The development of an SNP platform for rapid screening of the genomic constitution of xFestulolium hybrids and gene expression analysis

Štěpán Stočes[1], Tom Ruttink[2], David Kopecký[1], Jan Bartoš[1] and Jaroslav Doležel[1]

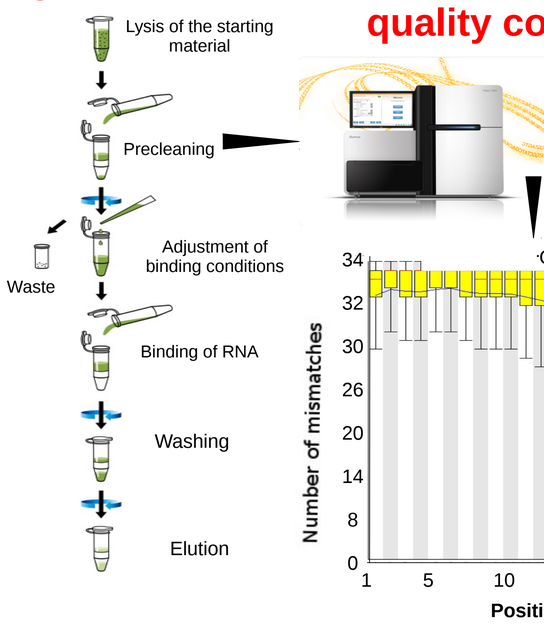
1) Centre of the Region Haná for Biotechnological and Agricultural Research, Institute of Experimental Botany, Šlechtitelů 31, CZ-783 71, Olomouc, Czech Republic
2) Institute for Agricultural and Fisheries Research, Caritasstraat 21, 9090, Melle, Belgium

Introduction

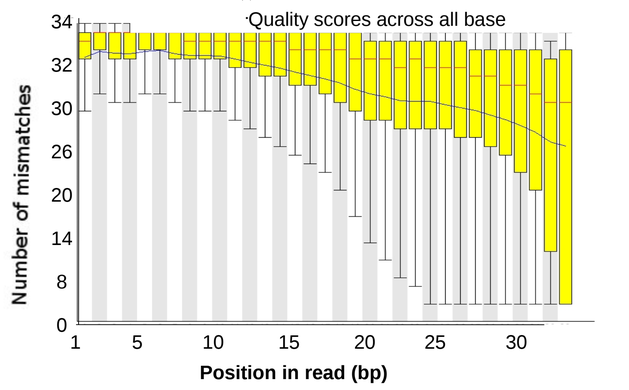
- xFestulolium is the superior grass developed by breeders using intergeneric cross of ryegrass and fescue species. It combines complementary agronomic attributes of both genera – high yield and nutrition of ryegrasses and tolerance to abiotic stresses of fescues.
- Despite the increasing popularity of these hybrids among seed companies and farmers, there is a lack of knowledge on the genomic constitution and gene expression. There are several methods, which have been used for discrimination of genome composition, but they generally suffer by low-throughput and limited resolution.
- On the other hand, Next Generation Sequencing technologies enables production of large sequence datasets, which can be used for the analysis of genome composition, molecular marker development, phylogenetic and ecological studies and analysis of transcriptomes using RNA-sequencing.
- The aim of our project is to study genomic constitution and gene expression in F1-F3 generations of xFestulolium hybrids using Illumina RNAseq technology.

