and Stain optical maps allow the generation of a high-quality genome assembly for the European barn swallow (*Hirundo rustica rustica*). Poster presented at the G10K meeting, Rockefeller University, New York, September 2018.

«A knowledge of sequences could contribute much
to our understanding of living matter.»
Frederick Sanger, 1980

SECTION A

Third-Generation Sequencing and Assembly of a High Quality Genome for the European Barn Swallow (*Hirundo rustica rustica*)

The present section of this Ph.D. thesis outlines the scientific work that I have accomplished during the last year of my graduate studies. The goal was to generate a reference genome for the European barn swallow (*Hirundo rustica rustica*) using state-of-the-art sequencing and computational approaches, including SMRT long-read sequencing by Pacific Biosciences and DLS optical mapping by Bionano Genomics. The barn swallow is a bird species subjected to hundreds of ecological studies in the past and the organism of choice of my Ph.D. thesis supervisor, Prof. Nicola Saino. This scientific endeavour culminated in a publication that I authored entitled *"SMRT long-read sequencing and Direct Label and Stain optical maps allow the generation of a high-quality genome assembly for the European barn swallow (*Hirundo rustica rustica*)"* that has already undergone the first round of review (minor revisions requested) in the peer-reviewed journal Gigascience (IF 7.5, 2016). The manuscript is available as preprint on BioRxiv (https://doi.org/10.1101/374512) and as **Appendix 2** of this section.

ABSTRACT

The barn swallow (*Hirundo rustica*) is a migratory bird that has been the focus of a large number of ecological, behavioural and genetic studies. To facilitate further population genetics and genomic studies, I have generated a high-quality genome for the European subspecies (*Hirundo rustica rustica*). In particular, I have assembled a highly contiguous genome sequence using third-generation Single Molecule Real-Time (SMRT) DNA sequencing from Pacific Biosciences (Menlo Park, California, USA) and optical mapping from Bionano Genomics (San Diego, California, USA).

For optical mapping, DNA molecules were labelled both with one of the original Nick, Label, Repair and Stain (NLRS) nickases (enzyme Nb.BssSI) and with the new Direct Label and Stain (DLS) approach (enzyme DLE-1). This allowed to compare and integrate optical maps derived both from NLRS and DLS technologies. The latter was officially released in February 2018 and avoids nicking and subsequent cleavage of DNA molecules upon staining. To my knowledge, this has been the first genome assembly to incorporate DLS data and this approach has more than doubled the assembly N50 with respect to the nickase system. Furthermore, the dual enzyme hybrid scaffold led to a marginal increase in scaffold N50 and an overall increase of confidence in scaffolds.

After removal of haplotigs, the final assembly is approximately 1.21 Gbp in size, with a N50 value of over 25.95 Mbp. The high genome contiguity achieved represents an improvement over 650 fold with respect to a previously reported assembly based on paired-end short read data, and it is well in excess of those specified for “Platinum genomes” by the Vertebrate Genomes Project. It can therefore constitute a valuable resource for studies concerning the evolution of avian genomes in general as well as for population genetics and genomics in the barn swallow, with the potential for boosting research on the barn swallow biology and ecology at unprecedented speed.

INTRODUCTION

Long Walk to Genomics:
A Brief History of DNA Sequencing and Genome Assembly

«Sequencing DNA now is one of the easiest jobs you could have besides sloppin' burgers»
Kary Mullis, 1998

In 1866, Gregor Mendel outlined is theory of inheritance, where heredity is determined by discrete "factors". These factors are present in couples: one from the father and one from the mother. His work laid the foundations of genetics, which in turn gave birth to one of the most memorable and inspiring scientific quests: nucleic acid sequencing. The following sections outline a chronicle of this ongoing quest.

1. Sequencing of nucleic acids in the XX century

1.1. The mistery of genes and the discovery of DNA structure

In 1869, while working at the University of Tübingen in Germany just a few years after Mendel experiments, the swiss physician and biologist Friedrich Miescher noted something that *"cannot belong among any of the protein substances known hitherto"* (Portugal and Cohen 1977). He had, for the first time in history, identified nucleic acids[1]. Between 1885 and 1901, Albrecht Kossel was able to isolate and name the five constituent organic components of nucleic acids: adenine, cytosine, guanine, thymine, and uracil[2]. In 1904, Walter Sutton and Theodor Boveri independently found that, as predicted by Mendel's theory of inheritance, chromosomes occur in matched pairs, one inherited from the mother and one from the father, and therefore proposed chromosomes as the substrates of heredity. This theory was further reinforced by the accurate work of Thomas Morgan at the Columbia University on *Drosophila melanogaster* (Morgan 1911). In 1913, Morgan's observations allowed one of his students, Alfred Sturtevant, to construct the first genetic map of a chromosome (Sturtevant 1913). However, chromosomes are organized in chromatin, which is made up of both nucleic acids and proteins, and in the first half of the XX century it was commonly believed that proteins held the genetic blueprint of inheritance, due to their higher degree of complex behaviours.

In 1941, George Beadle and Edward Tatum at Stanford University conducted a series of experiments using X rays on the fungus *Neurospora*. These experiments revealed that every enzyme required in a metabolic pathway is generally produced by a single gene, and each gene can be inactivated by exposure to X rays. In agreement with Mendel's theory of inheritance, this result suggested that genes are indeed discrete

[1] Nucleic acids are biopolymers composed of nucleotides. Nucleotides are made of a nitrogenous base, a five-carbon sugar (ribose or deoxyribose), and at least one phosphate group. In principle, the sequential addition of nucleotides to a DNA chain is limitless, making nucleic acids the largest known biomolecules.
[2] "Encyclopaedia Britannica - Alfred Kossel". https://www.britannica.com/biography/Albrecht-Kossel

units. In 1944, Oswald Avery, Colin MacLeod and Maclyn McCarty showed that purified DNA can change one bacterial strain into another, transforming the properties of a living cell. This result was strongly suggestive of a key role of DNA as substrate of genetic information.

In 1951, Erwin Chargaff realized that in the DNA the sum of Adenines and Guanines was equal to that of Cytosines and Thymines, a detail revealing of DNA structure (Chargaff et al. 1951; Chargaff, Lipshitz, and Green 1952). Two years later, Chargaff's observations along with the famous X ray picture 51 from Rosalind Franklin and Maurice Wilkins, were used by Francis Crick and James Watson to "discover the secret of life", i.e. the DNA structure (Watson and Crick 1953). In Crick and Watson's model of DNA molecules, double-stranded (ds) right-handed helix strands have antiparallel complementary base sequences, which readily explains Chargaff's observation of equimolarity of complementary bases. In their 1953 article, Crick and Watson commented *"so far as is known the sequence of bases along the chain is irregular"* and *"the sequence of bases on a single chain does not appear restricted in any way"*, two features entailing a role in the storage of genetic information.

1.2. Early sequencing methods and achievements

Year 1953 also marked the first "sequencing" of a biological molecule. While working at the University of Cambridge, Frederick Sanger was able to sequence the two chains of insulin protein by a refined partition chromatography method[3] (Sanger and Thompson 1953a, Sanger and Thompson 1953b). In this approach, the two chains are separated and fragmented, the fragments are individually read and sequences from each fragment overlapped to yield a complete sequence. Challenging the main view of proteins as amorphous biological molecules, with this work Sanger was able to definitively show that proteins are ordered chains amino acid residues.

Proteins were sequenced before nucleic acids as enzymes to fragment DNA had not been developed yet, and also because they were more abundant and stable than RNA. However, in many ways the 1950s paved the way to modern DNA sequencing. In particular, two major milestones of the decade were the production of isotopes and of radiolabeled biological molecules for staining and visualization.

When in 1965 it came to nucleic acids, it was first the turn of tRNA from *Saccharomyces cerevisiae*, as means to cleave RNA fragments were available since 1940s (Holley et al. 1965). As for the insulin protein sequence, RNA was first fragmented with bovine pancreatic ribonuclease (RNase A), the DNA fragments were separated by chromatography and their partial digestion with snake venom phosphodiesterase provided a mixture of degradation products from which the sequence could be deduced. Sequencing 76 nucleotides required five people working three years with one gram of pure material isolated from 140 kg of yeast (Shendure et al. 2017).

The first DNA to be sequenced was the cos-site of phage Lambda DNA in 1968 (Wu and Kaiser 1968). In the capsid, Lambda phage genome is made of linear dsDNA, with cohesive (or sticky) single-stranded (ss)

[3] Three years earlier, Pehr Edman had already published a paper demonstrating a label-cleavage method for protein sequencing (Edman 1950).

complementary extremities that are ligated when the genome is circularized in the cytoplasm. The ss nature of cohesive ends in the linearized form allowed Wu and Kaiser to determine their 12 base sequence through cycle extension of the 3' ends by polymerase-catalysed addition of nucleotides. While DNA sequencing approaches were still extremely laborious, RNA sequencing was developing relatively fast. The same year, a team including Sanger was able to determine the 120 bp-long sequence of 5s ribosomal RNA using 32P-labelled RNA and paper fractionation-based approach (Brownlee, Sanger, and Barrell 1968). In 1972, Walter Fiers from the Laboratory of Molecular Biology of the University of Ghent in Belgium sequenced the 510 bp of coat protein gene from Bacteriophage MS2, an RNA virus which infects the *Escherichia coli* bacterium (Min Jou et al. 1972). This was the first sequenced gene. In 1973, Walter Gilbert and Allan Maxam were able to report 24 bases of the *E. coli* lactose-repressor binding site by copying its DNA into RNA, at the pace of one base per month (Gilbert and Maxam 1973).

In 1975, whilst at the Laboratory of Molecular Biology in Cambridge, Frederick Sanger developed the "plus and minus" method for DNA sequencing and applied it to determine two short regions in bacteriophage φX174 single-stranded DNA (Sanger and Coulson 1975). In this method, a primer is extended by a polymerase to generate a population of newly synthesized DNA strands of different lengths. Polymerization continues afterwards in four pairs of "plus" and "minus" reaction mixtures: the minus mixtures have three NTPs and the plus mixtures have only one. The positions at which polymerization had terminated because of the absence of correct dNTPs in the minus mixtures allow to determine nucleotide composition, except for homopolymers (Wu 1994). Sanger demonstrated the strength of his new method by determining all 5,368 bp of the Bacteriophage PhiX174 genome (Sanger et al. 1977).

In the early 1970s, RNA sequencing was still ahead of DNA sequencing. Indeed, the first organism to have its genome completely sequenced was Bacteriophage MS2, with a 3,569 bp RNA genome (Fiers et al. 1976). However, in 1977 the development of two methods that could decode hundreds of bases in a day transformed the field (Shendure et al. 2017). Both methods were developed by the two pioneers in the field of DNA sequencing, Frederick Sanger and Walter Gilbert. The Sanger method, also known as dideoxy sequencing, relies on four separate polymerization reactions performed using labelled primers, where each reaction is supplied with small amounts of one chain-terminating nucleotide to produce fragments of different lengths (Sanger, Nicklen, and Coulson 1977). When the DNA polymerase incorporates a 2,3-dideoxynucleoside triphosphate (ddNTP) at the 3-end of the growing DNA strand, the newly synthesized strand the lacks a 3-hydroxyl group and chain elongation is terminated (J. Adams 2008).

In contrast with the polymerization-based approach developed by Sanger, Gilbert's method is purely chemical. It involves the use of four sets of terminally labelled deoxyoligonucleotides that are randomly cleaved at base-specific sites along the molecule by chemical compounds (Maxam and Gilbert 1977). Specifically, this approach takes advantage of the ability of dimethyl sulfate (DMS), formic acid and hydrazine to specifically modify bases within the DNA molecule. DMS methylases nitrogen 7 of G, which consequently opens between carbon 8 and nitrogen 9; formic acid weakens A and G glycosidic bonds by protonation of purine-ring nitrogens; and hydrazine splits T and C rings, but in the presence of NaCl is

selective for C rings. In all four reactions, Piperidine addiction displaces the modified nucleotides and catalyzes phosphodiester bond cleavage[4].

In both Sanger's and Gilbert's methods, DNA fragments were separated using the recently developed polyacrylamide gel electrophoresis (PAGE) (Maniatis, Jeffrey, and van deSande 1975). Gel slabs were subsequently exposed to X-rays to produce carbon-copy images of the radioactive labels where distances along DNA molecules could be used to determine the nucleotide order. Sanger's and Gilbert's radioactive methods for DNA sequencing allowed to read up to 400 bases in length.

1.3. Throughput and automation in DNA sequencing

To speed up the sequencing process, in 1979 Rodger Staden proposed the idea of "shotgun sequencing", where random bacterial vectors are sequenced in parallel and sequencing reads are assembled using the overlaps between sequences (Staden 1979). Two years later, a seminal paper by Smith and Waterman determined the rules to achieve the highest pairwise homology in a pool of sequences (T. F. Smith and Waterman 1981), laying the foundations of bioinformatics. In 1981, Joachim Messing developed the first shotgun sequencing method based on the single-stranded M13 phage vector (Messing, Crea, and Seeburg 1981). Only one year after, Sanger used the shotgun sequencing approach to assemble the entire 48,502 bp of bacteriophage Lambda genome (Sanger et al. 1982); and two years later, 172,282 bp representing the complete sequence of the Epstein–Barr virus B95-8 strain were determined using the dideoxynucleotide/M13 shotgun sequencing approach (Baer et al. 1984).

One issue at stake was the production of the required amount of starting material (F. K. Nelson et al. 2011). With respect to this, a progressively crucial role was played by the increasing availability of recombinant DNA technologies (Jackson and Symons 1972; Cohen et al. 1973) and cloning vectors throughout the 1980s (Slatko et al. 1993). After 1983, the development of Polymerase Chain Reaction by Kary Mullis revolutionized the field (Mullis 1990)[5].

In 1984, the Medical Research Council in the United Kingdom launched its official programme for the sequencing of full genomes. At the time, DNA sequencing was still performed with the time-consuming original approach developed by Sanger, which involved four sequencing reactions in four separate tubes using tritium-radiolabeled primers. In 1985, Lloyd Smith and Lee Hood were able to synthesize several fluorescent DNA primers (L. M. Smith et al. 1985). Next year, they set up a method for the partial

[4] Interestingly, while more laborious, Gilbert's method is the only direct method for DNA sequencing: it avoids issues related to polymerase synthesis of DNA associated with Sanger sequencing (i.e., premature termination due to hard DNA structures) and it can be employed in the absence of any prior sequence information for primer hybridization.

[5] Another powerful approach to the issue of template availability was PCR-free multiple displacement amplification (MDA) of the plasmid or DNA sample (J. R. Nelson et al. 2002). MDA involves a first step of random primers annealing followed by polymerase-mediated chain elongation at a constant temperature. Bacteriophage Φ29 DNA polymerase is usually employed, which can produce DNA amplicons greater than 70 kilobase pairs and has a very high fidelity and 3'–5' proofreading activity (Blanco et al. 1989). During extension, when the polymerase encounters another copying starting site, it displaces the DNA strand and continues the strand elongation. The strand displacement generates newly synthesized ssDNA template for more primers to anneal. Even when starting from tiny amounts of raw material (e.g. a single cell), the overall result of this process is a relatively high amount of "hyper-branched" DNA, with genome coverage up to 99% (Paez et al. 2004).

automation of DNA sequence analysis in Sanger sequencing (L. M. Smith et al. 1986). In this method, four dyes distinguished by their fluorescent emission *spectra* are covalently attached to the oligonucleotide primers and employed in distinct dideoxy extension reactions (dye primer sequencing). The products are then run together on a polyacrylamide gel and Fluorescent Energy Resonance Transfer (FRET) is used to read the sequence.

In 1986 Applied Biosystems, a company founded in 1981 that had so far focussed mostly on protein sequencing, switched to nucleic acids and was able release the first commercially available four-color fluorescence automated DNA sequencer (370A) based on Smith and Hood method. This machine was able to handle 32 samples per run. One year later, Prober and co-workers described a novel set of four chain-terminating dideoxynucleotides, each carrying a succinyl fluorescein dye with different emission spectra, thereby allowing reactions to be performed in a single tube (dye terminator sequencing) (Prober et al. 1987). A scanning system allowed multiple samples to be run simultaneously, with automatic computer-based base calling. The new approach, coupled with the commercialization of the technology, quadrupled the throughput compared to earlier methods. The same year, the sequencing approach was increasingly refined introducing an optimized T7 DNA polymerase called Sequenase. This enzyme was highly processive, lacked proofreading 3' to 5' exonuclease activity and efficiently used nucleotide analogs, allowing to generate around 1,000 bases per day (Tabor and Richardson 1987). Two years later, the introduction of the thermostable Taq DNA polymerase and of cycle sequencing, whereby the sequencing reaction is performed at 72 °C repeated multiple times in the same tube (linear amplification), greatly reduced template requirements and facilitated miniaturization (V. Murray 1989; Craxton 1991). By the end of the 1980s, even a completely new method for DNA sequencing by stepwise dNTP incorporation was developed (Hyman 1988). Subsequently refined (Nyrén, Pettersson, and Uhlén 1993; Ronaghi et al. 1996) and now known as "pyrosequencing", the method relies on measuring enzymatic luminometric inorganic pyrophosphate detection generated by pyrophosphate release during DNA polymerization (Nyrén 1987). This method has several advantages over Sanger's approach, including the use of natural nucleotides and the possibility of observing nucleotide synthesis in real time (Heather and Chain 2016).

"Sequences, Sequences, and Sequences" was Sanger to entitle a historical note on DNA sequencing in 1988, suggestive of the general hype (Sanger 1988). Many sequencing projects were launched and succeeded, as testified by the progressive availability of sequence data. Online repositories were created to archive those sequences. Genbank was founded in 1982 with about half a million bases, but already by the end of the decade contained over 40 million bases[6]. This impressive growth rate has never stopped since then, with almost 10-fold increases every 5 years.

2. The age of the Human Genome Project

The technological advancements of the 1980s were further reinforced and extended in the 1990s. Major milestones included the introduction of fluorescent boron-dipyrromethene dyes (bodipy) instead of labelled

[6] GenBank and WGS Statistics. https://www.ncbi.nlm.nih.gov/genbank/statistics/

primers and terminators (Lee et al. 1992); magnetic bead-based DNA purification methods that simplified the automation of pre-sequencing steps (DeAngelis, Wang, and Hawkins 1995); and capillary electrophoresis, which eliminated the pouring and loading of gels, while also simplifying the identification and interpretation of the fluorescent signal (J. Zhang et al. 1995). These breakthroughs combined with the adoption of industrial processes to maximize efficiencies and minimize errors (Shendure et al. 2017), allowed a 10-fold increase in speed over traditional slab gel technology, and by 1998 routine DNA sequencing of 1,000 bases was achieved in less than one hour (Salas-Solano et al. 1998). Moreover, as early as the 1990 the principles of "paired-end" sequencing were outlined, allowing to sequence dsDNA (Edwards et al. 1990). In paired-end sequencing both ends of a DNA fragment of known length are sequenced and linked together, increasing the accuracy of base calling. This also allows to link together sequence reads even when they are not complementary (i.e. if they have not been sequenced from end to end), a feature that would have turned extremely useful in the assembly of genomes derived from shotgun sequencing.

The advancements in sequencing technologies and the exponential increase in sequence data were paralleled by advancements in bioinformatics tools. In 1990, the National Center for Biotechnology Information (NCBI) released its Basic Local Alignment Search Tool (BLAST) (Altschul et al. 1990), which considerably sped up the process of sequence alignment. As the complexity of genomes was progressively understood, more refined tools were made available, including RepeatMasker in 1996, to deal with repeated genomic regions (Smit 1993), and GENSCAN in 1997, to predict gene structures (Burge and Karlin 1997). In 1998, Phil Green at the University of Washington developed phred, a fundamental bioinformatic algorithm, along with two associated softwares for phred output analysis, phrap and consed (Gordon, Abajian, and Green 1998). Phred introduced reliable quality metrics for base calling, providing support for automated read analyses, particularly in repeated sequences[7] (Ewing et al. 1998; Ewing and Green 1998).

2.1. Genomes, genomes, genomes

The 1990 probably marked the watershed of the genomic era, as it saw the initiation of a variety of extremely ambitious sequencing project, the foremost being the Human Genome Project (HGP)[8] (Shendure et al. 2017). The HGP was aimed to produce genetic maps, physical maps, and finally the complete nucleotide sequence map of human chromosomes. In 1987, the Department of Energy (DOE) of the United States had already established an early genome project to produce data on the mutagenic effects of radiation. In 1988, the DOE and the National Institute of Health (NIH) received funding from the Congress and formalized an agreement to "coordinate research and technical activities related to the human genome." DNA structure discoverer James Watson was appointed to lead the NIH component, the Office of Human Genome Research[9]. In 1990, the official kickoff was the publication by the DOE and the NIH of a joint research plan for the next five

[7] Phred is still widely used today. It assigns a Quality Value score to each base called using the formula $QV = -10*\log_{10}(P_e)$ where P_e is the probability that the base call is an error. For instance, a QV20 implies 99% accuracy while QV30 implies 99.9% accuracy.

[8] "News About the Human Genome Project". https://www.genome.gov/12011251/news-about-the-human-genome-project/

[9] This was renamed the National Center for Human Genome Research (NCHGR) the following year.

years. The deadline for project completion was set to the end of September 2005, that is about fifty years after Watson and Crick article on the structure of DNA. However, Watson resigned already by 1992 due to the persistent confrontation with NIH Director Bernadine Healy (L. Roberts 1992), and the following year Francis Collins was appointed in his place. In 1993, Collins traced a new five-year plan for the HGP (Collins and Galas 1993) and already by 1994 the HGP team published the first detailed linkage map of the human genome (J. C. Murray et al. 1994).

In the meantime, other projects were devoted to sequence the genomes of smaller model organisms, mostly viruses and bacteria. As early as 1990, the 192 kbp sequence of Vaccinia genome was published (Goebel et al. 1990). The 229 kbp DNA of the human Cytomegalovirus genome (Bankier et al. 1991), and the 186 kbp genome of smallpox (Massung et al. 1994) followed soon after.

Escherichia coli was considered the most promising candidate for the first bacterial genome to be sequenced, but in 1995 this record was achieved by *Haemophilus influenzae*[10]. A team headed by Craig Venter, who had founded The Institute for Genomic Research (TIGR) three years earlier, and Nobel laureate Hamilton Smith from Johns Hopkins University, sequenced the 1.8 Mb bacterial genome in only 13 months at a cost of only 50 cents per base (i.e. half the current costs at the time) (Fleischmann et al. 1995). To achieve this goal in such a short time frame, the TIGR team developed a new software, named TIGR Assembler, to assemble the massive amount of information resulting from shotgun sequencing of *H. influenzae* genome. Venter's *H. Influenzae* project had failed to win funding from the NIH, as at that time this assembly approach was considered unfeasible by most researchers in the field. Proving the research community wrong, the software assembled approximately 24,000 DNA fragments into the whole genome using 30 Central Processing Unit (CPU) hours with half a gigabyte of Random Access Memory (RAM) (Sutton et al. 1995). Later that year the same approach was used to determine the 0.58 Mbp of *Mycoplasma genitalium* genome, a bacterium associated with reproductive-tract infections and renowned for having the shortest genome of all free-living organisms (Fraser et al., 1995). This genome was sequenced in only 8 months between January and August of 1995. The *H. influenzae* and *M. genitalium* genomes were the first Whole-Genome Sequencing (WGS) efforts using a shotgun approach and an automated assembly pipeline, supporting the potential of the method. In the next two years TIGR rapidly added to the list the first Archaea genome, *Methanococcus jannaschii* (1.66 Mb) (Bult et al. 1996); the 2.2 Mb genome of the hyperthermophilic, sulphate-reducing archaeon *Archaeoglobus fulgidus* (Klenk et al. 1997); the *Helicobacter pylori* genome (1.7 Mb) (Tomb et al. 1997), and the 1.4 Mb genome of the Lyme disease spirochete, *Borrelia burgdorferi* (Fraser et al. 1997).

The first eukaryotic genome was assembled in 1996, with the 12 Mb of yeast (*S. cerevisiae*) genome containing about 6,000 genes (Goffeau et al. 1996). *S. cerevisiae* project was launched in 1989 by Andre Goffeau and involved over 100 laboratories and 600 scientists all over Europe, America and Japan. Each laboratory focussed on a specific portion of the genome so that in the end only 3.4% of the total sequencing efforts were duplicated among laboratories. In 1998, an international consortium led by the american

[10] The first complete *E. coli* genome sequence (4.6 Mb) came only in 1997 (Blattner et al. 1997).

geneticist Richard Wilson announced the sequencing of the 97 Mb of *Caenorhabditis elegans* genome (C. elegans Sequencing Consortium 1998). The same year also saw the first large-scale survey for Single Nucleotide Polymorphisms (SNPs) of the human genome, with 2.3 megabases of human genomic DNA determined by sequencing at the Whitehead Institute for Biomedical Research at Cambridge (D. G. Wang et al. 1998). The first prototypes for genotyping arrays were developed, allowing to simultaneously assess 500 SNPs and demonstrating the feasibility of large-scale identification of human variation (D. G. Wang et al. 1998).

2.2. The genome race

Despite the successes with genomes from lower organisms, formally halfway through the schedule for completing the HGP in 1998, only approximately 50 Mb of human sequence had been determined by the HGP team, representing less than 1.5% of the entire 3,3 Gbp genome. To the task, the HGP team had developed an accurate but laborious strategy named "hierarchical shotgun" sequencing. The pipeline involved the cloning of large fragments of the human genome into bacterial artificial chromosomes (BACs). BACs were then fragmented, size-selected and sub-cloned. Individual clones were picked and grown, purified DNA was isolated and used as a template for automated Sanger sequencing (Shendure et al. 2017). In 1998, the world's sequencing capacity was approximately 100 Mb per year. Nonetheless, HGP leaders announced the intent to complete the sequencing by the end of 2003, 2 years ahead of previous projections (Collins et al. 1998). Events were to take a different direction. A private effort led by Craig Venter's new company Celera decided to rival the public HGP. Building upon the experience gained at TIGR with bacterial genomes, Celera rapidly sequenced the 175 Mbases of *Drosophila melanogaster* genome using an internally-developed WGS approach and a new genome assembler (M. D. Adams et al. 2000; Myers et al. 2000). The new assembly process involved a more effective overlap–layout–consensus approach[11] (Shendure et al. 2017). The clear success with *D. melanogaster* pushed Venter to turn immediately to the human genome and the presence of a second aggressive player sped up the HGP effort. In March 1999, HGP leaders announced the successful completion of the pilot phase, and in September they announced the release of the first draft by spring 2000. The sequence of chromosome 22, the second smallest human chromosome, was published in December that year (Mayor 1999), and at the beginning of year 2000, HGP team had produced and deposited in Genbank two of the three billion bp of the human genome[12]. However, Venter's people at Celera had not been sitting on their hands. Counter to the HGP leaders, they believed direct (i.e. without BAC intermediates) shotgun sequencing to be the fastest and most effective approach (Weber and Myers 1997). Providing strong support to this view, Venter was able to produce a draft of the human genome less than two years after having entered the competition. The race then culminated in a joint announcement by the two group leaders, Craig Venter and Francis Collins, gathered at the White House with United States

[11] One year later Pevzner, Tang and Waterman would have introduced EULER, a radically new assembly approach based on de Bruijn graphs (Pevzner, Tang, and Waterman 2001). This new class of algorithms would have replaced previous overlap graphs.

[12] "Two Thirds of Human DNA Script Deciphered by Human Genome Project". https://www.genome.gov/10002080/2000-release-twothirds-human-dna-sequenced/

President Bill Clinton in June 2000. The HGP followed up on the announcement publishing the genome by February 2001 (Lander et al. 2001), and Venter's published his genome assembly one day later (Venter et al. 2001).

In their violent competition, both HGP and Celera teams had actually assembled and reported only working drafts of the human genome. The most complete of the two assemblies from the HGP team represented about 90% of the genome, with only 25% of the genome in its finished form. Due to the direct shotgun approach employed, Venter's genome was even less accurate (Shendure et al. 2017). A high-quality reference would have required three more years of work by the HGP[13]. Finally, by October 2004 the International Human Genome Sequencing Consortium published the description of the complete human genome sequence[14] (International Human Genome Sequencing Consortium 2004). In the following years and until today, the genome was continuously improved by the Genome Reference Consortium (Schneider et al. 2017).

The cost of the HGP was $2.7 billion in FY 1991 dollars[15]. Thanks to the development of progressively refined sequencing machine based on Sanger sequencing, per base sequencing costs dropped over 100-fold during the whole project. Undisputed market leader at the time was Applied Biosystems. In 1995, the company replaced the previous 370A model with ABI377, which could deal with up to 96 lanes at a time. One year later came the ABI 310, the first capillary DNA sequencer, and in 1998 the first capillary instrument, the ABI Prism 3700. In 2001 was the turn of the 16-capillary ABI Prism 3100, and in 2002 that of ABI 3730xl with 48 to 96 capillaries, where sequences are produced automatically with QVs[16]. These rapid advancements were prompted by the genome race. After the completion of the HGP, the NHGRI[17] started to award millions of dollars every year in research grants to further support the development of new sequencing approaches[18], paving the way to massively parallel sequencing.

3. NextGen: DNA sequencing in the third millennium

Nucleic acids sequencing methods experienced the greatest development after the turn of the millennium, exceeding over 4-fold microchip complexity and computing improvements described by Moore's law (Stein 2010; Heather and Chain 2016). From a technical standpoint, this impressive burst was allowed by advancements in microfabrication, high-resolution imaging and computational power (Heather and Chain 2016). In contrast with the earlier monopoly of Applied Biosystems, this time several companies competed

[13] Meanwhile, in 2002 two new online repositories had been created in addition to Genbank to host human genome data, the University of California Santa Cruz (UCSC) Genome Browser (Kent et al. 2002) and Ensembl (Hubbard et al. 2002).
[14] Actually, this genome did not represent a single individual. DNA from several donors were pooled together, with one individual of European and African ancestry that contributed the most (Green et al. 2010).
[15] "Human Genome Project - FAQ". https://www.genome.gov/11006943/human-genome-project-completion-frequently-asked-questions/
[16] Quality Values. See also footnote 7.
[17] In 1997, the United States Department of Health and Human Services (DHHS) renamed NCHGR the National Human Genome Research Institute (NHGRI), making it one of the 27 institutes and centers that constitute the NIH.
[18] Later turned into a stable plan called "Revolutionary DNA Sequencing Technologies program" that would have laid the foundation of the US$1,000 genome concept.

to prevail on the market (Shendure et al. 2017), including 454, Solexa, Agencourt (Brenner et al. 2000; McKernan et al. 2009), Helicos (Braslavsky et al. 2003; T. D. Harris et al. 2008), Complete Genomics (Drmanac et al. 2010) and Ion Torrent (J. M. Rothberg et al. 2016). The market competition gave birth to a plethora of sequencing technologies, collectively referred as Next-Generation Sequencing (NGS), and now also known as Second-Generation Sequencing (SGS).

3.1. High-throughput sequencing by synthesis

All SGS approaches rely on a library preparation from a source of DNA. In a classical protocol for NGS library generation, DNA fragmentation and size selection is followed by addition of adapters to the end of the DNA fragments[19] and sequencing is performed on the resulting library (usually prior amplification) (Head et al. 2014; Heather and Chain 2016). The most widespread sequencing method was developed by Shankar Balasubramanian and David Klenerman and involves the stepwise, polymerase-mediated incorporation of fluorescently labelled deoxynucleotides (Shendure et al. 2017). In the mid 1990s, Balasubramanian and Klenerman at Cambridge University had been able to observe the motion of single polymerase molecules as they synthesized DNA immobilized on a surface using fluorescently labelled nucleotides. In June 1998, Balasubramanian and Klenerman obtained seed funding from a venture capital firm and founded the company Solexa. In 2003, Solexa proposed a new sequencing approach based on solid phase sequencing (Balasubramanian, Klenerman, and Barnes 2003; Braslavsky et al. 2003; Mitra et al. 2003). Next year Solexa acquired from Manteia the colony sequencing technology (or bridge amplification), which was based on a process invented in 1997 by Pascal Mayer and Laurent Farinelli (Kawashima, Farinelli, and Mayer 1998). In this approach, tightly clustered copies of individual molecules, or "polonies" (Mitra and Church 1999), are produced on a surface from an immobilized template library[20] (C. P. Adams and Kron 1997; Adessi et al. 2000). The amplification of single DNA molecules into clusters enhanced the fidelity and accuracy of base calling, while reducing the cost of the system optics through generation of a stronger signal[21]. In 2005, Solexa added to its method the recently developed reversible terminators (Ruparel et al. 2005; T. S. Seo et al. 2005; Barnes et al. 2006) and an engineered DNA polymerase (Ost 2006). The resulting platform was able to image and determine each single nucleotide added to all the DNA fragments placed on the surface of a flow cell[22]. This approach would have come to be known worldwide as Sequencing by Synthesis (SBS).

[19] Step order may vary and often includes end repair and dA-Tailing as well as purification of the ligation products from the mixture with (often biotin) probes.

[20] Alternative methods introduced over the years include clonal emulsion PCR with copies of each template immobilized on beads that are then arrayed on a surface for sequencing (Dressman et al. 2003; Margulies et al. 2005; Shendure et al. 2005), or rolling circle amplification in solution to generate clonal 'nanoballs' that are arrayed and sequenced (Drmanac et al. 2010).

[21] "History of Illumina Sequencing and Solexa Technology". https://emea.illumina.com/science/technology/next-generation-sequencing/illumina-sequencing-history.html?langsel=/gb/

[22] The addition of a single nucleotide is guaranteed by reversible terminators. These, along with fluorescent groups, are then washed away for a next extension step.

3.2. Second-generation DNA sequencers

The 2005 was probably the *annus mirabilis* for SGS. That year, the first next-generation DNA sequencing machine, the GS20, was introduced in the market by 454 Life Sciences (Henson, Tischler, and Ning 2012). The GS20 used single-molecule template synthesis of small, bead-bound DNA fragments in a water-in-oil emulsion clonal PCR (emPCR) (Tawfik and Griffiths 1998), and dNTPs incorporation detection by Charge Coupled Device (CCD) sensors beneath the surface of about one million microwells in a refined version of pyrosequencing (Ronaghi et al. 1996; Toumazou and Purushothaman 2004; Margulies et al. 2005). The system produced reads around 400–500 base pairs, had 99% accuracy and could sequence up to 25 million bp in a single 4-hour run at less than one-sixth the cost of conventional methods.

Genome Analyzer, the first Solexa commercial sequencer, was launched in 2006. In contrast with GS20, Solexa sequencer had a higher throughput (1 Gbp in a single run) but read length of only 35 bp (Bentley et al. 2008). However, a great advantage was that these represented paired end reads, allowing to size the gap between relatively distant sequences in a DNA fragment. On January 2007, Solexa was acquired by Illumina, a company founded in April 1998 that was to dominate the market in the next decade.

The 2007 was the year of SOLiD from Applied Biosystems. This technology was based on a ligation strategy, relying on the specificity of DNA ligases to ligate fluorescent oligonucleotides to templates in a sequence-dependent manner (Brenner et al. 2000; Shendure et al. 2005). This method was later reported to have some issues in sequencing palindromic sequences (Y.-F. Huang et al. 2012) and subsequently abandoned.

A new sequencing approach based on proton detection in semiconductors was released in 2011 (S. Huang et al. 2010; J. M. Rothberg et al. 2011). This approach relies on the measurement of hydrogen ions release during nucleotide addition in DNA synthesis. Parallel measurements of multiple templates is carried out in microwell plates where each of the four nucleotides is added in succession. Complementary nucleotide incorporation and subsequent release of a hydrogen ion result in a pH variation that can be translated into a voltage change recorded by a semiconductor sensor. As this technology does not relay on imaging, the process of reading each nucleotide can occur in seconds and at considerably lower costs. In January 2012, Ion Torrent released a more powerful machine, called the Ion Proton, which the company claimed could have allowed a large sequencing facility to sequence a human genome in a single day for the long-sought-after price of $1,000. However, one limitation of the Ion Torrent, it turned out, is that it may mismeasure the length of homopolymers (Loman et al. 2012; Song et al. 2017).

By January 2014, Illumina appeared to have reached a position of near monopoly (Greenleaf and Sidow 2014), holding 70% of the market for genome-sequencing machines and accounting for more than 90% of all DNA data produced (Zimmerman 2014; Regalado 2014). The same year, the company announced the HiSeq X Ten, claiming that forty these machines would have been be able to sequence more genomes in one year

than had been produced by all other sequencers to date, allowing large-scale whole-genome sequencing for $1,000/genome[23] (Hayden 2014).

3.3. Genomic big data

The results from this blossom of high-throughput sequencing strategies and machines have been innumerable. The power of the new methods was proved by the cost-effective and rapid re-sequencing of many milestone genomes such as that of *E. coli* (Shendure et al. 2005) and *M. genitalium* (Margulies et al. 2005). At that point, the read output was so high that the 5 kbp-long PhiX174, the same genome Sanger first sequenced using his plus and minus method, became a standard control during Illumina sequencing runs (Mukherjee et al. 2015). At the end of 2000s, SGS also allowed extremely novel applications, as chromatin immunoprecipitation followed by sequencing (ChIP–seq) (Johnson et al. 2007), genome-wide epigenetic landscape determination (Lister et al. 2008), high-throughput RNA-seq (Cloonan et al. 2008; Mortazavi et al. 2008; Nagalakshmi et al. 2008; Wilhelm et al. 2008), chromatin accessibility (Boyle et al. 2008), whole-exome sequencing (Ng et al. 2009) and ribosome profiling (Ingolia et al. 2009), as well as many human genome and cancer genome re-sequencing projects. Human genotyping became popular, with millions of SNPs from hundreds of individuals produced by projects such as HapMap (Thorisson et al. 2005). Indeed, human genome re-sequencing also started to become affordable. The first individuals to have their genome fully re-sequenced were Craig Venter in 2007 (Levy et al. 2007) and Jim Watson in 2008 (Wheeler et al. 2008). In 2011, six-year-old Nicholas Volker was reported as the first patient saved by DNA sequencing, as his one in 1 billion genetic mutation of XIAP gene turned out to be treatable with cord transplant[24]. Efforts in human genome sequencing were resumed with great pomp by the 1000 Genomes Project (2007-2015)[25], and more and more large-scale genome projects have constantly been and are being proposed[26].

This tremendous flood of data was barely accompanied by advancements in bioinformatics tools to store, process, analyse and visualize them. A new series of integrated open source software and algorithms for bioinformatics were released, including R-based Bioconductor (2001); short-read and vertebrate-specific aligner BLAT (Kent 2002); the platform for integrated genome analysis Galaxy (Giardine et al. 2005); the NCBI Short Read Archive (SRA, 2005); the assembly algorithms ALLPATHS (Butler et al. 2008), Velvet (Zerbino and Birney 2008) and SOAPdenovo (R. Li et al. 2010); more efficient read alignment algorithms as Bowtie (Langmead et al. 2009) and BWA (H. Li and Durbin 2009); integrated tools for read data management as SAMtools (H. Li et al. 2009); algorithms for variant discovery as GATK (McKenna et al. 2010) for SNPs and BreakDancer (Chen et al. 2009) and Pindel (Ye et al. 2009) for structural variants; and genome data visualization software as the Integrated Genomics Viewer (J. T. Robinson et al. 2011).

[23] In January 2017, Illumina released NovaSeq, claiming that this new machine will pave the way to the $100 genome. The machine can output up to 3,000 Gbp in a single run.
[24] "One in a billion Foundation". http://www.oneinabillionic.com/our-history/
[25] "1000 Genomes - About". http://www.internationalgenome.org/about
[26] Among others: the Wellcome Trust UK10K in 2010, which aims to compare the genomes of 4,000 healthy people with those of 6,000 people living with a disease of suspected genetic cause; and the All of Us (previously known as the Precision Medicine Initiative) launched by Barack Obama to collect genetic and health data from one million subjects by 2022 (Reardon 2015).

During the great period of excitement that followed the introduction of SGS platforms, cost-effective genome drafts for most model species of eukaryotes were rapidly produced, including the genomes of the mouse (Mouse Genome Sequencing Consortium et al. 2002), the rat (Gibbs et al. 2004), the common chimpanzee (Chimpanzee Sequencing and Analysis Consortium 2005), the rice (International Rice Genome Sequencing Project 2005), the red alga *Cyanidioschyzon merolae* (Nozaki et al. 2007), the mais (*Zea mays*) (Schnable et al. 2009), two ancestral human genomes of Neanderthal (Green et al. 2010) Denisovan[27] (Meyer et al. 2012), the zebrafish (Howe et al. 2013), and the *Xenopus laevis* (Session et al. 2016). The commitment in reference genome sequencing and assembly has recently scaled up from single-species projects to multiple-species coordinated efforts (Genome 10K Community of Scientists 2009; G. Zhang et al. 2014; Jarvis et al. 2014), and several projects to produce high-quality genomes for most organisms are currently underway (Pennisi 2017; Koepfli et al. 2015; G. Zhang et al. 2015; Teeling et al. 2018; Lewin et al. 2018).

4. Genome sequencing in the Third Generation Sequencing era

Notwithstanding all these successes and its widespread usage, the main issue of SGS is the overall quality of the assembled genomes that often fail to map in low-complexity regions and to assemble in long continuous contigs. As more complex genomes are addressed, the impact of repeat elements increases exponentially whereas paired-end sequencing and more robust assembly algorithms based on de Bruijn graphs help only to a very limited extent. Moreover, limitations induced by short-read technology greatly reduce the potential to detect large structural variants (SVs), including length variation in highly repeated motifs, other large insertion and deletion events (indels), duplications, inversions and translocations (**Figure 1**) in re-sequencing efforts (Tattini, D'Aurizio, and Magi 2015).

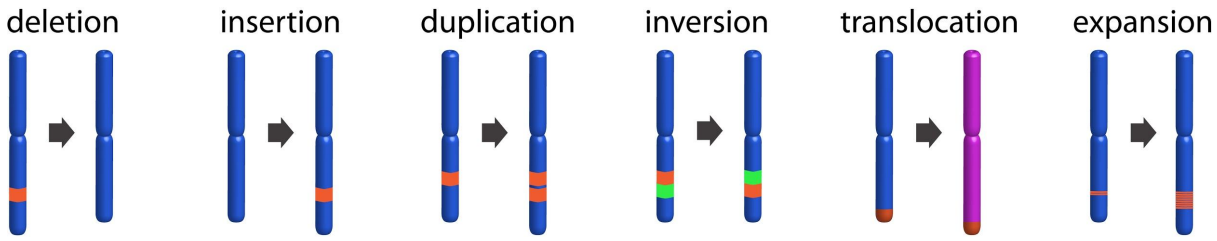

Figure 1: Genomic structural rearrangements.

Overall, short reads do not seem to be able to detect beyond 20% of SVs[28]. While SNPs were long regarded as the most relevant type of genetic variation, it is now clear that SVs also play a key biological role (Sudmant et al. 2015). It has been recently shown that human-chimpanzee genotypic differences are in the order of $3x10^7$ substitutions, $5x10^6$ indels (<80 bp) and $7x10^4$ SVs (>80 bp) (Chimpanzee Sequencing and Analysis Consortium 2005). However, when the number of bp affected is considered, the order of

[27] The branch of ancient DNA developed its own methods to deal with ancient DNA degradation, such as single-stranded library preparation.
[28] Presentation by Pacific Biosciences SV specialist Aaron Wenger, Barcelona PacBio meeting, 1-3 November 2017.

importance of genotypic differences is reversed, with structural variants accounting for 68 Mb of overall variation (57%), substitutions (1 bp) for 30 Mb (25%) and indels for 22 Mb (18%)[29]. This implies that the vast majority (57-75%) of genomic variation may so far have escaped detection by SGS. This in turn could account for much of the "missing heritability" problem (Eichler et al. 2010). Accordingly, evidence for the importance of SVs in determining both simple and complex phenotypic traits is constantly growing (Lamichhaney et al. 2016; Thomas et al. 2008; Horton, Moore, and Maney 2014; Joron et al. 2011; J. Wang et al. 2013; Kunte et al. 2014; Nishikawa et al. 2015). In birds, one striking example of SVs relevance is the ruff (*Philomachus pugnax*), where three extremely differentiated male morphs coexist (Hogan-Warburg 1966; Jukema and Piersma 2006). WGS has now allowed the solution of the scientific enigma of how such complex phenotypic variation can be maintained in a Mendelian fashion, revealing that the two morphs of minor frequency are determined by just two alleles, both associated with an evolutionary stable 4.5 Mb inversion that occurred about 3.8 million years ago (Lamichhaney et al. 2016).

The development of new SGS-based approaches and of radically new sequencing technologies has now opened the era of Third-Generation Sequencing, which promises to unveil the entirety of genome complexity at the population scale (Bleidorn 2016). TGS definition may vary, but it is generally attributed to technologies capable of sequencing single molecules without DNA amplification (Heather and Chain 2016). The first Single Molecule Sequencing (SMS) technology was developed by Stephen Quake (Braslavsky et al. 2003; T. D. Harris et al. 2008) and commercialized in 2009 by Helicos BioSciences. It worked broadly in the same manner that Illumina does, but without any bridge amplification. This was slow, expensive and produced relatively short reads, nonetheless it avoided all DNA amplification-associated biases and errors. Unfortunately, Helicos filed for bankruptcy early in 2012.

Two other TGS approaches arose in the 2010s. The first approach, known as long-read Single Molecule Real-Time (SMRT) sequencing was developed by Watt Webb and Harold Craighead at the Cornell University and was further refined and commercialized by Jonas Korlach and Pacific Biosciences (PacBio). The second approach is nanopore sequencing[30]. The first assemblies of the human genome using PacBio and Oxford Nanopore technologies were reported in 2016 (J.-S. Seo et al. 2016) and in 2018 (Jain et al. 2018), respectively. Another single-molecule DNA technology is Bionano optical mapping, though this approach does not involve sequencing. The two following sections outline in more detail the two cutting-edge TGS technologies employed in this study: PacBio long-read SMRT sequencing and Bionano optical mapping.

---

[29] *Ibidem.*

[30] Nanopore sequencing was first hypothesized in the 1980s but, since electric field-driven transport of DNA through a nanometre-scale pore is so fast that the number of ions per nucleotide is insufficient to yield an adequate signal, decades of work were required to develop the concept into a technology (Church et al. 1998; Bayley 2015; Deamer, Akeson, and Branton 2016). Oxford Nanopore Technologies (ONT), a company founded by Bayley in 2005, recently succeed (S. Huang et al. 2010; Manrao et al. 2012; Cherf et al. 2012). ONT uses genetically modified bacterial nanopores inserted into an artificial lipid bilayer, placed in individual microwells tens of micrometers wide, and arrayed on a sensor chip. As each nucleotide or single strand of DNA travels through a channel it disrupts a current running through the pore, and the change is measured by a semiconductor sensor. Because each base disrupts the electric field in a slightly different way, those current changes can then be translated into a DNA sequence. The longest reads so far obtained are in the order of 900 kilobases (Jain et al. 2018). Moreover, as they rely on the detection of electronic (rather than optical) signals, nanopore devices can be as small as a USB stick. In 2016, such portability allowed to sequence Ebola virus at field sites in West Africa in less than 60 minutes (Quick et al. 2016).

4.1. SMRT long reads for Whole Genome Sequencing

The principles of SMRT sequencing were originally put forward in 2011 and rely on Zero-Mode Waveguide (ZMW) nanowell arrays (Levene et al. 2003), where a single DNA polymerase is bound to the bottom of each well (Eid et al. 2009) (**Figure 2**). This provides volume confinement (100 nm holes, 20 zeptoliters each) where the incorporation of fluorophore-labeled nucleotides in thousands of parallelized single-molecule sequencing reactions is literally filmed in real time.

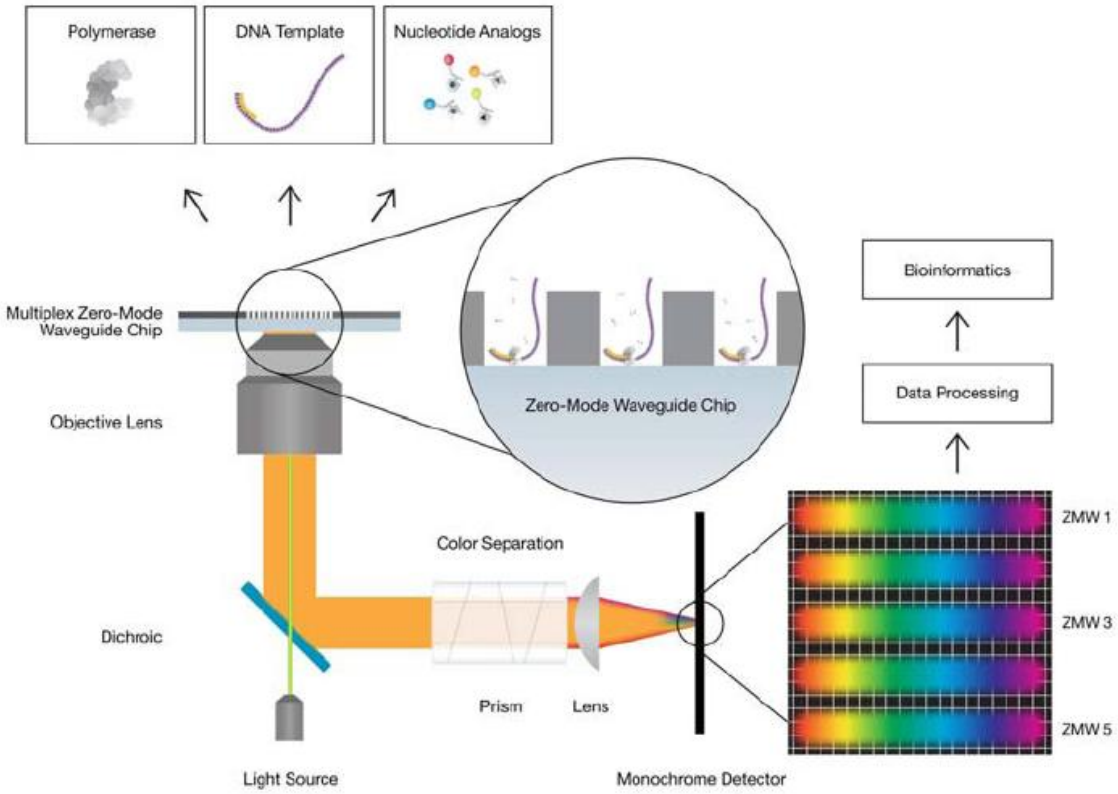

Figure 2: SMRT sequencing workflow. *Image courtesy of Pacific Biosciences.*

This setting allows SMRT sequencing to offer many advantages over SGS. These include ultralong read lengths (even beyond 200 kbp, >10 kbp on average) (R. J. Roberts, Carneiro, and Schatz 2013); high consensus accuracy (Rhoads and Au 2015); as well as low sequencing-context bias (either GC-content or low-complexity) and therefore uniform coverage along the genome. These technical advantages result in accurate mapping of sequencing reads, greatly facilitating SVs detection (Sudmant et al. 2015; Huddleston et al. 2017; Merker et al. 2018). Therefore, SMRT sequencing is considered the current optimum for generating *de novo* genome assemblies (Rhoads and Au 2015). Moreover, SMRT sequencing provides simultaneous capability of epigenetic characterization (Rhoads and Au 2015). In fact, SMRT sequencing is theoretically capable of simultaneously assessing not only the single nucleotides, but also 25 epigenetic marks associated with them (Flusberg et al. 2010), many of which have only been reported in prokaryotes (R. J. Roberts et al. 2015). This is allowed by the unique temporal pattern recorded when a methylated base is read by the

polymerase (Flusberg et al. 2010). SMRT sequencing would therefore represents the best solution to assess both the genome and the epigenome of an individual, potentially allowing to reveal previously unreported epigenetic signatures. Unfortunately, at least 200-250X coverage is required to detect with confidence the epigenetic modifications at bp resolution[31], although lower coverage (at least 20X) has been shown to reveal CpG islands (Suzuki et al. 2016). Finally, the amplification-free single-molecule resolution of SMRT sequencing provides good chances to simultaneously assess somatic mosaicism (Eid et al. 2009).

Another recent feature of PacBio platform is that with its SMRT Link software suite, can provide "joint calling", which simultaneously considers reads from multiple, related (i.e. parents and offspring) individuals. Join calling takes advantage of coverage in one individual to support variant calls in another, ultimately increasing sensitivity and allowing to identify shared variants. For example, going from *solo* variant calling at 5X in trios to joint 3×5X variant calling in trios led to a +22% in SV discovery in humans[32].

All the advantages of SMRT sequencing over SGS currently come at the price of a higher per base sequencing cost. In 2018, Pacific Biosciences announced the release by 2019 of new chemicals and platforms that will guarantee an up to 8-fold drop in sequencing costs on their platform. Should Pacific Biosciences fulfil this objective, the costs of SMRT sequencing and SGS will become similar, supporting the use of a long-read only sequencing strategy for many applications, including combined whole-genome and whole-epigenome projects.

4.2. Optical mapping for hybrid scaffolding

While SMRT sequencing outputs usually contain several reads >100 kbp long, the average N50 (i.e. the read median length over which lays 50% of total base pairs) is generally 10-15 kbp. Thus, massive direct sequencing of high molecular weight (>50 Kbp) DNA still proves challenging. In this context, optical maps, another single molecule technology, have already proven invaluable to assemble contigs in larger, often chromosome-level, scaffolds (Hastie et al. 2013; Gnerre et al. 2011; Pendleton et al. 2015; Mostovoy et al. 2016). Optical mapping also allows detection of SVs encompassing the length of individual reads, either NGS short reads or TGS long reads, simultaneously validating most SVs found with its high sensitivity to even relatively small (< 500 bp) indels (Cao et al. 2014; Mak et al. 2016).

Nowadays, the company leader in optical mapping is Bionano Genomics. Bionano optical mapping relies on nanoscale channels that, through progressively smaller sieves, come to accommodate thousands of individual, ultralong (>150-200 Kbp) double-stranded DNA filaments in parallel (Lam et al. 2012) (**Figure 3**).

[31] "DNA Modification Detection with SMRT Sequencing Using R". https://github.com/PacificBiosciences/R-kinetics
[32] *Ibidem* 25, 19.

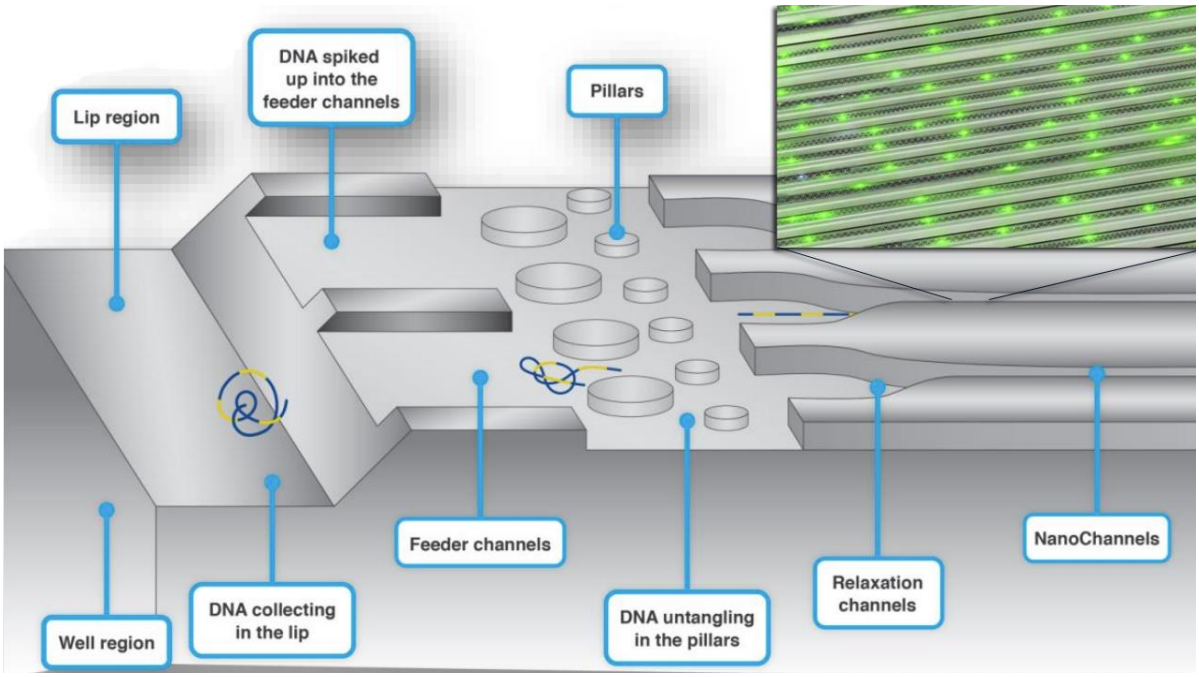

Figure 3: Scheme of nanochannels for Bionano optical mapping. *In a Bionano chip, DNA molecules are linearized while they flow through progressively finer sieves.*

Molecules are then stained to recognize specific 6-7 bp motifs and imaged. An example of optical maps produced by a run is provided in **Figure 4**.

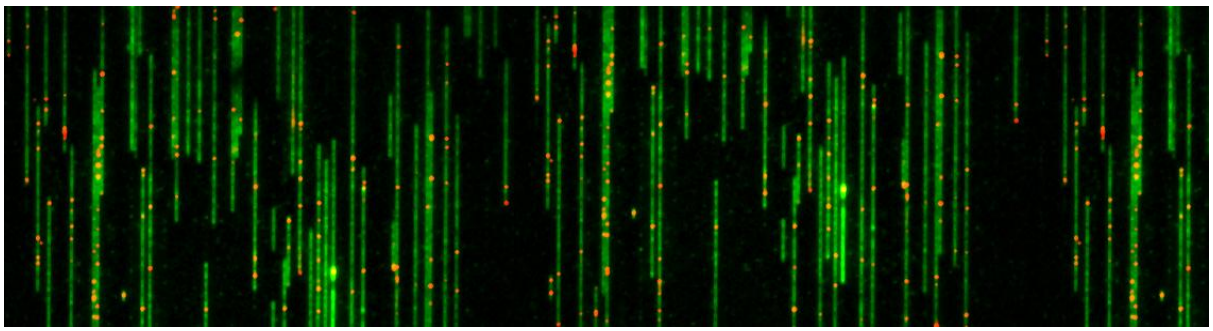

Figure 4: Staining of DNA molecules in a Bionano chip. *In Bionano optical mapping each individual DNA molecule is imaged while it runs through parallel nanochannels. Molecules are fluorescently labelled at different sites with several enzymes (using green and red labels in this image). Label-to-label distance can be measured afterwards, allowing for the production of the optical maps.*

Label to label distance is then measured to produce the optical maps. The resulting patterns are matched with available WGS data to either correct or improve previous draft assemblies and to validate/identify SVs by comparing the results to a reference genome (**Figure 5**).

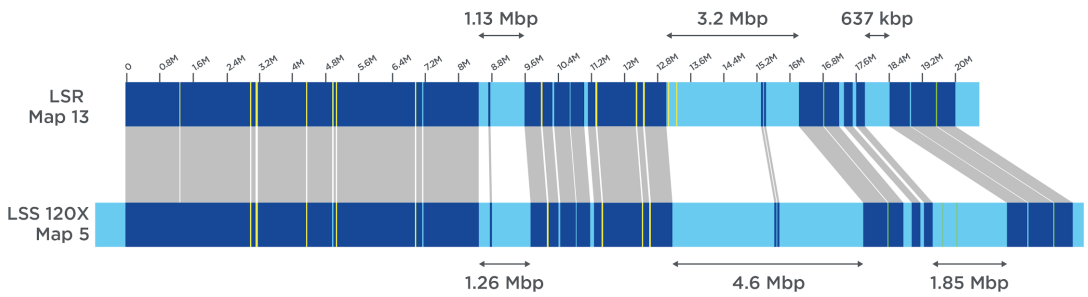

Figure 5: Example of a Bionano optical map. *The difference in length in certain regions of the map may highlight the presence of indels.*

Until 2018, two enzymes were available for the staining of the DNA molecules: BSPQI, with recognition sequence GCTCTTCN^, and Nb.BssSI with recognition sequence CACGA^G. These were both nicking enzyme, that is they labelled the molecules introducing the fluorescent tag via a nick on one of the two DNA strands. This approach was collectively referred as Nick, Label, Repair and Stain. In March 2018, Bionano Genomics[33] introduced the non-nicking enzyme DLE-1, with recognition sequence CTTAAG, therefore allowing the Direct Label and Stain (DLS) approach employed in this study.

[33] DLS Technology - Bionano Genomics. https://bionanogenomics.com/technology/dls-technology/

The Barn Swallow

«Doubtless when the swallows arrive in spring, they operate like clocks.»
René Descartes, 1646

Barn swallows are among the wild bird species with the closest relationship with human beings. They have been nesting on human artefacts for thousands of years. No ancient civilization around the Mediterranean was immune to their fascination: they were considered a minor deity by ancient Egyptians as well as by the Greeks, who invented the myth of Chelidonia. In the Japanese culture, they are still regarded as a sign of good fortune when they arrive in spring. Compared to other birds, barn swallows are very common in several rural habitats and relatively easy to study in large numbers. Indeed, the scientific interest for this species has been spurred by these synanthropic habits and by its abundance, as well as by its worldwide distribution, its philopatric migratory behaviour, its worrisome conservation status and cultural value. This interest is testified by the over 1,200 studies conducted in many populations (Europe, Israel, Japan and North America) since 1985. Many studies have also been carried out earlier in the XX century, making the barn swallow one of the most studied bird species. These studies have focused on its biology, life history, morphological, behavioural and physiological traits as well as on investigating its intraspecific sexual selection and response to climate change. This rapid build-up of scientific knowledge on fundamental aspects of the barn swallow biology has fostered further diverse and most interesting scientific questions. In this context, the availability of high-quality genomic resources, including a reference genome, is thus pivotal to further boost the study and conservation of this species.

1. General description

The barn swallow is a small (about 20 g), migratory, semi-colonial and socially monogamous, aerially insectivorous passerine bird belonging to the Hirundinidae family ('hirundines'), which includes some 83 species throughout the World (Møller 1994; Turner 2006; Cramp 1998; Turner 2004). The barn swallow is a polytypic species, which occurs with at least 8 subspecies with a Holarctic distribution in Europe, Asia and North America (**Figure 6**).

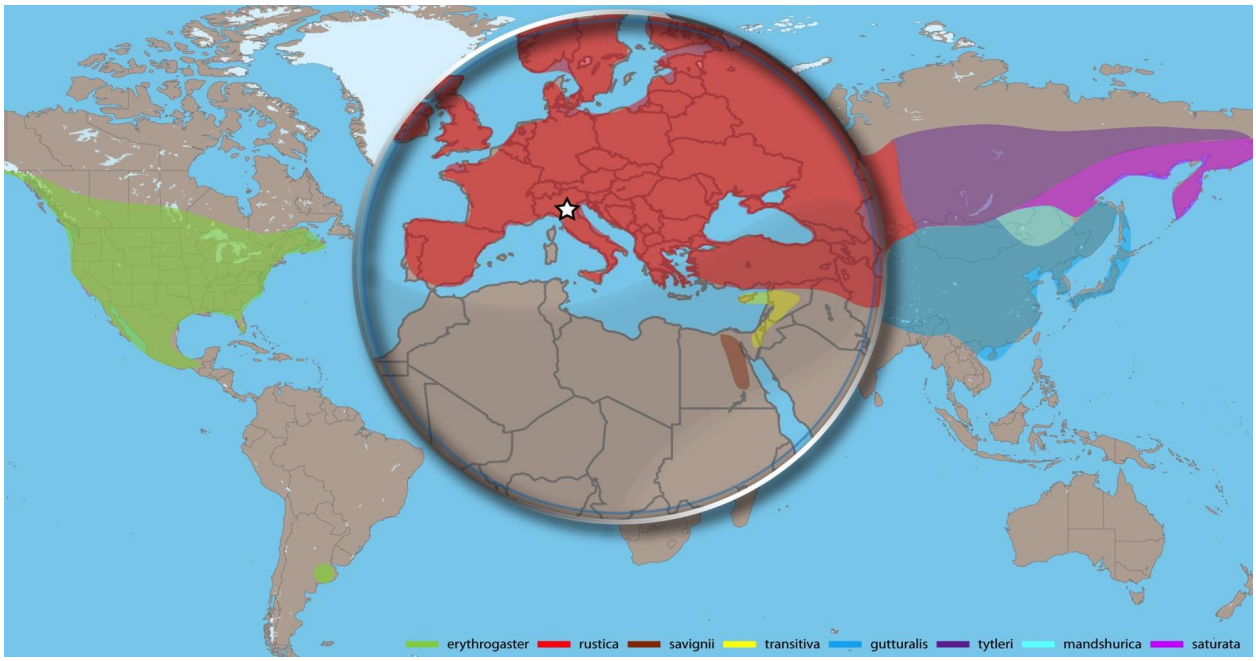

Figure 6: Barn swallow (Hirundo rustica) subspecies in the World. *The regions where the 8 subspecies of* Hirundo rustica *are found are highlighted with different colors as indicated in the legend. In the present work, I focussed on the European subspecies (*Hirundo rustica rustica*) indicated in red (the white star marks the sampling locality).*

The European barn swallow (*Hirundo rustica rustica*) (**Figure 7**) breeds in a broad latitudinal range, between 63-68°N and 20-30°N (Turner 2006).

Figure 7: The European barn swallow (*Hirundo rustica rustica*). *Courtesy of Chiara Scandolara.*

Taxonomic relationships between subspecies still remain debated as they have not yet been fully evaluated with high-resolution molecular tools.

