Comparative Analysis of Repetitive Sequences Reveals Genome Differences between Two Common Cultivated Vaccinium Species (V. corymbosum and V. macrocarpon)

Nusrat Sultana1, Sedat Serçe1*, Gerhard Menzel2, Tony Heitkam2, Thomas Schmidt2
1Department of Agricultural Genetic Engineering, Ayhan Şahenk Faculty of Agricultural Sciences and Technologies, Niğde Ömer Halisdemir University, 51240, Niğde, Turkey
2Institute of Botany, Technische Universität Dresden, D-01062 Dresden, Germany

*Corresponding Author Received: May 25, 2017
E-mail: sedatserce@gmail.com
Accepted: October 06, 2017

Abstract:

Diploid northern highbush blueberry (Vaccinium corymbosum) and American cranberry (Vaccinium macrocarpon) are the two most widely studied species of the economically important genus Vaccinium for which whole genome sequence data and draft assemblies have become recently available in the public database. In this study, the major genomic differences between these two species were discovered through comparative analysis of repetitive sequence using the RepeatExplorer software. Approximately 80% of the V. corymbosum and 90% of the V. macrocarpon genome are composed of repetitive DNA, the main portion of which belongs to Ty1/copia and Ty3/gypsy LTR-retrotransposons, followed by DNA transposon, non-LTR-retrotransposon, and satellite DNA fractions. Analysis of the retrotransposon’s key enzyme, the reverse transcriptase (RT), reveals that elements of the Ale and Ogre/Tat lineages have been predominant in the Ty1/copia lineages and Ty3/gypsy superfamilies, respectively. In addition, L1-type long interspersed nuclear elements (LINEs) were dominating the non-LTR retrotransposon fraction. However, V. macrocarpon contains remarkably higher amounts of LTR and non-LTR retrotransposon compared to V. corymbosum. In total, seven different putative satellite families were identified and characterized. Among them VaccSat1, VaccSat4, VaccSat5 and VaccSat6 only occur in V. corymbosum, whereas VaccSat2, VaccSat3 and VaccSat7 were common in both species. Therefore, the percentage of satellite repeats was significantly higher in V. corymbosum than in V. macrocarpon. Comparative phylogenetic analysis of these abundant repetitive sequences demonstrates the extent of sequence diversity and specificity between this two Vaccinium species. Data derived from this study are highly valuable for the further genomic characterization of wild and cultivated Vaccinium species.

Keywords: Vaccinium corymbosum, vaccinium macrocarpon, repeatexplorer, repetitive dna, satellite dna, reverse transcriptase, phylogenetics

INTRODUCTION:

Vaccinium is a widely distributed genus throughout the world and belongs to the family Ericaceae, comprising approximately 450 species dispersed in about 30 different sections [1, 2]. Blueberry and cranberry are the two most commonly cultivated Vaccinium species along with other economically important plant species such as whortleberry, bilberry, cow berry or liquor berry [3-7]. Nowadays these species of Vaccinium draw the attention of the scientific media because of numerous reported health benefits including high levels of antioxidant potential and preventive activity against cancer [8-10]. Nonetheless, both blueberry (Vaccinium ssp. section Cyanococcus) and cranberry (Vaccinium ssp. Section Oxyccoccus) are members of a diverse phylogenetic group of species with intraspecies ploidy levels ranging from diploid to hexaploid (2n=2x=24; 2n=3x=36; 2n=4x=48; 2n=6x=72) [11, 12].

Identification and characterization of Vaccinium species and the subsequent improvement of their genome is still a difficult task due to extensive polyploidization and overlapping morphologies in the inter- and intra-species level. Therefore, state-of-the-art advanced technological strategies have been employed to reveal the genomic complexity of this genus [13-15]. For instance, whole genome sequencing, RNA-sequencing, gene annotation, and molecular markers for the study of Vaccinium genetic diversity are now available or underway (Bian et al. 2014, Rowland et al. 2012). In addition, the publicly available Genome Database for Vaccinium (GDV) (http://www.vaccinium.org/) has been established for efficient manipulation of this vast amount of information by many plant scientists. Eventually, raw whole genome sequence data and draft genome assemblies for diploid northern highbush blueberry (V. corymbosum strain: W8520, 2n=24, genome size 500 Mbp) and diploid American cranberry (V. macrocarpon cultivar: Ben Lear CNJ99-125-1 inbred clone, 2n=24, 470 Mbp) have been released through GDV in late 2011 [12, 16, 17].

