SNP Discovery and High Density Infinium Chip Design for Sunflower Genotyping

Pegadaraju Venkatramana[1], Rick Nipper[2], Robert Bialozynski[1], Luke Score[1], Quentin Schultz[1]
& Benjamin Kauffman[1]
[1]BioDiagnostics Inc, 507 Highland Drive River Falls, WI-54022
[2]Floragenex Inc, 1900 Millrace Drive Eugene, OR 97403

Key challenges in sunflower Industry

Quality traits

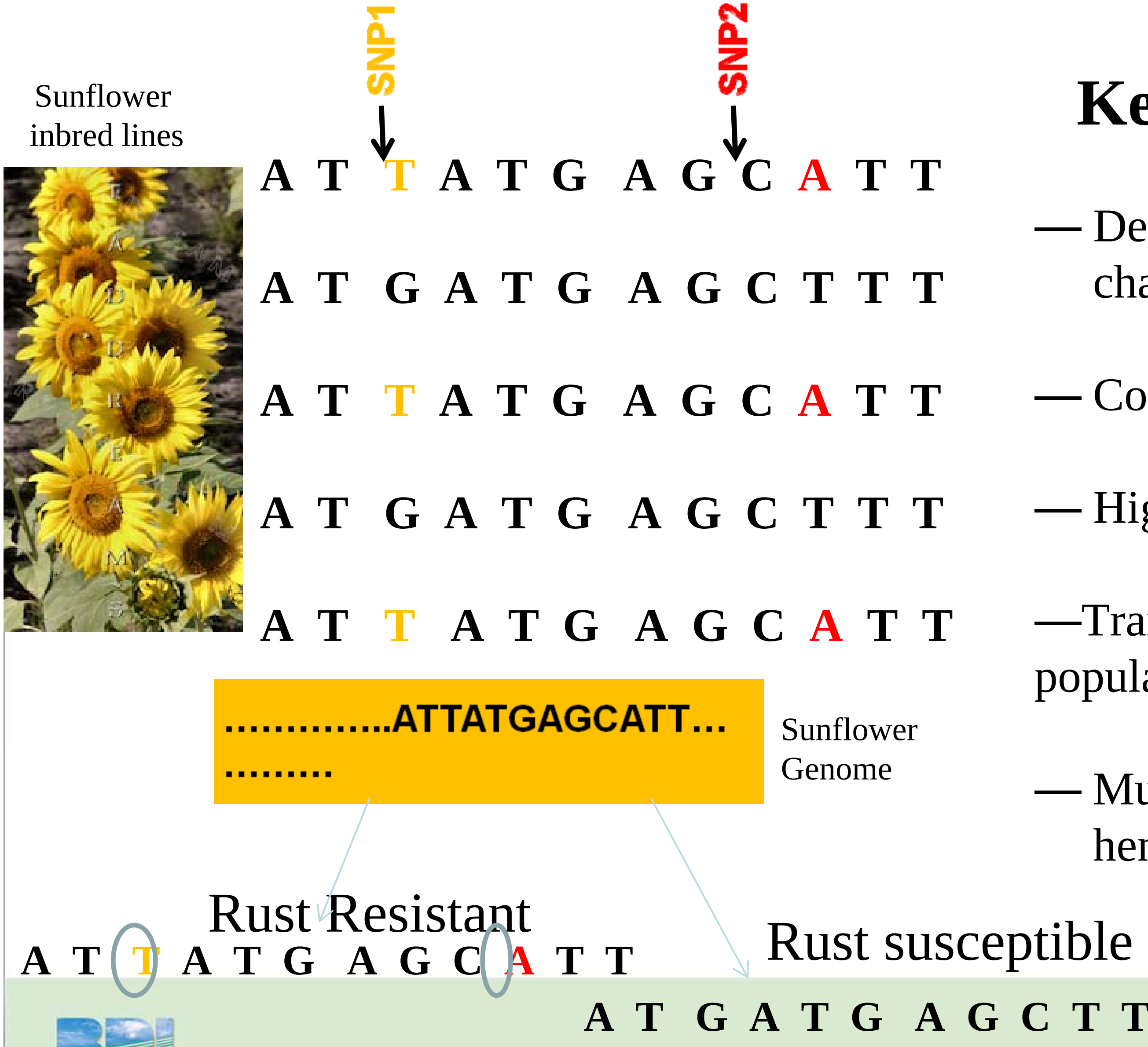
Disease
Downy Mildew
Rust
Sclerotinia
Verticillium
Alternaria
Stem Cancer
Rhizopus