Bioinformatics pipeline

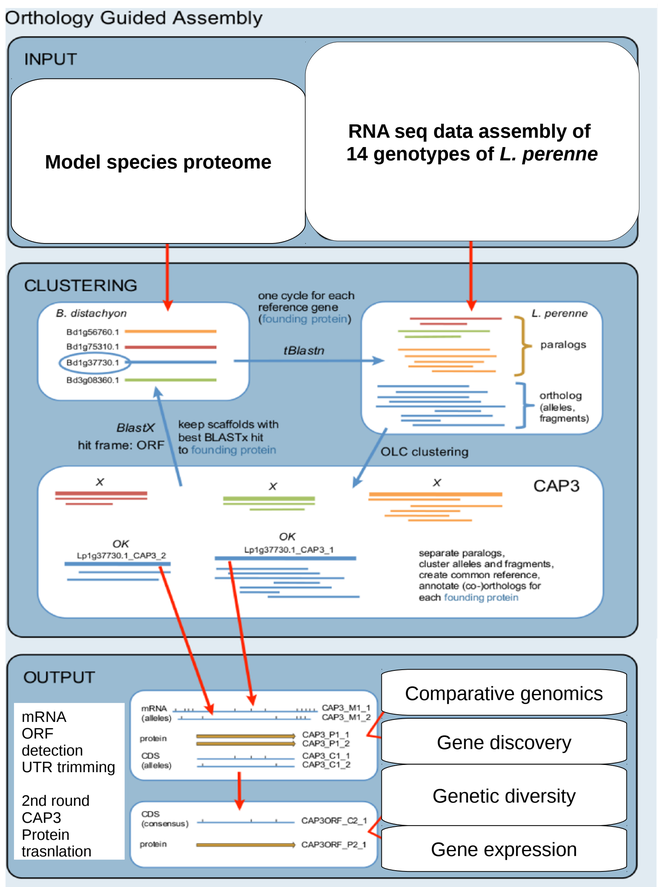
A) Isolation RNA, Illumina sequencing and quality control of data



- RNA was isolated using RNeasy Plant Mini Kit (6 genotypes of *F.pratensis* and 6 genotypes of *L.multiflorum*)
- RNAs were sequenced using Illumina RNA technology.
- Quality control of sequence data was performed using FastQC



B) Development of coding sequences reference

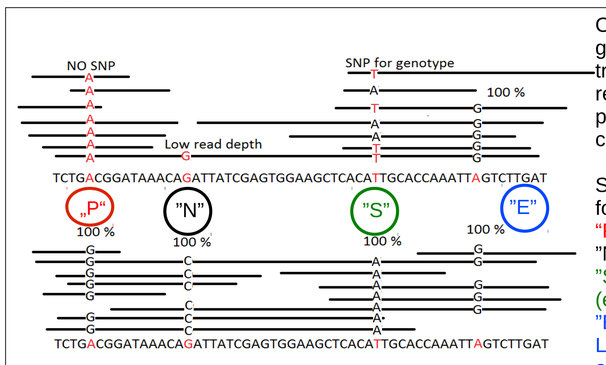


Since lack of fully annoted reference sequence of *L. multiflorum* (Lm) and *F. pratensis* (Fp), we used existing reference sequence of *L. perenne* (Lp) made from 14 genotypes of this species for *de novo* assembly of our RNA-seq data.

Lp reference sequence was developed using *Brachypodium distachyon* protein set for identification of orthologs and candidate genes. All contigs from 14 Lp genotypes with a significant similarity to the candidate *B.distachyon* proteins were selected and clustered with CAP3 software.

To distinguish between alleles and paralogs, it compared CAP3-contigs (BLASTx) against all *B.distachyon* proteins and retained only those with a best BLAST hit with the original candidate gene. Finally, were identified single orthologs and was constructed non-redundant reference transcriptome sequence.

