Plans for scaling vertebrate genome sequencing and assembly at the Sanger Institute

Shane A. McCarthy[1], Marcus Klarqvist[1], Milan Malinsky[1,2], Dirk-Dominik Dolle[1], Hannes Svardal[1], Karen Oliver[1], Michelle Smith[1], Kim Judge[1], William Chow[1], Mike Quail[1], Kerstin Howe[1], Thomas Keane[3], Richard Durbin[1]

[1]Wellcome Trust Sanger Institute, Hinxton, Cambridgeshire, CB10 1SA, UK
[2]Zoological Institute, Dept. of Environmental Sciences, University of Basel, 4051 Basel, Switzerland
[3]European Bioinformatics Institute, Hinxton, Cambridgeshire, CB10 1SD, UK

During 2016-17 the Wellcome Trust Sanger Institute is undertaking a pilot vertebrate genome sequencing project, with the aim 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 planning to sequence some **amphibian** and **mammalian** species.

Project overview

During 2016-17 the Wellcome Trust Sanger Institute is undertaking a pilot vertebrate genome sequencing project, with the aim 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 planning to sequence some **amphibian** and **mammalian** species. Specific targets include:

- ∼**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*
  - **cichliforms** related to the major haplochromine cichlid evolutionary radiation
  - **notothenioid** fish from the Antarctic radiation
  - **anabantoid** fish include gouramis
- several **caecilian** species representing the most basal branch of amphibians
- **multiple rodents** with extreme phenotypes providing evolutionary context to mice and rats

To collect these samples we are collaborating with multiple members of the Genome10K community and other evolutionary researchers. We are currently using **Pacific Biosciences Sequel**, **10X Genomics Chromium** and **BioNano Irys** as core technologies, and evaluating others including **Oxford Nanopore** and **Dovetail**. Our long term aim is to scale up to sequence hundreds then thousands of new species per year.

Project data flow

Data generation and production is being coordinated at the Sanger Insitute, tying in with existing production, R&D and analysis pipelines. 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.