Weeds
Biennials
Wormwood
Foxtail
Kochia
Redroot Pigweed
Wild sunflower
Wild Buckwheat

Insects
Sunflower Moth
Weevil
Midge
Longhorn Beetle
Wire Worms

High Oleic soybeans

SNP Markers as effective tools for breeding



Sunflower inbred lines

SNP1

SNP2

A T T A T G A G C A T T

A T G A T G A G C T T T

A T T A T G A G C A T T

A T G A T G A G C T T T

A T T A T G A G C A T T

..............ATTATGAGCATT... ....

Sunflower Genome

Rust Resistant

A T T A T G A G C A T T

Rust susceptible

A T G A T G A G C T T T

Key Features:

— Detects single nucleotides changes

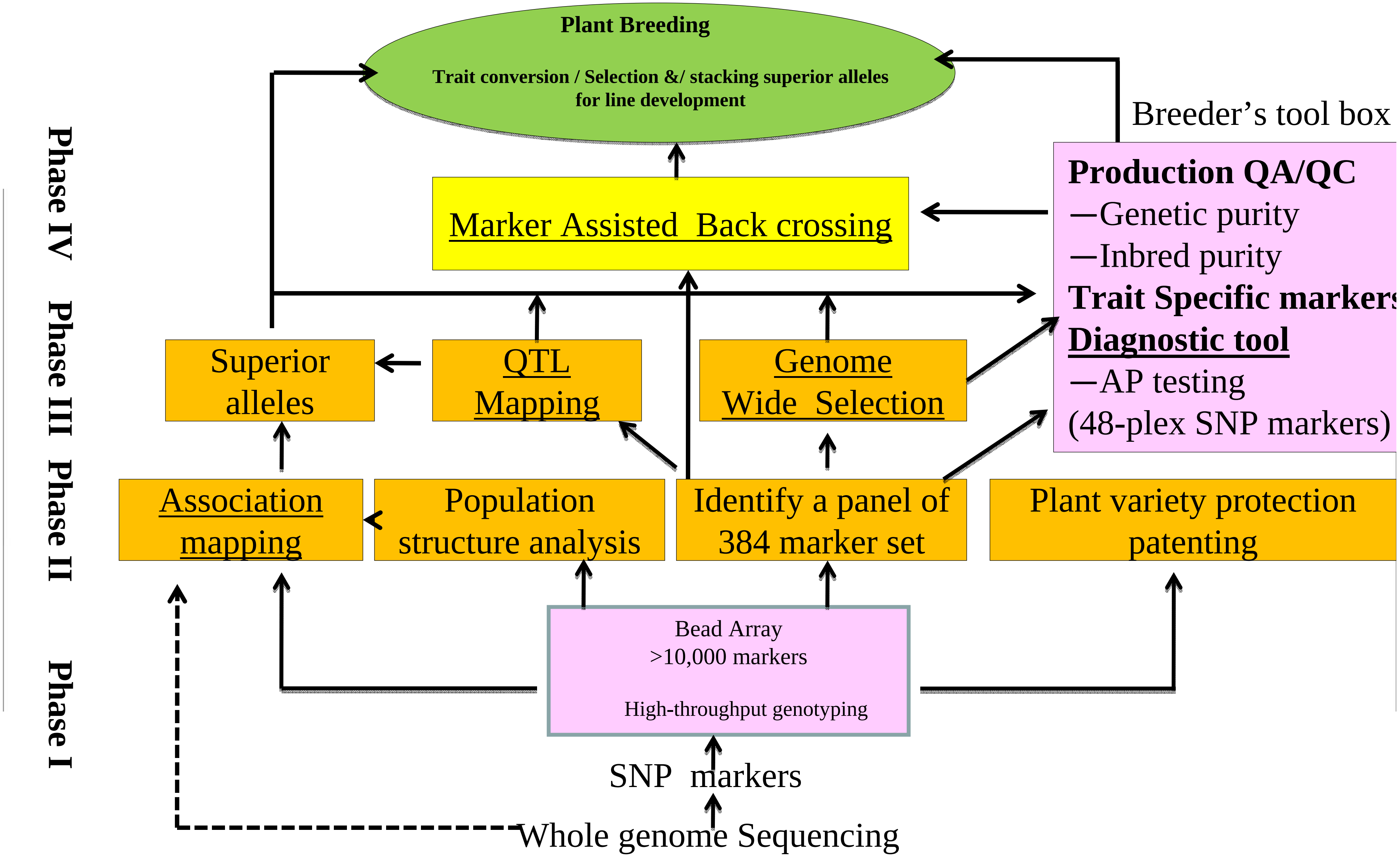
— Co-dominant in nature

— Highly abundant in number

—Transferable across populations

— Multiplexed effectively and hence, cost effective

Application of SNP markers in plant breeding



Plant Breeding

Trait conversion / Selection &/ stacking superior alleles for line development

Breeder's tool box