1.1. Distribution and migration

Almost all barn swallow populations are migratory. Only *H. r. savignii* (Nile valley) is mostly resident (Cramp 1998; Turner 2006). In Europe, fall migration to sub-Saharan Africa occurs in August-October. Barn swallows are then mostly sedentary while wintering from the Sahel to Southern Africa, when they accomplish a complete annual moult (Jenni and Winkler 1994; Turner 2006; Saino, Romano, Caprioli, Lardelli, Micheloni, et al. 2013). Spring migration occurs in March-May. Migration phenology varies between sexes, age classes and geographical populations (Møller 1994; Cramp 1998; Rubolini, Spina, and Saino 2004; Saino, Szép, Romano, et al. 2004; Turner 2006; Møller 2007; Saino et al. 2010; Liechti, Scandolara, and Rubolini 2015). European geographical populations show migratory connectivity and follow different migration routes (Møller 1994; Cramp 1998; Saino, Szép, Ambrosini, et al. 2004; Turner 2006; Saino and Ambrosini 2008; Ambrosini, Møller, and Saino 2009; Ambrosini, Rubolini, et al. 2011; Ambrosini et al. 2014). Studies on migration phenology are generally performed by deployment of geolocators which enable to record the location of the wintering site(s), the date of departure from the breeding colonies and the date of arrival to the stationary sites in sub-Saharan Africa, the duration of stationary periods in the wintering grounds, the date of departure from the wintering areas and the date of arrival to the breeding site, as well as the routes followed during autumn and spring migration. Several lines of evidence suggest that central components of the migration phenotype, such as migratory orientation or the propensity to migrate, are under genetic control (Berthold 1991; Helbig 1996; Pulido et al. 2001; Liedvogel, Åkesson, and Bensch 2011).

1.2. Dispersal and longevity

Barn swallows are particularly well suited for longitudinal (lifelong) studies because of their high breeding philopatry (Møller 1994; Saino, Bolzern, and Møller 1997; Cramp 1998; Saino, Calza, et al. 1999; Turner 2006; Saino et al. 2012; Romano et al. 2016). Exhaustive, yearly capture of all adult individuals (from 1 year of age onward) breeding in a particular colony thus allows accurate estimates of survival. On the other hand, natal philopatry is very low, with a local return rate of the young (mainly males) around 5% (Møller 1994; Cramp 1998; Turner 2006; Scandolara, Lardelli, et al. 2014). Annual adult survival (from age 1 year onwards) is 30-40% (Møller 1994; Turner 2006).

1.3. Ecology

In Europe, barn swallows mostly breed colonially in rural buildings, in close association with animal farming (Moller, de Lope, and Saino 1995; Ambrosini, Bolzern, Canova, Arieni, et al. 2002; Ambrosini et al. 2006; Ambrosini, Bani, et al. 2011; Ambrosini et al. 2012). Ecological conditions during wintering and migration

have major effects on survival, phenology and breeding performance as well as on population dynamics (Moller, de Lope, and Saino 1995; Ambrosini, Bolzern, Canova, and Saino 2002; Saino, Szép, Romano, et al. 2004; Balbontín et al. 2009; Saino, Romano, and Caprioli 2012; Sicurella et al. 2014; Saino, Rubolini, et al. 2015; Sicurella et al. 2016). Barn swallows harbour virulent ectoparasites and hematozoans, and susceptibility to parasite infection/infestation is heritable (Møller 1990, 1994; Møller, Martinelli, and Saino 2004).

1.4. Geographical variation

Extensive variation exists in phenological, migration and morphological traits both within and among geographical populations. Tail length, which is under inter-sexual selection in European populations (Cramp 1998), shows the largest dimorphism among morphological traits and considerable geographical variation (Møller 1994; Moller, de Lope, and Saino 1995). The melanin-based coloration of the ventral plumage regions varies continuously from white to dark chestnut within populations (Saino, Romano, Rubolini, Teplitsky, et al. 2013). Measures of morphological traits, coloration, feather growth rates and parasite load (i.e. ectoparasitic Diptera, Mallophaga and Acari or haemosporidian blood parasites as *Plasmodium*, *Haemoproteus* and *Leucocytozoon*) are routinely performed in the study populations. Morphological traits, including wing length, tail length and coloration, are heritable (Møller 1994; Saino, Martinelli, et al. 2003; Saino, Romano, Rubolini, Teplitsky, et al. 2013).

1.5. Breeding and sexual behaviour

The barn swallow attains sexual maturity at 1 year of age (Møller 1994; Cramp 1998; Turner 2006). Pairs have 1-3 clutches (2-7 eggs each) per breeding season (March-August). Offspring are attended for ca. 20 days before fledging and 1 week after fledging. In Europe, breeding most often occurs in colonies (up to tens of pairs) (Møller 1994; Cramp 1998; Turner 2006). Females choose both social and extra-pair mates based on a number of male phenotypic traits (Møller 1988; Møller 1994; Moller, de Lope, and Saino 1995; Møller, Saino, et al. 1998; Møller et al. 2003; Turner 2006) and differentially invest in reproduction based on those traits (Saino, Bertacche, et al. 2002; Saino, Ambrosini, et al. 2002; Møller et al. 2006; Saino et al. 2014; Romano et al. 2015). Depending on the geographical population, females prefer as social and extra-pair mates, males with long tail feathers, large white spots on the tail feathers, dark ventral coloration, dark chestnut forehead patch and producing more elaborated songs (Saino, Primmer, et al. 1997; Møller, Barbosa, et al. 1998; Vortman et al. 2011; Scordato and Safran 2014; Wilkins et al. 2016; Romano et al. 2017). Male secondary sexual traits preferred by females impose physiological or survival costs (Moller, de Lope, and Saino 1995; Saino, Cuervo, et al. 1997; Saino, Bolzern, and Møller 1997). Morphological traits under directional sexual selection, 'ordinary' traits and also phenological traits show additive genetic variation and moderate-to-high heritability (Møller 1994; Moller 2001; Saino, Martinelli, et al. 2003; Saino, Romano, Rubolini, Teplitsky, et al. 2013; Wilkins et al. 2016). Several fitness traits of males are correlated with traits that are targeted by female mate choice (Saino and Møller 1996; Saino, Galeotti, et al. 1997; Galeotti, Saino, and Sacchi 1997; Saino, Stradi, et al. 1999; Saino, Incagli, et al. 2002; Saino, Ferrari, et al. 2003; Saino et al.

2011; Saino, Romano, Rubolini, Ambrosini, et al. 2013; Saino, Canova, et al. 2013; Saino, Romano, et al. 2015; Romano, Saino, and Møller 2017). Studies of breeding phenology and performance generally assess the date of laying of first, second and third clutches, the number of eggs in all clutches, the fledging success and the lifetime reproductive success. More elaborate analyses have also included tests on personality traits as tonic immobility, boldness in reaction to handling or boldness towards a predator. Other physiological characters that have been the focus of several studies are telomere length (Parolini et al. 2015) and oxidative status, both potentially important components of bodily condition with repercussions on viability and performance and both measured in peripheral red blood cells.

2. Genetic studies in the barn swallow

The barn swallow has relatively recently become the focus of several genetic studies. Some of these studies investigated the divergence between populations and subspecies, while other focussed on the genetic control of phenological traits.

2.1 Demography and association mapping

Early analyses of population structure in the barn swallow involved a few microsatellite markers (Primmer, Møller, and Ellegren 1995; Tsyusko et al. 2007; Santure et al. 2010) and pointed to no broad-scale genetic differentiation, even between designated subspecies (Dor et al. 2012). More recently, three genomic studies were performed using genotyping-by-sequencing (ddRAD). In the first analysis, the authors identified 9,493 SNPs from 350 individuals belonging to 8 populations and 4 subspecies and performed an association study based on pairwise FST accounting for geographic distance, environmental context (i.e. altitude, temperature) and phenotype (wing length and breast colour) (Safran et al. 2016). This led to the identification of two divergent traits related to migratory behaviour and sexual signalling, that together with geographic distance, explained >70% of genome-wide divergence among populations. This work was also the first where a draft of the barn swallow genome was presented, as detailed in Section 3.

The same year another research group analysed separately eight microsatellite loci (Hru2, Hru7, Hir6, Hir10, Hir15, Hir19, Hir20 and Hir22) specifically developed for the barn swallow (Primmer et al. 1995; Tsyusko et al. 2007) in 452 individuals, a 1,023-bp stretch of the mitochondrial ND2 gene in 291 individuals and >20,000 ddRAD markers on a subset of 216 individuals along a migratory divide in Central Europe (von Rönn, Shafer, and Wolf 2016). Then, they analysed the population structure and found that population structure among breeding populations was essentially absent with results from microsatellites, mitochondrial DNA sequence and ddRAD sequencing were highly concordant with low and non-significant FST estimates between the three sampling areas and a single genetic cluster best explaining the genotypic data suggesting one panmictic population where gene flow is overwhelming. They also performed outlier analysis for sampling area and migratory type. One single outlier was observed among sampling areas and none were detected among the migratory phenotypes within the migratory divide. The single outlier locus mapped to the BUB1 gene which has a role in mitotic and meiotic organization. A low number of outliers is not unexpected given the moderate marker density of ~1 SNP every 58 kbp and assuming a genome size of 1.28

Gbp (Andrews, Mackenzie, and Gregory 2009), and the general fast decay of linkage disequilibrium in birds (Poelstra, Ellegren, and Wolf 2013). Unless a marker happens to hit a causal variant (which will be rare), we would expect only a very small number of SNPs to be part of an extended region of differentiation elevated by strong and recent selection. This led the authors to conclude that whole-genome re-sequencing would have be needed to characterize the genome-wide differentiation landscape.

Lastly, in 2017 Safran and co-workers performed a new analysis of 23,251 SNPs from 533 individuals along a transect in the Siberian range for *H. r. rustica, tytleri* and *gutturalis (Scordato et al. 2017)*. This allowed to assess the degree of gene flow among the three subspecies and to test whether the degree of divergence in ventral coloration and wing length was associated with the extent of hybridization in secondary contact. This also allowed them to identify genomic regions associated with throat brightness and wing length that are likely to contribute to the differentiation between the three barn swallow subspecies.

2.2. Genetic control of phenological traits

Perception of variation in day length is a major proximate mechanism beating the time of seasonal, periodic changes in physiology and behaviour. The 'circadian clock' senses temporal variation in light/dark cycles and produces a cascade of physiological processes that can ultimately cause adaptive behavioural shifts, such as breeding or preparing to and undertaking migration in birds. A large body of studies has led to the identification of several genes that are in control of the circadian clock and to the dissection of the molecular bases of circadian oscillations (Dvornyk, Vinogradova, and Nevo 2003; Hall 2003; Caprioli et al. 2012). In mice, variation in the number of CAG polyQ repeats in the carboxyl-terminal polyglutamine stretch (polyQ) of the *Clock* gene affects circadian rhythms (Vitaterna et al. 2006; O'Malley, Ford, and Hard 2010; Caprioli et al. 2012). Despite low polymorphism at the *Clock* gene in barn swallow populations from Italy (Caprioli et al. 2012) and other regions (Dor et al. 2011), the laboratory where this Ph.D. thesis work was conducted was able to show that an increasing number of CAG repeats in the polyQ tract is associated with delayed reproduction and molt (Caprioli et al. 2012; Saino, Romano, Caprioli, Fasola, Lardelli, et al. 2013) and delayed migration (Bazzi et al. 2015). Interestingly, *Clock* polyQ length appeared to be under negative viability selection because the frequency of 'long' polyQ alleles declined among older individuals compared to younger ones (Caprioli et al. 2012). Moreover, methylation at *Clock* predicted spring migration phenology and thus breeding success (Saino, Ambrosini, Albetti, et al. 2017).

3. A genome for the barn swallow

Several characteristics, shared by many bird species, make the barn swallow a good candidate for WGS and genomic studies in vertebrates. First, flying birds have the smallest genomes among amniotes, and the barn swallow genome size has been estimated in 1.28 Gbp (Andrews, Mackenzie, and Gregory 2009). Secondly, avian genomes are generally characterized by high stability and evolutionary stasis, determining long syntenic gene blocks that allow easier interspecific whole-genome comparisons. Importantly, especially for some TGS approaches requiring high amounts of High Molecular Weight (HMW) DNA, nucleated red blood cells in birds provide easy access to huge quantities of DNA. Finally, while relatively high recombination

rates in birds can easily disrupt linkage-disequilibrium between co-segregating molecular markers making reduced representation gene mapping less effective (Poelstra, Ellegren, and Wolf 2013; von Rönn, Shafer, and Wolf 2016), inexpensive WGS can overcome this issue as it potentially allows determining the causative *loci* directly. However, while the barn swallow is certainly a highly studied model species from the ecological point of view, very few genomic studies have been conducted to date, also due to the lack of a reference genome. A first draft of the genome for the subspecies *H. r. erythrogaster* was reported in 2016 by a research group from the University of Colorado (Safran et al. 2016). While it was not possible to analyze it as genome data are still undisclosed to date, this draft assembly is certainly affected by the limitations common to many SGS-based genome assemblies. First, it contains only 1.1 Gb of assembled sequences compared to the overall estimated genome size of 1.28 Gb (Andrews, Mackenzie, and Gregory 2009), implying missing information for approximately 15% of the genome. Second, the estimated coverage (47x), together with the use of SGS (Illumina Hiseq 101-base paired end), did not allow the assembly of contigs over the average length of 11,010 bp (N50 = 38,844 bp and N90 = 3,718 bp), thus preventing the assembly of long scaffolds for most of the genome (longest scaffold: 732,517 bp). Moreover, the genome was derived from a male, thus excluding information for the W chromosome because in birds females are the heterogametic (ZW) sex. As a consequence of the absence of a complete high-quality genome assembly, the authors could only putatively map contigs and genes on chromosomes by blasting them on another bird species (i.e. the collared flycatcher *Ficedula albicollis*) genome, under the theoretical assumption of synteny.

With the goal of overcoming the aforementioned issues, I employed TGS to produce a high-quality genome for the European subspecies (*Hirundo rustica rustica*). This genome sequencing and assembly has been carried out in compliance with the pipeline and guidelines for a Platinum standard reference genome of the Vertebrate Genome Project (VGP), of which my supervisor and I are the only partners in Italy. The VGP is an international endeavour aimed to generate and make available to the public through a digital open-access library called Genome Ark, near-gapless, chromosome-level, phased and annotated reference-quality genome assemblies of all the ~66,000 vertebrate species living on the planet Earth[34]. The VGP is led by former members and institutions from the 2009-established Genome 10K Consortium (G10K), a previous effort to produce 10,000 vertebrate genomes that merged with the VGP. These institutions include the Vertebrate Genome Laboratory (VGL) at the Rockefeller University (New York, US), the Wellcome Sanger Institute (Hinxton, UK) and the Max Planck Institute of Molecular Cell Biology and Genetics (Dresden, Ge). As stated by the the VGP "*high-quality error-free genome assemblies and annotations are necessary as current 1st and 2nd generation genome sequencing approaches generate numerous errors that cause a variety of problems in downstream analyses. Parts of genes are missing, and some are incorrectly assembled, while others are completely missing from the assemblies despite pieces found in the raw sequence reads. Due to these fragmented, error-prone assemblies, researchers have had to clone, re-sequence, and correct individual genes. In some cases, the gene structures are too complex, too long, or too*

[34] Vertebrate Genomes Project Plan. https://www.rockefeller.edu/research/vertebrate-genomes-project/vertebrate-genomes-project-plan/

closely related, preventing even the Sanger-based higher quality 1st generation methods from correcting genome assemblies. In many other instances, investigators do not even know that they are working with incorrect gene sequences and structures, impacting many scientific findings and scientific progress."[35]

The VGP has defined a "Platinum" standard for a genome assembly to be included in the VGP list. This standard is called a 3.4.2.QV40 phased metric and requires the following specific genome metrics to be simultaneously achieved:

- Contig N50 above 1 million bp, i.e. over 50% of assembled contigs with lengths above 1Mb;
- Scaffold N50 above 10 Mb;
- 90% of the genome assembled into chromosomes confirmed by 2 independent sources;
- Base-calling quality error of QV40, i.e. no more than 1 nucleotide error in 10,000 bp;
- The two haplotypes of the diploid genome correctly phased, i.e. haplotypes clearly distinguished from one another.

According to VGP guidelines, achieving the 3.4.2.QV40 phased metric using the current VGP pipeline includes the following technologies:

- Pacific Biosciences (Menlo Park, US) Single Molecule, Real-Time (SMRT) sequencing for the initial assembly of phased contigs at a final 60X coverage (30X/haplotype), which I performed relying of the FGCZ;
- 10X Genomics (Pleasanton, US) linked reads for intermediate-range scaffolding and further phasing at a final 70X coverage (35X/haplotype);
- Bionano (San Diego, US) Next-Generation-Mapping (NGM) at a final 80X coverage to adjust previous scaffolding errors and for further scaffolding, which I also performed at FGCZ;
- Hi-C linked reads from Arima Genomics (San Diego, US) at a final 70X for long-range scaffolding up to chromosome length, which we also performed in collaboration with the VGL.

Of the over 380 vertebrate genomes in the NCBI database on May 2018, only 16 meet these standards[36]. In September 2018 the Vertebrate Genomes Project has released 15 genome assemblies achieving the 3.4.2.QV40 metrics (Bioproject PRJNA489243)[37].

[35] Vertebrate Genomes Project - Technology pipeline and policies. https://www.rockefeller.edu/research/vertebrate-genomes-project/technology-pipeline-and-policies/
[36] Vertebrate Genomes Project - Technology Pipeline and Policies. https://www.rockefeller.edu/research/vertebrate-genomes-project/technology-pipeline-and-policies/
[37] Vertebrate Genomes Project - Phase 1 first data release. https://vertebrategenomesproject.org/news/

METHODS

«Progress in science depends on new techniques,
new discoveries and new ideas,
probably in that order.»
Sydney Brenner, 2002

Blood sample collection

The blood used as a source of DNA was derived from a minimally invasive sampling performed on a single female individual of approximately two years of age during May 2017 in a farm near Milan in Northern-Italy (45.4N 9.3E)[38]. Appropriate consent was obtained from the local authorities (Regione Lombardia). Blood was collected in heparinized capillary tubes. Three hours after collection, the sample was centrifuged to separate blood cells from plasma and then stored at -80°C.

DNA extraction and quality control for SMRT library preparation

DNA extraction was performed on blood cells portion of centrifuged whole blood containing nucleated erythrocytes and leukocytes using the Wizard genomic DNA purification kit (Promega, Cat. No. A1125). This kit employs a protocol similar to classical Phenol/Chloroform DNA extraction, with no vortexing steps after cell lysis. After purification, DNA quality and concentration was assessed by Nanodrop (Thermo Fisher Scientific, Cat. No. ND-1000) and subsequently by Pulsed Field Gel Electrophoresis (PFGE). Detectable DNA was over 23 kbp in size, with the vast majority over 50 kbp and even over 200 kbp (**Figure 8**). The presence of HMW DNA is relevant for SMRT library preparation as subsequent shearing of DNA at desired (10 to 15 kbp) fragment length is best achieved with DNA of > 100 kbp (and usually at least 50 kbp).

[38] The same individual was recaptured in 2018 upon return during the breeding season. This time the individual was sacrificed and organs were separately collected and stored at -80 °C. These included spleen, liver (with gallbladder), heart, pancreas, duodenum, glandular stomach, muscular stomach, blind along with fast and rectum, ovaries, oviduct, trachea, oesophagus, lungs, kidneys, upper and deep pectoral muscle, tongue, eyes, brain, cerebellum and bone marrow. Appropriate consent was obtained from the local authorities (Regione Lombardia). Curiously, this constituted a fortunate event as on-year survival of barn swallows is around 30% (Scandolara, Lardelli, et al. 2014; Scandolara, Rubolini, et al. 2014; Liechti, Scandolara, and Rubolini 2015; Dunn, Hobson, and Liechti 2015; Matyjasiak et al. 2016; Saino, Ambrosini, and Caprioli 2017; Saino, Ambrosini, Caprioli, et al. 2017).

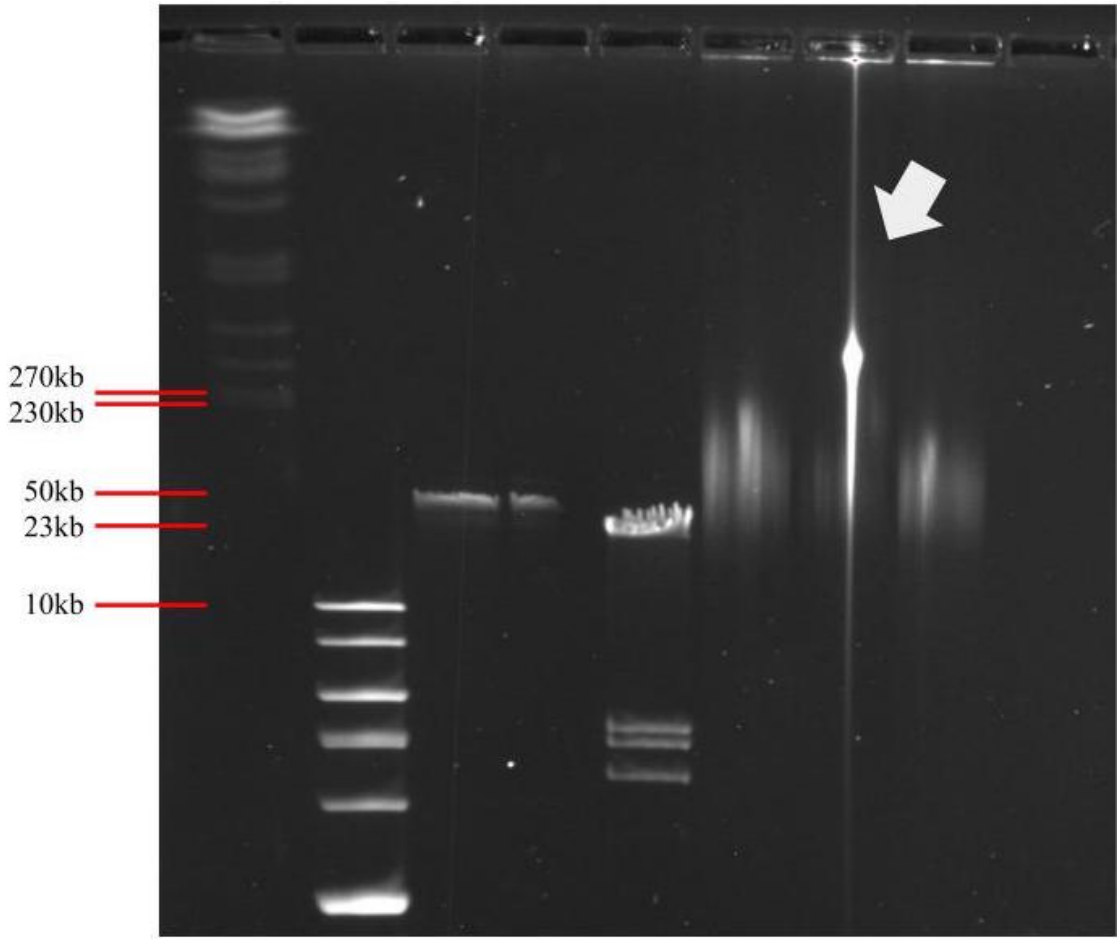

Figure 8: PFGE on a 1x agarose gel run for 18 hours at 160 mV. *The two lowest overlapping bands in lane 1 represent yeast chromosomes of 230 kbp and 270 kbp, respectively. Lane 2 contains 1kb DNA ladder (highest band 10 kbp), lane 3 and 4 the undigested lambda phage (50 kbp) and lane 5 digested lambda (upper band 23 kbp). Lane 7 contains the sample used in the study.*

PFGE quality results were further confirmed by capillary electrophoresis on FEMTO Pulse instrument (AATI, Cat. No. FP-1002-0275) (**Figure 9**). Released in 2016, The FEMTO Pulse is an analytical instrument that uses the same principle of PFGE changing by a 180 degree the direction of the field but in a capillary rather than in an agarose gel, thus allowing Pulsed-Field Capillary Electrophoresis (PFCE). In contrast with PFGE, PFCE only requires approximately one hour to efficiently separate nucleic acids.

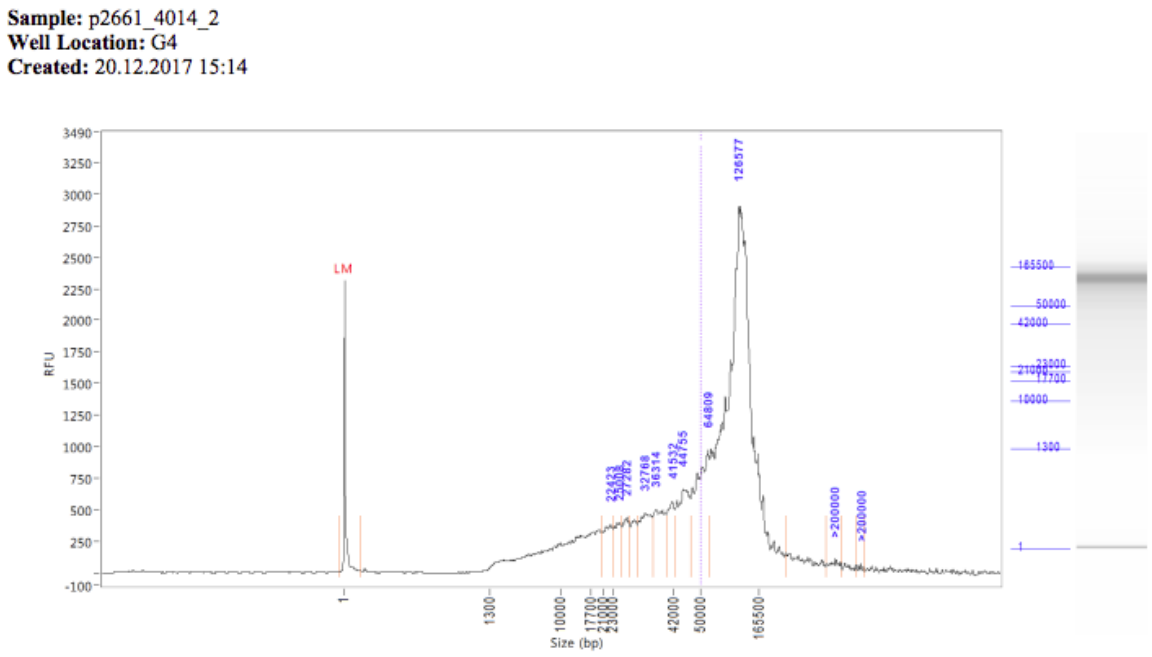

Figure 9: FEMTO Pulse capillary electrophoresis results for the DNA sample used in the study. *The sample modal length of DNA fragments peaked at about 127 kbp, few fragments were above 160 kbp and some sheared DNA between 1 and 50 kbp was present.*

DNA was stored at -80°C and shipped on dry ice to the Functional Genomics Center of Zurich (FGCZ), the core Genomics Facility of the ETH (Zurich, Switzerland)[39].

SMRT library preparation and sequencing

SMRT library preparation and sequencing was performed at the FGCZ. SMRTbell Express Template Prep Kit (Pacific Biosciences, Cat. No. 101-357-000) was used to produce the insert library. Input gDNA concentration was measured on a Qubit Fluorometer dsDNA Broad Range (Life Technologies, Cat. No. 32850). 10µg of gDNA was mechanically sheared to an average size distribution of 40-50 kbp, using a Megaruptor Device (Diagenode, Cat. No. B06010001). FEMTO Pulse capillary electrophoresis was employed to assess the size of the fragments. 5 µg of sheared gDNA was DNA-damage repaired and end-repaired using polishing enzymes. Blunt-end ligation was used to create the SMRTbell template. A Blue Pippin device (Sage Science, Cat. No. BLU0001) was used to size-select the SMRTbell template and enrich for fragments > 30 kbp, excluding the first two cells for which the library was enriched for fragments > 15 kbp. The size-selected library was checked using FEMTO Pulse and quantified on a Qubit Fluorometer. A ready to sequence SMRT bell-Polymerase Complex was created using the Sequel binding kit 2.0 (Pacific Biosciences, Cat. No. 100-862-200). The Pacific Biosciences Sequel instrument was programmed to sequence the library on 18 Sequel SMRT Cells 1M v2 (Pacific Biosciences, Cat. No. 101-008-000), taking one movie of 10 hours per cell, using the Sequel Sequencing Kit 2.1 (Pacific Biosciences, Cat. No. 101-310-

[39] To date (September 2018) there are no PacBio Sequel machines operating in Italy.

400). After the run, sequencing data quality was checked via the PacBio SMRT Link v5.0.1 software using the "run QC module".

Assembly of SMRT reads

The final assembly of long reads was conducted in collaboration with Dr. Matteo Chiara and Prof. David Horner with software CANU v1.7 (Koren et al. 2017) using default parameters except for the "correctedErrorRate" which was set at 0.075. The assembly processes (**Figure 10**) occupied 3,840 CPU hours and 2.2 Tb of RAM for read correction, 768 CPU hours and 1.1 Tb of RAM for the trimming steps, and 3280 CPU hours and 2.2 Tb of RAM for the assembly phase.

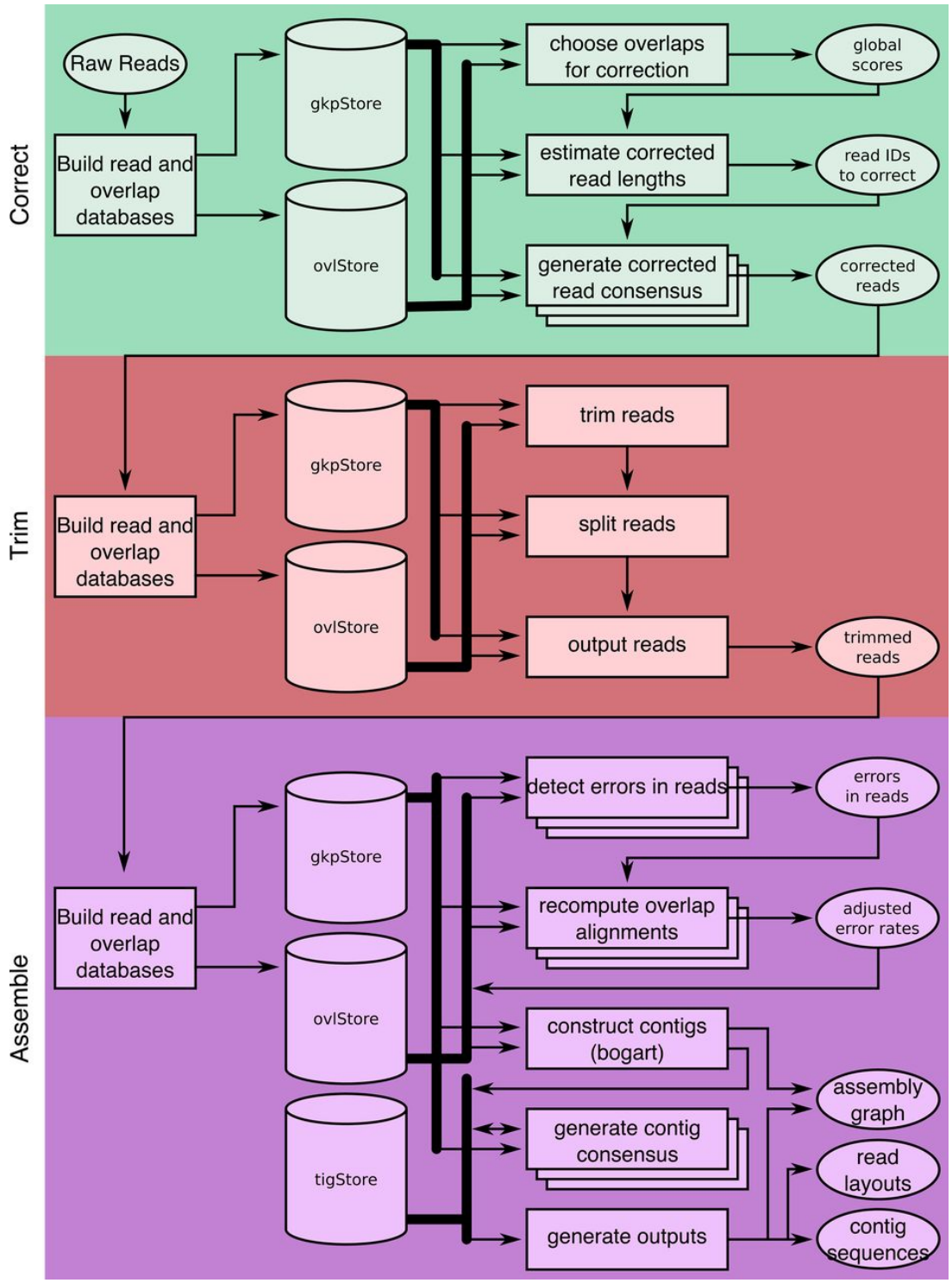

Figure 10: Pipeline for genome assembly using CANU *(Koren et al. 2017)**. The full CANU run includes three stages: correction (green), trimming (red), and assembly (purple). Canu stages share an interface for binary on-disk stores (databases), as well as parallel store construction. In all stages, the first step constructs an indexed store of input sequences, generates a k-mer histogram, constructs an indexed store of all-versus-all overlaps, and collates summary statistics. The correction stage (green) selects the best overlaps to use for correction, estimates corrected read lengths, and generates corrected reads. The trimming stage (red) identifies unsupported regions in the input and trims or splits reads to their longest*

supported range. The assembly stage (purple) makes a final pass to identify sequencing errors; constructs the best overlap graph (BOG); and outputs contigs, an assembly graph, and summary statistics (Koren et al. 2017).

Cell count and DNA extraction for optical mapping

For the production of optical maps[40], DNA above > 200 kbp was extracted at the FGCZ from 7-8 µl of the cell portion from the same blood sample used for SMRT sequencing (following a second shipment). Extraction was performed by Dr. Lucy Poveda sing the Blood and Cell Culture DNA Isolation kit (Bionano Genomics, Cat. No. RE-016-10). This very HMW Extraction was achieved by embedding cells in low melting temperature agarose plugs that were incubated with Proteinase K (Qiagen, Cat. No. 158920) and RNAseA (Qiagen, Cat. No. 158924). The plugs were washed and solubilized using Agarase Enzyme (Thermo Fisher Scientific, Cat. No. EO0461) to release HMW DNA and further purified by drop dialysis. DNA was homogenised overnight prior to quantification using a Qubit Fluorometer.

***In silico* digestion**

The genome assembly obtained with CANU was *in silico* digested using Bionano Access software to test whether the nicking enzyme (Nb.BssSI), with recognition sequence (CACGAG), and the non-nicking enzyme DLE-1, with recognition sequence (CTTAAG), were suitable for optical mapping in our bird genome. An average of 16.9 nicks/100 kbp with a nick-to-nick distance N50 of 11,708 bp were expected for Nb.BssSI, while DLE-1 was found to induce 19.1 nicks/100 kbp with a nick-to-nick distance N50 of 8,775 bp, both in line with manufacturer's requirements.

DNA labeling for optical mapping

For NLRS, DNA was labeled according to manufacturer's instructions using the Prep DNA Labeling Kit-NLRS (Bionano Genomics, Cat. No. 80001). 300 ng of purified genomic DNA was nicked with Nb.BssSI (New England Biolabs, Cat. No. R0681S) in NEB Buffer 3. The nicked DNA was labeled with a fluorescent-dUTP nucleotide analog using Taq DNA polymerase (New England BioLabs, Cat. No. M0267S). After labeling, nicks were ligated with Taq DNA ligase (New England BioLabs, Cat. No. M0208S) in the presence of dNTPs. The backbone of fluorescently labeled DNA was counterstained overnight with YOYO-1 (Bionano Genomics, Cat. No. 80001).

For DLS, DNA was labeled using the Bionano Prep DNA Labeling Kit-DLS (Cat. No. 80005) according to manufacturer's instructions. 750 ng of purified genomic DNA was labeled with DLE labeling Mix and subsequently incubated with Proteinase K (Qiagen, Cat. No. 158920) followed by drop dialysis. After the clean-up step, the DNA was pre-stained, homogenised and quantified using on a Qubit Fluorometer to

[40] As for the SMRT sequencing, the production of optical maps was carried out at the FGCZ since no operating Bionano Saphyr instrument is currently (September 2018) present in Italy.

establish the appropriate amount of backbone stain. The reaction was incubated at room temperature for at least 2 hours.

Generation of optical maps

NLRS and DLS labelled DNA were loaded into a nanochannel array of a Saphyr Chip (Bionano Genomics, Cat. No. FC-030-01) and run by electrophoresis each into a compartment. Linearized DNA molecules were imaged using the Saphyr system and associated software (Bionano Genomics, Cat. No. 90001 and CR-002-01).

Assembly of optical maps

The *de novo* assembly of the optical maps was performed using the Bionano Access v1.2.1 and Bionano Solve v3.2.1 software. The assembly type performed was the "non-haplotype" with "no extend split" and "no cut segdups". Default parameters were adjusted to accommodate the genomic properties of the barn swallow genome. Specifically, given the size of the genome, the minimal length for the molecules to be used in the assembly was reduced to 100 kbp, the "Initial P-value" cut off threshold was adjusted to 1×10^{-10} and the P-value cut off threshold for extension and refinement was set to 1×10^{-11} according to manufacturer's guidelines (default values are 150 kbp, 1×10^{-11} and 1×10^{-12} respectively).

Hybrid scaffolding

Single and dual enzyme Hybrid Scaffolding (HS) was performed using Bionano Access v1.2.1 and Bionano Solve v3.2.1. For the dual enzyme and DLE-1 scaffolding, default settings were used to perform the HS. For Nb.BssSI the "aggressive" settings were used without modification.

Annotation of repeats and genes

Repetitive sequences were soft masked by using software windowmasker (Morgulis et al. 2006) and RepeatMasker (Smit, Hubley, and Green 1996–2010) with defaults parameters. Repetitive sequences models were derived from the Repbase database (Bao, Kojima, and Kohany 2015). *De novo* gene prediction was performed using Augustus (Stanke et al. 2004) with *Gallus gallus* gene models.

BUSCO genes and **synteny with the Chicken genome**

Detection and annotation of single Universal Single-Copy Orthologs (BUSCO) was performed by applying software BUSCO v.3 (Simão et al. 2015) with default parameters. Software Dfenies (Cabanettes and Klopp 2018) was used to align the latest assemblies of the *Gallus gallus* genome with our assembly of the *H. rustica rustica* genome. Synteny was assessed by visual inspection of the final output, in the form of a dotplot graph.

RESULTS

Raw Sequencing Results

The SMRT sequencing effort employed 18 SMRT cells which yielded an average of 3.7 Gbp (standard deviation: 1.7) per SMRT cell (**Figure 11**, top panel). The max read length achieved was above 160 kbp but most SMRT cell produced reads of a maximum of about 80,000 bp (**Figure 11**, middle panel). The first two cells used a conventional 15 kbp library, and the switch to a 30 kbp library did not impact significantly on total Gbp yield nor on max read length, but led to a considerable improvement in average read length (average read length = 14,995 bp; average N50 read length = 25,622 bp) (**Figure 11**, bottom panel). The raw sequence data are available in the GenBank repository upon under Bioproject PRJNA481100, and specifically in Sequence Read Archives SRX4455045 and SRX4455046.

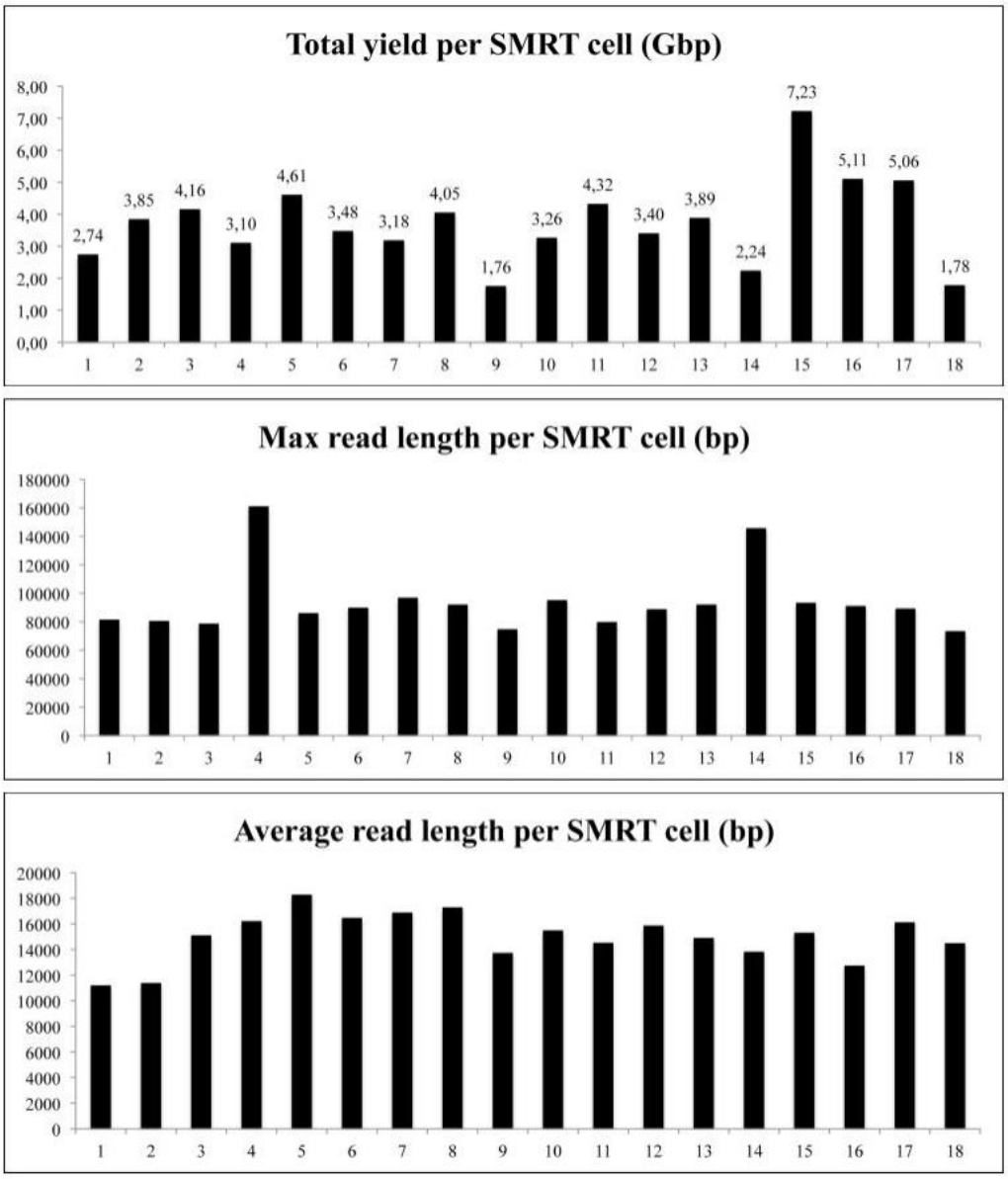

Figure 11: Summary statistics for each SMRT cell employed. *The total yield per SMRT cell (first panel) was variable, with a mean of 3.7 Gbp and a s.d. 1.7 Gbp. Only to SMRT cell achieved read length above 100 kbp (SMRT cells 4 and 14). Average read length (bottom panel) was about 14,000 bp with longest reads above 16,000 bp.*

GC Content Distribution in Sequencing Reads

The GC content distribution of reads was wide (**Figure 12**). This is likely explained by the presence in avian genomes of three classes of chromosomes: macrochromosomes (50-200 Mbp, 5 in chicken), intermediate chromosomes (20-40 Mbp, 5 in chicken) and microchromosomes (12 Mbp on average, 28 in chicken) (Axelsson et al. 2005). These last account for only 18% of the total genome but harbour ~31% of all chicken genes, have higher recombination rates and higher GC contents on average (Kadi et al. 1993).

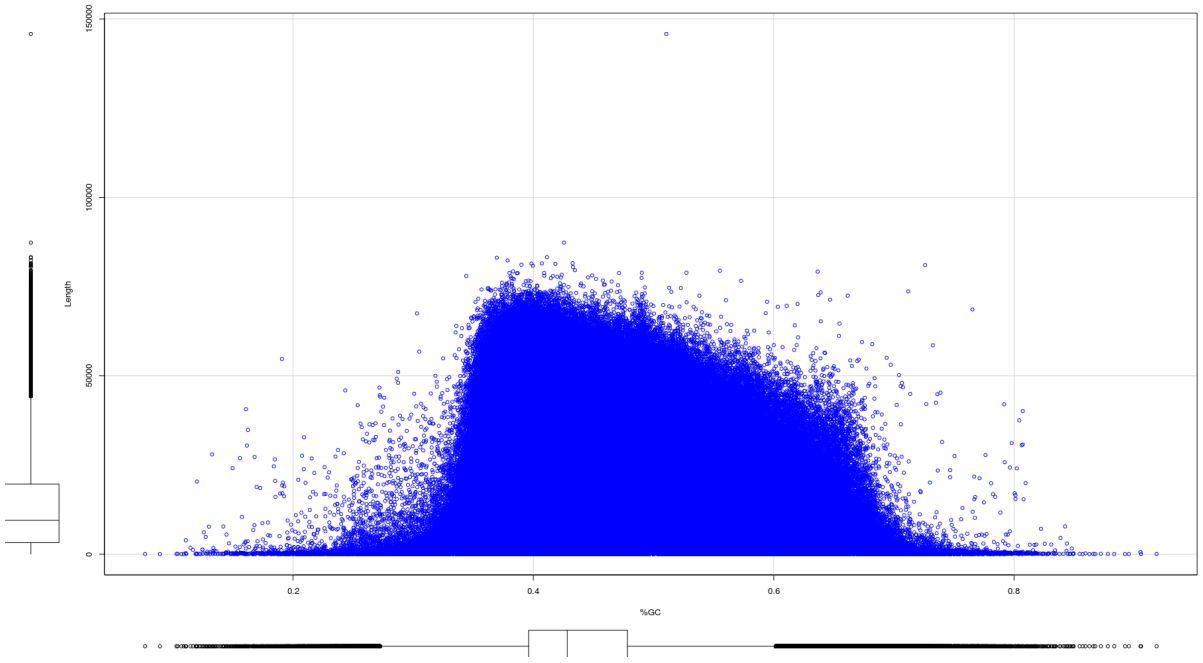

Figure 12: GC content distribution in all sequence reads. *The null hypothesis for the GC content distribution is a normal distribution. However, the GC content distribution is affected by sequencing bias or by the heterogeneity of GC content in genome sequences. SMRT sequencing is known to have a little GC content bias, if any. Moreover, GC content bias usually should be toward lower GC content sequences, that are easier to sequence. Here, the right-skewness of the curve could be explained by the presence of microchromosomes in birds, which are known to have a higher GC content.*

SMRT-based Assembly Summary Statistics

The SMRT-only assembly contained 3,872 contigs with a N50[41] of 5,2 Mbp for a total length of the assembly of 1311.7 Mbp (**Table 1**). Final polishing was performed using the Arrow v2.10 software (Pacific Biosciences) and resulted in final coverage of 45.4x. The raw read error rate could be estimated from the

[41] It should be noted that N50 here indicates N50 of SMRT contigs not that of sequencing reads (i.e. the N50 of consensus sequence derived from assembled reads). N50 contigs represents the length of the i-th contig when contigs are order by their length and their length is summed from the longest contig until the length of the 50% of the genome is reached. Thereby the same can apply to any collection of sequences, from raw sequencing reads to scaffolds or even to the length of optical maps.

final assembly and was similar to that declared by the manufacturer (i.e. around 15%) (Ardui et al. 2018) (**Figure 13**).

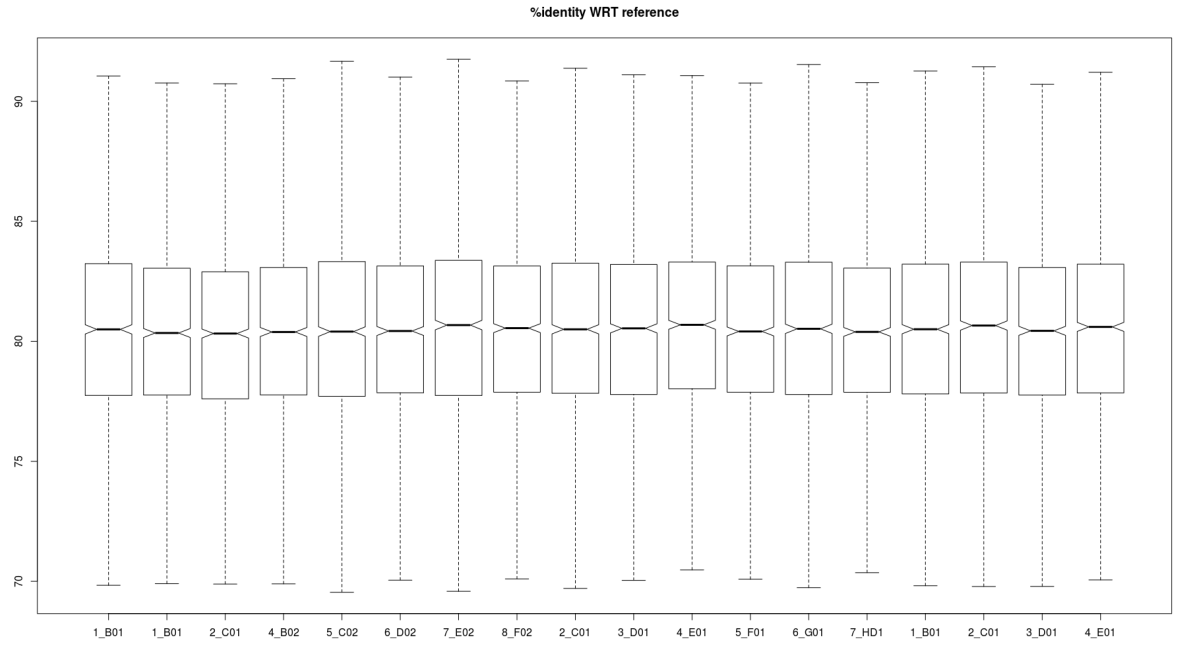

Figure 13: Mean error rate in sequencing reads. *When raw sequencing reads are mapped back to the assembly obtained from the reads the mean error rate was about 20%, with little or no deviation in different SMRT cells.*

Optical Mapping Results

In the experiment with Nb.BssSI, molecule N50 was 0.1298 Mbp for molecules above 20 kbp and 0.2336 Mbp for molecules above 150 kbp - with an average label density of 11.8/100 kbp for molecules above 150 kbp. Map rate was 38.9% for molecules above 150 kbp. Effective coverage was 28.2x. In the experiment with DLE-1, molecule N50 was 0.2475 Mbp for molecules above 20 kbp and 0.3641 Mbp for molecules above 150 kbp - with an average label density of 15.7/100 kbp for molecules above 150 kbp. Map rate was 56.4% for molecules above 150 kbp. Effective coverage was 30.6x. Using both DLE-1 and Nb.BssSI, label metrics were in line with the manufacturer's expectations.

Assembly of Optical Maps

A total of 233,450 (of 530,527) NLRS-labelled molecules (N50 = 0.2012 Mbp) were aligned to produce 2,384 map fragments with an N50 of 0.66 Mbp for a total length of 1338.6 Mbp (coverage = 32x). 108,307 (of 229,267) DLE-1 labelled input DNA molecules with a N50 of 0.3228 Mbp (theoretical coverage of the reference 48x) produced 555 maps with a N50 length of 12.1 Mbp for a total length 1299.3 Mbp (coverage = 23x).

Hybrid Scaffolding Results

The NLRS hybrid scaffold had an N50 of 8.3 Mbp (scaffold only N50 = 10.8 Mbp) for a total length of 1,338.6 Mbp (total length of scaffolded contigs = 1,175.3 Mbp) and consisted of 409 scaffolds and 2,899 un-scaffolded contigs. The DLS hybrid scaffold had N50 of 17.3 Mbp (scaffold only N50 = 25.9 Mbp) for a total length of 1,340.2 Mbp (total length of scaffolded contigs = 1,148.4 Mbp) and consisted of 211 scaffolds and 3,106 un-scaffolded contigs.

Dual Enzyme Hybrid Scaffolding Results

Dual enzyme HS (incorporating both DLS and NLRS maps) resulted in an assembly with N50 of 23.8 Mbp (scaffold only N50 = 28.4 Mbp) for a total length of 1,351.8 Mbp (total length of scaffolded contigs = 1,208.8 Mbp) and consisted of 273 scaffolds and 2,810 un-scaffolded contigs. During the automatic conflict resolution in the dual enzyme HS, 185 SMRT contigs were cut, as Bionano maps confidently indicated mis-assemblies of the SMRT reads. Conversely, 117 Bionano maps were cut indicating that the chimeric score did not provide sufficient confidence to cut the assembly based on SMRT contigs[42]. Of 3,872 SMRT contigs, 1,243 (32%) were anchored in the Bionano maps (of which 990 were anchored in both DLS and NLRS maps). 56 and 226 were anchored in DLS and NLRS maps respectively. 2,810 maps could not be anchored at all.

Assembly Results After Removal of Haplotigs

Notably, all hybrid assemblies were somewhat larger than the expected genome size, and in all cases, the N50 of un-scaffolded contigs was extremely low (0.06 Mbp for the dual enzyme hybrid assembly). Here, the most likely explanation is that a significant proportion of these small contigs might represent divergent homologous haplotigs that were assembled independently. Similarity searches were consistent with this possibility as almost 95% of the contigs that were not scaffolded in the dual enzyme hybrid assembly showed > 98% identity to scaffolded contigs over 75% of their length or more (**Figure 14**).

[42] Next to creating long and contiguous scaffolds, the Hybrid Scaffold pipeline also detects and resolves chimeric joins present in either input assembly (NGS or Bionanomaps). Chimeric joins may be formed when short reads, molecules, or paired-end inserts are unable to span across long DNA repeats. These errors would appear as conflicting junctions in the alignment between the two assemblies. Upon the detection of a conflict, the pipeline uses Bionano's long native molecules to determine which assembly has been likely constructed incorrectly. If the genome map does not have long molecule support at the conflict junction, then the map is cut, thus removing the putative chimeric join. If it does have molecule support, the sequence fragment is cut.

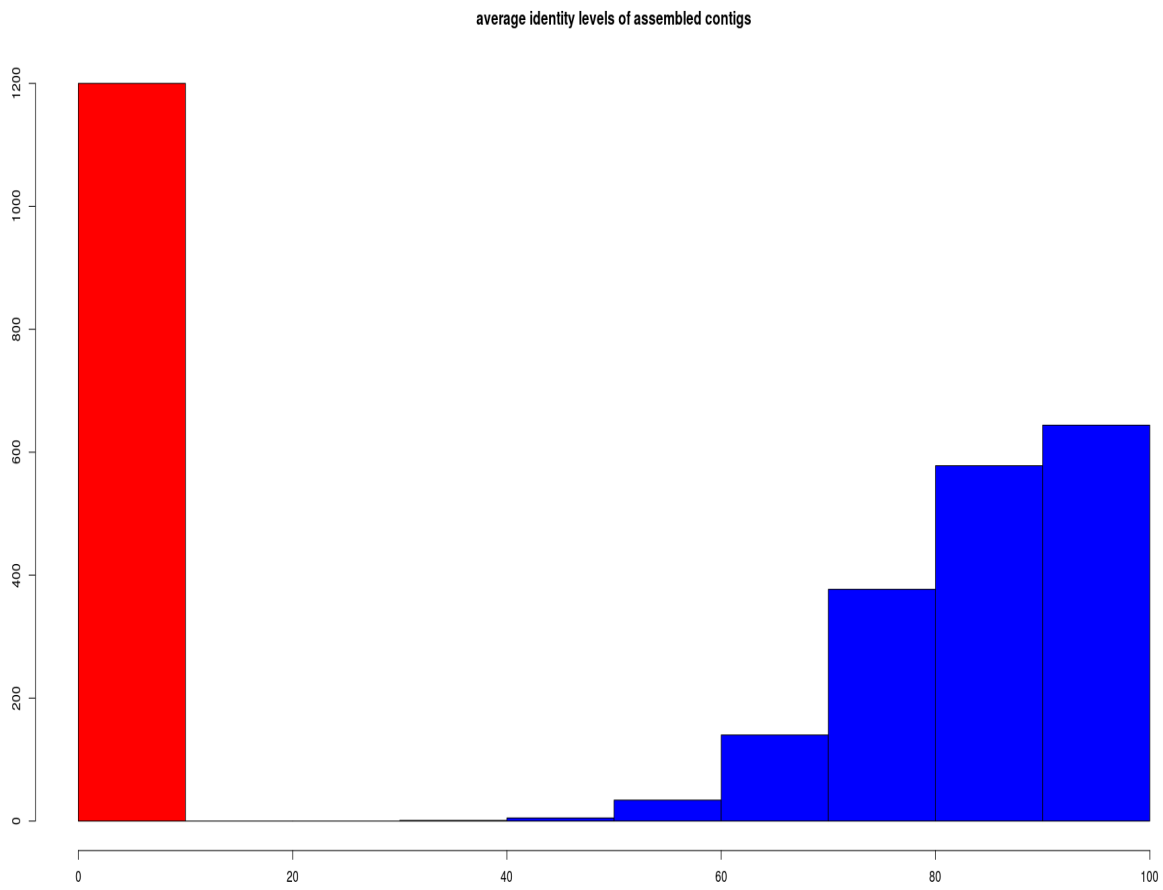

Figure 14: Identification of haplotigs and removal of unscaffolded contigs by identity to scaffolded contigs. *Unscaffolded contigs with very high identity (blue) - putative haplotigs - were clearly separated from contigs with low identity (red) based on similarity searches.*

These contigs were discarded, resulting in a final assembly (**Table 1,** and **Supplementary Table 1** in **Appendix 1** for detailed statistics) of 1.21 Gbp (N50 = 25.9 Mbp) made up of 273 dual enzyme hybrid scaffolds (N50 = 28.42 Mbp) and 91 unscaffolded contigs (N50 = 0.0644 Mbp).

	(Safran et al. 2016)[1]	**SMRT contigs[2]**	**Final assembly[3]**
Species	*H. r. erythrogaster*	*H. r. rustica*	
Starting raw data (Gbp)	61.7	66.4	59.6
N50 (bp)	38,844	5,189,284	25,954,216
N90 (bp)	3,718	85,340	2,002,624
Total size (Gbp)	1.1	1.31	1.21
Theoretical genome coverage*	47x	52x	47x

% genome coverage*	85.9	102.6	94.5
# of contigs/scaffolds	100,153	3,872	364
Avg contig/scaffold length (bp)	11,010	338,782	3,334,461
Longest contig/scaffold (bp)	732,517	33,230,000	98,053,015

Table 1: Assembly metrics for contigs and final scaffolds in our European barn swallow genome compared to the published American barn swallow genome. *[1] Illumina PE reads assembled using SOAPdenovo v2.04 (R. Li et al. 2010). [2] SMRT reads assembled using CANU v1.7 (Koren et al. 2017). [3] SMRT contigs assembled with CANU and scaffolded using Bionano dual enzyme HS, with haplotigs removed as detailed in the text. *Based on a barn swallow genome size estimate of 1.28 Gbp (Andrews, Mackenzie, and Gregory 2009).*

Coverage of the Final Assembly

The average read SMRT read coverage for the genome assembly was 34.15X, implying a theoretical QV of over 40 (**Figure 15**).

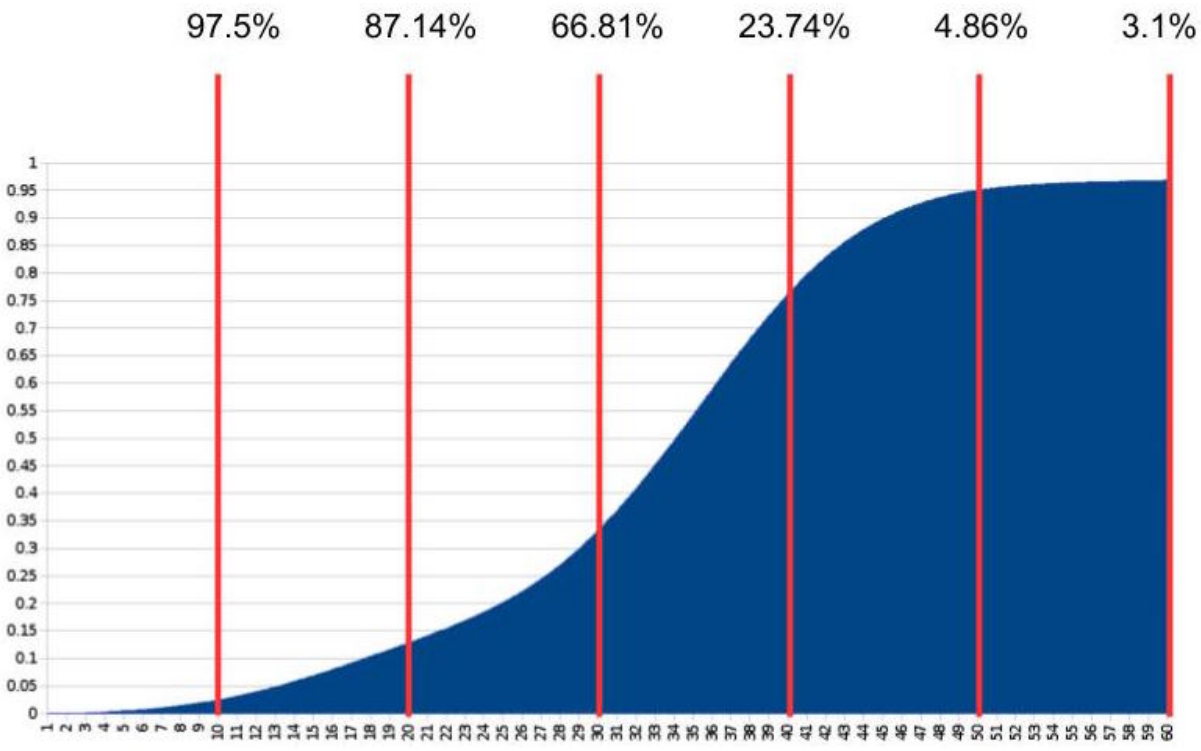

Figure 15: Cumulative coverage depth distribution observed in the final (de-haplotyped) assembly of the barn swallow genome. *Coverage is indicated on the X axis. Red lines are used to display the proportion of the genome covered by more than 10, 20, 30, 40, 50 or 60 reads respectively.*

Genome Annotation

7.11% of the final assembly was annotated as repetitive using RepeatMasker (Smit, Hubley, and Green 1996–2010), with the major contributions deriving from L2/CR1/Rex LINE elements (3.37%), retroviral LTRs (1.59%) and simple repeats (1.56%).

In all, 35,644 protein coding genes were predicted, of which 9,189 were overlapped by more than 30% of their size with repetitive genomic elements. Of the remaining 26,455 predicted protein coding genes, 24,331 harboured a PFAM protein domain. Simple similarity searches based on blastp (Altschul et al. 1990) (with default parameters) suggested that 17,895 of the predicted protein coding genes have a best reciprocal blast hit with gene models derived from *Gallus gallus* GRCg6a assembly[43], while 2,927 of the proteins predicted by Augustus did not show any significant match (e-value <= 1 x 10^{-15}, identity > 35%).

BUSCO Genes in the Final Genome Assembly

Of a total of 4915 conserved bird Benchmarking with Universal Single-Copy Orthologs (BUSCO) groups (Simão et al. 2015) sought, 4,598 (93.6%) were complete, 4,521 (92.0%) were complete and single-copy, 77 (1.6%) were complete and duplicated, 192 (3.9%) were fragmented and 125 (2.5%) were missing.

Synteny Map

The alignment of the final assembly with the published chromosome-level assembly of the chicken (*Gallus gallus*) genome GRCg6a using D-Genies (Cabanettes and Klopp 2018) indicates high levels of collinearity between these two genomes with a limited number of intra-chromosomal rearrangements (**Figure 16**). In particular, the three largest chromosomes (1-3) appear to have been assembled in less than 10 scaffolds in our genome assembly, while the other chromosomes have been assembled in less than 5 scaffolds (**Figure 17**). For instance, over 50% of chromosome 1 - the largest chicken chromosome - was represented by a single scaffold, and 90% of it was represented by as little as 5 scaffolds (**Figure 18**). These results were mirrored also in the other large chromosomes (data not shown). Only the Z chromosome appears highly fragmented, potentially as a consequence of having sequence the heterogametic sex (ZW) and therefore have halved coverage for this chromosome.