Although Vaccinium research has progressed, information about genome structure and organization, more precisely presence and organization of repetitive DNA, are very scarce. Repeats can reach high copy numbers, in some grasses even amounting to 85% of the plant nuclear DNA [18-20]. Even though Polashock et al. 2014 [17] reported that about 39.53% of cranberry genome (Vaccinium macrocarpon) is composed of transposable elements, their structural and organizational diversity is still not well explored. Nevertheless, structural and functional analysis of repetitive DNA is the crucial part of a successful
genome assembly, marker development, epigenetic effect identification, and study of evolutionary genome dynamics as closely related organisms with similar gene content can contain differing amounts of repetitive DNA [21-25]. Based on their genomic organization and structural features, repetitive DNA can be divided into the two major groups of tandemly organized and dispersed sequences [26]. Tandemly organized sequences include satellite DNA, micro- and minisatelites, telomeric repeats and ribosomal genes, which are characterized by a specific repeated sequence motif arranged in arrays of up to several kb within a specific location in the genome [20, 27]. On the other hand, dispersed repetitive DNA sequences are mainly composed of transposable elements, in particular retrotransposons and DNA transposons, which are scattered throughout the genome or interspersed within other sequences [28]. Among these different types of repetitive DNA, retrotransposons are most abundant within plant genomes and have a significant influence on plant genome size variation [29]. Repetitive DNA is important for centromere formation and genome integrity maintenance [25], as well as for regulation of genome transcription [30]. Due to this profound effect of different types of repetitive DNA on genome dynamics they could play a major role in plant evolution and speciation [31, 32].

Availability of next-generation sequencing data allow unprecedented opportunities to access, identify and quantify repeats in plant genomes [33]. “RepeatExplorer”, a graph-based clustering pipeline for next-generation sequence data, is especially useful for the identification and characterization of different repeat types [34] and has been applied already for a variety of plant genomes, including potato, camellia, spring onion and cucumber [35-36], as well as for comparative analyses within the Fabaceae and Musaceae [21, 34].

We performed a comparative analysis of repetitive DNA to reveal the structural and organizational genomic diversity within *V. corymbosum* and *V. macrocarpon* using phylogenetic frameworks and efficient bioinformatics tools. We identified and classified the major repeat fraction of both genomes, and subsequently focused on specific repeat classes. A total of seven satellite families were annotated in *V. corymbosum*, three of them were shared with *V. macrocarpon*. In order to assess the retrotransposon content, we extracted and compared sequence data of their core enzyme, the reverse transcriptase (RT) from both assembled *Vaccinium* genomes. Phylogenetic and similarity-based analyses of identified satellites and RTs have been employed to understand the key driving force of evolutionary genome dynamics within both *Vaccinium* species.

**MATERIALS AND METHODS:**

Whole genome sequence (paired-end reads) for *Vaccinium corymbosum* and *V. macrocarpon* were extracted from public database available from NCBI SRA, and detailed information about these reads are provided in Table 1. Reads were quality-filtered and only reads longer than 70 nt were retained: We removed Illumina TruSeq adapters using the Trimmomatic tool [37] with the parameters ILLUMINACLIP:TruSeq3-PE-2:2:30:10. Applying the Fastx_trimmer from FASTX-Toolkit (hannonlab.cshl.edu/ fastx_toolkit), all reads were trimmed to 70 bp, followed by removal of shorter sequences using seqtk (github.com/ lh3/seqtk). A custom script was used to check, which reads were paired, followed by interlacing and selection of random 10 M reads. These reads have been subjected to the RepeatExplorer pipeline [34] with parameters -t 39 (minimal overlap for clustering) and -o 30 (minimum overlap for assembly). RepeatExplorer clustering has been conducted both, individually for *V. corymbosum* and *V. macrocarpon* and also comparatively. Based on read similarity the pipeline produces read clusters which were subsequently classified automatically by the RepeatMasker software and manually as follows: Each cluster and representative supercluster has been characterized based on the cluster shape, size, protein domain hit and Repeatmasker hits and other detailed information obtained from the RepeatExplorer output. After complete characterization the quantification of different types of repetitive DNA was performed using Microsoft Excel software. Each cluster from the comparative clustering was plotted on scatter plot diagram and species-specific satellite cluster was identified.

### Table 1: Available public genome (DNA) database for *V. corymbosum* and *V. macrocarpon* (Source://www. vaccinium.org)