Production QA/QC

—Genetic purity

—Inbred purity

Trait Specific markers

Diagnostic tool

—AP testing

(48-plex SNP markers)

Marker Assisted Back crossing

Superior alleles

QTL Mapping

Genome Wide Selection

Association mapping

Population structure analysis

Identify a panel of 384 marker set

Plant variety protection patenting

Bead Array
>10,000 markers

High-throughput genotyping

SNP markers

Whole genome Sequencing

Phase IV

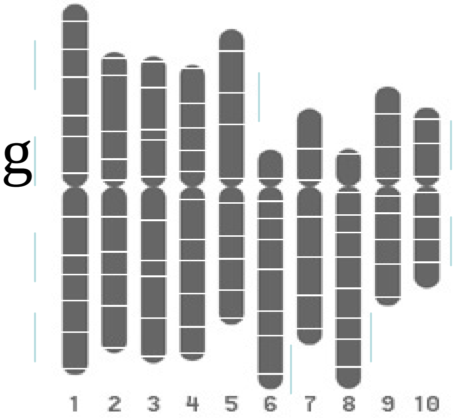
Phase III

Phase II

Phase I

Next Generation Sequencing platforms

Amplicon Sequencing



Whole Genome Sequencing

1
2
3
4
5
6
7
8
9
10

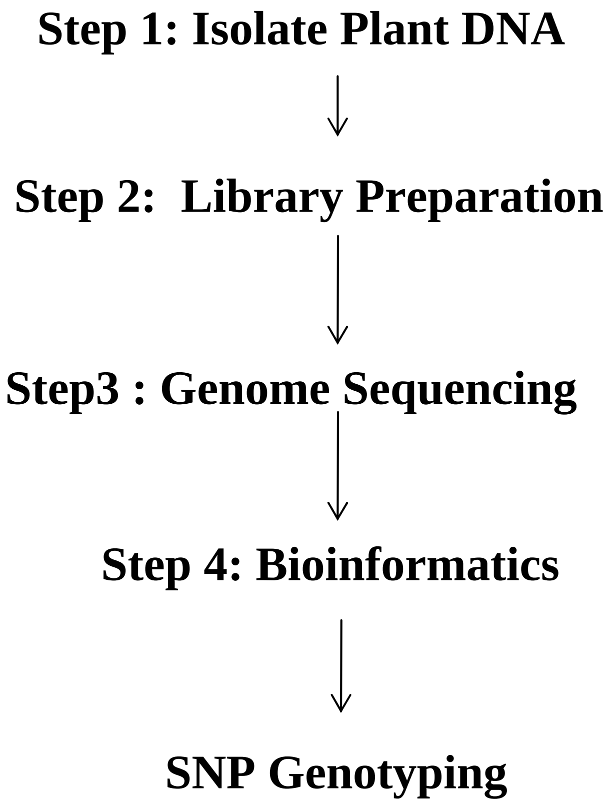
	Amplicon Sequencing	Whole Genome Sequencing		
Sequencing Machine	**ABI3730**	**Roche GSFLX**	**Solexa**	**SOLiD**
Read length bp	800	250	35-75	25-35
Reads per run	96	400k	130M	150M
Throughput per run	0.1MB	100MB	10GB	5GB
Cost per GB	>$2500K	$84K	$2K	$4K