[43] Available from: https://www.ncbi.nlm.nih.gov/genome/proteins/111?genome_assembly_id=374862

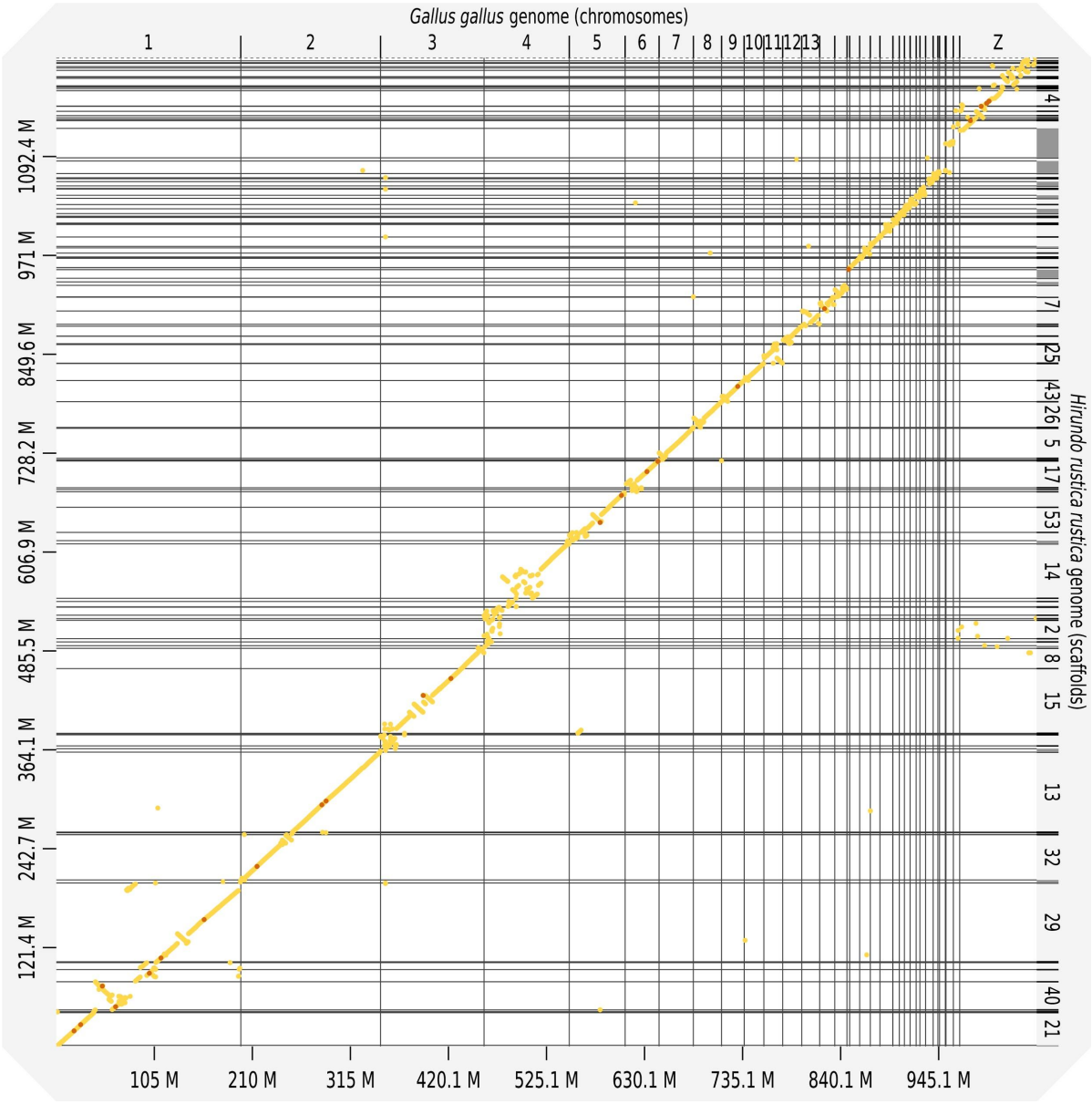

Figure 16: Alignment of the final assembly with the published chromosome-level assembly of the chicken (*Gallus gallus*) genome GRCg6a. *Light to dark yellow dots indicate progressively higher similarity between sequences. Alignment performed using D-Genies (Cabanettes and Klopp 2018).*

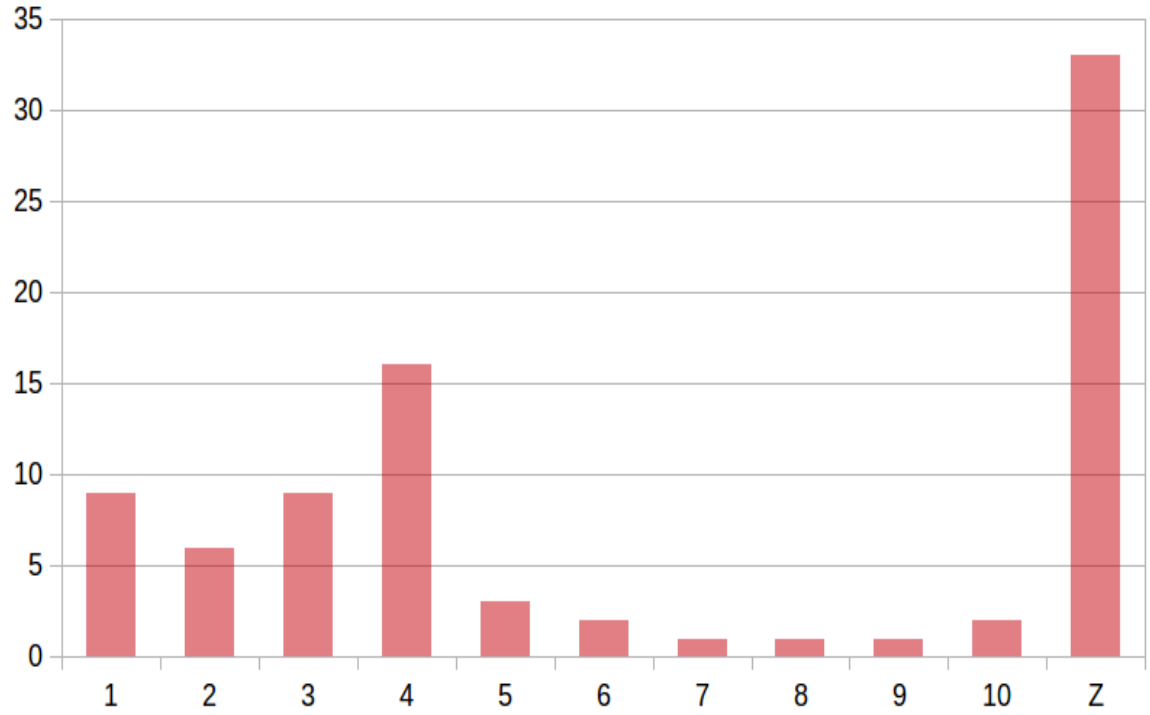

Figure 17: Number of scaffolds per chicken chromosomes (GRCg6a). *Chicken chromosomes on the x axis and scaffold counts on the y axis. The largest chicken chromosome (chr 1) in our genome assembly is represented by less than 10 scaffolds and the second largest by 6 scaffolds. The Z chromosome is assembled in over 30 scaffolds, potentially as a consequence of the halved coverage for this chromosome due to the sequencing of a female individual.*

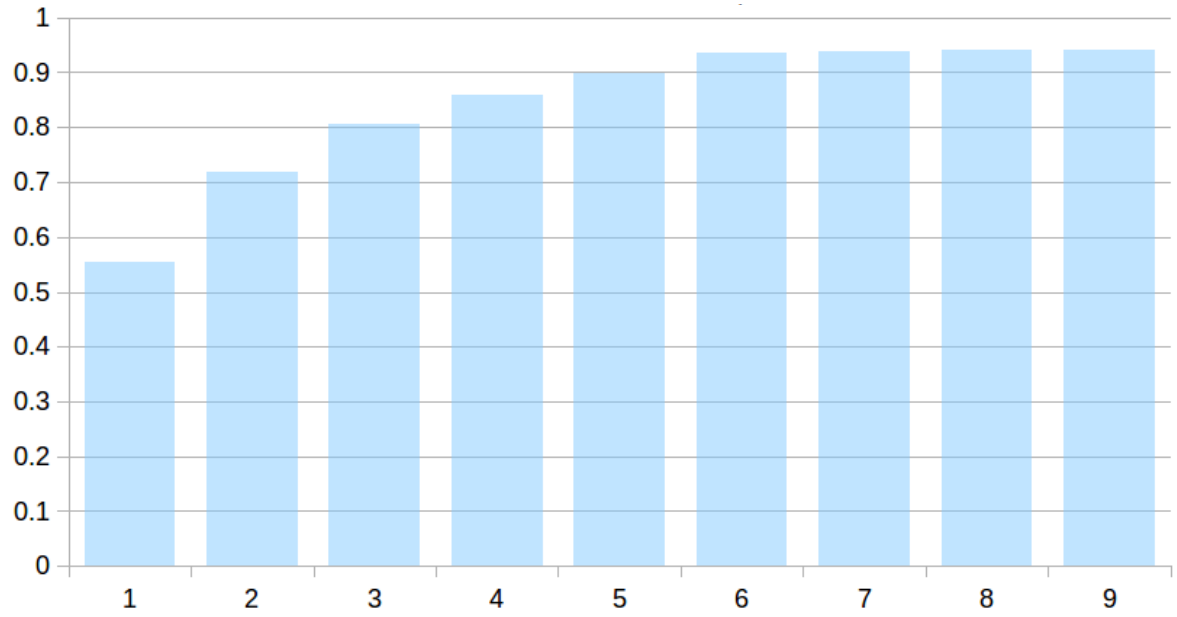

Figure 18: Cumulative scaffold size for chromosome 1. *The graph shows that over 50% of chromosome 1 in chicken is represented by a single scaffold in our genome assembly. Five scaffolds cover 90% of the sequence. Similar results were obtained for the other large chromosomes.*

DISCUSSION

The combination of long reads and optical maps has already proven invaluable to produce high-quality genome assemblies, even in the case of particularly complex genomes (Nowoshilow et al. 2018). In the present work, using only SMRT sequencing and Bionano optical maps our collaborators and I have produced a high-quality and contiguous genome for the European barn swallow. With respect to a previously reported SGS-based assembly of the American barn swallow genome using a comparable amount of raw data (Safran et al. 2016), even the contigs generated from long-read sequencing alone show an 134-fold increase in N50 (**Appendix 1**, **Supplementary Table 1**). Moreover, in terms of N50, number and average length, these contigs are similar to those recently obtained for the Anna's hummingbird (*Calypte anna*) and the Zebra Finch (*Taeniopygia guttata*) genomes using SMRT sequencing (Korlach et al. 2017).

As an alternative to scaffolding with long insert mate-pairs (Hunt et al. 2014) or to chromatin proximity ligation sequencing (Burton et al. 2013), contiguity and accuracy of long-read-based assemblies can be further improved by optical mapping. Here, the fold change in N50 attained by Bionano NLRS hybrid scaffolding of the European barn swallow genome (1.6 fold before removal of haplotigs) is comparable with results obtained by other genome assemblies that have employed this method (Gao et al. 2018). Strikingly, the new DLS method greatly outperformed the NLRS system, providing a 3.3 fold increase of N50 (before removal of haplotigs). Moreover, incorporation of both labelling systems into the hybrid scaffolding yielded a final assembly showing 5-fold improvement of the N50 with respect to the original SMRT assembly, simultaneously providing "independent" validation of many scaffold junctions. It should be noted that the presence of numerous microchromosomes in avian genomes restricts the final N50 value potentially attainable for the assembly, as the fully assembled karyotype would have an N50 of ~ 90 Mbp. Yet, after removal of putative haplotigs, this genome assembly contiguity metrics meet the high standards of the VGP consortium "Platinum Genome" criteria (contig N50 in excess of 1 Mbp and scaffold N50 above 10 Mbp) (Lewin et al. 2018).

The percentage of contiguously assembled BUSCO genes is consistent with recent results with Anna's Hummingbird and the Zebra Finch (Korlach et al. 2017) (**Figure 19**). However, 40 of the "missing" bird BUSCO genes are absent from at least 2 of the 54 available avian genome sequences, suggesting that, despite the potentially incomplete nature of some draft genomes, even some BUSCO genes may not be universally conserved among birds.

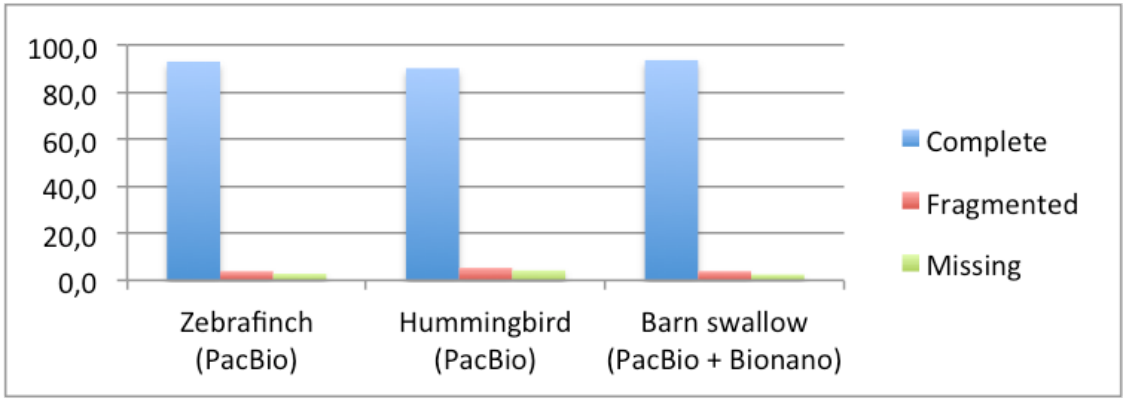

Figure 19: Comparison of BUSCO analysis results. *Results of BUSCO gene analysis in our genome assembly are in line with results in the recent assemblies of the Anna's hummingbird (*Calypte anna*) and the Zebra Finch (*Taeniopygia guttata*) genomes (Korlach et al. 2017)*, with slightly better results in terms of complete and fragmented genes in our assembly.

The high level of collinearity between independently assembled and scaffolded sequences for the barn swallow and the chicken genomes provides circumstantial support for the quality of both the contigs and the hybrid scaffolds and is consistent with previous observations of high levels of synteny and minimal inter-chromosomal rearrangements among birds (Ellegren 2010).

The final assembly is slightly smaller than the previously estimated genome size (1.28 Gbp) (Andrews, Mackenzie, and Gregory 2009), possibly reflecting an imprecise older estimate and/or the possibility that some poorly assembled repeats were discarded in the final step described above.

Given the inception of large scale sequencing initiatives aiming to produce genome assemblies for a wide range of organisms (Koepfli et al. 2015; G. Zhang et al. 2015; Pennisi 2017; Teeling et al. 2018), it is critical to identify combinations of sequencing and scaffolding approaches that allow the cost effective generation of genuinely high-quality genome assemblies (Lewin et al. 2018). While SGS technologies have allowed the production of cost-effective genome drafts for many birds and other vertebrate species (Genome 10K Community of Scientists 2009; G. Zhang et al. 2014; Jarvis et al. 2014), the reduction in genome sequencing costs has typically come at the price of compromises in contiguity and accuracy of assemblies with respect to earlier efforts based on Sanger reads and extensive physical mapping (Henson, Tischler, and Ning 2012). Many limitations of SGS-based assemblies stem from the occurrence of long sequence repeats. In many animal species, transposons are frequently located in introns (Sela, Kim, and Ast 2010) and the presence of large gene families of closely related paralogs can lead to the existence of long "genic" repeats. Accordingly, even apparently contiguous genic regions can feature juxtaposition of paralogous gene fragments (Korlach et al. 2017). With respect to mammals, avian genomes generally contain relatively low proportions of repetitive sequences (Ellegren 2010) and this appears to be the case also for the barn swallow genome.
While exhibiting higher rates of single-base errors than some SGS approaches, TGS technologies, including SMRT sequencing, offer read-lengths unparalleled by SGS or Sanger sequencing (Bleidorn 2016). Moreover, recent and on-going improvements in TGS methods are rapidly reducing the "per-base" cost of TGS data compared to that of SGS.

Usage of the barn swallow genome assembly and future perspectives

The genomic era is still in its infancy, and new sequencing approaches are unceasingly developed and proposed. However, frontier approaches and technologies consisting of high-throughput short- and long-read sequencing, and optical mapping now enable to tackle some of the major long-standing or emerging issues in biology in an unprecedentedly deep, quantitative and cost-effective way (Ellegren 2014). The potential of

these approaches was so far constrained by the lack of genome data on non-model organisms and was therefore exploited only to a very limited extent in ecological and evolutionary studies. In the recent past, genomic projects mostly focused on model species (e.g. mice, *Drosophila melanogaster*) as manageable representatives of the complexity of life. While model organisms are an invaluable tool for dissecting many fundamental biological issues, it is now clear that they do not represent the diversity of even their close relatives, let alone biological diversity as a whole and several studies have already demonstrated that non-model species allow researchers to address a wide spectrum of diverse key biological questions (Dill 2002). How much and which type of genetic variation is present in wild populations? What is the genetic causation and architecture of complex quantitative traits? How does natural selection affect underlying genetic variation? How do demography and population divergence shape genetic variation? The data presented here, as well as attesting to the effectiveness of SMRT sequencing combined with DLS optical mapping for the assembly of vertebrate genomes, will represent an invaluable asset for population genetics and genomics in the barn swallow as well as for comparative genomics in birds. Immediate future progress on the barn swallow genome include the phasing of the assembly to generate extended haplotypes, a more thorough gene annotation using RNA/IsoSeq sequencing data, a detailed comparison with genome data from *Hirundo rustica erythrogaster* genome sequencing and assembly project, the characterization of the epigenetic landscape and a re-evaluation of data from previous population genetics studies conducted in this species. However, the availability of a high-quality genome for the barn swallow is intended to be the starting point for a broad population genomics studies where this genome will serve as a reference. In 2001, the completion of the human genome (Venter et al. 2001; Lander et al. 2001) paved the way to human population genomics. The first effort in this direction was established in 2007 with the 1000 Genomes Project (1KGP), which has subsequently generated the most comprehensive and open catalogue of phased genetic variants for the human genome[44] (Birney and Soranzo 2015; Zheng-Bradley and Flicek 2017). Between 2008 and 2015, 1KGP has generated the largest public catalogue of human variation and genotype data, ultimately including 2,504 people from five continental regions (Birney and Soranzo 2015). These data allowed thousands of genetic studies on human populations (Birney and Soranzo 2015; Zheng-Bradley and Flicek 2017). They have been extensively employed in many GWAS for the imputation of previously overlooked and new causal variants (Bamshad et al. 2011; Khurana et al. 2013; Bowles et al. 2015). Combined with RNA-seq data and functional annotations of regulatory elements, human 1KGP data allowed mapping of expressed QT *loci* (eQTLs) and the study of their transcriptional regulation (Pickrell et al. 2010; Lappalainen et al. 2013). The analysis of the frequency of variants from re-sequencing studies allowed the identification of genomic regions under selective pressure in coding (Fu and Akey 2013), non-coding (Jha, Lu, and Xu 2015) and in regulatory regions (Ward and Kellis 2012). Genome scans for advantageous variants quickly fixated in the population allowed the detection of signatures of natural selection (Grossman et al. 2013; Fagny et al. 2014). Furthermore, data from 1KGP produced numerous insights into human origins (1000 Genomes Project Consortium et al. 2010, 2012, 2015), demography (Gravel et al. 2011; K.

[44] 1000 Genomes | A Deep Catalog of Human Genetic Variation. http://www.internationalgenome.org/

Harris and Nielsen 2013; Schiffels and Durbin 2014) and migration (Gravel et al. 2013). However, no comparable effort has been so far carried out in other species, particularly in any non-model species. Current development in sequencing technology and associated drop in sequencing costs now enable to design and conduct population-level whole-genome sequencing projects capable of providing extraordinary amounts of high-quality data for virtually any wild animal species. The availability of abundant, unbiased and accurate WGS data allows to perform a wide range of sophisticated statistical analyses (Ellegren 2014). These include precise determination of classical genetic metrics of individuals and populations as well as several newly-developed and extremely powerful methods that can greatly contribute to provide a detailed view of the natural history and biology of the species under study. Mirroring human studies, three major expanding research areas include the identification of the genes responsible for life history traits; and the assessment of selective sweeps that have left traces within the genomes along evolution; and the study of demography, especially in the light of the generalized decline of natural populations. The discovery of the genetic determinants of traits associated with fitness variation in the wild (trait mapping), and the dissection of the patterns of selection that act on them (selective sweep mapping), are fundamental to unveil the mechanisms of evolution. Whether a population can respond to ever-occurring environmental change, including human-driven change in ecological conditions, ultimately depends on phenotypic plasticity in traits relevant to the response to the environmental effects and the amount of standing genetic variation at those traits. In the past, studies focussing on the genetic architecture of quantitative traits (QT) have regarded populations with known pedigree relationships, limited population size and low heterozygosity (Slate et al. 2010; Schielzeth and Husby 2014). Still, association mapping revealed a high number of *loci* influencing QT (Visscher et al. 2017). More recently, larger sets of markers from wider populations have enabled the first Genome-Wide Association Studies (GWAS) in wild populations (M. R. Robinson et al. 2013; Husby et al. 2015; Scordato et al. 2017; Narum et al. 2018). Nonetheless, our knowledge of the genetics of phenotypic traits in non-model organisms in the wild is close to *nihil* (M. R. Robinson et al. 2013). This is true for virtually all sort of traits encompassing morphology, physiology and behaviour. Which are the genes involved in standing phenotypic variation? Are traits influenced by few genes with relatively large effect (oligogenicity) or by several small-effect genes (polygenicity)? What is the relative contribution of SNPs, indels and structural variants to standing phenotypic variation? Are GWAS findings reproducible among wild populations? These are just some of the major questions that still remain largely unanswered. The identification of the causative genes of phenotypic variability, especially for traits that are key for the conservation biology of this species such as life history, phenological, morphological, behavioural, physiological and parasitological (including the microbiota) traits that are known to be major fitness components currently under natural and/or sexual selection will guarantee detailed knowledge of the biology of the study populations. In the barn swallow, this effort will be greatly facilitated by the presence of large repositories of existing blood samples as a source of DNA and matched phenotypic information.

Another open research route involves the analysis of genetic variation along the entire genome to conducting genome scans to detect traces of recent selective pressures that have driven the evolution at specific *loci* to

understand how selection shapes trait variability (Josephs, Stinchcombe, and Wright 2017). Indeed, the heritable component of phenotypic variation is the target of natural and sexual selection and thus genetic variation represents the raw material for adaptation (Fisher 1930). Therefore, the signatures of recent selective processes can be often traced within the patterns of genetic diversity at the genome scale. However, these analyses have so far been constrained by the lack of empirical data. The increasing availability of genome sequences from non-model species provides the unprecedented opportunity to identify the genomic regions evolving under natural selection (Ellegren 2014). In particular, genome scans can unveil those regions of the genome that have been subject to recent selection without the need to *a priori* define traits that are putatively important to adaptation, ultimately deepening our understanding of the genetic architecture behind adaptive evolution (Ellegren 2014). A popular approach, particularly in NMS, is Fixation Index (FST) outlier analysis (Beaumont and Nichols 1996) where FST for individual *loci* is compared with a neutral model. Candidate *loci* for positive or divergent selection are expected to have a high FST (i.e. high levels of differentiation between populations), whereas low FST is associated *loci* with under balancing selection. Other analyses require measuring haplotype length and diversity to spot extended blocks of LD, a general reduction of nucleotide diversity or runs of homozygosity in specific genomic regions (Zheng-Bradley and Flicek 2017). Selective sweeps, whereby an advantageous variant is selected and drags neutral surrounding variants in LD to unexpectedly high frequencies as the piggyback with the advantageous variant generate long blocks of haplotypes of low diversity, (J. M. Smith and Haigh 1974; Sabeti et al. 2007; Tishkoff et al. 2007). Moreover, long stretches of consecutive homozygous genotypes often found in the genome can be regarded as the product of strong selective pressures determining local reduction of genetic diversity (Metzger et al. 2015).

Finally, the individual-based, long-term study of wild populations exposed to natural environmental conditions provides essential insights to the processes that govern population divergence, potentially allowing to predict the evolutionary destiny of populations. Patterns of genetic diversity within and between populations are shaped by demography, divergence and the extent of reproductive isolation (Ellegren 2014). All individuals and populations of a species are related to each other through common ancestry. Population genomics allows accurate reconstruction of current and past population size and structure, revealing the underlying complex demographic scenarios and assessing the effect of gene flow and introgression (Schiffels and Durbin 2014). In the absence of recombination and selection, their time of divergence can be directly estimated from the total number of differences between sequences (Crow and Kimura 1970). This information in turn can be used to infer the time of ancestral genetic separations, effective population sizes (N_e) over the generations and demographic events (e.g. bottlenecks). Today, one of the standard approaches for past demographic inference from genomic data is Allele Frequency Spectrum (AFS) analysis. In particular, Joint Allele Frequency Spectrum (JAFS), which represents the joint distribution of allele frequencies across two or more related populations, has been a commonly employed method for inferring demographic models for many species (Gutenkunst et al. 2009). Studying the demography and population divergence of declining barn swallow populations in conjunction with the ecological challenges that they are

facing, potentially allowing to dissect the relative impact of global environmental changes on species microevolution. On the other hand, speciation is a fundamental process responsible for the diversity of life. Progress has been made in detecting individual 'speciation genes' that cause reproductive isolation. In contrast, until recently, less attention has been given to genome-wide patterns of divergence during speciation. Thus, major questions remain concerning how individual speciation genes are arrayed within the genome, and how this affects speciation.

In conclusion, to all the aforementioned studies and goals, the establishment of a reference genome for the barn swallow achieved with the present work will represent the first, essential milestone.

Appendix 1

Comparison of barn swallow genome assemblies

	Safran	SMRT contigs	SMRT vs Safran	HS1 (BssSI)		HS1 vs SMRT	HS2 (DLE-1)		HS2 vs HS1	DE-HS		DE-HS vs HS2	Final	Final vs DE-HS	Final vs Safran
Strategy	**Illumina 101 PE**	**PacBio Sequel**	**Fold-change (x)**	**Bionano Saphyr**		**Fold-change (x)**	**Bionano Saphyr**		**Fold-change (x)**	**Bionano Saphyr**		**Fold-change (x)**	**No haplotigs**	**Fold-change (x)**	**Fold-change (x)**
Starting raw data (Gbp)	61,7	66,4	1,1		59,6	0,9		59,6	1,0		59,6	1,0	59,6	1,0	0,97
N50 (Mbp)	0,039	5,189	134	8,327	10,789	1,6	17,321	25,954	2,1	23,843	28,420	1,38	25,954	1,09	668
N90 (Mbp)	0,0037180	0,093	25	0,106	1,15	1,1	0,094	2,336	0,89	0,11	2,12	1,22	2,00	17,44	538
Total size (Gbp)	1,1	1,312	1,2	1,34	1,18	1,02	1,34	1,15	1,001	1,35	1,21	1,35	1,21	0,90	1,1
Theroretical genome coverage*	47	51,88	1,11		46,56	0,9		46,56	1,0		46,56	1,0	46,56	1,0	0,99
% genome coverage*	85,9	102,5	1,2	104,6	91,8	1,02	104,7	89,7	1,001	105,6	94,4	1,009	94,53	0,90	1,10
# of contigs/scaffolds	100153	3872	26	3308	409	1,2	3317	211	0,997	3083	273	1,08	364	8,47	275
Avg contig/scaffold length (Mbp)	0,011	0,339	31	0,405	2,874	1,2	0,404	5,442	0,998	0,438	4,428	1,09	3,334	7,60	303
Longest contig/scaffold (Mbp)	0,733	33,230	45	39,193	39,193	1,2	109,491	109,491	2,8	98,053	98,053	0,90	98,053	1,00	134
Expected # of genes per contig□	1	26	31	31	219	1,0	31	414	1,0	33,4	336,8	1,09	253,6	7,60	302

* Barn swallow estimated genome size: 1.28 Gbp (Andrews et al. 2009)

□ Estimate based on average protein coding gene size for latest available annotation of the *Gallus gallus* genome (GRCg6a) = 13146 bp

Safran = Safran et al. 2016
SMRT contigs = CANU assembly of SMRT long-reads
HS1 = Hybrid Scaffold with BssSI
HS2 = Hybrid Scaffold with DLE-1
DE-HS = Dual Enzyme Hybrid Scaffold with BssSI and DLE-1
Final = Final assembly from DE-HS after removal of haplotigs

For all HS metrics including (column 1) and excluding (column 2) unscaffolded contigs are reported

Supplementary Table 1: Assembly metrics comparison for contigs and scaffolds between different assemblies.

Appendix 2

BioRxiv version of July 23, 2018.

SMRT long-read sequencing and Direct Label and Stain optical maps allow the generation of a high-quality genome assembly for the European barn swallow (*Hirundo rustica rustica*)

Giulio Formenti* (giulio.formenti@unimi.it), Department of Environmental Science and Policy, University of Milan (Milan, Italy).

Matteo Chiara* (matteo.chiara@unimi.it), Department of Biosciences, University of Milan (Milan, Italy).

Lucy Poveda (lucy.poveda@fgcz.uzh.ch), Functional Genomics Center of Zurich, University of Zurich, (Zurich, Switzerland).

Kees-Jan Francoijs (kfrancoijs@bionanogenomics.com), Bionano Genomics (San Diego, CA, USA).

Andrea Bonisoli-Alquati (aalquati@cpp.edu), Department of Biological Sciences, California State Polytechnic University (Pomona, CA, USA).

Luca Canova (canova@unipv.it), Department of Biochemistry, University of Pavia (Pavia, Italy).

Luca Gianfranceschi (luca.gianfranceschi@unimi.it), Department of Biosciences, University of Milan (Milan, Italy).

David Stephen Horner (david.horner@unimi.it), Department of Biosciences, University of Milan (Milan, Italy).

Nicola Saino (nicola.saino@unimi.it), Department of Environmental Science and Policy, University of Milan (Milan, Italy).

*These authors contributed equally to the work.

ABSTRACT

Background:
The barn swallow (*Hirundo rustica*) is a migratory bird that has been the focus of a large number of ecological, behavioural and genetic studies. To facilitate further population genetics and genomic studies, here we present a high-quality genome for the European subspecies (*Hirundo rustica rustica*).

Findings:
We have assembled a highly contiguous genome sequence using Single Molecule Real-Time (SMRT) DNA sequencing and Bionano optical maps. We compared and integrated optical maps derived both from the Nick, Label, Repair and Stain and from the Direct Label and Stain technologies. For our SMRT-only assembly, the direct labelling system more than doubled the assembly N50 with respect to the nickase system. The dual enzyme hybrid scaffold led to a further marginal increase in scaffold N50 and an overall increase of confidence in scaffolds. After removal of haplotigs, the final assembly is approximately 1.21 Gbp in size, with an N50 value of over 25.95 Mbp, representing an improvement in N50 of over 650 fold with respect to a previously reported assembly based on paired-end short read data.

Conclusions:
This high-quality genome assembly represents a valuable resource for further studies of population genetics of the barn swallow and for studies concerning the evolution of avian genomes. It also represents the first

genome assembled combining SMRT sequencing with the new Bionano Direct Label and Stain technology for scaffolding, highlighting the potential of this methodology to contribute to substantial increases in the contiguity of genome assemblies.

Keywords: genome, barn swallow, third-generation sequencing, SMRT, long reads, Bionano, DLS, DLE-1, optical maps, single molecule.

Data Description

Context

The barn swallow is a passerine bird with at least 8 recognized subspecies in Europe, Asia and North America. The European barn swallow (*Hirundo rustica rustica*) (Figure 1) breeds in a broad latitudinal range, between 63-68°N and 20-30°N [1]. Numerous evolutionary and ecological studies have focussed on its biology, life history, sexual selection, and response to climate change. More recently, the barn swallow has become the focus of genetic studies on the divergence between subspecies [2–4] and on the control of phenological traits [5–8]. Due to its synanthropic habits and its cultural value, the barn swallow is also a flagship species in conservation biology [1]. The availability of high-quality genomic resources, including a reference genome, is thus pivotal to further boost the study and conservation of this species.

Figure 1: the European barn swallow (*Hirundo rustica rustica*). Courtesy of Chiara Scandolara.

In 2016, Safran and coworkers reported the first draft of the genome for the American subspecies (Hirundo rustica erythrogaster) constructed from Illumina paired-end reads at 47x coverage depth [2]. This assembly was described as containing 1.1 Gbp of assembled sequences (average contig length 11 kbp, contig N50 = 39 kbp, contig N90 = 3.8 kbp, longest scaffold: 732 kbp), compared to an estimated genome size of 1.28 Gbp [9]. Moreover, the assembly was derived from a male individual, excluding information for the W chromosome, as females are the heterogametic (ZW) sex in birds.

To address the aforementioned limitations, we have employed two single-molecule technologies, SMRT Third-Generation Sequencing (TGS) from Pacific Biosciences (Menlo Park, California, USA) and optical mapping from Bionano Genomics (San Diego, California, USA), to produce a state-of-the-art high-quality genome assembly for the European subspecies. For optical mapping we have labeled DNA molecules both with one of the original Nick, Label, Repair and Stain (NLRS) nickases (Nb.BssSI) and with the new Direct Label and Stain (DLS) approach (enzyme DLE-1). The latter technique was officially released in February

2018 and avoids nicking and subsequent cleavage of DNA molecules during staining [10]. We show that, at least with our data, DLS allows a considerable improvement of scaffold contiguity with respect to the nickase tested. Furthermore, the "dual enzyme" approach affords additional support for scaffold junctions. To our knowledge this genome assembly is the first to incorporate DLS data, and their integration with SMRT sequencing provided assembly contiguity metrics well in excess of those specified for "Platinum genomes" by the Vertebrate Genomes Project (VGP) [11].

Blood sample collection

The blood used as a source of DNA was derived from a minimally invasive sampling performed on a female individual of approximately two years of age during May 2017 in a farm near Milan in Northern-Italy (45.4N 9.3E). Blood was collected in heparinized capillary tubes. Three hours after collection, the sample was centrifuged to separate blood cells from plasma and then stored at -80°C.

DNA extraction and quality control for SMRT library preparation

DNA extraction was performed on blood cells portion of centrifuged whole blood containing nucleated erythrocytes and leukocytes using the Wizard genomic DNA purification kit (Promega, Cat. No. A1125). This kit employs a protocol similar to classical Phenol/Chloroform DNA extraction, with no vortexing steps after cell lysis. After purification, DNA quality and concentration was assessed by Nanodrop (Thermo Fisher Scientific, Cat. No. ND-1000) and subsequently by Pulsed Field Gel Electrophoresis (PFGE). Detectable DNA was over 23 kbp in size, with the vast majority over 50 kbp and even over 200 kbp (Supplementary Figure 1). PFGE quality results were further confirmed by capillary electrophoresis on FEMTO Pulse instrument (AATI, Cat. No. FP-1002-0275) (Supplementary Figure 2). DNA was stored at -80°C and shipped on dry ice.

SMRT library preparation and sequencing

SMRTbell Express Template Prep Kit (Pacific Biosciences, Cat. No. 101-357-000) was used to produce the insert library. Input gDNA concentration was measured on a Qubit Fluorometer dsDNA Broad Range (Life Technologies, Cat. No. 32850). 10µg of gDNA was mechanically sheared to an average size distribution of 40-50 kbp, using a Megaruptor Device (Diagenode, Cat. No. B06010001). FEMTO Pulse capillary

electrophoresis was employed to assess the size of the fragments. 5 µg of sheared gDNA was DNA-damage repaired and end-repaired using polishing enzymes. Blunt-end ligation was used to create the SMRTbell template. A Blue Pippin device (Sage Science, Cat. No. BLU0001) was used to size-select the SMRTbell template and enrich for fragments > 30 kbp, excluding the first two cells for which the library was enriched for fragments > 15 kbp. The size-selected library was checked using FEMTO Pulse and quantified on a Qubit Fluorometer. A ready to sequence SMRT bell-Polymerase Complex was created using the Sequel binding kit 2.0 (Pacific Biosciences, Cat. No. 100-862-200). The Pacific Biosciences Sequel instrument was programmed to sequence the library on 18 Sequel SMRT Cells 1M v2 (Pacific Biosciences, Cat. No. 101-008-000), taking one movie of 10 hours per cell, using the Sequel Sequencing Kit 2.1 (Pacific Biosciences, Cat. No. 101-310-400). After the run, sequencing data quality was checked via the PacBio SMRT Link v5.0.1 software using the "run QC module". An average of 3.7 Gbp (standard deviation: 1.7) were produced per SMRT cell (average N50 read length = 25,622 bp), with considerable improvements between the 15 kbp library and the 30 kbp library (see Supplementary Figure 3 for more detailed statistics). We observed a wide distribution in the GC content of reads (Supplementary Figure 4). This is likely explained by the presence in avian genomes of three classes of chromosomes: macrochromosomes (50-200 Mbp, 5 in chicken), intermediate chromosomes (20-40 Mbp, 5 in chicken) and microchromosomes (12 Mbp on average, 28 in chicken). These last account for only 18% of the total genome but harbor ~31% of all chicken genes, have higher recombination rates and higher GC contents on average [12].

Assembly of SMRT reads

The final assembly of long reads was conducted with software CANU v1.7 [13] using default parameters except for the "correctedErrorRate" which was set at 0.075. The assembly processes occupied 3,840 CPU hours and 2.2 Tb of RAM for read correction, 768 CPU hours and 1.1 Tb of RAM for the trimming steps, and 3280 CPU hours and 2.2 Tb of RAM for the assembly phase. The assembly contained 3,872 contigs with a N50 of 5,2 Mbp for a total length of the assembly of 1311.7 Mbp (Table 1 and Supplementary Table 1). Final polishing was performed using the Arrow v2.10 software (Pacific Biosciences) and resulted in final coverage of 45.4x.

Cell count and DNA extraction for optical mapping

High-molecular weight (HMW) DNA was extracted from 7-8 µl of the cell portion from the same blood sample used for SMRT sequencing using the Blood and Cell Culture DNA Isolation kit (Bionano Genomics, Cat. No. RE-016-10). HMW DNA was extracted by embedding cells in low melting temperature agarose plugs that were incubated with Proteinase K (Qiagen, Cat. No. 158920) and RNAseA (Qiagen, Cat. No. 158924). The plugs were washed and solubilized using Agarase Enzyme (Thermo Fisher Scientific, Cat. No. EO0461) to release HMW DNA and further purified by drop dialysis. DNA was homogenised overnight prior to quantification using a Qubit Fluorometer.

***In silico* digestion**

The genome assembly obtained with CANU was *in silico* digested using Bionano Access software to test whether the nicking enzyme (Nb.BssSI), with recognition sequence (CACGAG), and the non-nicking enzyme DLE-1, with recognition sequence (CTTAAG), were suitable for optical mapping in our bird genome. An average of 16.9 nicks/100 kbp with a nick-to-nick distance N50 of 11,708 bp were expected for Nb.BssSI, while DLE-1 was found to induce 19.1 nicks/100 kbp with a nick-to-nick distance N50 of 8,775 bp, both in line with manufacturer's requirements.

DNA labeling for optical mapping

For NLRS, DNA was labeled according to manufacturer's instructions using the Prep DNA Labeling Kit-NLRS (Bionano Genomics, Cat. No. 80001). 300 ng of purified genomic DNA was nicked with Nb.BssSI (New England Biolabs, Cat. No. R0681S) in NEB Buffer 3. The nicked DNA was labeled with a fluorescent-dUTP nucleotide analog using Taq DNA polymerase (New England BioLabs, Cat. No. M0267S). After labeling, nicks were ligated with Taq DNA ligase (New England BioLabs, Cat. No. M0208S) in the presence of dNTPs. The backbone of fluorescently labeled DNA was counterstained overnight with YOYO-1 (Bionano Genomics, Cat. No. 80001).

For DLS, DNA was labeled using the Bionano Prep DNA Labeling Kit-DLS (Cat. No. 80005) according to manufacturer's instructions. 750 ng of purified genomic DNA was labeled with DLE labeling Mix and subsequently incubated with Proteinase K (Qiagen, Cat. No. 158920) followed by drop dialysis. After the clean-up step, the DNA was pre-stained, homogenised and quantified using on a Qubit Fluorometer to establish the appropriate amount of backbone stain. The reaction was incubated at room temperature for at

least 2 hours.

Generation of optical maps

NLRS and DLS labeled DNA were loaded into a nanochannel array of a Saphyr Chip (Bionano Genomics, Cat. No. FC-030-01) and run by electrophoresis each into a compartment. Linearized DNA molecules were imaged using the Saphyr system and associated software (Bionano Genomics, Cat. No. 90001 and CR-002-01). In the experiment with DLE-1, molecule N50 was 0.2475 Mbp for molecules above 20 kbp and 0.3641 Mbp for molecules above 150 kbp - with an average label density of 15.7/100 kbp for molecules above 150 kbp. Map rate was 56.4% for molecules above 150 kbp. Effective coverage was 30.6x.

In the experiment with Nb.BssSI, molecule N50 was 0.1298 Mbp for molecules above 20 kbp and 0.2336 Mbp for molecules above 150 kbp - with an average label density of 11.8/100 kbp for molecules above 150 kbp. Map rate was 38.9% for molecules above 150 kbp. Effective coverage was 28.2x. Using both DLE-1 and Nb.BssSI, label metrics were in line with the manufacturer's expectations.

Assembly of optical maps

The *de novo* assembly of the optical maps was performed using the Bionano Access v1.2.1 and Bionano Solve v3.2.1 software. The assembly type performed was the "non-haplotype" with "no extend split" and "no cut segdups". Default parameters were adjusted to accommodate the genomic properties of the barn swallow genome. Specifically, given the size of the genome, the minimal length for the molecules to be used in the assembly was reduced to 100 kbp, the "Initial P-value" cut off threshold was adjusted to 1×10^{-10} and the P-value cut off threshold for extension and refinement was set to 1×10^{-11} according to manufacturer's guidelines (default values are 150 kbp, 1×10^{-11} and 1×10^{-12} respectively).

A total of 233,450 (of 530,527) NLRS-labelled molecules (N50 = 0.2012 Mbp) were aligned to produce 2,384 map fragments with an N50 of 0.66 Mbp for a total length of 1338.6 Mbp (coverage = 32x). 108,307 (of 229,267) DLE-1 labelled input DNA molecules with a N50 of 0.3228 Mbp (theoretical coverage of the reference 48x) produced 555 maps with a N50 length of 12.1 Mbp for a total length 1299.3 Mbp (coverage = 23x).

Hybrid scaffolding

Single and dual enzyme Hybrid Scaffolding (HS) was performed using Bionano Access v1.2.1 and Bionano Solve v3.2.1. For the dual enzyme and DLE-1 scaffolding, default settings were used to perform the HS. For Nb.BssSI the "aggressive" settings were used without modification. The NLRS hybrid assembly had an N50 of 8.3 Mbp (scaffold only N50 = 10.8 Mbp) for a total length of 1,338.6 Mbp (total length of scaffolded contigs = 1,175.3 Mbp) and consisted of 409 scaffolds and 2,899 un-scaffolded contigs. The DLS hybrid assembly had N50 of 17.3 Mbp (scaffold only N50 = 25.9 Mbp) for a total length of 1,340.2 Mbp (total length of scaffolded contigs = 1,148.4 Mbp) and consisted of 211 scaffolds and 3,106 un-scaffolded contigs. Dual enzyme HS (incorporating both DLS and NLRS maps) resulted in an assembly with N50 of 23.8 Mbp (scaffold only N50 = 28.4 Mbp) for a total length of 1,351.8 Mbp (total length of scaffolded contigs = 1,208.8 Mbp) and consisted of 273 scaffolds and 2,810 un-scaffolded contigs. During the automatic conflict resolution in the dual enzyme HS, 185 SMRT contigs were cut, as Bionano maps confidently indicated mis-assemblies of the SMRT reads. Conversely 117 bionano maps were cut indicating that the chimeric score did not provide sufficient confidence to cut the assembly based on SMRT contigs. Of 3,872 SMRT contigs, 1,243 (32%) were anchored in the Bionano maps (of which 990 were anchored in both DLS and NLRS maps). 56 and 226 were anchored in DLS and NLRS maps respectively. 2810 maps could not be anchored at all.

Notably, all hybrid assemblies were somewhat larger than the expected genome size, and in all cases, the N50 of un-scaffolded contigs was extremely low (0.06 Mbp for the dual enzyme hybrid assembly). We hypothesized that a significant proportion of these small contigs might represent divergent homologous haplotigs that were assembled independently. Similarity searches were consistent with this possibility as almost 95% of the contigs that were not scaffolded in the dual enzyme hybrid assembly showed > 98% identity to scaffolded contigs over 75% of their length or more. These contigs were discarded, resulting in a final assembly (Table 1 and Supplementary Table 1 for detailed statistics) of 1.21 Gbp (N50 = 25.9 Mbp) made up of 273 dual enzyme hybrid scaffolds (N50 = 28.42 Mbp) and 91 un-scaffolded contigs (N50 = 0.0644 Mbp). The final assembly is slightly smaller than the previously estimated genome size (1.28 Gbp) [9], possibly reflecting an imprecise older estimate and/or the possibility that some poorly assembled repeats were discarded in the final step described above. The average read SMRT read coverage for the genome assembly was 34.15X (implying a theoretical QV of over 40). Supplementary Figure 5 provides a

summary of observed sequence coverage depth.

	Safran et al. [2][1]	**SMRT contigs**[2]	**Final assembly**[3]
Species	*H. r. erythrogaster*	*H. r. rustica*	
Starting raw data (Gbp)	61.7	66.4	59.6
N50 (bp)	38844	5189284	25954216
N90 (bp)	3718	85340	2002624
Total size (Gbp)	1.1	1.31	1.21
Theoretical genome coverage*	47x	52x	47x
% genome coverage*	85.9	102.6	94.5
# of contigs/scaffolds	100153	3872	364
Avg contig/scaffold length (bp)	11010	338782	3334461
Longest contig/scaffold (bp)	732517	33230000	98053015

Table 1: Assembly metrics for contigs and final scaffolds in our European barn swallow genome compared to the published American barn swallow genome. [1] Illumina PE reads assembled using SOAPdenovo v2.04 [14]. [2] SMRT reads assembled using CANU v1.7 [13]. [3] SMRT contigs assembled with CANU and scaffolded using Bionano dual enzyme HS, with haplotigs removed as detailed in the text. *Based on a barn swallow genome size estimate of 1.28 Gbp [9].

Annotation of genes and repeats

With respect to mammals, avian genomes generally contain relatively low proportions of repetitive sequences and show strong mutual synteny [15]. This appears to be the case also for the barn swallow genome. In particular, 7.11% of the final assembly was annotated as repetitive using RepeatMasker [16], with the major contributions deriving from L2/CR1/Rex LINE elements (3.37%), retroviral LTRs (1.59%) and simple repeats (1.56%). These repeats were soft-masked prior to *de novo* gene prediction using Augustus [17] with *Gallus gallus* gene models.

In all, 35,644 protein coding genes were predicted, of which 9,189 were overlapped by more than 30% of their size with repetitive genomic elements. Of the remaining 26,455 predicted protein coding genes, 24,331 harbored a PFAM protein domain. Simple similarity searches based on blastp [18] (with default parameters)

suggested that 17,895 of the predicted protein coding genes have a best reciprocal blast hit with gene models derived from *Gallus gallus* GRCg6a assembly (as available from [19]), while 2,927 of the proteins predicted by Augustus did not show any significant match (e-value <= 1 x 10^{-15}, identity > 35%).

BUSCO genes

Of a total of 4915 conserved bird Benchmarking with Universal Single-Copy Orthologs (BUSCO) groups [20] sought, 4,598 (93.6%) were complete, 4,521 (92.0%) were complete and single-copy, 77 (1.6%) were complete and duplicated, 192 (3.9%) were fragmented and 125 (2.5%) were missing. The percentage of contiguously assembled BUSCO genes is consistent with recent results with Anna's Hummingbird (*Calypte anna*) and the Zebra Finch (*Taeniopygia guttata*) [21]. We note that 40 of the "missing" bird BUSCO genes are absent from at least 2 of the 54 available avian genome sequences, suggesting that, despite the potentially incomplete nature of some draft genomes, some of these genes may not be universally conserved among birds.

Synteny with the Chicken and Hummingbird genomes

Alignment of the final assembly with the most recent assembly of the chicken genome (GRCg6a) using D-Genies [22] indicates high levels of collinearity between these two genomes with a limited number of intra-chromosomal rearrangements (Figure 2). The high level of collinearity between independently assembled and scaffolded sequences provides circumstantial support for the quality of both the contigs and the hybrid scaffolds and is consistent with previous observations of high levels of synteny and minimal inter-chromosomal rearrangements among birds [15].

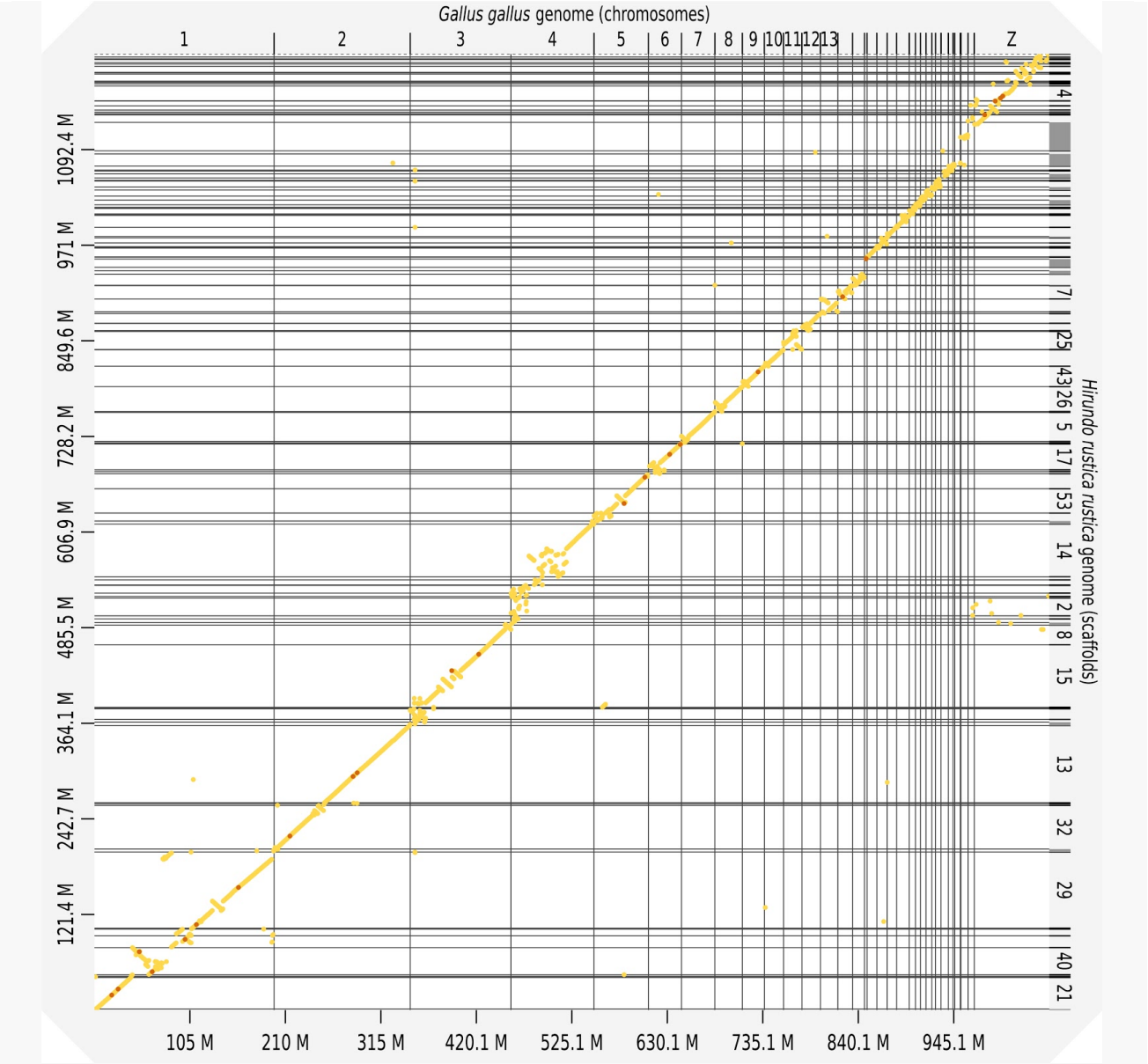

Figure 2: Alignment of the final assembly with the published chromosome-level assembly of the chicken (*Gallus gallus*) genome GRCg6a using D-Genies [22]. Light to dark yellow dots indicate progressively higher similarity between sequences.

Conclusion

During the last 20 years, nucleic acid sequencing technologies have developed over four times faster than improvements of microchip complexity predicted by Moore's law [23,24]. While short-read NGS (now known as Second Generation Sequencing - SGS) technologies have allowed the production of cost-effective genome drafts for many birds and other vertebrate species [25–27], the reduction in genome sequencing costs has typically come at the price of compromises in contiguity and accuracy of assemblies with respect

to earlier efforts based on Sanger reads and extensive physical mapping [28]. Many limitations of SGS-based assemblies stem from the occurrence of long sequence repeats. In many animal species, transposons are frequently located in introns [29] and the presence of large gene families of closely related paralogs can lead to the existence of long "genic" repeats. Accordingly, even apparently contiguous genic regions can feature juxtaposition of paralogous gene fragments [21]. Given the inception of large scale sequencing initiatives aiming to produce genome assemblies for a wide range of organisms [30–33], it is critical to identify combinations of sequencing and scaffolding approaches that allow the cost effective generation of genuinely high-quality genome assemblies [11]. While exhibiting higher rates of single-base errors than some SGS approaches, TGS technologies, including SMRT sequencing, offer read-lengths unparalleled by SGS or Sanger sequencing [34]. Moreover, recent and ongoing improvements in TGS methods are rapidly reducing the "per-base" cost of TGS data compared to that of SGS. On the other hand, as an alternative to scaffolding with long insert mate-pairs [35] or to chromatin proximity ligation sequencing [36], contiguity and accuracy of long-read-based assemblies can be further improved by optical mapping. This relies on nanoscale channels that can accommodate thousands of single, ultralong (>200 kbp) double-stranded DNA filaments in parallel, subsequently stained to recognize specific 6-7 bp long motifs [37]. The combination of long reads and optical maps has already proven invaluable to produce high-quality genome assemblies, even in the case of particularly complex genomes [38]. Here, using only SMRT sequencing and Bionano optical maps we have produced a high-quality and contiguous genome for the barn swallow. With respect to a previously reported SGS-based assembly of the American barn swallow genome using a comparable amount of raw data [2], even the contigs generated from long-read sequencing alone show an 134-fold increase in N50. In terms of N50, number and average length, these contigs are similar to those recently obtained for the Anna's hummingbird genome using the same technology [21]. Furthermore, the fold change in N50 attained by Bionano NLRS hybrid scaffolding of the European barn swallow genome (1.6 fold before removal of haplotigs) is comparable with results obtained by other genome assemblies that have employed this method [39]. Strikingly, the new DLS method greatly outperformed the NLRS system, providing a 3.3 fold increase of N50 (before removal of haplotigs). Moreover, incorporation of both labelling systems into the hybrid scaffolding yielded a final assembly showing 5-fold improvement of the N50 with respect to the original SMRT assembly, simultaneously providing "independent" validation of

many scaffold junctions. We note that the presence of numerous microchromosomes in avian genomes restricts the final N50 value potentially attainable for the assembly, as the fully assembled karyotype would have an N50 of ~ 90 Mbp. Yet, after removal of putative haplotigs, our genome assembly contiguity metrics meet the high standards of the VGP consortium "Platinum Genome" criteria (contig N50 in excess of 1 Mbp and scaffold N50 above 10 Mbp) [11]. Accordingly, we believe that the data presented here, as well as attesting to the effectiveness of SMRT sequencing combined with DLS optical mapping for the assembly of vertebrate genomes, will provide an invaluable asset for population genetics studies in the barn swallow and for comparative genomics in birds.

Re-use Potential

Future directions for the barn swallow genome include the phasing of the assembly to generate extended haplotypes, a more thorough gene annotation using RNA/IsoSeq sequencing data, detailed comparisons with genome data from *Hirundo rustica erythrogaster,* re-evaluation of data from previous population genetics studies conducted in this species, as well as characterization of the epigenetic landscape.

Availability of supporting data

The data sets supporting the results of this article will available in the GenBank repository upon acceptance, under Bioproject PRJNA481100.

Competing interests

Kees-Jan Francoijs is currently employed at Bionano Genomics (San Diego, CA, USA). All other authors declare no competing interest.

Funding

Funding to A.B.-A. was provided by Cal Poly Pomona College of Science.

Authors' contributions

G.F, N.S., A.B.-A., L.G., D.S.H, M.C. and L.C. conceived the project and designed the experiments; G.F. performed DNA extraction and quality control; M.C. carried out CANU assemblies, gene and repeat annotation. D.S.H., M.C. and L.G. performed other bioinformatics analyses; L.P. conducted the optical

mapping; K.J.F. produced the hybrid scaffolds; G.F., D.S.H, M.C., N.S. and L.C. drafted the manuscript. All authors edited and contributed to the manuscript.

Acknowledgements

We thank Manuela Caprioli for support in field work, sample collection, DNA extraction and quality control as well as Dr. Elena Galati for support in PFGE quality control. We also thank The Functional Genomics Center of Zurich, where SMRT sequencing and optical mapping were carried out, and particularly Andrea Patrignani for SMRT sequencing. We thank Chiara Scandolara for the barn swallow picture used for Figure 1. We acknowledge the support of ELIXIR-IT and CINECA (HPC@CINECA) for provision of computational resources for SMRT read assembly.

Ethics approval

The blood sample used to generate the genomic data derived from a minimally invasive sampling on a single individual. Appropriate consent was obtained from the local authorities (Regione Lombardia).

Additional files

Supplementary Figure 1 (Supplementary Figure 1.png)

PFGE on a 1x agarose gel run for 18 hours at 160 mV. The two lowest overlapping bands in lane 1 represent yeast chromosomes of 230 kbp and 270 kbp, respectively. Lane 2 contains 1kb DNA ladder (highest 10 kbp), lane 3 and 4 the undigested lambda phage (50 kbp) and lane 5 digested lambda (upper band 23 kbp). Lane 7 contains the sample used in the study.

Supplementary Figure 2 (Supplementary Figure 2.tif)

FEMTO Pulse capillary electrophoresis results for the DNA sample used in the study.

Supplementary Figure 3 (Supplementary Figure 3.png)

Summary statistics for each SMRT cell employed.

Supplementary Figure 4 (Supplementary Figure 4.png)

GC content distribution in all sequence reads.

Supplementary Figure 5 (Supplementary Figure 5.png)

Cumulative coverage distribution of the final (de-haplotyped) assembly of the barn swallow genome. Coverage is indicated on the X axis. Red lines are used to display the proportion of the genome covered by more than 10, 20, 30, 40, 50 or 60 reads respectively.

Supplementary Table 1 (Supplementary Table 1.xlsx)

Comparison of assembly metrics for contigs and scaffolds between different assemblies. In hybrid scaffolds, the first column refers to assemblies including the un-scaffolded contigs while the second column only includes scaffolded contigs metrics. The estimated genome size of 1.28 Gbp is from [9]. Average gene size was estimated according to the latest available annotation of the *Gallus gallus* genome (GRCg6a).

List of abbreviations

DLS, Direct Label and Stain; HMW, High Molecular Weight; HS, Hybrid Scaffold; NGS, Next Generation Sequencing; NLRS, Nick, Label, Repair and Stain; N50, the shortest sequence length at 50% of the genome; N90, the shortest sequence length at 90% of the genome; PFGE, Pulsed Field Gel Electrophoresis; QV, Quality Value; SGS, Second Generation Sequencing; SMRT, Single Molecule Real-Time; TGS, Third Generation Sequencing; VGP, Vertebrate Genomes Project.

References

1. Turner A. The barn swallow. T & AD Poyser, London; 2006.

2. Safran RJ, Scordato ESC, Wilkins MR, Hubbard JK, Jenkins BR, Albrecht T, et al. Genome-wide differentiation in closely related populations: the roles of selection and geographic isolation. Mol Ecol. 2016;25:3865–83.

3. von Rönn JAC, Shafer ABA, Wolf JBW. Disruptive selection without genome-wide evolution across a migratory divide. Mol Ecol. 2016;25:2529–41.

4. Scordato ESC, Wilkins MR, Semenov G, Rubtsov AS, Kane NC, Safran RJ. Genomic variation across two barn swallow hybrid zones reveals traits associated with divergence in sympatry and allopatry. Mol Ecol. 2017;26:5676–91.

5. Caprioli M, Ambrosini R, Boncoraglio G, Gatti E, Romano A, Romano M, et al. Clock gene variation is associated with breeding phenology and maybe under directional selection in the migratory barn swallow. PLoS One. 2012;7:e35140.

6. Saino N, Romano M, Caprioli M, Fasola M, Lardelli R, Micheloni P, et al. Timing of molt of barn swallows is delayed in a rare Clock genotype. PeerJ. 2013;1:e17.

7. Bazzi G, Ambrosini R, Caprioli M, Costanzo A, Liechti F, Gatti E, et al. Clock gene polymorphism and scheduling of migration: a geolocator study of the barn swallow Hirundo rustica. Sci Rep. 2015;5:12443.

8. Saino N, Ambrosini R, Albetti B, Caprioli M, De Giorgio B, Gatti E, et al. Migration phenology and

breeding success are predicted by methylation of a photoperiodic gene in the barn swallow. Sci Rep. 2017;7:45412.

9. Andrews CB, Mackenzie SA, Gregory TR. Genome size and wing parameters in passerine birds. Proc Biol Sci. 2009;276:55–61.

10. DLS announcement by Bionano Genomics at AGBT [Internet]. Available from: https://bionanogenomics.com/wp-content/uploads/2018/02/Bionano-AGBT2018-DLS_launch_final.pdf

11. Lewin HA, Robinson GE, Kress WJ, Baker WJ, Coddington J, Crandall KA, et al. Earth BioGenome Project: Sequencing life for the future of life. Proc Natl Acad Sci U S A. 2018;115:4325–33.

12. Kadi F, Mouchiroud D, Sabeur G, Bernardi G. The compositional patterns of the avian genomes and their evolutionary implications. J Mol Evol. 1993;37:544–51.

13. Koren S, Walenz BP, Berlin K, Miller JR, Bergman NH, Phillippy AM. Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation. Genome Res. 2017;27:722–36.

14. Li R, Zhu H, Ruan J, Qian W, Fang X, Shi Z, et al. De novo assembly of human genomes with massively parallel short read sequencing. Genome Res. 2010;20:265–72.

15. Ellegren H. Evolutionary stasis: the stable chromosomes of birds. Trends Ecol Evol. 2010;25:283–91.

16. Smit AF, Hubley R, Green P. RepeatMasker Open-3.0 [Internet]. 1996–2010. Available from: http://www.repeatmasker.org

17. Stanke M, Steinkamp R, Waack S, Morgenstern B. AUGUSTUS: a web server for gene finding in eukaryotes. Nucleic Acids Res. 2004;32:W309–12.

18. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. J Mol Biol. 1990;215:403–10.

19. Gallus gallus Proteins [Internet]. NCBI. Available from: https://www.ncbi.nlm.nih.gov/genome/proteins/111?genome_assembly_id=374862

20. Simão FA, Waterhouse RM, Ioannidis P, Kriventseva EV, Zdobnov EM. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. Bioinformatics. 2015;31:3210–2.

21. Korlach J, Gedman G, Kingan SB, Chin C-S, Howard JT, Audet J-N, et al. De novo PacBio long-read and phased avian genome assemblies correct and add to reference genes generated with intermediate and short reads. Gigascience. 2017;6:1–16.

22. Cabanettes F, Klopp C. D-GENIES: dot plot large genomes in an interactive, efficient and simple way. PeerJ. 2018;6:e4958.

23. Heather JM, Chain B. The sequence of sequencers: The history of sequencing DNA. Genomics. 2016;107:1–8.

24. Stein LD. The case for cloud computing in genome informatics. Genome Biol. 2010;11:207.

25. Genome 10K Community of Scientists. Genome 10K: a proposal to obtain whole-genome sequence for 10,000 vertebrate species. J Hered. 2009;100:659–74.

26. Zhang G, Li C, Li Q, Li B, Larkin DM, Lee C, et al. Comparative genomics reveals insights into avian genome evolution and adaptation. Science. 2014;346:1311–20.

27. Jarvis ED, Mirarab S, Aberer AJ, Li B, Houde P, Li C, et al. Whole-genome analyses resolve early branches in the tree of life of modern birds. Science. 2014;346:1320–31.

28. Henson J, Tischler G, Ning Z. Next-generation sequencing and large genome assemblies. Pharmacogenomics. 2012;13:901–15.

29. Sela N, Kim E, Ast G. The role of transposable elements in the evolution of non-mammalian vertebrates and invertebrates. Genome Biol. 2010;11:R59.

30. Koepfli K-P, Paten B, Genome 10K Community of Scientists, O'Brien SJ. The Genome 10K Project: a way forward. Annu Rev Anim Biosci. 2015;3:57–111.

31. Zhang G, Rahbek C, Graves GR, Lei F, Jarvis ED, Gilbert MTP. Genomics: Bird sequencing project takes off. Nature. 2015;522:34.

32. Pennisi E. Sequencing all life captivates biologists. Science. 2017;355:894–5.

33. Teeling EC, Vernes SC, Dávalos LM, Ray DA, Gilbert MTP, Myers E, et al. Bat Biology, Genomes, and the Bat1K Project: To Generate Chromosome-Level Genomes for All Living Bat Species. Annu Rev Anim Biosci. 2018;6:23–46.

34. Bleidorn C. Third generation sequencing: technology and its potential impact on evolutionary biodiversity research. System Biodivers. Taylor & Francis; 2016;14:1–8.

35. Hunt M, Newbold C, Berriman M, Otto TD. A comprehensive evaluation of assembly scaffolding tools. Genome Biol. 2014;15:R42.

36. Burton JN, Adey A, Patwardhan RP, Qiu R, Kitzman JO, Shendure J. Chromosome-scale scaffolding of de novo genome assemblies based on chromatin interactions. Nat Biotechnol. 2013;31:1119–25.

37. Lam ET, Hastie A, Lin C, Ehrlich D, Das SK, Austin MD, et al. Genome mapping on nanochannel arrays for structural variation analysis and sequence assembly. Nat Biotechnol. 2012;30:771–6.

38. Nowoshilow S, Schloissnig S, Fei J-F, Dahl A, Pang AWC, Pippel M, et al. The axolotl genome and the evolution of key tissue formation regulators. Nature. 2018;554:50–5.

39. Gao Y, Wang H, Liu C, Chu H, Dai D, Song S, et al. De novo genome assembly of the red silk cotton tree (Bombax ceiba). Gigascience [Internet]. 2018;7. Available from: http://dx.doi.org/10.1093/gigascience/giy051

References

1000 Genomes Project Consortium, Gonçalo R. Abecasis, David Altshuler, Adam Auton, Lisa D. Brooks, Richard M. Durbin, Richard A. Gibbs, Matt E. Hurles, and Gil A. McVean. 2010. "A Map of Human Genome Variation from Population-Scale Sequencing." *Nature* 467 (7319): 1061–73.

1000 Genomes Project Consortium, Goncalo R. Abecasis, Adam Auton, Lisa D. Brooks, Mark A. DePristo, Richard M. Durbin, Robert E. Handsaker, Hyun Min Kang, Gabor T. Marth, and Gil A. McVean. 2012. "An Integrated Map of Genetic Variation from 1,092 Human Genomes." *Nature* 491 (7422): 56–65.

1000 Genomes Project Consortium, Adam Auton, Lisa D. Brooks, Richard M. Durbin, Erik P. Garrison, Hyun Min Kang, Jan O. Korbel, et al. 2015. "A Global Reference for Human Genetic Variation." *Nature* 526 (7571): 68–74.

