

RESEARCH ARTICLE

Open Access

De novo sequencing of sunflower genome for SNP discovery using RAD (Restriction site Associated DNA) approach

Venkatramana Pegadaraju^{1*}, Rick Nipper², Brent Hulke³, Lili Qi³ and Quentin Schultz¹

Abstract

Background: Application of Single Nucleotide Polymorphism (SNP) marker technology as a tool in sunflower breeding programs offers enormous potential to improve sunflower genetics, and facilitate faster release of sunflower hybrids to the market place. Through a National Sunflower Association (NSA) funded initiative, we report on the process of SNP discovery through reductive genome sequencing and local assembly of six diverse sunflower inbred lines that represent oil as well as confection types.

Results: A combination of Restriction site Associated DNA Sequencing (RAD-Seq) protocols and Illumina paired-end sequencing chemistry generated high quality 89.4 M paired end reads from the six lines which represent 5.3 GB of the sequencing data. Raw reads from the sunflower line, RHA 464 were assembled *de novo* to serve as a framework reference genome. About 15.2 Mb of sunflower genome distributed over 42,267 contigs were obtained upon assembly of RHA 464 sequencing data, the contig lengths ranged from 200 to 950 bp with an N₅₀ length of 393 bp. SNP calling was performed by aligning sequencing data from the six sunflower lines to the assembled reference RHA 464. On average, 1 SNP was located every 143 bp of the sunflower genome sequence. Based on several filtering criteria, a final set of 16,467 putative sequence variants with characteristics favorable for Illumina Infinium Genotyping Technology (IGT) were mined from the sequence data generated across six diverse sunflower lines.

Conclusion: Here we report the molecular and computational methodology involved in SNP development for a complex genome like sunflower lacking reference assembly, offering an attractive tool for molecular breeding purposes in sunflower.

Keywords: Single nucleotide polymorphism (SNP), Restriction site associated DNA sequencing (RAD-Seq)

Background

Domestic sunflower (*Helianthus annuus L.*, $2n = 2x = 34$, haploid genome size ~3.5 Gbp) is native to North America [1] and widely cultivated as oilseed and confection crop types. Besides being economically important, sunflower also serves as model in ecological and evolutionary studies [2-4]. A major focus in both public and private sunflower breeding programs has been to develop sunflower hybrid varieties with improved yield, oil content and resistance to a wide range of diseases. Breeding new hybrids by conventional practices mostly is slow and uncertain;

however, application of molecular markers can improve efficiency of plant selection, saving time and providing accuracy in a breeding program [5-8]. A wide range of molecular markers such as RFLP, AFLP, SSR and TRAP developed in sunflower have successfully enabled construction of high density genetic maps [9-15] and led to identification of molecular markers linked to disease resistance genes [16-19]. However, in general, practical usage of these markers for routine breeding purposes is limited due to high assay cost, low reproducibility, and lack of QTL validation studies [20].

In recent years, SNP markers have gained popularity in crop breeding programs due to their low cost, high throughput efficiency, and abundance. Particularly in association mapping studies, SNPs are the preferred

* Correspondence: peg_ven@yahoo.com

¹BioDiagnostics, Inc, 507 Highland Drive, River Falls, WI 54022, USA
Full list of author information is available at the end of the article

marker type since they involves scanning whole genomes with extremely high marker densities to identify closely linked markers to causal polymorphisms [21,22]. It is estimated that due to low linkage disequilibrium and high haplotype diversity, SNPs in the order of several thousand would be needed to successfully conduct genome wide association analysis in sunflower [23].

Large-scale discovery of genome-wide distributed SNPs can be effectively conducted with the aid of massively parallel, next-generation sequencing (NGS) technologies [24]. Several studies that involve whole genome sequencing (WGS) efforts have led to the successful discovery of SNPs in *Arabidopsis* [25], humans [26], and *Medicago* [27]. NGS technologies have also been extended for SNP discovery in large and complex genomes that lack an assembled reference genome [28,29]. A common approach in these situations is the use of a complexity reduction strategy that is designed to selectively interrogate a small percentage of the target genome [30,31]. By restricting sequencing on a smaller fraction of the genome, overall sequencing costs are reduced compared to WGS strategies, while still identifying a large amount of genetic variation. For instance, by using an RNA sequencing approach (RNA-Seq) on tissues from two diverse maize inbred lines, more than 4900 SNPs associated with 2,400 genes were identified and validated [32]. Similarly, previous work in sunflower produced nearly 10,000 SNPs with RNA-Seq [29].

In addition, a cadre of methods have been developed that involve the usage of restriction enzymes on genomic DNA for complexity reduction. These strategies can be at the nucleotide level and are viewed as simple and highly efficient methods in plant and animal genome sequencing studies. One such method, CRoPS (complexity reduction of polymorphic sequences) can overcome the problems associated with highly duplicated regions in complex genomes that hamper the process of SNP identification [33]. Restriction site-associated DNA sequencing (RAD-Seq) is an emerging method for SNP detection in genomes and is based on identifying polymorphic variants adjacent to restriction enzyme digestion sites [34,35]. Application of RAD-Seq for identifying genetic variants has been demonstrated in a variety of species with and without an available reference genome [36,37]. More recently paired-end RAD-Seq (RAD-PE) has been used in a variety of efforts for both genome assembly and SNP marker development [38,39].

Furthermore, RAD-Seq approach has also been exploited in wide range of other studies such as association mapping [40], population genetics inferences [41-44], genetic mapping [35,45,46] and in estimation of allele frequencies [47]. RAD-Seq differs from RNA-Seq in that non-transcribed loci are also sequenced, thus affording us an opportunity to broaden the known SNPs in sunflower to include those

outside of transcribed regions. Here we demonstrate the use of paired-end RAD sequencing to enable efficient, cost-effective, high throughput marker development in *H. annuus*, a major oil crop without an assembled genome sequence. Results on the use of this sequence resource for detection of sequence variation and design of SNP marker panels for Illumina Infinium Genotyping Technology (IIGT) are also discussed.

