Get off on the right foot. The importance of initial steps in NGS data analysis

Winant V.[1], Sonet G.[2], Virgilio M.[3], De Meyer M.[1]

[1] Department of Biology, Royal Museum for Central Africa, Leuvensesteenweg 13, 3080 Tervuren, Belgium
[2] Joint Experimental Molecular Unit (JEMU), Royal Belgian Institute of Natural Sciences (RBINS), Vautierstraat 29, 1000 Brussels, Belgium
[3] Joint Experimental Molecular Unit (JEMU), Royal Museum for Central Africa (RMCA), Leuvensesteenweg 13, 3080 Tervuren, Belgium

Introduction

Next-generation sequencing (NGS) is possibly the fastest evolving sector of molecular biology, allowing higher and higher throughput DNA sequencing at more and more affordable costs. The shift from "traditional" molecular techniques to NGS opens up a great deal of new opportunities, particularly in the study of non-model organisms. Yet, NGS data processing is still far from being standardized as it often relies on combinations of different software packages and on custom-made scripts incorporated at various stages of a pipeline. Here we present the results of a pilot study that aimed at evaluating the suitability of restriction-site associated DNA sequencing (RAD-seq) in resolving the phylogeny of a small species complex (*Ceratitis fascinventris*, *C. anonae* and *C. rosa*, Diptera, Tephritidae). Males can be differentiated by secondary sexual characters on the mid leg (**Figure 1**) but females cannot be easily separated. Currently five different genotypic groups are recognized within the complex based on microsatellite analysis.