Criteria for choosing a panel of lines for sequencing

Sunflower Line
TX1612
CR29
Seeds 2000 Confection B Line
HA467
RHA468 699-10
RHA464 09 098-4

- Lines identified for sequencing should not be redundant with earlier publically sequenced lines for SNP discovery
- Selected lines should be genetically diverse & must posses least amount of heterozygosity
- Both public and propriety lines should be included
- Representation of A-, B- R-lines and wild germplasm.

Key Steps in Whole Genome SNP Identification



Step 1: Isolate Plant DNA

↓

Step 2: Library Preparation

↓

Step3 : Genome Sequencing

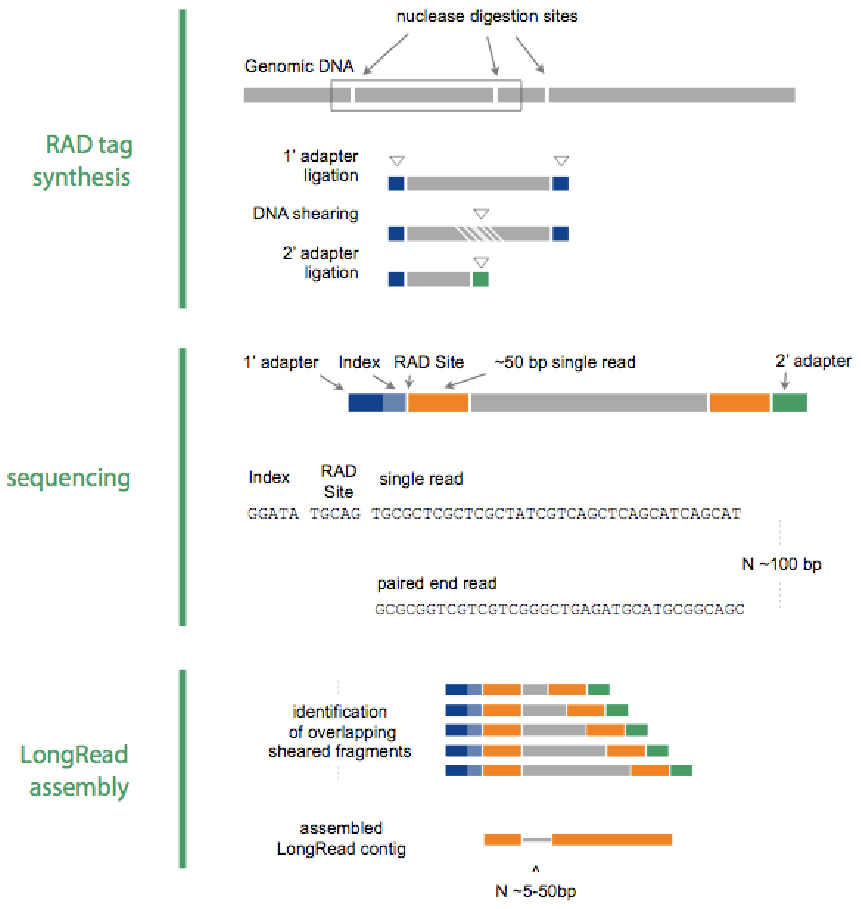
↓

Step 4: Bioinformatics

↓

SNP Genotyping

RAD LongRead – A local assembly approach