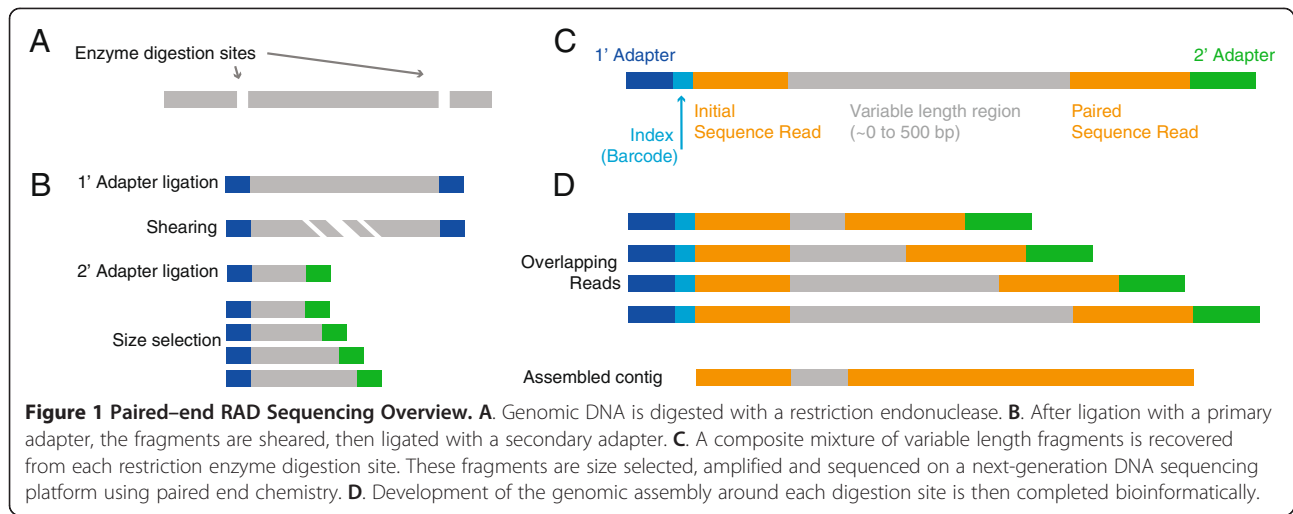
Results and discussion

Paired-end RAD-Seq and *de novo* assembly

Unlike randomized short-insert NGS sequencing methods, RAD genome fragments share a unique architecture: a sequence anchored by the restriction enzyme cleavage site and a variable sequence end generated from a shearing step during library construction (Figure 1). When RAD is coupled with paired-end sequencing approaches now available on NGS platforms, the opposite ends of the RAD fragment are linked in *cis* and the fragment can then be interrogated. Mate-pairs with identical single read sequences can then be readily assembled into contigs spanning hundreds of base pairs (Figure 1C).

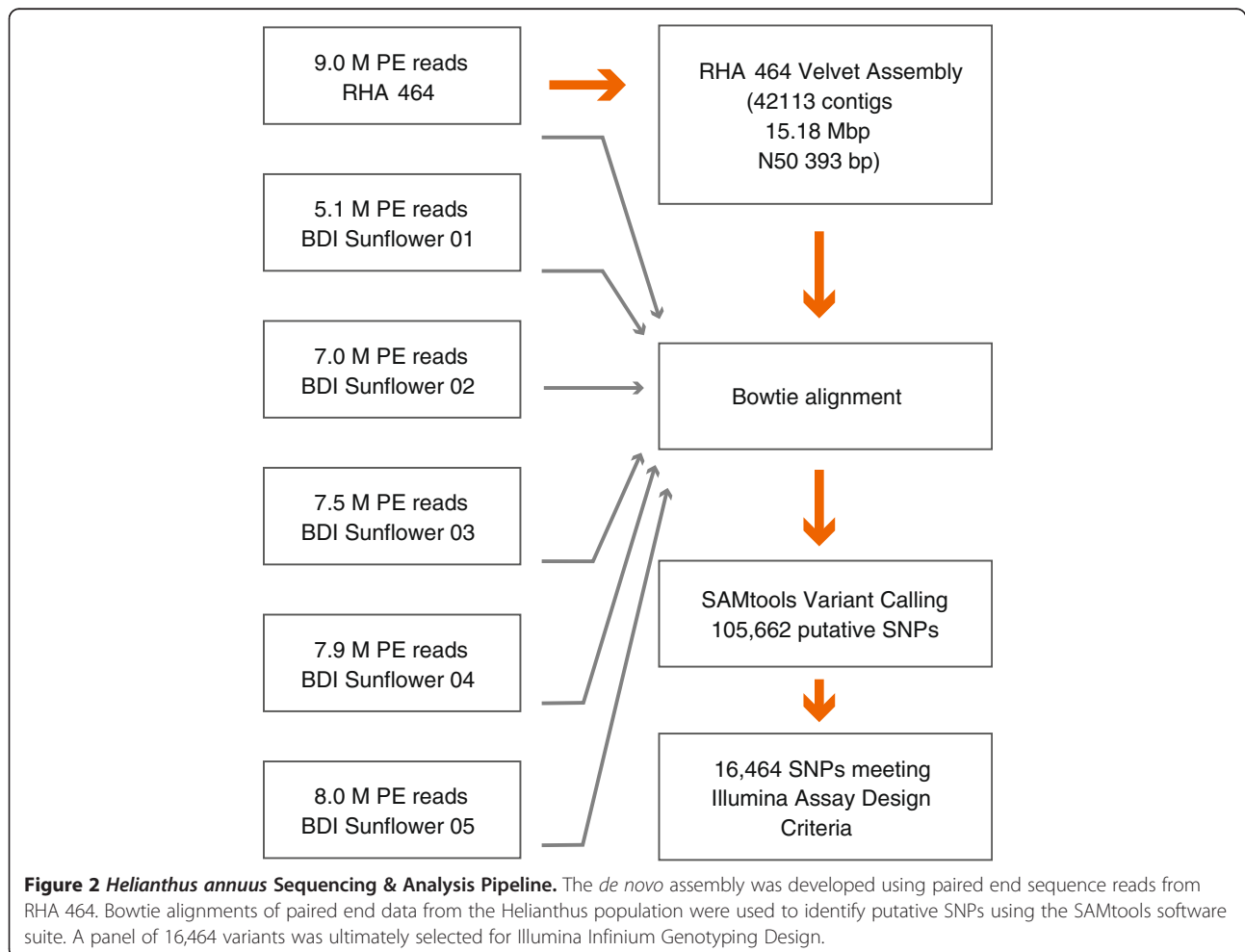
To promote SNP identification in low-copy, gene-rich regions of the 3.5 Gbp sunflower genome, a species expected to contain upwards of 80% retrotransposon content [48], the 5-methylcytosine (5mC) sensitive type II nuclease *PstI* (5'CTGCA/G'3) was selected for RAD-Seq in each of the six lines. Numerous studies have documented retroelement-dense regions of plant genome are often subjected to cytosine methylation of CpG, CpNpG and CpNpN nucleotides [49-51]. Restriction enzymes such as *PstI*, which do not cleave 5mC-modified DNA, have been shown to specifically sample the hypomethylated genomic fraction of plant genomes [52].