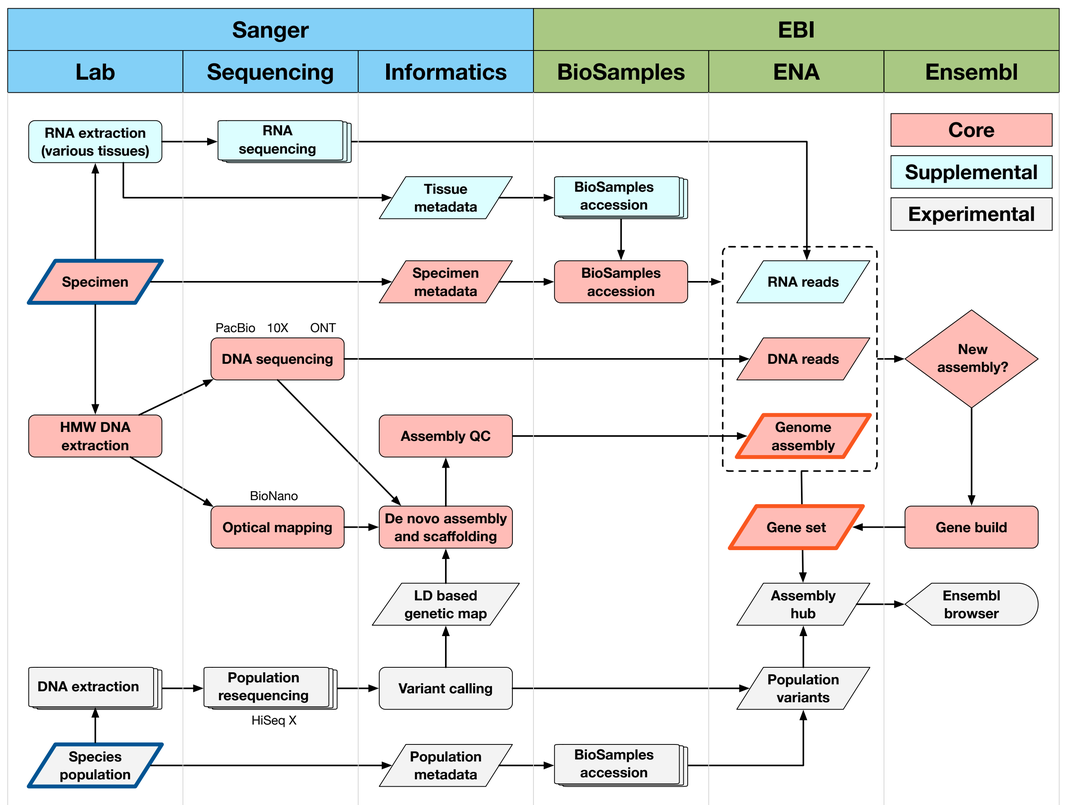

Figure 1: Flow diagram for data production for the Vertebrate Genomes Project.

Assembly scaffolding using linkage disequilibrium

Credit: Marcus Klarqvist

A key problem in reference genome assembly is to connect together initial contigs or scaffolds, which may have typical size 100kb-10Mb, into chromosome-scale contiguous sequences. As an alternative to physically based methods that use long range restriction maps, read pairs or read sets, we have been investigating using patterns of linkage disequilibrium (LD) between SNPs obtained by low coverage population sequencing. Here we show preliminary results from experiments to re-assemble human chromosomes, including a reconstituted map of chromosome 20 and examples of chains of nearby loci in high LD. We are currently exploring applying this approach to scaffold the *Astatotilapia calliptera* assembly.

