Scaling up reference quality assembly of vertebrate genomes

Shane A. McCarthy[1,2], Iliana Bista[1,2], Dirk-Dominik Dolle[1], Francesca Giordano[1], William Chow[1], Petr Danecek[1], Hannes Svardal[1,2], Milan Malinsky[1,3], Jingtao Lilue[1], Michelle Smith[1], Kim Judge[1], Karen Oliver[1], Mike Quail[1], Zemin Ning[1], Kerstin Howe[1], Richard Durbin[1,2]

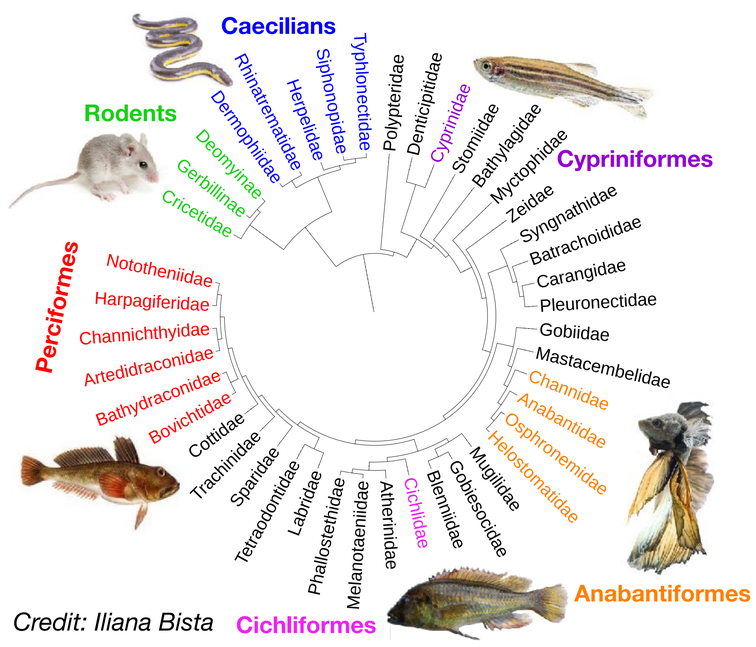
[1]Wellcome Trust Sanger Institute, Hinxton, CB10 1SA, UK [2]Department of Genetics, University of Cambridge, Cambridge, CB2 3EH, UK
[3]Zoological Institute, Department of Environmental Sciences, University of Basel, 4051 Basel, Switzerland

Increased read length and throughput of long read sequencing technologies such as PacBio and Oxford Nanopore are now enabling high contiguity and accuracy de novo reference assemblies for vertebrate scale genomes. During 2017 the Wellcome Trust Sanger Institute has undertaken a pilot project aiming to obtain **reference quality genome sequences** for 50-100 species and additional genome data for a similar number of related species. Our primary focus is on **fish genomes**, but we are also sequencing some **amphibian** and **mammalian** species. This initiative is part of the broader international **Vertebrate Genomes Project (VGP)** being led by the Genome 10k Consortium in collaboration with a number of other genome sequencing projects. The initial aim of the VGP is to obtain reference quality genome assemblies for **one species from every vertebrate order** over the next year or two.

- ∼**30 representatives of fish orders** for which adequate references do not currently exist.
- multiple samples (6 references plus up to 10-20 others) within each of the following fish groups:
  - **cyprinids** related to the zebrafish *Danio rerio*.
  - **cichlid** fish part of the major cichlid evolutionary radiations; recently completed sequencing of >1000 genomes using Illumina HiSeqX.
  - **notothenioid** fish from the Antarctic radiation.
  - **anabantoid** fish including gouramis.
- several **caecilian** species representing the most basal branch of amphibians.
- multiple **rodents** with extreme phenotypes providing evolutionary context to mice and rats.

Project overview

Species targetted as part of this project are summarised in the box to the left and in Figure 1. To collect these samples we are collaborating with multiple members of the Genome10k community and other evolutionary researchers.



Credit: Iliana Bista

Figure 1: Highlighted are groups where we will sequence additional species for more in depth analyses.

We are currently using **Pacific Biosciences Sequel**, **10X Genomics Chromium** and **BioNano Saphyr** as core technologies within existing production and R&D pipelines at the Sanger Institute. We are evaluating others technologies including **Oxford Nanopore**. For the VGP orders, we will also generate HiC data to aid chromosome assignment and polishing.

