agronomy

Article

Association Mapping of Fertility Restorer Gene for CMS PET1 in Sunflower

Denis V. Goryunov [1,2,*], Irina N. Anisimova [3], Vera A. Gavrilova [3], Alina I. Chernova [1], Evgeniia A. Sotnikova [4], Elena U. Martynova [1], Stepan V. Boldyrev [1,5], Asiya F. Ayupova [1], Rim F. Gubaev [1], Pavel V. Mazin [1], Elena A. Gurchenko [1], Artemy A. Shumskiy [1], Daria A. Petrova [1], Sergey V. Garkusha [6], Zhanna M. Mukhina [6], Nikolai I. Benko [7], Yakov N. Demurin [8], Philipp E. Khaitovich [1] and Svetlana V. Goryunova [1,5,*]

[1] Skolkovo Institute of Science and Technology, Moscow 121205, Russia; alin.chernova@gmail.com (A.I.C.); elenamartynovaster@gmail.com (E.U.M.); beibaraban34@gmail.com (S.V.B.); A.Ayupova@skoltech.ru (A.F.A.); rimgubaev@gmail.com (R.F.G.); iaa.aka@gmail.com (P.V.M.); Elena.Gurchenko@skoltech.ru (E.A.G.); artemy.shumskiy@gmail.com (A.A.S.); D.Petrova@skoltech.ru (D.A.P.); P.Khaitovich@skoltech.ru (P.E.K.)

[2] Belozersky Institute of Physico-Chemical Biology, Lomonosov Moscow State University, Moscow 119992, Russia

[3] N.I. Vavilov All-Russian Research Institute of Plant Genetic Resources, Saint Petersburg (ex Leningrad) 190000, Russia; irina_anisimova@inbox.ru (I.N.A.); v.gavrilova@vir.nw.ru (V.A.G.)

[4] Department of Computer Science and Control Systems, Bauman Moscow State Technical University, Moscow 105005, Russia; sotnikova.evgeniya@gmail.com

[5] Institute of General Genetics, Russian Academy of Science, Moscow 119333, Russia

[6] All-Russia Rice Research Institute, Krasnodar 350921, Russia; arrri_kub@mail.ru (S.V.G.); agroplazma@gmail.com (Z.M.M.)

[7] Breeding and Seed Production Company "Agroplazma", Krasnodar 350012, Russia; agroplasma@rambler.ru

[8] Pustovoit All-Russia Research Institute of Oil Crops, Krasnodar 350038, Russia; yakdemurin@yandex.ru

[*] Correspondence: D.Goryunov@skoltech.ru (D.V.G.); S.Goryunova@skoltech.ru (S.V.G.); Tel.: +7-926-151-2745 (D.V.G.); +7-926-218-0858 (S.V.G.)

Received: 20 December 2018; Accepted: 23 January 2019; Published: 24 January 2019

Abstract: The phenomenon of cytoplasmic male sterility (CMS), consisting in the inability to produce functional pollen due to mutations in mitochondrial genome, has been described in more than 150 plant species. With the discovery of nuclear fertility restorer (*Rf*) genes capable of suppressing the CMS phenotype, it became possible to use the CMS-*Rf* genetic systems as the basis for practical utilization of heterosis effect in various crops. Seed production of sunflower hybrids all over the world is based on the extensive use of the PET1 CMS combined with the *Rf1* gene. At the same time, data on *Rf1* localization, sequence, and molecular basis for the CMS PET1 type restoration of fertility remain unknown. Searching for candidate genes of the *Rf1* gene has great fundamental and practical value. Therefore, in this study, association mapping of fertility restorer gene for CMS PET1 in sunflower was performed. The genome-wide association study (GWAS) results made it possible to isolate a segment 7.72 Mb in length on chromosome 13, in which 21 candidates for *Rf1* fertility restorer gene were identified, including 20 pentatricopeptide repeat (PPR)family genes and one Probable aldehyde dehydrogenase gene. The results will serve as a basis for further study of the genetic nature and molecular mechanisms of pollen fertility restoration in sunflower, as well as for further intensification of sunflower breeding.

Keywords: cytoplasmic male sterility; fertility restoration; sunflower; *Rf1* gene; GWAS; Pentatricopeptide Repeats; *PPR* genes; association mapping; candidate genes

1. Introduction

The phenomenon of cytoplasmic male sterility, consisting in the inability to produce functional (viable) pollen due to mutations in mitochondrial genome, has been described in more than 150 plant species [1–3]. With the discovery of nuclear *Rf* genes capable of suppressing the cytoplasmic male sterility (CMS) phenotype, it became possible to use the CMS-*Rf* genetic systems as the basis for the practical utilization of heterosis effect in various crops (maize, sunflower, rice, sorghum, sugar beet, rapeseed, and others), and also as models to study the interaction mechanisms between the nuclear and mitochondrial genomes. The use of CMS lines as the female parent eliminates manual labor during crossing, and the use of fertile paternal lines carrying the gene (genes) for the restoration of pollen fertility (*Rf*) allows the production of highly fertile offspring that exhibit heterosis effect. When crossing the lines of annual cultivated sunflower (*Helianthus annuus* L.), the heterosis effect ranges from 28 to 40% according to different authors [4]. Sunflower has more than 70 sources of CMS, but in commercial hybrids breeding, the PET1 CMS obtained by P. Leclercq from the interspecific hybrid *H. petiolaris* Nutt. × *H. annuus* is used predominantly [5]. The PET1 mtDNA differs from the mtDNA of fertile forms by the presence of an inversion of 11 kb and insertion of 5 kb that leads to the appearance of a new open reading frame *orfH522*, which is co-transcribed with the atpA gene and encodes a 16 kDa protein [6,7]. Literature describes various types of inheritance of pollen fertility restoration in sunflower with CMS PET1-type and reports on the different number of genes that determine this character. Based on hybridological analysis, various authors distinguish from one to five genes responsible for the restoration of pollen fertility in CMS PET1 [8,9]. Seed production of sunflower hybrids is based on the extensive use of the *Rf1* and *Rf2* genes, which by interacting with each other give the effect of restoring pollen fertility. It is believed that the main gene is *Rf1*, which is responsible for the fertility restoration and is present in the vast majority of CMS PET1 fertility restoring lines [10]. Sometimes, as a result of hybridological analysis in the second generation of hybrids, monogenic segregation occurs, in such cases it is assumed that the second gene is present in both crossed forms.