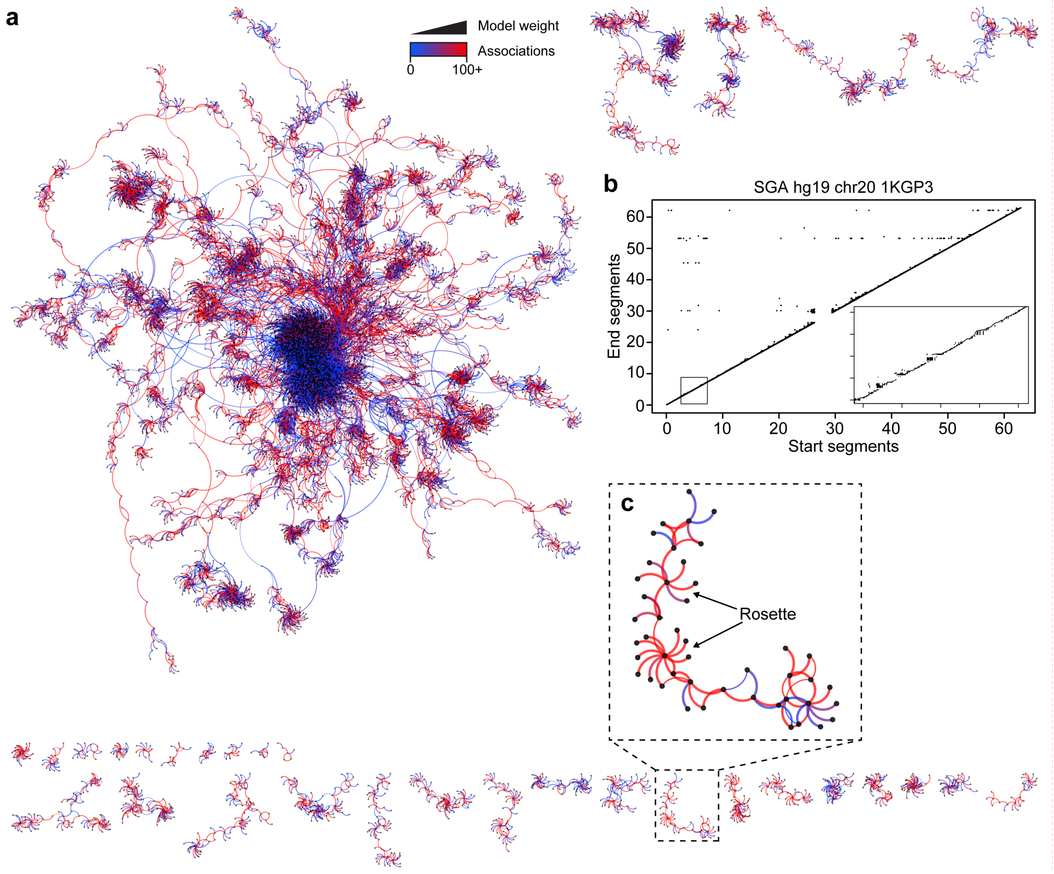

Figure 2: An SGA assembly was made using 30× Illumina HiSeq X data for NA12878. **(a)** Graph nodes are the contigs from this SGA assembly with edges indicating an LD association between contigs. The colour of the edges indicates the number of SNPs in association, while the width of the edges indicates the strength of the association. **(b)** For chromosome 20, we compare the predicted position of the contigs with the truth showing very good agreement. **(c)** Details of one of the connected components. A rosette is indicative of a number of smaller contigs contained within a larger contig.

Preliminary cichlid assemblies

Credit: Milan Malinsky

We have recently completed preliminary assemblies for two cichlid species *Astatotilapia calliptera* (fAstCal1, Figure 3) and *Simochromis diagramma* (fSimDia1). We reached a **contig N50 of 2.47Mb (fAstCal1) and 1.25Mb (fSimDia1)**, with **assembly sizes of 890Mb and 850Mb** respectively, in agreement with expectations. N50 of fAstCal1 was increased to **4.2Mb after scaffolding** with BioNano Irys data and **4.5Mb after scaffolding** with Dovetail Chicago data.

Figure 3: Eastern happy *Astatotilapia calliptera* from Lake Masoko, Tanzania.

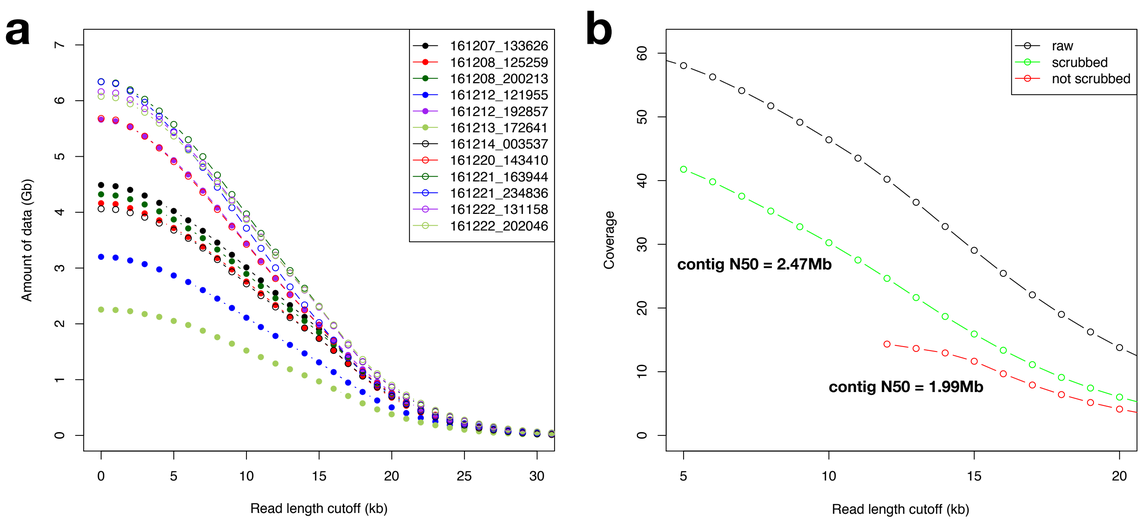

Figure 4: (a) Read length distribution per SMRT cell shows improvement in yield over time. **(b)** Read scrubbing with Gene Myers' Dazzler pipeline, although it initially removes material and breaks chimeric reads, results in higher coverage after error correction and a resulting improved assembly.

Contact **sm15@sanger.ac.uk** **@mcshane**