<table>
<thead>
<tr>
<th>Types of data</th>
<th><em>V. corymbosum</em></th>
<th><em>V. macrocarpon</em></th>
</tr>
</thead>
<tbody>
<tr>
<td>General information</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Name of the studied organism</td>
<td><em>V. corymbosum</em></td>
<td><em>V. macrocarpon</em></td>
</tr>
<tr>
<td>Strain</td>
<td>W8520</td>
<td>cultivar:Ben Lear</td>
</tr>
<tr>
<td>Chromosome number</td>
<td>2n=24</td>
<td>2n=24</td>
</tr>
<tr>
<td>Genome size</td>
<td>500Mb (Reference; Brown et al. 2011)</td>
<td>470Mb (Reference; Zdep'ski et al. 2011)</td>
</tr>
<tr>
<td>Whole Genome Sequencing (NCBI-SRA data)</td>
<td>PRJNA170639</td>
<td>PRJNA245813</td>
</tr>
<tr>
<td>Instrument</td>
<td>Illumina HiSeq 1000</td>
<td>Illumina Genome Analyzer Ix</td>
</tr>
<tr>
<td>Library layout</td>
<td>Paired</td>
<td>Paired</td>
</tr>
<tr>
<td>Sra accession</td>
<td>SRA035499</td>
<td>SRA161994</td>
</tr>
<tr>
<td>Assembly Name</td>
<td><em>V. corymbosum</em>, Aug_2015</td>
<td>ASM77533v1</td>
</tr>
<tr>
<td>Assembly or available from</td>
<td>IGBQuickloadsite</td>
<td>GCA_000775353.1</td>
</tr>
<tr>
<td></td>
<td><a href="http://www">http://www</a>. igbquickload.org/blueberry</td>
<td>(NCBI)</td>
</tr>
</tbody>
</table>

Species-specific clusters and clusters having a star-like or circular graph typical for satellite repeats have been selected for further analysis using Geneious software platform version 6.8.1 (http://www.geneious.com, [38]). The contigs with the highest genome representation has been used to delimit the satellite’s monomer sequence. In order to derive a more robust monomer consensus, iterative read mapping has been performed, until the consensus remains stable. Eventually, artificial monomer trimers were used as a template and aligned against all contigs to ensure that the monomer consensus is equally covered by reads. Characterization of each prominent satellite family and monomer sequences was determined based on the repetitive sequence alignment of the tentative satellite cluster, and dot plot. The final satellite consensus monomers were subjected to phylogenetic comparisons.

Transposable element-derived reverse transcriptase protein sequences were extracted using protein domain
finder tools implemented in the RepeatExplorer Galaxy tool using the database for *V. corymbosum* and *V. macrocarpon* (Table 1). The sequences were filtered with minimum identity: 0.3, minimum similarity: 0.4, minimum alignment length: 0.8, interruptions: 3. Remaining RT sequences was subjected for clustering with 90% similarity hit with Cd-hit [39]. After clustering the sequences were subjected for pairwise similarity analysis and phylogenetic analysis using Geneious software platform.

RESULTS AND DISCUSSIONS:

Genomic proportion of repetitive DNA:

We performed graph-based sequence clustering for both blueberry and cranberry using publicly available whole genome sequence data. In order to obtain consistent results from RepeatExplorer runs, four different genome coverages ranging from 0.02x to 2.04x were tested individually. It was found that identified repeat proportions were significantly influenced by the genome coverage used for repeat identification. Therefore, a genome coverage of 2.04x was finally considered for the identification of the repeat content.

The results show that 79% of the blueberry and 91.4% of cranberry genomes belong to repetitive DNA. The rest of the genome for both of the species were constituted by single reads most likely representing the genic portion. Bar graph generated from Superclusters analysis reveals that the most abundant repeat superclusters in blueberry and cranberry belong to the tandem repeat group (blueberry) and the Ty3/gypsy LTR lineage (cranberry), respectively (Figure 1).

Moreover, we found that 24% and 49% were LTR-retrotransposon, 2% and 6% belong to non-LTR retrotransposon, 8% and 11% to DNA transposon, and 3% and 0.06% were satellite families of the blueberry and cranberry genomes, respectively. Other characterized reads group into rDNA, mitochondrial DNA and plastid DNA in both of the species. The total amount of characterized repetitive DNA with 49% in cranberry was much higher compared to blueberry with 37% (Figure 2).

COMPARATIVE CLUSTERING:

In order to identify the overall genomic differences and species-specific clusters, a comparative sequence clustering with equal genome coverage of 0.3x for both of the species was performed. This comparative sequence clustering identified 482 clusters with a minimal genome proportion of 0.009%. The scatterplot (Figure 3) shows separation of read clusters into three categories: Sequences similarly abundant in both species, enriched sequences and species-specific sequences. Even though most of the clusters are shared by both of the species with different genome proportion, there are some species-specific or species-enriched clusters, which can be used to differentiate these two species. For instance, the two blueberry-specific clusters CL8 and CL9 (genome proportion 0.46%) were found to be satellite families with a monomer size of 147 bp. This satellite was named VaccSat 1, with the representative individual clustering for *V. corymbosum* in supercluster SCL1. On the
other hand, CL102 with a genome proportion of 0.016% of the cranberry genome was specific to cranberry reads, most likely representing a satellite family with a monomer size of 565 bp. In addition, there were some small, genome specific uncharacterized clusters for both the blueberry and the cranberry genome (Table 2 and Figure 3).