RAD tag synthesis

nuclease digestion sites

Genomic DNA

1' adapter ligation

DNA shearing

2' adapter ligation

sequencing

1' adapter

Index

RAD Site

~50 bp single read

2' adapter

Index

RAD Site

single read

```
GGATA TGCAG TGCGCTCGCTCGCTATCGTCAGCTCAGCATCAGCAT
```

N ~100 bp

paired end read

```
GCGCGGTCGTCGTCGGGCTGAGATGCATGCGGCAGC
```

LongRead assembly

identification of overlapping sheared fragments

assembled LongRead contig

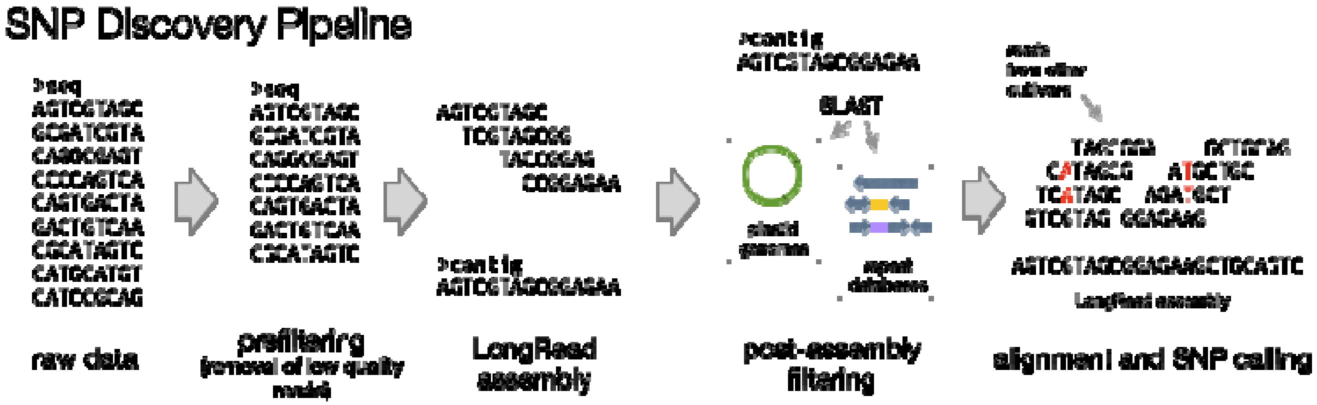
N ~5-50bp

Sunflower has a complex genome of 3.5Gb and remains to be sequenced

Discovery of SNPs variants in sunflower would require development & assembly of large island of DNA sequence to detect SNPs