Sequencing results for each of the six lines are detailed in Additional file 1. A total of 44.7 M reads (89.4 M paired-end) were obtained from all six lines, representing ~5.3 Gb of sequence data. An overall workflow of sequence analysis performed in this study is encapsulated in Figure 2. To construct a skeleton sunflower reference, reads from line RHA 464 [53] were first coalesced into contigs using the Velvet assembler [54]. Initial *de novo* assembly produced ~15.2 Mb of sunflower genome sequence distributed over 42,267 individual contigs. This quantity of assembled sequence data is approximately half of the sequence content developed from a contemporary whole transcriptome sequencing project [55]. Contig lengths for the RHA assembly ranged between 200 and 920 bp with an N_{50} length of 393 bp for all RHA 464 assemblies (Figure 3A and Table 1). The contig length distribution is in line with the fragment size range selected during RAD-Seq library preparation. After initial assembly, contigs were aligned against a custom sequence database to remove sequences with significant plastid homology



with 42,113 contigs spanning 15.18 Mbp of the sunflower genome remaining. Contigs passing these filters were then evaluated for the presence of repetitive elements using the RepeatMasker web server with the Arabidopsis Repbase library. The percentage of the RAD-Seq RHA 464 assembly

classified as repetitive by RepeatMasker was 1.75%. This is consistent with a genome assembly principally from low-copy regions, as the 3.5 gigabasepair sunflower genome is expected to contain over 80% repetitive nucleotide content. The major classes of repetitive DNA elements that were



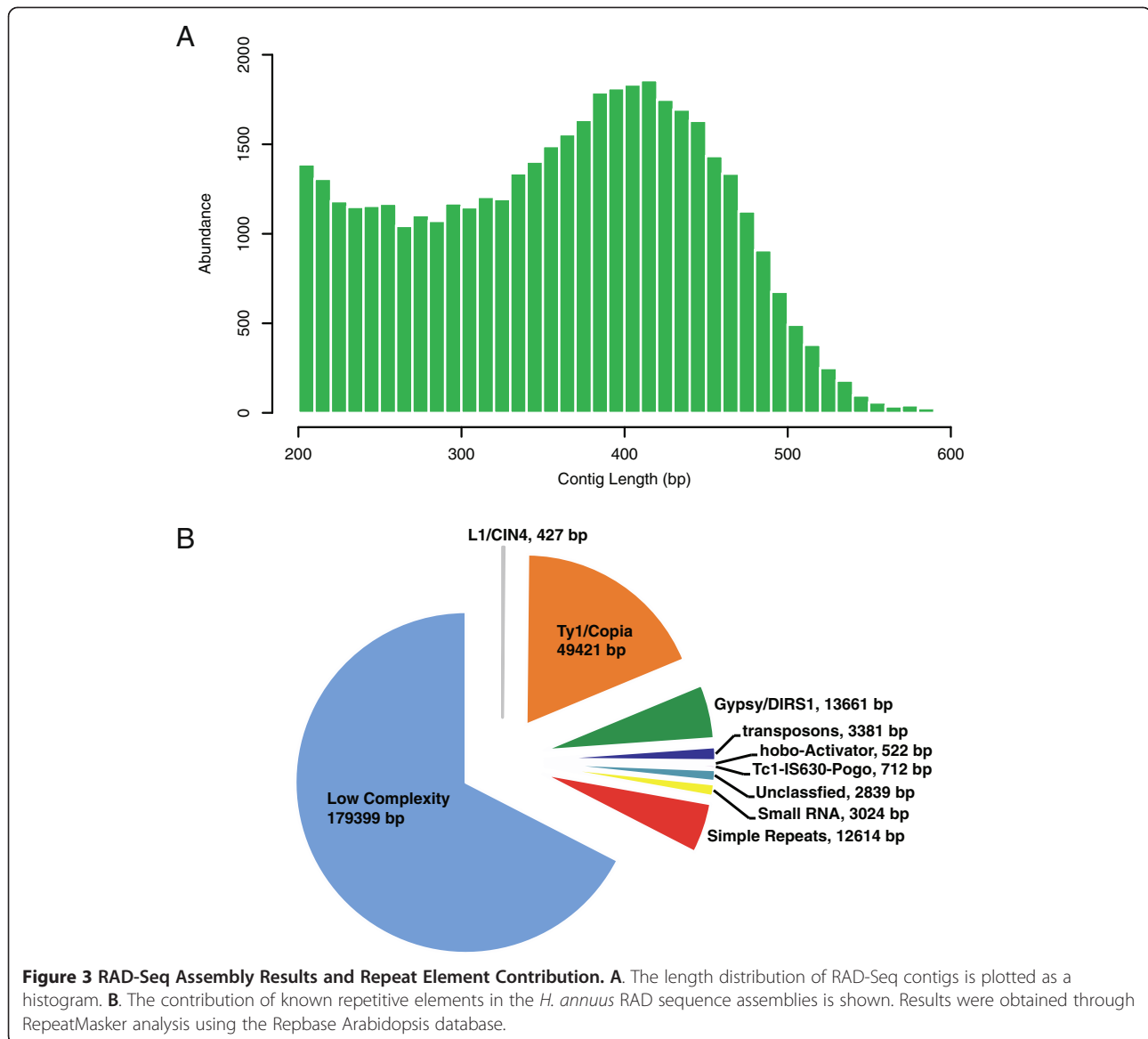


Table 1 Paired-end RAD-Seq assembly statistics

Feature	Value
Number of RHA 464 contigs assembled	42267
Contigs removed due to plastid homology	154
Number of contigs retained	42113
Total assembly length (bp)	15181868
Minimum contig length (bp)	200
Maximum contig length (bp)	920
GC%	36.1
N50 Contig Length (bp)	393
N90 Contig Length (bp)	254

identified belonged to low complexity sequences and Ty/Copia long-terminal repeat (LTR) retroelement families (Figure 3B and Additional file 2). The GC dinucleotide content for RHA 464 sunflower assemblies was 36.2% (Table 1), which is consistent with results from paired-end RAD-Seq studies in other plant genomes [56].