Satellite DNA:
In total, seven putative satellite families were identified and characterized in the blueberry and cranberry genomes, designated as VaccSat 1 to VaccSat 7 (Figure 4).

Three of which are common to both species, whereas four are specifically enriched in blueberry. A dendrogram constituted from monomer consensus sequences reveals that VaccSat 1 and VaccSat 4 belong to the same satellite group which is highly heterogeneous and specific for the blueberry genome (Figure 5).

Whereas VaccSat 1 belongs to one supercluster (SCL1) is split into two individual clusters CL1 and CL3 with a total genome proportion of 0.99%, a monomer size of 146-147bp and an average GC content of 18.4%, VaccSat 4 belongs to CL 65 with a monomer size of 101 bp, an average GC content of 17.4% and a genome proportion of 0.05%. The other two blueberry-specific satellite families, VaccSat 5 and VaccSat 6, have a monomer size of 36-36 bp and 49 bp, an average GC content of 19.9% and 22.4%, and with the same genome proportion of 0.015% for blueberry which is highly underrepresented in cranberry genome.

VaccSat 2, 3 and 7 are common for both of the species. VaccSat 2 belongs to CL34 and CL 234 with a genome proportion 0.12% and 0.0665% for blueberry and cranberry, respectively. In addition, the monomer size was 238 bp with a comparatively higher GC content of 40.9% than the other satellite families.

Moreover, VaccSat 3 was similarly heterogeneous like VaccSat 1 and distributed in two individual clusters for both of the species. In case of blueberry the cluster number was CL6 and CL64, whereas for cranberry it were CL29 and CL 144, with the same monomer size of 154 bp, an average GC content of 21.1% but with a different total genome proportion of 0.36% and 0.284%, respectively.

VaccSat7 is another satellite family shared by both of the species and a highly heterogeneous monomer size ranging from 49 bp to 70bp, an average GC content of 31.4% and constituted within the single clusters CL 104 and CL76 for blueberry and cranberry, respectively (Table 2).
**Figure 4:** Satellite-specific circular graphs of read clusters identified in *V. corymbosum* and *V. macrocarpon*.

**Figure 5:** Phylogenetic dendrogram showing the relationship among the seven identified satellite families.
Transposable elements:

To gain detailed insight about transposable elements and their diversity between the two species, extracted reverse transcriptase (RT) protein domain were analyzed with phylogenetic frameworks. It was found that Ty3/gypsy retrotransposons make up the most abundant portion of repetitive DNA for both of the species (Figure 2). All major lineages such as Ogre/Tat, Athila and Chromoviruses occur, with Ogre/Tat being the most diverse lineage, containing 123 and 183 individual RT sequence clusters detected in blueberry and cranberry genomes respectively. The Ty3/gypsy-related pararetroviruses have been also detected in both genomes. Also, for Ty1/copia retrotransposons, all published lineages including Ale/Retrofit, Angela/Tork, TAR, SIRE/Maximus, Ivana/Oryco and Bianca have been identified. The lineage of Ale/Retrofit was the most diverse with sequence numbers of 82 and 109, followed by TAR with 17 and 22 sequences for blueberry and cranberry, respectively.

Even though they contribute only 2 % to the genome, LINEs are highly diverse with 249 and 269 individual RT sequence clusters detected in blueberry and cranberry, respectively.

Pairwise sequence similarity analysis between the two Vaccinium species reveals that pararetroviruses are highly heterogeneous in the blueberry genome compared to cranberry, whereas LINEs, Ikeros, Reina, SIRE and Orge/Tat are more heterogeneous in the cranberry genome compared to blueberry. On the other hand, element like Tekay, CRM, Athila, TAR, Ivana, Tork, Ale, Bianca and Galadriel were almost equally heterogeneous in both of the species (Table 3, Figure 6 and 7).

CONCLUSION:

In this study, two genomes of the genus Vaccinium have been characterized based on the analysis of repetitive DNA. LTR retrotransposons and satellite repeats were studied in detail using phylogenetic frameworks. Our results prove that the genus Vaccinium has three most prominent satellite families with satellite-typical monomer sizes, whereas their
abundance in the genome are mostly species-specific. We also find that the Ogre/Tat-like Ty3/gypsy lineage and Ty1/copia lineage Ale/Retrofit are the most diversifying LTR retrotransposons for the genus *Vaccinium*, even though the nature of the heterogeneity of these TE elements could be species-specific.

ACKNOWLEDGMENT:

This research work is a part of the PhD study of Nusrat Sultana, who is supported by The Scientific and Technological Research Council of Turkey (TUBITAK), provided TUBITAK-2215 graduate scholarship program, and Scientific Research Projects Unit (BAP) of Niğde Ömer Halisdemir University (BAP provided FEB 2017/18 DOKTEP project.

REFERENCES:


