RNAseq as a source of genetic polymorphisms in *Epipactis* for molecular marker development

Daniel PRAT[1,*], **Jérôme BRIOLAY**[2], **Vincent NAVRATIL**[3]

1 Université Lyon 1, UMR 5023 LEHNA, 69622 Villeurbanne Cedex, France
2 Université Lyon 1, DTAMB, 69622 Villeurbanne Cedex, France
3 Université Lyon 1, PRABI, 69622 Villeurbanne Cedex, France
* **daniel.prat@univ-lyon1.fr**

Abstract – Most orchids have large genomes and consequently the development of molecular markers to analyze their genetic diversity is rather difficult. Isozymes have firstly been used, but now due to the evolution of techniques and the hazardous chemicals required, this type of genetic markers is no longer considered. Microsatellite markers are codominant markers and suitable for genetic analyses but their development is difficult and costly especially for large genome species. Consequently, genetic analyses in orchids have been restricted to few species and with a limited number of markers. New genome sequencing methods became very efficient and of reduced cost. In this context, in order to limit the sequencing effort, in *Epipactis*, we have investigated sequencing of coding and expressed sequences by RNAseq. cDNA fragments have been standardized to a length of 450 pb and then sequenced. In order to increase the level of polymorphisms, three plants, two from *Epipactis helleborine* and one from *E. placentina* were investigated. Both plants of *E. helleborine* were taken in separated stands, one at low elevation (200 m) and one at higher elevation (2300 m). We analyzed bulked flowers within a large developmental gradient from buds to pollinated flower in order to increase the number of expressed genes. Sequences of approximatively 200 DNA bases at each ends were obtained by 454 sequencing. About 6 billion of sequences per plant have been assembled into 100 000 sequences using trinity software. Obtained sequences are almost similar to already known sequences obtained in Monocots. Assembled gene sequences from the three different plants have been aligned in order to find sequence polymorphisms. Polymorphic sequences have been detected and are suitable to design primer pairs in order to reveal genetic variation within and among species in *Epipactis*. The next steps will consist of these primer pairs test and genetic analysis within and among populations.

INTRODUCTION

Genetic studies are required to provide suitable information for orchid conservation but the development of molecular markers is rather difficult in orchids due to their large genome size. Recent evolution of molecular biology techniques with new generation sequencing provides tools for analyzing polymorphism. Among them, RNA-Seq allows to focus analysis on expressed sequences and to reduce amount of sequencing effort. Gene sequences can be reconstructed by assembling short sequences. Gene variation can be obtained by sequence comparison of both alleles of heterozygous genes. In order to increase the level of detected polymorphisms, three different plants belonging to two different species and from ecologically different stands are investigated. Molecular markers will be then developed and applied for genetic studies.

MATERIALS AND METHODS

- Plant materials: two plants of *Epipactis helleborine* (210 m asl, poplar plantation; 1300 m asl, meadow) and one plant of *E. placentina*; flowers (flower buds to wilted flowers) were collected and bulked separately for each plant (immediate storage in liquid nitrogen).

- Messenger RNAs were extracted and fragmented (average of 400 bases) prior to reverse transcription, amplification and Illumina sequencing (150 bases from each end; Figure 1).