The *Rf1* gene was originally assigned to the sixth linkage group [11] and was subsequently reassigned to the second linkage group [12]. On the integrated genetic map of sunflower, the genetic factor *Rf1* responsible for the restoration of pollen fertility is localized in the linkage group 13 [13,14]. To identify the *Rf1* gene alleles in the genotype, the closely linked molecular markers were developed [15], the diagnostic value of which was confirmed by several researchers [4,16]. However, approximately 10% of clones from the N.I. Vavilov All-Russian Research Institute of Plant Genetic Resources (VIR) collection lacked markers, although the presence of the *Rf1* gene dominant allele in their genotypes was confirmed using other methods [17].

Markers designed based on the information on the primary nucleotide sequence of the *Rf* genes themselves are considered to be the most effective for the selection of the functional *Rf* alleles carriers [18]. The absence of such markers in sunflower is explained by the lack of information about the nature of the *Rf* gene (genes). To identify the *Rf1* gene, positional cloning method was used [19], and molecular markers based on polymorphic fragments of *PPR* genes have also been developed [20,21]. However, to this day, the sequence of the *Rf1* gene remains unknown.

Most of the *Rf* genes described so far encode PPR proteins that contain repetitive degenerate motifs of 35 amino acid residues (Pentatricopeptide Repeats, PPR), with a few exceptions—for example, the *Rf2* gene, which encodes maize aldehyde dehydrogenase [22] and the rice *Rf2* gene, the product of which is a protein containing a glycine-rich domain [23]. RF-PPR proteins have from 15 to 20 PPR motifs in their sequences [24]. In angiosperm plants, *PPR* genes belong to two main types, P and PLS, which differ in the structure of PPR domains [25,26]. In flowering plants, the family of *PPR* genes, containing up to 600 members, is involved in anterograde/retrograde regulation, which ensures the coordinated work of nuclear and organelle genomes. Genes the products of which have the function of restoring fertility are included into the *RFL-PPR* (Restoration of Fertility Like-PPR) subfamily. Unlike the numerous families of conservative *PPR* genes that regulate processing, as well as participate in the splicing, editing, stabilization and translation of organelle RNAs, *RFL-PPR* genes are organized into clusters and are characterized by an exceptionally high level of variability [27]. It is believed that allele-specific markers of *RFL-PPR* genes can be used for positional cloning of fertility restorer genes, as well as for efficient selection of the carriers of functional alleles of *Rf* loci [18,28]. In addition, an exceptionally high rate of evolution of the subfamily of *RFL-PPR* genes [29,30], as well as the important role of CMS and restoration of fertility in the formation of new species [31] allow us to consider them as a source of molecular markers for phylogenetic research.

Thus, the *Rf1* gene is a key element in obtaining heterotic sunflower hybrids based on CMS. At the same time, data on its localization in the genome remain controversial. In addition, the gene sequence and molecular basis for the CMS PET1 type restoration of fertility remains unknown. Therefore, searching for candidate genes and mapping of the *Rf1* gene may have a great fundamental and practical value. Therefore, in this study, association mapping of fertility restorer gene for CMS PET1 in sunflower was performed.

2. Materials and Methods

134 *Helianthus annuus* L. elite lines from "Agroplazma" breeding and seed production company (Krasnodar, Russia) were taken into the study (Table S1, Table S2). They included 74 restorer lines carrying dominant allele of the *Rf1* gene and 60 sterility maintainer lines with the recessive allele of the gene.

Genomic DNA was extracted from the etiolated seedlings after one week of germination in the dark. 100 mg of tissue for each sample was grounded using the FastPrep-96™ Automated Homogenizer (MP Biomedicals, Santa Ana, CA, USA). DNA was extracted using the NucleoSpin® Plant II plant DNA extraction kit (Macherey-Nagel, Düren, Germany) according to the manufacturer's recommendations and stored at −20 °C. The quality of the purified DNA samples and DNA concentration were assessed by gel electrophoresis and Qubit 3.0 Fluorometer (ThermoFisher Sscientific, Waltham, MA, USA). Restriction site associated DNA sequencing (RAD) libraries were prepared using HindIII and NlaIll endonucleases as previously described [32] with minor modifications and sequenced in Illumina HiSeq4000 (San Diego, CA, USA). Raw sequence data are available on NCBI SRA under project number PRJNA515598.

Preprocessing of raw reads was performed with the aid of the Trimmomatic software (version 0.30) [33]. Genome variants were called in Tassel 5 GBS v2 pipeline [34] with the following command line arguments: -kmerLength 65, -minMAPQ 20, and -mnQS 20. Bowtie2 [35] was used to map tags to the HanXRQr1.0 reference genome [36] with —very-sensitive—very-sensitive preset. Principal component analysis (PCA) and linkage disequilibrium (LD) analyses were accomplished in Tassel 5 software and visualized by means of ggplot2 R library (version 3.1.0) [37]. Statistical analysis using the mixed linear models (MLMs) [38] implemented in the Tassel 5 software was performed for association mapping with PCA and kinship matrixes as covariates. Multiallelic variants and those with the high missing call rates, MAF below 0.01 as well as the samples with many missing calls were filtered out in PLINK 1.9 [39,40] before genome-wide association study (GWAS) analysis. Significant loci were identified based on Bonferroni and FDR adjusted q-values with 0.01 alpha significance level.

Genome-wide association study results were visualized with the aid of the qqman R package (version 0.1.4) [41].

3. Results and Discussion

3.1. Genotyping and GWAS Analysis

Sequencing of RAD-libraries and subsequent analysis has identified 28,153 SNP (Single nucleotide polymorphisms) in 134 sunflower accessions. Overall transitions to transversions ratio was 1.83.

PCA analysis revealed significant population structure. Restorer lines and sterility maintainers form separate groups on the scatterplot (Figure 1).