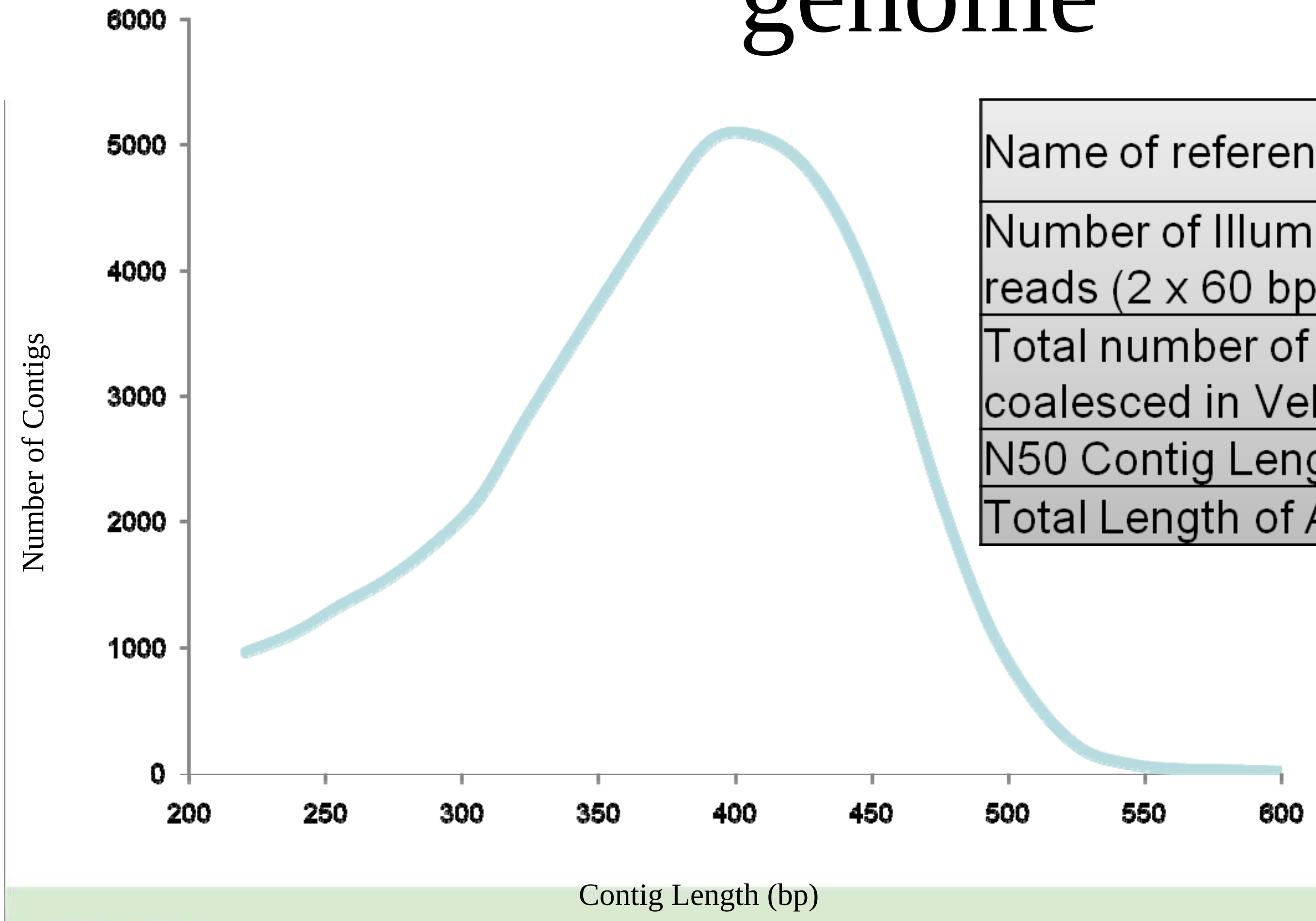
RAD LongRead technology coupled with bioinformatics analysis was adapted to *de novo* assemble of sunflower genome and identify SNP markers

