The tomato genome sequence

The GBF Bioinformatics and Biostatistics Team : Eli Maza, Anis Djari, Margo Zahm, Clément Folgoas

Leader: Mohamed Zouine

The goal of the GBF-Bioinformatics team is to generate digital tools and resources for the need of all GBF projects.

Through Local, National, EU and International collaborations, significant and useful tools and resources for the tomato community have been generated.

The Tomato Genome Sequence
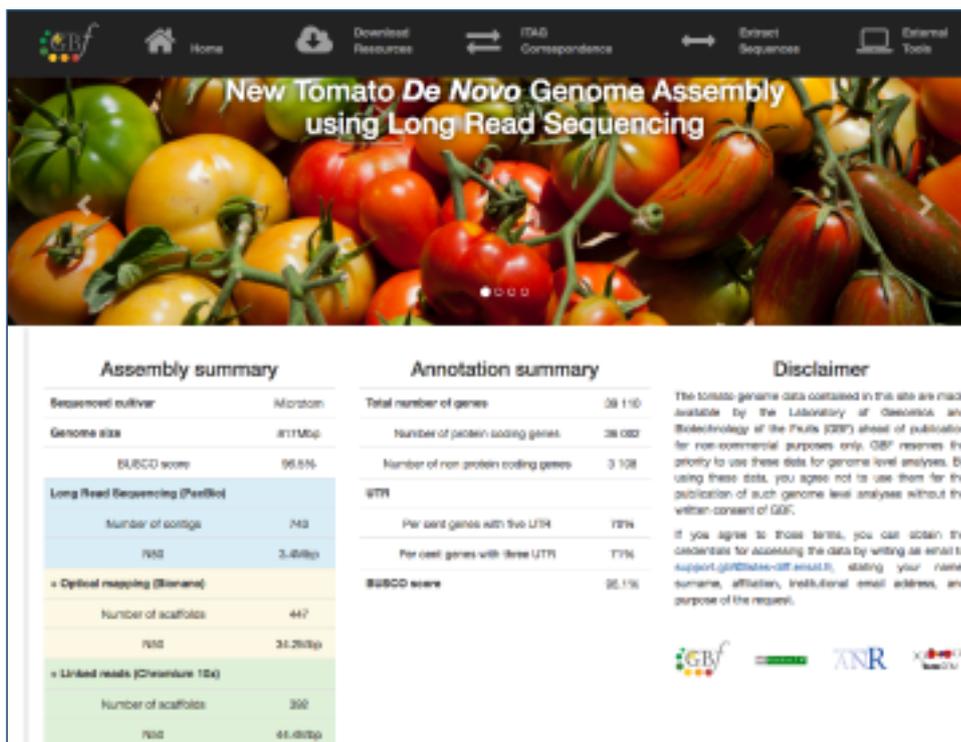
The new genome assembly and annotation are accessible   here

Recently, GBF lab have designed and funded throug French (ANR) and EU programs a project to improve the actual tomato reference genome sequence by long read sequencing technologies.

 The integration of these three approaches allowed to reach a genome size of **~830 Mb** with an **N50 of 45 Mb**. The assembly contiguity reached chromosome-arm-levels. Also one full chromosome (Ch12) has been fully assembled in one scaffold. The integration of the genetic map allowed to generate the 12 pseudomolecules corresponding to the 12 tomato

chromosomes.Several regions corresponding to chromosome zero in the SL3.0 reference genome were included in the current assembly.

This new reference genome was annotated using Eugen-EP giving a high busco score >96%.

Previous assembly
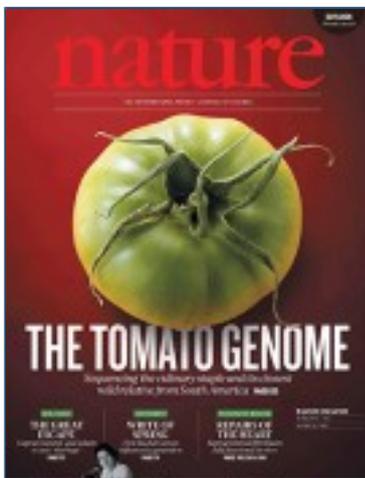
The first version of the tomato reference genome has been published in Nature on May 31, 2012, culminating years of work by the Tomato Genome Consortium, a multi-national team of scientists from 14 countries.

The GBF lab was actively involved in all steps of its production and mainly in the genome assembly part.
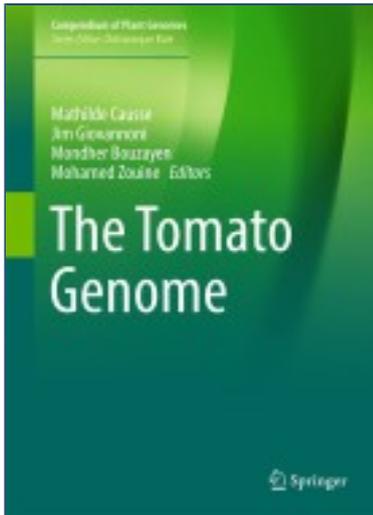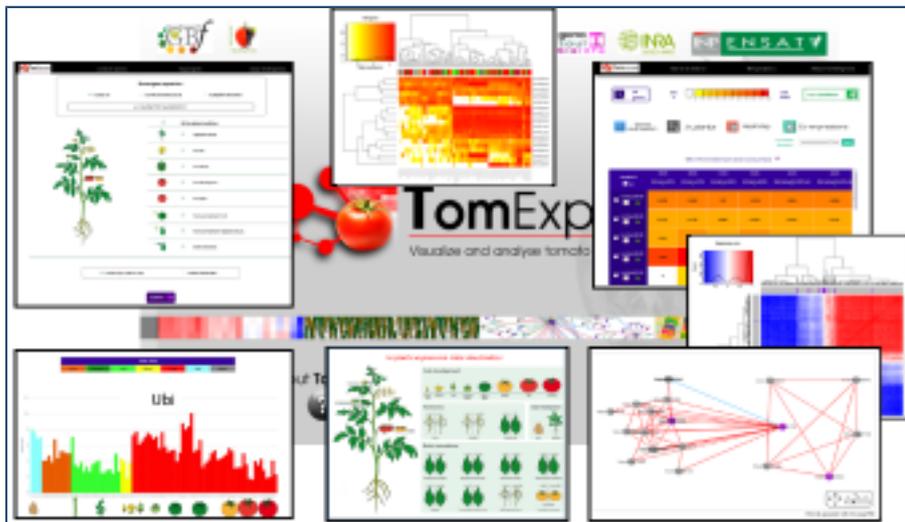
The whole story of the generation of this first sequence as a reference genome for tomato is detailed in this book co-edited by the GBF Lab

The TomExpress Platform

 TomExpress: a unified tomato RNA-Seq resource providing a tool for visualization of expression data, clustering and correlation networks

The TomExpress RNA-Seq resource was developed to provide the tomato community with a dedicated browser and tools for public RNA-seq data handling, visualization and mining. To avoid major biases resulting from the use of different mapping and statistical processing in each project, all RNA-Seq row sequence data available in public database (EMBL-EBI-ENA) were mapped *de novo* on a unique tomato reference genome using the most popular mapping software (TopHat2-Bowtie) and accurate mapping parameters. Following calculation of the number of counts per gene in each RNA-seq project, the same normalization method was applied to all counts available for the tomato as described in Maza el al 2012. This unifies the whole set of expression data and makes them comparable.

A database has been designed where each expression value is associated with the corresponding experimental details. To make the data searchable, a friendly web interface has been developed which also provides versatile data mining tools . These include output graphics showing histograms, heat-maps of hierarchically clustered expression data and identification and visualization of correlation networks of co-regulated gene groups