Universidad de los Andes
Facultad de Ciencias

Data Mining of the Coffee Rust Genome

David Botero-Rozo, William Giraldo, Álvaro Gaitán, Marco Cristancho, Diego M. Riaño-Pachón & Silvia Restrepo.

E-Mail: do.botero29@uniandes.edu.co

URLs: http://lamfu.uniandes.edu.co/
http://bce.uniandes.edu.co/
http://bioinformatics.cenicafe.org/index.php/wiki

WHY DO WE STUDY COFFEE RUST?

Coffee leaf rust (caused by the fungus *Hemileia vastatrix*) is the most limiting disease wherever coffee is cultivated. In Colombia, coffee represents 16% of the country's agricultural GDP and since 2008 high incidence of coffee leaf rust in crops established with susceptible varieties has caused significant reduction in yield. Genome and transcriptome sequencing and bioinformatics analysis tools have been applied for the understanding of this organism and its interaction with the plant and the environmental variations that result in epidemics.

WHAT DID WE GET?

A total of 73GB of NGS data was generated. An assembly of 396.264 contigs (N50 of 1590 and 841 of mean length) was obtained; this assembly is highly fragmented. Nevertheless, we obtained very large contigs (the largest of 85Kb and coverage of 148x). After filtering out contigs with putative contaminants with MEGAN (coffee and bacterial sequences that could be into the rust samples), we obtained 31.376 contigs that showed similarities to reported fungal sequences. Forty four putative mitochondrial contigs were identified through blast homologies using *Puccinia* genome sequences. We also assembled three transcriptomes with a length of 55,791 and 64,752 contigs for non-normalized libraries and 44,297 contigs for a normalized library.

WHAT DID WE DO?

We sampled isolates from field crops and collected eight different races of *H. vastatrix*. One isolate per race was sequenced by Illumina and 454 technologies. The data was subjected to pre-quality control with FastQC and was cleaned. After testing several assemblers with different combinations of data, we decided to do a hybrid assembly using the CLC assembler. The assembly was analyzed with MEGAN to evaluate the level of contamination and do a first approximation to the biological communities associated to *H. vastatrix* on the coffee leaf. BLAST was used to search the *H. vastatrix* mitochondria, comparing our contigs against the Puccinia mitochondrial genome (the closest sequenced organism to *H. vastatrix*). The mitochondrial contigs identified were annotated with MAKER. Finally, using Trinity we assembled three transcriptomes from different races of *H. vastatrix*.

This is the first approach to study the genome of the causal agent of the coffee rust, *H. vastatrix*. Further data mining will allow the identification of virulence and aggressiveness factors, important in the characterization of new races of the pathogen and the detection of isolates that might infect resistant coffee varieties.

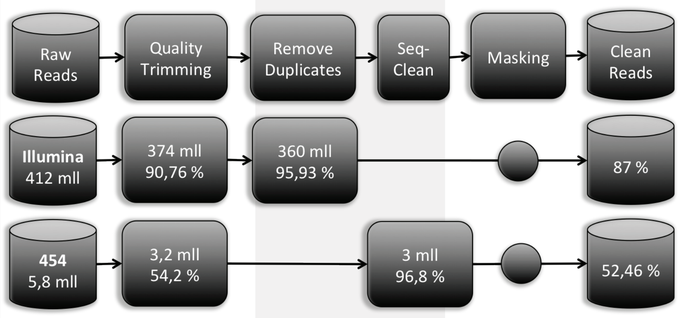

Figure 1. Data cleaning. Here the process of cleaning is illustrated. Top: Every cleaning step is shown. Middle and bottom: Illumina and 454 inputs and outputs reads given in million of reads and fraction of reads with references to the last step. Note that Illumina reads were not subjected to seq-clean tool (too short for low complexity filtering). The 454 reads were not subjected to duplication reads removal. Both data sets were subjected to low complexity masking.

Ensamblaje		454	Illumina	Hybrid I	Hybrid II	Hybrid III
Nº Sequences		96620	533113	539797	552497	396264
Residues		42983726	39939422401	340157447	331278615	333481311
Lenght	Max	13171	73636	73632	65139	85126
	Average	444,87	601,91	630,16	599,6	841,56
	N50	476	1157	1199	1030	1590
Reads	Total	2256160	374326150	376582310	335831208	336649188
	Unassembled	253786	9840637	9380494	17439547	19788611
	Assembled	2002374	364485513	367201816	318391661	316860577
	Multihit	82701	38766177	44468504	37896198	37520793
	Potential pairs		187163075	187163075	166787524	
	Paired		40391454	42942190	39253515	78105740
	Not Paired		146771621	144220885	127534009	255469308

Figure 2. Assembly results. The results of five different assemblies are shown. The first and second hybrid assemblies differ by the duplicate removal and low complexity masking. In the third assembly we added two plates of 454. In the last assembly it is shown that including 454 data reduce the number of contigs and increase the contig size.
The reads mapping results are shown: multi hit reads, potential pairs, not paired and successful pairs.

