Development of a Galaxy workflow for SNP detection in Grapevine and Poplar whole genome Illumina resequencing data

Nathalie Choixne1, Marc Bras4, Nacer Mohellibi1, Sandie Arnoux2, Hadi Quesneville3, Juliette Goarin2, Jean-Michel Boursiquot2, Vanina Guérin2, Marie-Christine Le Paslier2, Aurélie Berard3, Stéphane Schult2, Dominique Brunet5, Rémi Bouzon6, Frédérique Bilton2, Patricia Faivre-Rampant7, Anne-Françoise Adam-Blondon1

1 INRA-URGI (Unité de Recherche en Génomique-Info), UR1164, Versailles, France
2 INRA-DIAPC (Diversité et Adaptation des Plantes Cultivées), UMR1197, Montpellier, France
3 INRA-AOFP (Amélioration, Génétique et Physiologie Forestières), UR1168, Orléans, France
4 INRA-EGPV (Étude du Polymorphisme des Génotypes Végétaux), UR1129 CEA-IGCNG, Evry, France
5 INRA-URGV (Unité de Recherche en Génomique Végétale) UMR1165, Evry, France

Introduction

Large SNPs discovery projects are underway in poplar and grapevine using the illumina sequencing technology. In grapevine, 30 genotypes from different Vitis species are currently being resequenced and the reads obtained will be aligned along the grapevine reference genome sequence. In poplar, resequencing on P. nigra is divided in two steps. The first one is the deep resequencing of a few individuals (i) to construct a reference genome of the species and (ii) to identify SNPs. The second one consists of the resequencing of several genotypes at low coverage (2x) to maximize SNP discovery. Libraries and paired-ends sequencing (2x75bp and 2x100bp) on GAIX were performed by EPGV group and CNG (Centre National de Génotypage) Biological resources and Sequencing platforms.

Sequencing data are being analysed using MAPHiTS (Mapping Analysis Pipeline for High-Throughput Sequences), a pipeline for SNPs detection developed by the URGI platform using the Galaxy workflow manager [1]. MAPHiTS is currently running with the following public tools BWA [2,3], SAMtools [3], Tablet [4] and VarScan [5]. MAPHiTS workflow is able to deliver all SNPs and small indels found in the data set and to filter them according to various parameters such as the genome coverage, the allele frequency and pValue.

MAPHiTS : Mapping Analysis Pipeline for High-Throughput Sequences

Reference genomes: Vitis vinifera PN40024 (12X version) Populus trichocarpa V2

Visualization of SNPs and Indels on Gbrowse (GnpGenome)
http://urgi.versailles.inra.fr/index.php urgi/Data/Genome-Run-GnpGenome

References


Acknowledgments: We thank IGA for helpful discussion and INRA – AIP Bioresources, the PLANT - KBBE2008 and Eobiove projects for the financial support.