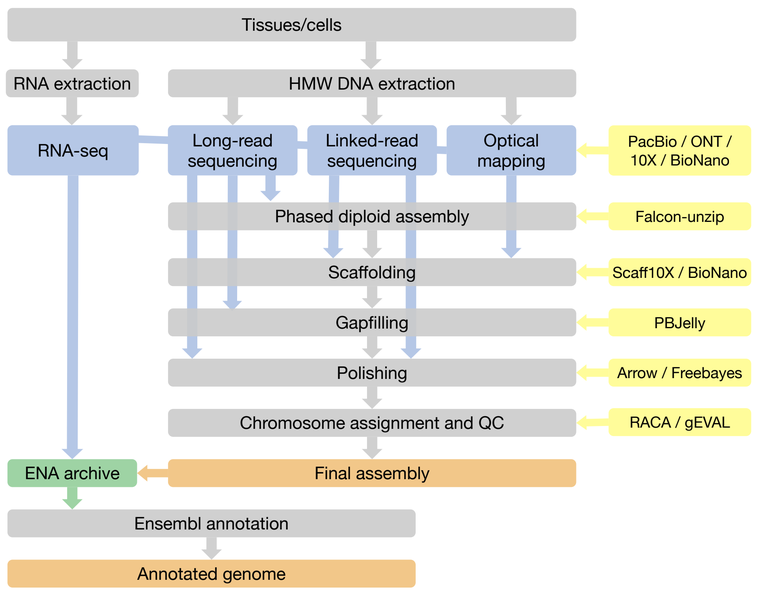


Figure 2: Data production at the Sanger for long-read reference quality de novo assemblies. We are collaborating with the European Bioinformatics Institute (EBI) to streamline data deposition into the relevant archives to enable efficient gene annotation and presentation in Ensembl.

10X Genomics Chromium linked-read QC

Credit: Petr Danecek and Zemin Ning

We are using 10X Genomics Chromium linked-read data for scaffolding and polishing (for details about scaffolding, see Zemin Ning's poster about **scaff10x** here at #gi2017). To measure the quality of the 10X data, we are developing a tool, **bxcheck**, to produce summary QC data. Plots below show some examples of data we have generated so far for this project. Scaffolding is most dependent on having a large number of fragments of a good length ($>10^5$) to give enough information to join contigs reliably.

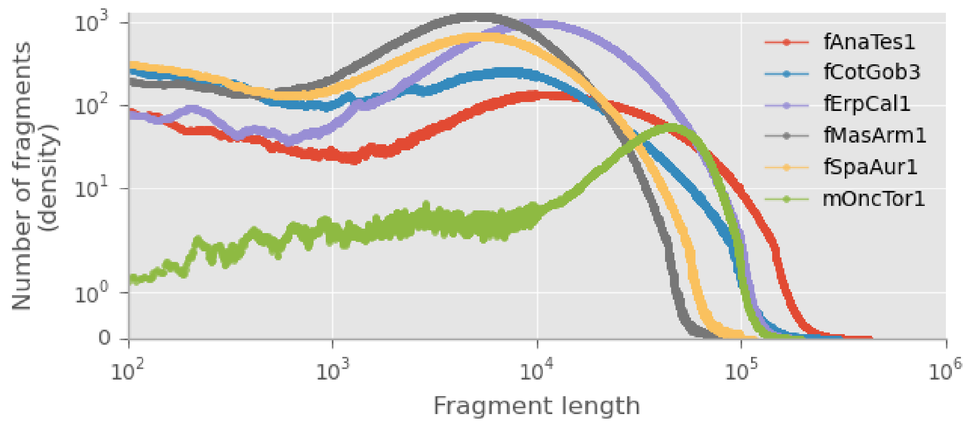


Figure 3: Density of fragments at longest fragment length (molecule length).