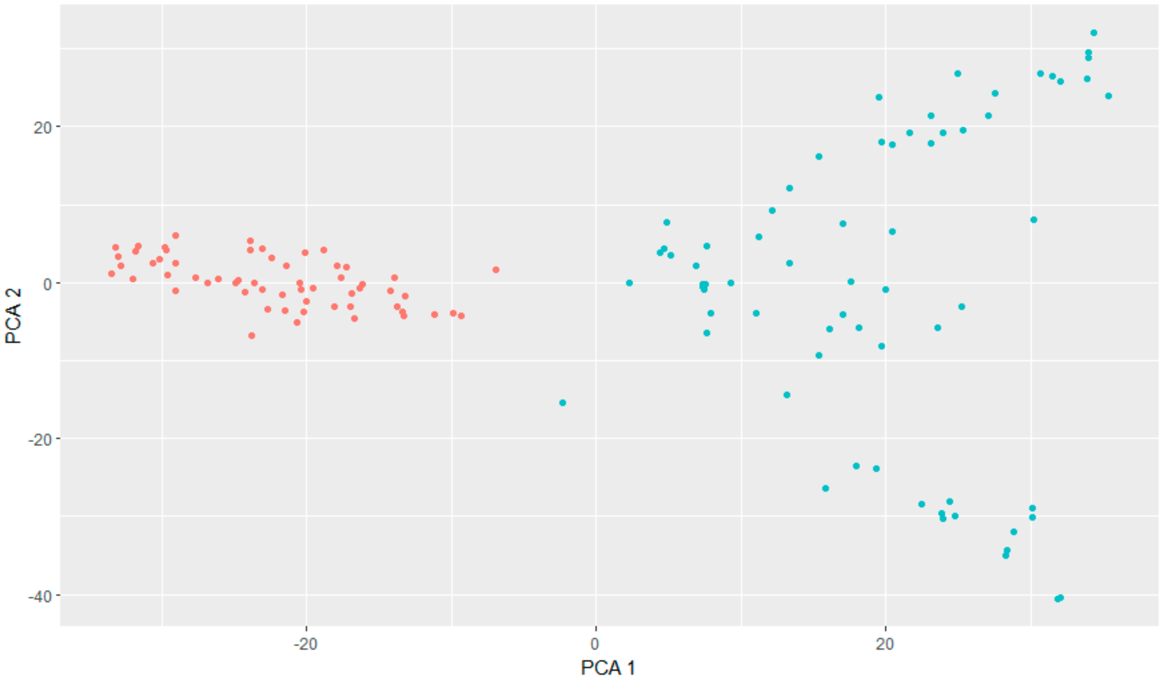

Figure 1. Principal component analysis (PCA) plot of sunflower lines based on Restriction site associated DNA (RAD) sequencing data. Pink dots–sterility maintainers, blue dots–restorer lines.

GWAS analysis revealed four loci associated with the ability to suppress CMS phenotype. Single significant marker was revealed at both 8 and 17 linkage groups. Most of the markers significantly associated with the trait under study, as well as the markers with the highest *p*-values, were located at 10 and 13 LG (Figure 2, Figure S1, Table S3).

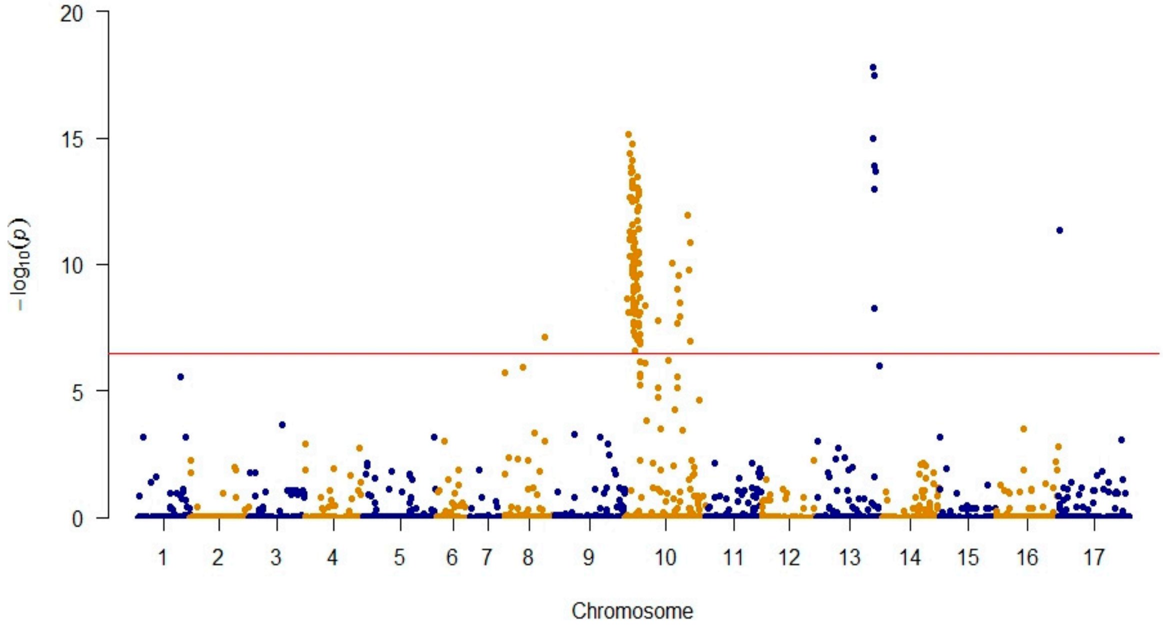

Figure 2. Manhattan plot of associations with the ability to suppress CMS (cytoplasmic male sterility) phenotype. Red line indicates the significance threshold based on the Bonferroni multiple testing correction (alpha = 0.01).

In addition to the difference in the ability to restore pollen fertility in the crosses with sterile lines with PET1-type cytoplasm, the analyzed sunflower lines differed by the presence (restorer lines) or absence (sterility maintainer lines) of plant branching. This is due to the fact that to obtain F1 hybrids, non-branched lines with a single large apical head are most often used as female parents, and lines with a recessive type of branching, with multiple small heads located on the lateral branches, are used as male parents. This approach allows an increase in the length of the flowering period of male parents due to the difference in the flowering times of the heads on the plant, and at the same time to get F1 plants with a large single head. It is known from the literature that branching locus is localized on chromosome 10 [42,43]. Therefore, the associations identified on chromosome 10 seem to be linked to this trait.

At the same time, the associations identified on chromosome 13 correspond with the data obtained in the previous studies. For instance, Yu et al. combined RFLP, RFLP-SSR, and SSR maps and obtained data for localization of *Rf1* in LG13 [44]. One year later Kusterer et al. [45] map *Rf1* based on cosegregation with SSR markers ORS388 and ORS1030 belonging to LG 13 Tang et al. [14]. Further Kusterer et al. obtained saturated map of the fertility restoration region *Rf1* [13]. Mapping data have confirmed the location of *Rf1* on LG13 near marker ORS1030. According to Yue et al. *Rf1* is in the interval between markers ORS511 and ORS799 of linkage group 13 [20]. Based on this, the most likely location for the candidate *Rf1* genes appears to be chromosome 13.