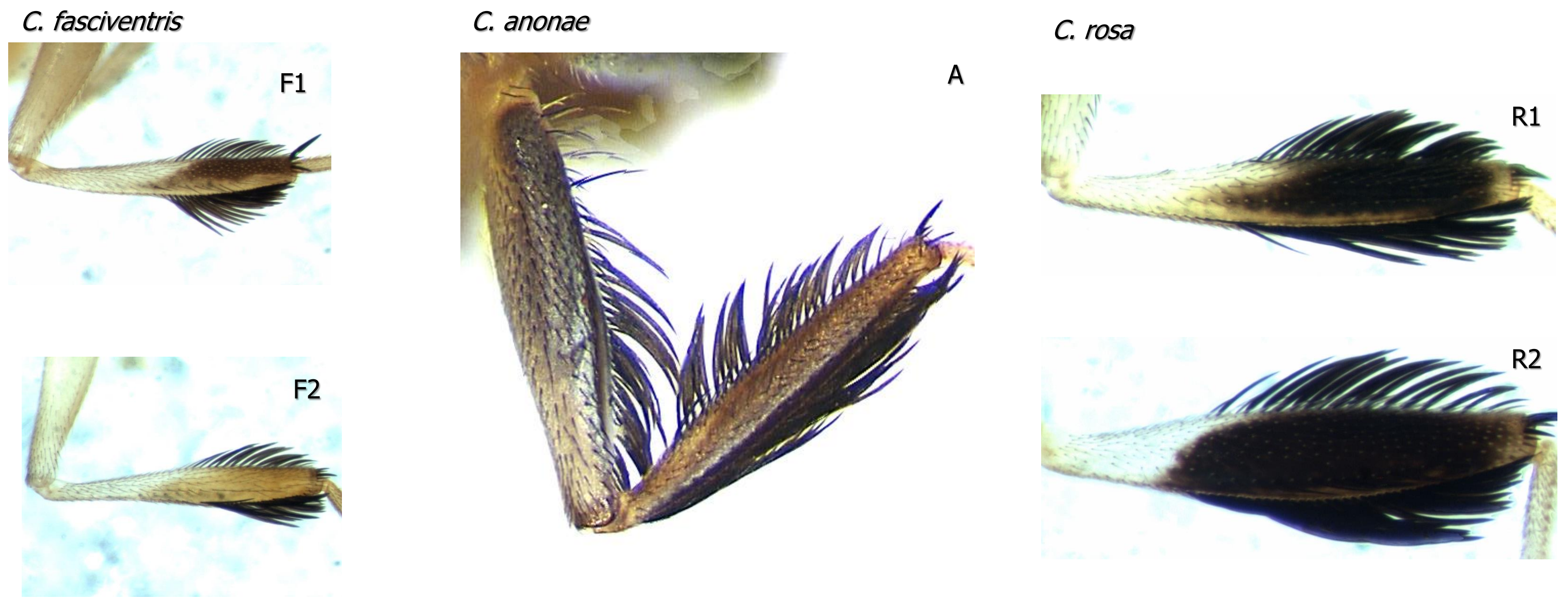


Figure 1 : morphological differences in male leg ornamentation patterns of *Ceratitis fasciventris* (F1, F2), *C. anonae* (A), *C. rosa* (R1, R2)

Materials & Methods

The effects of different cleaning and filtering procedures were evaluated by testing dedicated software for RAD data analysis (PyRAD and Stacks), a general trimmer for Illumina sequence data (Trimmomatic) as well as custom Linux and R scripts (**Figure 2**). DNA extraction was performed with the "DNeasy Blood & Tissue Kit" (Qiagen) using a non-destructive approach. Individual DNA samples were first digested with a restriction enzyme (Sbf1). DNA fragments were ligated P1 adapters then were multiplexed to a single library. The library of 16 individuals was randomly sheared, size-selected and finally ligated to a second adapter (P2 adapter). Only P1 adapter-ligated RAD tags was-amplified during the final PCR amplification step. This last step was followed by purification and gel extraction. The library was sequenced by the GenePool Genomics Facility of Edinburgh. The resulting reads were filtered by PyRAD and Stacks (**Figure 2**).

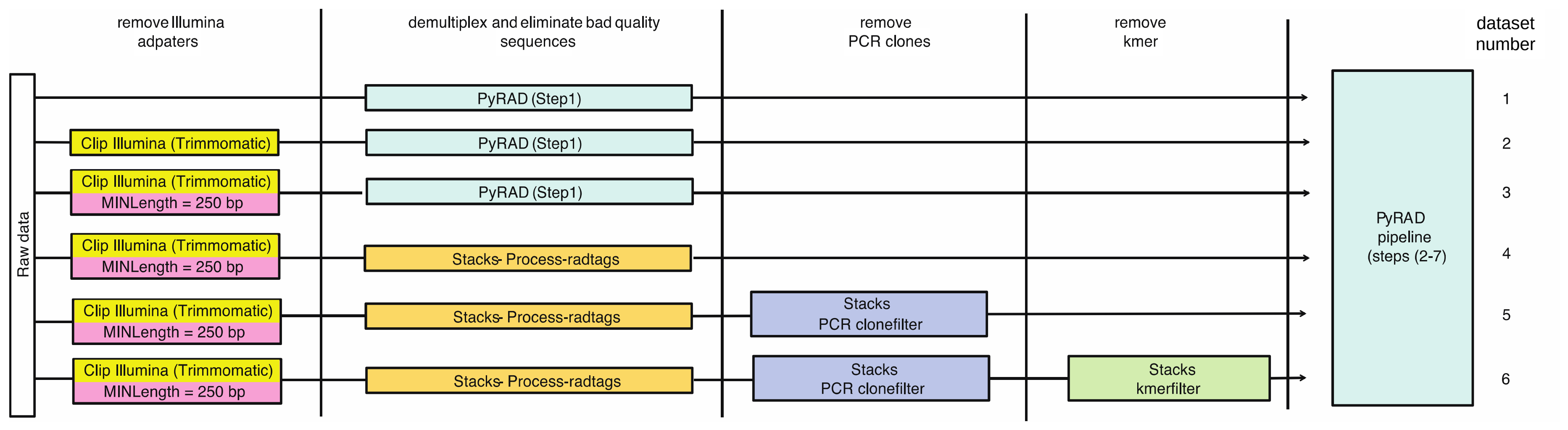


Figure 2 : Illustration of 6 cleaning and filtering procedures (providing 6 sets of reads). Raw data were-first clipped to remove Illumina adapters and in-line barcode remnants of all reads with Trimmomatic then de-multiplexed using PyRAD (datasets #1-3) or Stacks (datasets #4-6). The last two steps aimed at removing PCR clones and reads containing kmers.