Overview of SNP Detection Process



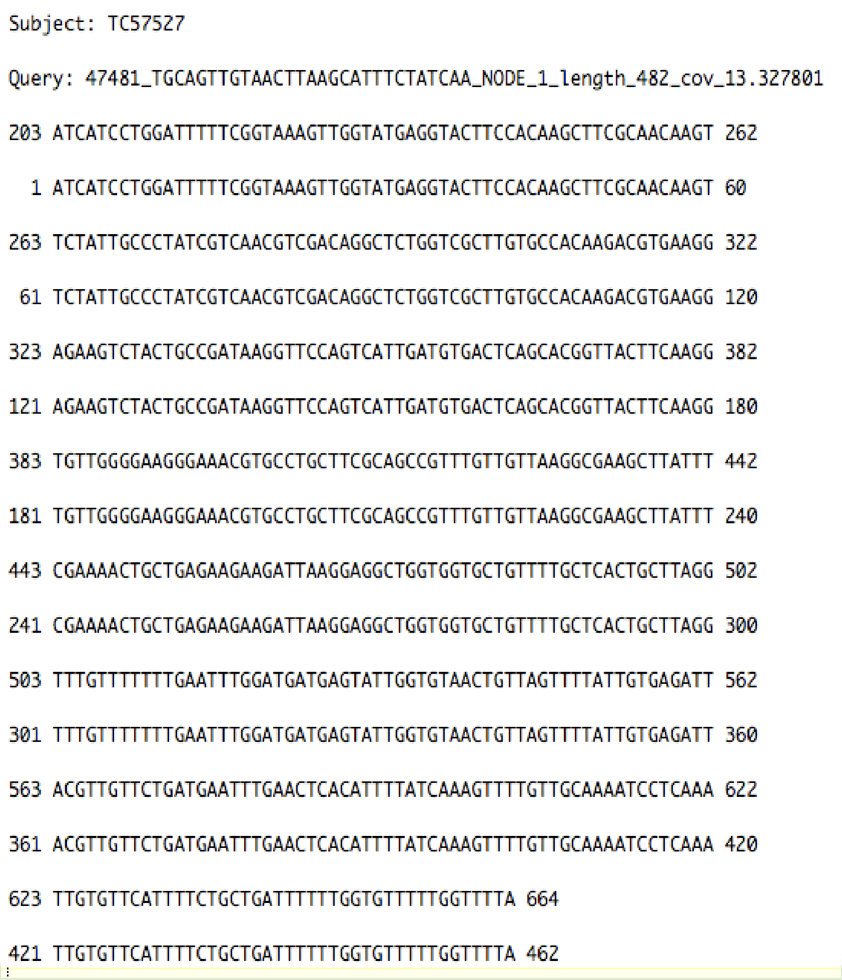
Genomic DNA from six selected sunflower isolines was digested with the endonuclease PstI and transformed into RAD libraries using methods similar to (Baird, et al . 2008 PLoS ONE 3(10)). Libraries were sequenced on an Illumina Genome Analyzer IIx at the University of Oregon High Throughput Sequencing Facility. Sequences from BDI_Sunflower_06 were coalesced in LongRead contigs using the program Velvet (Zerbino and Birney. 2008 Genome Research 18: 821:829). After alignment of assembled contigs against a custom database to remove sequences with significant plastid homology, 50,726 contigs covering 18 Mbp of the sunflower genome remained. These served as a reference scaffold for sequence alignment of Illumina data from the other cultivars. Sequence alignment and variant calling was accomplished though use of internal Floragenex tools.

Assembly of Raw Illumina read to generate contigs of the reference genome



Name of reference cultivar	BDI_Sunflower_06
Number of Illumina paired end reads (2 x 60 bp) to obtain ref	9,016,941
Total number of contigs coalesced in Velvet assembly	50,726
N50 Contig Length	379 bp
Total Length of Assembly	18.87 Mbp