Adams, Christopher P., and Stephen Joseph Kron. 1997. Method for performing amplification of nucleic acid with two primers bound to a single solid support. USPTO 5641658. *US Patent*, filed August 3, 1994, and issued June 24, 1997. https://patentimages.storage.googleapis.com/11/8d/27/1e9724e1e2d015/US5641658.pdf.

Adams, J. 2008. "DNA Sequencing Technologies." *Nature Education* 1 (1): 193.

Adams, M. D., S. E. Celniker, R. A. Holt, C. A. Evans, J. D. Gocayne, P. G. Amanatides, S. E. Scherer, et al. 2000. "The Genome Sequence of Drosophila Melanogaster." *Science* 287 (5461): 2185–95.

Adessi, C., G. Matton, G. Ayala, G. Turcatti, J. J. Mermod, P. Mayer, and E. Kawashima. 2000. "Solid Phase DNA Amplification: Characterisation of Primer Attachment and Amplification Mechanisms." *Nucleic Acids Research* 28 (20): E87.

Altschul, S. F., W. Gish, W. Miller, E. W. Myers, and D. J. Lipman. 1990. "Basic Local Alignment Search Tool." *Journal of Molecular Biology* 215 (3): 403–10.

Ambrosini, R., Luciano Bani, Dario Massimino, Lorenzo Fornasari, and Nicola Saino. 2011. "Large-Scale Spatial Distribution of Breeding Barn Swallows Hirundo Rustica in Relation to Cattle Farming." *Bird Study: The Journal of the British Trust for Ornithology* 58 (4): 495–505.

Ambrosini, R., Anna Maria Bolzern, Luca Canova, Silvia Arieni, Anders Pape Moller, and Nicola Saino. 2002. "The Distribution and Colony Size of Barn Swallows in Relation to Agricultural Land Use." *The Journal of Applied Ecology* 39 (3): 524–34.

Ambrosini, R., Anna Maria Bolzern, Luca Canova, and Nicola Saino. 2002. "Latency in Response of Barn Swallow Hirundo Rustica Populations to Changes in Breeding Habitat Conditions." *Ecology Letters* 5 (5): 640–47.

Ambrosini, R., Riccardo Borgoni, Diego Rubolini, Beatrice Sicurella, Wolfgang Fiedler, Franz Bairlein, Stephen R. Baillie, et al. 2014. "Modelling the Progression of Bird Migration with Conditional Autoregressive Models Applied to Ringing Data." *PloS One* 9 (7): e102440.

Ambrosini, R., Raffaella Paola Ferrari, Roberta Martinelli, M. Romano, and Nicola Saino. 2006. "Seasonal, Meteorological, and Microhabitat Effects on Breeding Success and Offspring Phenotype in the Barn Swallow, Hirundo Rustica." *Écoscience* 13 (3): 298–307.

Ambrosini, R., A. P. Møller, and N. Saino. 2009. "A Quantitative Measure of Migratory Connectivity." *Journal of Theoretical Biology* 257 (2): 203–11.

Ambrosini, R., Diego Rubolini, Paola Trovò, Giovanni Liberini, Marco Bandini, A. Romano, Beatrice Sicurella, Chiara Scandolara, M. Romano, and Nicola Saino. 2012. "Maintenance of Livestock Farming May Buffer Population Decline of the Barn Swallow Hirundo Rustica." *Bird Conservation International* 22 (4): 411–28.

Ambrosini, R., D. Rubolini, A. P. Møller, L. Bani, J. Clark, Z. Karcza, D. Vangeluwe, C. du Feu, F. Spina, and N. Saino. 2011. "Climate Change and the Long-Term Northward Shift in the African Wintering

Range of the Barn Swallow Hirundo Rustica." *Climate Research* 49 (2): 131–41.

Andrews, Chandler B., Stuart A. Mackenzie, and T. Ryan Gregory. 2009. "Genome Size and Wing Parameters in Passerine Birds." *Proceedings. Biological Sciences / The Royal Society* 276 (1654): 55–61.

Ardui, Simon, Adam Ameur, Joris R. Vermeesch, and Matthew S. Hestand. 2018. "Single Molecule Real-Time (SMRT) Sequencing Comes of Age: Applications and Utilities for Medical Diagnostics." *Nucleic Acids Research* 46 (5): 2159–68.

Axelsson, Erik, Matthew T. Webster, Nick G. C. Smith, David W. Burt, and Hans Ellegren. 2005. "Comparison of the Chicken and Turkey Genomes Reveals a Higher Rate of Nucleotide Divergence on Microchromosomes than Macrochromosomes." *Genome Research* 15 (1): 120–25.

Baer, R., A. T. Bankier, M. D. Biggin, P. L. Deininger, P. J. Farrell, T. J. Gibson, G. Hatfull, et al. 1984. "DNA Sequence and Expression of the B95-8 Epstein—Barr Virus Genome." *Nature* 310 (July): 207.

Balasubramanian, Shankar, David Klenerman, and Colin Barnes. 2003. Arrayed polynucleotides and their use in genome analysis. USPTO 20030022207:A1. *US Patent*, filed May 22, 2002, and issued January 30, 2003. https://patentimages.storage.googleapis.com/28/85/94/2ce5b9b4e6eed4/US20030022207A1.pdf.

Balbontín, Javier, Anders Pape Møller, Ignacio G. Hermosell, Alfonso Marzal, Maribel Reviriego, and Florentino de Lope. 2009. "Individual Responses in Spring Arrival Date to Ecological Conditions during Winter and Migration in a Migratory Bird." *The Journal of Animal Ecology* 78 (5): 981–89.

Bamshad, Michael J., Sarah B. Ng, Abigail W. Bigham, Holly K. Tabor, Mary J. Emond, Deborah A. Nickerson, and Jay Shendure. 2011. "Exome Sequencing as a Tool for Mendelian Disease Gene Discovery." *Nature Reviews. Genetics* 12 (11): 745–55.

Bankier, A. T., S. Beck, R. Bohni, C. M. Brown, R. Cerny, M. S. Chee, C. A. Hutchison 3rd, T. Kouzarides, J. A. Martignetti, and E. Preddie. 1991. "The DNA Sequence of the Human Cytomegalovirus Genome." *DNA Sequence: The Journal of DNA Sequencing and Mapping* 2 (1): 1–12.

Bao, Weidong, Kenji K. Kojima, and Oleksiy Kohany. 2015. "Repbase Update, a Database of Repetitive Elements in Eukaryotic Genomes." *Mobile DNA* 6 (June): 11.

Barnes, C., S. Balasubramanian, X. Liu, and H. Swerdlow. 2006. Labelled nucleotides. *Patent US7057026*, issued 2006.

Bayley, Hagan. 2015. "Nanopore Sequencing: From Imagination to Reality." *Clinical Chemistry* 61 (1): 25–31.

Bazzi, G., R. Ambrosini, M. Caprioli, A. Costanzo, F. Liechti, E. Gatti, L. Gianfranceschi, et al. 2015. "Clock Gene Polymorphism and Scheduling of Migration: A Geolocator Study of the Barn Swallow Hirundo Rustica." *Scientific Reports* 5 (July): 12443.

Beaumont, M. A., and R. A. Nichols. 1996. "Evaluating Loci for Use in the Genetic Analysis of Population Structure." *Proceedings of the Royal Society of London. Series B, Containing Papers of a Biological Character. Royal Society* . http://rspb.royalsocietypublishing.org/content/263/1377/1619.short.

Bentley, David R., Shankar Balasubramanian, Harold P. Swerdlow, Geoffrey P. Smith, John Milton, Clive G. Brown, Kevin P. Hall, et al. 2008. "Accurate Whole Human Genome Sequencing Using Reversible Terminator Chemistry." *Nature* 456 (7218): 53–59.

Berthold, P. 1991. "Orientation in Birds. Spatiotemporal Programmes and Genetics of Orientation." *EXS* 60: 86–105.

Birney, Ewan, and Nicole Soranzo. 2015. "Human Genomics: The End of the Start for Population Sequencing." *Nature* 526 (7571): 52–53.

Blanco, L., A. Bernad, J. M. Lázaro, G. Martín, C. Garmendia, and M. Salas. 1989. "Highly Efficient DNA Synthesis by the Phage Phi 29 DNA Polymerase. Symmetrical Mode of DNA Replication." *The Journal of Biological Chemistry* 264 (15): 8935–40.

Blattner, F. R., G. Plunkett 3rd, C. A. Bloch, N. T. Perna, V. Burland, M. Riley, J. Collado-Vides, et al. 1997. "The Complete Genome Sequence of Escherichia Coli K-12." *Science* 277 (5331): 1453–62.

Bleidorn, Christoph. 2016. "Third Generation Sequencing: Technology and Its Potential Impact on Evolutionary Biodiversity Research." *Systematics and Biodiversity* 14 (1): 1–8.

Bowles, Neil E., Chuanchau J. Jou, Cammon B. Arrington, Brett J. Kennedy, Aubree Earl, Norisada Matsunami, Lindsay L. Meyers, et al. 2015. "Exome Analysis of a Family with Wolff-Parkinson-White Syndrome Identifies a Novel Disease Locus." *American Journal of Medical Genetics. Part A* 167A (12): 2975–84.

Boyle, Alan P., Sean Davis, Hennady P. Shulha, Paul Meltzer, Elliott H. Margulies, Zhiping Weng, Terrence S. Furey, and Gregory E. Crawford. 2008. "High-Resolution Mapping and Characterization of Open Chromatin across the Genome." *Cell* 132 (2): 311–22.

Braslavsky, Ido, Benedict Hebert, Emil Kartalov, and Stephen R. Quake. 2003. "Sequence Information Can Be Obtained from Single DNA Molecules." *Proceedings of the National Academy of Sciences of the United States of America* 100 (7): 3960–64.

Brenner, S., M. Johnson, J. Bridgham, G. Golda, D. H. Lloyd, D. Johnson, S. Luo, et al. 2000. "Gene Expression Analysis by Massively Parallel Signature Sequencing (MPSS) on Microbead Arrays." *Nature Biotechnology* 18 (6): 630–34.

Brownlee, G. G., F. Sanger, and B. G. Barrell. 1968. "The Sequence of 5 S Ribosomal Ribonucleic Acid." *Journal of Molecular Biology* 34 (3): 379–412.

Bult, C. J., O. White, G. J. Olsen, L. Zhou, R. D. Fleischmann, G. G. Sutton, J. A. Blake, et al. 1996. "Complete Genome Sequence of the Methanogenic Archaeon, Methanococcus Jannaschii." *Science* 273 (5278): 1058–73.

Burge, C., and S. Karlin. 1997. "Prediction of Complete Gene Structures in Human Genomic DNA." *Journal of Molecular Biology* 268 (1): 78–94.

Burton, Joshua N., Andrew Adey, Rupali P. Patwardhan, Ruolan Qiu, Jacob O. Kitzman, and Jay Shendure. 2013. "Chromosome-Scale Scaffolding of de Novo Genome Assemblies Based on Chromatin Interactions." *Nature Biotechnology* 31 (12): 1119–25.

Butler, Jonathan, Iain MacCallum, Michael Kleber, Ilya A. Shlyakhter, Matthew K. Belmonte, Eric S. Lander, Chad Nusbaum, and David B. Jaffe. 2008. "ALLPATHS: De Novo Assembly of Whole-Genome Shotgun Microreads." *Genome Research* 18 (5): 810–20.

Cabanettes, Floréal, and Christophe Klopp. 2018. "D-GENIES: Dot Plot Large Genomes in an Interactive, Efficient and Simple Way." *PeerJ* 6 (June): e4958.

Cao, Hongzhi, Alex R. Hastie, Dandan Cao, Ernest T. Lam, Yuhui Sun, Haodong Huang, Xiao Liu, et al. 2014. "Rapid Detection of Structural Variation in a Human Genome Using Nanochannel-Based Genome Mapping Technology." *GigaScience* 3 (1): 34.

Caprioli, M., R. Ambrosini, G. Boncoraglio, E. Gatti, A. Romano, M. Romano, D. Rubolini, L. Gianfranceschi, and N. Saino. 2012. "Clock Gene Variation Is Associated with Breeding Phenology and Maybe under Directional Selection in the Migratory Barn Swallow." *PloS One* 7 (4): e35140.

C. elegans Sequencing Consortium. 1998. "Genome Sequence of the Nematode C. Elegans: A Platform for Investigating Biology." *Science* 282 (5396): 2012–18.

Chargaff, E., R. Lipshitz, and C. Green. 1952. "Composition of the Desoxypentose Nucleic Acids of Four Genera of Sea-Urchin." *The Journal of Biological Chemistry* 195 (1): 155–60.

Chargaff, E., R. Lipshitz, C. Green, and M. E. Hodes. 1951. "The Composition of the Desoxyribonucleic Acid of Salmon Sperm." *The Journal of Biological Chemistry* 192 (1): 223–30.

Chen, Ken, John W. Wallis, Michael D. McLellan, David E. Larson, Joelle M. Kalicki, Craig S. Pohl, Sean D. McGrath, et al. 2009. "BreakDancer: An Algorithm for High-Resolution Mapping of Genomic Structural Variation." *Nature Methods* 6 (9): 677–81.

Cherf, Gerald M., Kate R. Lieberman, Hytham Rashid, Christopher E. Lam, Kevin Karplus, and Mark Akeson. 2012. "Automated Forward and Reverse Ratcheting of DNA in a Nanopore at 5-Å Precision." *Nature Biotechnology* 30 (4): 344–48.

Chimpanzee Sequencing and Analysis Consortium. 2005. "Initial Sequence of the Chimpanzee Genome and Comparison with the Human Genome." *Nature* 437 (7055): 69–87.

Church, G., D. W. Deamer, D. Branton, R. Baldarelli, and J. Kasianowicz. 1998. "Characterization of Individual Polymer Molecules Based on Monomer-Interface Interactions. Patent US5795782." *The Concept of ssDNA Modulating an Electronic Signal While Moving through a Membrane Pore Led Eventually to Practical Nanopore Sequencing*.

Cloonan, Nicole, Alistair R. R. Forrest, Gabriel Kolle, Brooke B. A. Gardiner, Geoffrey J. Faulkner, Mellissa K. Brown, Darrin F. Taylor, et al. 2008. "Stem Cell Transcriptome Profiling via Massive-Scale mRNA Sequencing." *Nature Methods* 5 (7): 613–19.

Cohen, S. N., A. C. Chang, H. W. Boyer, and R. B. Helling. 1973. "Construction of Biologically Functional Bacterial Plasmids in Vitro." *Proceedings of the National Academy of Sciences of the United States of America* 70 (11): 3240–44.

Collins, F., and D. Galas. 1993. "A New Five-Year Plan for the U.S. Human Genome Project." *Science* 262 (5130): 43–46.

Collins, F., A. Patrinos, E. Jordan, A. Chakravarti, R. Gesteland, and L. Walters. 1998. "New Goals for the U.S. Human Genome Project: 1998-2003." *Science* 282 (5389): 682–89.

Cramp, S. 1998. *The Complete Birds of the Western Palearctic on CD-ROM*. Oxford University Press, Oxford.

Craxton, Molly. 1991. "Linear Amplification Sequencing, a Powerful Method for Sequencing DNA." *Methods* 3 (1): 20–26.

Crow, J. F., and M. Kimura. 1970. *An Introduction to Population Genetics Theory*. Edited by Ny: Harper And Row.

Deamer, David, Mark Akeson, and Daniel Branton. 2016. "Three Decades of Nanopore Sequencing." *Nature Biotechnology* 34 (5): 518–24.

DeAngelis, M. M., D. G. Wang, and T. L. Hawkins. 1995. "Solid-Phase Reversible Immobilization for the Isolation of PCR Products." *Nucleic Acids Research* 23 (22): 4742–43.

Dill, Lawrence M. 2002. "Model Systems in Behavioral Ecology: Integrating Conceptual, Theoretical, and Empirical Approaches . Monographs in Behavior and Ecology. Edited by Lee Alan Dugatkin. Princeton (New Jersey): Princeton University Press . 79.50 (hardcover); 35.00 (paper). Xxiii + 551 P; Ill.; Index. ISBN: 0–691–00652–0 (hc); 0–691–00653–9 (pb). 2001." *The Quarterly Review of Biology* 77 (3): 361–62.

Dor, Roi, Irby J. Lovette, Rebecca J. Safran, Shawn M. Billerman, Gernot H. Huber, Yoni Vortman, Arnon Lotem, et al. 2011. "Low Variation in the Polymorphic Clock Gene Poly-Q Region despite Population Genetic Structure across Barn Swallow (Hirundo Rustica) Populations." *PloS One* 6 (12): e28843.

Dor, Roi, Rebecca J. Safran, Yoni Vortman, Arnon Lotem, Andrew McGowan, Matthew R. Evans, and Irby J. Lovette. 2012. "Population Genetics and Morphological Comparisons of Migratory European (Hirundo Rustica Rustica) and Sedentary East-Mediterranean (Hirundo Rustica Transitiva) Barn Swallows." *The Journal of Heredity* 103 (1): 55–63.

Dressman, Devin, Hai Yan, Giovanni Traverso, Kenneth W. Kinzler, and Bert Vogelstein. 2003. "Transforming Single DNA Molecules into Fluorescent Magnetic Particles for Detection and Enumeration of Genetic Variations." *Proceedings of the National Academy of Sciences of the United States of America* 100 (15): 8817–22.

Drmanac, Radoje, Andrew B. Sparks, Matthew J. Callow, Aaron L. Halpern, Norman L. Burns, Bahram G. Kermani, Paolo Carnevali, et al. 2010. "Human Genome Sequencing Using Unchained Base Reads on

Self-Assembling DNA Nanoarrays." *Science* 327 (5961): 78–81.

Dunn, P. O., K. A. Hobson, and F. Liechti. 2015. "Assessing Costs of Carrying Geolocators Using Feather Corticosterone in Two Species of Aerial Insectivore." *Royal Society of Health Journal*. http://rsos.royalsocietypublishing.org/content/2/5/150004.abstract.

Dvornyk, Volodymyr, Oxana Vinogradova, and Eviatar Nevo. 2003. "Origin and Evolution of Circadian Clock Genes in Prokaryotes." *Proceedings of the National Academy of Sciences of the United States of America* 100 (5): 2495–2500.

Edman, Pehr. 1950. "Method for Determination of the Amino Acid Sequence in Peptides." *Acta Chemica Scandinavica* 4 (7): 283–93.

Edwards, A., H. Voss, P. Rice, A. Civitello, J. Stegemann, C. Schwager, J. Zimmermann, H. Erfle, C. T. Caskey, and W. Ansorge. 1990. "Automated DNA Sequencing of the Human HPRT Locus." *Genomics* 6 (4): 593–608.

Eichler, Evan E., Jonathan Flint, Greg Gibson, Augustine Kong, Suzanne M. Leal, Jason H. Moore, and Joseph H. Nadeau. 2010. "Missing Heritability and Strategies for Finding the Underlying Causes of Complex Disease." *Nature Reviews. Genetics* 11 (6): 446–50.

Eid, John, Adrian Fehr, Jeremy Gray, Khai Luong, John Lyle, Geoff Otto, Paul Peluso, et al. 2009. "Real-Time DNA Sequencing from Single Polymerase Molecules." *Science* 323 (5910): 133–38.

Ellegren, Hans. 2010. "Evolutionary Stasis: The Stable Chromosomes of Birds." *Trends in Ecology & Evolution* 25 (5): 283–91.

———. 2014. "Genome Sequencing and Population Genomics in Non-Model Organisms." *Trends in Ecology & Evolution* 29 (1): 51–63.

Ewing, B., and P. Green. 1998. "Base-Calling of Automated Sequencer Traces Using Phred. II. Error Probabilities." *Genome Research* 8 (3): 186–94.

Ewing, B., L. D. Hillier, M. C. Wendl, and P. Green. 1998. "Base-Calling of Automated Sequencer Traces usingPhred. I. Accuracy Assessment." *Genome Research*. http://genome.cshlp.org/content/8/3/175.short.

Fagny, Maud, Etienne Patin, David Enard, Luis B. Barreiro, Lluis Quintana-Murci, and Guillaume Laval. 2014. "Exploring the Occurrence of Classic Selective Sweeps in Humans Using Whole-Genome Sequencing Data Sets." *Molecular Biology and Evolution* 31 (7): 1850–68.

Fiers, W., R. Contreras, F. Duerinck, G. Haegeman, D. Iserentant, J. Merregaert, W. Min Jou, et al. 1976. "Complete Nucleotide Sequence of Bacteriophage MS2 RNA: Primary and Secondary Structure of the Replicase Gene." *Nature* 260 (5551): 500–507.

Fisher, R. A. 1930. *The Genetical Theory Of Natural Selection*. Edited by Oxford At The.

Fleischmann, R. D., M. D. Adams, O. White, R. A. Clayton, E. F. Kirkness, A. R. Kerlavage, C. J. Bult, J. F. Tomb, B. A. Dougherty, and J. M. Merrick. 1995. "Whole-Genome Random Sequencing and Assembly of Haemophilus Influenzae Rd." *Science* 269 (5223): 496–512.

Flusberg, Benjamin A., Dale R. Webster, Jessica H. Lee, Kevin J. Travers, Eric C. Olivares, Tyson A. Clark, Jonas Korlach, and Stephen W. Turner. 2010. "Direct Detection of DNA Methylation during Single-Molecule, Real-Time Sequencing." *Nature Methods* 7 (6): 461–65.

Fraser, C. M., S. Casjens, W. M. Huang, G. G. Sutton, R. Clayton, R. Lathigra, O. White, et al. 1997. "Genomic Sequence of a Lyme Disease Spirochaete, Borrelia Burgdorferi." *Nature* 390 (6660): 580–86.

Fu, Wenqing, and Joshua M. Akey. 2013. "Selection and Adaptation in the Human Genome." *Annual Review of Genomics and Human Genetics* 14 (July): 467–89.

Galeotti, P., N. Saino, and R. Sacchi. 1997. "Song Correlates with Social Context, Testosterone and Body Condition in Male Barn Swallows." *Animal: An International Journal of Animal Bioscience* 53 (4): 687–700.

Gao, Yong, Haibo Wang, Chao Liu, Honglong Chu, Dongqin Dai, Shengnan Song, Long Yu, et al. 2018. "De Novo Genome Assembly of the Red Silk Cotton Tree (Bombax Ceiba)." *GigaScience* 7 (5). https://doi.org/10.1093/gigascience/giy051.

Genome 10K Community of Scientists. 2009. "Genome 10K: A Proposal to Obtain Whole-Genome Sequence for 10,000 Vertebrate Species." *The Journal of Heredity* 100 (6): 659–74.

Giardine, Belinda, Cathy Riemer, Ross C. Hardison, Richard Burhans, Laura Elnitski, Prachi Shah, Yi Zhang, et al. 2005. "Galaxy: A Platform for Interactive Large-Scale Genome Analysis." *Genome Research* 15 (10): 1451–55.

Gibbs, Richard A., George M. Weinstock, Michael L. Metzker, Donna M. Muzny, Erica J. Sodergren, Steven Scherer, Graham Scott, et al. 2004. "Genome Sequence of the Brown Norway Rat Yields Insights into Mammalian Evolution." *Nature* 428 (6982): 493–521.

Gilbert, W., and A. Maxam. 1973. "The Nucleotide Sequence of the Lac Operator." *Proceedings of the National Academy of Sciences of the United States of America* 70 (12): 3581–84.

Gnerre, Sante, Iain Maccallum, Dariusz Przybylski, Filipe J. Ribeiro, Joshua N. Burton, Bruce J. Walker, Ted Sharpe, et al. 2011. "High-Quality Draft Assemblies of Mammalian Genomes from Massively Parallel Sequence Data." *Proceedings of the National Academy of Sciences of the United States of America* 108 (4): 1513–18.

Goebel, S. J., G. P. Johnson, M. E. Perkus, S. W. Davis, J. P. Winslow, and E. Paoletti. 1990. "The Complete DNA Sequence of Vaccinia Virus." *Virology* 179 (1): 247–66, 517–63.

Goffeau, A., B. G. Barrell, H. Bussey, R. W. Davis, B. Dujon, H. Feldmann, F. Galibert, et al. 1996. "Life with 6000 Genes." *Science* 274 (5287): 546, 563–67.

Gordon, D., C. Abajian, and P. Green. 1998. "Consed: A Graphical Tool for Sequence Finishing." *Genome Research* 8 (3): 195–202.

Gravel, Simon, Brenna M. Henn, Ryan N. Gutenkunst, Amit R. Indap, Gabor T. Marth, Andrew G. Clark, Fuli Yu, Richard A. Gibbs, 1000 Genomes Project, and Carlos D. Bustamante. 2011. "Demographic History and Rare Allele Sharing among Human Populations." *Proceedings of the National Academy of Sciences of the United States of America* 108 (29): 11983–88.

Gravel, Simon, Fouad Zakharia, Andres Moreno-Estrada, Jake K. Byrnes, Marina Muzzio, Juan L. Rodriguez-Flores, Eimear E. Kenny, et al. 2013. "Reconstructing Native American Migrations from Whole-Genome and Whole-Exome Data." *PLoS Genetics* 9 (12): e1004023.

Greenleaf, William J., and Arend Sidow. 2014. "The Future of Sequencing: Convergence of Intelligent Design and Market Darwinism." *Genome Biology* 15 (3): 303.

Green, Richard E., Johannes Krause, Adrian W. Briggs, Tomislav Maricic, Udo Stenzel, Martin Kircher, Nick Patterson, et al. 2010. "A Draft Sequence of the Neandertal Genome." *Science* 328 (5979): 710–22.

Grossman, Sharon R., Kristian G. Andersen, Ilya Shlyakhter, Shervin Tabrizi, Sarah Winnicki, Angela Yen, Daniel J. Park, et al. 2013. "Identifying Recent Adaptations in Large-Scale Genomic Data." *Cell* 152 (4): 703–13.

Gutenkunst, Ryan N., Ryan D. Hernandez, Scott H. Williamson, and Carlos D. Bustamante. 2009. "Inferring the Joint Demographic History of Multiple Populations from Multidimensional SNP Frequency Data." *PLoS Genetics* 5 (10): e1000695.

Hall, Jeffrey C. 2003. "Genetics and Molecular Biology of Rhythms in Drosophila and Other Insects." *Advances in Genetics* 48: 1–280.

Harris, Kelley, and Rasmus Nielsen. 2013. "Inferring Demographic History from a Spectrum of Shared Haplotype Lengths." *PLoS Genetics* 9 (6): e1003521.

Harris, Timothy D., Phillip R. Buzby, Hazen Babcock, Eric Beer, Jayson Bowers, Ido Braslavsky, Marie Causey, et al. 2008. "Single-Molecule DNA Sequencing of a Viral Genome." *Science* 320 (5872): 106–

9.

Hastie, Alex R., Lingli Dong, Alexis Smith, Jeff Finklestein, Ernest T. Lam, Naxin Huo, Han Cao, et al. 2013. "Rapid Genome Mapping in Nanochannel Arrays for Highly Complete and Accurate de Novo Sequence Assembly of the Complex Aegilops Tauschii Genome." *PloS One* 8 (2): e55864.

Hayden, Erika Check. 2014. "Is the $1,000 Genome for Real?" *Nature News*. https://www.nature.com/articles/nature.2014.14530.

Head, Steven R., H. Kiyomi Komori, Sarah A. LaMere, Thomas Whisenant, Filip Van Nieuwerburgh, Daniel R. Salomon, and Phillip Ordoukhanian. 2014. "Library Construction for next-Generation Sequencing: Overviews and Challenges." *BioTechniques* 56 (2): 61–64, 66, 68, passim.

Heather, James M., and Benjamin Chain. 2016. "The Sequence of Sequencers: The History of Sequencing DNA." *Genomics* 107 (1): 1–8.

Helbig, A. 1996. "Genetic Basis, Mode of Inheritance and Evolutionary Changes of Migratory Directions in Palaearctic Warblers (Aves: Sylviidae)." *The Journal of Experimental Biology* 199 (Pt 1): 49–55.

Henson, Joseph, German Tischler, and Zemin Ning. 2012. "Next-Generation Sequencing and Large Genome Assemblies." *Pharmacogenomics* 13 (8): 901–15.

Hogan-Warburg, Alida Johanna. 1966. *Social Behavior of the Ruff, Philomachus Pugnax (L.)*. Brill Archive.

Holley, R. W., J. Apgar, G. A. Everett, J. T. Madison, M. Marquisee, S. H. Merrill, J. R. Penswick, and A. Zamir. 1965. "STRUCTURE OF A RIBONUCLEIC ACID." *Science* 147 (3664): 1462–65.

Horton, Brent M., Ignacio T. Moore, and Donna L. Maney. 2014. "New Insights into the Hormonal and Behavioural Correlates of Polymorphism in White-Throated Sparrows, Zonotrichia Albicollis." *Animal Behaviour* 93 (July): 207–19.

Howe, Kerstin, Matthew D. Clark, Carlos F. Torroja, James Torrance, Camille Berthelot, Matthieu Muffato, John E. Collins, et al. 2013. "The Zebrafish Reference Genome Sequence and Its Relationship to the Human Genome." *Nature* 496 (7446): 498–503.

Huang, Shuo, Jin He, Shuai Chang, Peiming Zhang, Feng Liang, Shengqin Li, Michael Tuchband, Alexander Fuhrmann, Robert Ros, and Stuart Lindsay. 2010. "Identifying Single Bases in a DNA Oligomer with Electron Tunnelling." *Nature Nanotechnology* 5 (12): 868–73.

Huang, Yu-Feng, Sheng-Chung Chen, Yih-Shien Chiang, Tzu-Han Chen, and Kuo-Ping Chiu. 2012. "Palindromic Sequence Impedes Sequencing-by-Ligation Mechanism." *BMC Systems Biology* 6 Suppl 2 (December): S10.

Hubbard, T., D. Barker, E. Birney, G. Cameron, Y. Chen, L. Clark, T. Cox, et al. 2002. "The Ensembl Genome Database Project." *Nucleic Acids Research* 30 (1): 38–41.

Huddleston, John, Mark J. P. Chaisson, Karyn Meltz Steinberg, Wes Warren, Kendra Hoekzema, David Gordon, Tina A. Graves-Lindsay, et al. 2017. "Discovery and Genotyping of Structural Variation from Long-Read Haploid Genome Sequence Data." *Genome Research* 27 (5): 677–85.

Hunt, Martin, Chris Newbold, Matthew Berriman, and Thomas D. Otto. 2014. "A Comprehensive Evaluation of Assembly Scaffolding Tools." *Genome Biology* 15 (3): R42.

Husby, Arild, Takeshi Kawakami, Lars Rönnegård, Linnéa Smeds, Hans Ellegren, and Anna Qvarnström. 2015. "Genome-Wide Association Mapping in a Wild Avian Population Identifies a Link between Genetic and Phenotypic Variation in a Life-History Trait." *Proceedings. Biological Sciences / The Royal Society* 282 (1806): 20150156.

Hyman, E. D. 1988. "A New Method of Sequencing DNA." *Analytical Biochemistry* 174 (2): 423–36.

Ingolia, Nicholas T., Sina Ghaemmaghami, John R. S. Newman, and Jonathan S. Weissman. 2009. "Genome-Wide Analysis in Vivo of Translation with Nucleotide Resolution Using Ribosome Profiling." *Science* 324 (5924): 218–23.

International Human Genome Sequencing Consortium. 2004. "Finishing the Euchromatic Sequence of the Human Genome." *Nature* 431 (7011): 931–45.

International Rice Genome Sequencing Project. 2005. “The Map-Based Sequence of the Rice Genome.” *Nature* 436 (7052): 793–800.

Jackson, D. A., and R. H. Symons. 1972. “Biochemical Method for Inserting New Genetic Information into DNA of Simian Virus 40: Circular SV40 DNA Molecules Containing Lambda Phage Genes and the ….” *Proceedings of the*. http://www.pnas.org/content/69/10/2904.short.

Jain, Miten, Sergey Koren, Karen H. Miga, Josh Quick, Arthur C. Rand, Thomas A. Sasani, John R. Tyson, et al. 2018. “Nanopore Sequencing and Assembly of a Human Genome with Ultra-Long Reads.” *Nature Biotechnology* 36 (4): 338–45.

Jarvis, Erich D., Siavash Mirarab, Andre J. Aberer, Bo Li, Peter Houde, Cai Li, Simon Y. W. Ho, et al. 2014. “Whole-Genome Analyses Resolve Early Branches in the Tree of Life of Modern Birds.” *Science* 346 (6215): 1320–31.

Jenni, L., and R. Winkler. 1994. “Moult and Ageing of European Passerines.–224 S.” *Academic Press, London*.

Jha, Pankaj, Dongsheng Lu, and Shuhua Xu. 2015. “Natural Selection and Functional Potentials of Human Noncoding Elements Revealed by Analysis of Next Generation Sequencing Data.” *PloS One* 10 (6): e0129023.

Johnson, David S., Ali Mortazavi, Richard M. Myers, and Barbara Wold. 2007. “Genome-Wide Mapping of in Vivo Protein-DNA Interactions.” *Science* 316 (5830): 1497–1502.

Joron, Mathieu, Lise Frezal, Robert T. Jones, Nicola L. Chamberlain, Siu F. Lee, Christoph R. Haag, Annabel Whibley, et al. 2011. “Chromosomal Rearrangements Maintain a Polymorphic Supergene Controlling Butterfly Mimicry.” *Nature* 477 (7363): 203–6.

Josephs, Emily B., John R. Stinchcombe, and Stephen I. Wright. 2017. “What Can Genome-Wide Association Studies Tell Us about the Evolutionary Forces Maintaining Genetic Variation for Quantitative Traits?” *The New Phytologist* 214 (1): 21–33.

Jukema, Joop, and Theunis Piersma. 2006. “Permanent Female Mimics in a Lekking Shorebird.” *Biology Letters* 2 (2): 161–64.

Kadi, Farida, Dominique Mouchiroud, Georgette Sabeur, and Giorgio Bernardi. 1993. “The Compositional Patterns of the Avian Genomes and Their Evolutionary Implications.” *Journal of Molecular Evolution* 37 (5): 544–51.

Kawashima, Eric, Laurent Farinelli, and Pascal Mayer. 1998. Method of nucleic acid amplification. WIPO 1998044151:A1. *World Patent*, filed April 1, 1998, and issued October 8, 1998.

Kent, W. James. 2002. “BLAT—The BLAST-Like Alignment Tool.” *Genome Research* 12 (4): 656–64.

Kent, W. James, Charles W. Sugnet, Terrence S. Furey, Krishna M. Roskin, Tom H. Pringle, Alan M. Zahler, and David Haussler. 2002. “The Human Genome Browser at UCSC.” *Genome Research* 12 (6): 996–1006.

Khurana, Ekta, Yao Fu, Vincenza Colonna, Xinmeng Jasmine Mu, Hyun Min Kang, Tuuli Lappalainen, Andrea Sboner, et al. 2013. “Integrative Annotation of Variants from 1092 Humans: Application to Cancer Genomics.” *Science* 342 (6154): 1235587.

Klenk, H. P., R. A. Clayton, J. F. Tomb, O. White, K. E. Nelson, K. A. Ketchum, R. J. Dodson, et al. 1997. “The Complete Genome Sequence of the Hyperthermophilic, Sulphate-Reducing Archaeon Archaeoglobus Fulgidus.” *Nature* 390 (6658): 364–70.

Koepfli, Klaus-Peter, Benedict Paten, Genome 10K Community of Scientists, and Stephen J. O’Brien. 2015. “The Genome 10K Project: A Way Forward.” *Annual Review of Animal Biosciences* 3: 57–111.

Koren, Sergey, Brian P. Walenz, Konstantin Berlin, Jason R. Miller, Nicholas H. Bergman, and Adam M. Phillippy. 2017. “Canu: Scalable and Accurate Long-Read Assembly via Adaptive K-Mer Weighting and Repeat Separation.” *Genome Research* 27 (5): 722–36.

Korlach, Jonas, Gregory Gedman, Sarah B. Kingan, Chen-Shan Chin, Jason T. Howard, Jean-Nicolas Audet,

Lindsey Cantin, and Erich D. Jarvis. 2017. “De Novo PacBio Long-Read and Phased Avian Genome Assemblies Correct and Add to Reference Genes Generated with Intermediate and Short Reads.” *GigaScience* 6 (10): 1–16.

Kunte, K., W. Zhang, A. Tenger-Trolander, D. H. Palmer, A. Martin, R. D. Reed, S. P. Mullen, and M. R. Kronforst. 2014. “Doublesex Is a Mimicry Supergene.” *Nature* 507 (7491): 229–32.

Lam, Ernest T., Alex Hastie, Chin Lin, Dean Ehrlich, Somes K. Das, Michael D. Austin, Paru Deshpande, et al. 2012. “Genome Mapping on Nanochannel Arrays for Structural Variation Analysis and Sequence Assembly.” *Nature Biotechnology* 30 (8): 771–76.

Lamichhaney, Sangeet, Guangyi Fan, Fredrik Widemo, Ulrika Gunnarsson, Doreen Schwochow Thalmann, Marc P. Hoeppner, Susanne Kerje, et al. 2016. “Structural Genomic Changes Underlie Alternative Reproductive Strategies in the Ruff (Philomachus Pugnax).” *Nature Genetics* 48 (1): 84–88.

Lander, E. S., L. M. Linton, B. Birren, C. Nusbaum, M. C. Zody, J. Baldwin, K. Devon, et al. 2001. “Initial Sequencing and Analysis of the Human Genome.” *Nature* 409 (6822): 860–921.

Langmead, Ben, Cole Trapnell, Mihai Pop, and Steven L. Salzberg. 2009. “Ultrafast and Memory-Efficient Alignment of Short DNA Sequences to the Human Genome.” *Genome Biology* 10 (3): R25.

Lappalainen, Tuuli, Michael Sammeth, Marc R. Friedländer, Peter A. C. ’t Hoen, Jean Monlong, Manuel A. Rivas, Mar Gonzàlez-Porta, et al. 2013. “Transcriptome and Genome Sequencing Uncovers Functional Variation in Humans.” *Nature* 501 (7468): 506–11.

Lee, L. G., C. R. Connell, S. L. Woo, R. D. Cheng, B. F. McArdle, C. W. Fuller, N. D. Halloran, and R. K. Wilson. 1992. “DNA Sequencing with Dye-Labeled Terminators and T7 DNA Polymerase: Effect of Dyes and dNTPs on Incorporation of Dye-Terminators and Probability Analysis of Termination Fragments.” *Nucleic Acids Research* 20 (10): 2471–83.

Levene, M. J., J. Korlach, S. W. Turner, M. Foquet, H. G. Craighead, and W. W. Webb. 2003. “Zero-Mode Waveguides for Single-Molecule Analysis at High Concentrations.” *Science* 299 (5607): 682–86.

Levy, Samuel, Granger Sutton, Pauline C. Ng, Lars Feuk, Aaron L. Halpern, Brian P. Walenz, Nelson Axelrod, et al. 2007. “The Diploid Genome Sequence of an Individual Human.” *PLoS Biology* 5 (10): e254.

Lewin, Harris A., Gene E. Robinson, W. John Kress, William J. Baker, Jonathan Coddington, Keith A. Crandall, Richard Durbin, et al. 2018a. “Earth BioGenome Project: Sequencing Life for the Future of Life.” *Proceedings of the National Academy of Sciences of the United States of America* 115 (17): 4325–33.

Liechti, F., C. Scandolara, and D. Rubolini. 2015. “Timing of Migration and Residence Areas during the Non-breeding Period of Barn Swallows Hirundo Rustica in Relation to Sex and Population.” *Journal of Avian*. https://onlinelibrary.wiley.com/doi/abs/10.1111/jav.00485.

Liedvogel, Miriam, Susanne Åkesson, and Staffan Bensch. 2011. “The Genetics of Migration on the Move.” *Trends in Ecology & Evolution* 26 (11): 561–69.

Li, Heng, and Richard Durbin. 2009. “Fast and Accurate Short Read Alignment with Burrows–Wheeler Transform.” *Bioinformatics* 25 (14): 1754–60.

Li, Heng, Bob Handsaker, Alec Wysoker, Tim Fennell, Jue Ruan, Nils Homer, Gabor Marth, Goncalo Abecasis, Richard Durbin, and 1000 Genome Project Data Processing Subgroup. 2009. “The Sequence Alignment/Map Format and SAMtools.” *Bioinformatics* 25 (16): 2078–79.

Li, Ruiqiang, Hongmei Zhu, Jue Ruan, Wubin Qian, Xiaodong Fang, Zhongbin Shi, Yingrui Li, et al. 2010. “De Novo Assembly of Human Genomes with Massively Parallel Short Read Sequencing.” *Genome Research* 20 (2): 265–72.

Lister, Ryan, Ronan C. O’Malley, Julian Tonti-Filippini, Brian D. Gregory, Charles C. Berry, A. Harvey Millar, and Joseph R. Ecker. 2008. “Highly Integrated Single-Base Resolution Maps of the Epigenome in Arabidopsis.” *Cell* 133 (3): 523–36.

Loman, Nicholas J., Raju V. Misra, Timothy J. Dallman, Chrystala Constantinidou, Saheer E. Gharbia, John Wain, and Mark J. Pallen. 2012. "Performance Comparison of Benchtop High-Throughput Sequencing Platforms." *Nature Biotechnology* 30 (5): 434–39.

Mak, Angel C. Y., Yvonne Y. Y. Lai, Ernest T. Lam, Tsz-Piu Kwok, Alden K. Y. Leung, Annie Poon, Yulia Mostovoy, et al. 2016. "Genome-Wide Structural Variation Detection by Genome Mapping on Nanochannel Arrays." *Genetics* 202 (1): 351–62.

Maniatis, T., A. Jeffrey, and H. van deSande. 1975. "Chain Length Determination of Small Double- and Single-Stranded DNA Molecules by Polyacrylamide Gel Electrophoresis." *Biochemistry* 14 (17): 3787–94.

Manrao, Elizabeth A., Ian M. Derrington, Andrew H. Laszlo, Kyle W. Langford, Matthew K. Hopper, Nathaniel Gillgren, Mikhail Pavlenok, Michael Niederweis, and Jens H. Gundlach. 2012. "Reading DNA at Single-Nucleotide Resolution with a Mutant MspA Nanopore and phi29 DNA Polymerase." *Nature Biotechnology* 30 (4): 349–53.

Margulies, Marcel, Michael Egholm, William E. Altman, Said Attiya, Joel S. Bader, Lisa A. Bemben, Jan Berka, et al. 2005. "Genome Sequencing in Microfabricated High-Density Picolitre Reactors." *Nature* 437 (7057): 376–80.

Massung, R. F., L. I. Liu, J. Qi, J. C. Knight, T. E. Yuran, A. R. Kerlavage, J. M. Parsons, J. C. Venter, and J. J. Esposito. 1994. "Analysis of the Complete Genome of Smallpox Variola Major Virus Strain Bangladesh-1975." *Virology* 201 (2): 215–40.

Matyjasiak, P., D. Rubolini, M. Romano, and N. Saino. 2016. "No Short-Term Effects of Geolocators on Flight Performance of an Aerial Insectivorous Bird, the Barn Swallow (Hirundo Rustica)." *Journal of Ornithology / DO-G* 157 (3): 653–61.

Maxam, A. M., and W. Gilbert. 1977. "A New Method for Sequencing DNA." *Proceedings of the National Academy of Sciences of the United States of America* 74 (2): 560–64.

Mayor, S. 1999. "First Human Chromosome Is Sequenced." *BMJ* 319 (7223): 1453A.

McKenna, Aaron, Matthew Hanna, Eric Banks, Andrey Sivachenko, Kristian Cibulskis, Andrew Kernytsky, Kiran Garimella, et al. 2010. "The Genome Analysis Toolkit: A MapReduce Framework for Analyzing next-Generation DNA Sequencing Data." *Genome Research* 20 (9): 1297–1303.

McKernan, Kevin Judd, Heather E. Peckham, Gina L. Costa, Stephen F. McLaughlin, Yutao Fu, Eric F. Tsung, Christopher R. Clouser, et al. 2009. "Sequence and Structural Variation in a Human Genome Uncovered by Short-Read, Massively Parallel Ligation Sequencing Using Two-Base Encoding." *Genome Research* 19 (9): 1527–41.

Merker, Jason D., Aaron M. Wenger, Tam Sneddon, Megan Grove, Zachary Zappala, Laure Fresard, Daryl Waggott, et al. 2018. "Long-Read Genome Sequencing Identifies Causal Structural Variation in a Mendelian Disease." *Genetics in Medicine: Official Journal of the American College of Medical Genetics* 20 (1): 159–63.

Messing, J., R. Crea, and P. H. Seeburg. 1981. "A System for Shotgun DNA Sequencing." *Nucleic Acids Research* 9 (2): 309–21.

Metzger, Julia, Matthias Karwath, Raul Tonda, Sergi Beltran, Lídia Águeda, Marta Gut, Ivo Glynne Gut, and Ottmar Distl. 2015. "Runs of Homozygosity Reveal Signatures of Positive Selection for Reproduction Traits in Breed and Non-Breed Horses." *BMC Genomics* 16 (October): 764.

Meyer, Matthias, Martin Kircher, Marie-Theres Gansauge, Heng Li, Fernando Racimo, Swapan Mallick, Joshua G. Schraiber, et al. 2012. "A High-Coverage Genome Sequence from an Archaic Denisovan Individual." *Science* 338 (6104): 222–26.

Min Jou, W., G. Haegeman, M. Ysebaert, and W. Fiers. 1972. "Nucleotide Sequence of the Gene Coding for the Bacteriophage MS2 Coat Protein." *Nature* 237 (5350): 82–88.

Mitra, R. D., and G. M. Church. 1999. "In Situ Localized Amplification and Contact Replication of Many

Individual DNA Molecules." *Nucleic Acids Research* 27 (24): e34.
Mitra, R. D., Jay Shendure, Jerzy Olejnik, Edyta-Krzymanska-Olejnik, and George M. Church. 2003. "Fluorescent in Situ Sequencing on Polymerase Colonies." *Analytical Biochemistry* 320 (1): 55–65.
Møller, A. P. 1988. "Female Choice Selects for Male Sexual Tail Ornaments in the Monogamous Swallow." *Nature* 332 (April): 640.
———. 1990. "Parasites and Sexual Selection: Current Status of the Hamilton and Zuk Hypothesis." *Journal of Evolutionary Biology* 3 (5-6): 319–28.
———. 1994. *Sexual Selection and the Barn Swallow*. Oxford University Press, Oxford.
Moller, A. P. 2001. "Heritability of Arrival Date in a Migratory Bird." *Proceedings of the Royal Society of London B: Biological Sciences* 268 (1463): 203–6.
Møller, A. P. 2007. "Tardy Females, Impatient Males: Protandry and Divergent Selection on Arrival Date in the Two Sexes of the Barn Swallow." *Behavioral Ecology and Sociobiology* 61 (8): 1311–19.
Møller, A. P., A. Barbosa, J. J. Cuervo, F. de Lope, S. Merino, and N. Saino. 1998. "Sexual Selection and Tail Streamers in the Barn Swallow." *Proceedings of the Royal Society of London B: Biological Sciences* 265 (1394): 409–14.
Møller, A. P., J. Brohede, J. J. Cuervo, F. de Lope, and C. Primmer. 2003. "Extrapair Paternity in Relation to Sexual Ornamentation, Arrival Date, and Condition in a Migratory Bird." *Behavioral Ecology: Official Journal of the International Society for Behavioral Ecology* 14 (5): 707–12.
Møller, A. P., Y. Chabi, J. J. Cuervo, F. Lope, J. Kilpimaa, M. Kose, P. Matyjasiak, et al. 2006. "An Analysis of Continent-Wide Patterns of Sexual Selection in a Passerine Bird." *Evolution; International Journal of Organic Evolution* 60 (4): 856–68.
Moller, A. P., F. de Lope, and N. Saino. 1995. "Sexual Selection in the Barn Swallow Hirundo Rustica. VI. Aerodynamic Adaptations." *Journal of Evolutionary Biology* 8 (6): 671–87.
Møller, A. P., R. Martinelli, and N. Saino. 2004. "Genetic Variation in Infestation with a Directly Transmitted Ectoparasite." *Journal of Evolutionary Biology* 17 (1): 41–47.
Møller, A. P., N. Saino, G. Taramino, P. Galeotti, and S. Ferrario. 1998. "Paternity and Multiple Signaling: Effects of a Secondary Sexual Character and Song on Paternity in the Barn Swallow." *The American Naturalist* 151 (3): 236–42.
Morgan, T. H. 1911. "THE ORIGIN OF FIVE MUTATIONS IN EYE COLOR IN DROSOPHILA AND THEIR MODES OF INHERITANCE." *Science* 33 (849): 534–37.
Morgulis, Aleksandr, E. Michael Gertz, Alejandro A. Schäffer, and Richa Agarwala. 2006. "WindowMasker: Window-Based Masker for Sequenced Genomes." *Bioinformatics* 22 (2): 134–41.
Mortazavi, Ali, Brian A. Williams, Kenneth McCue, Lorian Schaeffer, and Barbara Wold. 2008. "Mapping and Quantifying Mammalian Transcriptomes by RNA-Seq." *Nature Methods* 5 (7): 621–28.
Mostovoy, Yulia, Michal Levy-Sakin, Jessica Lam, Ernest T. Lam, Alex R. Hastie, Patrick Marks, Joyce Lee, et al. 2016. "A Hybrid Approach for de Novo Human Genome Sequence Assembly and Phasing." *Nature Methods* 13 (7): 587–90.
Mouse Genome Sequencing Consortium, Robert H. Waterston, Kerstin Lindblad-Toh, Ewan Birney, Jane Rogers, Josep F. Abril, Pankaj Agarwal, et al. 2002. "Initial Sequencing and Comparative Analysis of the Mouse Genome." *Nature* 420 (6915): 520–62.
Mukherjee, Supratim, Marcel Huntemann, Natalia Ivanova, Nikos C. Kyrpides, and Amrita Pati. 2015. "Large-Scale Contamination of Microbial Isolate Genomes by Illumina PhiX Control." *Standards in Genomic Sciences* 10 (March): 18.
Mullis, K. B. 1990. "The Unusual Origin of the Polymerase Chain Reaction." *Scientific American* 262 (4): 56–61, 64–65.
Murray, J. C., K. H. Buetow, J. L. Weber, S. Ludwigsen, T. Scherpbier-Heddema, F. Manion, J. Quillen, V. C. Sheffield, S. Sunden, and G. M. Duyk. 1994. "A Comprehensive Human Linkage Map with

Centimorgan Density. Cooperative Human Linkage Center (CHLC).” *Science* 265 (5181): 2049–54.
Murray, V. 1989. “Improved Double-Stranded DNA Sequencing Using the Linear Polymerase Chain Reaction.” *Nucleic Acids Research* 17 (21): 8889.
Myers, E. W., G. G. Sutton, A. L. Delcher, I. M. Dew, D. P. Fasulo, M. J. Flanigan, S. A. Kravitz, et al. 2000. “A Whole-Genome Assembly of Drosophila.” *Science* 287 (5461): 2196–2204.
Nagalakshmi, Ugrappa, Zhong Wang, Karl Waern, Chong Shou, Debasish Raha, Mark Gerstein, and Michael Snyder. 2008. “The Transcriptional Landscape of the Yeast Genome Defined by RNA Sequencing.” *Science* 320 (5881): 1344–49.
Narum, Shawn R., Alex Di Genova, Steven J. Micheletti, and Alejandro Maass. 2018. “Genomic Variation Underlying Complex Life-History Traits Revealed by Genome Sequencing in Chinook Salmon.” *Proc. R. Soc. B* 285 (1883): 20180935.
Nelson, F. Kenneth, Michael Snyder, Andrew F. Gardner, Cynthia L. Hendrickson, Jay A. Shendure, Gregory J. Porreca, George M. Church, et al. 2011. “Introduction and Historical Overview of DNA Sequencing.” *Current Protocols in Molecular Biology / Edited by Frederick M. Ausubel ... [et Al.]* 96 (1): 7–0.
Nelson, John R., Yuyang Christine Cai, Theresa L. Giesler, Joseph W. Farchaus, Shanmuuga T. Sundaram, Maria Ortiz-Rivera, Lou P. Hosta, et al. 2002. “TempliPhi, phi29 DNA Polymerase Based Rolling Circle Amplification of Templates for DNA Sequencing.” *BioTechniques* Suppl (June): 44–47.
Ng, Sarah B., Emily H. Turner, Peggy D. Robertson, Steven D. Flygare, Abigail W. Bigham, Choli Lee, Tristan Shaffer, et al. 2009. “Targeted Capture and Massively Parallel Sequencing of 12 Human Exomes.” *Nature* 461 (7261): 272–76.
Nishikawa, Hideki, Takuro Iijima, Rei Kajitani, Junichi Yamaguchi, Toshiya Ando, Yutaka Suzuki, Sumio Sugano, et al. 2015. “A Genetic Mechanism for Female-Limited Batesian Mimicry in Papilio Butterfly.” *Nature Genetics* 47 (4): 405–9.
Nowoshilow, Sergej, Siegfried Schloissnig, Ji-Feng Fei, Andreas Dahl, Andy W. C. Pang, Martin Pippel, Sylke Winkler, et al. 2018. “The Axolotl Genome and the Evolution of Key Tissue Formation Regulators.” *Nature* 554 (7690): 50–55.
Nozaki, Hisayoshi, Hiroyoshi Takano, Osami Misumi, Kimihiro Terasawa, Motomichi Matsuzaki, Shinichiro Maruyama, Keiji Nishida, et al. 2007. “A 100%-Complete Sequence Reveals Unusually Simple Genomic Features in the Hot-Spring Red Alga Cyanidioschyzon Merolae.” *BMC Biology* 5 (July): 28.
Nyrén, P. 1987. “Enzymatic Method for Continuous Monitoring of DNA Polymerase Activity.” *Analytical Biochemistry* 167 (2): 235–38.
Nyrén, P., B. Pettersson, and M. Uhlén. 1993. “Solid Phase DNA Minisequencing by an Enzymatic Luminometric Inorganic Pyrophosphate Detection Assay.” *Analytical Biochemistry* 208 (1): 171–75.
O’Malley, Kathleen G., Michael J. Ford, and Jeffrey J. Hard. 2010. “Clock Polymorphism in Pacific Salmon: Evidence for Variable Selection along a Latitudinal Gradient.” *Proceedings. Biological Sciences / The Royal Society* 277 (1701): 3703–14.
Ost, T. B. 2006. Improved polymerases. *Patent WO2006120433*, issued 2006.
Paez, J. Guillermo, Ming Lin, Rameen Beroukhim, Jeffrey C. Lee, Xiaojun Zhao, Daniel J. Richter, Stacey Gabriel, et al. 2004. “Genome Coverage and Sequence Fidelity of phi29 Polymerase-Based Multiple Strand Displacement Whole Genome Amplification.” *Nucleic Acids Research* 32 (9): e71.
Parolini, Marco, A. Romano, Lela Khoriauli, Solomon G. Nergadze, Manuela Caprioli, Diego Rubolini, Marco Santagostino, Nicola Saino, and Elena Giulotto. 2015. “Early-Life Telomere Dynamics Differ between the Sexes and Predict Growth in the Barn Swallow (Hirundo Rustica).” *PloS One* 10 (11): e0142530.
Pendleton, Matthew, Robert Sebra, Andy Wing Chun Pang, Ajay Ummat, Oscar Franzen, Tobias Rausch,

Adrian M. Stütz, et al. 2015. “Assembly and Diploid Architecture of an Individual Human Genome via Single-Molecule Technologies.” *Nature Methods* 12 (8): 780–86.

Pennisi, Elizabeth. 2017. “Sequencing All Life Captivates Biologists.” *Science* 355 (6328): 894–95.

Pevzner, P. A., H. Tang, and M. S. Waterman. 2001. “An Eulerian Path Approach to DNA Fragment Assembly.” *Proceedings of the National Academy of Sciences of the United States of America* 98 (17): 9748–53.

Pickrell, Joseph K., John C. Marioni, Athma A. Pai, Jacob F. Degner, Barbara E. Engelhardt, Everlyne Nkadori, Jean-Baptiste Veyrieras, Matthew Stephens, Yoav Gilad, and Jonathan K. Pritchard. 2010. “Understanding Mechanisms Underlying Human Gene Expression Variation with RNA Sequencing.” *Nature* 464 (7289): 768–72.

Poelstra, J. W., H. Ellegren, and J. B. W. Wolf. 2013. “An Extensive Candidate Gene Approach to Speciation: Diversity, Divergence and Linkage Disequilibrium in Candidate Pigmentation Genes across the European Crow Hybrid Zone.” *Heredity* 111 (6): 467–73.

Portugal, Franklin H., and Jack S. Cohen. 1977. *A Century of DNA : A History of the Discovery of the Structure and Function of the Genetic Substance*. MIT Press.

Primmer, C. R., A. P. Møller, and H. Ellegren. 1995. “Resolving Genetic Relationships with Microsatellite Markers: A Parentage Testing System for the Swallow Hirundo Rustica.” *Molecular Ecology* 4 (4): 493–98.

Prober, J. M., G. L. Trainor, R. J. Dam, F. W. Hobbs, C. W. Robertson, R. J. Zagursky, A. J. Cocuzza, M. A. Jensen, and K. Baumeister. 1987. “A System for Rapid DNA Sequencing with Fluorescent Chain-Terminating Dideoxynucleotides.” *Science* 238 (4825): 336–41.

Pulido, F., P. Berthold, G. Mohr, and U. Querner. 2001. “Heritability of the Timing of Autumn Migration in a Natural Bird Population.” *Proceedings. Biological Sciences / The Royal Society* 268 (1470): 953–59.

Quick, Joshua, Nicholas J. Loman, Sophie Duraffour, Jared T. Simpson, Ettore Severi, Lauren Cowley, Joseph Akoi Bore, et al. 2016. “Real-Time, Portable Genome Sequencing for Ebola Surveillance.” *Nature* 530 (7589): 228–32.

Reardon, Sara. 2015. “Giant Study Poses DNA Data-Sharing Dilemma.” *Nature* 525 (7567): 16–17.

Regalado, Antonio. 2014. “EmTech: Illumina Says 228,000 Human Genomes Will Be Sequenced This Year.” *Technology Review* 24.

Rhoads, Anthony, and Kin Fai Au. 2015. “PacBio Sequencing and Its Applications.” *Genomics, Proteomics & Bioinformatics* 13 (5): 278–89.

Roberts, L. 1992. “Why Watson Quit as Project Head.” *Science* 256 (5055): 301–2.

Roberts, Richard J., Mauricio O. Carneiro, and Michael C. Schatz. 2013. “The Advantages of SMRT Sequencing.” *Genome Biology* 14 (7): 405.

Roberts, Richard J., Tamas Vincze, Janos Posfai, and Dana Macelis. 2015. “REBASE--a Database for DNA Restriction and Modification: Enzymes, Genes and Genomes.” *Nucleic Acids Research* 43 (Database issue): D298–99.

Robinson, James T., Helga Thorvaldsdóttir, Wendy Winckler, Mitchell Guttman, Eric S. Lander, Gad Getz, and Jill P. Mesirov. 2011. “Integrative Genomics Viewer.” *Nature Biotechnology* 29 (1): 24–26.

Robinson, Matthew R., Anna W. Santure, Isabelle Decauwer, Ben C. Sheldon, and Jon Slate. 2013. “Partitioning of Genetic Variation across the Genome Using Multimarker Methods in a Wild Bird Population.” *Molecular Ecology* 22 (15): 3963–80.

Romano, A., Alessandra Costanzo, Manuela Caprioli, Marco Parolini, Roberto Ambrosini, Diego Rubolini, and Nicola Saino. 2016. “Better-Surviving Barn Swallow Mothers Produce More and Better-Surviving Sons.” *Evolution; International Journal of Organic Evolution* 70 (5): 1120–28.

Romano, A., Alessandra Costanzo, Diego Rubolini, Nicola Saino, and Anders Pape Møller. 2017. “Geographical and Seasonal Variation in the Intensity of Sexual Selection in the Barn Swallow Hirundo

Rustica: A Meta-Analysis." *Biological Reviews of the Cambridge Philosophical Society* 92 (3): 1582–1600.

Romano, A., M. Romano, M. Caprioli, A. Costanzo, M. Parolini, D. Rubolini, and N. Saino. 2015. "Sex Allocation according to Multiple Sexually Dimorphic Traits of Both Parents in the Barn Swallow (Hirundo Rustica)." *Journal of Evolutionary Biology* 28 (6): 1234–47.

Romano, A., N. Saino, and A. P. Møller. 2017. "Viability and Expression of Sexual Ornaments in the Barn Swallow Hirundo Rustica: A Meta-Analysis." *Journal of Evolutionary Biology* 30 (10): 1929–35.

Ronaghi, M., S. Karamohamed, B. Pettersson, M. Uhlén, and P. Nyrén. 1996. "Real-Time DNA Sequencing Using Detection of Pyrophosphate Release." *Analytical Biochemistry* 242 (1): 84–89.

Rönn, Jan A. C. von, Aaron B. A. Shafer, and Jochen B. W. Wolf. 2016. "Disruptive Selection without Genome-Wide Evolution across a Migratory Divide." *Molecular Ecology* 25 (11): 2529–41.

Rothberg, J. M., W. Hinz, K. L. Johnson, and J. Bustillo. 2016. Apparatus for measuring analytes using large scale FET arrays. *Patent EP2639579*, issued 2016.

Rothberg, Jonathan M., Wolfgang Hinz, Todd M. Rearick, Jonathan Schultz, William Mileski, Mel Davey, John H. Leamon, et al. 2011. "An Integrated Semiconductor Device Enabling Non-Optical Genome Sequencing." *Nature* 475 (7356): 348–52.

Rubolini, D., F. Spina, and N. Saino. 2004. "Protandry and Sexual Dimorphism in Trans-Saharan Migratory Birds." *Behavioral Ecology: Official Journal of the International Society for Behavioral Ecology* 15 (4): 592–601.

Ruparel, Hameer, Lanrong Bi, Zengmin Li, Xiaopeng Bai, Dae Hyun Kim, Nicholas J. Turro, and Jingyue Ju. 2005. "Design and Synthesis of a 3′-O-Allyl Photocleavable Fluorescent Nucleotide as a Reversible Terminator for DNA Sequencing by Synthesis." *Proceedings of the National Academy of Sciences of the United States of America* 102 (17): 5932–37.

Sabeti, Pardis C., Patrick Varilly, Ben Fry, Jason Lohmueller, Elizabeth Hostetter, Chris Cotsapas, Xiaohui Xie, et al. 2007. "Genome-Wide Detection and Characterization of Positive Selection in Human Populations." *Nature* 449 (7164): 913–18.

Safran, R. J., E. S. C. Scordato, M. R. Wilkins, J. K. Hubbard, B. R. Jenkins, T. Albrecht, S. M. Flaxman, et al. 2016. "Genome-Wide Differentiation in Closely Related Populations: The Roles of Selection and Geographic Isolation." *Molecular Ecology* 25 (16): 3865–83.

Saino, N., and R. Ambrosini. 2008. "Climatic Connectivity between Africa and Europe May Serve as a Basis for Phenotypic Adjustment of Migration Schedules of trans-Saharan Migratory Birds." *Global Change Biology*. https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1365-2486.2007.01488.x.

Saino, N., R. Ambrosini, B. Albetti, M. Caprioli, B. De Giorgio, E. Gatti, F. Liechti, et al. 2017. "Migration Phenology and Breeding Success Are Predicted by Methylation of a Photoperiodic Gene in the Barn Swallow." *Scientific Reports* 7 (March): 45412.

Saino, N., R. Ambrosini, and M. Caprioli. 2017. "Sex-dependent Carry-over Effects on Timing of Reproduction and Fecundity of a Migratory Bird." *Journal of Animal Ecology* 86 (2): 239–49.

Saino, N., R. Ambrosini, M. Caprioli, F. Liechti, A. Romano, D. Rubolini, and C. Scandolara. 2017. "Wing Morphology, Winter Ecology, and Fecundity Selection: Evidence for Sex-Dependence in Barn Swallows (Hirundo Rustica)." *Oecologia* 184 (4): 799–812.

Saino, N., R. Ambrosini, R. Martinelli, and S. Calza. 2002. "Offspring Sexual Dimorphism and Sex-allocation in Relation to Parental Age and Paternal Ornamentation in the Barn Swallow." *Molecular*. https://onlinelibrary.wiley.com/doi/abs/10.1046/j.1365-294X.2002.01542.x.

Saino, N., V. Bertacche, R. P. Ferrari, R. Martinelli, A. P. Møller, and R. Stradi. 2002. "Carotenoid Concentration in Barn Swallow Eggs Is Influenced by Laying Order, Maternal Infection and Paternal Ornamentation." *Proceedings. Biological Sciences / The Royal Society* 269 (1501): 1729–33.

Saino, N., A. M. Bolzern, and A. P. Møller. 1997. "Immunocompetence, Ornamentation, and Viability of

Male Barn Swallows (Hirundo Rustica)." *Proceedings of the National Academy of Sciences of the United States of America* 94 (2): 549–52.

Saino, N., Stefano Calza, Paola Ninni, and Anders Pape Moller. 1999. "Barn Swallows Trade Survival against Offspring Condition and Immunocompetence." *The Journal of Animal Ecology* 68 (5): 999–1009.

Saino, N., Luca Canova, Alessandra Costanzo, Diego Rubolini, Alexandre Roulin, and Anders Pape Møller. 2013. "Immune and Stress Responses Covary with Melanin-Based Coloration in the Barn Swallow." *Evolutionary Biology* 40 (4): 521–31.

Saino, N., Manuela Caprioli, M. Romano, Giuseppe Boncoraglio, Diego Rubolini, Roberto Ambrosini, Andrea Bonisoli-Alquati, and A. Romano. 2011. "Antioxidant Defenses Predict Long-Term Survival in a Passerine Bird." *PloS One* 6 (5): e19593.

Saino, N., José Javier Cuervo, Marco Krivacek, Florentino de Lope, and Anders Pape Møller. 1997. "Experimental Manipulation of Tail Ornament Size Affects the Hematocrit of Male Barn Swallows (Hirundo Rustica)." *Oecologia* 110 (2): 186–90.