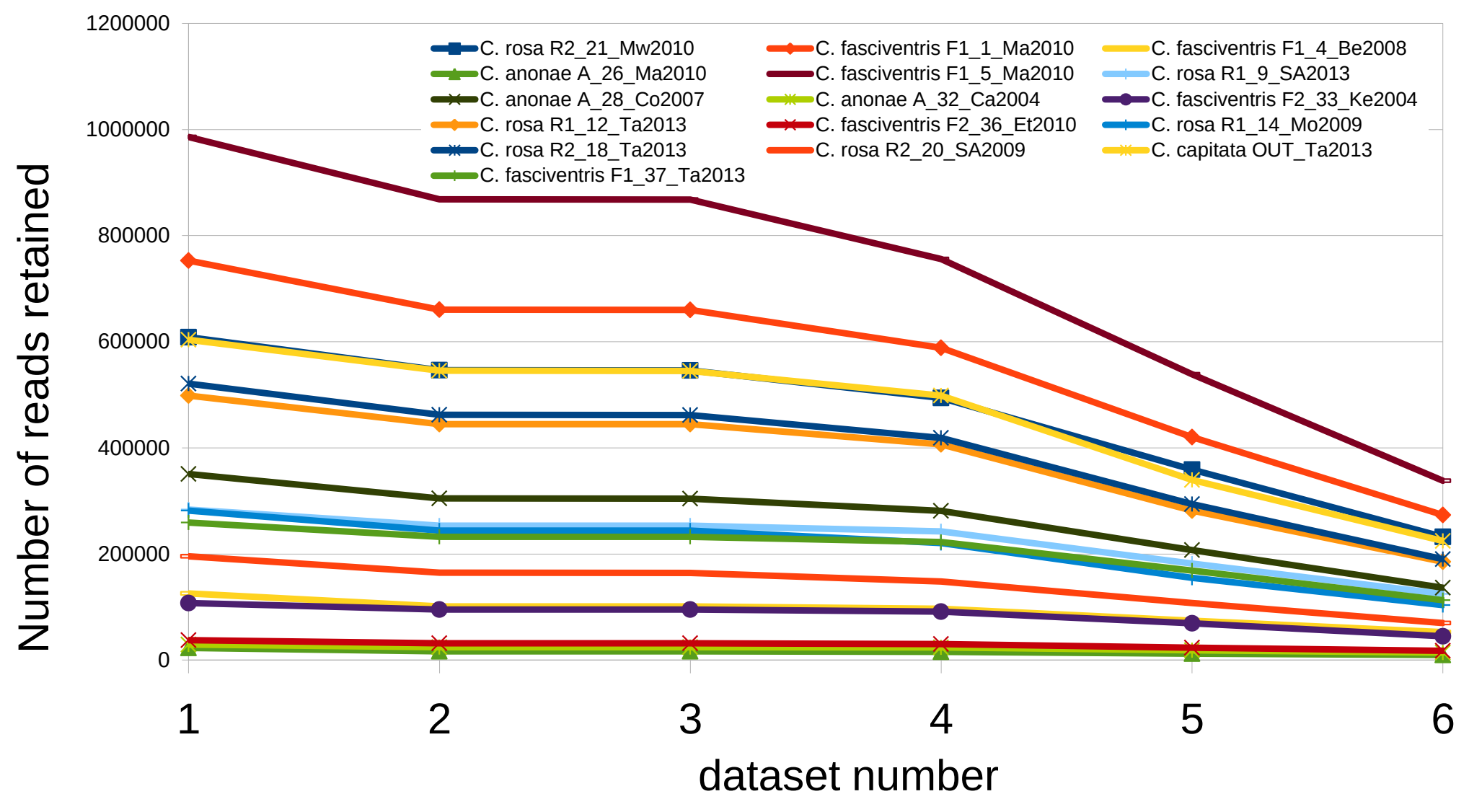
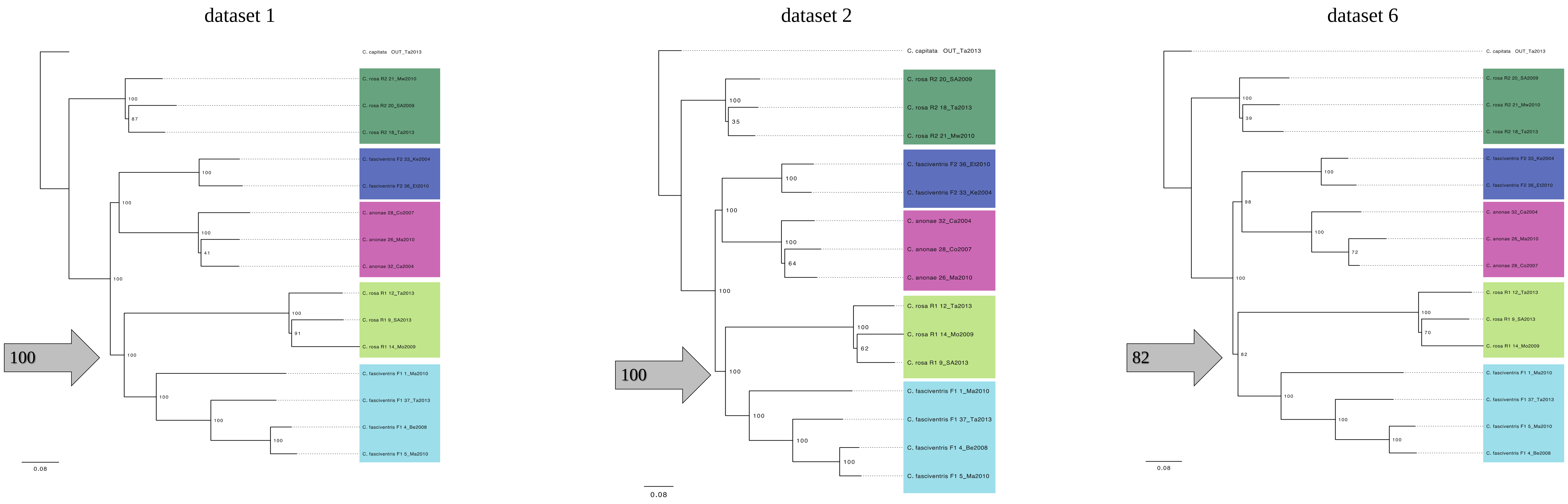


Figure 3 : number of reads obtained for each sample (see Fig. 2)

Results and Discussion


dataset 1

dataset 2

dataset 6

Figure 4: Maximum Likelihood trees (bootstrap values are indicated) obtained from datasets 1,2 and 6 (see Fig. 2)

RAD-sequencing allowed recovering five main clades corresponding to five genotypic clusters previously identified by microsatellite markers (i.e. F1, F2, A, R1, R2). As expected, sequential cleaning and filtering reduced the number of reads retained at each step (Fig. 3). Regardless of that, the ML trees obtained (here shown for sets 1, 2 and 6, see Fig. 4) showed largely comparable topologies. Node support varied for each of the different datasets. Yet, the only relevant difference between trees could be observed for the critical clade including *C. rosa* R1 and *C. fasciventris* F1 (100% in sets 1 and 2, 82% in set 6, arrow in Fig. 4). These results suggest the initial cleaning and filtering strategies adopted during the pipeline are crucial for optimal resolution and accuracy of data interpretation.

Acknowledgments

This project was funded by the Belgian Science Policy Office (BELSPO, as part of the AGORA-MOLCOL project) and carried out in collaboration with the Joint Experimental Molecular Unit (JEMU).

For more information, please visit http://jemu.myspecies.info/