Comparison of RAD-Seq assemblies from RHA 464 to preexisting sunflower unigenes at the Dana Farber Cancer Institute (DFCI) confirmed high sequence identity between RAD assemblies and known sunflower genomic sequences. A representative alignment is shown in Additional file 3 illustrating the match of a single paired-end RAD-Seq contig with tentative consensus EST TC57527 from the *H. annuus* DFCI EST database. The high sequence coverage inherent in paired-end RAD-Seq minimizes sequencing and assembly errors, as each

nucleotide in the contig is derived from the consensus of many overlapping Illumina sequence reads. Experimental paired-end RAD-Seq studies using the sequenced B73 maize cultivar have placed the single base accuracy for paired-end RAD-Seq contigs at approximately 99.95% (unpublished data). This accuracy in RAD-Seq contig assembly is a key advantage for applications involving genome assembly and downstream marker development.

SNP discovery

Alignment of sequence data from all six sunflower inbred lines to the RHA 464 reference used the Bowtie/SAMtools [57,58] variant detection pipeline and revealed the presence of 105,662 putative sunflower SNPs (Table 2). The calculated polymorphism rate in this study is approximately one SNP observed per 143 bp of sunflower genomic sequence. This rate of nucleotide variation agrees well with the levels of genetic diversity reported in recent sunflower studies [29,59,60]. Several additional filtering steps were implemented to pare down the initial SNP dataset to a panel of markers suitable for Infinium genotyping. The full set of 105,662 potential variant loci was scored over the sunflower population. Although SNP genotypes could be reliably called for over 79% of loci ($\geq 4x$ sequence coverage), 11,614 SNPs with missing genotype calls in three or more sunflower lines were not considered for further analysis (Table 2). Genotypes for the remaining 94,048 alleles indicated 89.2% of genotype calls were homozygous in the target population, in line with the inbred nature of the selected lines. The large tracts of sequence landscape generated around candidate SNPs foster the conversion of variants identified from RAD-Seq into downstream genotyping platforms such as the Illumina GoldenGate, Infinium and Sequenom iPLEX systems. Of the 94,048 candidate SNPs, less than 4% had to be removed due to the absence of insufficient

flanking genomic sequence (minimum limit of 50 bp flanking candidate SNP). The vast majority of discarded SNPs were rejected due to the presence of flanking polymorphisms within 50 bp that can interfere with oligo hybridization during genotyping (Table 2). A final set of 16,467 SNPs was identified and considered for Infinium genotyping design.

Analysis of the sequence variation identified in RAD-Seq aligns well with values reported in other plant studies. First, the observed SNP transition/transversion ratio of 1.72 (Figure 4A) in this study is very similar to the distributions reported in other studies of sunflower and eggplant [29,36]. Analysis of the location of variants in assembled RHA 464 contigs indicated no significant bias in allele detection: most SNPs were identified between 100 and 300 bp from the start of each contig (Figure 4B). Finally, approximately 57% of all assembled contigs (23984 out of 42113) contained at least one detectable polymorphism, reflecting the high degree of variation present in sunflower (Figure 4C).

Conclusions

The application of next-generation DNA sequencing to generate large numbers of genetic markers has revolutionized plant breeding, facilitating both molecular genetic research and marker-assisted selection efforts. We have demonstrated paired-end RAD-Seq is an efficient and cost-effective means of SNP discovery in a species with a complex, highly repetitive genome. From less than a flowcell of Illumina paired-end sequence data we sequenced six diverse sunflower lines, assembled over 40,000 high-quality sequence contigs with an N_{50} contig length of 393 nucleotides, mined more than 100,000 sequence variations from the sunflower genome and identified 16,467 candidate SNPs suitable for downstream genotyping. The RAD-Seq method is appropriate for studies where many thousands of SNPs need to be rapidly identified at a low cost, in a format suitable for high-throughput genotyping.

Methods

Plant material and DNA extraction

Sunflower inbred lines (TX16R, CR29, SEEDS2000 B-Line, HA 467, RHA 468, and RHA 464) were grown under laboratory greenhouse conditions for four weeks, all true leaves were harvested and lyophilized prior to DNA extraction. DNA was extracted from 40 mg of each inbred line with the DNeasy 96 Plant Kit (Qiagen) using a modified protocol. Tissue was pulverized with 3 mm beads in a Harbil shaker. Buffer AP1 with DX and RNaseA was added to the tissue, 500 μ L per sample, and incubated at 55°C for 60 min. Buffer AP2 was added, 150 μ L per well, and incubated at -20°C for 15 min. AP3/E was combined with supernatant, 600 μ L and 400 μ L respectively,

Table 2 *Helianthus* SNP filtering and statistics

Feature	Value
<i>Helianthus</i> samples sequenced:	6
Total number of SNP variants identified:	105662
Total possible SNP genotypes in population:	633972
SNP genotypes with high confidence call:	502837 (79.3%)
SNP genotypes with missing or low quality data:	131135 (20.7%)
SNP loci with < 50% genotype data:	11614
SNPs passing initial filters:	94048
Number of fixed genotype calls:	448289 (89.2%)
Number of heterozygous genotype calls:	54548 (10.8%)
SNP Loci with insufficient flanking sequence for IIGT*:	3445
SNP Loci with nearby polymorphism (< 50 bp):	74136
SNP Loci meeting all defined IIGT assay design criteria:	16467

* Illumina Infinium Genotyping Technology.