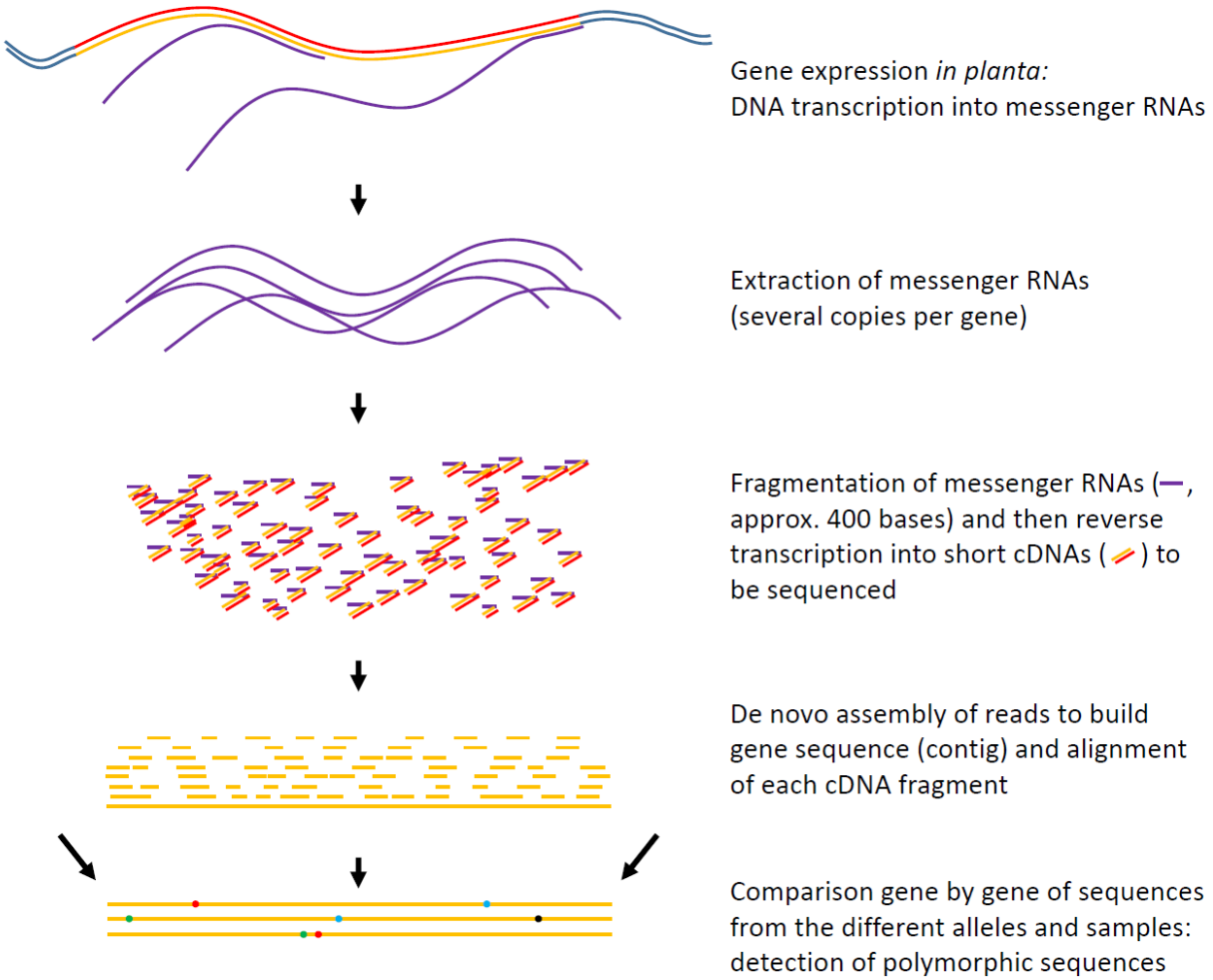

Figure 1. Principle of the present study.

- Sequences were assembled with Trinity (Galaxy platform). Reads of each plant were aligned with BWA on complete assembly. Sequences and polymorphisms were analyzed with Galaxy platform tools (Mpileup, varscan…).

RESULTS

- More than 4 000 000 fragments were sequenced for each plant (Table 1); their cumulated length represented more than 1 450 000 000 pb.

- The complete assembly consists of about 200 000 contigs, of an average size of 605 bp and median size of 345 pb stretching on 120 000 000 aligned bases. Sequencing depth reached thus about 12x.

- Sequence similarity analyses by blast*n* and Diamond reveal long assembled chloroplast genome sequences (up to 10 054 bp) and the identification of genes for more than 11 000 proteins, some hits being related to retrotransposons

- Polymorphisms was detected at the level of individual plant (related to heterozygosity) and at the complete sample (Table1). A large amount of SNPs was detected, more than 1.6 SNP/kbp. Indels were less abundant.