Within chromosome 13, based on the results of the GWAS analysis, a 7.72 Mb long section (coordinates170494693–178217103), can be distinguished, in which eight significant SNPs are located with *p*-values ranging from 5.69×10^{-9} to 1.53×10^{-18} (Table 1, Figure S2).

Table 1. List of Single nucleotide polymorphisms at linkage group 13, significantly associated with the ability to suppress CMS phenotype after Bonferroni correction.

Marker	Position	*p*-Value
S13_170494693	170494693	1.01×10^{-15}
S13_171053833	171053833	1.53×10^{-18}
S13_173268042	173268042	3.46×10^{-18}
S13_173832391	173832391	5.69×10^{-9}
S13_174474103	174474103	1.22×10^{-14}
S13_174474122	174474122	1.22×10^{-14}
S13_174809087	174809087	1.10×10^{-13}
S13_178217103	178217103	2.03×10^{-14}

To compare localization of 7.72 Mb region identified in this study with previously reported data we blasted PCR primer sequences of the ORS511, ORS799 and ORS1030 markers against the reference genome. ORS511 and ORS1030 were mapped in close proximity to each other on LG13 in according to Tang et al. [14]. Complete sequences of ORS1030 forward and reverse primers were mapped with 100% identity twice in the genome. Forward primer mapped to the positions 169535691–169535666 and 169655088–169655063 and reverse primer to the positions 169535262–169535287 and 169654659–169654684 of LG 13. For ORS511 complete sequences of forward primer have no hits on the 100% identity threshold. Reverse primer of ORS511 was mapped at 169733686–169733704 of LG13. For the ORS799 marker complete sequences of forward and reverse primers were uniquely mapped to the genome in position 186516272–186516291 and 186516418–186516399 of LG13 respectively.

These data suggest that identified 7.72 Mb region (coordinates170494693–178217103) is located within segment of chromosome 13 flanked by SSR markers ORS799 and ORS1030 (coordinates 169535262–186516418).

3.2. Identification of Rf1 Candidate Genes

Within identified 7.72 Mb region in the HanXRQr1.0 reference genome sequence [36], 11 *PPR* genes are located, which are the most likely candidate genes for the fertility restorer gene *Rf1*. Almost all *Rf* genes in various plant species that have been identified so far belong to this family [46–49].

PPR genes are thought to be present in all eukaryotes, but they are most common in the genomes of terrestrial plants, where they form one of the largest gene families [50]. For example, in the genome of *Arabidopsis thaliana* L. there are about 450 genes of this family [51,52], about 500 in the maize genome [53], and more than 600 in the genome of *Oryza sativa* L. [25]. Proteins of this family are characterized by the presence of multiple helix-turn-helix domains, forming a supercoil with a central groove [50,52]. This allows the protein to bind to RNA and participate in the RNA-protein interactions. Pentatricopeptide repeat (PPR) proteins play a significant role in regulating gene expression in the organelles at the RNA level [50,54,55].

The total number of annotated *PPR* genes in the sunflower genome HanXRQr1.0 is 333. Therefore, the identified region of 7.72 Mb (comprising 0.214% of the genome length) contains 3.3% of all annotated *PPR* genes and is rich in *PPR* genes. In addition, within this region, 10 genes of the *TPR* family are annotated. It is known that the sequences of PPR proteins are similar to the sequences of the TPR-family proteins and it is assumed that the tetratricopeptide repeat (*TPR*)- family genes gave rise to *PPR* genes at the early stages of the evolution of eukaryotes [56].

Therefore, it was decided to include the gene sequence of both the *PPR* and *TPR* families in the further analysis. It should be noted that genome sections 7.72 Mb in length, flanking the region of the chromosome 13 mentioned above, did not contain any annotated sequence of the *PPR* family, and only a single sequence belonging to the *TPR* family (HanXRQChr13g0421851).

The analysis of the translated amino acid sequences of 22 genes of the *PPR* and *TPR* families located in the identified region and its flanking regions was conducted using ScanProsite tool of ExPASy SIB Bioinformatics Resource Portal (SIB Swiss Institute of Bioinformatics, Lausanne, Switzerland). As a result, in all 11 amino acid sequences of the PPR family and in 10 of the 11 sequences of the TPR family, Pentatricopeptide (PPR) repeats were identified. Therefore, within the 7.72 Mb region and the flanking regions, 21 genes were detected, their protein products demonstrating the primary structure characteristic of the sequences of the *PPR* family. Meanwhile, in addition to PPR repeats, the amino acid sequence of the protein product of one of the genes revealed a region of homology with UDP-glycosyltransferases, and therefore this gene was excluded from the list of possible candidate genes for *Rf1*.

In addition to *PPR* genes, a gene annotated as Probable aldehyde dehydrogenase 5F1 was detected in the 7.72 Mb region of chromosome 13. It was previously shown that *Rf2* gene of maize is the gene encoding aldehyde dehydrogenase [57]. Therefore, this gene is also a possible candidate *Rf* gene. The list of identified candidate genes is shown in Table 2, and their arrangement within the 7.72 Mb region is shown in Figure 3.

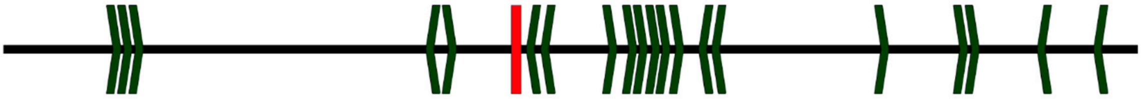

Figure 3. Schematic localization of the candidate *Rf1* genes within the 7.72 Mb region. Green arrows indicate the gene sequences of the *PPR* family. The direction of the arrow reflects the orientation of the sequence in the genome. Red box indicates the location of the Probable aldehyde dehydrogenase 5F1gene.

The number of PPR repeats in the sequence and the length of the protein products of the candidate PPR family genes varied from 2 to 15 and from 110 to 756 amino acids, respectively.