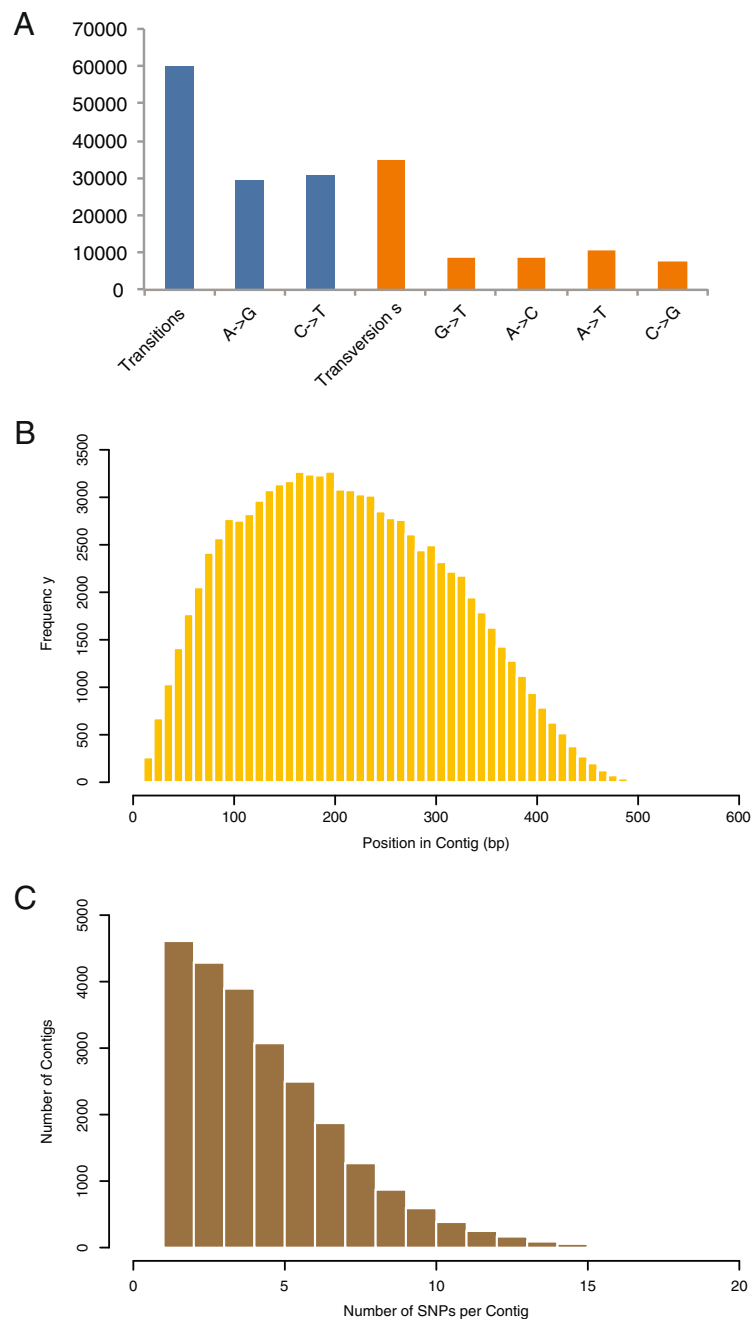


Figure 4 Helianthus SNP Discovery. **A.** The number and ratio of SNP transitions and transversions observed in the Helianthus population is graphed. **B.** The frequency of SNPs by position in each respective contig is plotted. **C.** The number of sequence variations observed across each RAD-Seq contig is shown.

and then added to the binding plates. The rest of the extraction was carried out according to kit instructions. DNA was eluted in a final volume of 50 μ L.

DNA was quantified using the PicoGreen kit (Molecular Probes) according to the kit instructions. A standard curve was made using quantified λ DNA from 100 to 0 ng/ μ L. A 1/200 dilution of Picogreen reagent in 1x TE (provided in kit) was mixed with 2 μ L of isolated DNA, briefly vortexed,

and incubated in the dark for 5 min. Assays were performed in black 96-well Fluotrac plates and fluorescence was measured with a Spectramax Gemini XPS (Molecular Devices) using 485 nm excitation and 538 nm emission.

RAD library preparation protocols

Genomic DNA from six selected sunflower inbred lines (TX16R, CR29, SEEDS2000 B-Line, HA 467, RHA 468,

and RHA 464) was digested with the restriction endonuclease *PstI* and processed into RAD libraries similar to the method of Baird, et al., 2008. Briefly, ~300 ng of genomic DNA was digested for 60 min at 37°C in a 50 µL reaction with 20 units (U) of *PstI* (New England Biolabs [NEB]). After digestion, samples were heat-inactivated for 20 min at 65°C followed by addition of 2.0 µL of 100 nM P1 Adapter(s), a modified Solexa® adapter (Illumina, Inc.). *PstI* P1 adapters each contained a unique multiplex sequence index (barcode) which is read during the first four nucleotides of the Illumina sequence read. 100 nM P1adaptors were added to each sample along with 1 µL of 10 mM rATP (Promega), 1 µL 10× NEB Buffer 4, 1.0 µL (1000 U) T4 DNA Ligase (high concentration, Enzymatics, Inc), and 5 µL H₂O which was then incubated at room temperature (RT) for 20 min. Samples were again heat-inactivated for 20 min at 65°C, pooled and randomly sheared with a Bioruptor (Diagenode) to an average size of 500 bp. Samples were then run out on a 1.5% agarose (Sigma), 0.5X TBE gel, and DNA 300 bp to 800 bp was isolated using a MinElute Gel Extraction Kit (Qiagen). End blunting enzymes (Enzymatics, Inc) were then used to polish the ends of the DNA. Samples were then purified using a MinElute column (Qiagen) and 15 U of Klenow exo⁻ (Enzymatics) was used to add adenine (Fermentas) overhangs on the 3' end of the DNA at 37°C. After subsequent purification, 1 µL of 10 µM P2 adapter, a divergent modified Solexa® adapter (Illumina, Inc.), was ligated to the obtained DNA fragments at 18°C. Samples were again purified and eluted in 50 µL. The eluate was quantified using a Qubit fluorimeter and 20 ng of this product was used in a PCR amplification with 20 µL Phusion Master Mix (NEB), 5 µL of 10 µM modified Solexa® Amplification primer mix (Illumina, Inc.) and up to 100 µL H₂O. Phusion PCR settings followed product guidelines for a total of 18 cycles. Again, samples were gel purified, excising DNA from the 300 to 700 bp size range, and diluted to 1 nM.

Illumina sequencing

A set of RAD libraries generated from lines TX16R, CR29, SEEDS2000 B-Line, HA 467, RHA 468, and RHA 464 was run on an Illumina Genome Analyzer IIx at the University of Oregon High Throughput Sequencing Facility in Eugene, Oregon. Illumina protocols were followed for an asymmetric length paired end sequencing run, with an initial 40 bp read and second 80 bp read.