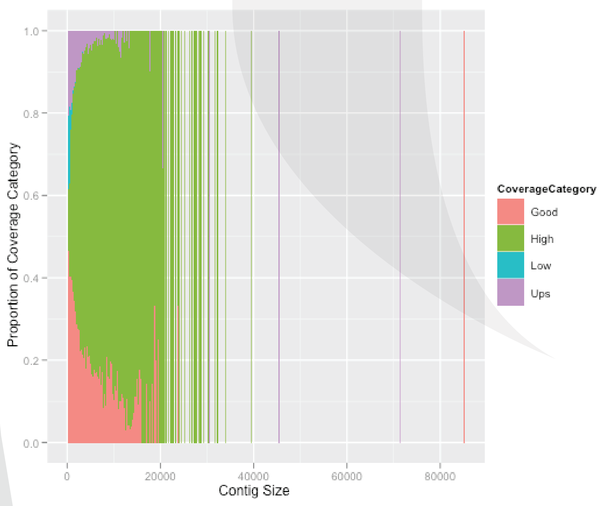

Figure 3. Contigs size fraction of coverage. Results of coverage for the last assembly. Low means coverage smaller than 5x, Good is coverage between 5x and 45x, High is coverage between 45x and 100x and Ups is coverage higher than 100x, these categories are quite arbitrary. The bins for the contig size are 150bp wide. The average coverage of a considerable fraction of the contigs of this assembly is acceptable (those with High and Good values in the Figure).

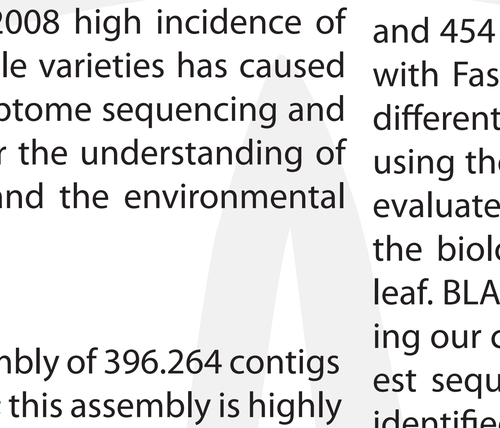

Figure 4. Distribution of number of reads in contigs. Most contigs have few reads. The assembly is very fragmented.

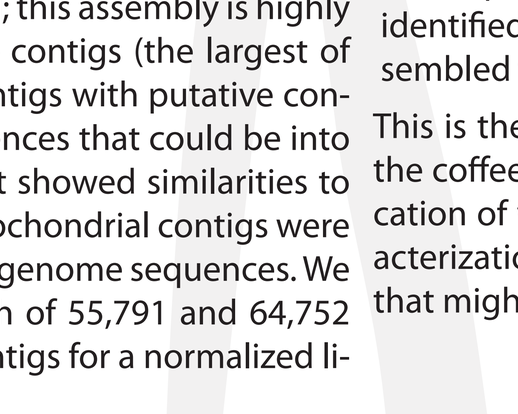

Figure 5. Distribution of contig size. Most contigs are very small.

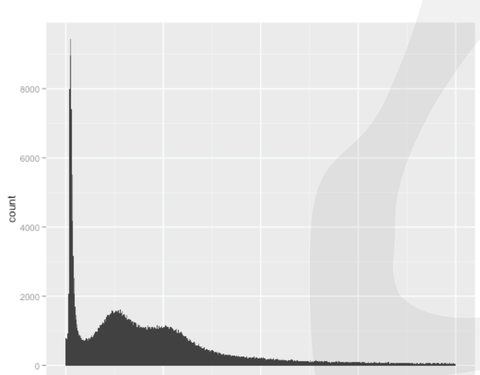

Figure 6. Distribution of coverage per contig. Most contigs have a reasonable coverage value.