Genomic regions with increased LD could be recognized as signatures of strong selection pressure on the traits encoded within these regions. The results of the analysis showed the presence of an extended section of elevated LD in 13 LG (Figure 4), of which the identified 7.72 Mb region forms part. This fact is an indirect proof of the localization of candidate genes in this region of the genome.

Table 2. The list of candidate *Rf1* genes identified within the 7.72 Mb region.

Gene	Start	End	Strand	Product	Gene Bank Accession Number of Translated Protein	Hits for All PROSITE (Release 2018_11) Motifs
HanXRQChr13g0418841	170850155	170852002	+	Putative pentatricopeptide repeat	OTG02960	PS51375 Pentatricopeptide (PPR) repeat
HanXRQChr13g0418861	170908019	170909110	+	Putative pentatricopeptide repeat	OTG02962	PS51375 Pentatricopeptide (PPR) repeat
HanXRQChr13g0419621	173473487	173475525	–	Probable pentatricopeptide repeat-containing protein At2g41080	OTG03034	PS51375 Pentatricopeptide (PPR) repeat
HanXRQChr13g0419631	173484455	173500401	+	Putative pentatricopeptide repeat	OTG03035	PS51375 Pentatricopeptide (PPR) repeat
HanXRQChr13g0419931	174209661	174217234	–	Putative pentatricopeptide repeat	OTG03064	PS51375 Pentatricopeptide (PPR) repeat
HanXRQChr13g0420121	174799667	174801481	+	Probable pentatricopeptide repeat (PPR) superfamily protein	OTG03081	PS51375 Pentatricopeptide (PPR) repeat
HanXRQChr13g0420241	174944047	174945506	+	Putative pentatricopeptide repeat	OTG03093	PS51375 Pentatricopeptide (PPR) repeat
HanXRQChr13g0420261	174962084	174962512	+	Putative pentatricopeptide repeat	OTG03095	PS51375 Pentatricopeptide (PPR) repeat
HanXRQChr13g0420351	175219425	175219886	–	Putative pentatricopeptide repeat	OTG03099	PS51375 Pentatricopeptide (PPR) repeat
HanXRQChr13g0420811	176970038	176972308	+	Probable pentatricopeptide repeat (PPR) superfamily protein	OTG03141	PS51375 Pentatricopeptide (PPR) repeat
HanXRQChr13g0421081	178216563	178219635	–	Probable putative pentatricopeptide repeat-containing protein At4g17915	OTG03166	PS51375 Pentatricopeptide (PPR) repeat
HanXRQChr13g0418851	170877322	170879307	+	Putative tetratricopeptide-like helical domain	OTG02961	PS51375 Pentatricopeptide (PPR) repeat
HanXRQChr13g0419881	174159006	174160682	–	Putative tetratricopeptide-like helical domain	OTG03060	PS51375 Pentatricopeptide (PPR) repeat
HanXRQChr13g0420271	175002640	175003793	+	Putative tetratricopeptide-like helical domain	OTG03096	PS51375 Pentatricopeptide (PPR) repeat
HanXRQChr13g0420281	175016437	175018065	+	Putative tetratricopeptide-like helical domain	OTG03097	PS51375 Pentatricopeptide (PPR) repeat
HanXRQChr13g0420301	175055952	175057826	+	Putative tetratricopeptide-like helical domain	OTG03098	PS51375 Pentatricopeptide (PPR) repeat
HanXRQChr13g0420371	175253986	175294219	–	Putative tetratricopeptide-like helical domain	OTG03101	PS51375 Pentatricopeptide (PPR) repeat
HanXRQChr13g0420861	177597409	177599211	+	Putative tetratricopeptide-like helical domain	OTG03145	PS51375 Pentatricopeptide (PPR) repeat
HanXRQChr13g0420881	177609240	177611054	+	Putative tetratricopeptide-like helical domain	OTG03147	PS51375 Pentatricopeptide (PPR) repeat
HanXRQChr13g0421271	178655189	178657150	–	Probable tetratricopeptide repeat (TPR)-like superfamily protein	OTG03183	PS51375 Pentatricopeptide (PPR) repeat
HanXRQChr13g0419821	174082899	174091500	–	Probable aldehyde dehydrogenase 5F1	OTG03054	NA

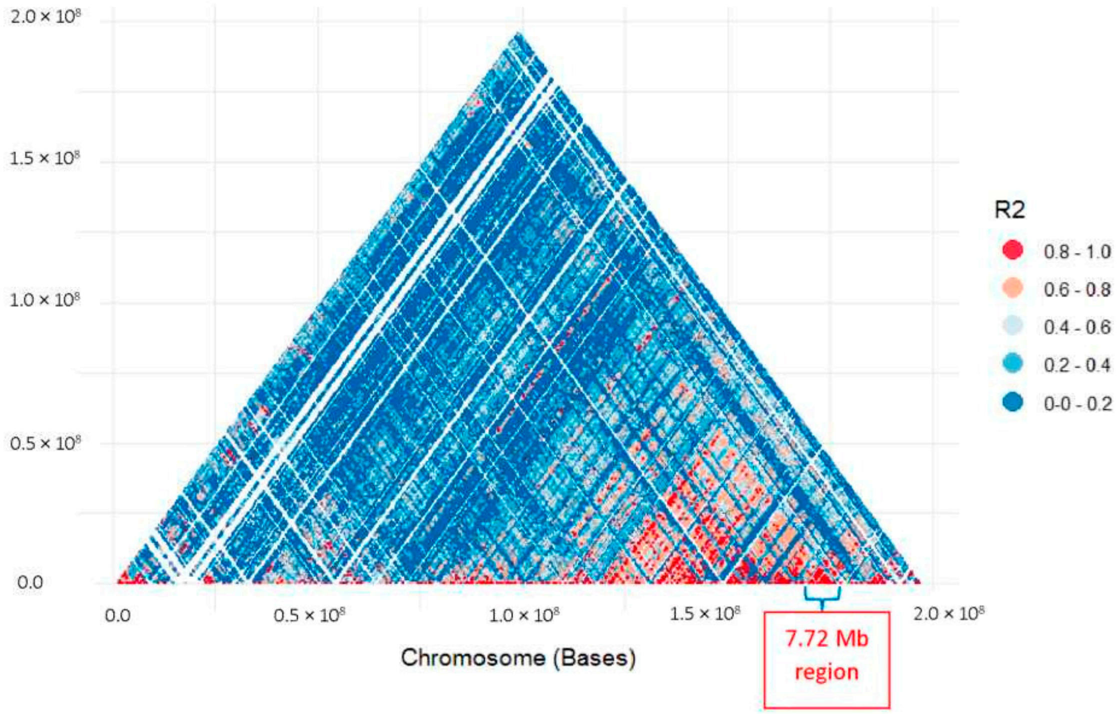

Figure 4. Pairwise Linkage Disequilibrium (LD) Plot of the LG13. Individual data points reflect squared allele frequency correlations (R2) for all possible pairs of polymorphic SNP markers of LG13. The x- and y-axes correspond to the coordinates within 13 LG. Location of 7.72 Mb indicated by curly bracket.