Saino, N., R. P. Ferrari, M. Romano, D. Rubolini, and A. P. Møller. 2003. "Humoral Immune Response in Relation to Senescence, Sex and Sexual Ornamentation in the Barn Swallow (Hirundo Rustica)." *Journal of Evolutionary Biology* 16 (6): 1127–34.

Saino, N., Paolo Galeotti, Roberto Sacchi, and Anders Pape Møller. 1997. "Song and Immunological Condition in Male Barn Swallows (Hirundo Rustica)." *Behavioral Ecology: Official Journal of the International Society for Behavioral Ecology* 8 (4): 364–71.

Saino, N., Michele Incagli, Roberta Martinelli, and Anders Pape Møller. 2002. "Immune Response of Male Barn Swallows in Relation to Parental Effort, Corticosterone Plasma Levels, and Sexual Ornamentation." *Behavioral Ecology: Official Journal of the International Society for Behavioral Ecology* 13 (2): 169–74.

Saino, N., Roberta Martinelli, M. Romano, and Anders Pape Møller. 2003. "High Heritable Variation of a Male Secondary Sexual Character Revealed by Extra-pair Fertilization in the Barn Swallow." *Italian Journal of Zoology* 70 (2): 167–74.

Saino, N., and Anders Pape Møller. 1996. "Sexual Ornamentation and Immunocompetence in the Barn Swallow." *Behavioral Ecology: Official Journal of the International Society for Behavioral Ecology* 7 (2): 227–32.

Saino, N., C. R. Primmer, H. Ellegren, and A. P. Møller. 1997. "An Experimental Study of Paternity and Tail Ornamentation in the Barn Swallow (Hirundo Rustica)." *Evolution; International Journal of Organic Evolution* 51 (2): 562–70.

Saino, N., M. Romano, Roberto Ambrosini, Diego Rubolini, Giuseppe Boncoraglio, Manuela Caprioli, and A. Romano. 2012. "Longevity and Lifetime Reproductive Success of Barn Swallow Offspring Are Predicted by Their Hatching Date and Phenotypic Quality." *The Journal of Animal Ecology* 81 (5): 1004–12.

Saino, N., M. Romano, and M. Caprioli. 2012. "A Ptilochronological Study of Carry-over Effects of Conditions during Wintering on Breeding Performance in the Barn Swallow Hirundo Rustica." *Journal of Avian*. https://besjournals.onlinelibrary.wiley.com/doi/full/10.1111/j.1600-048X.2012.05622.x.

Saino, N., M. Romano, Manuela Caprioli, Roberto Lardelli, Pierfrancesco Micheloni, Chiara Scandolara, Diego Rubolini, and Mauro Fasola. 2013. "Molt, Feather Growth Rate and Body Condition of Male and Female Barn Swallows." *Journal of Ornithology / DO-G* 154 (2): 537–47.

Saino, N., M. Romano, M. Caprioli, M. Fasola, R. Lardelli, P. Micheloni, C. Scandolara, D. Rubolini, and L. Gianfranceschi. 2013. "Timing of Molt of Barn Swallows Is Delayed in a Rare Clock Genotype." *PeerJ* 1 (February): e17.

Saino, N., M. Romano, A. Romano, Diego Rubolini, Roberto Ambrosini, Manuela Caprioli, Marco Parolini,

Chiara Scandolara, Gaia Bazzi, and Alessandra Costanzo. 2015. "White Tail Spots in Breeding Barn Swallows Hirundo Rustica Signal Body Condition during Winter Moult." Edited by Javier Perez-Tris. *The Ibis* 157 (4): 722–30.

Saino, N., M. Romano, Diego Rubolini, Roberto Ambrosini, Manuela Caprioli, Aldo Milzani, Alessandra Costanzo, Graziano Colombo, Luca Canova, and Kazumasa Wakamatsu. 2013. "Viability Is Associated with Melanin-Based Coloration in the Barn Swallow (Hirundo Rustica)." *PloS One* 8 (4): e60426.

Saino, N., M. Romano, Diego Rubolini, Manuela Caprioli, Alessandra Costanzo, Luca Canova, and Anders Pape Møller. 2014. "Melanic Coloration Differentially Predicts Transfer of Immune Factors to Eggs with Daughters or Sons." *Behavioral Ecology: Official Journal of the International Society for Behavioral Ecology* 25 (5): 1248–55.

Saino, N., M. Romano, Diego Rubolini, Celine Teplitsky, Roberto Ambrosini, Manuela Caprioli, Luca Canova, and Kazumasa Wakamatsu. 2013. "Sexual Dimorphism in Melanin Pigmentation, Feather Coloration and Its Heritability in the Barn Swallow (Hirundo Rustica)." *PloS One* 8 (2): e58024.

Saino, N., D. Rubolini, R. Ambrosini, M. Romano, C. Scandolara, G. D. Fairhurst, M. Caprioli, A. Romano, B. Sicurella, and F. Liechti. 2015. "Light-Level Geolocators Reveal Covariation between Winter Plumage Molt and Phenology in a Trans-Saharan Migratory Bird." *Oecologia* 178 (4): 1105–12.

Saino, N., D. Rubolini, L. Serra, M. Caprioli, M. Morganti, R. Ambrosini, and F. Spina. 2010. "Sex-Related Variation in Migration Phenology in Relation to Sexual Dimorphism: A Test of Competing Hypotheses for the Evolution of Protandry." *Journal of Evolutionary Biology* 23 (10): 2054–65.

Saino, N., R. Stradi, P. Ninni, E. Pini, and A. P. Møller. 1999. "Carotenoid Plasma Concentration, Immune Profile, and Plumage Ornamentation of Male Barn Swallows (Hirundo Rustica)." *The American Naturalist* 154 (4): 441–48.

Saino, N., Tibor Szép, Roberto Ambrosini, M. Romano, and Anders Pape Møller. 2004. "Ecological Conditions during Winter Affect Sexual Selection and Breeding in a Migratory Bird." *Proceedings. Biological Sciences / The Royal Society* 271 (1540): 681–86.

Saino, N., T. Szép, M. Romano, and D. Rubolini. 2004. "Ecological Conditions during Winter Predict Arrival Date at the Breeding Quarters in a trans-Saharan Migratory Bird." *Ecology*. https://besjournals.onlinelibrary.wiley.com/doi/abs/10.1046/j.1461-0248.2003.00553.x.

Salas-Solano, O., E. Carrilho, L. Kotler, A. W. Miller, W. Goetzinger, Z. Sosic, and B. L. Karger. 1998. "Routine DNA Sequencing of 1000 Bases in Less than One Hour by Capillary Electrophoresis with Replaceable Linear Polyacrylamide Solutions." *Analytical Chemistry* 70 (19): 3996–4003.

Sanger, F. 1988. "Sequences, Sequences, and Sequences." *Annual Review of Biochemistry* 57: 1–28.

Sanger, F., G. M. Air, B. G. Barrell, N. L. Brown, A. R. Coulson, J. C. Fiddes, C. A. Hutchison III, P. M. Slocombe, and M. Smith. 1977. "Nucleotide Sequence of Bacteriophage φX174 DNA." *Nature* 265 (February): 687.

Sanger, F., and A. R. Coulson. 1975. "A Rapid Method for Determining Sequences in DNA by Primed Synthesis with DNA Polymerase." *Journal of Molecular Biology* 94 (3): 441–48.

Sanger, F., A. R. Coulson, G. F. Hong, D. F. Hill, and G. B. Petersen. 1982. "Nucleotide Sequence of Bacteriophage λ DNA." *Journal of Molecular Biology* 162 (4): 729–73.

Sanger, F., S. Nicklen, and A. R. Coulson. 1977. "DNA Sequencing with Chain-Terminating Inhibitors." *Proceedings of the National Academy of Sciences of the United States of America* 74 (12): 5463–67.

Sanger, F., and E. O. P. Thompson. 1953a. "The Amino-Acid Sequence in the Glycyl Chain of Insulin. II. The Investigation of Peptides from Enzymic Hydrolysates." *Biochemical Journal* 53 (3): 366–74.

———. 1953b. "The Amino-Acid Sequence in the Glycyl Chain of Insulin. I. The Identification of Lower Peptides from Partial Hydrolysates." *Biochemical Journal* 53 (3): 353–66.

Santure, Anna W., John G. Ewen, Delphine Sicard, Derek A. Roff, and Anders P. Møller. 2010. "Population Structure in the Barn Swallow, Hirundo Rustica: A Comparison between Neutral DNA Markers and

Quantitative Traits." *Biological Journal of the Linnean Society. Linnean Society of London* 99 (2): 306–14.

Scandolara, C., Roberto Lardelli, Giovanni Sgarbi, Manuela Caprioli, Roberto Ambrosini, Diego Rubolini, and Nicola Saino. 2014. "Context-, Phenotype-, and Kin-Dependent Natal Dispersal of Barn Swallows (Hirundo Rustica)." *Behavioral Ecology: Official Journal of the International Society for Behavioral Ecology* 25 (1): 180–90.

Scandolara, C., D. Rubolini, R. Ambrosini, M. Caprioli, S. Hahn, F. Liechti, A. Romano, M. Romano, B. Sicurella, and N. Saino. 2014. "Impact of Miniaturized Geolocators on Barn Swallow Hirundo Rustica Fitness Traits." *Journal of Avian Biology* 45 (5): 417–23.

Schielzeth, Holger, and Arild Husby. 2014. "Challenges and Prospects in Genome-Wide Quantitative Trait Loci Mapping of Standing Genetic Variation in Natural Populations." *Annals of the New York Academy of Sciences* 1320 (July): 35–57.

Schiffels, Stephan, and Richard Durbin. 2014. "Inferring Human Population Size and Separation History from Multiple Genome Sequences." *Nature Genetics* 46 (8): 919–25.

Schnable, Patrick S., Doreen Ware, Robert S. Fulton, Joshua C. Stein, Fusheng Wei, Shiran Pasternak, Chengzhi Liang, et al. 2009. "The B73 Maize Genome: Complexity, Diversity, and Dynamics." *Science* 326 (5956): 1112–15.

Schneider, Valerie A., Tina Graves-Lindsay, Kerstin Howe, Nathan Bouk, Hsiu-Chuan Chen, Paul A. Kitts, Terence D. Murphy, et al. 2017. "Evaluation of GRCh38 and de Novo Haploid Genome Assemblies Demonstrates the Enduring Quality of the Reference Assembly." *Genome Research* 27 (5): 849–64.

Scordato, Elizabeth S. C., and Rebecca J. Safran. 2014. "Geographic Variation in Sexual Selection and Implications for Speciation in the Barn Swallow." *Avian Research* 5 (1): 8.

Scordato, Elizabeth S. C., Matthew R. Wilkins, Georgy Semenov, Alexander S. Rubtsov, Nolan C. Kane, and Rebecca J. Safran. 2017. "Genomic Variation across Two Barn Swallow Hybrid Zones Reveals Traits Associated with Divergence in Sympatry and Allopatry." *Molecular Ecology* 26 (20): 5676–91.

Sela, Noa, Eddo Kim, and Gil Ast. 2010. "The Role of Transposable Elements in the Evolution of Non-Mammalian Vertebrates and Invertebrates." *Genome Biology* 11 (6): R59.

Seo, Jeong-Sun, Arang Rhie, Junsoo Kim, Sangjin Lee, Min-Hwan Sohn, Chang-Uk Kim, Alex Hastie, et al. 2016. "De Novo Assembly and Phasing of a Korean Human Genome." *Nature* 538 (7624): 243–47.

Seo, Tae Seok, Xiaopeng Bai, Dae Hyun Kim, Qinglin Meng, Shundi Shi, Hameer Ruparel, Zengmin Li, Nicholas J. Turro, and Jingyue Ju. 2005. "Four-Color DNA Sequencing by Synthesis on a Chip Using Photocleavable Fluorescent Nucleotides." *Proceedings of the National Academy of Sciences of the United States of America* 102 (17): 5926–31.

Session, Adam M., Yoshinobu Uno, Taejoon Kwon, Jarrod A. Chapman, Atsushi Toyoda, Shuji Takahashi, Akimasa Fukui, et al. 2016. "Genome Evolution in the Allotetraploid Frog Xenopus Laevis." *Nature* 538 (7625): 336–43.

Shendure, Jay, Shankar Balasubramanian, George M. Church, Walter Gilbert, Jane Rogers, Jeffery A. Schloss, and Robert H. Waterston. 2017. "DNA Sequencing at 40: Past, Present and Future." *Nature* 550 (7676): 345–53.

Shendure, Jay, Gregory J. Porreca, Nikos B. Reppas, Xiaoxia Lin, John P. McCutcheon, Abraham M. Rosenbaum, Michael D. Wang, Kun Zhang, Robi D. Mitra, and George M. Church. 2005. "Accurate Multiplex Polony Sequencing of an Evolved Bacterial Genome." *Science* 309 (5741): 1728–32.

Sicurella, Beatrice, Manuela Caprioli, A. Romano, M. Romano, Diego Rubolini, Nicola Saino, and Roberto Ambrosini. 2014. "Hayfields Enhance Colony Size of the Barn Swallow Hirundo Rustica in Northern Italy." *Bird Conservation International* 24 (1): 17–31.

Sicurella, Beatrice, Federica Musitelli, Diego Rubolini, Nicola Saino, and Roberto Ambrosini. 2016. "Environmental Conditions at Arrival to the Wintering Grounds and during Spring Migration Affect

Population Dynamics of Barn Swallows Hirundo Rustica Breeding in Northern Italy." *Population Ecology* 58 (1): 135–45.

Simão, Felipe A., Robert M. Waterhouse, Panagiotis Ioannidis, Evgenia V. Kriventseva, and Evgeny M. Zdobnov. 2015. "BUSCO: Assessing Genome Assembly and Annotation Completeness with Single-Copy Orthologs." *Bioinformatics* 31 (19): 3210–12.

Slate, Jon, Anna W. Santure, Philine G. D. Feulner, Emily A. Brown, Alex D. Ball, Susan E. Johnston, and Jake Gratten. 2010. "Genome Mapping in Intensively Studied Wild Vertebrate Populations." *Trends in Genetics: TIG* 26 (6): 275–84.

Slatko, Barton E., Peter Heinrich, B. Tracy Nixon, and Richard L. Eckert. 1993. "Preparation of Templates for DNA Sequencing." *Current Protocols in Molecular Biology / Edited by Frederick M. Ausubel ... [et Al.]* 21 (1): 7.3.1–7.3.10.

Smit, A. F. 1993. "Identification of a New, Abundant Superfamily of Mammalian LTR-Transposons." *Nucleic Acids Research* 21 (8): 1863–72.

Smit, A. F., R. Hubley, and P. Green. 1996–2010. "RepeatMasker Open-3.0." 1996–2010. http://www.repeatmasker.org.

Smith, J. M., and J. Haigh. 1974. "The Hitch-Hiking Effect of a Favourable Gene." *Genetical Research* 23 (1): 23–35.

Smith, L. M., Steven Fung, Michael W. Hunkapiller, Tim J. Hunkapiller, and Leroy E. Hood. 1985. "The Synthesis of Oligonucleotides Containing an Aliphatic Amino Group at the 5′ Terminus: Synthesis of Fluorescent DNA Primers for Use in DNA Sequence Analysis." *Nucleic Acids Research* 13 (7): 2399–2412.

Smith, L. M., J. Z. Sanders, R. J. Kaiser, P. Hughes, C. Dodd, C. R. Connell, C. Heiner, S. B. Kent, and L. E. Hood. 1986. "Fluorescence Detection in Automated DNA Sequence Analysis." *Nature* 321 (6071): 674–79.

Smith, T. F., and M. S. Waterman. 1981. "Identification of Common Molecular Subsequences." *Journal of Molecular Biology* 147 (1): 195–97.

Song, Liting, Wenxun Huang, Juan Kang, Yuan Huang, Hong Ren, and Keyue Ding. 2017. "Comparison of Error Correction Algorithms for Ion Torrent PGM Data: Application to Hepatitis B Virus." *Scientific Reports* 7 (1): 8106.

Staden, R. 1979. "A Strategy of DNA Sequencing Employing Computer Programs." *Nucleic Acids Research* 6 (7): 2601–10.

Stanke, Mario, Rasmus Steinkamp, Stephan Waack, and Burkhard Morgenstern. 2004. "AUGUSTUS: A Web Server for Gene Finding in Eukaryotes." *Nucleic Acids Research* 32 (Web Server issue): W309–12.

Stein, Lincoln D. 2010. "The Case for Cloud Computing in Genome Informatics." *Genome Biology* 11 (5): 207.

Sturtevant, Alfred H. 1913. "The Linear Arrangement of Six Sex-Linked Factors in Drosophila, as Shown by Their Mode of Association." *The Journal of Experimental Zoology* 14 (1): 43–59.

Sudmant, Peter H., Tobias Rausch, Eugene J. Gardner, Robert E. Handsaker, Alexej Abyzov, John Huddleston, Yan Zhang, et al. 2015. "An Integrated Map of Structural Variation in 2,504 Human Genomes." *Nature* 526 (7571): 75–81.

Sutton, Granger G., Owen White, Mark D. Adams, and Anthony R. Kerlavage. 1995. "TIGR Assembler: A New Tool for Assembling Large Shotgun Sequencing Projects." *Genome Science and Technology* 1 (1): 9–19.

Suzuki, Yuta, Jonas Korlach, Stephen W. Turner, Tatsuya Tsukahara, Junko Taniguchi, Wei Qu, Kazuki Ichikawa, et al. 2016. "AgIn: Measuring the Landscape of CpG Methylation of Individual Repetitive Elements." *Bioinformatics* 32 (19): 2911–19.

Tabor, S., and C. C. Richardson. 1987. "DNA Sequence Analysis with a Modified Bacteriophage T7 DNA Polymerase." *Proceedings of the National Academy of Sciences of the United States of America* 84 (14): 4767–71.

Tattini, Lorenzo, Romina D'Aurizio, and Alberto Magi. 2015. "Detection of Genomic Structural Variants from Next-Generation Sequencing Data." *Frontiers in Bioengineering and Biotechnology* 3 (June): 92.

Tawfik, D. S., and A. D. Griffiths. 1998. "Man-Made Cell-like Compartments for Molecular Evolution." *Nature Biotechnology* 16 (7): 652–56.

Teeling, Emma C., Sonja C. Vernes, Liliana M. Dávalos, David A. Ray, M. Thomas P. Gilbert, Eugene Myers, and Bat1K Consortium. 2018. "Bat Biology, Genomes, and the Bat1K Project: To Generate Chromosome-Level Genomes for All Living Bat Species." *Annual Review of Animal Biosciences* 6 (February): 23–46.

Thomas, James W., Mario Cáceres, Joshua J. Lowman, Caroline B. Morehouse, Meghan E. Short, Erin L. Baldwin, Donna L. Maney, and Christa L. Martin. 2008. "The Chromosomal Polymorphism Linked to Variation in Social Behavior in the White-Throated Sparrow (Zonotrichia Albicollis) Is a Complex Rearrangement and Suppressor of Recombination." *Genetics* 179 (3): 1455–68.

Thorisson, Gudmundur A., Albert V. Smith, Lalitha Krishnan, and Lincoln D. Stein. 2005. "The International HapMap Project Web Site." *Genome Research* 15 (11): 1592–93.

Tishkoff, Sarah A., Floyd A. Reed, Alessia Ranciaro, Benjamin F. Voight, Courtney C. Babbitt, Jesse S. Silverman, Kweli Powell, et al. 2007. "Convergent Adaptation of Human Lactase Persistence in Africa and Europe." *Nature Genetics* 39 (1): 31–40.

Tomb, J. F., O. White, A. R. Kerlavage, R. A. Clayton, G. G. Sutton, R. D. Fleischmann, K. A. Ketchum, et al. 1997. "The Complete Genome Sequence of the Gastric Pathogen Helicobacter Pylori." *Nature* 388 (6642): 539–47.

Toumazou, C., and S. Purushothaman. 2004. Sensing apparatus and method. *Patent US7686929*, issued 2004.

Tsyusko, Olga V., Maureen B. Peters, Cris Hagen, Tracey D. Tuberville, Timothy A. Mousseau, Anders P. Møller, and Travis C. Glenn. 2007. "Microsatellite Markers Isolated from Barn Swallows (Hirundo Rustica)." *Molecular Ecology Notes* 7 (5): 833–35.

Turner, A. 2004. "Family Hirundinidae (Swallows and Martins)." In *Handbook of the Birds of the World*, edited by J. Hoyo, A. Elliott, and D. A. Christie, Vol. 9, pp. 602–85. Lynx Edicions, Barcelona.

———. 2006. *The Barn Swallow*. T & AD Poyser, London.

Venter, J. C., M. D. Adams, E. W. Myers, P. W. Li, R. J. Mural, G. G. Sutton, H. O. Smith, et al. 2001. "The Sequence of the Human Genome." *Science* 291 (5507): 1304–51.

Visscher, Peter M., Naomi R. Wray, Qian Zhang, Pamela Sklar, Mark I. McCarthy, Matthew A. Brown, and Jian Yang. 2017. "10 Years of GWAS Discovery: Biology, Function, and Translation." *American Journal of Human Genetics* 101 (1): 5–22.

Vitaterna, Martha Hotz, Caroline H. Ko, Anne-Marie Chang, Ethan D. Buhr, Ethan M. Fruechte, Andrew Schook, Marina P. Antoch, Fred W. Turek, and Joseph S. Takahashi. 2006. "The Mouse Clock Mutation Reduces Circadian Pacemaker Amplitude and Enhances Efficacy of Resetting Stimuli and Phase-Response Curve Amplitude." *Proceedings of the National Academy of Sciences of the United States of America* 103 (24): 9327–32.

Vortman, Yoni, Arnon Lotem, Roi Dor, Irby J. Lovette, and Rebecca J. Safran. 2011. "The Sexual Signals of the East-Mediterranean Barn Swallow: A Different Swallow Tale." *Behavioral Ecology: Official Journal of the International Society for Behavioral Ecology* 22 (6): 1344–52.

Wang, D. G., J. B. Fan, C. J. Siao, A. Berno, P. Young, R. Sapolsky, G. Ghandour, et al. 1998. "Large-Scale Identification, Mapping, and Genotyping of Single-Nucleotide Polymorphisms in the Human Genome." *Science* 280 (5366): 1077–82.

Wang, John, Yannick Wurm, Mingkwan Nipitwattanaphon, Oksana Riba-Grognuz, Yu-Ching Huang, Dewayne Shoemaker, and Laurent Keller. 2013. "A Y-like Social Chromosome Causes Alternative Colony Organization in Fire Ants." *Nature* 493 (7434): 664–68.

Ward, Lucas D., and Manolis Kellis. 2012. "Evidence of Abundant Purifying Selection in Humans for Recently Acquired Regulatory Functions." *Science* 337 (6102): 1675–78.

Watson, J. D., and F. H. C. Crick. 1953. "Molecular Structure of Nucleic Acids." *Nature.*

Weber, J. L., and E. W. Myers. 1997. "Human Whole-Genome Shotgun Sequencing." *Genome Research* 7 (5): 401–9.

Wheeler, David A., Maithreyan Srinivasan, Michael Egholm, Yufeng Shen, Lei Chen, Amy McGuire, Wen He, et al. 2008. "The Complete Genome of an Individual by Massively Parallel DNA Sequencing." *Nature* 452 (7189): 872–76.

Wilhelm, Brian T., Samuel Marguerat, Stephen Watt, Falk Schubert, Valerie Wood, Ian Goodhead, Christopher J. Penkett, Jane Rogers, and Jürg Bähler. 2008. "Dynamic Repertoire of a Eukaryotic Transcriptome Surveyed at Single-Nucleotide Resolution." *Nature* 453 (7199): 1239–43.

Wilkins, M. R., H. Karaardıç, Y. Vortman, T. L. Parchman, T. Albrecht, A. Petrželková, L. Özkan, et al. 2016. "Phenotypic Differentiation Is Associated with Divergent Sexual Selection among Closely Related Barn Swallow Populations." *Journal of Evolutionary Biology* 29 (12): 2410–21.

Wu, R. 1994. "Development of the Primer-Extension Approach: A Key Role in DNA Sequencing." *Trends in Biochemical Sciences* 19 (10): 429–33.

Wu, R., and A. D. Kaiser. 1968. "Structure and Base Sequence in the Cohesive Ends of Bacteriophage Lambda DNA." *Journal of Molecular Biology* 35 (3): 523–37.

Ye, Kai, Marcel H. Schulz, Quan Long, Rolf Apweiler, and Zemin Ning. 2009. "Pindel: A Pattern Growth Approach to Detect Break Points of Large Deletions and Medium Sized Insertions from Paired-End Short Reads." *Bioinformatics* 25 (21): 2865–71.

Zerbino, Daniel R., and Ewan Birney. 2008. "Velvet: Algorithms for de Novo Short Read Assembly Using de Bruijn Graphs." *Genome Research* 18 (5): 821–29.

Zhang, Guojie, Cai Li, Qiye Li, Bo Li, Denis M. Larkin, Chul Lee, Jay F. Storz, et al. 2014. "Comparative Genomics Reveals Insights into Avian Genome Evolution and Adaptation." *Science* 346 (6215): 1311–20.

Zhang, Guojie, Carsten Rahbek, Gary R. Graves, Fumin Lei, Erich D. Jarvis, and M. Thomas P. Gilbert. 2015. "Genomics: Bird Sequencing Project Takes off." *Nature* 522 (7554): 34.

Zhang, J., Y. Fang, J. Y. Hou, H. J. Ren, R. Jiang, P. Roos, and N. J. Dovichi. 1995. "Use of Non-Cross-Linked Polyacrylamide for Four-Color DNA Sequencing by Capillary Electrophoresis Separation of Fragments up to 640 Bases in Length in Two Hours." *Analytical Chemistry* 67 (24): 4589–93.

Zheng-Bradley, Xiangqun, and Paul Flicek. 2017. "Applications of the 1000 Genomes Project Resources." *Briefings in Functional Genomics* 16 (3): 163–70.

Zimmerman, Eilene. 2014. "Illumina." *MIT Technology Review*, February 18, 2014. https://www.technologyreview.com/s/524531/why-illumina-is-no-1/.

SECTION B

Evolutionary study of Huntington's Disease-causing CAG repeats along vertebrate phylogenetic history

This section presents the methodological work and the conclusions drawn from my - and other collaborators - work on the study of the evolutionary origins of Huntington's Disease, a genetic neurodegenerative disorder. The study was conducted in the Laboratory of Stem Cell Biology and Pharmacology of Neurodegenerative Diseases directed by Prof. Elena Cattaneo at the University of Milan where I worked for the first two years of my PhD (and also during my Master Thesis work) and whose research effort is on the phylogenetic and biological investigation of HD causative gene. The study was made possible also thanks to a collaboration between Prof. Cattaneo and my Ph.D. thesis supervisor Prof. Nicola Saino. A manuscript is in preparation reporting part of the data from this work together with other data obtained in the Cattaneo lab.

INTRODUCTION

1. Huntington's Disease

1.1. The disease.

Huntington's Disease (HD) is a neurodegenerative disorder with a fully autosomal dominant inheritance pattern (MacDonald et al. 1993; Lee et al. 2012). Prevalence (i.e. the actual number of cases) is estimated to be one in about 7,300 individuals in western populations (Almqvist et al. 2001; Morrison, Harding-Lester, and Bradley 2011; Fisher and Hayden 2014). In 1872, the physician George Huntington first described HD as chorea on the basis of its symptoms characterized by involuntary body movements (Wiener and Lang 1989). In 1993, after extensive research (Gusella et al. 1983), the gene responsible for the disease (IT15 or Htt) was identified on chromosome 4 short arm (region 4p16.3) and the protein it encodes was named Huntingtin (HTT) (MacDonald et al. 1993). HTT is a highly conserved, ubiquitously expressed protein (Strong et al. 1993). The mutation associated with HD is found in 5' terminal of the HD gene within exon 1 and takes the form of an over-threshold CAG repeats run, which encodes for a polyGlutamine (polyQ) stretch (MacDonald et al. 1993). In humans this locus is highly polymorphic and is known to vary between 9 and 35 repeats in the healthy range (Kremer et al. 1994; Rubinsztein et al. 1994). Several studies have indicated that the CAG tract distribution in the healthy range appears positively skewed, with a modal length of 18 repeats in European populations showing higher HD incidence, and 17 and 16 in East Asia and Africa respectively (Bates, Harper, and Jones 2002; Fisher and Hayden 2014). In the British Columbia population, the most frequent CAG repeats number is 17 (Semaka et al. 2013) (see **Figure 1**).

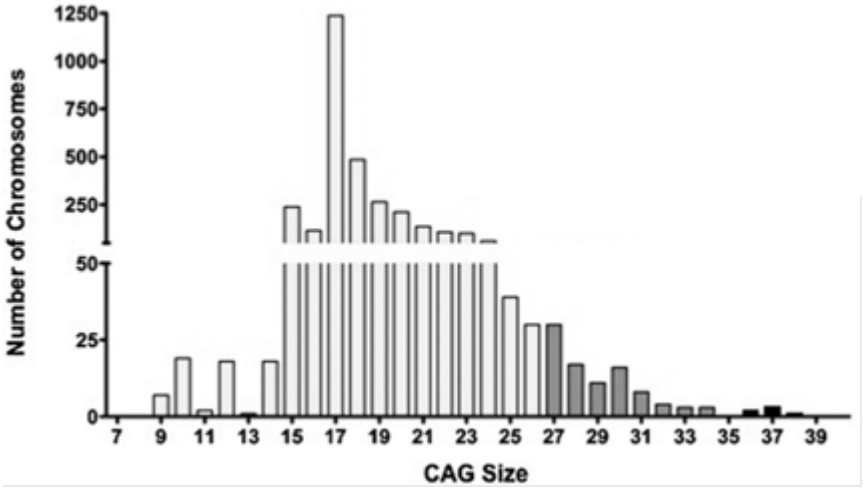

Figure 1: CAG size distribution of 3,188 alleles ascertained in a sample of individuals from British Columbia unrelated to HD. *Grey bars represent intermediate (27–35 CAG) alleles, while reduced penetrance alleles (36–39 CAG) are dark-shaded. Adapted from Semaka et al. 2013.*

The mean age at onset (i.e. when choreic movements start to develop) is 45 years and it is correlated with repeats number, with the length of the repeats accounting for approximately 60% of the variation observed in the age at motor onset (Gusella, MacDonald, and Lee 2014). A fraction of the remaining variance can

potentially be attributed to genetic variations within the genome of HD patients (Gusella, MacDonald, and Lee 2014). A number of genetic modifiers have been proposed over the years (Weydt et al. 2009), and some have been identified in Genome-Wide Association Studies (Genetic Modifiers of Huntington's Disease (GeM-HD) Consortium 2015). After disease onset, symptoms progress and deteriorate until death that usually occurs within 15-20 years, often due to infections (Albin and Tagle 1995).

Borderline individuals with a number of repeats between 36 and 40 are characterized by incomplete penetrance, that is they do not necessarily develop HD. However, these individuals can transmit disease alleles to childhood through a mechanism defined genetic anticipation, where germline mutations favor the transmission of longer alleles to the offspring (Pearson 2003). This phenomenon occurs especially during paternal transmission and can affect also HD alleles.

HD belongs to the 'CAG repeats neurological disorders group' (McMurray 2010), which also includes Spinocerebellar Ataxias (SCA, types 1, 2, 3, 6, 7, 17), Spinobulbar Muscular Atrophy (SBMA), and Dentatorubral-pallidoluysian atrophy (DRPLA). These neurodegenerative disorders affect specific areas of the brain and specific neuronal subpopulations. This specificity may suggest that the expanded CAG repeat itself is not sufficient to cause selective neuronal death and the cause of the disease is linked to the loss of function of the involved protein (Cattaneo et al. 2001; Cattaneo, Zuccato, and Tartari 2005).

1.2. Htt gene structure and expression.

The Htt gene is a 67 exon-long gene located on the short arm (p) of human chromosome 4. It encodes for a 3,144 amino acid-long protein named HTT (**Figure 2**). In rodents as well as in humans, there are two alternative splicing mRNA products of transcription. The main transcript, about 10 kbp in size, is ubiquitously expressed but predominantly found in the neurons of the central nervous system (DiFiglia et al. 1995; Ferrante et al. 2000), but also in testes (Strong et al. 1993). In the brain, expression occurs specifically in cortical neurons of layers III and IV. More recent findings have revealed the presence of a natural Htt antisense mRNA limited to the Exon 1 capable of regulating gene expression (Chung et al. 2011). In mammalian cells, HTT is found mainly in cytoplasm, associated to the nucleus, endoplasmic reticulum and Golgi apparatus (Zuccato, Valenza, and Cattaneo 2010). HTT has also been found associated with microtubules, which may suggest an implication in cellular transport of vesicles and organelles (Gutekunst et al. 1995; Gauthier et al. 2004).

Given its high molecular weight (380 kDa), for years little was known about HTT structure as it was hard to identify its tridimensional structure by X-ray crystallography. In 2018, cryo-electron microscopy was able to reveal the structure of full-length human HTT in a complex with HTT-associated protein 40 (HAP40) to a very high resolution (Guo et al. 2018). This study has shown that HTT is largely α-helical, with three major domains: the amino- and carboxy-terminal domains and smaller bridge domain containing different types of tandem repeats between them. The amino- and carboxy-terminal domains harbour multiple HEAT (huntingtin, elongation factor 3, protein phosphatase 2A and lipid kinase TOR) repeats (Takano and Gusella 2002; W. Li et al. 2006) arranged in a solenoid fashion (Guo et al. 2018). These 40 residue-long HEAT sequences can fold in helix-turn-helix motifs or α-RODs (Andrade, Petosa, and O'Donoghue 2001) and are

involved in protein-protein interactions. Four HEAT repeats clusters have been identified in HTT (Palidwor et al. 2009). It is also commonly accepted that the polyP tract found in mammals after the polyQ tract in the amino-terminal domain has evolved as a polyQ stabilizer, which acts by raising its solubility (Steffan et al. 2004).

The presence of NES sequences (*Nuclear Export Signal*) and NLS (*Nuclear Localization Signal*) argues in favour of a protein role in molecules exchange between nucleus and cytoplasm (Xia et al. 2003). This hypothesis is also supported by the perinuclear localization of HTT and by demonstration that residue 17°, immediately preceding the polyQ tract, interacts with the nuclear pore protein TRP (*Translocated Promoter Region*) which enables HTT to migrate out of the nucleus. Removal of this residue causes HTT accumulation in the nucleus (Cornett et al. 2005).

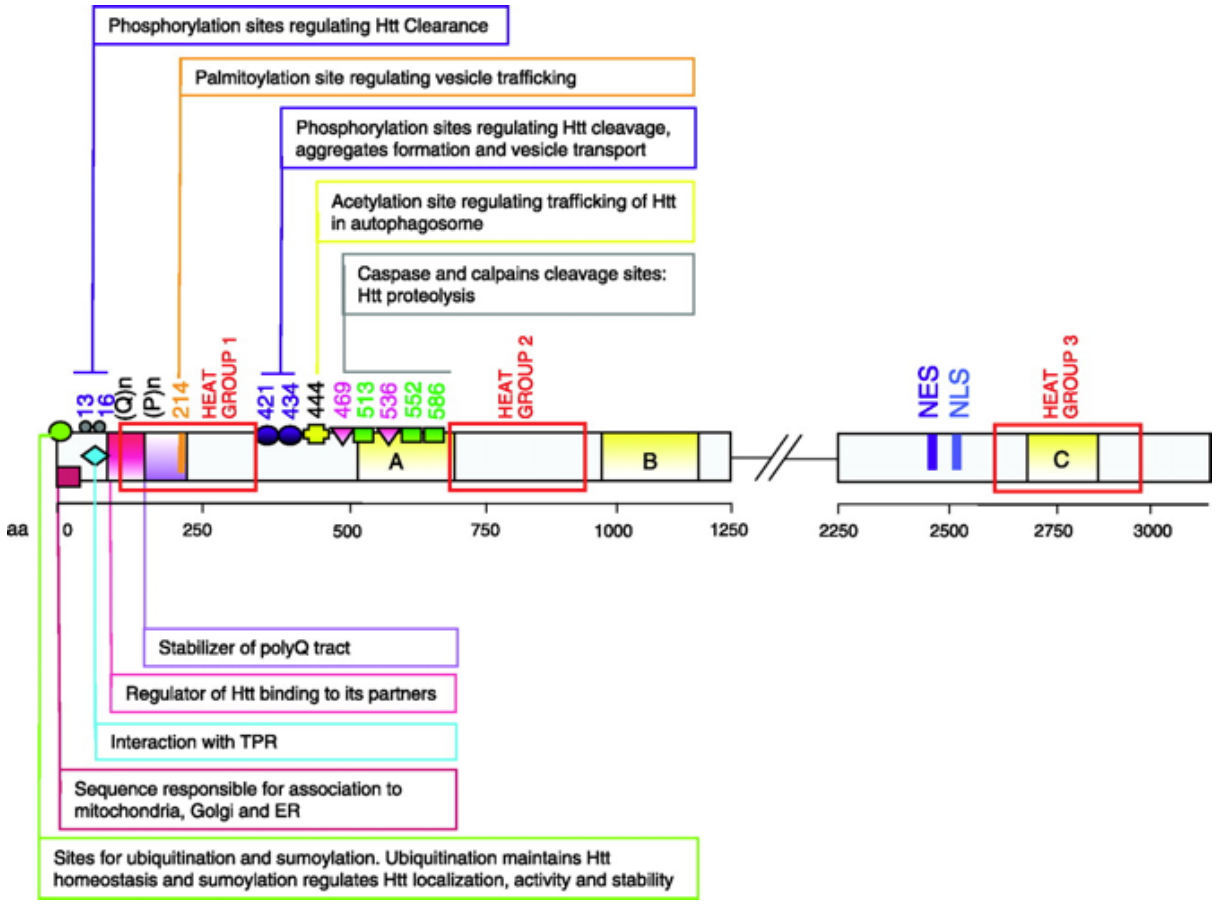

Figure 2: HTT protein schematic structure. *Adapted from Zuccato et al. 2010.*

HTT undergoes several post-translational modifications, including ubiquitination (lysine 6, 9 and 15), sumoylation (N-17) and acetylation of lysine 444 (Jeong et al. 2009) that are believed to affect protein cellular localization and homeostasis (Kalchman et al. 1996; Steffan et al. 2004). Other modifications include phosphorylation of specific serine residues (13, 16, 421, 434, 1181, 1201), and palmitoylation of cysteine 214, which appears to be connected with vesicular trafficking (Yanai et al. 2006).

1.3. HTT gain and loss of function.

Two models, positing a *gain* and a *loss* of function respectively, have been competing over the years to account for the relationship between HTT with an expanded polyQ and disease development and progression. The first model (gain of function model) predicts that pathology is engendered by a non-physiological function of the expanded glutamine residues stretch that causes HTT to acquire new functions that are toxic for neurons. In general, several dysfunctions have been identified in association with an acquired pathogenicity of the mutated protein. However, the gain of function model fails to explain selective striatal neuronal loss (Cattaneo et al. 2001; Cattaneo, Zuccato, and Tartari 2005) suggesting that also a loss of normal HTT function could be related to the disease. The loss of function model posits that HTT and its polyQ repeat play a normal physiological role that is altered by the over-threshold expansion (Cattaneo et al. 2001; Cattaneo, Zuccato, and Tartari 2005). Supporting this view, over the years many physiological activities of the protein during development and in the adulthood have been identified that might be compromised in the disease.

2. Huntington's Disease gene evolution

HD-causing gene is extensively found in living organisms and its CAG repeat region is a feature of all vertebrate lineages so far investigated. The present chapter highlights the mechanisms through which this repeat might have evolved in the first place, how it keeps evolving, the most prominent mechanisms that might affect its changes and what was known of its evolutionary history prior to this work.

2.1. Repeats within the genome

In relatively recent years it has become apparent that the genomes of extant organisms are, to a great extent, interspersed by short repetitive elements. The human genome contains simple sequence repeats (SSRs, also called microsatellites) of 1-6 nucleotide repeats in the order of 3% (Tóth, Gáspári, and Jurka 2000). In the long history of life, it is very likely that specific repeats have not always been present, or possibly, they must have appeared and disappeared during genetic evolution (Ellegren 2004). When a region of repeats in an extant organism is observed, it is reasonable to conceive that this region has evolved along generations potentially spanning millions of years from an ancestral condition characterized by the total absence of the region, or by a low number of repeats engendered by random mutation. It has long been recognized that on a genome scale transitions and transversions[45] alone do not account for the overall presence of repeats (Durrett and Kruglyak 1999) and replication slippage has been identified as the most common cause of SSRs variation in the genomes (McMurray 2010).

2.1.1. Replication slippage and repeat mutation

Although many other factors certainly play a role, replication slippage is believed to be caused mainly by mismatches between DNA strands while being replicated during meiosis (see **Figure 3** caption for a brief description of this mechanism) (Wheeler et al. 1999), mitosis and repair (Tautz and Schlötterer 1994; Pearson, Nichol Edamura, and Cleary 2005). Repeats expansions by slippage mutation rely on the formation of DNA secondary structures called hairpins, and it has been argued that rate of formation, abundance and length of simple repeats in genomes can be explained by the differential capacity of their specific sequence to form hairpins (Bacolla et al. 2008). Mean error rate in humans has been quantified in about one single variation per 1,000 generations (Weber and Wong 1993) and the magnitude of slippage mutation is much higher than that of point mutations (Jarne and Lagoda 1996). Most slippage events result in a change of just one repeat unit and slippage rates vary for different repeat unit sizes (Kruglyak et al. 1998). Repetitive elements can be present in coding as well as in non-coding regions. When considering protein coding regions, single nucleotide insertions or deletions alone cannot explain the presence of repeats as they would turn into frameshift mutations (i.e. changes in the codon reading frame), a deleterious condition in most cases (Watson 2014). However, trinucleotide variation in coding regions does not cause frameshift (Metzgar and Wills 2000), implying that triplet insertions are generally more tolerated than single, double or any non-

[45] Transitions are interchanges of two-ring purines (A <> G) or of one-ring pyrimidines (C <> T). Transversions are interchanges of purine for pyrimidine bases. Transitions are generally favoured over transversions, perhaps since they involve less structure modifications (Futuyma 2013).

multiple of three insertions and deletions. These repeats are referred as homorepeats or polyX regions (where is X stands for the amino acid they encode for) (Pablo Mier and Andrade-Navarro 2017), and have been described as more abundant in eukaryotic than prokaryotic genomes (Faux et al. 2005; Jorda and Kajava 2010).

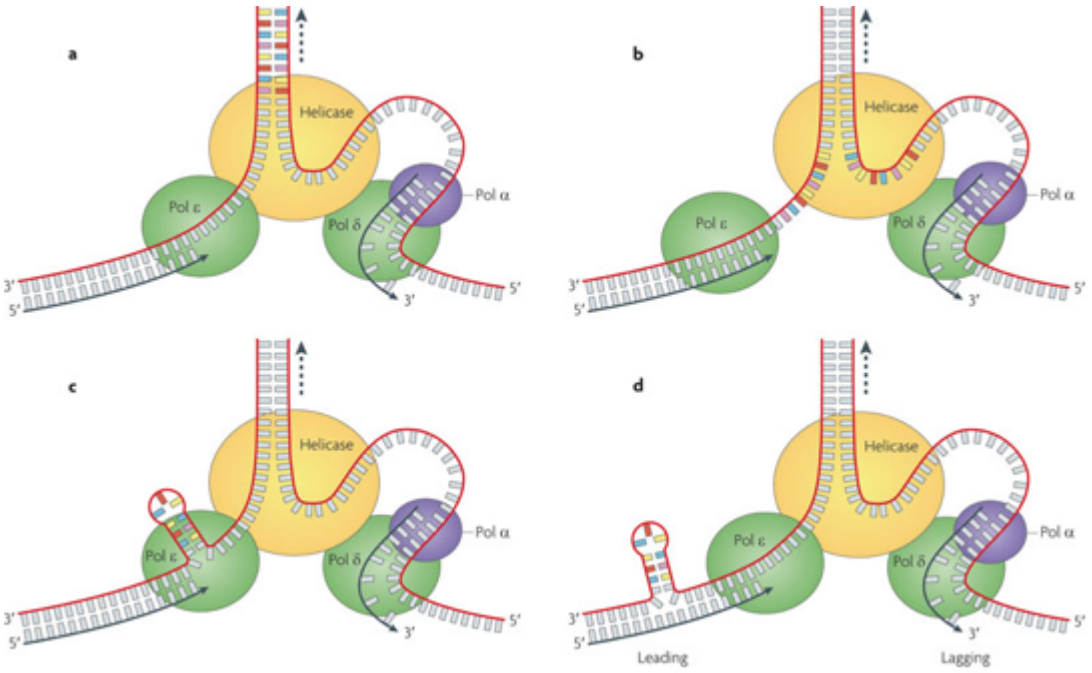

Figure 3: Replication slippage. *Nucleotides representing repeats are coloured. In (b) the polymerase encounters the repeat tract and progression is slowed but the helicase speed is unaffected (space between green and yellow ovals). To avoid uncoupling of the polymerase and the helicase, the leading strand polymerase bypasses a segment of unreplicated repeat template on the leading strand, which has formed a hairpin (c). Deletions occur if the loops form on the template strand (d) and insertions occur when loop formation is on the daughter strand (not shown). Adapted from McMurray et al. 2010.*

2.2.2. *Cis* and *trans* acting modifiers

Repeats might have constant or variable mutation rates. In the case the mutation rate is variable, this variability can be due to external conditions (environmental stressors that affect the probability of mutation), genetic factors (e. g. length dependency, *cis* or *trans*-acting factors, see below) or a combination of both. Length dependency of mutation rates can be regarded as a predictor of propensity to mutation (Xu, Peng, and Fang 2000). However, lengths being equal, certain repeats appear to be more prone to mutation than others (Brock, Anderson, and Monckton 1999), implying that expandability is influenced by factors other than length. In HD, single sperm analysis of individuals carrying a constitutive allele of the same length showed that sex, mode of transmission and genomic context were found to be relevant predictors of expandability (Wheeler et al. 2007). The latter element includes *cis* and *trans* acting modifiers (Richards and Sutherland 1994). *Cis* acting modifiers are neighbouring genomic sequences (Martins et al. 2008), while *trans* acting

factors represent other genomic regions (coding and non-coding). These factors assume a special importance in the case that specific modifiers could be identified for either screening or treatment purposes (Coles, Leggo, and Rubinsztein 1997). To understand whether long pathological CAG alleles were caused by a general tendency of specific genotypes to give rise to new mutations, several studies have been carried out over the years. For example, it was noted that CAG/CTG repeats are generally found in CpG islands[46] (Brock, Anderson, and Monckton 1999). The null-hypothesis for this occurrence is that as CpG have higher GC content they are more easily turned into CAG/CTG repeats than other genomic regions. Alternatively, since CpG islands are generally associated with transcriptional regulation (Deaton and Bird 2011), their co-occurrence with CAG/CTG repeats might have a functional explanation.

Several authors have also suggested that specific haplogroups (defined on the basis of SNPs) are associated with the expandability propriety of repeats (Warby, Montpetit, et al. 2009; Warby et al. 2011), although debate exists on the subject (Falush 2009; Warby, Visscher, et al. 2009).

2.3.3. Evolutionary explanations for repeat expansion

Unless repeat expansions represent an almost neutral or even negative side effect of DNA replication that can by no means be totally overcome, the sole fact that repeats are present (also in coding regions), and that the mechanism producing them has been conserved throughout the 3.5-billion-years history of life, calls for an explanation. Despite this, for many years repeats were regarded as 'junk' DNA (Orgel and Crick 1980) and therefore almost completely ignored (Shapiro 2011). However, many new data are challenging this view and point to a positive function of repeats in the genomes (Kashi and King 2006), which makes them suitable for positive natural selection (Haasl and Payseur 2013). In particular, given their length variation, several authors have proposed that they might work as 'tuning knobs' of gene function (Nithianantharajah and Hannan 2007), and thereby potentially involved in shaping the phenotype (Johnsen et al. 2007). If repeats truly work as 'tuning knobs', then their units of mutation are not single nucleotides but rather entire triplets (Richards and Sutherland 1994; Richards 2001).

In evolutionary studies, HD is often regarded as a model for its peculiar tendency to remain present in the population in spite of its potential detrimental effects in the adulthood. This finding has been either explained by the disease onset after reproductive age or by a direct increase in fitness of longer/expanded repeats (Eskenazi, Wilson-Rich, and Starks 2007). In the first scenario, healthy alleles are thought as neutral, having no consequences for the phenotype (Kashi and King 2006), while pathological alleles have serious phenotypic consequences (i.e. the disease in the adulthood), but may escape natural selection due to their late age-at-onset (Haldane 1941; Partridge and Gems 2002). According to this scenario, a simple mechanistic bias toward repeat expansion was invoked as early as 1994, especially for expansions above the pathological threshold (Rubinsztein et al. 1994). Indeed, if replication machinery errors favour expansions over contractions, the bias would be readily explained.

[46] CpG islands are genomic regions of at least 200 bp where CG content exceeds 50-60% (Gardiner-Garden and Frommer 1987).

In the second scenario, fitness and natural selection have a prominent role in favouring the spread of longer alleles. Therefore, the bias toward longer alleles would be produced by positive selection rather than by drift or a molecular mechanism intrinsic to DNA replication. Indeed, an extensive survey of human genome has revealed that trinucleotide (TNRs) repeats distribution in coding region is not random (Kozlowski, de Mezer, and Krzyzosiak 2010). The general mechanism of slippage mutations alone cannot explain the relative abundance of specific trinucleotide repeats. Exons are greatly enriched with specific TNRs and there is a preferential codon usage for these TNRs. Intriguingly for the case of Htt gene, CNG[47] repeats are the most represented TNRs in exons (**Figure 4A**), with a 10.4-fold CAG overrepresentation. Glutamine is by far the preferred translation (61%) (**Figure 4B**), and although this last finding could theoretically be explained in terms of a lower degree of toxicity (i. e. Glutamine stretches are more tolerated than others), the overrepresentation in exons of CAG tracts clearly points toward a biological role in protein function.

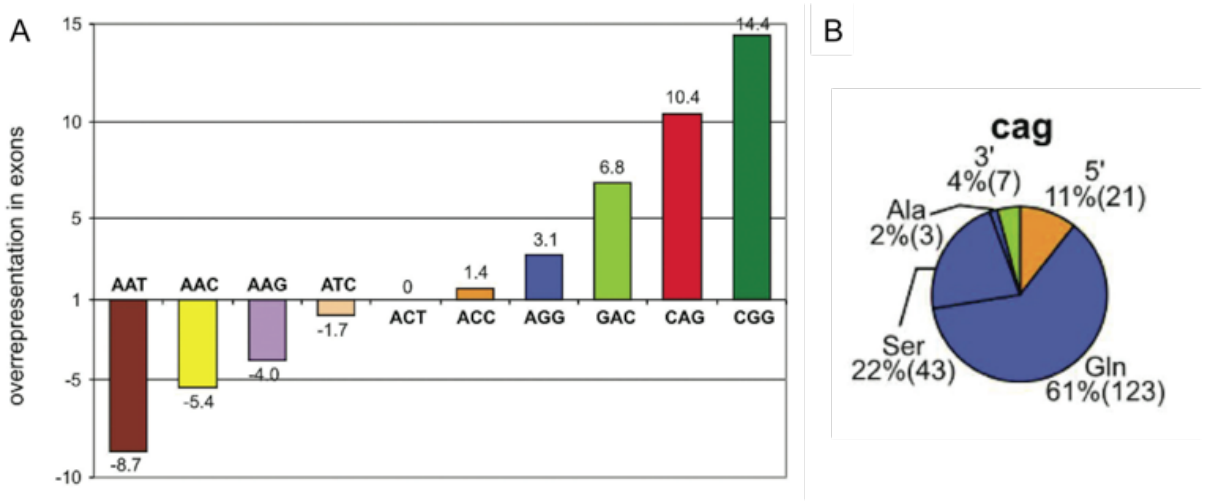

Figure 4: TNR in the human genome. *Panel A shows the overrepresentation of specific triplet repeats in human genome. CAG is the most represented TNR in exons (Kozlowski et al. 2010) and the second in terms of overrepresentation (10.4-folds with respect to ACT used as benchmark). Panel B shows that glutamine is also overrepresented as codon translation for CAG triplets (61%). Adapted from Kozlowski et al. 2010.*

In humans, Glutamine-encoding repeats are by CAG codons by approximately 72% (Athey et al. 2017). Specifically, CAG triplets are consistently more abundant in short homorepeats (4-8 residues) than in shorter glutamine stretches (1-3 residues) (Mier and Andrade-Navarro 2018). Short Glutamine stretches are generally stable and they are present in approximately similar ratios in proteins[48] (Mier and Andrade-Navarro 2018). This pattern is observed in most vertebrate species (Mier and Andrade-Navarro 2018). However, in primates for longer repeats (>8) the CAG triplet usage is reduced as compared short repeats (1-

[47] N = A, T, C or G

[48] In particular, there is on average about 10 "Q", one "QQ" and 0.1 "QQQ" per protein (Mier and Andrade-Navarro 2018).

8). Short stretches (1-8) also show a general length similarity in primates, especially in the shorter repeats. Moreover, Glutamine homorepeats do not appear to be significantly longer in humans than in the other non-human primates (Mier and Andrade-Navarro 2018), as suggested in the past (Rubinsztein et al. 1995), therefore additional studies are required to verify CAG repeat lengths in primates. Interestingly, there are almost no CAA pure regions encoding for polyQ stretches (Mier and Andrade-Navarro 2018). Furthermore, stretches with more than ten repeats are rarely stable in length among primate species, while 4Q stretches show extreme levels of stability. Moreover, when only CAG regions associated with diseases are considered, human polyQ region tends to be the longest one, with unexpectedly high proportions of CAG codons (mean value >90%). Non-human primates evolutionary closer to human also tend to have longer repeats in those regions (Mier and Andrade-Navarro 2018).

According to the above results, CAG repeat distribution and conservation in exons could be due to alternative or additional forms of *a posteriori* natural selection on the expansion events. In this scenario, it is not immediate to understand why they would favour the appearance of long, pathological repeats. However, for natural selection to occur the key factor is fitness, interpreted as the relative contribution of the individual to the next generation (Maynard-Smith 1989; Orr 2009). Whenever a condition enhancing fitness is present, this will also allow natural selection to act, regardless of the detrimental effects potentially exerted on the individual by the condition. Intriguingly, brain magnetic resonance imaging analysis conducted on subjects carrying below-threshold alleles demonstrated that specific trinucleotide repeats in the Htt gene influence 'normal' brain structure in humans (Mühlau et al. 2012). Specifically, longer repeats in the normal range are associated with increases in grey matter content of the *Globus pallidus* subcortical brain structure (also known as Paleostriatum). Consistently, in an in vitro stem cell assay, heterologous Htt from species with progressively longer CAG repeats was found capable of promoting neural formation in a CAG dependent manner (Lo Sardo et al. 2012). Additional evidence in favour of a neuronal role of HTT polyQ come from a Htt CAG-depleted murine model (Clabough and Zeitlin 2006). Although no evident developmental defects were reported, these mice show impairment in memory and learning probably linked to synaptic activity and plasticity.

2.2. Evolutionary history of Huntingtin gene

Before the genomic era, DNA cloning and sequencing techniques laboriously allowed make available few Htt gene homologues, including that of mouse (Lin, Nasir, and MacDonald 1994; Barnes et al. 1994), rat (Schmitt et al. 1995), pupperfish (Baxendale et al. 1995), zebrafish (Karlovich et al. 1998), vinegar fruit fly (Z. Li et al. 1999) and pig (Matsuyama et al. 2000). The study of these homologues has clarified that the Htt gene appeared very early in the history of life and even amoebas as *Dictyostelium discoideum* do carry a copy of the gene homologous to that of vertebrates (Myre et al. 2011). The experimental removal of *D. discoideum* Htt ortholog has shown that, although absence of HTT does not compromise viability, it causes cell defects by impairing the response to osmolarity variations (Myre et al. 2011). Moreover, it has been shown to impair chemotaxis and cytokinesis (Wang et al. 2011).

Three additional papers on Htt evolutionary history have been published by the laboratory where this part of the present Ph.D. thesis work has been developed. Evidence on Htt evolution came from homologous genes characterization in two basal chordates as *Ciona intestinalis* and *Ciona savignyi* (Gissi et al. 2006). This study shed light on the first steps of vertebrate Htt orthologous genes evolution. In tunicates, which are separated from vertebrates by about 520 million years of independent evolution, despite the presence of a Htt orthologous gene encoding for a 2,946 amino acid-long protein product, CAG/CCG repeats in the exon 1 of the gene homologous region are totally absent. Moreover, a conservation analysis spanning the entire gene revealed that the 5'-terminal portion of the gene was less conserved than its central and 3'-terminal region. This suggests that the first part of the gene (i.e. the one which bears the CAG repeats tract) underwent diversification along vertebrate evolution, while the rest of the gene is more likely connected to some ancestral function common to all chordates (Gissi et al. 2006). Moreover, cloning of amphioxus (*Branchiostoma floridae*) Htt ortholog revealed the presence of the CAG tract represented by two CAG repeats and no CCG region (Candiani et al. 2007). The first Q in *B. floridae* occupies the 18° amino acid position as in humans. Moreover, amphioxus Htt gene structure showed a high level of intron-exon conservation. Gene expression is detectable in the early neurula stage, and it localizes in the neural plate cells population. At subsequent stages, Htt expression is retained in the neural compartment, strongly suggesting a function during neural development. At subsequent larval stages, HTT is detected in the neural tube, with the strongest signal being present in its most anterior part. Finally, characterization of Sea urchin (*Strongylocentrotus purpuratus*) Htt gene structure revealed that *S. purpuratus* Htt gene is more similar to that of vertebrates than it is to that of *Ciona* genus, implying that CAG region in *Ciona* was likely lost following its separation from other deuterostomes (Tartari et al. 2008). Sea urchin Htt bears the beginning of a CAG tract (encoding 2Q) but in an amino acid environment different from that of cephalochordates and vertebrates. Moreover, Htt expression in echinoderms is predominantly confined to non-neural tissues (Kauffman et al. 2003). A first comparative analysis of the entire gene in several species (18 sequences, **Figure 5**) of protostomes and deuterostomes showed that N-17, the amino acid sequence before the polyQ tract, appears very highly conserved in vertebrates (Tartari et al. 2008). This comparative analysis was also the first report that a full trend of appearance of Glutamine residues can be traced from the very beginning of metazoan evolution (Tartari et al. 2008). Specifically, at the base of protostome-deuterostome divergence there probably was one or no Q and only deuterostome homologues show the appearance of a Q tract with 2Q in echinoderms and cephalochordates (Tartari et al. 2008). It is only with vertebrates that a real polyQ stretch (QQQQ) comes into existence, a feature that is extremely conserved thereafter. Moreover, CCG repeats (encoding the polyP region) appear to have originated after the separation of mammals from other vertebrates (Tartari et al. 2008) and it is only with mammals that a further increase in the length of the polyQ stretch is observed.

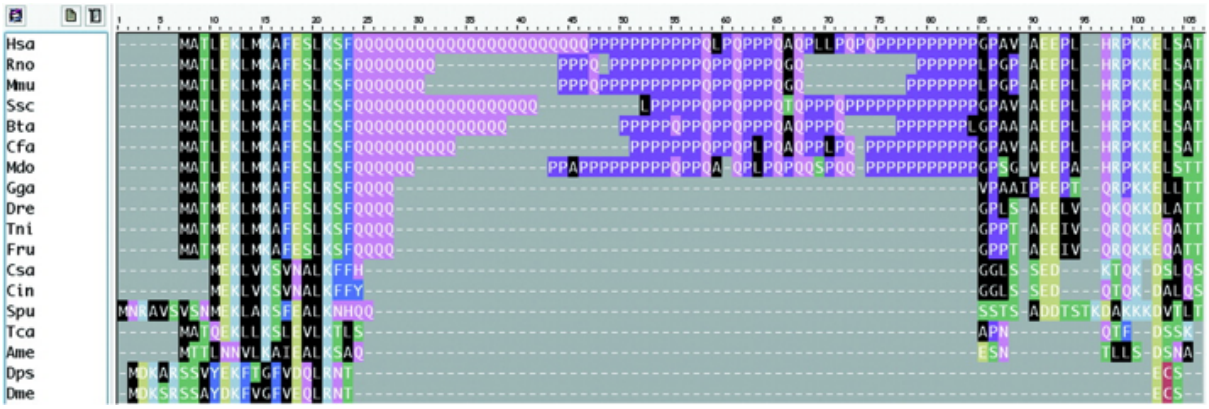

Figure 5: Multiple alignment of HTT protein sequences from different organisms. *Sequences are ordered from top to bottom according to their phylogenetic distance from* Homo sapiens *(Hsa, top). The CAG-encoded Glutamines (Q) are highlighted in purple, Prolines (P) are highlighted in blue (Prolines are encoded by CCN, where N can be any of the A,C,G,T DNA bases).* Rattus norvegicus *(Rno);* Mus musculus *(Mmu);* Sus scrofa *(Ssc);* Bos taurus *(Bta);* Canis familiaris *(Cfa);* Monodelphis domestica *(Mdo);* Gallus gallus *(Gga);* Danio rerio *(Dre);* Tetraodon nigroviridis *(Tni);* Fugu rubripes *(Fru);* Ciona savignyi *(Csa);* Ciona intestinalis *(Cin);* Strongylocentrotus purpuratus *(Spu);* Tribolium castaneum *(Tca);* Apis mellifera *(Ame);* Drosophila pseudoobscura *(Dps); and* Drosophila melanogaster *(Dme). Mammals show a polyP region (interspersed of glutamines and other amino acids) and sometimes particularly long polyQ regions (e.g.* Homo sapiens *and* Sus scrofa*). Adapted from Tartari et al. 2008.*

AIMS

The goal that I wished to achieve with this study, as part of an on-going effort in the host laboratory aimed at tracing Huntington's Disease-causing gene throughout evolution, was to reconstruct and understand the evolutionary origins of the CAG repeat embedded into the exon 1 of the Htt gene. This goal could be achieved by collecting DNA sequences from orthologous genes in order to allow a comparative analysis of the differences and similarities between the human sequence and that of other animal species. More specifically, existing sequences could be retrieved from public databases and/or assessed directly by sequencing from biological samples. These samples could be made available from already in place or newly established collaborations. Htt exon 1 sequences could then be aligned to each other in a multiple alignment, resulting in a detailed picture of Htt exon 1 CAG repeats along the tree of life. The multiple alignment, when subjected to a bioinformatics analysis of the selective pressures, could be used to elucidate the evolutionary features of this simple repeat.

RESULTS

1. The kick-off: database search

The amount of sequencing data publicly available for many organisms has kept accumulating at an unceasing pace[49], especially in the last twenty years after the advent of Next Generation Sequencing (NGS). Nucleic acids sequencing methods experienced the greatest and fastest development after the turn of the millennium, when first-generation dideoxy chain-termination method (Sanger sequencing), was rapidly replaced by cost-effective shotgun whole-genome sequencing (WGS) to generate the early drafts of the human genome (Lander et al. 2001; Venter et al. 2001; Heather and Chain 2016). With the advent of NGS, genome drafts for several vertebrate model species were then rapidly produced (Mouse Genome Sequencing Consortium et al. 2002; Genome 10K Community of Scientists 2009; Zhang et al. 2014; Jarvis et al. 2014) and several projects to produce high-quality genomes for most organisms are underway (Koepfli et al. 2015; Zhang et al. 2015; Pennisi 2017; Teeling et al. 2018; Lewin et al. 2018).

By constraining costs and time requests, NGS allows the rapid production of a great amount of genomic data. These data are usually released upon the publishing of the genomic analyses and stored into dedicated databases. Any study of genetic sequences should first attempt to rely on existing data for experimental planning. For this reason, starting from a multiple alignment of HTT orthologues available in the laboratory where I conducted this thesis work (Candiani et al. 2007; Tartari et al. 2008), I have searched the main publicly available DNA sequence databases for Htt orthologous genes. These sequences were mainly available from databases storing WGS data such as Genbank and Ensembl databases. Other databases, as the UCSC Genome Browser (Tyner et al. 2017), were sometimes used for cross validation. In most cases, sequence data were often retrieved and stored using the FASTA file format. An outline of the databases most relevant for this study (Genbank, Ensembl and UCSC) is provided in **Appendix 1**.

1.1. Retrieved DNA sequences

For the preliminary analysis in public databases, I decided to focus on vertebrates as previous results from the laboratory where I conducted this Ph.D. thesis work suggested that the CAG repeat originated at the bottom of vertebrate radiation (**Figure 5**) (Candiani et al. 2007). I also included few vertebrate outgroups (i.e. taxa that serve as external reference with respect to the set of organisms under study) in the analysis to further confirm this view.

While the degree of representation in sequence data may differ considerably among taxonomic group, they are nonetheless available for all major clades of vertebrates, including mammals, birds, reptiles, amphibians and fishes (i.e. lungfishes, bony fishes and cartilaginous fishes). Among the orthologous genes resulting from direct gene search by name or by BLAST, a total of 100 DNA sequences were retrieved from public databases and were considered reliable for being included in the final multiple alignment. These selected Htt sequences belonged to 96 vertebrate species, 2 vertebrate outgroups both belonging to cephalochordates (genus *Branchiostoma*) and also included 2 chordate outgroups, the acorn worm (*Saccoglossus kowalevskii*) and the purple sea urchin (*Strongylocentrotus purpuratus*). In the vast majority

[49] "JGI GOLD | Statistics." Accessed July 07, 2018. https://gold.jgi.doe.gov/statistics

of cases, these were the reference sequences annotated in the databases for a given species, implying one sequence per species (i.e. the sequence from the Primary Assembly, see **Appendix 1** for a definition of Primary Assembly). Genbank was always used as primary source, and Ensembl as secondary source for validation or for alternative searches in a few genome assemblies. In particular, for two sequences (*Macropus eugenii* and *Pteropus vampyrus*), the record was present only in Ensembl, which was then used as primary source. Four more sequences (*Takifugu rubripes, Branchiostoma lanceolatum, Branchiostoma floridae* and *Homo neanderthalensis*) were retrieved directly from the raw data associated with the relative publications (Baxendale et al. 1995; Candiani et al. 2007; Prüfer et al. 2014). An excel file reporting the database accession numbers for all sequences was produced and is reported in **Appendix 2**. A concise summary of the species included in the analysis, grouped by their taxonomic relationships, is reported in **Table 1**.