- *E. placentina* sample was not particularly differentiated.

- Position of SNPs and indels are localized along complete assembled sequences (Figure 2).

CONCLUSION

RNA-Seq analysis reveals a large amount of polymorphisms suitable for genetic studies. It provides also information on expressed genes.

Polymorphisms observed on coding sequences could be more related to plant phenotype and more reliable to morphological variation among plants and species.

Next and important steps would be design of primer pairs and DNA amplification in order to test their ability to reveal gene diversity. Selected primers will be then applied in *Epipactis* genetic studies.

Table 1. Main features of RNA-Seq analysis.

	E. helleborine 1	*E. helleborine* 2	*E. placentina*	Complete dataset
Reads	5 651 111	4 641 189	4 373 442	14 665 742
Number of contigs	102 452	100 463	96 786	199 373
Coding DNA sequences	68 276	63 968	61 778	118 200
Single Nucleotide Polymorphisms	17 157	16 311	9 586	71 131
Polymorphic contigs (counts)	5 478	4 812	3 240	17 739
Polymorphic contigs (%)	5.3	3.3	4.8	8.9
Contigs with indels	900	866	633	3 283

Parameter *Minimum read depth* set to 25

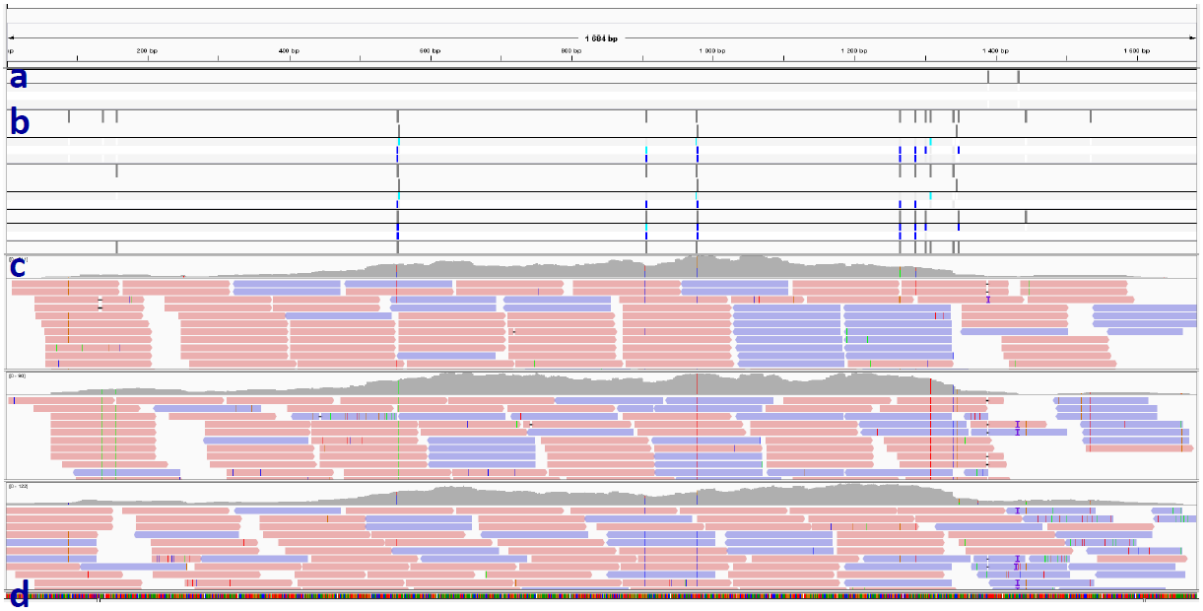

Figure 2. Position of indels (a), SNPs (b) and of reads for the three plants (c) along assembled sequence (d) of aminoacyl tRNA synthase complex-interacting multifunctional protein 1 (1684 pb).