It should be noted that the reference genome used in the analysis was obtained by sequencing the XRQ line, which is a cytoplasmic male sterility maintainer (PET1 type) [36]. At the same time, it is known that the *Rf* locus may undergo complex evolutionary events [46] and the structure of the identified site may differ in the genome of the fertility restorer lines. Therefore, to identify the *Rf1* gene, to determine the sequence of the dominant alleles of the *Rf1* gene, and to understand the evolution of the sunflower *Rf1* locus, the additional analysis of the structure of the 7.72 Mb region in the genome of fertility restorer lines is required.

4. Conclusions

In this work, high-throughput genotyping of sunflower lines, differing by the ability to suppress CMS phenotype was carried out and a genome-wide association study was performed. The GWAS results made it possible to isolate a segment 7.72 Mb in length on chromosome 13, in which 21 candidate *Rf1* fertility restorer genes were identified, including 20 *PPR*-family genes and one Probable aldehyde dehydrogenase gene. The results will serve as a basis for further study of the genetic nature and molecular mechanisms for pollen fertility restoration in sunflower, as well as for the search of selection markers.

Supplementary Materials: The following are available online at http://www.mdpi.com/2073-4395/9/2/49/s1, Table S1: List of the sunflower lines selected for the study, Table S2: List of the sequenced sunflower samples, corresponding flowcell lanes and barcodes, Table S3: List of SNPs, significantly associated with the ability to suppress CMS phenotype after FDR correction, Figure S1: Quantile-quantile plot of associations with the ability to suppress CMS phenotype, Figure S2: Distribution of the ability to suppress CMS phenotype across sunflower samples with different allelic states for 8 statistically significant markers: (A) S13_170494693, (B) S13_171053833, (C) S13_173268042, (D) S13_173832391, (E) S13_174474103, (F) S13_174474122, (G) S13_174809087, (H) S13_178217103.

Author Contributions: Conceptualization, S.V.G. (Svetlana V. Goryunova), I.N.A.; methodology, N.I.B., Y.N.D., P.E.K.; formal analysis, D.V.G., S.V.G. (Svetlana V. Goryunova), E.A.S.; investigation, A.I.C., E.U.M., S.V.B., A.F.A., E.A.G.; writing—original draft preparation, D.V.G., S.V.G. (Svetlana V. Goryunova), I.N.A., V.A.G.; writing—review and editing, A.A.S., P.V.M., R.F.G., S.V.G. (Sergey V. Garkusha); supervision, P.E.K.; project administration, D.A.P.; funding acquisition, D.V.G., S.V.G. (Svetlana V. Goryunova), P.E.K., Z.M.M.

Funding: This research was funded by the Ministry of Science and Higher Education of the Russian Federation (Grant No. 14.609.21.0099, Identification No. RFMEFI60916X0099).

Acknowledgments: We are grateful to S.V. Nuzhdin, S.A. Spirin and A.V. Alexeyevsky for valuable discussion. We would also like to thank I.N. Benko and A. Yu. Donchenko for technical support.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Rhoades, M.M. Cytoplasmic inheritance of male sterility in *Zea mays*. *Science* **1931**, *73*, 340–341. [CrossRef] [PubMed]
2. Schnable, P. The molecular basis of cytoplasmic male sterility and fertility restoration. *Trends Plant Sci.* **1998**, *3*, 175–180. [CrossRef]
3. Ivanov, M.K.; Dymshits, G.M. Cytoplasmic male sterility and restoration of pollen fertility in higher plants. *Russ. J. Genet.* **2007**, *43*, 354–368. [CrossRef]
4. Anisimova, I.N.; Gavrilova, V.A.; Rozhkova, V.T.; Timofeeva, G.I.; Tikhonova, M.A. Molecular markers in identification of sunflower pollen fertility restorer genes. *Russ. Agric. Sci.* **2009**, *35*, 367–370. [CrossRef]
5. Leclercq, P. Une sterilite male cytoplasmique chez le tournesol. *Ann. Amel. Plantes* **1969**, *19*, 99–106.
6. Horn, R.; Köhler, R.H.; Zetsche, K. A mitochondrial 16 kDa protein is associated with cytoplasmic male sterility in sunflower. *Plant Mol. Biol.* **1991**, *17*, 29–36. [CrossRef] [PubMed]
7. Hans Köhler, R.; Horn, R.; Lössl, A.; Zetsche, K. Cytoplasmic male sterility in sunflower is correlated with the co-transcription of a new open reading frame with the *atpA* gene. *MGG Mol. Gen. Genet.* **1991**, *227*, 369–376. [CrossRef]
8. Miller, J.F.; Fick, G.N. Genetics of sunflower. In *Sunflower Technology and Production*; American Society of Agronomy, Crop Science Society of America, Soil Science Society of America: Madison, WI, USA, 1997; pp. 441–495. ISBN 978-0-89118-135-4.
9. Anaschenko, A.V.; Duca, M.V. Studies of sunflower (*Helianthus annuus* L.) CMS–*Rf* genetic system: II. Male fertility restoration in hybrids based on CMSP. *Russ. J. Genet.* **1985**, *21*, 1999–2004.
10. Serieys, H. Identification, study and utilization in breeding programs of new CMS sources. *Helia* **1996**, *19*, 144–158.
11. Gentzbittel, L.; Vear, F.; Zhang, Y.-X.; Bervillé, A.; Nicolas, P. Development of a consensus linkage RFLP map of cultivated sunflower (*Helianthus annuus* L.). *Theor. Appl. Genet.* **1995**, *90*, 1079–1086. [CrossRef]
12. Jan, C.C.; Vick, B.A.; Miller, J.F.; Kahler, A.L.; Butler, E.T. Construction of an RFLP linkage map for cultivated sunflower. *Theor. Appl. Genet.* **1998**, *96*, 15–22. [CrossRef]
13. Kusterer, B.; Horn, R.; Friedt, W. Molecular mapping of the fertility restoration locus *Rf1* in sunflower and development of diagnostic markers for the restorer gene. *Euphytica* **2005**, *143*, 35–42. [CrossRef]
14. Tang, S.; Yu, J.-K.; Slabaugh, B.; Shintani, K.; Knapp, J. Simple sequence repeat map of the sunflower genome. *Theor. Angew. Genet.* **2002**, *105*, 1124–1136. [CrossRef]
15. Horn, R.; Kusterer, B.; Lazarescu, E.; Prüfe, M.; Friedt, W. Molecular mapping of the *Rf1* gene restoring pollen fertility in PET1-based F1 hybrids in sunflower (*Helianthus annuus* L.). *Theor. Appl. Genet.* **2003**, *106*, 599–606. [CrossRef] [PubMed]
16. Markin, N.; Usatov, A.; Makarenko, M.; Azarin, K.; Gorbachenko, O.; Kolokolova, N.; Usatenko, T.; Markina, O.; Gavrilova, V. Study of informative DNA markers of the *Rf1* gene in sunflower for breeding practice. *Czech J. Genet. Plant Breed.* **2017**, *53*, 69–75. [CrossRef]
17. Anisimova, I.N.; Gavrilova, V.A.; Rozhkova, V.T.; Port, A.I.; Timofeeva, G.I.; Duka, M.V. Genetic diversity of sources of sunflower pollen fertility restorer genes. *Russ. Agric. Sci.* **2011**, *37*, 192–196. [CrossRef]
18. Sykes, T.; Yates, S.; Nagy, I.; Asp, T.; Small, I.; Studer, B. In-silico identification of candidate genes for fertility restoration in cytoplasmic male sterile perennial ryegrass (*Lolium perenne* L.). *Genome Biol. Evol.* **2016**, *9*, 351–362.
19. Horn, R.; Hamrit, S. Gene cloning and characterization. In *Genetics, Genomics and Breeding of Sunflower*; CRC Press: Boca Raton, FL, USA, 2010; pp. 173–219. ISBN 978-1-138-11513-2.
20. Yue, B.; Vick, B.A.; Cai, X.; Hu, J. Genetic mapping for the *Rf1* (fertility restoration) gene in sunflower (*Helianthus annuus* L.) by SSR and TRAP markers. *Plant Breed.* **2010**, *129*, 24–28. [CrossRef]