Phylum	**Subphylum**	**Class**	**Count**
Chordata	Vertebrata	Mammalia	46 (of which 18 Primates)
		Aves	18
		Reptilia	5
		Amphibia	3
		Actinopterygii	22
		Sarcopterygii	1 (*Latimeria chalumnae*)
		Chondrichthyes	1 (*Callorhinchus milii*)
	Cephalochordata	Leptocardii	2 (order Amphioxiformes)
Hemichordata		Enteropneusta	1 (*Saccoglossus kowalevskii*)
Echinodermata		Echinoidea	1 (*Strongylocentrotus purpuratus*)

Table 1: Summary of the 100 Htt exon 1 orthologous sequences retrieved from public databases. *The count refers to the number of species included in the final dataset for each taxonomic group.*

The availability of whole-genome sequencing (WGS) data, and particularly of those vertebrate genomes, reflects the interest for model organisms or endangered species in genetic research. However, it also reflects the challenges associated with genome sequencing. These challenges arise from the size and the complexity of the genomes themselves, which determine the amount of raw sequencing data required for a reliable

genome assembly. Genome sizes are normally reported for the haploid genome, as no genome assembly is currently phased into haplotigs (i.e. a genome assembly with distinguished homologous chromosomes). These haploid genome sizes are usually expressed in C values, which represent the total weight of DNA in picograms (the conversion factor picograms/base pairs is 1 pg = 978 Mbp). Today, C value information is available for many species[50]. The lowest C values are found in birds, which have been shown to have the smallest genomes among vertebrates (mean C value of about 1.36 pg)[51]. Interestingly, this feature appears to be associated with flight, as non-flying birds have higher C values on average (the highest reported being that of the ostrich *Struthio camelus,* 2.16 pg) (Zhang et al. 2014). As a result, plenty of bird WGS data are available (Zhang et al. 2014), and this made birds very good candidates for the search of gene orthologs.

For mammals, C values are narrowed around the average of 3.2 pg (s.d. 0.8 pg, ranging from 1.6 to 8.4 pg)[52], while average C values for reptiles are about 2.3 pg (s.d. 0.7 pg, ranging from 1 to 5,4 pg)[53]. By contrast, amphibians show extremely variable genome sizes, with mean C values in the order of 18.9 pg and a s.d. of 19 pg[54]. Among them, C values can peak up to 120 pg in the family Proteidae, but can be as little as 1 pg in some frogs and toads. Extra-large genome sizes constitute a practically insurmountable barrier for NGS-based WGS, and readily explains why there are so few amphibians genomes available. This is also the reason why I had been able to retrieve the sequence for only three amphibian species (**Table 1**).

Fish genome sizes are also variable, with a mean C value of 2.3 pg (s.d. 2.1 pg, ranging from 0.4 to 17.5 pg)[55] determined by the relatively small genomes of bony fishes, whereas cartilaginous fishes (Chondrichthyes) and lungfishes (Sarcopterygii) show the highest C values on average (up to 17 pg).

It should be noted that the major challenge with large genome sizes is not represented by the size of the genome *per se* (which might eventually turn into higher sequencing cost to obtain enough sequence coverage), but from the widespread presence of low-complexity regions (i.e. tracts which are highly enriched in tandem nucleotide repeats). Indeed, these low-complexity regions account for most of the 'extra' DNA (Biscotti, Olmo, and Heslop-Harrison 2015) and genomes assemblers often fail to map reads in these genomic regions, which are consequently missing. For this reason, the majority of huge genomes still defy WGS. Yet, this situation appears to have reached a turning point thanks to the recently developed long read-based approaches (collectively called Third-Generation Sequencing - TGS, as opposed to NGS). This improvement is exemplified by the very recent genome assembly of the Mexican axolotl (*Ambystoma mexicanum*), with a C value of 48 pg (Nowoshilow et al. 2018).

1.1.3 Disregarded sequences in our Htt gene search

In my analyses, not all Htt exon 1 orthologous sequences initially retrieved were used to produce the final multiple alignment. There were multiple reasons for exclusion, and in general the entire process required

[50] "Animal Genome Size Database." Accessed July 07, 2018. http://www.genomesize.com/.
[51] *Ibidem.*
[52] *Ibidem*.
[53] *Ibidem.*
[54] *Ibidem.*
[55] *Ibidem.*

accurate manual curation. Sequences that were disregarded fell into two main categories: 1) records showing complete absence of recognizable Htt exon 1 sequences; and 2) records with apparently incomplete Htt exon 1 sequences. Indeed, several genes annotated as Htt did not show the DNA sequence expected from the comparison with their relatives in the phylogeny. This finding, rather than suggesting abrupt changes in gene content for those species, is to be interpreted in the first place as an issue due to NGS technology and the assembly process of NGS data. In fact, it has been shown that NGS suffers from context-induced sequencing bias, which means that some genomic regions tend to be underrepresented in sequencing reads (including GC-rich regions and low-complexity, repeated regions). If these genomic regions are missing from the raw data, they will be then missing in the genome assembly and also in the annotated genes. Moreover, most of NGS genomes available are missing an important portion of the genomic DNA sequences as a result of the context-induced poor quality of genomes assembly based on short-read sequencing. This happens because even when low-complexity regions are present in the raw reads, it is often impossible to align them into the assembly given the short length of NGS reads (100-300 bp). The consequence of the absence of the region in the raw data or of its absence in the final genome assembly is that the automated process of gene annotation by the algorithms fails to correctly identify and annotate coding regions of the genome, and wrong gene sequences are ultimately reported in the databases. A report suggests that about 30% genes might be completely missing from even high-quality NGS genomes (Zhang et al. 2014). Of the remaining genes, many are missing several portions of the gene. This happens because NGS genomes are not only incomplete but also highly fragmented (i.e. they lack contiguity along the chromosome), and if by chance part of a gene falls into a different scaffold, that portion (in our case the exon 1 of Htt) will likely be missing from the final sequence. Still, the rest of the gene can be annotated by the automated pipelines, but the missing parts will be excluded and sometimes even replaced by the surrounding intronic context and erroneously interpreted as coding sequence.

Htt exon 1 might be a notable example of the aforementioned assembly errors, as its repeated regions greatly reduce the probability of automatically assembling that entire genomic region into a contiguous scaffold, often producing a breakpoint in the genome assembly at that particular position. Consequently, while sometimes the start of exon 1 sequence could be identified, part of exon 1 appeared to be missing as a result of errors in the assembly and annotation process. Moreover, the resulting absence of natural contiguity, and eventually of a transcriptional start site, often turns into part of the 5' untranslated region (UTR) or of other upstream genomic regions being interpreted by the algorithms producing the annotations as transcriptional start sites and being merged into the Htt coding sequence. In these cases, the interruption often coincided with the repeated regions, once more pointing at the unreliability of NGS approaches to correctly assemble sequence reads in contigs and to annotate low-complexity regions.

1.1.4 Sequences that required manual inspection of the raw data

The short length of NGS reads can be challenging even if the genomic region of interest is correctly annotated (He et al. 2011). In some cases reads may have correctly aligned, however the resulting consensus sequence may not be reliable if no read was able to encompass the length of the repeat. This issue is

widespread and explains why all NGS genome assemblies produced over the years are substantially incomplete (Schatz, Delcher, and Salzberg 2010; Ye et al. 2011; Treangen and Salzberg 2011). As genome assembly relies on the alignment of individual sequencing reads using partially overlapping portions of the sequences, when a low-complexity region extends beyond the length of the sequencing reads there is no mathematical mean to determine a meaningful alignment between the reads belonging to that particular region (Treangen and Salzberg 2011). In this case, the only possibility is to rely on some 'impurity' within the repeat, usually represented by SNPs, to be used as anchor points for the alignment (He et al. 2011). Yet, this allows to solve the issue only in a limited number of cases. In particular, as NGS reads usually span 100-300 bp, if the repeat length is around this limit or extends beyond its correct alignment is always challenging and in many cases just impossible. This case is often found in NGS data from public database, where short reads have been the standard in the last twenty years. While the relatively short CAG repeats of most vertebrate species analysed did not make this problematic in our dataset, an interesting example of this issue, in which I have also run into, is that of the hominid genomes published since 2010. Paleogenomics results suffer from the combination of low-quality starting material and DNA amplification-based shotgun sequencing, and these genomes are no exception to this. In the case of the Denisova genome (an archaic human of the genus *Homo*), the longest read aligning over the Htt CAG repeat region reported 19 repeats, however no Solexa read encompassed the CAG repeat from both sides (**Figure 3**) (Reich et al. 2010). Even for the genome for the Neanderthal man (*Homo neanderthalensis*) published in 2014 and derived from a very well-preserved (but still highly-degraded) finger phalanx discovered in 2008 during excavation in the east gallery of Denisova Cave in the Altai Mountains (Siberia) (Prüfer et al. 2014), it was only possible to putatively attribute at least 23 repeats.

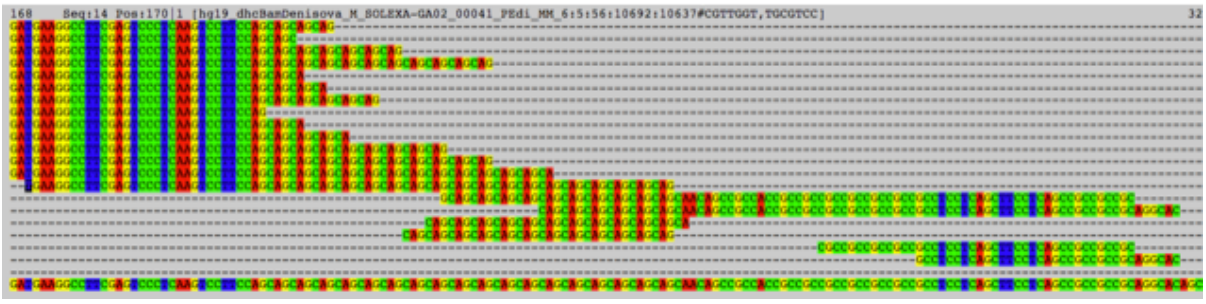

Figure 6: Multiple alignment of individual reads mapped to the Htt CAG repeat region within the BAM file of the 2010 Denisova genome assembly. *In order to assess the repeat length at least one read encompassing the CAG repeat must be present in the raw data. In this case, no sequencing read has been capable of reading through the anchor points on both sides of the repeat. Consensus is shown in the bottom row. The multiple alignment is displayed using software SeaView v4.5.4 (Gouy, Guindon, and Gascuel 2010).*

As to July 2018, there were 484 publicly available vertebrate genome projects on Genbank[56]. However, this number has experienced an exponential growth associated to the drop in WGS costs in the last few years. At the end of 2015 when this preliminary inspection was concluded, this number was limited to 267[57]. Given all the aforementioned issues and uncertainties, the 100 manually curated Htt exon 1 sequences selected for this study may still represent a good subset of the ones currently available in public databases. Importantly, they were highly consistent with results from our own sequencing effort (see Section 3).

[56] "Genbank: search keyword 'vertebrata'" Accessed July 25, 2018 https://www.ncbi.nlm.nih.gov/genome/?term=(vertebrata)

[57] https://www.ncbi.nlm.nih.gov/genome?term=(vertebrata)%20AND%20(%222000%2F1%2F1%22%5BCreate%20Date%5D%20%3A%20%222015%2F31%2F12%22%5BCreate%20Date%5D)

2. Sampling

2.1. Rationale and the choice of the species

The rationale behind the extensive sampling carried out in the present work posits that the more sequences available from closely-to-distantly related organisms, the more likely it is to obtain a detailed picture of the origin and evolution of a DNA sequence, in our case the Htt CAG repeat. This is the guiding principle behind comparative genetics and, as stated by one of the most prominent living evolutionary biologist, *"the comparative method is the gold standard for testing evolutionary hypotheses. [...] The comparative method can address the adaptive significance of traits"* (Nesse 2011). Accordingly, samples were collected from as many vertebrate species as possible, in order to maximize the biodiversity represented in the pool of samples. They were collected in a variety of forms, including different biological tissues, museal skins and already extracted DNA. As for sequence data from public databases, the choice of focusing on vertebrates relied on previous results suggesting that the CAG repeat originated at the bottom of vertebrate radiation (Candiani et al. 2007). All major clades of vertebrates were included in this study, i.e. mammals, birds, reptiles, amphibians and fishes (including cartilaginous fishes). Within-clade sampling was highly dependent on the availability of samples from collaborators, but was also informed by preliminary results.

2.2. Network of suppliers

A large network of collaborators and sample suppliers from all over the world was established for the present work (**Figure 7**).

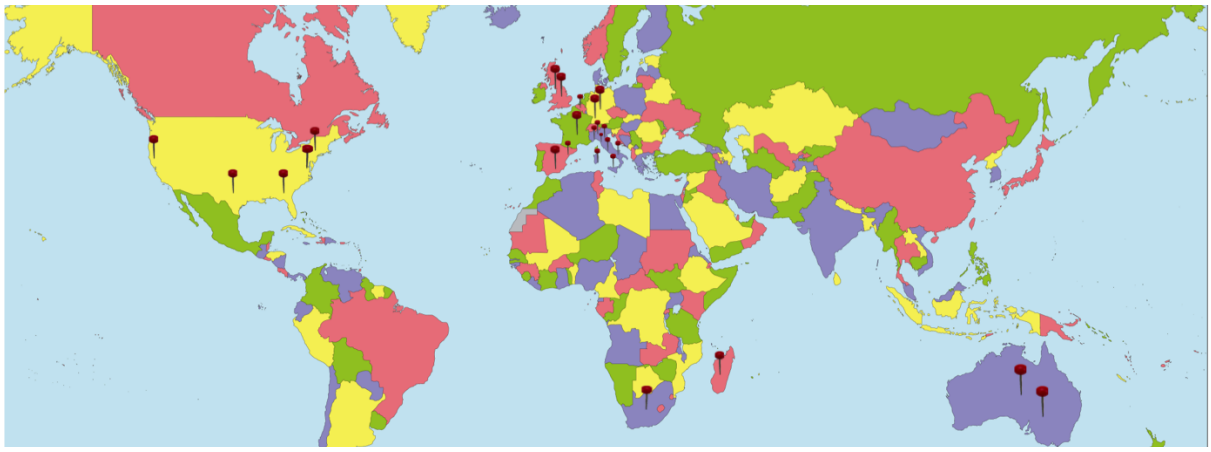

Figure 7: Location of suppliers. *Most collaborations for this project were established with groups and Institutions in Europe and North America, yet some valuable samples came also from Africa and Australia.*

This work was made possible thanks to the connections already available in the laboratory and through the work of other colleagues in the laboratory, and also thanks to the crucial support of my Ph.D. supervisor, Prof. Nicola Saino (University of Milan) and his collaborators worldwide. Other collaborations were also established throughout the project development. Overall, over 43 researchers belonging to 34 Institutions

were involved and enquired to provide samples to the project. Most of our collaborators were able to provide samples, and many of these samples were included in the analyses. Specifically, a total of 1,307 samples from 33 collaborators were collected, precisely catalogued and stored in the laboratory. Not all these samples were included in the analysis, as detailed in Section 3. A detailed list of providers can be found in **Appendix 3**. These providers included research institutions such as biology and zoology departments (e.g. the Dept. of Zoology at the University of Sassari in Sardinia, which provided several mammalian samples, but also that of Bicocca and Insubria, which provided hard-to-collect bat samples), veterinary clinics and associated institutions (including the "Banca Tessuti Cetacei del Mediterraneo", the cetacean tissue biobank in Padua) several museums of Natural History (especially the Natural History Museum of Milan, but also the Natural History Museum of Victoria in Australia that provided several invaluable samples of monotremes and marsupials) and bioparks where monkeys and apes were kept in captivity (in particular the Biopark of Rome, which was able to provide several important samples, as detailed below). Some providers required a special mention, thanks to the quantity and quality of samples provided, to the peculiar nature of the samples and to the efforts devoted to this project.

2.2.1 Retrieving the Htt gene exon 1 from museum collections

Of particular interest and value were the specimens that it was possible to obtain from museum collections. Two collections of special interest were represented by the Natural History Museum of Milan, to which we could have access thanks to the contribution of Dr. Giorgio Bardelli and Dr. Michela Podestà (both employed at the museum), and by Natural History Museum of Pavia, through its curator Dr. Edoardo Razzetti. From museum specimens preserved by various means (tanning, ethanol) it was sometimes possible, with the approval and supervision of the curators, to obtain small tissue samples. These represented minimally invasive tissue cuts. Attention was put on these samples, due to the risk of contamination in PCR-based assessment of DNA sequences. This is especially true when it comes to museum collections, where preparation of the samples and age has got rid of most DNA and where handling of the samples and inappropriate storage may allow for the greatest variety of contamination sources to get in contact with the samples (bacteria, dust, curators and visitors DNA among others). Due to the low DNA quality or the presence of contaminants inhibiting the PCR reaction, many of these samples could not be amplified. On the other hand, some samples provided valuable genetic information and results were often consistent with the expectations. Attention was paid to manually inspect all sequences deriving from these sources, in order to reduce the risk of false positive results. A sample worth recalling that was included in the analysis is the pangolin (*Phataginus tricuspis*[58]). This sample was retrieved from an ethanol-preserved foetus during a visit at the Natural History Museum of Milan. Two cuts were made directly in the tissue by the curator, Dr. Giorgio Bardelli, and the resulting tissues pieces were quickly sealed within eppendorf tubes. Both cuts amplified, revealing a peculiar sequence (ID68), as it is expected in the case of a species that has highly

[58] This sample was originally determined as *Manis sp*, without further specification. It was brought to the museum on March 3rd, 1969 from the Central African Republic (near Bagandou, Mbaiki Prefecture) as a donation from Dr. Silvio Pampiglione. It was once indicated as *Manis tricuspis*. However, after subsequent closer inspection, Dr. Bardelli could trace back its story and accordingly the sample was renamed as *Phataginus tricuspis* (Rafinesque, 1821).

diverged from the rest of mammalian species, such as that they deserve a dedicated taxonomic order, Pholidota. Another striking example is that of the Museums Victoria in Australia. Thanks to their prompt answer and shipment of samples, I was able to sequence the Htt exon 1 of the echidna (*Tachyglossus aculeatus*) (ID76), an extremely rare and valuable sample for a species at the root of mammalian radiation. Specifically, they provided two tissue samples belonging to two distinct individuals, which independently confirmed the sequence. In all, 7 out of 32 samples, belonging to 4 species, from the Natural History Museum of Milan were included in the final dataset[59] (189 samples overall).

Overall, these collaborations confirm that Natural History museums can represent important sources of biological samples also for DNA studies. They can act not only as repositories of phenotypic information for rare, endangered or even extinct species — as they did for centuries in the past — but they can also act as potential sources of genetic information, especially for species otherwise hard or even nearly impossible to sample *de novo*.

2.2.2 The importance of primates

A specific focus on primates was conducted at various stages. This was primarily because of their close relationship with human beings, as the extant primate species represent the closest outgroups of our species. In a first step, several samples from the Biopark of Rome were provided in 2013 by a collaborator of the laboratory (Dr. Cristina Martinez-Labarga, University of Rome-Tor Vergata). Preliminary results from some of these informed a visit to the The Kyoto University Primate Research Institute (PRI) in 2015 (followed by a second visit to PRI in 2017) to perform sequencing on samples that was hard to ship to Italy, as detailed further on in Section 5. Furthermore, a second shipment from Dr. Martinez-Labarga occurred in 2018. A detailed account of this part of the project is provided in the relative section of results.

2.5. Sample collection

Unless otherwise stated, samples were directly retrieved by Dr. Michela Pacifico, a research fellow in the laboratory, and me or were delivered to the laboratory by mail. Samples may have consisted of small tissue fragments from the widest variety of organs (i.e. muscle, brain, entrails) or other biological tissues such as blood, and sometimes even hairs or feathers. In few cases already extracted DNA was available. The amount of sample available was also highly dependent on the supplier and on the species. Indeed, only very tiny amounts of tissue were available for the smallest vertebrate species (e.g. lizards). Depending on the sample source and type, collection has been carried out using different strategies. For example, samples already stored frozen by the supplier were collected on dry ice, or were sometimes thawed and preserved in ethanol (70° or absolute) during the transport or during mail shipment. Whenever available, the information on the original sample collection date was recorded. Upon reception in the laboratory, samples were initially stored frozen at -20 °C. Long-term storage of tissue samples was achieved at -80°C.

[59] Preliminary results from this work were reported in poster presentation to the congress of Società Italiana di Evoluzione Biologica (SIBE) in 2015:
https://air.unimi.it/handle/2434/465175?mode=simple.878#.W40J-pMzYdV

2.6. Sampling results

Sampling was conducted mainly over the first three months of the project, yet many samples were delivered to the laboratory also in the following months. Of the total of 1,307 samples collected, 534 were mammals, 304 birds, 233 reptiles, 136 amphibians, 99 fishes (of which 10 cartilaginous fishes) and one Agnata (*Lampreda sp.*). **Table 3** summarizes the number of samples per species collected, grouped in major clades. A detailed excel file listing all samples delivered and their specifications is available upon request to the Cattaneo laboratory. The samples that were successfully processed and included in the analyses were all drawn from this list.

Phylum	**Subphylum**	**Class**	**N. of orders**		**N. of families**		**N. of species**	
Chordata	Vertebrata	Mammalia	21	*29*	62	*158*	156	*5,513*
		Aves	28	*42*	51	*227*	93	*10,425*
		Reptilia	7*	*4*	25	*81*	139	*10,711*
		Amphibia	2	*3*	17	*59*	125	*7,302*
		Actinopterygii	21	*44*	46	*446*	69	*33,600*
		Chondrichthyes	4	*15*	5	*59*	6	*1,100*
		Agnatha	1	*2*	1	*4*	1	*110*

Table 2: Summary of the samples collected during the project. *Numbers refer to the count of taxonomic groups for each category present in the final sample. *Including suborders. Numbers in italic represent the count of extant taxonomic units for the specific taxa, as detailed in text.*

The number of samples for each class of vertebrates partially reflects the network of collaborators established for this work and the intrinsic difficulties associated with sampling of some taxonomic groups (e.g. lungfishes, for which no sample was available). It also reflects the phylogenetic history of the different taxonomic groups, as in the case of jawless fishes (Agnatha), a clade that comprises approximately only 110 extant species[60]. Yet all major clades had similar degrees of representation in the sample pool. Specifically, 0.5% of all 1,100 living species of cartilaginous fishes[61], 0.2% of about 33,600 living species of ray-finned

[60] "Britannica | Agnathan." Accessed July 18, 2018. https://www.britannica.com/animal/agnathan

[61] "Biodiversity Explorer: Chondrichthyes (cartilaginous fish, including sharks, rays & chimaeras)." Accessed: July 17, 2018. http://www.biodiversityexplorer.org/chondrichthyes/

fishes (Actinopterygii)[62], 1.7% of 7,302 living species of amphibians (Frost 2016), 1.3% of 10,711 living species of reptiles[63], 0.9% of 10,425 living bird species[64] and 2,8% of 5,513 living mammalian species (Wilson and Reeder 2005) were represented at the end of the sampling effort. Importantly, a greater representation was obtained at the level of families and orders. In terms of families, cartilaginous fishes were represented by 8%, bony fishes by 10%, amphibians by 29%, reptiles by 31%, birds by 22% and mammals by 39%[65]. In terms of orders, cartilaginous fishes were represented by 27%, ray-finned fishes by 48%, amphibians by 67%, reptiles by 100%, birds by 67% and mammals by 72%[66].

As detailed in the following sections, not all the samples were ultimately processed or included in the analyses, depending also on the quality of the sample or depending on the interest of the sample for the analysis. However, for most of the samples the original tissue and/or the extracted DNA is conserved in Cattaneo laboratory and available for future studies.

[62] From IUCN Red List citing "Fishbase." Accessed: October 20, 2014. http://www.fishbase.org
[63] From IUCN Red List citing "The Reptile Database compiled by Peter Uetz and Jirí Hošek."
[64] From IUCN Red List citing "BirdLife International. 2014" Accessed July 24, 2014. http://datazone.birdlife.org/home
[65] "Integrated Taxonomic Information System." https://www.itis.gov/
[66] Ibidem.

3. Development of protocols for Htt exon 1 amplification and sequencing

In order to PCR-amplify and then sequence as many as possible Htt exon 1 orthologs I tested and established several PCR protocols and conducted PCR reactions together with Dr. Michela Pacifico. All protocols relied on the amplification of relatively short DNA sequences, that is usually producing fragments between 100 and 1000 bp suitable for cloning or direct sequencing. Differences in PCR protocols were mainly requested to adapt them to a specific taxonomic group. Overall, 6 PCR protocols were employed to produce the final dataset.

3.1. Issues with Htt exon 1 amplification

Since the development of DNA amplification techniques, amplifying repeated DNA sequences has proven challenging. Rather than being a technical issue, this in part arises from a truly biological problem, as the DNA-copying machinery often fails in low complexity regions that are widely present in most genomes. Intriguingly, these issues are also believed to be at the root of the uncontrolled expansion events that give origin to human HD alleles (see Introduction, Section 2.1.1). Other PCR-related issues are common to most comparative genetics studies, and arise from the uniqueness of every genome. Some of the major PCR-related issues encountered during the development of the PCR protocols are discussed in the following sections.

3.1.1. Issues caused by the polymerase

Issues similar to that of biological processes can occur whenever the polymerase attempts to replicate DNA *in vitro*, as during PCR amplification (Kunkel 2004; Garcia-Diaz and Kunkel 2006). Particularly, slippage (see Introduction, Section 2.1.1.) can often occur and can even escape the correcting capacity of proofreading polymerases (Lujan, Clark, and Kunkel 2015). However, this issue appears to be highly dependent on the length of the repeat itself. With the relatively short (4-15 copies) CAG repeats found in most non-human vertebrates the risk of slippage is limited, especially if high-quality, proofreading TAQ polymerases are used (Lujan, Clark, and Kunkel 2015). Yet, in principle the random amplification of unintended slippage products can occur (Lujan, Clark, and Kunkel 2015). To avoid false positive results, especially in case of unexpected results (i.e. results that largely differ from that of other species of the same taxonomic group), it can be important to validate the length of the repeat by independent amplification and sequencing. Though non-zero, the probability of a perfect match between two independent amplifications resulting in a false positive (which would imply an identical slippage error) is relatively low.

Interestingly, while the polymerase can be an issue during PCR amplification due the potential preferential amplification of slippage products, the subsequent Sanger sequencing — although still relying on a polymerase — is far less error-prone. This is because Sanger sequencing relies on a linear (not exponential) amplification of the input DNA, which should normally result in a relatively fair representation of all PCR products present in the sequencing reaction (Slatko et al. 2001). Accordingly, chromatograms containing overlapping peaks that represent the spurious PCR products of polymerase slippage can sometimes be found (**Figure 8**), especially when assessing relatively long repeats (20 copies and above).

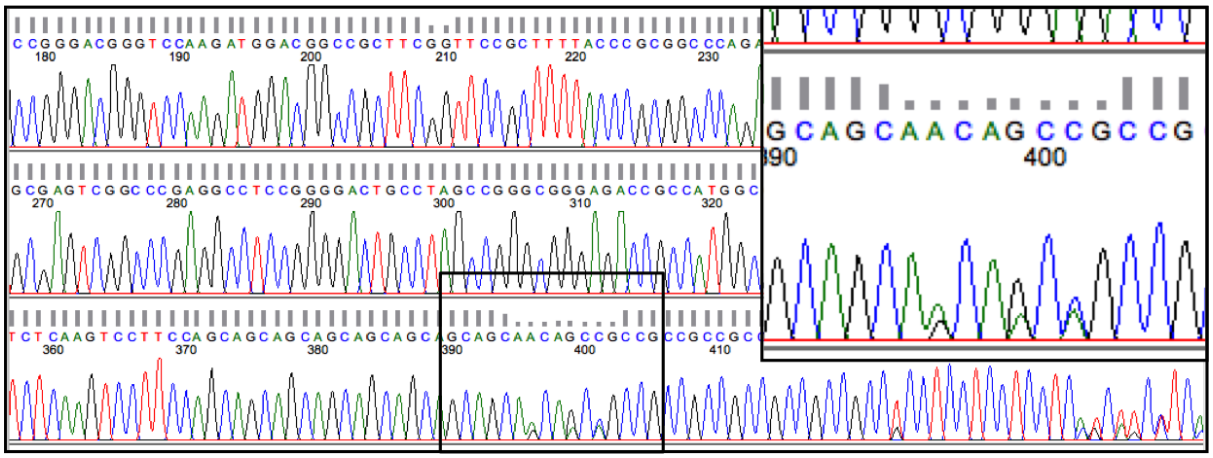

Figure 8: Sanger chromatogram showing the presence of two alternative PCR amplification products differing for one CAG repeat. *Given that the PCR fragment is read using a forward primer, two amplicons overlap until the end of the CAG repeat. Specifically, the first amplicon is TTC(CAG)$_8$CAACAG while the second amplicon is TTC(CAG)$_9$CAACAG. The PCR product belongs to a Japanese macaque with a CAG repeat length of 10/11 codons, and primates (including humans) were shown to have a CAA codon before the last CAG. This results in a mixed CAA/CAG codon (starting from position 384) followed by a mixed CAG/CAA codon (position 387) and by a mixed CCG/CAG codon (position 400). The sequence is displayed using software FinchTV*[67].

3.1.2. Issues caused by repeat length

Another issue when analysing a repeated DNA region is the length of the repeated region compared to the length of the sequencing read, as recalled in Section 1.1.3. Still, in the case of non-pathological Htt CAG repeat lengths, this limitation mainly applies to very few species, as the human and pork (*Sus scrofa*), the only animals where CAG repeats beyond 20 were so far reported.

3.1.3. Issues caused by differences in gene orthologs

A main issue when attempting to analyse gene orthologs stems from the evolutionary process itself. Even in highly conserved genes as the Htt gene, mutations disrupt sequence similarity among species, ultimately resulting in PCR primers that fail to map in the homologous regions. Indeed, due to the redundancy of the genetic code, gene conservation can only get rid of non-synonymous substitutions, while synonymous substitutions can accumulate at a normal rate. While a certain amount of mismatch (1-2 bases) can be tolerated by relatively permissive PCR conditions, strong differences between template and primer sequences simply turn into failed amplification or in the amplification of unintended targets. The latter possibility is sometimes even worse, as the case of amplicons of similar length requires sequencing to ascertain a false positive result. Intriguingly, the amplification of unintended products has sometimes the potential for discovery, as it turned out to be the case with my discovery of a non-coding Htt pseudogene in a primate lineage (see Section 5).

[67] "Geospiza. FinchTV." March 05, 2018 http://jblseqdat.bioc.cam.ac.uk/gnmweb/download/soft/FinchTV_1.4/doc/

3.2. Potential strategies and the strategy of choice

Several strategies can be employed to overcome the aforementioned issues. For polymerase-related issues, one important factor is the choice of a proofreading polymerase. Proofreading polymerases are enzymes capable of sensing perturbations in the structure of the double helix. These perturbations are determined by the mispairing of the two DNA complementary strands that occurs in the presence of nucleotide mis-incorporation or of hairpin structures (Kroutil et al. 1996). When the polymerases sense the perturbation in the structure, it stalls and its 3' → 5' exonuclease activity removes the misincorporated nucleotide[68]. Then the extension can proceed again. However, the power of these polymerases in counteracting the phenomenon is limited and applies only to single bases mismatches. Therefore, DNA amplification can sometimes result in mutated DNA strands (Kroutil et al. 1996).

On the other hand, overcoming the issues related to read length depends on the extent of the repeat to be amplified. In the case of very short repeats, NGS can be sufficient whenever the repeat is consistently shorter than the repeat length. However, it has to be noted here that short read sequencing does not overcome the slippage issue, as these technologies mostly rely on PCR product amplification and on polymerases freely floating in the sequencing reaction (Kebschull and Zador 2015). Intriguingly, long-read Third-Generation Sequencing (TGS) approaches now appear to be able to overcome these issues thanks to the absence of DNA amplification steps and thanks to the much longer reads that they provide (Roberts, Carneiro, and Schatz 2013). The higher error rates of long reads with respect to short reads are still a challenge, but algorithms for the automatic detection of repeat length are rapidly being established (Liu et al. 2017). When this thesis work was planned and carried out, TGS strategies were still under development and highly experimental, discouraging us from their employment. These being unavailable, Sanger sequencing — with its reads slightly less than a kbp-long on average (Heather and Chain 2016) — appeared to offer the most scalable, cost-effective and reliable strategy. With respect to scalability, Sanger is still more advantageous for several kind of projects, as it allows to process samples individually upon necessity, but also in multiples of 8 for moderate throughput. It should also be noted that the issue of mispairing of primers can be more easily handled with Sanger sequencing rather than with NGS by designing different primer pairs for different groups of species using the information available from close relatives in sequence databases. With respect to cost, amplicon sequencing with Illumina platforms is still more expensive than Sanger, and the extra cost is only justified when samples are pooled together or when a very high level of heterozygosity is expected within the amplicon[69] (e.g. in the presence of elevated levels of somatic mosaicism). In terms of reliability, both approaches would do, but thanks to the linear amplification as previously recalled Sanger sequencing can be considered more reliable for some applications.

[68] "NEB: Polymerase Fidelity." https://www.neb.com/tools-and-resources/feature-articles/polymerase-fidelity-what-is-it-and-what-does-it-mean-for-your-pcr

[69] The presence of two amplicons at approximately equimolar ratio can normally be detected in Sanger traces. However, in this case base calling has to be performed manually or using dedicated softwares. Moreover, if the proportion of the two amplicon differs considerably (e.g. underrepresentation of one of the two amplicons of <30%), even eye judgement might be difficult. If more than two amplicons are present (e.g. in the case of 3n and above genomes), the sequence readability is definitely compromised and different approaches are required.

3.3. Genomic DNA extraction.

Genomic DNA (gDNA) extraction represents the first challenge of a successful project aimed at assessing genetic sequences. Relevant here is the overall quality of the DNA, which may depend on many factors. The first key factor is represented by the preservation of the starting tissue/blood sample. Purity is affected when the storing agents are not appropriate for subsequent DNA extraction, as is often the case when the original purpose was not that of collecting DNA (e.g. histological tissues). Integrity of the molecules can instead be compromised by storage conditions. These can be highly variable between samples and biobanks/collections, depending on the original sampling purpose. Samples are often preserved frozen (at -20°C or at -80°C) or in ethanol (etOH) at various concentrations (normally from 70% to 90%). Other means, including proprietary storing buffers, can also be employed. In case of cold-preservation approaches, consequent cellular dehydration, mechanical stresses, thermal shocks and/or formation of ice crystals may all be harmful (Bakhach 2009). Specifically, cell membrane damages allow DNA-degrading enzymes (DNAses) that are normally present into the cells to reach the DNA no longer protected by the nucleus (Shao, Khin, and Kopp 2012). EtOh is by far the most frequent alternative for preservation at room temperature. While EtOH (or preserving agents with similar properties) offers the advantage of denaturing and inactivating DNAses, the thermal energy is still capable of inducing DNA nicking and double-stranded breaks (Nagy 2010).

Another factor that is intimately associated with DNA quality is the time elapsed from sample collection and DNA extraction, as over long timeframes even tiny forces can act on DNA molecules. Theoretically, cryogenic temperature (−196°C) may conserve tissues indefinitely (Bakhach 2009). However, in practice even the extremely high stability of double-stranded nucleic acids is insufficient to fully preserve DNA integrity for more than few years. All these factors played a major role in our experiments, as the samples to which we had access had been previously preserved in many different ways, and thus showed a wide spectrum of degradation levels. However, the importance of DNA integrity alone was limited with respect to DNA quality, as Htt exon 1 is relatively short (usually from 100 to 300 bp), and can thus be often PCR-amplified even in samples characterized by highly sheared DNA. By contrast, the presence of contaminants that could not be removed by canonical DNA extraction protocols, particularly polymerase inhibitors, was likely to be a highly detrimental factor for success in PCR amplification.

3.3.1 The protocol of choice for genomic DNA extraction.

As detailed in **Table 3**, 189 DNA samples were employed overall for genetic analyses in the final dataset. Of these, 18 DNA samples were already extracted by me before October 2015 during my Master Thesis work, using a Phenol/Chloroform-based protocol internally developed or a commercially available column-based kit from Macherey-Nagel (Nucleospin Tissue ID: 740952). I then extracted another 139 out of 189 samples between March and September 2015, before the start of my graduate studies. For a series of DNA extractions I also successfully employed a high-throughput DNA extraction kit (ZR-96 Genomic DNA - Tissue MiniPrep 2x96 Preps. ID: D3055). A total of 65 samples that were used for the analyses were processed with this kit. Other 29 DNA samples were delivered as already extracted, and the extraction methods were not reported in the shipment (i.e. the one sample of the Japanese macaque used in the analyses, the healthy

human sample from the Besta Institute, the samples from the Biopark of Rome and the samples from the collaborator Dr. Adriana Bellati). The remaining 3 DNA samples were extracted during my graduate studies.

DNA extraction method	**N. of samples**
Nucleospin Tissue	89
ZR-96 Genomic DNA - Tissue MiniPrep 2x96	65
Phenol/Chloroform-based extraction	6
Unknown*	29

Table 3: Samples subjected to DNA extraction that were included in the final dataset. **Already extracted DNA delivered to the laboratory, without specification of the DNA extraction protocol employed.*

DNA samples were immediately stored at 4°C for short-term downstream sequencing, or -20 °C for long-term storage. 4 °C storage should avoid DNA degradation if the DNA was appropriately extracted in the first place (i.e. if DNAses and other harmful enzymes were successfully removed). For 161 samples, long-term storage was achieved in numbered 96 well-plates sealed with plastic sealers, and the relative position of all DNA samples in the plates was recorded in a separate excel file. For final storage, these plates were ultimately covered with aluminium sealers, as some plastic sealers easily detach upon freezing. The rest of samples (28, of which 6 were used up by the end of the project and one was in Japan) were stored using Eppendorf tubes in a box at -20 °C.

3.4. gDNA quality control.

In order to ascertain the overall quality of the DNA samples employed in PCR, most of 189 DNA samples were run on agarose gels (usually between at 1-2X) with different stainings, often Ethidium Bromide, but also with other DNA intercalating agents as GelStar (Lonza ID: 50535) and Eurosafe (Euroclone ID: EMR440001). The most frequently employed ladder was 1 Kb plus DNA ladder (ThermoFisher Scientific ID: 10787018). For 136 of these samples, the relative picture was recorded on a GelDoc XR instrument (Bio-Rad) or on similar instruments. For 27 samples the little DNA quantity available (<10 µL) did not recommend to check DNA quality on gel (which usually requires at least the use of one µL). For 25 samples the picture was not taken or was not saved due to technical issues. For one sample (Japanese monkey) this was not obtained since extraction had already been performed in Prof. Hiroo Imai's laboratory (see also Section 5).

Samples showed the widest spectrum of DNA conservation and degradation (**Figure 9**):

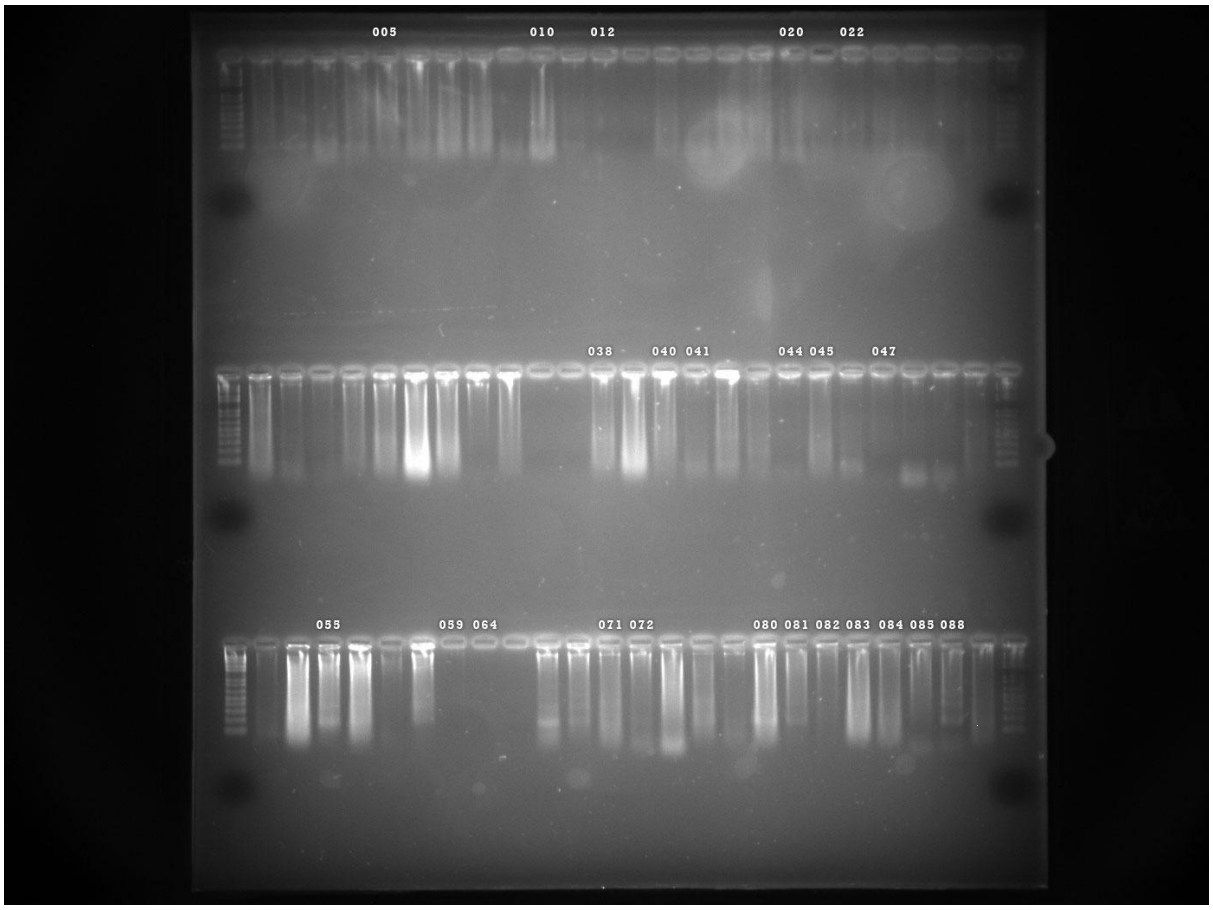

Figure 9: Different degradation level of some DNA sample employed in the study. *IDs for samples included in the final data set are reported above lanes. 1 Kb plus DNA ladder is shown for comparison in the first lane of each row.*

For example, the highly concentrated human sample conserved for diagnostic purposes, showed almost no sign of degradation. At the other end of the spectrum, very old samples, or samples that were suspected of incorrect storage prior to delivery to laboratory, could show clear signs of degradation, up to the point of being mostly sheared to less than few hundred bp fragments.

Sometimes, and especially in the beginning of the project, the DNA quality was also assessed by a Nanodrop1000 spectrophotometer (ThermoFisher Scientific). Specifically, this was conducted on 41 samples included in the work due to the early decision of not measuring the DNA concentration. As for agarose gel quality control, for 8 samples spectrophotometer assessment was avoided in order to prevent waste of little DNA quantities available. For the 23 samples extracted before 2015 and for the *M. fuscata* the information was already available.

The initial decision of not measuring the DNA concentration was dictated by several reasons. In particular, as samples have to be measured and then diluted to the final working concentration one by one, this have considerably increased the risk of contamination associated with handling samples, especially when stored in plates. In addition, samples were often of relatively low-quality and thus the DNA was likely degraded to the extent that in some cases the concentration of the genomic DNA was not detectable or unreliable in Nanodrop quantification. A measurement by spectrophotometer would then have constituted a

poor indicator of the overall number of target DNA molecules available for PCR amplification. Moreover, it is known that singleplex PCR reactions are usually not too sensitive to DNA concentration, which can vary from single molecules to some hundreds of nanograms without necessarily compromising the final result (Innis, Gelfand, and Sninsky 1999). Also, given that the extraction columns have limits in terms of the overall DNA quantity that they can retain, this DNA extraction approach to some extent already normalizes DNA concentration in the upper range, avoiding PCR inhibition. All PCR protocols were thus established using the µL as a measure of the DNA to load in the amplification reaction (arguably, genomic DNA concentrations were between few ng/µL and up to 500 ng/µL).

3.5. PCR reactions

3.5.1. Preliminary considerations on primer design

In order to amplify the gene of interest, it was necessary to develop and test specific PCR protocols capable of consistently amplifying the exon 1 portion of the Htt gene. As previously recalled, this gene portion is relatively hard to amplify due to the presence of the CAG repeat. This is especially true in mammals and in some reptilian species, where the CAG repeat is relatively long and another repeat (CCN, with N being mostly G o A) is present. The bonds formed by these high-GC content repeats require higher temperatures to be broken (Mamedov et al. 2008). A protocol that I had developed during my Master Thesis work was available and had already been tested on few mammalian species. This protocol relies on PCR primers designed at the very beginning and at the end of exon 1 coding sequence. These two regions, and the 5' end in particular, are extremely conserved in vertebrates, implying that the two primers designed here can potentially work for many species. Yet, as previously recalled, the redundancy of the genetic code allows for third-codon position mutations to be present even in conserved portions of the gene. One alternative is then to design and employ degenerate primers, harbouring synthetic nucleotides in third-codon positions capable of bonding to more than a single nucleotide (Linhart and Shamir 2007). However, this strategy can be expensive and not necessarily essential where the degree of concordance between species is high, as is the case of Htt exon 1. Moreover, the conserved portions of the gene are not only highly conserved but also apparently unique in vertebrate genomes (with the exception of the pseudogene found in a primate family, see Section 5), reducing the potential for mispairing of the primers to unintended targets in other genomic regions.

3.5.3. Primer design for the major taxonomic groups

Given the wide range of clades involved in the project, I first conducted several group-specific comparisons of the homologous regions based on the DNA sequences available from online databases in the same 5' and 3' ends of Htt Exon 1 of the available protocol (**Figure 10**). Based on consensus sequences, I then designed primers that could potentially allow for perfect pairing in several species belonging to multiple lineages.

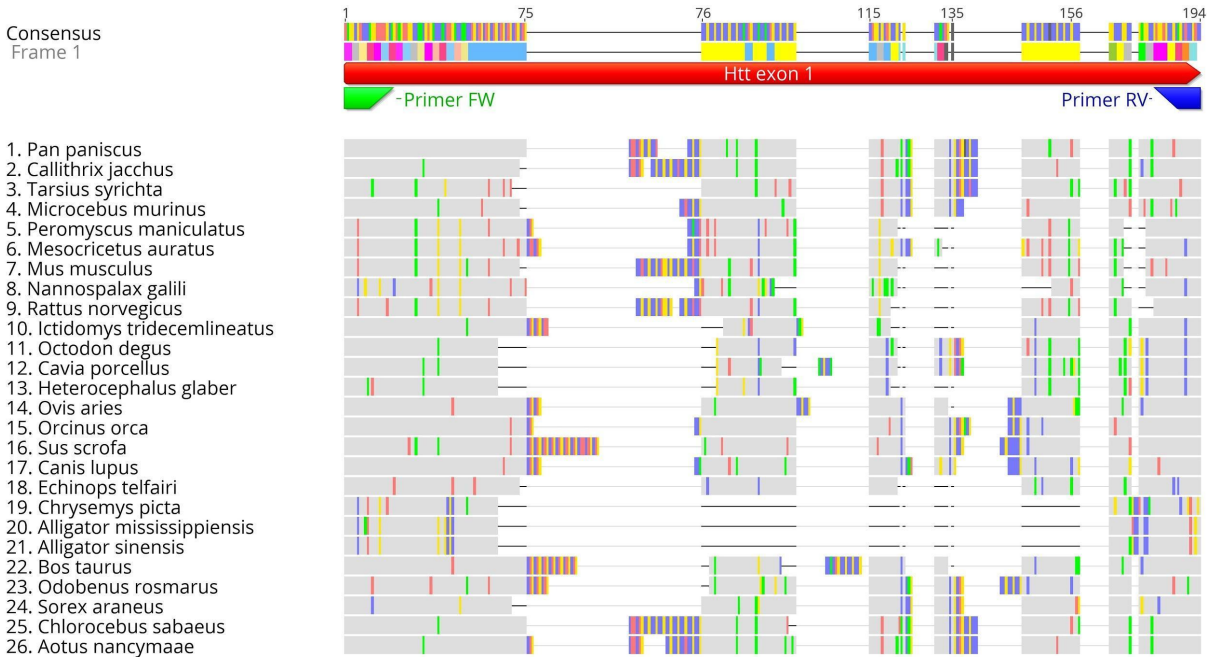

Figure 10: Illustrative multiple alignment for primer design. *Given a list of aligned orthologous sequences (1-26) the consensus sequence (top) allows to design primers that have the highest chances of priming in different species. Here the consensus sequence reports both the nucleotide sequence (first row) and the protein product (second row). The nucleotide sequence highlights the colours of the four nucleotides (C in blue, A in yellow, G in red, and T in green) while in the protein sequence the polyQ region is highlighted in light blue. The primer binding regions were designed immediately at the beginning of Htt exon 1 (annotated in red) for primer forward (annotated in green) and a the end of the exon for primer reverse (annotated in blue). Image drawn using Geneious v9.1.4 (Kearse et al. 2012).*

Overall, I produced 6 primer pairs that were used for PCR included in the final dataset, as detailed in **Table 4**. The original PCR protocol was adjusted for every taxonomic group and primer pairs based on tests on few DNA samples available in abundance.

Group	**Primer FW**	**Primer RV**
Human	5'-ttgctgtgtgaggcagaacc-3'	5'-gcagttaaaagaaccccgc-3'
NHPs	5'-tggctctgtgaggcagaaca-3'	5'-caacacagttaaaccccgc-3'
Mammals	5'-atggcgaccctggagaagctg-3'	5'-ggtcggtgcagcggctcct-3'
Birds	5'-atggccaccatggagaagctg-3'	5'-ggtctctggagcggctcct-3'
Reptiles	5'-atggccaccatggagaagctg-3'	5'-ggtctctggagcggctcct-3'
Fishes	5'-atggccaccatggagaaattg-3'	5'-atggccaccatggagaaattg-3'

Table 4: primer pairs employed to produce the DNA sequences. *Annealing temperature is highly primer dependent while extension time depends on the length of the target template region.*

3.5.4. Scaling-down of PCR reaction volumes

In setting up the conditions for a moderate-throughput (96-well) PCR pipeline, reaction volumes were scaled down to 15 µL also to reduce cost of reagents. The pipeline was proven to be robust and particularly helpful with DNA stored in plates, as it allowed loading the template DNA in the PCR buffer using multichannel pipettes (0.5-10 µL). The pipeline (see **Figure A2.5** in Section 5) was used for a preliminary screening of plates to determine which DNA samples amplified easily. Moreover, this setup allowed for a more rapid screening of PCR products, relying on large agarose gels loaded with multichannel pipettes into 96-well electrophoresis cells (Bio-Rad Subcell Model 96).

When the number of samples to be processed at once was low, or when some particular samples required to be treated independently or repeated, the PCR reactions were prepared in 0.2 µL PCR tubes.

3.5.5. PCR reaction reagents and conditions

Concentrations and volumes were principally determined by the scale down of the quantities already tested in the protocol previously developed, with few adjustments between taxonomic groups. When the combination of primers/protocol did not appear to work for a particular taxonomic group (e.g. in fishes), additional primer pairs and amplification conditions were tested.

Individual PCR protocols for the specific taxonomic groups have been produced and recorded for future reference. An example of PCR reaction reagent concentrations and volumes, used to amplify Htt exon 1 sequence in birds, is provided below.

Reaction protocol	
Buffer HF 10x	1.5 µL
Mg_2SO_4 50 mM	0.24 µL
dNTPs 10 mM	0.3 µL
Primer FW 10 uM	0.6 µL
Primer RV 10 uM	0.6 µL
Platinum Taq DNA Polymerase High Fidelity (Cat. no.11304-011)	0.15 µL
H_2O to volume (15 µL)	
Template DNA	3 µL

The amplification protocols were based on the Platinum Taq Polymerase datasheet, which in particular reported a working temperature of the polymerase of 68 °C. The thermocycler was set to use hot start[70] (to avoid mispriming during initial heating) and hot lid (to avoid condensation below the lid).

There were three main factors that appeared to be influencing PCR amplification success. First, the initial denaturation temperature, that for many protocols could not be reduced below 10' as this prevented amplification, potentially because the Htt locus is hardly accessible by the polymerase. A second key factor

[70] It should be noted that the Platinum Taq Polymerase employed is also "hot start".

was the annealing temperature, which often required adjustment to the primer pairs. However, as shown in **Table 5** most of primers worked with annealing temperature between 58-60 °C. The third factor was the extension time. Here, the principal need was that of adjusting it according to the predicted length of the amplicon, given that the manufacturer specifications reported an extension time of 1 kbp/minute, which however was to be increased considerably in order for the reaction to succeed. The necessity of increasing extension time was likely due to the problematic region of the genome to be amplified, which might lead the polymerase to stall multiple times along the DNA strand. The amplification protocols developed are reported below.

	Human	NHPs	Mammals	Birds	Reptiles	Fishes
Initial denaturation	10’ , 96 °C	10’ , 96 °C	10’ , 96 °C	10’ , 96 °C	10’ , 96 °C	10’ , 96 °C
36 cycles:						
Denaturation	45’’, 96 °C	45’’, 96 °C	45’’, 96 °C	45’’, 96 °C	45’’, 96 °C	45’’, 96 °C
Annealing	45’’, 60 °C	45’’, 60 °C	45’’, 58 °C	45’’, 58 °C	45’’, 58 °C	45’’, 55 °C
Extension	1.5’, 68 °C	1’, 68 °C	1’, 68 °C	1’, 68 °C	1’, 68 °C	30’’, 68 °C
Final extension	10’, 68 °C	10’, 68 °C	10’, 68 °C	10’, 68 °C	10’, 68 °C	10’, 68 °C

Table 5: PCR conditions of the six protocols developed. *Differences are in the annealing temperature and in the length of extension time.*

3.5.6. PCR results

PCR amplification was not always successful and this accounts for the 189 samples included in the final analysis (of 1307 initially collected). This was due to a variety of reasons, including the presence of too many sequence mismatches between the primers and the template, the poor quality of the template DNA and the presence of contaminants in the sample inhibiting PCR reaction. In some cases, also the low quantity of the DNA extracted might have acted as limiting factor. Importantly, samples tended to show high reproducibility, pointing to factors other than the operator variability affecting the success rate.

Amplicons with lengths in the range predicted from the sequences of close relatives present in public databases were interpreted as positive results and were further processed for detailed inspection and sequencing. True positive amplicons (i.e. verified by sequencing) were then indicated in each gel image and in the excel file of positive results. An example of this process is shown in **Figure 11**.

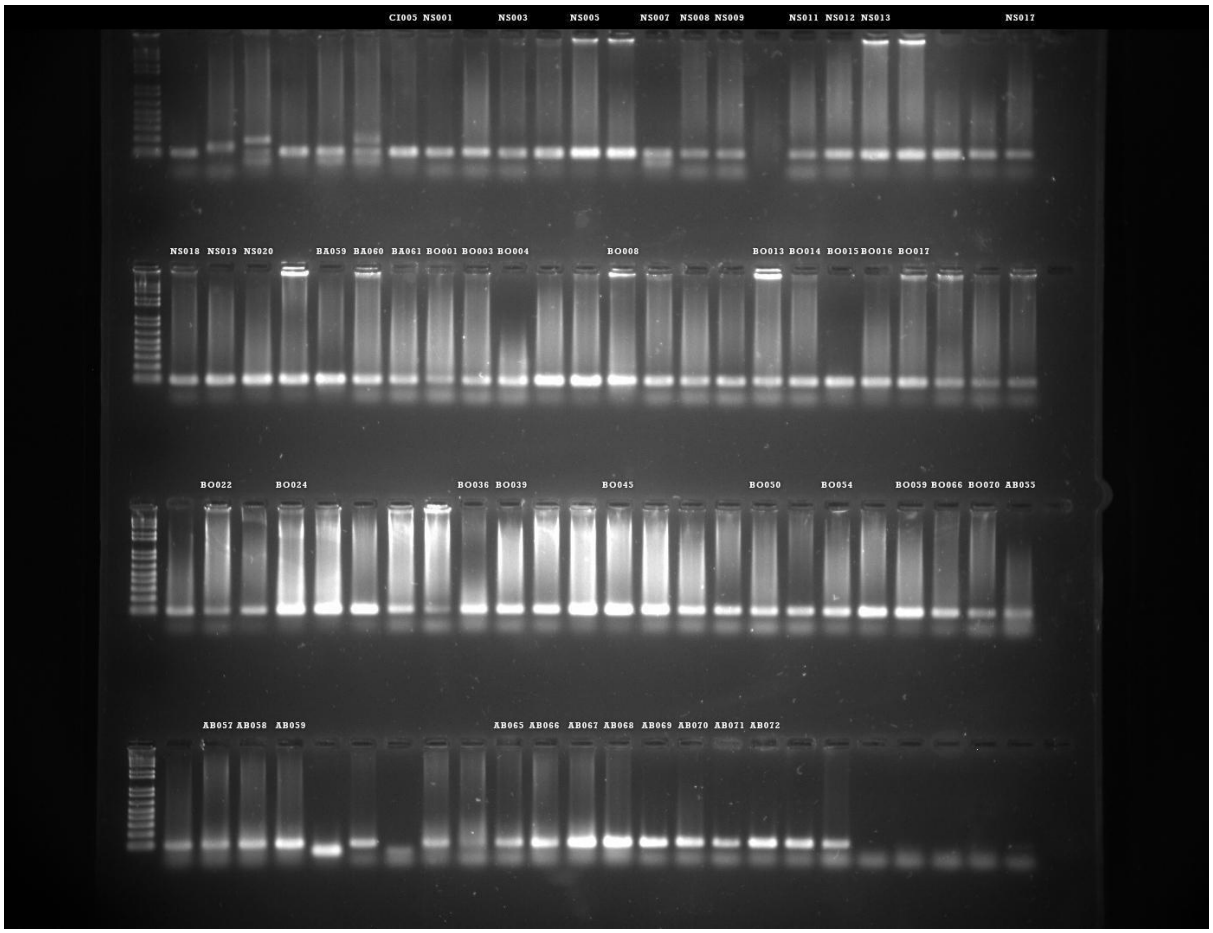

Figure 11: Illustrative gel picture for a 96-well gel electrophoresis. *Marker is 1 kb Plus DNA ladder. Expected amplicon size is 100 bp (birds) and is observed in most samples as a sharp and bright band right above primers and primer dimers. Residual genomic DNA is observed in many samples. Labels refer to unique internal IDs and are reported for samples included in the final dataset. Last five lanes are PCR negative controls.*

PCR products were stored at 4 °C to be subsequently sequenced or cloned, or at -20 °C for longer storage.

3.6. Sequencing of PCR products

3.6.1. Direct sequencing of PCR products

Of the 189 samples included in the final dataset, 158 were assessed by direct Sanger sequencing. 86 samples were sequenced once (either forward or reverse), 47 samples were sequenced twice (both forward and reverse with the exception of one sample that was sequenced forward twice) and 25 were sequenced three times. Overall, 189 sequences were obtained by direct sequencing and included in the final dataset.

3.6.2. Cloning of PCR products

Of the 189 samples included in the final dataset, 31 were cloned prior to Sanger sequencing. The moderate-throughput protocol for cloning of PCR products is shown in **Figure 15** in Section 5. The protocol requires cloning of PCR products in the vector of choice (in my case TOPO-TA pCR 4.0) according to datasheet

specifications (usually use 1:1 proportions) and the transformation of chemically competent bacterial cells (in my case TOP10 chemically competent cells) according to datasheet specifications using the above vector. Bacterial cells are then selected in the presence of Ampicillin on LB-agar plates that are incubated overnight at 37° C. A number of colonies is then picked and grown in LB+Ampicillin medium. A colony-PCR can be optionally performed to screen the clones for the insert using internal primers. Plasmids are then extracted with QIAprep Spin Miniprep Kit (Cat. No. 27104) or using a 96-well workstation (Qiagen QIAvac 96 Cat. No. 19504). Correct insert size can be further tested prior to sequencing by EcoRI digestion according to manufacturer's specifications. After digestion, canonical sequencing primers (e.g. M13, T7, T3) that match vector priming sites are then used for sequencing. Sequencing reactions were carried out in the same way as for direct sequencing, with sequencing primers added in Eppendorf tubes to the purified plasmids. Between 1-8 sequences for each sample were obtained and used in the final dataset (93 sequences overall, approximately 3 sequences per sample on average).

4. Sequencing results

The final dataset comprised 189 sequences from 108 species. A summary is provided in **Table 6**.

Group	Species	N. of individuals	N. sequences	Repeat n.
Mammals	*Callosciurus erythraeus*	1	5	7
	Gerbillus sp.	1	2	9
	Homo sapiens	1	2	12
	Hypsugo savii	1	2	9
	Hystrix cristata	3	6	4
	Macaca fuscata	1	8	11
	Mandrillus sphinx	4	10	11
	Manis sp.	1	6	4
	Martes foina	1	2	9
	Meles meles	1	2	10
	Miniopterus schreibersii	1	2	8
	Myocastor coypus	3	6	4
	Myotis emarginatus	3	6	9
	Myotis nattereri	3	6	9
	Papio hamadryas	3	10	12

	Pipistrellus kuhlii	1	4	6
	Plecotus macrobullaris	1	4	8
	Procavia capensis	4	7	15
	Rhinolophus ferrumequinum	3	6	6
	Rhinolophus hipposideros	1	2	6
	Rousettus aegyptiacus	1	2	5
	Rupicapra rupicapra	5	8	12
	Saguinus imperator	1	2	9
	Saguinus oedipus	1	3	13
	Sciurus carolinensis	7	16	8
	Sciurus vulgaris	5	7	7
	Stenella coeruleoalba	3	4	11
	Sylvilagus floridanus	5	6	5
	Tachyglossus aculeatus	2	4	5
	Tursiops truncatus	2	10	11
Birds	*Accipiter nisus*	3	8	4
	Aegithalos caudatus	1	1	4
	Anas platyrhynchos	1	1	4
	Apus apus	3	5	4
	Aquila chrysaetos	1	3	4
	Asio otus	3	5	4
	Athene noctua	3	6	4
	Burhinus oedicnemus	2	4	4
	Buteo buteo	1	1	4
	Calidris pugnax	1	2	4
	Caprimulgus europaeus	1	2	4
	Certhia brachydactyla	1	2	4
	Certhia familiaris	1	2	4
	Chalcopsitta duivenbodei	1	1	4

	Ciconia ciconia	1	2	4
	Cinclus cinclus	1	2	4
	Clamator glandarius	1	1	4
	Columba palumbus	1	1	4
	Corvus cornix	3	5	4
	Corvus monedula	1	2	4
	Corvus scapularis	1	2	4
	Coturnix coturnix	1	1	4
	Cuculus canorus	1	3	4
	Dacelo novaeguineae	1	2	4
	Dromas ardeola	3	4	4
	Emberiza citrinella	1	1	4
	Falco subbuteo	1	3	4
	Falco tinnunculus	2	6	4
	Gallus gallus	1	2	4
	Garrulus glandarius	2	2	4
	Gavia arctica	1	1	4
	Geronticus calvus	1	2	4
	Gypaetus barbatus	1	3	4
	Gyps fulvus	1	3	4
	Hirundo rustica	1	1	4
	Lanius collurio	1	1	4
	Otus scops	1	3	4
	Passer domesticus	1	1	4
	Passer montanus	1	2	4
	Phalacrocorax carbo	2	6	4
	Phasianus colchicus	1	3	4
	Phoenicopterus chilensis	1	2	4
	Phoenicopterus ruber	1	3	4

	Pica pica	1	1	4
	Picus viridis	2	3	4
	Platalea leucorodia	11	23	4
	Pluvialis squatarola	1	3	4
	Podiceps cristatus	1	2	4
	Rallus aquaticus	1	3	4
	Rhea americana	1	2	4
	Saxicola torquata	1	1	4
	Strix aluco	1	3	4
	Sturnus vulgaris	7	13	4
	Sylvia atricapilla	1	1	4
	Sylvia communis	1	1	4
	Tachybaptus ruficollis	1	2	4
	Tauraco persa	1	2	4
	Threskiornis aethiopicus	1	2	4
	Tragopan temminckii	1	2	4
	Tringa erythropus	1	3	4
	Turdus merula	4	8	4
Reptiles	*Algyroides nigropunctatus*	1	1	4
	Alluaudina bellyi	1	1	4
	Astrochelys radiata	1	1	4
	Coronella austriaca	1	1	4
	Hemidactylus turcicus	2	2	5
	Hierophis viridiflavus	1	1	4
	Leioheterodon modestus	1	1	4
	Lycodryas granuliceps	1	3	4
	Madascincus melanopleura	1	1	4
	Phelsuma serraticauda	1	2	6
	Podarcis siculus	1	2	4

	Podarcis tiliguerta	1	1	4
	Tarentola mauritanica	5	6	4
Fishes	*Capros aper*	1	2	4
	Diplodus sargus	1	2	4
	Thalassoma pavo	1	2	4
	Trachinotus ovatus	1	5	4

Table 6: List of all species included in the final dataset*. The number of sequences refers to the sum of all sequences available for all individuals. The number of repeats was calculated from the sequence reads and includes both CAG and CAA codons.*

These data are currently under analysis. Specifically, they are used in the attempt to demonstrate that the CAG tract has been under strong purifying selection throughout vertebrate evolution. This hypothesis is suggested by the general absence of non-synonymous (i.e. encoding for a different amino acid) nucleotide substitutions in the vast majority of species sequenced and by the simultaneous presence of several synonymous (CAA) substitutions in several species. All the work described in sections 2 to 3 was performed with the contribution of Dr. Michela Pacifico.