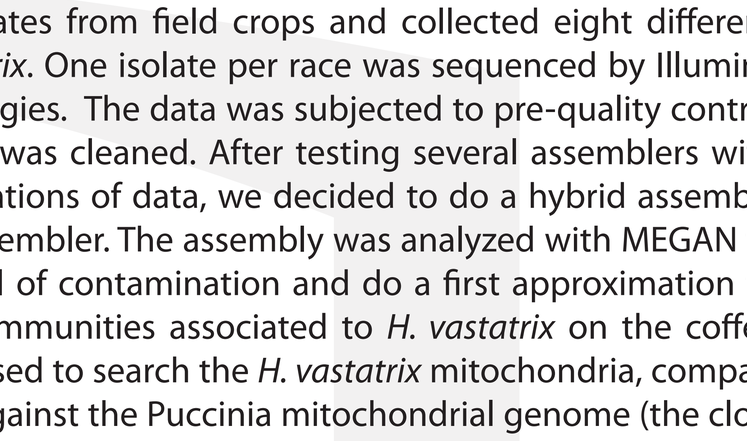

Figure 7. Taxonomy distribution of data (Megan - Reduced View). We ran blast of all contigs from the third assembly against nr database. Results were loaded in Megan to filter out contaminants: Viridiplantae and Bacteria. We found that almost all contigs hit a fungi sequence, showing that the *H. vastatrix* sequence data had very little contamination from other organisms.

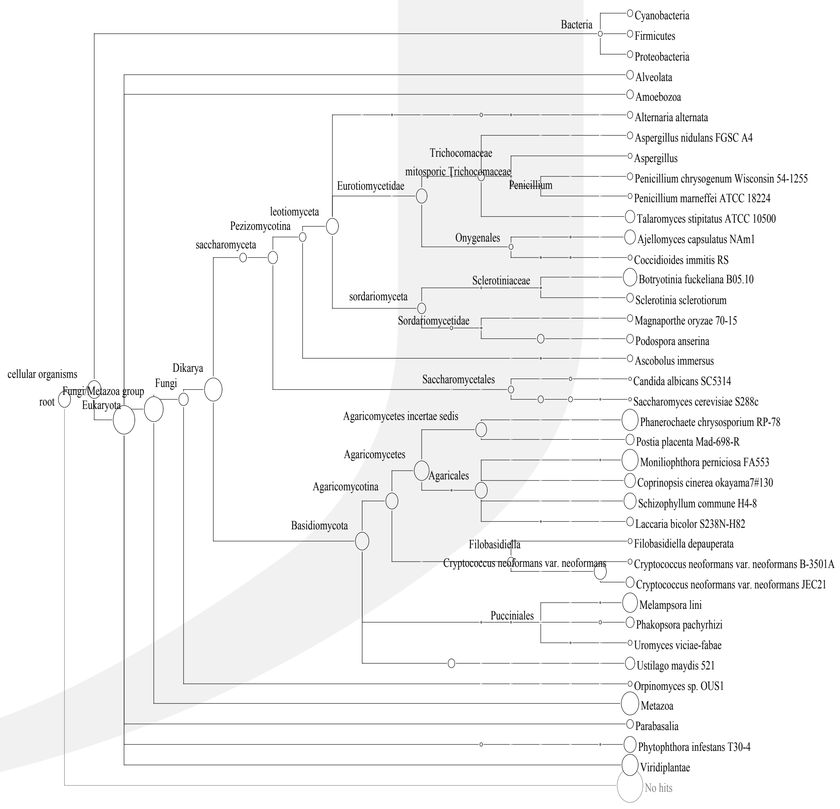

Figure 8. Taxonomy distribution of data (Megan - Expanded View).

Acknowledgment

We greatly acknowledge funding from Faculty of Sciences of the Universidad de los Andes through the program Funding Support for Assistant Graduate, to Ministerio de Agricultura and Colciencias for funding support for all the project.

References

AJJAMADA C. KUSHALAPPA & ALBERTUS B. ESKES. Coffee Rust: Epidemiology, Resistance and Management. 1989.

BRANDI L. CANTAREL, IAN KORF, SOFIA M.C. ROBB, GENIS PARRA, ERIC ROSS, BARRY MOORE, CARSON HOLT, ALEJANDRO SÁNCHEZ ALVARADO, AND MARK YANDELL. MAKER: an easy-to-use annotation pipeline designed for emerging model organism genomes. Genome Res. 2008 January; 18(1): 188–196.

HUSON DANIEL H., AUCH ALEXANDER F., JI QI AND SCHUSTER STEPHAN C. MEGAN analysis of metagenomic data. Genome Res. 2007; January; 17: 377-386.

MICHAEL C. SCHATZ, ARTHUR L. DELCHER AND STEVEN SALZBERG. Assembly of large genomes using second-generation sequencing. Genome Res. May 27, 2010.

Gerencia técnica, Programa de investigación científica. Centro Nacional de Investigaciones del Café "Pedro Uribe Mejía". Boletín técnico N°19, Cenicafé, recomendaciones para el manejo de la roya del cafeto en Colombia.1999.