21. Anisimova, I.N.; Alpatieva, N.V.; Rozhkova, V.T.; Kuznetsova, E.B.; Pinaev, A.G.; Gavrilova, V.A. Polymorphism among RFL-PPR homologs in sunflower (*Helianthus annuus* L.) lines with varying ability for the suppression of the cytoplasmic male sterility phenotype. *Russ. J. Genet.* **2014**, *50*, 712–721. [CrossRef]
22. Cui, X.; Wise, R.P.; Schnable, P.S. The *rf2* nuclear restorer gene of male-sterile T-cytoplasm maize. *Science* **1996**, *272*, 1334–1336. [CrossRef]
23. Itabashi, E.; Iwata, N.; Fujii, S.; Kazama, T.; Toriyama, K. The fertility restorer gene, *Rf2*, for Lead Rice-type cytoplasmic male sterility of rice encodes a mitochondrial glycine-rich protein: *Rf2* for CMS rice encodes a glycine-rich protein. *Plant J.* **2011**, *65*, 359–367. [CrossRef] [PubMed]
24. Melonek, J.; Stone, J.D.; Small, I. Evolutionary plasticity of restorer-of-fertility-like proteins in rice. *Sci. Rep.* **2016**, *6*, 35152. [CrossRef] [PubMed]
25. Lurin, C. Genome-wide analysis of *Arabidopsis* pentatricopeptide repeat proteins reveals their essential role in organelle biogenesis. *Plant Cell* **2004**, *16*, 2089–2103. [CrossRef] [PubMed]
26. Cheng, S.; Gutmann, B.; Zhong, X.; Ye, Y.; Fisher, M.F.; Bai, F.; Castleden, I.; Song, Y.; Song, B.; Huang, J.; et al. Redefining the structural motifs that determine RNA binding and RNA editing by pentatricopeptide repeat proteins in land plants. *Plant J. Cell Mol. Biol.* **2016**, *85*, 532–547. [CrossRef] [PubMed]
27. Fujii, S.; Bond, C.S.; Small, I.D. Selection patterns on restorer-like genes reveal a conflict between nuclear and mitochondrial genomes throughout angiosperm evolution. *Proc. Natl. Acad. Sci. USA* **2011**, *108*, 1723–1728. [CrossRef] [PubMed]
28. Kaur, P.; Verma, M. Insights into *PPR* Gene Family in *Cajanus cajan* and other legume species. *J. Data Min. Genom. Proteom.* **2016**, *7*. [CrossRef]
29. Dahan, J.; Mireau, H. The *Rf* and *Rf*-like PPR in higher plants, a fast-evolving subclass of *PPR* genes. *RNA Biol.* **2013**, *10*, 1469–1476. [CrossRef]
30. Gaborieau, L.; Brown, G.G.; Mireau, H. The propensity of pentatricopeptide repeat genes to evolve into restorers of cytoplasmic male sterility. *Front. Plant Sci.* **2016**, *7*, 1816. [CrossRef]
31. Rieseberg, L.H.; Blackman, B.K. Speciation genes in plants. *Ann. Bot.* **2010**, *106*, 439–455. [CrossRef]
32. Zhigunov, A.V.; Ulianich, P.S.; Lebedeva, M.V.; Chang, P.L.; Nuzhdin, S.V.; Potokina, E.K. Development of F1 hybrid population and the high-density linkage map for European aspen (*Populus tremula* L.) using RADseq technology. *BMC Plant Biol.* **2017**, *17*, 180. [CrossRef]
33. Bolger, A.M.; Lohse, M.; Usadel, B. Trimmomatic: A flexible trimmer for Illumina sequence data. *Bioinformatics* **2014**, *30*, 2114–2120. [CrossRef] [PubMed]
34. Bradbury, P.J.; Zhang, Z.; Kroon, D.E.; Casstevens, T.M.; Ramdoss, Y.; Buckler, E.S. TASSEL: Software for association mapping of complex traits in diverse samples. *Bioinform. Oxf. Engl.* **2007**, *23*, 2633–2635. [CrossRef] [PubMed]
35. Langmead, B.; Trapnell, C.; Pop, M.; Salzberg, S.L. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol.* **2009**, *10*, R25. [CrossRef] [PubMed]
36. Badouin, H.; Gouzy, J.; Grassa, C.J.; Murat, F.; Staton, S.E.; Cottret, L.; Lelandais-Brière, C.; Owens, G.L.; Carrère, S.; Mayjonade, B.; et al. The sunflower genome provides insights into oil metabolism, flowering and Asterid evolution. *Nature* **2017**, *546*, 148–152. [CrossRef]
37. Wickham, H. *ggplot2*; Springer: New York, NY, USA, 2009; ISBN 978-0-387-98140-6.
38. Zhang, Z.; Ersoz, E.; Lai, C.-Q.; Todhunter, R.J.; Tiwari, H.K.; Gore, M.A.; Bradbury, P.J.; Yu, J.; Arnett, D.K.; Ordovas, J.M.; et al. Mixed linear model approach adapted for genome-wide association studies. *Nat. Genet.* **2010**, *42*, 355–360. [CrossRef] [PubMed]
39. Chang, C.C.; Chow, C.C.; Tellier, L.C.; Vattikuti, S.; Purcell, S.M.; Lee, J.J. Second-generation PLINK: Rising to the challenge of larger and richer datasets. *GigaScience* **2015**, *4*, 7. [CrossRef] [PubMed]
40. Purcell, S.; Neale, B.; Todd-Brown, K.; Thomas, L.; Ferreira, M.A.R.; Bender, D.; Maller, J.; Sklar, P.; de Bakker, P.I.W.; Daly, M.J.; et al. PLINK: A tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet.* **2007**, *81*, 559–575. [CrossRef]
41. Turner, S.D. QQman: An R package for visualizing GWAS results using QQ and manhattan plots. *BioRxiv* **2014**. [CrossRef]
42. Tang, S.; Leon, A.; Bridges, W.C.; Knapp, S.J. Quantitative trait loci for genetically correlated seed traits are tightly linked to branching and pericarp pigment loci in sunflower. *Crop Sci.* **2006**, *46*, 721. [CrossRef]