Bioinformatics – Sequence processing, paired-end RAD-Seq assembly and SNP detection

A combination of open source and proprietary bioinformatics tools was used for processing and sequence analysis. A list of open source programs, versions, and commands used in sequence analysis can be found in a supplemental

file (Additional file 4). Initially, raw sequence data produced on two GAIx sequence lanes were sorted by the appropriate multiplex index (MID) or “barcode” assigned to each sunflower line during RAD-Seq library construction. During de-multiplexing, indexes were trimmed from reads and the remaining sequence segregated to individual sample files. Reads from RHA 464 were then processed to extract low quality sequences. Any sequence with an average phred-scaled quality score below 20 (Q20) over the last 5 base pairs of the read was discarded. Remaining reads were then collapsed into RAD sequence clusters sharing 100% sequence identity across the single end Illumina read. To maximize efficient assembly of sequences we imposed a minimum of 50x and maximum 750x sequence coverage at any RAD sequence cluster. These thresholds were selected for this effort, because single loci with coverage under 50x would be expected to suffer from low sequence coverage ((80 bp × 50)/ 400 bp = 10.0x) resulting in short and fragmented contig assemblies, while loci with greater than 500 identical SE (Single-end) reads may be composed of high-copy contaminant DNA (plastids) or dosage from multiple genomic loci (e.g. retrotransposon derived sequences). The paired end sequences for each RAD locus were extracted from these selected loci and passed to the Velvet sequence assembler (version 1.0.18) for contig assembly [54]. Contigs not reaching a minimum length of 200 bp were excluded from the assembly.

Sequence reads from TX16R, CR29, Seeds 2000 B-line, HA 467, and RHA 468, were aligned to the reference RHA 464 assembly using the short-read aligner Bowtie (version 0.12.5) [57]. Alignment thresholds were specified which allowed up to three base pair mismatches between the 80 bp Illumina read and the reference (>95% identity). Reads not uniquely mapping (e.g. aligning to more than one contig in the RHA 464 reference) were discarded and not considered in the analysis. Bowtie alignments were piped to SAM tools (version 0.1.14) and reformatted into BAM and pileup files for SNP identification [58]. Sequence variants from pileups were then condensed into a variant call format (VCF) file using custom perl scripts. To be considered for genotyping design, a SNP had to have a minimum sequencing coverage of 4x in at least three lines, with at least 50 bp of flanking genomic sequence surrounding the target SNP. Variants with nearby flanking polymorphisms (within 50 bp of the candidate marker) were also excluded from further consideration for Infinium genotyping design.

Additional files

Additional file 1: Illumina Sequencing Results.

Additional file 2: *Helianthus* RAD-Seq RepeatMasker reports.

Additional file 3: Alignment of RAD-Seq contig with Sunflower EST collection at DFCL.

Additional file 4: List of open source programs and commands used in analysis.

Competing interests

This project was conducted in accordance with the NSA SNP Consortium, a public-private partnership between the non-profit National Sunflower Association, public researchers at USDA, and private seed companies. It was formed with the intent of having widespread, economically feasible application of SNP markers for the advancement of sunflower germplasm. NSA encourages public research entities to use the proprietary markers through NSA approved laboratories for research that meets the public research criteria. Such use will not incur any fees for membership or royalties. For-profit researchers may join the consortium.

Authors' contributions

VP designed, coordinated and oversaw the sunflower sequencing, data analysis and drafted the manuscript. RN created RAD libraries for sunflower sequencing and conducted the bioinformatics analysis and assisted in manuscript. BH & LQ were involved in planning the experiments, provided the material for sequencing and preparing the manuscript. QS helped in project planning and provided help in preparing the manuscript. All the authors read and approved the final manuscript.

Acknowledgements

This project was supported by the National Sunflower Association. The authors are grateful to Chondra Carlson, Robert Bialozynski, Mark Blackstad and Ruhii Rai for their technical support in the project. Authors would like to thank Nuseed Americas Inc. & USDA for providing CR29, Seeds 2000-B, and TX16R sunflower lines for the sequencing project.

Author details

¹BioDiagnostics, Inc, 507 Highland Drive, River Falls, WI 54022, USA.

²Floragenex, Inc. 44 West Broadway, Eugene, OR 97401, USA.

³USDA-Agricultural Research Service, Northern Crop Science Laboratory, 1307 18th St. N., Fargo, ND 58102-2765, USA.

Received: 19 February 2013 Accepted: 9 August 2013

Published: 15 August 2013

Reference:

- Harter AV, Gardner KA, Falush D, Lentz DL, Bye RA, Rieseberg LH: **Origin of extant domesticated sunflowers in eastern North America.** *Nature* 2004, **430**:201–205.
- Rieseberg LH, Raymond O, Rosenthal DM, Lai Z, Livingstone K, Nakazato T, Durphy JL, Schwarzbach AE, Donovan LA, Lexer C: **Major ecological transitions in wild sunflowers facilitated by hybridization.** *Science* 2003, **301**:1211–1216.
- Snow AA, Pilson D, Rieseberg LH, Paulsen M, Pleskac N, Reagon MR, Wolf DE: **A Bt transgene reduces herbivory and enhances fecundity in wild sunflowers.** *Ecol Appl* 2003, **13**:279–286.
- Lexer C, Lai Z, Rieseberg LH: **Candidate gene polymorphisms associated with salt tolerance in wild sunflower hybrids: Implications for the origin of *Helianthus paradoxus*, a diploid hybrid species.** *New Phytol* 2004, **161**:225–233.
- Mackill DJ, Nguyen HT, Zhan J: **Use of molecular markers in plant improvement programs for rain fed lowland rice.** *Field Crops Res* 1999, **64**:177–185.
- McCouch SR, Doerge RW: **QTL mapping in rice.** *Trends Genet* 1995, **11**:482–487.
- Snowdon RJ, Friedt W: **Molecular markers in Brassica oilseed breeding: Current status and future possibilities.** *Plant Breeding* 2004, **123**:1–8.
- Collard BCY, Jahufer MZZ, Brouwer JB, Pang ECK: **An introduction to markers, quantitative trait loci (QTL) mapping and marker-assisted selection for crop improvement: The basic concepts.** *Euphytica* 2005, **142**:169–196.
- Berry ST, Leon AJ, Hanfrey CC, Challis P, Burkholz A: **Molecular marker analysis of *Helianthus-annuus* L. 2. Construction of an RFLP linkage map for cultivated sunflower.** *Theor Appl Genet* 1995, **91**:195–199.
- Gentzbittel LF, Vear YX, Zhang A, Berville, Nicolas P: **Development of a consensus linkage RFLP map of cultivated sunflower (*Helianthus-annuus* L).** *Theor Appl Genet* 1995, **90**:1079–1086.
- Rieseberg LH, Choi HC, Chan R, Spore C: **Genomic map of a diploid hybrid species.** *Heredity* 1993, **70**:285–293.
- Gedil MA, Wye C, Berry S, Segers B, Peleman J: **An integrated restriction fragment length polymorphism - amplified fragment length polymorphism linkage map for cultivated sunflower.** *Genome* 2001, **44**:213–221.
- Heesacker A, Kishore VK, Gao W, Tang S, Kolkman JM: **SSRs and INDELs mined from the sunflower EST database: Abundance, polymorphisms, and cross-taxa utility.** *Theor Appl Genet* 2008, **117**:1021–1029.
- Tang S, Yu JK, Slabaugh MB, Shintani DK, Knapp SJ: **Simple sequence repeat map of the sunflower genome.** *Theor Appl Genet* 2002, **105**:1124–1136.
- Yu JK, Tang S, Slabaugh MB, Heesacker A, Cole G: **Towards a saturated molecular genetic linkage map for cultivated sunflower.** *Crop Sci* 2003, **43**:367–387.
- Tang S, Kishore VK, Knapp SJ: **PCR-multiplexes for a genome-wide framework of simple sequence repeat marker loci in cultivated sunflower.** *Theor Appl Genet* 2003, **107**:6–19.
- Qi LL, Hulke BS, Vick BA, Gulya TJ: **Molecular mapping of the rust resistance gene R4 to a large NBS-LRR cluster on linkage group 13 of sunflower.** *Theor Appl Genet* 2011, **123**:351–358.
- Lawson WR, Jan CC, Shatte T, Smith L, Kong GA, Kochman JK: **DNA markers linked to the R2 rust resistance gene in sunflower (*Helianthus annuus* L.) Facilitate anticipatory breeding for this disease variant.** *Mol Breeding* 2011, **28**:569–576.
- Bachlava E, Radwan OE, Abratti G, Tang S, Gao W, Heesacker AF, Bazzalo ME, Zambelli A, Leon AJ, Knapp SJ: **Downy mildew (PI8 and PI14) and rust (RADv) resistance genes reside in close proximity to tandemly duplicated clusters of non-TIR-like NBS-LRR-encoding genes on sunflower Chromosomes 1 and 13.** *Theor Appl Genet* 2011, **122**:1211–1221.
- Bernardo R: **Molecular markers and selection for complex traits in plants: Learning from the Last 20 Years.** *Crop Sci* 2008, **48**:1649–1664.
- Ganal MW, Altmann T, Röder MS: **SNP identification in crop plants.** *Curr Opin Plant Biol* 2009, **12**:211–217.
- Rafalski A: **Applications of single nucleotide polymorphisms in crop genetics.** *Curr Opin Plant Biol* 2002, **5**:94–100.
- Zhu C, Gore M, Buckler ES, Yu J: **Status and prospects of association mapping in plants.** *Plant Gen* 2008, **1**:5–20.
- Davey JW, Hohenlohe PA, Etter PD, Boone JQ, Catchen JM, Blaxter ML: **Genomewide genetic marker discovery and genotyping using next-generation sequencing.** *Nat Rev Genet* 2011, **12**:499–510.
- Weber AP, Weber KL, Carr K, Wilkerson C, Ohlrogge JB: **Sampling the *Arabidopsis* transcriptome with massively parallel pyrosequencing.** *Plant Physiol* 2007, **144**:32–42.
- Wheeler DA, Srinivasan M, Egholm M, Shen Y, Chen L, McGuire A, He W, Chen YJ, Makhijani V, Roth GT, Gomes X, Tartaro K, Niazi F, Turcotte CL, Izzyk GP L Jr, Chinault C, Song XZ, Liu Y, Yuan Y, Nazareth L, Qin X, Muzny DM, Margulies M, Weinstock GM, Gibbs RA, Rothberg JM: **The complete genome of an individual by massively parallel DNA sequencing.** *Nature* 2008, **452**:872–876.
- Cheung F, Haas BJ, Goldberg SMD, May GD, Xiao Y, Town CD: **Sequencing *Medicago truncatula* expressed sequenced tags using 454 Life Sciences technology.** *BMC Genomics* 2006, **7**:272.
- You FM, Huo N, Deal KR, Gu YQ, Luo M-C, McGuire PE, Dvorak J, Anderson OD: **Annotation-based genome-wide SNP discovery in the large and complex *Aegilops tauschii* genome using next-generation sequencing without a reference genome sequence.** *BMC Genomics* 2011, **12**:59.
- Bachlava E, Taylor CA, Tang S, Bowers JE M Jr, Burke JM, Knapp SJ: **SNP discovery and development of a high density genotyping array for sunflower.** *PLoS ONE* 2012, **7**:e29814.
- Wu X, Ren C, Joshi T, Vuong T, Xu D, Nguyen HT: **SNP discovery by high throughput sequencing in soybean.** *BMC Genomics* 2010, **11**:469.
- Varshney RK, Nayak SN, May GD, Jackson SA: **Next-generation sequencing technologies and their implications for crop genetics and breeding.** *Trends Biotechnol* 2009, **27**:522–530.
- Barbazuk WB, Emrich SJ, Chen HD, Li L, Schnable PS: **SNP discovery via 454 transcriptome sequencing.** *Plant J* 2007, **51**:910–918.
- Orsouw NJV, Hogers R, Janssen A, Yalcin F, Snoeiijers S, Verstege E, Schneiders H, Van Der PH, Van OJ, Versteegen H: **Complexity reduction of**