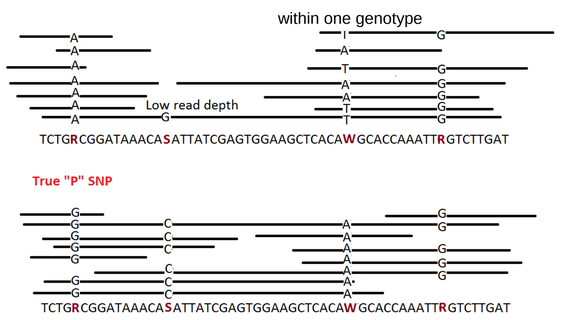
C) SNPs detection



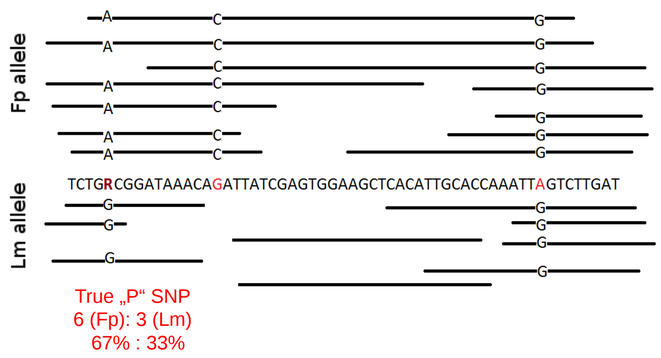
Our RNAseq data from 6 and 6 genotypes of parental species were trimmed and mapped on CDS reference. Thereafter, we detected all possible SNPs under low stringency conditions.

SNPs were cathegorized into the following groups:
“P” - polymorphic between Lm and Fp
”N” - not detected, low read depth
”S” - polymorphic within a species (either Lm or Fp)
”E” - nonpolymorphic between Fp and Lm, but polymorphic between samples and reference (Lp)

Reference was annotated with detected SNPs and transformed into the IUPAC code in location of these SNPs.
Reads from Fp and Lm were mapped on the reference including IUPAC codes and another round of SNPs detection was achieved. True polymorphic SNPs were selected for next steps. These true SNPs (1) include polymorphism between parental species, (2) are nonpolymorphic within one species and (3) have sufficient read depth.
IUPAC codes in reference sequence were changed to the original bases except true polymorphic SNPs.



D) Gene expression in F1 generations



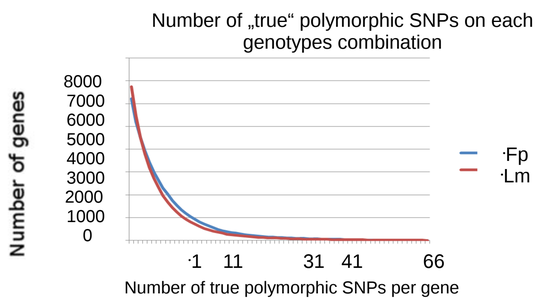
Reads from F1 hybrids were mapped on the reference with the IUPAC codes and SNPs were detected under low stringency conditions. The ratio of parental variants (Lm and Fp) for each SNP was obtained. This provided the insight on the change of gene expression in F1 hybrids. The analysis of F2 and F3 hybrids are in progress.

Results

Out of the total number of 600,000 to 900,000 SNPs, we identified 30,000 to 90,000 highly species-specific SNPs based on the combination of *F. pratensis* × *L. multiflorum* genotypes. These SNPs are localized on 6000 to 9000 genes distributed more or less evenly over all chromosomes.

Tab.1 - Number of all and potentially polymorphic SNPs between reference sequence and our genotypes

Genotypes	total SNP	potential "P" SNP
Fp51	696 555	436867
Fp52	845 168	629 622
Fp53	587 351	452 110
Fp54	795 953	546 360
Fp55	814 781	572 569
Fp56	745 826	483 592
Lm51	638 082	313 567
Lm52	664 712	490 588
Lm53	900 556	696 801
Lm54	800 478	524 964
Lm55	758 922	469 496
Lm56	622 605	474 160
total	9 184 194	6 090 696



Tab.2 - Number of true polymorphic SNPs for the combinations of parental genotypes used for development of F1 hybrids

Genotypes	Fp51	Fp52	Fp53	Fp54	Fp55	Fp56
Lm51		37926				
Lm52			64397			
Lm53			127364			
Lm54						
Lm55						92726
Lm56						

Conclusions

- Bioinformatics pipeline for the study of gene expression in intergeneric grass hybrids was developed
- This pipeline was found as an excellent tool for the analysis of gene expression in intergeneric hybrids.
- It offers a versatile tool for the gene expression studies of various intergeneric and interspecific hybrids.

This work was supported by grants P501/11/0504 and CZ.1.05/2.1.00/01.0007 and IGA PrF/2012/001