The repetitive landscape of Wheat Chromosome 5A. A preliminary study based on low-coverage NGS technologies

Lamontanara A[1], Vitulo N[2], Albiero A[2,3] , Forcato C[2], Campagna D[2], Dal Pero F[3], Cattivelli L[1], Bagnaresi P[1], Colaiacovo M[1], Faccioli P[1], Simkova H[6], Dolezel J[6], Perrotta G[5], Giuliano G[4], Valle G[2], Stanca M[1]

Motivation

Next generation sequencing (NGS) technologies are evolving at a very quick pace. While for most small genomes accurate characterization of genome landscape is no longer a challenging task in both terms of time and costs, complex eukaryotic genomes as cereal plant genomes pose dramatic constraints due to both their size and abundance of repeats and transposable elements (TE) which especially hamper final assembly steps. Nonetheless, acquiring preliminary information concerning the repetitive landscape of complex genomes is of obvious interest in order to gain insights in composition and dynamics of this sizeable genome fraction. Furthermore, gaining an early insight with respect to TE composition may prove useful in order to counteract technical issues which can potentially arise when finalizing genome assembly. Undertaking a low-coverage NGS sequencing approach on discrete genome fractions (as different arms of a given chromosome) can provide a first insight on these issues while keeping experimental efforts and costs under a desirable threshold.

Methods

Wheat (Triticum Aestivum) chromosome 5A short and long arms (5AS and 5AL, respectively) were independently isolated by flow cytometry. The DNA from the sorted chromosome arms was amplified by GenomiPhi amplification Kit (GE Healthcare), processed for DNA fragment analysis and run on a Roche 454-Titanium sequencer. A coverage of about 2X was produced. For 5AS and 5AL, 2,407,89 and 3,324,512 reads were respectively obtained. Long and short arm reads were independently blasted against Triticeae genomic repeat sequences (TREP complete database, BLASTN, Expect value < 10e-6; [1]) and matching reads were assigned to the “known TE families” group. In order to identify candidate novel TEs (novel TEs, while bearing scarce homology at the DNA level to known TEs may nonetheless exhibit substantial homologies in the CDS to

[1] Genomics Research Centre, Italian Agricultural Research Council, via S.Protaso 302, I-29017 Fiorenzuola d’Arda (Pc), Italy [2] CRIBI Biotechnology Center, University of Padova via U.Bassi 58/b, 35131 Padova, Italy [3] Bmr-genomics srl via Redipuglia 21/A, 35131 Padova, Italy [4] ENEA, Research Center CASACCIA, S.M. Galeria, 00163 Rome, Italy [5] ENEA, TRISAIA Research Center, S.S. 106 Ionica, 75026 Rotondella (Matera), Italy [6] Laboratory of Molecular Cytogenetics and Cytometry, Institute of Experimental Botany, Sokolovska 6, CZ-77200 Olomouc, Czech Republic

known TEs) the leftover reads were further screened by blasting against TREP protein division (PTREP, BLASTX, Expect value <10e-6). The resulting hits were grouped in the “novel TE families” fraction. Within the group of remaining reads, a further class of ill-defined “repeats” was identified and an approximate quantification was attempted on the basis of their participation to contigs with high coverage (>20-fold). In fact, given the 2-fold coverage in this study, a > 20-fold coverage should by definition represent repeats [1]. When populating this further class of repeats, in-house Python scripts were developed to parse ACE files and subsequently select only reads participating to contigs devoid of reads assigned to already classified TE or genes. The “others” group fraction refers to the leftover reads and should include, among various fractions, nuclear genes, organellar DNA, low-complexity DNA and further components.

Results

Known TE families (identity at the DNA level) reached 72.67% and 71.14% for 5AS and 5AL, respectively. Novel TE families (i.e. identity detected solely at protein level) amounted to 2.48% for 5AS and 2.60% for 5AL. Uncharacterized repeats were 10.35% and 7.97% for 5AS and 5AL, respectively, leaving an “others” fraction summing up to of 14.49% for 5AS and 18.29% for 5AL. TE family quantitative distribution was substantially uniform along the two 5A arms, apart some discrete families several-fold more abundant in the long arm. Few minor families were only detectable in one of the two arms, possibly reflecting recent bursts of transposition or, alternatively, classification artifacts. [1] Wicker T, Taudien S, Houben A, Keller B, Graner A, Platzer M, Stein N: A whole-genome snapshot of 454 sequences exposes the composition of the barley genome and provides evidence for parallel evolution of genome size in wheat and barley. Plant J 2009, 59 (5) :712-722.

Contact e-mail

antonella.lamontanara@entecra.it