- polymorphic sequences (CROPS): A novel approach for large-scale polymorphism discovery in complex genomes. *PLoS ONE* 2007, **14**:e1172.
34. Miller M, Dunham J, Amores A, Cresko W, Johnson E: **Rapid and cost effective polymorphism identification and genotyping using restriction site associated DNA (RAD) markers.** *Genome Res* 2007, **17**:240–248.
 35. Baird NA, Etter PD, Atwood TS, Currey MC, Shiver AL, Lewis ZA, Selker EU, Cresko WA, Johnson EA: **Rapid SNP discovery and genetic mapping using sequenced RAD markers.** *PLoS ONE* 2008, **3**:e3376.
 36. Barchi L, Lanteri S, Portis E, Acquadro A, Vale G, Toppino L, Rotino GL: **Identification of SNP and SSR markers in eggplant using RAD tag sequencing.** *BMC Genomics* 2011, **12**:304.
 37. Chutimanitsakun Y, Nipper R, Cuesta-Marcos A, Cistue L, Corey A, Filichkina T, Johnson E, Hayes P: **Construction and application for QTL analysis of a Restriction Site Associated DNA (RAD) linkage map in barley.** *BMC Genomics* 2011, **12**:4.
 38. Etter PD, Preston JL, Bassham S, Cresko WA, Johnson EA: **Local de novo assembly of RAD paired-end contigs using short sequencing reads.** *PLoS ONE* 2011, **6**:e18561.
 39. Nelson JC, Wang S, Wu Y, Li X, Antony G, White FF, Yu J: **Single-nucleotide polymorphism discovery by high-throughput sequencing in sorghum.** *BMC Genomics* 2011, **12**:352.
 40. Parchman TL, Gompert Z, Mudge J, Schilkey FD, Benkman CW, Buerkle CA: **Genome-wide association genetics of an adaptive trait in lodgepole pine.** *Mol Ecol* 2012, **21**:2991–3005.
 41. Emerson KJ, Merz CR, Catchen JM: **Resolving postglacial phylogeography using high-throughput sequencing.** *Proc Natl Acad Sci USA* 2010, **107**:16196–16200.
 42. Gompert Z, Forister ML, Fordyce JA, Nice CC, Williamson RJ, Buerkle CA: **Bayesian analysis of molecular variance in pyrosequences quantifies population genetic structure across the genome of *Lycaeides* butterflies.** *Mol Ecol* 2010, **19**:2455–2473.
 43. Hohenlohe PA, Bassham S, Etter PD, Stiffler N, Johnson EA, Cresko WA: **Population genomics of parallel adaptation in threespine stickleback using sequenced RAD tags.** *PLoS Genetics* 2010, **6**:e1000862.
 44. Arnold B, Corbett-Detig RB, Hartl D, Bomblies K: **RADseq underestimates diversity and introduces genealogical biases due to nonrandom haplotype sampling.** *Mol Ecol* 2013, **22**:3179–3190.
 45. Andolfatto P, Davison D, Erezylmaz D: **Multiplexed shotgun genotyping for rapid and efficient genetic mapping.** *Genome Res* 2011, **21**:610–617.
 46. Pfender W, Saha M, Johnson E, Slabaugh M: **Mapping with RAD (restriction site associated DNA) markers to rapidly identify QTL for stem rust resistance in *Lolium perenne*.** *TAG Theor Appl Genet* 2011, **122**:1467–1480.
 47. Tassell CPV, Smith TPL, Matukumalli LK: **SNP discovery and allele frequency estimation by deep sequencing of reduced representation libraries.** *Nat Methods* 2008, **5**:247–252.
 48. Kane NC, Gill N, King MG, Bowers JE, Berges H, Gouzy J, Bachlava E, Langlade NB, Lai Z, Stewart M, Burke JM, Vincourt P, Knapp SJ, Rieseberg LH: **Progress towards a reference genome for sunflower.** *Botany* 2011, **89**:429–437.
 49. Palmer LE, Rabinowicz PD, O'Shaughnessy AL, Balija VS, Nascimento LU, Dike S, De La BM, Martienssen RA, McCombie WR: **Maize genome sequencing by methylation filtration.** *Science* 2003, **302**:2115–2117.
 50. Rabinowicz PD, Citek R, Budiman MA, Nunberg A, Bedell JA, Lakey N, O'Shaughnessy AL, Nascimento LU, McCombie WR, Martienssen RA: **Differential methylation of genes and repeats in land plants.** *Genome Res* 2005, **15**:1431–1440.
 51. Ma XF, Gustafson JP: **Timing and rate of genome variation in triticales following allopolyploidization.** *Genome* 2006, **49**:950–958.
 52. Reyna-Lopez GE, Simpson J, Ruiz-Herrera J: **Differences in DNA methylation patterns are detectable during the dimorphic transition of fungi by amplification of restriction polymorphisms.** *Mol Gen Genet* 1997, **253**:703–710.
 53. Hulke BS, Miller JF, Gulya TJ: **Registration of the restorer oilseed sunflower germplasm RHA 464 possessing genes for resistance to downy mildew and sunflower rust.** *J Plant Reg* 2010, **4**:249–254.
 54. Zerbino DR, Birney E: **Velvet: Algorithms for de novo short read assembly using de Bruijn graphs.** *Genome Res* 2008, **18**:821–829.
 55. Wei W, Qi X, Wang L, Zhang Y, Hua W, Li D, Lv H, Zhang X: **Characterization of the sesame (*Sesamum indicum* L.) Global transcriptome using Illumina paired-end sequencing and development of EST-SSR markers.** *BMC Genomics* 2011, **12**:451.
 56. Scaglione D, Acquadro A, Portis E, Tirone M, Knapp SJ, Lanteri S: **RAD tag sequencing as a source of SNP markers in *Cynara cardunculus* L.** *BMC Genomics* 2012, **13**:3.
 57. Langmead B, Trapnell C, Pop M, Salzberg SL: **Ultrafast and memory-efficient alignment of short DNA sequences to the human genome.** *Genome Biol* 2009, **10**:R25.
 58. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R: **1000 Genome Project Data Processing Subgroup: The Sequence Alignment/Map format and SAMtools.** *Bioinformatics* 2009, **25**:2078–2079.
 59. Liu A, Burke JM: **Patterns of nucleotide diversity in wild and cultivated sunflower.** *Genetics* 2006, **173**:321–330.
 60. Kolkman JM, Berry ST, Leon AJ, Slabaugh MB, Tang S, Gao W, Shintani DK, Burke JM, Knapp SJ: **Single nucleotide polymorphisms and linkage disequilibrium in sunflower.** *Genetics* 2007, **177**:457–468.

doi:10.1186/1471-2164-14-556

Cite this article as: Pegadaraju et al.: *De novo* sequencing of sunflower genome for SNP discovery using RAD (Restriction site Associated DNA) approach. *BMC Genomics* 2013 **14**:556.

Submit your next manuscript to BioMed Central and take full advantage of:

- Convenient online submission
- Thorough peer review
- No space constraints or color figure charges
- Immediate publication on acceptance
- Inclusion in PubMed, CAS, Scopus and Google Scholar
- Research which is freely available for redistribution

Submit your manuscript at
www.biomedcentral.com/submit