5. A focus on primates

Genetic studies in human populations show that the distribution of the CAG repeat length in healthy human populations is biased toward longer CAG repeats with a mean of about 18 repeats and a modal length of 17 repeats in most populations (Kremer et al. 1994; Rubinsztein et al. 1994; Bates, Harper, and Jones 2002) (see also Introduction). The difference between the mean and the modal length arises from the skewness of the distribution toward longer alleles. About 80% of the CAG repeats span between 15 and 20 repetitions and another 17% between 21 and 26. Accordingly, intermediate alleles (IAs) between 27 and 35 CAG repeats are extremely rare (i.e. <3%), although one study has reported IAs incidence of 1 in 17 individuals in British Columbia general population (Semaka 2016). The bias toward longer alleles (i.e. elongation favoured over shortening) appears to shed light on the origin and permanence of Htt CAG repeats and can be explained by invoking a beneficial role of repeats below disease threshold (Lo Sardo et al. 2012; Zuccato and Cattaneo 2014) or by random mutation favouring those alleles (Rubinsztein et al. 1994). Either way, it is predicted that, unless hidden constraints are present, species other than humans can as well have developed IAs during the course of their evolutionary history. Notably, it has been already established that pigs bear CAG repeats in the human range (Matsuyama et al. 2000). However, naturally occurring Non-Human Primate (NHP) species with CAG repeats in the pathological range, or at least in the range of intermediate HD alleles, have never been so far identified.

The importance of the availability of a NHP model for the research on the disease is testified by the production of NHP laboratory models of HD as early as 1990, with HD-like brain lesions induced in baboons (Hantraye et al. 1990). Almost twenty years later, the first transgenic rhesus macaque (*Macaca mulatta*) was also produced, and appeared to show several histological features of HD, such as nuclear inclusions and neuropil aggregates, as well as clinical features of HD, including dystonia and chorea (Yang et al. 2008). Naturally occurring HD alleles in NHPs could represent an important milestone in the research on the disease as a more natural laboratory disease model could be developed. For this reason, a specific focus on primates was conducted over several years of the project, and a detailed account of this endeavour is reported in this section.

5.1. Preliminary results on NHP samples obtained from the Biopark of Rome

The first study on primate variability at Htt exon 1 CAG *locus* was conducted in 1994 on 48 samples belonging to 10 species and including 25 chimpanzees (*Pan troglodytes*), 3 gorillas (*Gorilla gorilla*), 3 baboons (*Papio sp.*), 3 orangutans (*Pongo pygmaeus*), 2 crab-eating macaques (*Macaca fascicularis*), 2 rhesus macaques (*Macaca mulatta*), 2 marmosets (*Callithrix sp.*), 1 olive baboon (*Papio anubis*), 1 gibbon (*Hylobates sp.*) and 2 pooled samples, one from six male and another from six female talapoins (*Miopithecus sp.*) (Rubinsztein et al. 1994). Repeat lengths were assessed by PCR and acrylamide gel electrophoresis. The outcome of this analysis suggested that NHPs lay at the bottom of human allelic length distribution, with a CAG range spanning from 7 to 16 repeats. On this basis, Rubinsztein and colleagues posited that humans inherited a shorter ancestral allele from the common ancestor of all primates. Those ancestral alleles independently expanded in humans to the current healthy length range (9-35 CAG repeats). By contrast HD

alleles are of much more recent origin and in most cases originate by mutation from IAs (see also Introduction).

The laboratory decided to set up collaborations with various groups to try to expand the knowledge on Htt exon 1 CAG variability in NHPs, which is – even now – incomplete, difficult and poorly reliable for reasons that will be explained later. In 2013, 68 buccal swabs samples belonging to 16 different species sampled at the Biopark of Rome by a collaborator, Dr. Cristina Martinez-Labarga, were delivered to the laboratory. This added to a previous shipment of 17 samples from NHP species, for a total of 85 samples. The second shipment included also 18 samples of Japanese macaque (*Macaca fuscata*). In a first step the samples were analyzed at Besta Institute (Milan, Italy) by capillary electrophoresis, as performed in Rubinsztein et al (1994). Capillary electrophoresis relies on the amplification of the *locus* of interest with Fluorescein amidite (FAM) labelled primers. Amplicon length is subsequently assessed by running the amplicons on a capillary and recording the fluorescence when they pass through a photocathode (Kimpton et al. 1993). The length of the repeat is determined by comparing the size of the amplification fragment to a ladder of known length (Kimpton et al. 1993) (**Figure 12**).

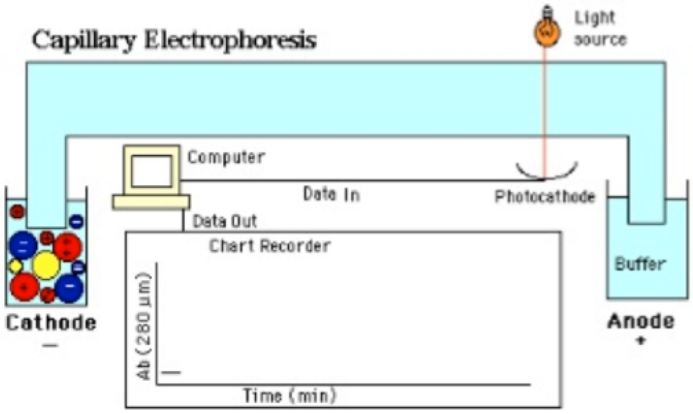

Figure 12: The principle of capillary electrophoresis. *During capillary electrophoresis the electron gradient allows migration of PCR amplicons through the capillary. When fluoresceinated amplicons reach the photocathode their fluorescence is recorded. A measure of time elapsed since the start of the run allows to determine the size of the amplicon with a high level of precision.*

The protocol employed, which is the same used for HD diagnosis in the hospital (Warner, Barron, and Brock 1993; Gellera et al. 1996), is provided below:

Reaction mix

$MgCl_2$ 1.5 mM

DMSO 1%

dNTPs 250 µM

FW primer -5' FAM labelled (sequence: CCTTCGAGTCCCTCAAGTCCTTC) 10 pmoles

RV primer (sequence: GCGGCGGTGGCGGCTGTTG) 10 pmoles
Taq Gold DNA Polymerase
100 ng di DNA
Final volume: 20 µl

PCR reaction conditions

35 cycles:

Denaturation	1' at 94 °C
Annealing	1' at 65 °C
Extension	2' at 72 °C

The amplicons were then run on a ABI PRISM 3130 Genetic Analyzer (Applied Biosystems, Foster City, CA, USA) where LIZ labeled ladder (GeneScan 500 LIZ) was used to determine tract length and CAG number. Repeat length varied considerably in the sample, with shortest repeats (7) represented in *Hylobates*, *Saguinus* and *Lemur* genera (**Table 7**).

Species	**N. of chr.**	**N. of alleles**	**N. of genotypes**	**N. of repeats**
Nycticebus pygmaeus	2	1	10-10	10
Eulemur fulvus albifrons	2	1	10-10	10
Lemur Variegatus	2	2	10-11	10, 11
Lemus catta	8	6	8-8, 7-9, 12-16, 10-10	7, 8, 9, 10,12,16
Sanguinus imperator	2	2	7-9	7, 9
Cebus apella	50	3	9-9 (3), 9-10 (11), 9-13 (4), 10-10 (4), 10-13 (2), 13-13	9, 10, 13
Macaca fuscata	38	10	9-9, 9-12, 9-14, 9-16, 9-17 (3), 9-19, 10-10, 14-15, 14-17, 16-17 (2), 17-17, 17-19, 17-22, 17-26 (2), 26-26	9, 10, 12, 14, 15, 16, 17, 19, 22, 27
Macaca mulatta	2	2	10-17	10, 17
Papio hamadryas	4	2	9-10, 10-10	9, 10
Cercopithecus sp.	2	1	8-8	8
Cercocebus sp.	6	4	9-9, 17-17, 13-14	9, 13, 14, 17
Mandrillus sphinx	14	4	9-10 (4), 9-11, 10-10, 10-16	9, 10, 11, 16

Hylobates lar	2	1	7-7	7
Pongo sp.	10	1	8-8 (5)	8
Gorilla gorilla	16	5	8-9, 9-9 (4), 9-10, 10-10, 16-17	8, 9, 10, 16, 17
Pan troglodytes	10	5	8-9, 9-10, 9-11, 9-12, 10-11	8, 9, 10, 11, 12

Table 7: Summary of the Htt exon 1 CAG repeat lengths measured in NHPs from the Biopark of Rome using capillary electrophoresis. *Number of chromosomes is the count of chromosomes (two per individual) analyzed. In the genotype column, numbers between brackets indicate the count of individuals with the same genotype.*

Surprisingly, in these first analyses, CAG lengths from 9 to 27 repeats were observed in the Japanese macaque (*Macaca fuscata*), a situation extremely similar to that of humans. Specifically, 3 out of 19 samples belonging to one primate species pointed to representatives of this species harbouring exceedingly long CAG repeats (26 and 27 repeats). Principal Component Analysis (PCA) was consistent with the Japanese macaque representing an outlier among all NHP species analyzed (**Figure 13**). This result was interesting and unexpected but, unfortunately, it turned out to be a technical artifact, as I was able to ascertain with a closer inspection only few years later. This finding is detailed in the following section.

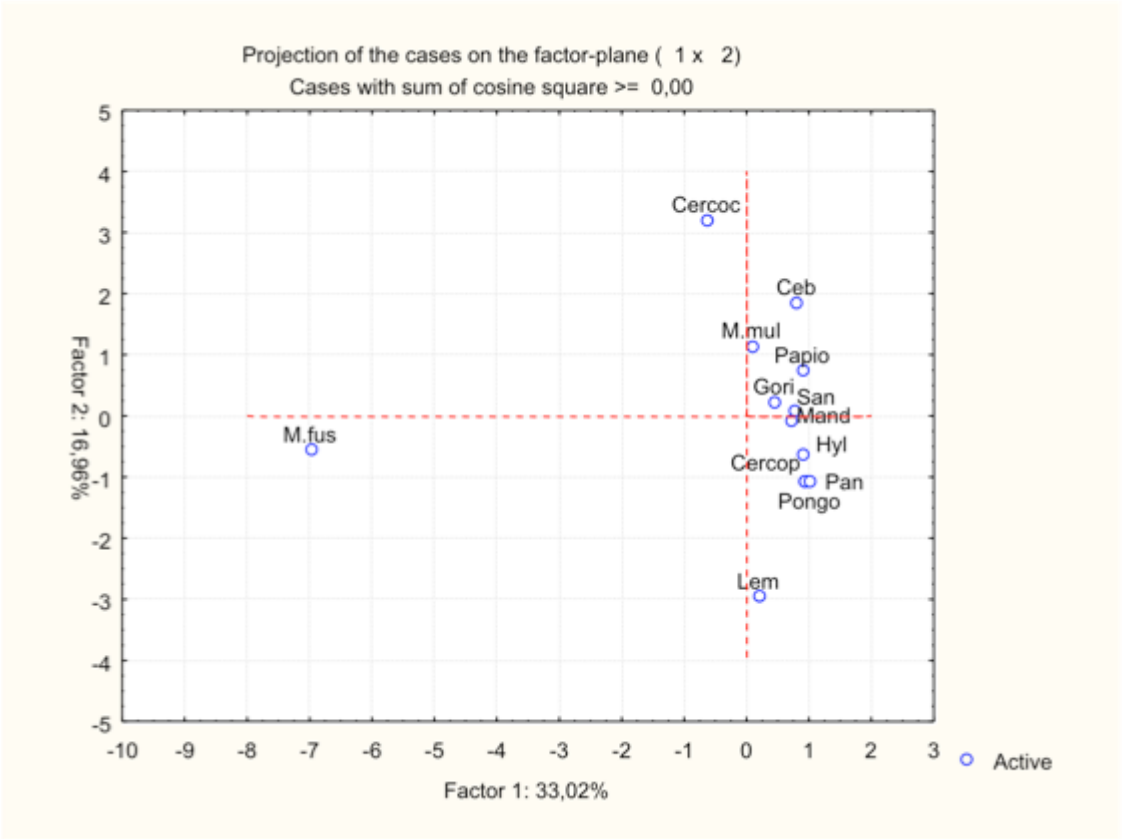

Figure 13: PCA of the repeat length in 85 primates belonging to 16 species. *For Factor 1, the M. fuscata clearly represented an outlier among all NHP species analyzed.*

5.2. The first visit to the Kyoto University Primate Research Institute

At the same time, there has been a growing interest in the laboratory on the possibility to expand the knowledge of the CAG size in a large population of NHP. We therefore tried to set up collaborations to obtain NHP DNA from overseas. However, all attempts made in the different years and through several foreign entities and Italian institutional bodies, failed due to the complex regulatory issues associated with the shipment of biological samples from NHP. To pursue the project we therefore established a collaboration with a Japanese laboratory at the Kyoto University Primate Research Institute (PRI), which was interested at hosting us for the development of the project and the on site preparation and processing of the NHP DNA. In short, the collaboration begun in 2015 and continued until 2017. To support the work on the assessment of Htt CAG in NHPs - which was unfunded – we also worked on a proposal to which I have largely contributed and entitled *"Intermediate allele identification in non-human primates through Htt Exon1 sequencing"* which was ultimately submitted and accepted for funding by the Cure Huntington's Disease foundation (CHDI, NY). However the work included in the proposal was not performed, the funding assigned has not been used and discontinued and the two stay in Japan did not lead to conclusive results on the CAG size in the NHP we planned to test. Yet, further assessment of the NHP DNA samples available in the Milano laboratory led to a serendipitous discovery as detailed in the next section.

The first visit to PRI in Inuyama was in mid 2015, right before the start of my graduate studies. PRI was chosen as Japanese macaques have been studied there for long time also by my guest Dr. Hiroo Imai (Associate Professor of Molecular and Cell Biology at Kyoto University). Our interest there was in the Japanese macaque (M. fuscata), and in other NHP, as we wanted to verify the putative long alleles found in the colony held in captivity at the Biopark of Rome. The general aim of this visit (3 weeks) to the PRI was to sequence as many Htt exon 1 orthologs in Japanese macaques as possible, in order to confirm the presence of high levels of variability, especially in terms of long alleles. However, only few days before my departure, Sanger sequencing performed through GATC (Konstanz, Germany) on one of the sample held in the laboratory of Milan provided results in conflict with the early report from the Besta Institute. In particular, the individual estimated as having a 17-26 CAG genotype, showed instead a 8-9 CAG genotype. In order to understand which result was correct, on August 2015 I moved to the PRI, where a fully equipped laboratory of molecular biology was available at the research center along with a collection of 503 DNA samples (**Figure 14**) that was put together on purpose from different Japanese macaque individuals.

	Population Name			Prefecture			N
Mafu	Numata	NMT	沼田	Gunma	群馬	Wild	20
	Annaka		安中	Gunma	群馬	Wild	1
	Takahama	THM	高浜	Fukui	福井	PRI	32
	Jigokudani	JGD	地獄谷	Nagano	長野	NBR	40
	Hagachi	HGC	波勝	Shizuoka	静岡	NBR	18
	Okazaki	OKZ	岡崎	Aichi	愛知	Wild	4
	Shiga	SHG	滋賀	Shiga	滋賀	NBR	37
	Arashiyama	ARY	嵐山	Kyoto	京都	PRI	31
	Minoo	MNO	箕面	Osaka	大阪	NBR	41
	Kii	KII	紀伊	Wakayama	和歌山	NBR	40
	Wakasa	WKS	若桜	Tottori	鳥取	PRI	41
	Shodoshima	SHD	小豆島	Kagawa	香川	NBR	17
	Kami	KAM	香美	Kochi	高知	Wild	12
	Takaoka	TKO	高岡	Kochi	高知	Wild	29
	Imabari		今治	Ehime	愛媛	Wild	1
	Koshima	KOS	幸島	Miyazaki	宮崎	Wild	83
Mamu	India		インド				28
	China		中国				28
	Total						503

Figure 14: Japanese macaque DNA samples available at PRI. *Japanese macaques are endemic in Japan and the sample was highly representative of the entire country.*

Assessment of the Htt Exon 1 sequence was carried out with a protocol specifically established by me for *M. fuscata* that included PCR amplification of the Region of Interest (ROI), cloning of the PCR product and Sanger sequencing (**Figure 2.5**). In this protocol, PCR amplification relies on two primers designed against the 5'-UTR of the Htt gene (5'-tggctctgtgaggcagaaca-3') and in the Intron 1 (5'-caacacagttaaaccccgc-3'), respectively. The PCR was carried out as follows (PCR volumes have been scaled down to 15 µL in order to avoid waste of reagents):

<u>Reaction mix</u>

Buffer HF 10X	1.5 µL
Mg_2SO_4 50 mM	0.4 µL
dNTPs 2.5 mM	1.2 µL
Primer FW 10 uM	0.6 µL
Primer RV 10 uM	0.6 µL
Platinum Taq DNA Polymerase High Fidelity (Cat. no.11304-011)	0.06 µL
H_2O to volume (15 µL)	
Template DNA	5 µL*

*concentrations in the order of 5 ng/µL

<u>PCR conditions</u>

Initial denaturation	3' at 96 °C
35 cycles:	
Denaturation	1' at 96 °C
Annealing	1' at 60 °C

Extension	1.5’ at 68 °C
Final extension	10’ at 68 °C

PCR was followed by moderate-throughput cloning and sequencing using 12 and 96 well plates (**Figure 15**) in a pCR4-TOPO Vector (Thermofisher Scientific) using One Shot TOP10 Chemically Competent *E. coli* cells (Thermofisher Scientific). TOPO-TA cloning v4.0 is a very efficient vector for rapid cloning of PCR products. The v4.0 employs a gene that is lethal to *E. coli* whenever the plasmid self-ligates in the absence of a PCR amplicon (Bernard and Couturier 1992; Bernard et al. 1993, 1994).

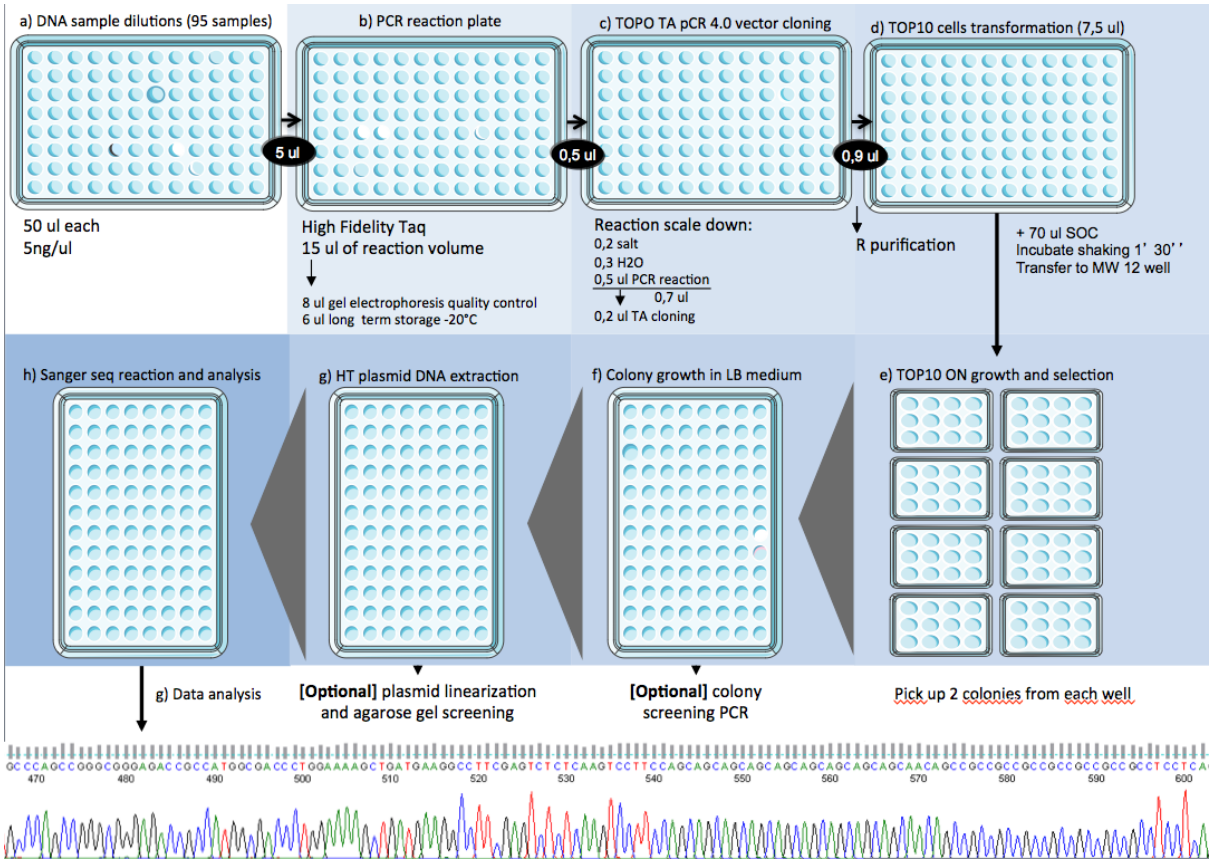

Figure 15: Moderate-throughput protocol for assessing Htt exon 1 sequence applied at the PRI on Japanese macaques DNA samples. *DNA dilutions were first prepared in a 96 well plate that was subsequently employed for PCR amplification. PCR products were then cloned using TOPO-TA vector 4.0. Competent cells were transformed using this vector and transferred to 12 well plates for bacterial growth. Colonies were isolated in 96 well plates were plasmid DNA extraction was carried out. Optionally, a PCR on growing bacterial colonies or plasmid linearization in the next step could be carried out to screen for correct insert size. Extracted TOPO-TA plasmids could be finally sequenced by Sanger according to manufacturer guidelines.*

Out of 503 samples available for testing, I performed a total of 661 PCR reactions, of which 245 were successful.

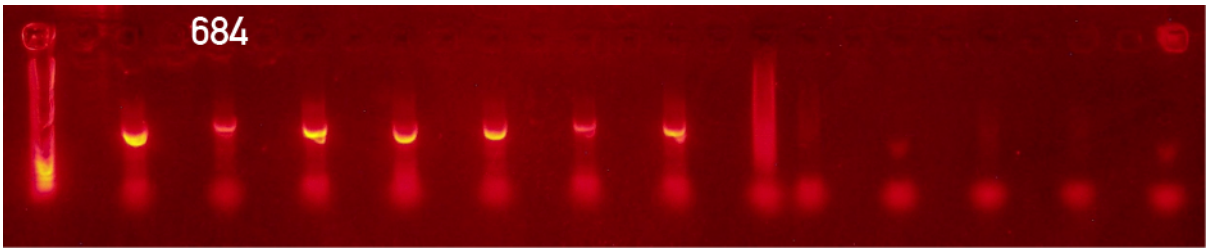

Figure 16: PCR on Japanese macaque DNA. *A macaque sample from Wakasa population in the Tottori prefecture (ID:684) was included in the multiple alignment (see Section 4).*

Of the 245 successful PCR reactions, I was able to clone and sequence 206 bacterial clones. As several bacterial clones belonged to the same PCR reaction and individual, I was ultimately able to assess with confidence 83 Htt CAG alleles (**Appendix 4**). As suggested by the preliminary sequencing of the suspicious sample, all samples sequenced in Japan turned out to have relatively short repeats. In particular, the longest alleles only reached 12 repeats, while the shortest was 6 repeats, with a modal length of 9 repeats (**Figure 17**). Repeat length distribution was right-skewed, similar to what is observed in humans, but with a shorter modal length.

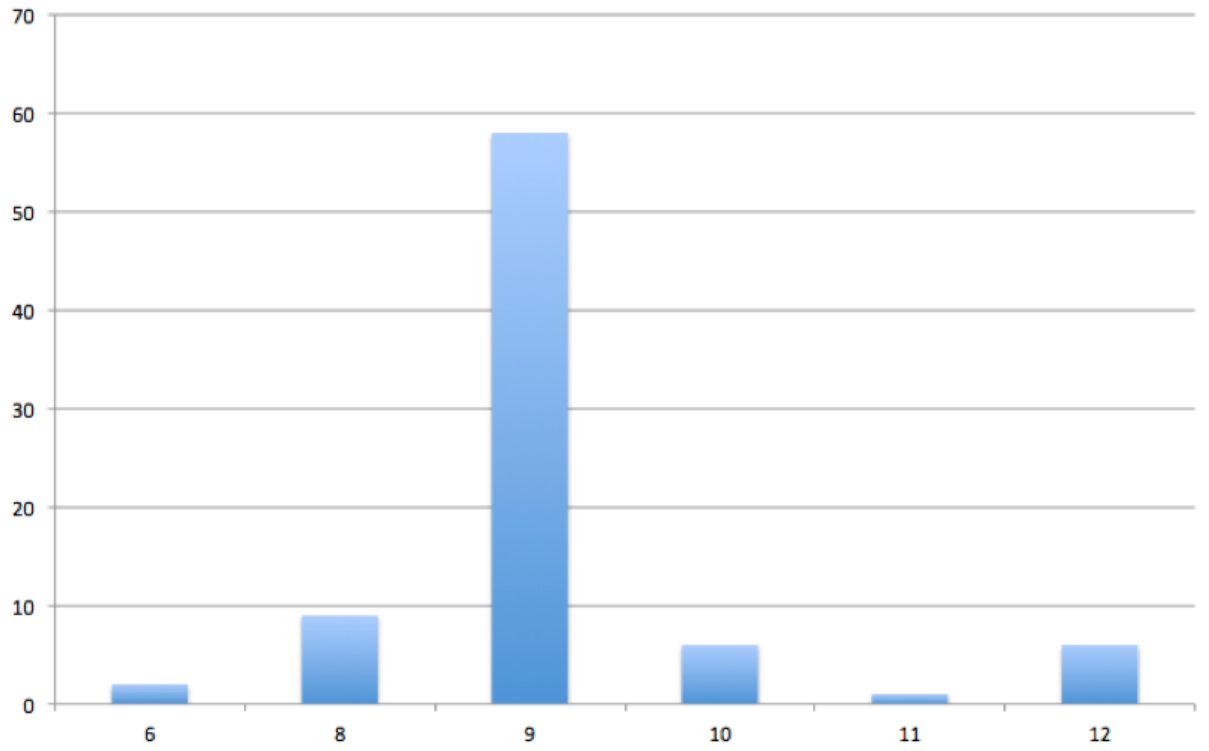

Figure 17: Htt exon 1 CAG length distribution in 83 alleles from Japanese macaques analyzed at PRI. *The distribution appears right-skewed, similar to what is observed in humans. However the modal length is only 9 repeats, while in humans the modal length is 17. Alleles with 8 and 9 repeats were also found in the sample from the colony breeding at the Biopark of Rome.*

We then realized that the explanation for the discrepancy observed on the *M. fuscata* samples tested in Japan and those tested in Milano likely relies in the human primers used in the Milano Hospital protocol to amplify Htt CAG which when used in other species may cause primer mispairing (see also Section 3.5). Our conclusion adds uncertainties also to the data reported by Rubinsztein et al. (1994) in which the 16 CAG detected in some primates have been measured by electrophoresis on acrylamide gels[71]. Yet, the focus on primates was not totally unfruitful, as it pushed us to carefully interrogate the samples available in our hands, ultimately leading to a serendipitous and insightful discovery, which is detailed in the next section.

5.3. Htt pseudogene in Callitrichidae

Early after my return from Japan and after the start of my Ph.D. project, we started to sequence Htt exon 1 orthologs in the primate DNA collection held at the laboratory. I rapidly identified an intriguing amplicon using DNA sample from a Emperor tamarin (*Saguinus imperator*) individual, which was yielding a PCR product longer than the average for similar samples (about 300 bp over the ~230 bp usually found in non-human primates). At that time I hypothesized a PCR artifact and I placed the result aside. However, one month later I received sequencing results from another sequencing reaction on a marmoset (*Callithrix jacchus*) from Labarga the Biopark of Rome. This was found different in many ways from the marmoset Htt sequence already available from public database and included a frameshift mutation (**Figure 18**).

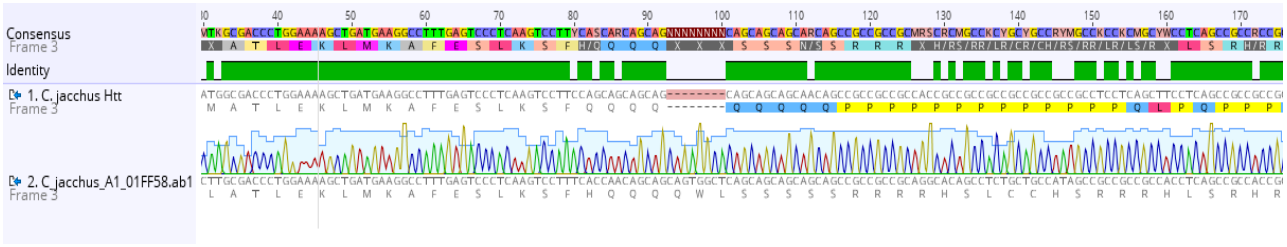

Figure 18: Marmoset Htt vs unidentified PCR amplification product from Marmoset. *The sequence result from the marmoset (bottom row) shows a CAG repeat interrupted by TGGCT after the first four codons (CAACAGCAGCAG). This causes a frameshift mutation.*

The sequencing results from the Emperor tamarin were also quite compelling in that, once again, Htt exon 1 was familiar in the sequence, but it highly differed with respect to that of other primates and/or mammals and included a stop codon in one of the two reads (**Figure 19**).

[71] This early approach requires a lot of subjective assessment by the operator and may be regarded as less reliable than Sanger sequencing.

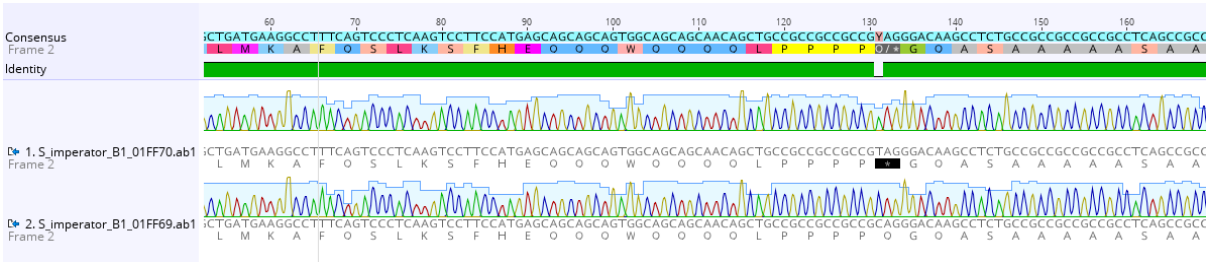

Figure 19: Marmoset Htt vs unidentified PCR amplification product from Emperor tamarin. *In the emperor tamarin Sanger reads (last two rows), the first two CAG codons in the CAG repeat are replaced by a CAT and a GAG, potentially resulted from two nucleotide substitutions per codon (G → T and C → G, respectively). Furthermore, they show a TGG substitution in the CAG repeat. The first read also shows a stop codon mutation (indicated by the *), potentially as a result of a PCR artifact causing a C → T mutation.*

Moreover, two alleles of slightly different length (9 bp difference between the two) were detected in the Emperor tamarin (**Figure 20**). The difference in length was due to an extra GCCGCCGCC in one of the two alleles, suggestive of a slippage mutation originating the secondary allele. However, as the secondary allele could not be replicated independently, it could also be the product of slippage during PCR amplification.

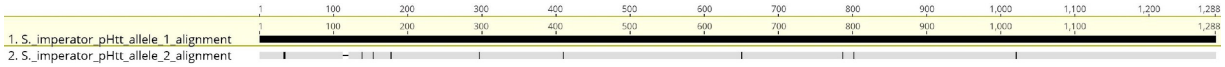

Figure 20: Potential alleles detected in the Emperor tamarin. *If putative allele 1 is taken as reference, allele 2 shows a GCCGCCGCC at the 5' end. Several point substitutions are also observed (98.4% identity between the two sequences), suggestive of the presence of two distinct alleles rather than of a PCR artifact.*

I then realized that these results could be explained by the presence of a pseudogene copy of Htt in these primates. This hypothesis was corroborated by the fact that *S. imperator* and *C. jacchus* are relatively close primates species belonging to the same family (Callitrichidae). I therefore conducted a bioinformatic search for the "alternative" Htt found in *C. jacchus* in the raw read archive of its WGS project and I ultimately spotted the same sequence but on a different chromosome with respect to Htt, spanning exons 1-9. The procedure is described in **Figure 21**.

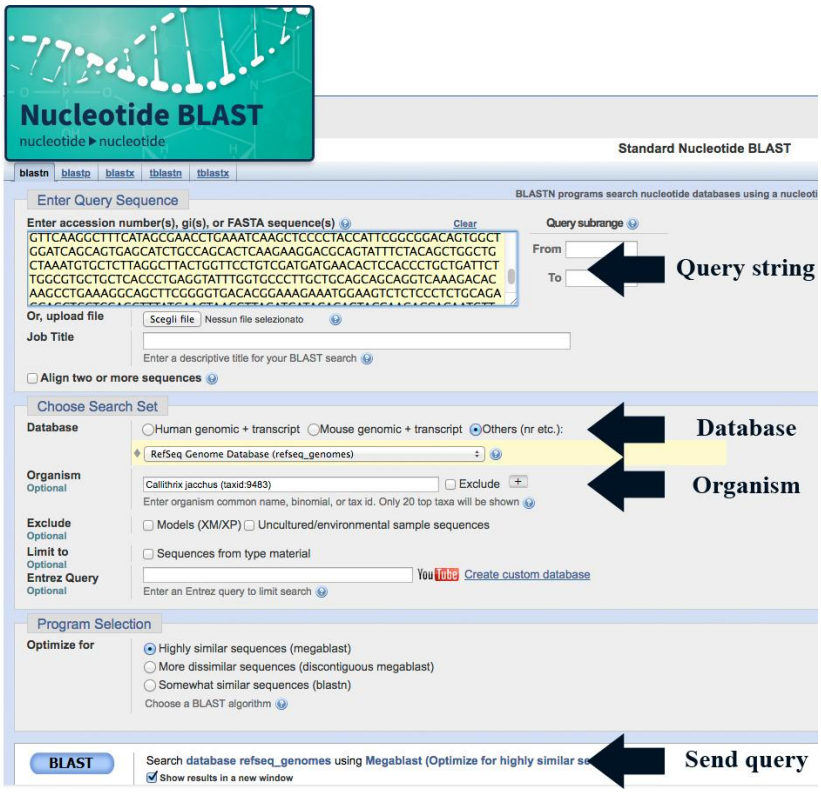

Figure 21: The pipeline to enquiry the NCBI database. *The sequence to be matched is copied in to the query form of BLAST software*[72]*. The database is set to "Others" and "refseq_genomes" is selected in the dropdown menu. The organism name (Callithrix jacchus, taxid:9483) is entered in the "Organism" field.*

The result of the query is shown in **Figure 22** and the pseudogene is represented in **Figure 23**.

[72] "NCBI - BLAST". https://blast.ncbi.nlm.nih.gov/Blast.cgi?PROGRAM=blastn&PAGE_TYPE=BlastSearch&LINK_LOC=blasthome

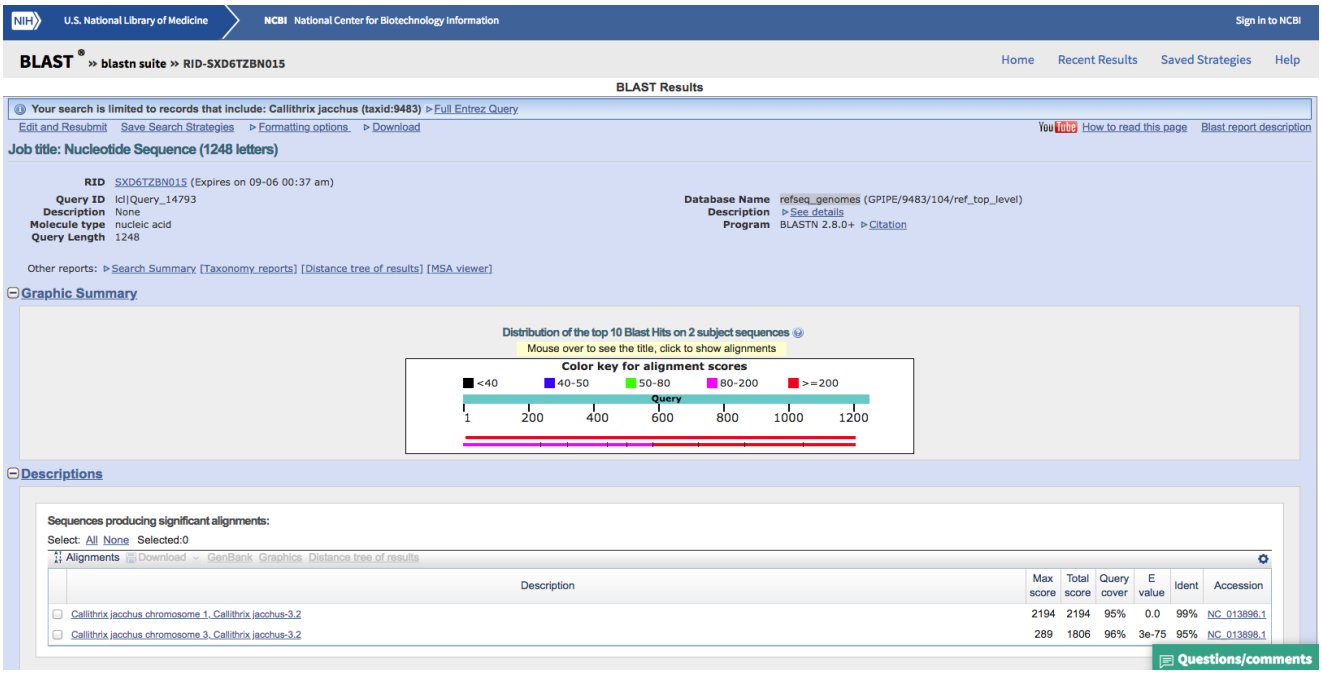

Figure 22: BLAST of the Marmoset unidentified amplification product. *As it can be seen in the coloured box, two very high score (>80 and >200) matches are found, one of the two has 8 breakpoints (vertical black lines) and indeed it corresponds to the genomic sequence of Htt in the Marmoset (the interruptions representing the intronic regions). The first sequence lays on chromosome 1, while the second sequence lays on chromosome 3. The first sequence has 1194/1197(99%) identity, strongly suggestive of belonging to the exact same genome region. The second sequence has about 90% identity depending on the exon assessed, suggestive of its paralog nature.*

Figure 23: Schematic representation of the pseudogene. *The first 9 exons composing HTT mRNA (first black stripe, exons highlighted by the red boxes above) are found in the pseudogene sequence (second black stripe). The main differences (indels) are observed in the first exon.*

I could then design new primers based on this sequence:

<u>Forward primers</u>

5'-GCGACCCTGGAAAAGCTGAT-3'

5'-CCTGGAAAAGCTGATGAAGGC-3'

<u>Reverse primers</u>

5'-TGGTCAGGGCTTGCAGAAG-3'

5'-GACTCATCCTTAGCCTTGGTG-3'

5'-TACTCCCACTACGGCTTCGG-3'

The expected amplicon in *C. jacchus* with these primers was about 1,100 bp. Prof. Nicola Saino then provided more samples from other species of Callitrichidae and I performed an optimized PCR reaction on *C. Jacchus* (1 individual), *S. Imperator* (1 individual) and *S. oedipus* (2 individuals) using the following protocol:

Reaction protocol	
Buffer HF 10x	1.5 µL
Mg_2SO_4 50 mM	0.24 µL
dNTPs 10 mM	0.3 µL
Primer FW 10 uM	0.5 µL
Primer RV 10 uM	0.5 µL
Platinum Taq DNA Polymerase High Fidelity (Cat. no.11304-011)	0.15 µL
H_2O to volume (15 µL)	
Template DNA	60 ng

PCR conditions	
Initial denaturation	10’ at 96 °C
36 cycles:	
Denaturation	45’’ at 96 °C
Annealing	30’’ at 58 °C
Extension	1.5’ at 68 °C
Final extension	10’ at 68 °C

According to PCR (**Figure 24** and **Figure 25**) and sequencing results, primers worked in the following combinations:

FW1&RV1 → *C. jacchus*, *S. imperator*;

FW1&RV2 → *S. oedipus;*

FW1&RV3 → *S. oedipus*, *C. jacchus;*

FW2&RV2 → *C. jacchus;*

FW2&RV3 → *S. imperator*, *S. oedipus.*

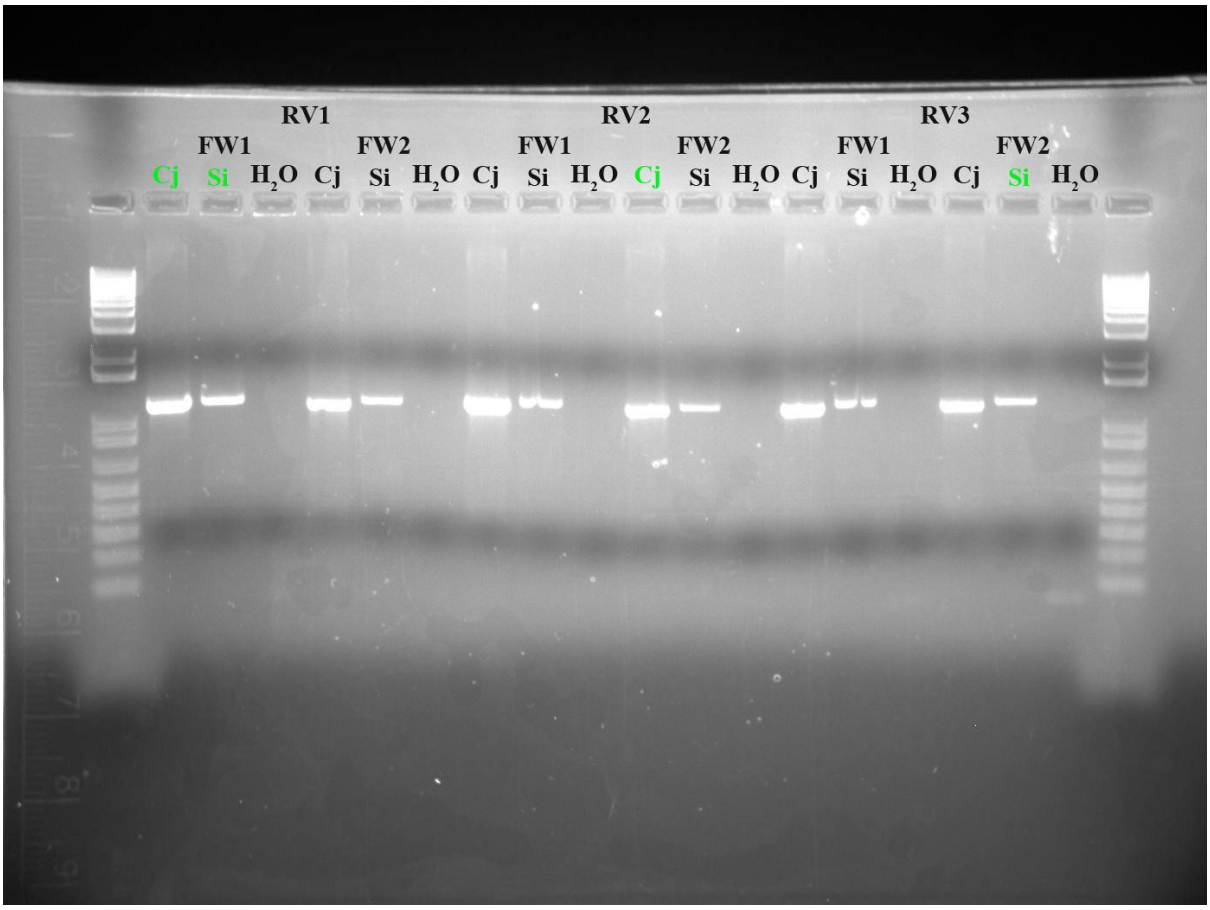

***Figure 24: PCR results for* C. jacchus *and* S.imperator.** *The PCR appeared to have produced the expected amplicons in all combinations of primers for both DNA samples. However, only amplicons highlighted in green were subsequently confirmed by sequencing.*

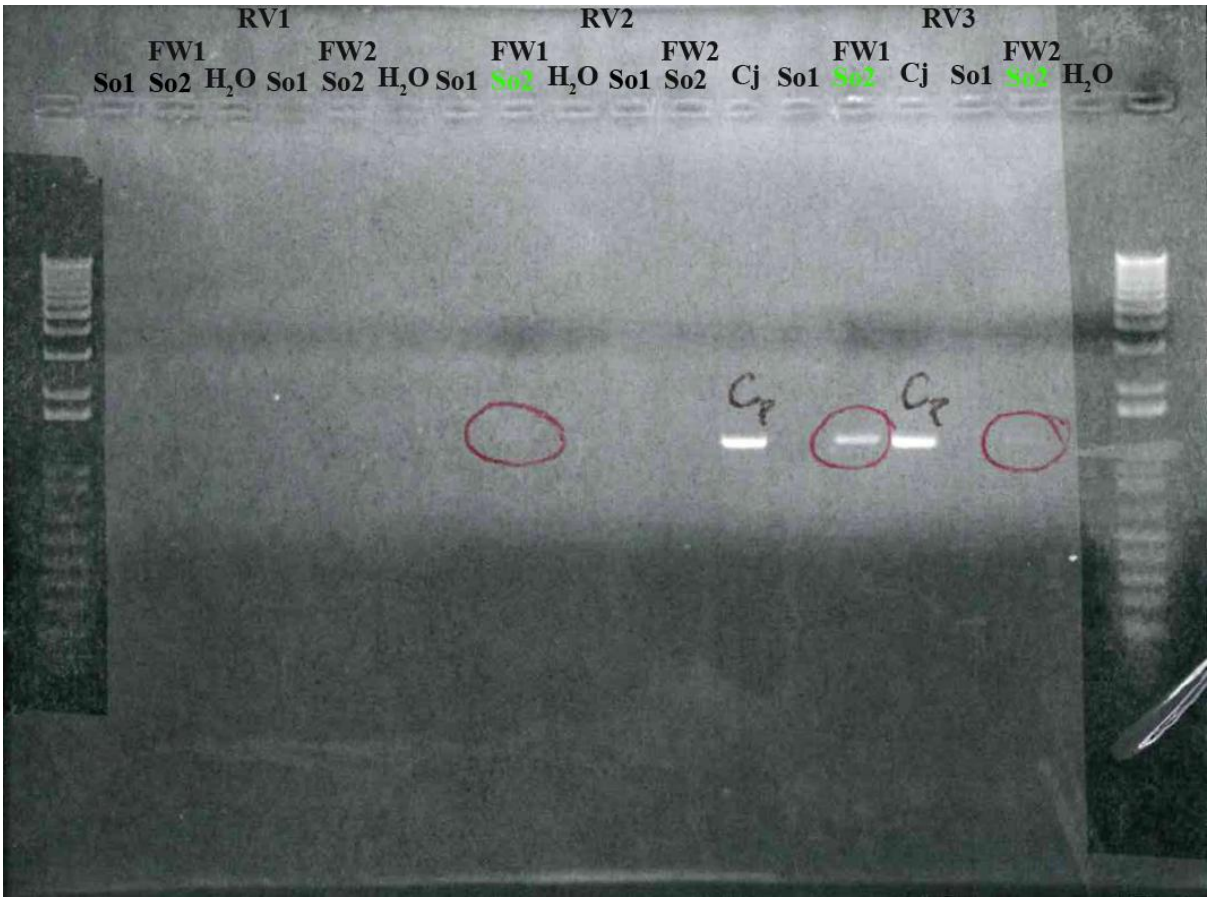

***Figure 25: PCR results for* S. oedipus.** *The PCR appeared to have produced the expected amplicons in the combinations of primers* FW1&RV2, FW1&RV3 FW2&RV3 *for one of the two DNA extractions (So2). Two positive controls (DNA from* C. jacchus*, Cj) were introduced in the experiment.*

Sequencing results for PCR reactions in *S. oedipus* are shown in **Figure 26**.

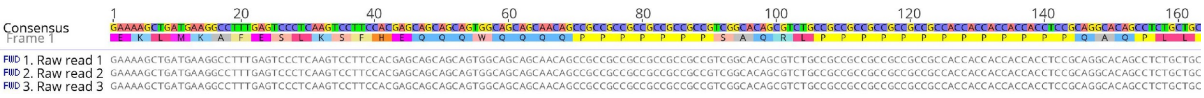

Figure 26: Alignment of sequencing reads for* S. oedipus *pseudogene. *Consensus and potential protein translation are shown above the reads.*

Furthermore, sequence reads for *C. jacchus* and *S. imperator* were aligned with previous reads from the same species and results are shown in **Figure 27** and **Figure 28**, respectively.

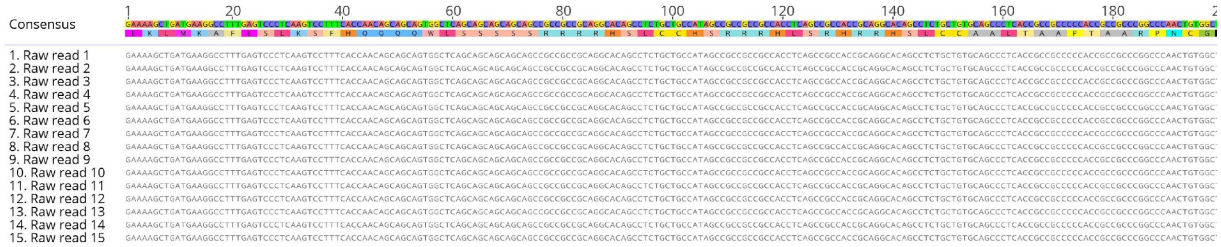

Figure 27: Alignment of sequencing reads for* C. jacchus *pseudogene. *Consensus and potential protein translation are shown above the reads.*

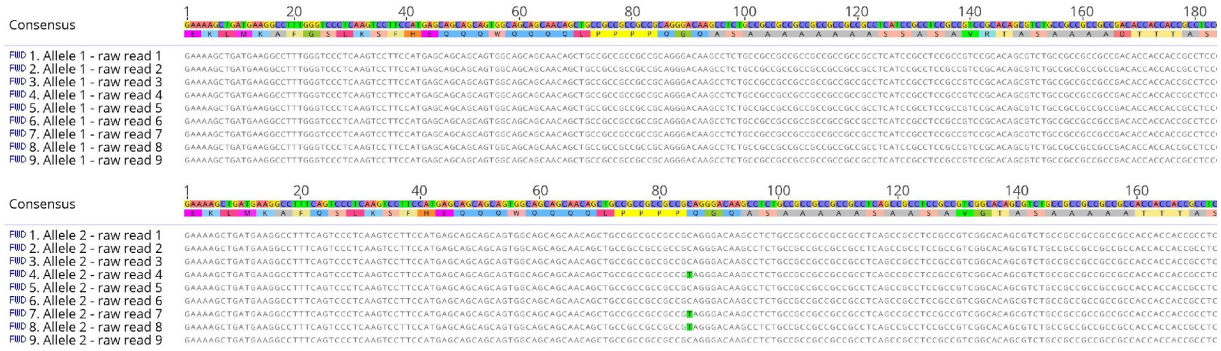

Figure 28: Alignment of sequencing reads for both alleles of* S. imperator *pseudogene. *Consensus and potential protein translation are shown above the reads.*

These results further confirmed the existence of a pseudogene in this primate family as well as the high variability of Htt exon 1 paralog sequence represented by the pseudogene, which also explained the abnormal length of *Emperor tamarin* amplicons. Sequencing results, while ultimately confirming pseudogene presence in Callitrichidae, could be aligned to reconstruct individual pseudogene sequences for *C. jacchus*, *S. imperator* (two alleles[73]) and *S. oedipus*. Independent validation of these findings on additional samples are ongoing in the laboratory.

5.4. Second visit to the PRI

To continue the investigation of Htt orthologs and paralogs (pHtt) in NHP, in 2017 I went back to the PRI in Inuyama. Also in this case the laboratory has supported the travel and the funding to purchase the reagents and perform the work. Samples were available through the PRI and its collaborators, and in particular through Dr. Takashi Hayakawa from the Japanese Monkey Center (JMC). A total of 111 samples belonging to different individuals from 35 NHP species were preliminarily individuated, **Table 8** provides a summary of the species. However, this preliminary list was further updated once on site according to the real availability of samples for processing (**Tables 9-11**).

[73] As recalled earlier, the second allele could not be independently amplified and could therefore result from a PCR artifact.

Species	**Source**
Aotus trivirgatus	PRI
Callithrix geoffroyi	JMC
Callithrix humeralifer	JMC
Callithrix jacchus	PRI
Callithrix penicillata	JMC
Callithrix pygmaea (= Cebuella pygmaea)	JMC
Cebus capucinus	PRI
Cercopithecus aethiops	PRI
Cercopithecus diana	PRI
Cercopithecus mitis	JMC
Galago senegalensis	JMC
Gorilla gorilla gorilla	PRI
Hylobates lar	PRI
Hylobates moloch	PRI
Hylobates pileatus	JMC
Macaca cyclopis	PRI
Macaca fuscata fuscata	PRI
Macaca mulatta	PRI
Macaca nigra	PRI
Macaca radiata	PRI
Macaca silenus	PRI
Pan paniscus	PRI
Pan troglodytes	PRI
Pan troglodytes verus	PRI

Papio hamadryas	PRI
Pongo abelii	PRI
Pongo pygmaeus	PRI
Saguinus imperator	JMC
Saguinus labiatus	JMC
Saguinus midas	JMC
Saguinus mystax	JMC
Saguinus oedipus	JMC
Saguinus oedipus	PRI
Saimiri sciureus	PRI
Symphalangus syndactylus	PRI/JMC

Table 8: Summary of the samples and species available for sequencing at PRI. *More samples were available for several species, however sampling was limited to up to five samples per species in order to maximize the chances of discovering long Htt alleles (see text for a detailed explanation).*

The choice of limiting sampling to five individuals per species was taken according to the following rationale and computer simulations. When we look at human CAG length distribution, we find that this tends to follow a "the longer the rarer" rule. Since this is likely to apply also to NHPs, in order to find very long alleles in an unbiased experimental design extensive sampling would probably be required. However, if we can find a species that appears to have a distribution of alleles identical or similar to humans (i.e. with a mean around 18.5 CAG repeats) then sampling 100 alleles from that species would allow to find at least 1 IA with P = 0.95 (bootstrap simulations n = 10^5). In order to find such a species, we need to sample as many different species as possible, under the fair assumption that they distribute approximately as humans (i.e. a distribution skewed toward longer alleles). To our knowledge this was a good working assumption since the same mechanisms generating human allelic length variation are likely to be at stake in other primate species and results in the Japanese macaque showed a right-skewed distribution. It turns out (bootstrap simulation n = 10^5) that the probability of assessing the true mean±3 in the human population from a sample with n = 5 is 95% (using the same confidence level the standard error is reduced to ±1 by a sampling of n = 40, while ±2 requires only n = 10). Under the aforementioned assumptions, to maximize the probability of finding AIs in NHPs, a reasonable approach would be to sample up to 5 alleles from any given species in order to determine whether that particular species has a mean length comparable to human and then sample within

that species, where possible, 100 alleles (50 individuals). It should be noted that the distribution could have been altered by inbreeding or drift and, in order to find the IAs, once the species with long CAG repeats has been identified a more extensive sampling could be required. However, once a species with mean and modal Htt CAG repeat lengths similar to humans was identified, it would have only been a matter of sampling. Hence, the plan at PRI was to try to assess the first five individuals for all the species available.
At PRI, I was made available by the master student Akihiro Itoigawa 7 already extracted DNA samples (**Table 9**).

ID	Tissue	Species
Unkown	Liver	*Aotus trivirgatus*
33-516	Liver	*Saimiri sciureus*
132-1903	Liver	*Saguinus oedipus*
Unkown	Liver	*Ateles belzebuth*
133-1933	Liver	*Aotus azarae*
Unkown	Liver	*Cebus apella*
104-135010	Liver	*Callithrix jacchus*

A first batch of 11 tissues was then obtained from Dr. Nagume Tani for DNA extraction (**Table 10**).

ID	Tissue	Species	Date
196	Liver	*S. oedipus*	31.10.16
18	Liver	*S. oedipus*	2.6.2010
130	Liver	*S. oedipus*	Unavailable
173	Liver	*Owl monkey*	Unavailable
165	Liver	*Owl monkey*	15.8.15
23	Liver	*C. jacchus*	2.9.2010
97	Liver	*C. jacchus*	2.11.2012
30	Liver	*C. jacchus*	Unavailable
32	Liver	*C. jacchus*	12.2010
198	Liver	*C. jacchus*	21.12.16
108	Liver	*C. jacchus*	Unavailable

Table 10: JMC samples where DNA extraction was performed.

DNA from samples listed in Table 10 was extracted and quantified at Nanodrop. I then received from JMC 100 samples of New World Monkeys. They represent a group of usually small primates, potentially suitable for disease modelling. I thus started DNA extraction from the first 22 of those tissue samples (**Table 11**).

ID	Family	Genus	Species	sex	Date
5610	Cebidae	*Callithrix*	*geoffroyi*	M	2003.01.12
5612	Atelidae	*Ateles*	*belzebuth*	F	2003.01.14
5625	Cebidae	*Saguinus*	*mystax*	F	2003.04.06
5642	Cebidae	*Callithrix*	*geoffroyi*	F	2003.06.21
5750	Cebidae	*Aotus*	*trivirgatus*	M	2005.03.12
5751	Atelidae	*Ateles*	*paniscus*	M	2005.03.26
5765	Cebidae	*Aotus*	*trivirgatus*	M	2005.06.03
5778	Cebidae	*Aotus*	*trivirgatus*	F	2005.07.20
5780	Atelidae	*Ateles*	*geoffroyi*	F	2005.08.08
5835	Cebidae	*Callithrix*	*geoffroyi*	F	2006.05.02
5871	Atelidae	*Ateles*	*geoffroyi*	F	2006.12.16
5890	Atelidae	*Lagothrix*	*lagothricha*	M	2007.04.05
5902	Cebidae	*Aotus*	*trivirgatus*	M	2007.05.05
5907	Atelidae	*Ateles*	*geoffroyi*	F	2007.05.15
5981	Cebidae	*Callithrix*	*geoffroyi*	M	2008.03.07
5982	Cebidae	*Callithrix*	*geoffroyi*	F	2008.03.07
5999	Atelidae	*Ateles*	*belzebuth*	M	2008.05.31
6099	Cebidae	*Aotus*	*trivirgatus*	U	2010.03.06
6128	Atelidae	*Ateles*	*geoffroyi*	F	2010.05.14
6162	Atelidae	*Ateles*	*belzebuth*	F	2010.11.10
6171	Cebidae	*Aotus*	*trivirgatus*	M	2011.01.18
6202	Atelidae	*Lagothrix*	*lagothricha*	M	2011.06.16

Table 11: JMC samples where DNA extraction was performed.

However, I have noticed that when amplifying the Callitrichidae samples, where both Htt gene and pHtt are present, Htt pseudogene is always preferentially amplified over the Htt gene. Results from sequencing of the samples from first 11 PRI samples suffered this issue. This requested me to design a new experimental strategy to selectively amplify the Htt gene in those species (see next section). Moreover, in JMC samples pHtt amplified also in species were it should not have amplified, suggesting that there could have been some DNA contamination in the sample. This is possible since Dr. Hayakawa had reported that many of these samples were very old. This is why it was ultimately not possible to draw solid conclusions from this effort.

5.5. Htt in Callitrichidae

The newly developed strategy for Htt exon 1 sequencing in Callitrichidae was successfully verified in Callitrichidae samples from the Biopark of Rome (*C. jacchus*, *S. imperator*) and to the Cotton-top tamarin (*S. oedipus*) provided by Prof. Saino where I had previously assessed the pHtt sequence. The protocol relies on a forward primer designed to be common to most mammals (5'-ATGGCGACCTGGAAAAGCTG-3') and on three reverse primers within the intronic region following the exon 1 (5'-CTGCTGGGTCACCCTGTC-3', 5'-GGGTGTCCCTACGGGTTT-3', and 5'-GAAGTGGGGGAGGGTCTC-3'). Since the pseudogene derives from retrotranscription of mRNA, it lacks intronic regions and would therefore not amplify. PCR conditions employed were as follows:

Reaction protocol	
Buffer HF 10x	1.5 µL
Mg_2SO_4 50 mM	0.24 µL
dNTPs 10 mM	0.3 µL
Primer FW 10 uM	0.5 µL
Primer RV 10 uM	0.5 µL
Platinum Taq DNA Polymerase High Fidelity (Cat. no.11304-011)	0.15 µL
H_2O to volume (15 µL)	
Template DNA	0.3 uL[74]
PCR conditions	
Initial denaturation	10' at 96 °C
36 cycles:	
Denaturation	45'' at 96 °C
Annealing	30'' at 58 °C
Extension	1' at 68 °C
Final extension	10' at 68 °C

Figure 29 shows the result of the PCR reaction.

[74] DNA concentration assessed at Nanodrop is not reliable for highly degraded DNA. Therefore an aliquot of the sample was employed.

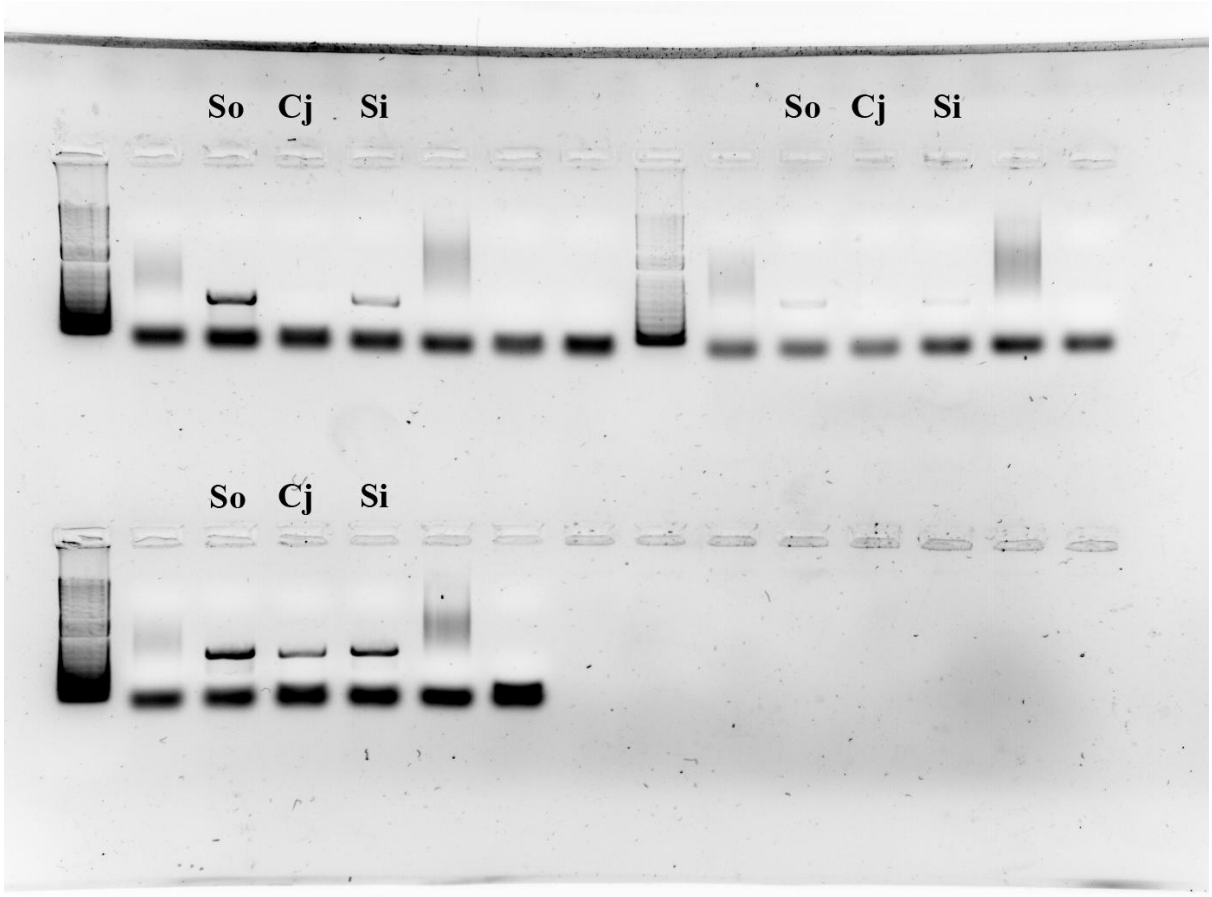

Figure 29: PCR reaction designed to target only the Htt gene in Callithricidae. S. oedipus *is found in lanes 3, 11 (top) and 3 (bottom).* C. jacchus *is found in lanes 4, 12 (top) and 4 (bottom).* S. imperator *is found in lanes 5, 13 (top) and 5 (bottom). 1 kb plus DNA ladder is present for comparison.*

Sequencing of the relative amplicons yielded results for *S. oedipus* and *S.* imperator, which are shown in **Figure 30** and **Figure 31** respectively, in agreement with the expected sequence for the Htt gene in primates, while cloning failed for *C. jacchus* (however its sequence is present in the NCBI RefSeq database).

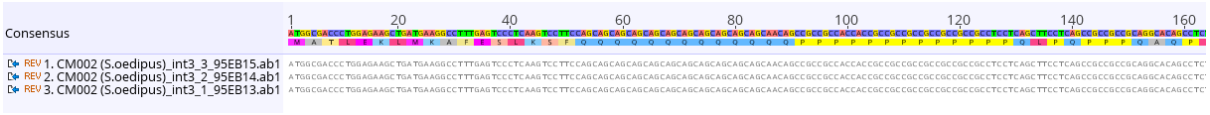

Figure 30: Alignment of sequencing reads for* S. oedipus *Htt gene.