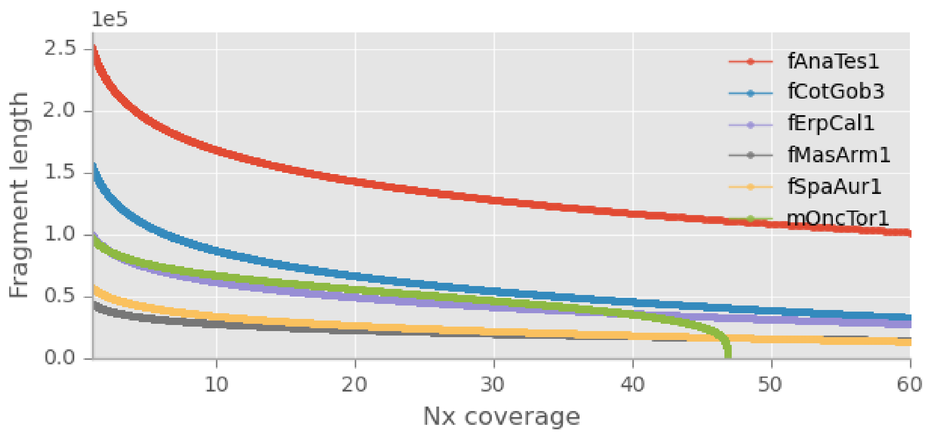


Figure 4: Longest fragment at Nx coverage.

www.github.com/pd3/bxcheck

3.4.2q40 assembly standard

For the VGP ordinal project and other "reference quality" assemblies, we are aiming for a **3.4.2q40** standard consisting of:

3	>1Mb contig N50	4	>10Mb scaffold N50
2	Chromosome assignment of >90% through synteny or genetic maps, where possible	q40	Average base quality >Q40

Assembly progress

We have reached our contiguity, scaffolding and base quality goals for 3 of our species so far using just **PacBio** and **10X** data. Other species will benefit from additional orthogonal data from **BioNano Saphyr** and HiC.

VGP ID	Species	CtgN50 (Mb)	#Ctgs	ScaffN50 (Mb)	#Scaffs	Length (Mb)	
fAnaTes1	***Anabas testudineus,* climbing perch**	3.49	1,150	16.56	612	573	Falcon, scaff10x, gap-fill, polish
fAstCal1	*Astatotilapia calliptera* Eastern Happy	3.33	914	11.27	352	883	Falcon-unzip, scaff10x, gap-fill, polish
fCotGob3	***Cottoperca gobio,* Channel bull blenny**	1.23	1,561	3.96	753	598	Falcon, scaff10x
fErpCal1	***Erpetoichthys calabaricus,* reedfish**	0.925	12,813	1.26	10,860	3,478	Falcon, scaff10x
fGouWil2	***Gouania willdenowi,* blunt-snouted clingfish**	1.36	1,833	8.76	535	938	Falcon-unzip, scaff10x
fMasArm1	***Mastacembelus armatus,* tire track eel**	6.25	349	12.75	196	578	Falcon-unzip, scaff10x, gap-fill, polish
fSimDia1	*Simochromis diagramma,* Tanganyika cichlid	1.81	2,827	6.16	1,593	864	Falcon, scaff10x, gap-fill, polish
fSpaAur1	***Sparus aurata,* gilt head sea bream**	1.39	3,455	-	-	877	Falcon-unzip
fZeuFab1	***Zeus faber,* John Dory**	0.349	9,677	-	-	943	Falcon-unzip
mAcoRus1	*Acomys russatus,* golden spiny mouse	0.365	13,155	16.39	6,065	2,416	Falcon, BioNano Saphyr hybrid scaffolding
mOncTor1	*Onychomys torridus,* southern grasshopper mouse	1.00	7,678	-	-	2,401	Falcon

Table 1: Current status of various PacBio assemblies, showing contig N50, number of contigs, scaffold N50, number of scaffolds and assembly size. Species in bold are those being sequenced as part of the VGP ordinal level project.

Acknowledgements: Byrappa Venkatesh, Maximilian Wagner **Caecilians**: Mark Wilkinson; **Cichlids**: George Turner, Martin Genner, Axel Meyer, Walter Salzburger, Milan Malinsky, Alexandra Tyers, Antonia Ford, Craig Albertson; **Cyprinids**: Lukas Ruber, Ralf Britz, Braedan McCluskey, Andrew Whiteley, Bill Trevarrow, Uwe Irion, Elisabeth Busch-Nentwich, Christiane Nuesslein Vollhard; **Notothenioids**: Melody Clark, Christina Cheng, Bill Detrich, John Postlethwait, Thomas Desvignes; **Rodents**: David Thybert, Thomas Keane.

www.sanger.ac.uk/science/data/vertebrate-genomes-project sm15@sanger.ac.uk @mcshane