43. Nambeesan, S.U.; Mandel, J.R.; Bowers, J.E.; Marek, L.F.; Ebert, D.; Corbi, J.; Rieseberg, L.H.; Knapp, S.J.; Burke, J.M. Association mapping in sunflower (*Helianthus annuus* L.) reveals independent control of apical vs. basal branching. *BMC Plant Biol.* **2015**, *15*, 84. [CrossRef]
44. Yu, J.-K.; Tang, S.; Slabaugh, M.B.; Heesacker, A.; Cole, G.; Herring, M.; Soper, J.; Han, F.; Chu, W.-C.; Webb, D.M.; et al. Towards a saturated molecular genetic linkage map for cultivated sunflower. *Crop Sci.* **2003**, *43*, 367. [CrossRef]
45. Kusterer, B.; Rozynek, B.; Brahm, L.; Prüfe, M.; Tzigos, S.; Horn, R.; Friedt, W. Construction of a genetic map and localization of major traits in sunflower (*Helianthus annuus* L.). *Helia* **2004**, *27*, 15–23. [CrossRef]
46. Hernandez Mora, J.R.; Rivals, E.; Mireau, H.; Budar, F. Sequence analysis of two alleles reveals that intra-and intergenic recombination played a role in the evolution of the radish fertility restorer (*Rfo*). *BMC Plant Biol.* **2010**, *10*, 35. [CrossRef] [PubMed]
47. Kazama, T.; Toriyama, K. A fertility restorer gene, *Rf4*, widely used for hybrid rice breeding encodes a pentatricopeptide repeat protein. *Rice* **2014**, *7*, 28. [CrossRef] [PubMed]
48. Madugula, P.; Uttam, A.G.; Tonapi, V.A.; Ragimasalawada, M. Fine mapping of *Rf2*, a major locus controlling pollen fertility restoration in sorghum A1 cytoplasm, encodes a PPR gene and its validation through expression analysis. *Plant Breed.* **2018**, *137*, 148–161. [CrossRef]
49. Bentolila, S.; Alfonso, A.A.; Hanson, M.R. A pentatricopeptide repeat-containing gene restores fertility to cytoplasmic male-sterile plants. *Proc. Natl. Acad. Sci. USA* **2002**, *99*, 10887–10892. [CrossRef]
50. Schmitz-Linneweber, C.; Small, I. Pentatricopeptide repeat proteins: A socket set for organelle gene expression. *Trends Plant Sci.* **2008**, *13*, 663–670. [CrossRef]
51. Aubourg, S.; Boudet, N.; Kreis, M.; Lecharny, A. In *Arabidopsis thaliana*, 1% of the genome codes for a novel protein family unique to plants. *Plant Mol. Biol.* **2000**, *42*, 603–613. [CrossRef]
52. Small, I.D.; Peeters, N. The PPR motif—A TPR-related motif prevalent in plant organellar proteins. *Trends Biochem. Sci.* **2000**, *25*, 46–47. [CrossRef]
53. Wei, K.; Han, P. Pentatricopeptide repeat proteins in maize. *Mol. Breed.* **2016**, *36*. [CrossRef]
54. Manna, S. An overview of pentatricopeptide repeat proteins and their applications. *Biochimie* **2015**, *113*, 93–99. [CrossRef] [PubMed]
55. Small, I.D.; Rackham, O.; Filipovska, A. Organelle transcriptomes: Products of a deconstructed genome. *Curr. Opin. Microbiol.* **2013**, *16*, 652–658. [CrossRef] [PubMed]
56. Barkan, A.; Small, I. Pentatricopeptide repeat proteins in plants. *Annu. Rev. Plant Biol.* **2014**, *65*, 415–442. [CrossRef] [PubMed]
57. Liu, Z.; Wang, D.; Feng, J.; Seiler, G.J.; Cai, X.; Jan, C.-C. Diversifying sunflower germplasm by integration and mapping of a novel male fertility restoration gene. *Genetics* **2013**, *193*, 727–737. [CrossRef] [PubMed]