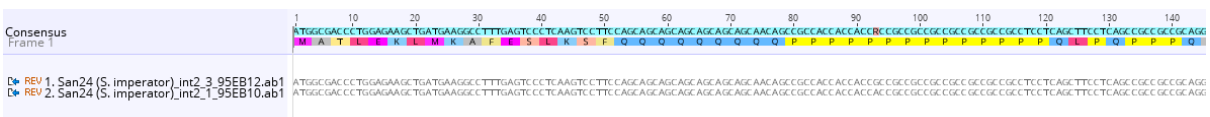

Figure 31: Alignment of sequencing reads for* S. imperator *Htt gene.

5.6. Data analysis

The serendipitously identified processed Htt pseudogene (pHtt) had escaped previous identification (Hohjoh et al. 2009) and ultimately turned out to be the most intriguing results in the NHP effort. Interestingly, the pHtt should exemplify the dynamics of this DNA sequence in the absence of selective constraints. I have found that the pHtt is composed of the paralog sequence of the first 9 Htt exons and lays within the third intron of the CNTNAP3 gene. Several point mutations and indels are present in the CAG repeat region in the four pHtt sequences collected from *Callithrix jacchus*, *Saguinus oedipus* (yet to be verified) and *Saguinus imperator* (two different alleles for the latter, second allele yet to be verified) as compared to the *HTT* gene from *Callithrix jacchus*. These include a G→A transition, at least three independent transversions (G→C, G→T and C→G) and all show either TGG or TGGCT insertions. This finding is in sharp contrast with what we documented in the Htt gene where we almost always observe an uninterrupted CAG tract, reinforcing the idea that natural selection acts on the Htt gene to preserve a pristine CAG repeat.

Another feature highlighted by the pHtt is that, as a whole, the exon 1 counterpart in the pHtt is far more prone to instability than the Htt gene exon 1. However, the CAG repeats show limited length variation (one extra CAG in *C. jacchus*) and this could reasonably be the consequence of the interruptions within the repeat itself, which are known to reduce the instability of repeats (see also Introduction). On the other hand, in the pHtt GC-rich region downstream the CAG repeat (homologous to the CCN/GCN region in the Htt gene) at least five insertions and two deletion events are highlighted when compared to the consensus sequence derived from all Htt and pHtt sequences available in primates.

This was the first attempt to evaluate pHtt sequences differences as new analyses are ongoing and some have suggested that a within-species comparison should be attempted. Nonetheless, these results suggest that in the absence of selective pressures Htt exon 1 sequence is highly unstable. The purity of CAG repeats is rapidly altered and the length of the repeats may vary considerably, with the repeats themselves being the major driver of indel mutations. Indirectly, these findings also point toward the presence of strong selective pressures to maintain the purity of CAG repeats and constrain the length of Htt exon 1.

Appendix 1

A1.1. Genbank genome browser

In 1982, the National Institute of Health (NIH) of the United States of America has launched a biological information repository called Genbank (Benson et al. 2005). This is now among the largest publicly available databases of sequence data in the world, containing the widest variety of annotated and often validated DNA sequences[75]. Genbank is constantly updated and maintained by the National Center for Biotechnology Information (NCBI), which is part of the United States National Library of Medicine (NLM), a branch of the NIH[76]. There are essentially two main direct ways to access DNA sequence data from the Genbank web interface: the 'gene' database and the 'nucleotide' database. These databases can be queried using dedicated search engines and user-defined query strings. The 'gene' database contains an organized list of all annotated and deposited genes from all organisms. For each gene, which is reported with its scientific name and abbreviation, a summary of all available information is provided in a single page, including the genomic region of the gene, the published literature related to the gene and all the relevant information in terms of annotated DNA sequences, including predicted and/or validated intron-exon boundaries of messenger RNA (mRNA) and protein sequences. The web page of the human Htt gene (Genbank ID: 3064)[77], a gene being at the focus of intensive studies since its discovery in 1993, contains an extensive record of sequence data and associated biological information. More specifically, it initially shows its chromosomal position in the context of the surrounding genes. This section is followed by a summary of the most recent literature published, together with a focus on the functional annotations that have been reported so far (GeneRIF[78]). The web page further shows a section dedicated to the phenotypic information that has been associated to the gene or to some of its specific variants (i.e. in this case its association with HD). It also connects to other databases of genetic variants associated to this gene (e.g. ClinVar, DbVar). Sections dedicated to its protein product localization, pathways and processes in which it is involved, interactions with other genes and proteins as well as the presence of recognized homologs are also displayed. Finally, a complete list of DNA sequences for this gene available on Genbank is reported. These can be distinguished depending on the source. The first source generally points to the genome assembly chosen as 'reference', often indicated as 'Primary Assembly'.

A genome assembly is the entire genomic consensus sequence derived through the assembly process of the raw sequencing reads and released by the curators of the genome in the database upon publication. It is important to note that until now the golden rule for genomic databases has been to report the genomic sequence for a single 'reference' individual. This relatively arbitrarily chosen individual(s) simply constitutes the individual(s) from which the DNA of the genome assembly derived (Nature Methods Editorial Board 2010). This has been mostly due to the cost of genome sequencing and assembly, which has

[75] "GenBank Overview." Accessed July 07, 2018. https://www.ncbi.nlm.nih.gov/genbank/
[76] "GenBank" Accessed July 07, 2018. https://en.wikipedia.org/wiki/GenBank
[77] "HTT Huntingtin [Homo Sapiens (human)] - Gene - NCBI." Accessed July 07, 2018. https://www.ncbi.nlm.nih.gov/gene/3064
[78] "About Gene RIF - Gene - NCBI." Accessed July 07, 2018. https://www.ncbi.nlm.nih.gov/gene/about-generif

not allowed for more than a genome assembly per species to be produced. Moreover, NGS approaches only enabled to define virtual 'haploid' genomes in the vast majority of cases (i.e. only one of the two alleles is reported for the reference). In very few cases the alternative alleles were reported as variants but the two haplotigs were often completely shuffled as it is unfeasible to determine long blocks of haplotype with NGS. For a given species, sometimes more than one assembly is present. This might result from serial attempts by the same or different research groups to increase the quality of the genome assembly from the same starting raw data, or might represent the result of subsequent sequencing efforts.

Generally, the Primary Assembly constitutes what Genbank curators consider the most up-to-date source for the genomic sequence for this reference. When available, this is the first result being shown. In the case of Htt, the Primary Assembly currently (June 2018) refers to the human genome assembly 'GRCh38' (Genome Reference Consortium human build n. 38[79]), the latest release of the long list of high-quality assemblies for the human genome generated since 2001[80]. As is the case for Htt, other validated sources may also be present (e.g. NCBI RefSeq genes[81]), together with their annotations (usually mRNA and protein). All these information and annotations normally result from automated algorithms and pipelines rather than from manual curation. On the contrary, when gene annotations have been manually reviewed by a human operator this is explicitly indicated. The Primary Assembly section further points at the Genbank sequence. It uses directly the coordinates for this gene that have been stored in Genbank database. A direct link to DNA sequence is also available, which represents a direct link to the 'nucleotide' database. This latter source, as well as RefSeq validated annotations when available, have been extensively employed in the present work as primary sources of DNA sequences.

A1.1.1. Genbank Primary Assembly record

The record headings for Htt gene in Genbank Primary Assembly, like all other genes present on Genbank, contain information on the authors of the sequencing (usually the members of a consortium aimed at establishing the WGS project), on the curation of the genome assembly and the article where the sequence/genome was originally published[82]. Importantly, they also report the genomic coordinates of the gene within the genome assembly, the coordinates on the gene for intron-exon boundaries, the coordinates of the coding sequence (which differs from the mRNA sequence because of the absence of untranslated regions), the entire encoded protein sequence and finally the DNA sequence. The latter is reported in rows, with spacing every ten nucleotides. A link to the FASTA file, one of the most exchangeable DNA sequence file formats, is also present.

[79] "GRCh38 - hg38 - Genome - Assembly - NCBI." Accessed July 07, 2018. https://www.ncbi.nlm.nih.gov/assembly/GCF_000001405.26/

[80] "Genome Browser FAQ." Accessed July 07, 2018. https://genome.ucsc.edu/FAQ/FAQreleases.html

[81] "About the NCBI RefSeqGene Project." Accessed July 07, 2018. https://www.ncbi.nlm.nih.gov/refseq/rsg/about/

[82] "Homo Sapiens Chromosome 4, GRCh38.p12 Primary Assembly - Nucleotide - NCBI." Accessed July 07, 2018. https://www.ncbi.nlm.nih.gov/nuccore/NC_000004.12?from=3074681&to=3243960&report=genbank

A1.1.2. FASTA file format

The text-based FASTA file format is among the most frequently employed file formats to store DNA sequence data[83]. Each sequence embedded into a FASTA file starts with the character '>' followed by the name of the sequence (without spaces). The sequence itself is then reported on new lines. A FASTA file can contain any number of sequence names and DNA sequences separated by new blank lines. In the present work, sequence data were often retrieved and stored using this file format.

A1.1.3. Genbank file format

The FASTA format is not the only way to store DNA information. Another relevant file format is the Genbank file format, which is capable of handling also most of the information on the annotations associated with the sequence. Genbank files can be downloaded directly from Genbank website, using the tool 'Send to:' present on top of its web pages. This type of file can subsequently be handled with specific softwares as Geneious (Kearse et al. 2012) and CLC Sequence Viewer[84].

A1.1.4. BLAST

The most popular algorithm/software for sequence search in Genbank database is called BLAST. BLAST, which stands for Basic Local Alignment Search Tool, was developed at NIH in 1990 (Altschul et al. 1990). The homonym algorithm on which it relies is generally much faster than other approaches and makes it practical to scan the huge genome databases currently available on Genbank (Elizabeth Cha and Rouchka 2005). The input query of this heuristic algorithm is represented by a query sequence in FASTA or Genbank format and by a matrix of weights. The latter represents the default or user-defined parameters that are thought to produce the best results given the specific query. A wide variety of adjustments can be made to optimize the search in the attempt to raise the chances of a true positive match. BLAST relies on the general principle of trying to match short sequences between the query and the sequences in the database, which are compared one-by-one in a parallel. Generally, BLAST outputs a long list of results, ordered by their distance with respect to the original query given the parameters of the matrix of weights. These results, also depending on the search settings, include genes from the 'gene' databases and nucleotide sequences from the 'nucleotide' database. BLAST and its evolutions have proven effective in retrieving annotated and unannotated orthologous genes using DNA sequences from closely related organisms as queries. Sometimes even distantly related organisms can do, provided that sequence conservation is high. This tool is present on both Genbank and Ensembl websites to query their databases and it has been widely applied in this study.

A1.2. Ensembl genome browser

Ensembl web portal was launched in 1999 as a joint scientific project between the European Bioinformatics Institute and the Wellcome Trust Sanger Institute, to store the sequence data deriving from the upcoming completion of the Human Genome Project (Hubbard et al. 2002). On average, while the two datasets look

[83] "FASTA Format." Accessed July 07, 2018. https://zhanglab.ccmb.med.umich.edu/FASTA/
[84] "CLC Sequence viewer - QIAGEN."
Accessed July 08, 2018. https://www.qiagenbioinformatics.com/products/clc-sequence-viewer/

very similar in terms of available genome data, Genbank is often more complete, at least with respect to the presence of Htt orthologous genes. Nonetheless, the two sources can sometimes be complementary to each other.

A typical search on Ensembl starts with the gene name from the website main page. For example, with 'Htt' as keyword, the first result displayed is the human Htt gene (Ensembl ID: ENSG00000197386). Similar to Genbank, the web page dedicated to Htt contains a summary of the information for this gene, gene ontologies, resources related to genetic variation and gene expression, a section on comparative genomics and a highlight of the gene sequence and its relative annotation. From this latter source, the DNA sequences can be downloaded in FASTA format.

A1.3. UCSC genome browser

Conceived at the University of California in Santa Cruz and released starting from 2000, the UCSC Genome Browser is an online resource to visualize and download genomic information on many vertebrate and invertebrate species (Tyner et al. 2017). It is similar to Genbank and Ensembl in many features, but it usually allows a more detailed and integrated view of sequence data. On the other hand, it generally harbours information from a more limited number of organisms than the other two databases. Nonetheless, it can sometimes be helpful for cross-validation, or for a closer inspection of sequence data.

A1.4. Ortholog gene search

In order to identify and retrieve the information of interest on a single gene or on a portion of a gene from these public databases, several — often complementary — approaches are possible. One way is to search through the annotations for genes that were already identified and marked as gene orthologs. Another approach is to use a known DNA sequence string as query to search directly into the annotated sequences or even in the raw sequence data. To this end, a range of solutions has been developed over the years. This approach is particularly helpful when the genomic information has already been uploaded in the database, but gene annotation is still incomplete or missing.

Appendix 2

ID	Class	Species	Source	
2	Mammalia (Primates)	*Homo neanderthalensis*	Publication	Prüfer et al. (2014) - raw data
3	Mammalia (Primates)	*Pan troglodytes*	GenBank	XM_016951204.1
4	Mammalia (Primates)	*Pan paniscus*	GenBank	XM_003812979.3
5	Mammalia (Primates)	*Gorilla gorilla*	GenBank	Alignment of 4 sequences: EU797062, Y07988, EU797063, L49364
6	Mammalia (Primates)	*Pongo pygmaeus*	GenBank	ENSPPYT00000016917.1
7	Mammalia (Primates)	*Aotus nancymaae*	GenBank	XM_012466834
8	Mammalia (Primates)	*Chlorocebus sabaeus*	GenBank	XM_008018062
9	Mammalia (Primates)	*Hylobates lar*	GenBank	EU797068
10	Mammalia (Primates)	*Cercocebus atys*	GenBank	XM_012059172
13	Mammalia (Primates)	*Macaca nemestrina*	GenBank	XM_011744916
14	Mammalia (Primates)	*Nomascus leucogenys*	GenBank	XM_012499669.1
15	Mammalia (Primates)	*Colobus guereza*	GenBank	EU797073
17	Mammalia (Primates)	*Papio anubis*	GenBank	XM_021938205.1
18	Mammalia (Primates)	*Macaca mulatta*	GenBank	XM_015137840.1
20	Mammalia (Primates)	*Macaca fascicularis*	GenBank	XM_015449989.1
21	Mammalia (Primates)	*Callithrix jacchus*	GenBank	NM_001267745.1
24	Mammalia (Primates)	*Tarsius syrichta*	GenBank	XM_021711235.1
25	Mammalia (Primates)	*Microcebus murinus*	GenBank	XM_012737354.2
26	Mammalia	*Tupaia belangeri*	GenBank	SRX198023
27	Mammalia	*Peromyscus maniculatus*	GenBank	XM_016004012.1
28	Mammalia	*Microtus ochrogaster*	GenBank	XM_005365914.2
29	Mammalia	*Mesocricetus auratus*	GenBank	XM_021226467.1
30	Mammalia	*Mus musculus*	GenBank	NM_010414.3
31	Mammalia	*Nannospalax galili*	GenBank	XM_008841869
32	Mammalia	*Rattus norvegicus*	GenBank	NM_024357.3
34	Mammalia	*Ictidomys tridecemlineatus*	GenBank	XM_005319002.3
38	Mammalia	*Octodon degus*	GenBank	XM_004624601.1
39	Mammalia	*Chinchilla lanigera*	GenBank	XM_013506814.1
40	Mammalia	*Cavia porcellus*	GenBank	XM_013157718.1
43	Mammalia	*Heterocephalus glaber*	GenBank	XM_021260364.1
45	Mammalia	*Bos taurus*	GenBank	XM_002688430.4
46	Mammalia	*Ovis aries*	GenBank	NM_001142638.1
48	Mammalia	*Orcinus orca*	GenBank	XM_004265165.2

51	Mammalia	*Sus scrofa*	GenBank	XM_013978501.2
54	Mammalia	*Odobenus rosmarus*	GenBank	XM_004396198.1
55	Mammalia	*Canis lupus*	GenBank	XM_536221.5
56	Mammalia	*Sorex araneus*	GenBank	XM_004617465.1
57	Mammalia	*Equus caballus*	Publication	Wade et al. (2009) - raw data
59	Mammalia	*Pteropus vampyrus*	Ensembl	ENSPVAT00000011321.1
69	Mammalia	*Loxodonta africana*	GenBank	XM_003411334.2
71	Mammalia	*Chrysochloris asiatica*	GenBank	XM_006875329.1
72	Mammalia	*Echinops telfairi*	GenBank	XM_004715096.1
73	Mammalia	*Dasypus novemcinctus*	GenBank	XM_012520645.1
74	Mammalia	*Monodelphis domestica*	GenBank	XM_016423341.1
75	Mammalia	*Macropus eugenii*	Ensembl	ENSMEUT00000004503.1 and raw data
77	Mammalia	*Ornithorhynchus anatinus*	GenBank	GCF_000002275.2 Ornithorhynchus_anatinus-5.0.1
86	Reptilia	*Anolis carolinensis*	GenBank	XM_008111310.2
91	Reptilia	*Chrysemys picta*	GenBank	XM_005304083.2
92	Reptilia	*Pelodiscus sinensis*	GenBank	XM_006128261.2
94	Reptilia	*Alligator mississippiensis*	GenBank	XM_006262647.3
95	Reptilia	*Alligator sinensis*	GenBank	XM_014524035.1
154	Aves	*Melopsittacus undulatus*	GenBank	XM_002195229.2
155	Aves	*Taeniopygia guttata*	GenBank	XM_005045496.1
156	Aves	*Ficedula albicollis*	GenBank	XM_005229813.1
157	Aves	*Falco peregrinus*	GenBank	XM_010583518
158	Aves	*Haliaeetus leucocephalus*	GenBank	XM_009099101.1
159	Aves	*Serinus canaria*	GenBank	XM_009471152.1
160	Aves	*Nipponia nippon*	GenBank	XM_010193794.1
161	Aves	*Mesitornis unicolor*	GenBank	XM_009285380.1
162	Aves	*Aptenodytes forsteri*	GenBank	XM_009984323.1
163	Aves	*Tauraco erythrolophus*	GenBank	XM_009708844.1
164	Aves	*Cariama cristata*	GenBank	XM_010133191.1
165	Aves	*Buceros rhinoceros*	GenBank	XM_005517876.1
166	Aves	*Pseudopodoces humilis*	GenBank	XM_010290303.1
167	Aves	*Phaethon lepturus*	GenBank	XM_009686173.1
168	Aves	*Struthio camelus*	GenBank	XM_010210577.1
169	Aves	*Colius striatus*	GenBank	XM_009948046.1
170	Aves	*Leptosomus discolor*	GenBank	XM_009948046.1
174	Aves	*Picoides pubescens*	GenBank	XM_009910100.1
175	Amphibia	*Xenopus tropicalis*	GenBank	XM_012955737.2
176	Amphibia	*Nanorana parkeri*	GenBank	XM_018554978.1
177	Amphibia	*Ambystoma mexicanum*	GenBank	XM_015604698.1
178	Sarcopterygii	*Latimeria chalumnae*	GenBank	XM_014489756.1
179	Actinopterygii	*Fundulus heteroclitus*	GenBank	XM_021320330.1
181	Actinopterygii	*Oryzias latipes*	GenBank	XM_020706268.1
182	Actinopterygii	*Oreochromis niloticus*	GenBank	XM_013276665.2
183	Actinopterygii	*Neolamprologus brichardi*	GenBank	XM_014341304.1
184	Actinopterygii	*Haplochromis burtoni*	GenBank	XM_014341304.1

185	Actinopterygii	*Xiphophorus maculatus*	GenBank	XM_005797294.2
186	Actinopterygii	*Poecilia orri*	GenBank	XM_016667560.1
187	Actinopterygii	*Poecilia reticulata*	GenBank	XM_017304689.1
188	Actinopterygii	*Tetraodon nigroviridis*	GenBank	ENSTNIT00000015214.1
189	Actinopterygii	*Stegastes partitus*	GenBank	XM_008277913.1
190	Actinopterygii	*Gasterosteus aculeatus*	GenBank	ENSGACT00000024315.1
191	Actinopterygii	*Takifugu rubripes*	GenBank	XM_011612986.1
192	Actinopterygii	*Maylandia zebra*	GenBank	XM_004538882.1
194	Actinopterygii	*Pundamilia nyererei*	GenBank	XM_013911515.1
195	Actinopterygii	*Esox lucius*	GenBank	XM_020043694.1
196	Actinopterygii	*Gadus morhua*	GenBank	ENSGMOT00000014002.1
197	Actinopterygii	*Astyanax mexicanus*	GenBank	XM_015604698.1
198	Actinopterygii	*Lepisosteus oculatus*	GenBank	XM_015344842.1
199	Actinopterygii	*Clupea harengus*	GenBank	XM_012836874.1
200	Actinopterygii	*Larimichthys crocea*	GenBank	XM_019278160.1
202	Actinopterygii	*Cynoglossus semilaevis*	GenBank	XM_017035042.1
204	Actinopterygii	*Danio rerio*	GenBank	NM_131018.1
205	Chondrichthyes	*Callorhinchus milii*	GenBank	XM_007905263.1
206	Leptocardii	*Branchiostoma lanceolatum*	Publication	Candiani et al. (2007)
207	Leptocardii	*Branchiostoma floridae*	Publication	Candiani et al. (2007)
208	Enteropneusta	*Saccoglossus kowalevskii*	GenBank	XM_006822922.1
209	Echinoidea	*Strongylocentrotus purpuratus*	GenBank	AM422557.1

Table 1: List of Htt exon 1 sequences retrieved from databases. *A total of 100 were included in the final dataset. The original source (Genbank, Ensembl, publication) is reported with the corresponding accession number.*

Appendix 3

Supplier	**Institute**
Adriana Bellati	Dipartimento di Scienze della Terra e dell'Ambiente, Università di Pavia (Pavia, IT)
Adriano Martinoli	Dipartimento di scienze teoriche e applicate, Università dell'Insubria (Varese, IT)
Alessandro Balestrieri	Dipartimento di Scienze e Politiche Ambientali, Università degli Studi di Milano (Milano, IT)
Alessandro Bianchi	Istituto Zooprofilattico Sperimentale "Bruno Ubertini" (Sondrio, IT)
Ana Rubio Garcia	Seal Rehabilitation and Research Centre (Pieterburen, NL)
Andrea Sforzi	Museo di Storia Naturale della Maremma (Grosseto, IT)
Andres Barbosa	Museo Nacional de Ciencias Naturales (Madrid, SP)
Angelica Crottini	Dipartimento di Biotecnologie e Bioscienze, Università di Milano-Bicocca (Milano, IT)
Bill Randall*	Coriell Institute (New York, USA)
Chang Anthony*	Yerkes National Primate research center, Emory University (Lawrenceville, Georgia, USA)
Christian Abee*	MD Anderson Cancer Center (Houston, Texas, USA)
Claudia Romeo	Dipartimento di Medicina Veterinaria, Università degli Studi di Milano (Milano, IT)
Cristina LaBarga	Bioparco di Roma, Università di Roma "Tor Vergata" (Roma, IT)
Daniel Berkowic	Department of Zoology, Tel Aviv University (Tel Aviv, IL)
Edoardo Razzetti	Museo di Storia Naturale di Pavia (Pavia, IT)
Enrico Merli	Ente forestale Provincia di Piacenza (Piacenza, IT)
Giorgia Tessa	Università di Torino (Torino, IT)
Giorgio Bardelli	Museo di Storia Naturale di Milano (Milano, IT)
Giorgio De Giorgi	SafariPark di Pombia (Novara, IT)
Giovanni Amori*	Dipartimento di Biologia e Biotecnologie "Charles Darwin", Università La Sapienza di Roma (Roma, IT) e Consiglio Nazionale delle Ricerche
Guoping Feng*	Massachusetts Institute of Technology (Boston, Massachusetts, USA)
Hirai Hirohisa	Primate Research Institute, Kyoto University (Inuyama, Japan)
Hiroo Imai	Primate Research Institute, Kyoto University (Inuyama, Japan)
Irene Bertoletti	Istituto Zooprofilattico Sperimentale "Bruno Ubertini" (Sondrio, IT)
Jerilyn Pecotte*	Southwestern National Primate research center (San Antonio, Texas, USA)

Joanna Sumner	Museums Victoria (Carlton, AUS)
Jodi McBride*	Oregon Health and Science University (Portland, USA) and Oregon national primate Research Center (Beaverton, Oregon, USA)
John Vandeberg*	Southwestern National Primate research center (San Antonio, Texas, USA)
Mahmoud Pouladi*	Translational Laboratory in Genetic Medicine (Singapore)
Marco Apollonio	Dipartimento di Zoologia, Università di Sassari (Sassari, IT)
Maria Rasotto	Dipartimento di Biologia, Università di Padova (Padova, IT)
Maristella Giurisato	Dipartimento di Biomedicina Comparata e Alimentazione, Università di Padova (Padova, IT) and Banca Tessuti Cetacei del Mediterraneo
Massimo Scandura	Dipartimento di Zoologia, Università di Sassari (Sassari, IT)
Maurizio Casiraghi	Dipartimento di Biotecnologie e Bioscienze, Università di Milano-Bicocca (Milano, IT)
Michela Podestà	Museo di Storia Naturale di Milano (Milano, IT)
Nicola Saino	Dipartimento di Scienze e Politiche Ambientali, Università degli Studi di Milano (Milano, IT), Oasi di Sant'Alessio Con vialone (Pavia, IT)
Nicoletta Ancona	Acquario Civico di Milano (Milano, IT)
Olga Rickards	Bioparco di Roma, Università di Roma "Tor Vergata" (Roma, IT)
Paolo Ciucci	Dipartimento di Biologia e Biotecnologie "Charles Darwin", Università La Sapienza di Roma (Roma, IT
Riccardo Castiglia	Dipartimento di Biologia e Biotecnologie "Charles Darwin", Università La Sapienza di Roma (Roma, IT) e Consiglio Nazionale delle Ricerche
Roberta Castiglioni	Parco Faunistico Le Cornelle (Bergamo, IT)
Shai Meiri	Department of Zoology, Tel Aviv University (Tel Aviv, IL)
Silvia Colmegna	Istituto Zooprofilattico Sperimentale "Bruno Ubertini" (Milano, IT)

Table 1: List of people and Institutions that were enquired for samples and eventually provided samples for this project. *A total of 43 researchers belonging to 34 Institutions were involved in the sampling phase. Contact persons that ultimately did not provide samples to the laboratory are indicated by *.*

Appendix 4

Individual No.	Allele 1	Allele 2
684	9	8
755	9	
1730	9	
1770	10	8
1841	9	
1224	8	
1304	8	
1308	8	
1316	8	
1549	9	
1554	8	
1628	8	
1668	8	
1673	9	
654	9	
772	9	12
1818	9	
1887	9	
1394	11	12
1466	12	10
1848	10	
1882	9	
1916	10	
1954	10	
1969	10	
1991	9	
2037	9	
2073	9	
1997	9	
1	9	
2	9	
3	9	
11	9	
17	9	
19	9	
21	9	
23	9	

26	9	
27	9	
28	9	
35	9	
43	9	
50	9	
52	9	
53	9	
59	9	
65	9	
69	9	
N-195	12	
N-211	12	
2284	9	
2253	9	
2285	9	
2262	12	
2263	9	
2268	9	
2114	9	
2117	9	
N-182	6	
N-185	9	
N-186	9	
H19-149	9	
H20-091	9	
H20-107	9	
H20-108	9	
H20-112	9	
H20-113	9	
H20-117	9	
H20-118	9	
H21-104	9	
VM08-241	9	
VM09-449	9	
VM10-131	9	
VM10-263	9	
VM10-272	9	
VM10-273	9	
VM10-311	6	

VM10-388	9	

Table 1: ID and CAG assessed in* M. fuscata *samples from the PRI. *When secondary alleles were detected (multiple lengths from different sequencing reactions) these are reported.*

References

Albin, R. L., and D. A. Tagle. 1995. "Genetics and Molecular Biology of Huntington's Disease." *Trends in Neurosciences* 18 (1): 11–14.

Almqvist, E. W., D. S. Elterman, P. M. MacLeod, and M. R. Hayden. 2001. "High Incidence Rate and Absent Family Histories in One Quarter of Patients Newly Diagnosed with Huntington Disease in British Columbia." *Clinical Genetics* 60 (3): 198–205.

Altschul, S. F., W. Gish, W. Miller, E. W. Myers, and D. J. Lipman. 1990. "Basic Local Alignment Search Tool." *Journal of Molecular Biology* 215 (3): 403–10.

Andrade, M. A., C. Petosa, and S. I. O'Donoghue. 2001. "Comparison of ARM and HEAT Protein repeats." *Journal of Molecular*. https://www.sciencedirect.com/science/article/pii/S0022283601946248.

Athey, John, Aikaterini Alexaki, Ekaterina Osipova, Alexandre Rostovtsev, Luis V. Santana-Quintero, Upendra Katneni, Vahan Simonyan, and Chava Kimchi-Sarfaty. 2017. "A New and Updated Resource for Codon Usage Tables." *BMC Bioinformatics* 18 (1): 391.

Bacolla, Albino, Jacquelynn E. Larson, Jack R. Collins, Jian Li, Aleksandar Milosavljevic, Peter D. Stenson, David N. Cooper, and Robert D. Wells. 2008. "Abundance and Length of Simple Repeats in Vertebrate Genomes Are Determined by Their Structural Properties." *Genome Research* 18 (10): 1545–53.

Bakhach, Joseph. 2009. "The Cryopreservation of Composite Tissues: Principles and Recent Advancement on Cryopreservation of Different Type of Tissues." *Organogenesis* 5 (3): 119–26.

Barnes, G. T., M. P. Duyao, C. M. Ambrose, S. McNeil, F. Persichetti, J. Srinidhi, J. F. Gusella, and M. E. MacDonald. 1994. "Mouse Huntington's Disease Gene Homolog (Hdh)." *Somatic Cell and Molecular Genetics* 20 (2): 87–97.

Bates, G., P. S. Harper, and L. Jones. 2002. *Huntington's Disease*. Oxford University Press.

Baxendale, S., S. Abdulla, G. Elgar, D. Buck, M. Berks, G. Micklem, R. Durbin, G. Bates, S. Brenner, and S. Beck. 1995. "Comparative Sequence Analysis of the Human and Pufferfish Huntington's Disease Genes." *Nature Genetics* 10 (1): 67–76.

Benson, Dennis A., Ilene Karsch-Mizrachi, David J. Lipman, James Ostell, and David L. Wheeler. 2005. "GenBank." *Nucleic Acids Research* 33 (Database issue): D34–38.

Bernard, P., and M. Couturier. 1992. "Cell Killing by the F Plasmid CcdB Protein Involves Poisoning of DNA-Topoisomerase II Complexes." *Journal of Molecular Biology* 226 (3): 735–45.

Bernard, P., P. Gabant, E. M. Bahassi, and M. Couturier. 1994. "Positive-Selection Vectors Using the F Plasmid ccdB Killer Gene." *Gene* 148 (1): 71–74.

Bernard, P., K. E. Kézdy, L. Van Melderen, J. Steyaert, L. Wyns, M. L. Pato, P. N. Higgins, and M. Couturier. 1993. "The F Plasmid CcdB Protein Induces Efficient ATP-Dependent DNA Cleavage by Gyrase." *Journal of Molecular Biology* 234 (3): 534–41.

Biscotti, Maria Assunta, Ettore Olmo, and J. S. Pat Heslop-Harrison. 2015. "Repetitive DNA in Eukaryotic Genomes." *Chromosome Research: An International Journal on the Molecular, Supramolecular and Evolutionary Aspects of Chromosome Biology* 23 (3): 415–20.

Brock, G. J., N. H. Anderson, and D. G. Monckton. 1999. "Cis-Acting Modifiers of Expanded CAG/CTG Triplet Repeat Expandability: Associations with Flanking GC Content and Proximity to CpG Islands." *Human Molecular Genetics* 8 (6): 1061–67.

Candiani, Simona, Mario Pestarino, Elena Cattaneo, and Marzia Tartari. 2007. "Characterization, Developmental Expression and Evolutionary Features of the Huntingtin Gene in the Amphioxus Branchiostoma Floridae." *BMC Developmental Biology* 7 (November): 127.

Cattaneo, E., D. Rigamonti, D. Goffredo, C. Zuccato, F. Squitieri, and S. Sipione. 2001. "Loss of Normal Huntingtin Function: New Developments in Huntington's Disease Research." *Trends in Neurosciences*

24 (3): 182–88.

Cattaneo, E., C. Zuccato, and M. Tartari. 2005. "Normal Huntingtin Function: An Alternative Approach to Huntington's Disease." *Nature Reviews. Neuroscience* 6 (12): 919–30.

Chung, Daniel W., Dobrila D. Rudnicki, Lan Yu, and Russell L. Margolis. 2011. "A Natural Antisense Transcript at the Huntington's Disease Repeat Locus Regulates HTT Expression." *Human Molecular Genetics* 20 (17): 3467–77.

Clabough, Erin B. D., and Scott O. Zeitlin. 2006. "Deletion of the Triplet Repeat Encoding Polyglutamine within the Mouse Huntington's Disease Gene Results in Subtle Behavioral/motor Phenotypes in Vivo and Elevated Levels of ATP with Cellular Senescence in Vitro." *Human Molecular Genetics* 15 (4): 607–23.

Coles, R., J. Leggo, and D. C. Rubinsztein. 1997. "Analysis of the 5'upstream Sequence of the Huntington's Disease (HD) Gene Shows Six New Rare Alleles Which Are Unrelated to the Age at Onset of HD." *Journal of Medical Genetics*. http://jmg.bmj.com/content/34/5/371.short.

Cornett, Jonathan, Fengli Cao, Chuan-En Wang, Christopher A. Ross, Gillian P. Bates, Shi-Hua Li, and Xiao-Jiang Li. 2005. "Polyglutamine Expansion of Huntingtin Impairs Its Nuclear Export." *Nature Genetics* 37 (2): 198–204.

Deaton, Aimée M., and Adrian Bird. 2011. "CpG Islands and the Regulation of Transcription." *Genes & Development* 25 (10): 1010–22.

DiFiglia, M., E. Sapp, K. Chase, C. Schwarz, A. Meloni, C. Young, E. Martin, J. P. Vonsattel, R. Carraway, and S. A. Reeves. 1995. "Huntingtin Is a Cytoplasmic Protein Associated with Vesicles in Human and Rat Brain Neurons." *Neuron* 14 (5): 1075–81.

Durrett, Richard, and Semyon Kruglyak. 1999. "A New Stochastic Model of Microsatellite Evolution." *Journal of Applied Probability* 36 (3): 621–31.

Elizabeth Cha, I., and Eric C. Rouchka. 2005. "Comparison of Current BLAST Software on Nucleotide Sequences." *IPDPS ... / International Parallel and Distributed Processing Symposium. IPDPS* 19 (April): 8.

Ellegren, Hans. 2004. "Microsatellites: Simple Sequences with Complex Evolution." *Nature Reviews. Genetics* 5 (6): 435–45.

Eskenazi, Benjamin R., Noah S. Wilson-Rich, and Philip T. Starks. 2007. "A Darwinian Approach to Huntington's Disease: Subtle Health Benefits of a Neurological Disorder." *Medical Hypotheses* 69 (6): 1183–89.

Falush, D. 2009. "Haplotype Background, Repeat Length Evolution, and Huntington's Disease." *American Journal of Human Genetics* 85 (6): 939–42.

Faux, Noel G., Stephen P. Bottomley, Arthur M. Lesk, James A. Irving, John R. Morrison, Maria Garcia de la Banda, and James C. Whisstock. 2005. "Functional Insights from the Distribution and Role of Homopeptide Repeat-Containing Proteins." *Genome Research* 15 (4): 537–51.

Ferrante, R. J., O. A. Andreassen, B. G. Jenkins, A. Dedeoglu, S. Kuemmerle, J. K. Kubilus, R. Kaddurah-Daouk, S. M. Hersch, and M. F. Beal. 2000. "Neuroprotective Effects of Creatine in a Transgenic Mouse Model of Huntington's Disease." *The Journal of Neuroscience: The Official Journal of the Society for Neuroscience* 20 (12): 4389–97.

Fisher, Emily R., and Michael R. Hayden. 2014. "Multisource Ascertainment of Huntington Disease in Canada: Prevalence and Population at Risk." *Movement Disorders: Official Journal of the Movement Disorder Society* 29 (1): 105–14.

Frost, D. R. 2016. "Amphibian Species of the World: An Online Reference. Version 6.0. New York: American Museum of Natural History."

Futuyma, D. J. 2013. *Evolution*. Sinauer.

Garcia-Diaz, Miguel, and Thomas A. Kunkel. 2006. "Mechanism of a Genetic Glissando: Structural Biology

of Indel Mutations." *Trends in Biochemical Sciences* 31 (4): 206–14.

Gardiner-Garden, M., and M. Frommer. 1987. "CpG Islands in Vertebrate Genomes." *Journal of Molecular Biology* 196 (2): 261–82.

Gauthier, Laurent R., Bénédicte C. Charrin, Maria Borrell-Pagès, Jim P. Dompierre, Hélène Rangone, Fabrice P. Cordelières, Jan De Mey, et al. 2004. "Huntingtin Controls Neurotrophic Support and Survival of Neurons by Enhancing BDNF Vesicular Transport along Microtubules." *Cell* 118 (1): 127–38.

Gellera, C., C. Meoni, B. Castellotti, B. Zappacosta, F. Girotti, F. Taroni, and S. DiDonato. 1996. "Errors in Huntington Disease Diagnostic Test Caused by Trinucleotide Deletion in the IT15 Gene." *American Journal of Human Genetics* 59 (2): 475–77.

Genetic Modifiers of Huntington's Disease (GeM-HD) Consortium. 2015. "Identification of Genetic Factors That Modify Clinical Onset of Huntington's Disease." *Cell* 162 (3): 516–26.

Genome 10K Community of Scientists. 2009. "Genome 10K: A Proposal to Obtain Whole-Genome Sequence for 10,000 Vertebrate Species." *The Journal of Heredity* 100 (6): 659–74.

Gissi, C., G. Pesole, E. Cattaneo, and M. Tartari. 2006. "Huntingtin Gene Evolution in Chordata and Its Peculiar Features in the Ascidian Ciona Genus." *BMC Genomics* 7 (November): 288.

Gouy, Manolo, Stéphane Guindon, and Olivier Gascuel. 2010. "SeaView Version 4: A Multiplatform Graphical User Interface for Sequence Alignment and Phylogenetic Tree Building." *Molecular Biology and Evolution* 27 (2): 221–24.

Guo, Qiang, Bin Huang, Jingdong Cheng, Manuel Seefelder, Tatjana Engler, Günter Pfeifer, Patrick Oeckl, et al. 2018. "The Cryo-Electron Microscopy Structure of Huntingtin." *Nature* 555 (7694): 117–20.

Gusella, J. F., M. E. MacDonald, and J. M. Lee. 2014. "Genetic Modifiers of Huntington's Disease." *Movement Disorders: Official Journal of the Movement Disorder Society* 29 (11): 1359–65.

Gusella, J. F., N. S. Wexler, P. M. Conneally, S. L. Naylor, M. A. Anderson, R. E. Tanzi, P. C. Watkins, K. Ottina, M. R. Wallace, and A. Y. Sakaguchi. 1983. "A Polymorphic DNA Marker Genetically Linked to Huntington's Disease." *Nature* 306 (5940): 234–38.

Gutekunst, C. A., A. I. Levey, C. J. Heilman, W. L. Whaley, H. Yi, N. R. Nash, H. D. Rees, J. J. Madden, and S. M. Hersch. 1995. "Identification and Localization of Huntingtin in Brain and Human Lymphoblastoid Cell Lines with Anti-Fusion Protein Antibodies." *Proceedings of the National Academy of Sciences of the United States of America* 92 (19): 8710–14.

Haasl, Ryan J., and Bret A. Payseur. 2013. "Microsatellites as Targets of Natural Selection." *Molecular Biology and Evolution* 30 (2): 285–98.

Haldane, J. B. S. 1941. *New Paths in Genetics*. Allen & Unwin.

Hantraye, P., D. Riche, M. Maziere, and O. Isacson. 1990. "A Primate Model of Huntington's Disease: Behavioral and Anatomical Studies of Unilateral Excitotoxic Lesions of the Caudate-Putamen in the Baboon." *Experimental Neurology* 108 (2): 91–104.

Heather, James M., and Benjamin Chain. 2016. "The Sequence of Sequencers: The History of Sequencing DNA." *Genomics* 107 (1): 1–8.

He, Dan, Farhad Hormozdiari, Nicholas Furlotte, and Eleazar Eskin. 2011. "Efficient Algorithms for Tandem Copy Number Variation Reconstruction in Repeat-Rich Regions." *Bioinformatics* 27 (11): 1513–20.

Hohjoh, Hirohiko, Hirofumi Akari, Yuko Fujiwara, Yoshiko Tamura, Hirohisa Hirai, and Keiji Wada. 2009. "Molecular Cloning and Characterization of the Common Marmoset Huntingtin Gene." *Gene* 432 (1-2): 60–66.

Hubbard, T., D. Barker, E. Birney, G. Cameron, Y. Chen, L. Clark, T. Cox, et al. 2002. "The Ensembl Genome Database Project." *Nucleic Acids Research* 30 (1): 38–41.

Innis, Michael A., David H. Gelfand, and John J. Sninsky. 1999. *PCR Applications: Protocols for*

Functional Genomics. Academic Press.

Jarne, P., and P. J. Lagoda. 1996. "Microsatellites, from Molecules to Populations and Back." *Trends in Ecology & Evolution* 11 (10): 424–29.

Jarvis, Erich D., Siavash Mirarab, Andre J. Aberer, Bo Li, Peter Houde, Cai Li, Simon Y. W. Ho, et al. 2014. "Whole-Genome Analyses Resolve Early Branches in the Tree of Life of Modern Birds." *Science* 346 (6215): 1320–31.

Jeong, Hyunkyung, Florian Then, Thomas J. Melia Jr, Joseph R. Mazzulli, Libin Cui, Jeffrey N. Savas, Cindy Voisine, et al. 2009. "Acetylation Targets Mutant Huntingtin to Autophagosomes for Degradation." *Cell* 137 (1): 60–72.

Johnsen, A., A. E. Fidler, S. Kuhn, K. L. Carter, A. Hoffmann, I. R. Barr, C. Biard, et al. 2007. "Avian Clock Gene Polymorphism: Evidence for a Latitudinal Cline in Allele Frequencies." *Molecular Ecology* 16 (22): 4867–80.

Jorda, Julien, and Andrey V. Kajava. 2010. "Protein Homorepeats Sequences, Structures, Evolution, and Functions." *Advances in Protein Chemistry and Structural Biology* 79: 59–88.

Kalchman, M. A., R. K. Graham, G. Xia, H. B. Koide, J. G. Hodgson, K. C. Graham, Y. P. Goldberg, R. D. Gietz, C. M. Pickart, and M. R. Hayden. 1996. "Huntingtin Is Ubiquitinated and Interacts with a Specific Ubiquitin-Conjugating Enzyme." *The Journal of Biological Chemistry* 271 (32): 19385–94.

Karlovich, Chris A., Rosalind M. John, Lucia Ramirez, Didier Y. R. Stainier, and Richard M. Myers. 1998. "Characterization of the Huntington's Disease (HD) Gene Homolog in the Zebrafish Danio Rerio." *Gene* 217 (1): 117–25.

Kashi, Yechezkel, and David G. King. 2006. "Simple Sequence Repeats as Advantageous Mutators in Evolution." *Trends in Genetics: TIG* 22 (5): 253–59.

Kauffman, Jeffrey S., Anna Zinovyeva, Kasumi Yagi, Kazuhiro W. Makabe, and Rudolf A. Raff. 2003. "Neural Expression of the Huntington's Disease Gene as a Chordate Evolutionary Novelty." *Journal of Experimental Zoology. Part B, Molecular and Developmental Evolution* 297 (1): 57–64.

Kearse, Matthew, Richard Moir, Amy Wilson, Steven Stones-Havas, Matthew Cheung, Shane Sturrock, Simon Buxton, et al. 2012. "Geneious Basic: An Integrated and Extendable Desktop Software Platform for the Organization and Analysis of Sequence Data." *Bioinformatics* 28 (12): 1647–49.

Kebschull, Justus M., and Anthony M. Zador. 2015. "Sources of PCR-Induced Distortions in High-Throughput Sequencing Data Sets." *Nucleic Acids Research* 43 (21): e143.

Kimpton, C. P., P. Gill, A. Walton, A. Urquhart, E. S. Millican, and M. Adams. 1993. "Automated DNA Profiling Employing Multiplex Amplification of Short Tandem Repeat Loci." *PCR Methods and Applications* 3 (1): 13–22.

Koepfli, Klaus-Peter, Benedict Paten, Genome 10K Community of Scientists, and Stephen J. O'Brien. 2015. "The Genome 10K Project: A Way Forward." *Annual Review of Animal Biosciences* 3: 57–111.

Kozlowski, Piotr, Mateusz de Mezer, and Wlodzimierz J. Krzyzosiak. 2010. "Trinucleotide Repeats in Human Genome and Exome." *Nucleic Acids Research* 38 (12): 4027–39.

Kremer, Berry, Paul Goldberg, Susan E. Andrew, Jane Theilmann, Hakan Telenius, Jutta Zeisler, Ferdinando Squitieri, et al. 1994. "A Worldwide Study of the Huntington's Disease Mutation: The Sensitivity and Specificity of Measuring CAG Repeats." *The New England Journal of Medicine* 330 (20): 1401–6.

Kroutil, L. C., K. Register, K. Bebenek, and T. A. Kunkel. 1996. "Exonucleolytic Proofreading during Replication of Repetitive DNA." *Biochemistry* 35 (3): 1046–53.

Kruglyak, S., R. T. Durrett, M. D. Schug, and C. F. Aquadro. 1998. "Equilibrium Distributions of Microsatellite Repeat Length Resulting from a Balance between Slippage Events and Point Mutations." *Proceedings of the National Academy of Sciences of the United States of America* 95 (18): 10774–78.

Kunkel, Thomas A. 2004. "DNA Replication Fidelity." *The Journal of Biological Chemistry* 279 (17): 16895–98.

Lander, E. S., L. M. Linton, B. Birren, C. Nusbaum, M. C. Zody, J. Baldwin, K. Devon, et al. 2001. "Initial Sequencing and Analysis of the Human Genome." *Nature* 409 (6822): 860–921.

Lee, J-M, E. M. Ramos, J-H Lee, T. Gillis, J. S. Mysore, M. R. Hayden, S. C. Warby, et al. 2012. "CAG Repeat Expansion in Huntington Disease Determines Age at Onset in a Fully Dominant Fashion." *Neurology* 78 (10): 690–95.

Lewin, Harris A., Gene E. Robinson, W. John Kress, William J. Baker, Jonathan Coddington, Keith A. Crandall, Richard Durbin, et al. 2018. "Earth BioGenome Project: Sequencing Life for the Future of Life." *Proceedings of the National Academy of Sciences of the United States of America* 115 (17): 4325–33.

Lin, B., J. Nasir, and H. MacDonald. 1994. "Sequence of the Murine Huntington Disease Gene: Evidence for Conservation, Alternate Splicing and Polymorphism in a Triplet (CCG) Repeat." *Human Molecular Genetics*. https://academic.oup.com/hmg/article-abstract/3/3/530/605583.

Linhart, Chaim, and Ron Shamir. 2007. "Degenerate Primer Design: Theoretical Analysis and the HYDEN Program." *Methods in Molecular Biology* 402: 221–44.

Liu, Qian, Peng Zhang, Depeng Wang, Weihong Gu, and Kai Wang. 2017. "Interrogating the 'Unsequenceable' Genomic Trinucleotide Repeat Disorders by Long-Read Sequencing." *Genome Medicine* 9 (1): 65.

Li, Wei, Louise C. Serpell, Wendy J. Carter, David C. Rubinsztein, and James A. Huntington. 2006. "Expression and Characterization of Full-Length Human Huntingtin, an Elongated HEAT Repeat Protein." *The Journal of Biological Chemistry* 281 (23): 15916–22.

Li, Z., C. A. Karlovich, M. P. Fish, M. P. Scott, and R. M. Myers. 1999. "A Putative Drosophila Homolog of the Huntington's Disease Gene." *Human Molecular Genetics* 8 (9): 1807–15.

Lo Sardo, Valentina, C. Zuccato, Germano Gaudenzi, Barbara Vitali, Catarina Ramos, Marzia Tartari, Michael A. Myre, et al. 2012. "An Evolutionary Recent Neuroepithelial Cell Adhesion Function of Huntingtin Implicates ADAM10-Ncadherin." *Nature Neuroscience* 15 (April): 713.

Lujan, Scott A., Alan B. Clark, and Thomas A. Kunkel. 2015. "Differences in Genome-Wide Repeat Sequence Instability Conferred by Proofreading and Mismatch Repair Defects." *Nucleic Acids Research* 43 (8): 4067–74.

MacDonald, Marcy E., Christine M. Ambrose, Mabel P. Duyao, Richard H. Myers, Carol Lin, Lakshmi Srinidhi, Glenn Barnes, et al. 1993. "A Novel Gene Containing a Trinucleotide Repeat That Is Expanded and Unstable on Huntington's Disease Chromosomes." *Cell* 72 (6): 971–83.

Mamedov, T. G., E. Pienaar, S. E. Whitney, J. R. TerMaat, G. Carvill, R. Goliath, A. Subramanian, and H. J. Viljoen. 2008. "A Fundamental Study of the PCR Amplification of GC-Rich DNA Templates." *Computational Biology and Chemistry* 32 (6): 452–57.

Martins, S., P. Coutinho, I. Silveira, and P. Giunti. 2008. "Cis-acting Factors Promoting the CAG Intergenerational Instability in Machado–Joseph Disease." *American Journal of.* https://onlinelibrary.wiley.com/doi/abs/10.1002/ajmg.b.30624.

Matsuyama, N., S. Hadano, K. Onoe, H. Osuga, J. Showguchi-Miyata, Y. Gondo, and J. E. Ikeda. 2000. "Identification and Characterization of the Miniature Pig Huntington's Disease Gene Homolog: Evidence for Conservation and Polymorphism in the CAG Triplet Repeat." *Genomics* 69 (1): 72–85.

Maynard-Smith, J. 1989. *Evolutionary Genetics*. Oxford University Press.

McMurray, Cynthia T. 2010. "Mechanisms of Trinucleotide Repeat Instability during Human Development." *Nature Reviews. Genetics* 11 (11): 786–99.

Metzgar, D., and C. Wills. 2000. "Evidence for the Adaptive Evolution of Mutation Rates." *Cell* 101 (6): 581–84.

Mier, Pablo, and Miguel A. Andrade-Navarro. 2017. "dAPE: A Web Server to Detect Homorepeats and Follow Their Evolution." *Bioinformatics* 33 (8): 1221–23.

Mier, P., and Miguel A. Andrade-Navarro. 2018. "Glutamine Codon Usage and polyQ Evolution in Primates Depend on the Q Stretch Length." *Genome Biology and Evolution* 10 (3): 816–25.

Morrison, P. J., S. Harding-Lester, and A. Bradley. 2011. "Uptake of Huntington Disease Predictive Testing in a Complete Population." *Clinical Genetics* 80 (3): 281–86.

Mouse Genome Sequencing Consortium, Robert H. Waterston, Kerstin Lindblad-Toh, Ewan Birney, Jane Rogers, Josep F. Abril, Pankaj Agarwal, et al. 2002. "Initial Sequencing and Comparative Analysis of the Mouse Genome." *Nature* 420 (6915): 520–62.

Mühlau, Mark, Juliane Winkelmann, Dan Rujescu, Ina Giegling, Nikolaos Koutsouleris, Christian Gaser, Milan Arsic, Adolph Weindl, Maximilian Reiser, and Eva M. Meisenzahl. 2012. "Variation within the Huntington's Disease Gene Influences Normal Brain Structure." *PloS One* 7 (1): e29809.

Myre, Michael A., Amanda L. Lumsden, Morgan N. Thompson, Wilma Wasco, Marcy E. MacDonald, and James F. Gusella. 2011. "Deficiency of Huntingtin Has Pleiotropic Effects in the Social Amoeba Dictyostelium Discoideum." *PLoS Genetics* 7 (4): e1002052.

Nagy, Zoltán Tamás. 2010. "A Hands-on Overview of Tissue Preservation Methods for Molecular Genetic Analyses." *Organisms, Diversity & Evolution* 10 (1): 91–105.

Nature Methods Editorial Board. 2010. "E Pluribus Unum." *Nature Methods* 7 (May): 331.

Nesse, Randolph M. 2011. "Ten Questions for Evolutionary Studies of Disease Vulnerability." *Evolutionary Applications* 4 (2): 264–77.

Nithianantharajah, Jess, and Anthony J. Hannan. 2007. "Dynamic Mutations as Digital Genetic Modulators of Brain Development, Function and Dysfunction." *BioEssays: News and Reviews in Molecular, Cellular and Developmental Biology* 29 (6): 525–35.

Nowoshilow, Sergej, Siegfried Schloissnig, Ji-Feng Fei, Andreas Dahl, Andy W. C. Pang, Martin Pippel, Sylke Winkler, et al. 2018. "The Axolotl Genome and the Evolution of Key Tissue Formation Regulators." *Nature* 554 (7690): 50–55.

Orgel, L. E., and F. H. Crick. 1980. "Selfish DNA: The Ultimate Parasite." *Nature* 284 (5757): 604–7.

Orr, H. Allen. 2009. "Fitness and Its Role in Evolutionary Genetics." *Nature Reviews. Genetics* 10 (8): 531–39.

Palidwor, Gareth A., Sergey Shcherbinin, Matthew R. Huska, Tamas Rasko, Ulrich Stelzl, Anup Arumughan, Raphaele Foulle, et al. 2009. "Detection of Alpha-Rod Protein Repeats Using a Neural Network and Application to Huntingtin." *PLoS Computational Biology* 5 (3): e1000304.

Partridge, Linda, and David Gems. 2002. "Mechanisms of Ageing: Public or Private?" *Nature Reviews. Genetics* 3 (3): 165–75.

Pearson, Christopher E. 2003. "Slipping While Sleeping? Trinucleotide Repeat Expansions in Germ Cells." *Trends in Molecular Medicine* 9 (11): 490–95.

Pearson, Christopher E., Kerrie Nichol Edamura, and John D. Cleary. 2005. "Repeat Instability: Mechanisms of Dynamic Mutations." *Nature Reviews. Genetics* 6 (10): 729–42.

Pennisi, Elizabeth. 2017. "Sequencing All Life Captivates Biologists." *Science* 355 (6328): 894–95.

Prüfer, Kay, Fernando Racimo, Nick Patterson, Flora Jay, Sriram Sankararaman, Susanna Sawyer, Anja Heinze, et al. 2014. "The Complete Genome Sequence of a Neanderthal from the Altai Mountains." *Nature* 505 (7481): 43–49.

Reich, David, Richard E. Green, Martin Kircher, Johannes Krause, Nick Patterson, Eric Y. Durand, Bence Viola, et al. 2010. "Genetic History of an Archaic Hominin Group from Denisova Cave in Siberia." *Nature* 468 (7327): 1053–60.

Richards, R. I. 2001. "Dynamic Mutations: A Decade of Unstable Expanded Repeats in Human Genetic Disease." *Human Molecular Genetics* 10 (20): 2187–94.

Richards, R. I., and G. R. Sutherland. 1994. "Simple Repeat DNA Is Not Replicated Simply." *Nature Genetics* 6 (2): 114–16.

Roberts, Richard J., Mauricio O. Carneiro, and Michael C. Schatz. 2013. "The Advantages of SMRT Sequencing." *Genome Biology* 14 (7): 405.

Rubinsztein, D. C., W. Amos, J. Leggo, S. Goodburn, S. Jain, S. H. Li, R. L. Margolis, C. A. Ross, and M. A. Ferguson-Smith. 1995. "Microsatellite Evolution--Evidence for Directionality and Variation in Rate between Species." *Nature Genetics* 10 (3): 337–43.

Rubinsztein, D. C., W. Amos, J. Leggo, S. Goodburn, R. S. Ramesar, J. Old, R. Bontrop, R. McMahon, D. E. Barton, and M. A. Ferguson-Smith. 1994. "Mutational Bias Provides a Model for the Evolution of Huntington's Disease and Predicts a General Increase in Disease Prevalence." *Nature Genetics* 7 (4): 525–30.

Schatz, Michael C., Arthur L. Delcher, and Steven L. Salzberg. 2010. "Assembly of Large Genomes Using Second-Generation Sequencing." *Genome Research* 20 (9): 1165–73.

Schmitt, I., D. Bächner, D. Megow, P. Henklein, H. Hameister, J. T. Epplen, and O. Riess. 1995. "Expression of the Huntington Disease Gene in Rodents: Cloning the Rat Homologue and Evidence for Downregulation in Non-Neuronal Tissues during Development." *Human Molecular Genetics* 4 (7): 1173–82.

Semaka, A., C. Kay, C. N. Doty, and J. A. Collins. 2013. "High Frequency of Intermediate Alleles on Huntington Disease-associated Haplotypes in British Columbia's General Population." *American Journal of*. https://onlinelibrary.wiley.com/doi/abs/10.1002/ajmg.b.32193.

Shao, Wen, Sonny Khin, and William C. Kopp. 2012. "Characterization of Effect of Repeated Freeze and Thaw Cycles on Stability of Genomic DNA Using Pulsed Field Gel Electrophoresis." *Biopreservation and Biobanking* 10 (1): 4–11.

Shapiro, J. A. 2011. *Evolution: A View from the 21st Century*. Upper Saddle River, NJ FT Press.

Slatko, B. E., L. M. Albright, S. Tabor, and J. Ju. 2001. "DNA Sequencing by the Dideoxy Method." *Current Protocols in Molecular Biology / Edited by Frederick M. Ausubel ... [et Al.]* Chapter 7 (May): Unit7.4A.

Steffan, Joan S., Namita Agrawal, Judit Pallos, Erica Rockabrand, Lloyd C. Trotman, Natalia Slepko, Katalin Illes, et al. 2004. "SUMO Modification of Huntingtin and Huntington's Disease Pathology." *Science* 304 (5667): 100–104.

Strong, T. V., D. A. Tagle, J. M. Valdes, L. W. Elmer, K. Boehm, M. Swaroop, K. W. Kaatz, F. S. Collins, and R. L. Albin. 1993. "Widespread Expression of the Human and Rat Huntington's Disease Gene in Brain and Nonneural Tissues." *Nature Genetics* 5 (3): 259–65.

Takano, Hiroki, and James F. Gusella. 2002. "The Predominantly HEAT-like Motif Structure of Huntingtin and Its Association and Coincident Nuclear Entry with Dorsal, an NF-kB/Rel/dorsal Family Transcription Factor." *BMC Neuroscience* 3 (October): 15.

Tartari, Marzia, Carmela Gissi, Valentina Lo Sardo, C. Zuccato, Ernesto Picardi, Graziano Pesole, and Elena Cattaneo. 2008. "Phylogenetic Comparison of Huntingtin Homologues Reveals the Appearance of a Primitive polyQ in Sea Urchin." *Molecular Biology and Evolution* 25 (2): 330–38.

Tautz, D., and Schlötterer. 1994. "Simple Sequences." *Current Opinion in Genetics & Development* 4 (6): 832–37.

Teeling, Emma C., Sonja C. Vernes, Liliana M. Dávalos, David A. Ray, M. Thomas P. Gilbert, Eugene Myers, and Bat1K Consortium. 2018. "Bat Biology, Genomes, and the Bat1K Project: To Generate Chromosome-Level Genomes for All Living Bat Species." *Annual Review of Animal Biosciences* 6 (February): 23–46.

Tóth, G., Z. Gáspári, and J. Jurka. 2000. "Microsatellites in Different Eukaryotic Genomes: Survey and Analysis." *Genome Research* 10 (7): 967–81.

Treangen, Todd J., and Steven L. Salzberg. 2011. "Repetitive DNA and next-Generation Sequencing: Computational Challenges and Solutions." *Nature Reviews. Genetics* 13 (1): 36–46.

Tyner, Cath, Galt P. Barber, Jonathan Casper, Hiram Clawson, Mark Diekhans, Christopher Eisenhart, Clayton M. Fischer, et al. 2017. "The UCSC Genome Browser Database: 2017 Update." *Nucleic Acids Research* 45 (D1): D626–34.

Venter, J. C., M. D. Adams, E. W. Myers, P. W. Li, R. J. Mural, G. G. Sutton, H. O. Smith, et al. 2001. "The Sequence of the Human Genome." *Science* 291 (5507): 1304–51.

Wang, Yu, Paul A. Steimle, Yixin Ren, Christopher A. Ross, Douglas N. Robinson, Thomas T. Egelhoff, Hiromi Sesaki, and Miho Iijima. 2011. "Dictyostelium Huntingtin Controls Chemotaxis and Cytokinesis through the Regulation of Myosin II Phosphorylation." *Molecular Biology of the Cell* 22 (13): 2270–81.

Warby, S. C., A. Montpetit, A. R. Hayden, J. B. Carroll, S. L. Butland, H. Visscher, J. A. Collins, A. Semaka, T. J. Hudson, and M. R. Hayden. 2009. "CAG Expansion in the Huntington Disease Gene Is Associated with a Specific and Targetable Predisposing Haplogroup." *American Journal of Human Genetics* 84 (3): 351–66.

Warby, S. C., H. Visscher, S. Butland, C. E. Pearson, and M. R. Hayden. 2009. "Response to Falush: A Role for Cis-Element Polymorphisms in HD." *American Journal of Human Genetics* 85 (6): 942–45.

Warby, Simon C., Henk Visscher, Jennifer A. Collins, Crystal N. Doty, Catherine Carter, Stefanie L. Butland, Anna R. Hayden, Ichiro Kanazawa, Colin J. Ross, and Michael R. Hayden. 2011. "HTT Haplotypes Contribute to Differences in Huntington Disease Prevalence between Europe and East Asia." *European Journal of Human Genetics: EJHG* 19 (5): 561–66.

Warner, J. P., L. H. Barron, and D. J. Brock. 1993. "A New Polymerase Chain Reaction (PCR) Assay for the Trinucleotide Repeat That Is Unstable and Expanded on Huntington's Disease Chromosomes." *Molecular and Cellular Probes* 7 (3): 235–39.

Watson, James D. 2014. *Molecular Biology of the Gene*. Pearson.

Weber, J. L., and C. Wong. 1993. "Mutation of Human Short Tandem Repeats." *Human Molecular Genetics* 2 (8): 1123–28.

Weydt, P., S. M. Soyal, C. Gellera, S. DiDonato, C. Weidinger, H. Oberkofler, B. Landwehrmeyer, and W. Patsch. 2009. "The Gene Coding for PGC-1α Modifies Age at Onset in Huntington's Disease." *Aktuelle Neurologie* 36 (S 02): P765.

Wheeler, V. C., W. Auerbach, J. K. White, J. Srinidhi, A. Auerbach, A. Ryan, M. P. Duyao, et al. 1999. "Length-Dependent Gametic CAG Repeat Instability in the Huntington's Disease Knock-in Mouse." *Human Molecular Genetics* 8 (1): 115–22.

Wheeler, V. C., F. Persichetti, S. M. McNeil, J. S. Mysore, S. S. Mysore, M. E. MacDonald, R. H. Myers, J. F. Gusella, N. S. Wexler, and US-Venezuela Collaborative Research Group. 2007. "Factors Associated with HD CAG Repeat Instability in Huntington Disease." *Journal of Medical Genetics* 44 (11): 695–701.

Wiener, W. J., and A. E. Lang. 1989. *Movement Disorders: A Comprehensive Survey*. Mount. Kisco, New York.

Wilson, Don E., and Deeann M. Reeder. 2005. *Mammal Species of the World: A Taxonomic and Geographic Reference*. JHU Press.

Xia, Jianrun, Denise H. Lee, Jillian Taylor, Mark Vandelft, and Ray Truant. 2003. "Huntingtin Contains a Highly Conserved Nuclear Export Signal." *Human Molecular Genetics* 12 (12): 1393–1403.

Xu, X., M. Peng, and Z. Fang. 2000. "The Direction of Microsatellite Mutations Is Dependent upon Allele Length." *Nature Genetics* 24 (4): 396–99.

Yanai, Anat, Kun Huang, Rujun Kang, Roshni R. Singaraja, Pamela Arstikaitis, Lu Gan, Paul C. Orban, et al. 2006. "Palmitoylation of Huntingtin by HIP14 Is Essential for Its Trafficking and Function." *Nature Neuroscience* 9 (6): 824–31.

Yang, Shang-Hsun, Pei-Hsun Cheng, Heather Banta, Karolina Piotrowska-Nitsche, Jin-Jing Yang, Eric C. H.

Cheng, Brooke Snyder, et al. 2008. “Towards a Transgenic Model of Huntington’s Disease in a Non-Human Primate.” *Nature* 453 (7197): 921–24.

Ye, Liang, Ladeana W. Hillier, Patrick Minx, Nay Thane, Devin P. Locke, John C. Martin, Lei Chen, et al. 2011. “A Vertebrate Case Study of the Quality of Assemblies Derived from next-Generation Sequences.” *Genome Biology* 12 (3): R31.

Zhang, Guojie, Cai Li, Qiye Li, Bo Li, Denis M. Larkin, Chul Lee, Jay F. Storz, et al. 2014. “Comparative Genomics Reveals Insights into Avian Genome Evolution and Adaptation.” *Science* 346 (6215): 1311–20.

Zhang, Guojie, Carsten Rahbek, Gary R. Graves, Fumin Lei, Erich D. Jarvis, and M. Thomas P. Gilbert. 2015. “Genomics: Bird Sequencing Project Takes off.” *Nature* 522 (7554): 34.

Zuccato, C., and E. Cattaneo. 2014. “The Normal Function of Huntingtin.” In *Huntington’s Disease*, edited by G. Bates, S. Tabrizi, and L. Jones. Oxford University Press.

Zuccato, C., Marta Valenza, and Elena Cattaneo. 2010. “Molecular Mechanisms and Potential Therapeutical Targets in Huntington’s Disease.” *Physiological Reviews* 90 (3): 905–81.