Contig Assembly



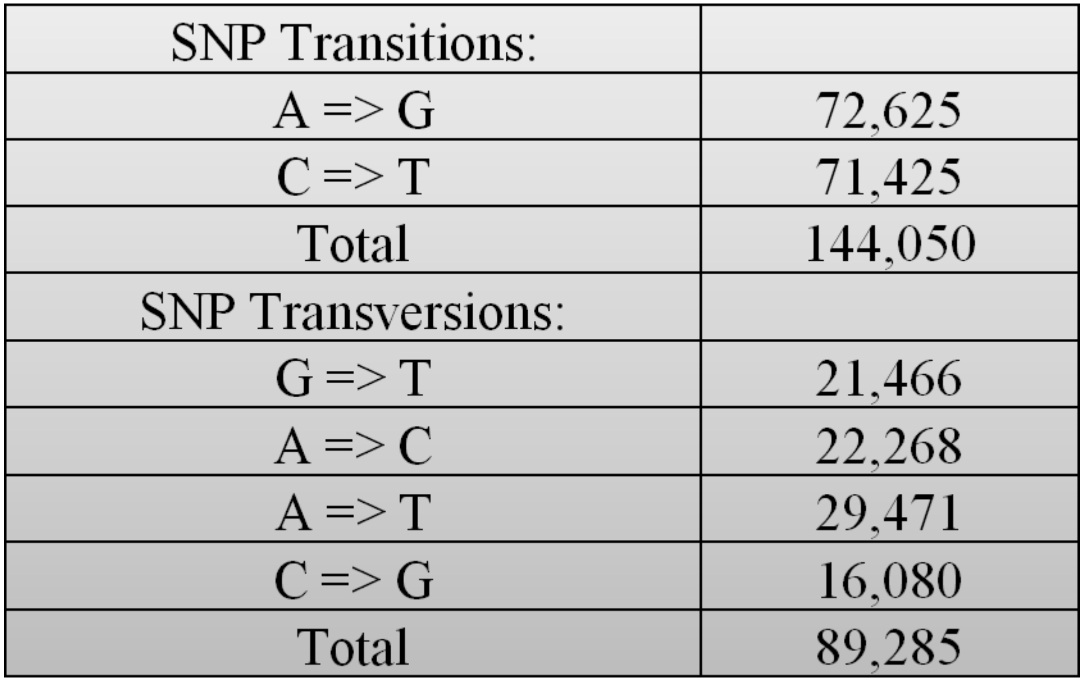
```
Subject: TC57527

Query: 47481_TGCAGTTGTAACTTAAGCATTTCTATCAA_NODE_1_length_482_cov_13.327801

203 ATCATCCTGGATTTTTCGGTAAAGTTGGTATGAGGTACTTCCACAAGCTTCGCAACAAGT 262

  1 ATCATCCTGGATTTTTCGGTAAAGTTGGTATGAGGTACTTCCACAAGCTTCGCAACAAGT 60

263 TCTATTGCCCTATCGTCAACGTCGACAGGCTCTGGTCGCTTGTGCCACAAGACGTGAAGG 322

 61 TCTATTGCCCTATCGTCAACGTCGACAGGCTCTGGTCGCTTGTGCCACAAGACGTGAAGG 120

323 AGAAGTCTACTGCCGATAAGGTTCCAGTCATTGATGTGACTCAGCACGGTTACTTCAAGG 382

121 AGAAGTCTACTGCCGATAAGGTTCCAGTCATTGATGTGACTCAGCACGGTTACTTCAAGG 180

383 TGTTGGGGAAGGGAAACGTGCCTGCTTCGCAGCCGTTTGTTGTTAAGGCGAAGCTTATTT 442

181 TGTTGGGGAAGGGAAACGTGCCTGCTTCGCAGCCGTTTGTTGTTAAGGCGAAGCTTATTT 240

443 CGAAAACTGCTGAGAAGAAGATTAAGGAGGCTGGTGGTGCTGTTTTGCTCACTGCTTAGG 502

241 CGAAAACTGCTGAGAAGAAGATTAAGGAGGCTGGTGGTGCTGTTTTGCTCACTGCTTAGG 300

503 TTTGTTTTTTTGAATTTGGATGATGAGTATTGGTGTAACTGTTAGTTTTATTGTGAGATT 562

301 TTTGTTTTTTTGAATTTGGATGATGAGTATTGGTGTAACTGTTAGTTTTATTGTGAGATT 360

563 ACGTTGTTCTGATGAATTTGAACTCACATTTTATCAAAGTTTTGTTGCAAAATCCTCAAA 622

361 ACGTTGTTCTGATGAATTTGAACTCACATTTTATCAAAGTTTTGTTGCAAAATCCTCAAA 420

623 TTGTGTTCATTTTCTGCTGATTTTTTGGTGTTTTTGGTTTTA 664

421 TTGTGTTCATTTTCTGCTGATTTTTTGGTGTTTTTGGTTTTA 462
```

Alignment of RAD LongRead contig from BDI_Sunflower_06 against the DFCI Sunflower EST sequence repository (HaGI_release_6). The contig shows 100% nucleotide identity with Tentative Consensus Sequence TC57527. The alignment spans the entire distance of the LongRead contig and suggests sunflower data was properly ordered and assembled by Velvet.

Variant Detection Summary Table

Number of contigs scanned for variants	50,726
Total sunflower genomic sequence in contigs	18.87 Mbp
Number of contigs with at least one polymorphism present	24,202
Average number of variants identified per contig	5
Total number of SNPs identified in six lines	233,335
Total number of InDels detected in six lines	5,280
Calculated SNP polymorphism rate	1 SNP / 81 bp
Calculated InDel rate	1 InDel/ 3,574 bp

SNP Transitions & Transversions



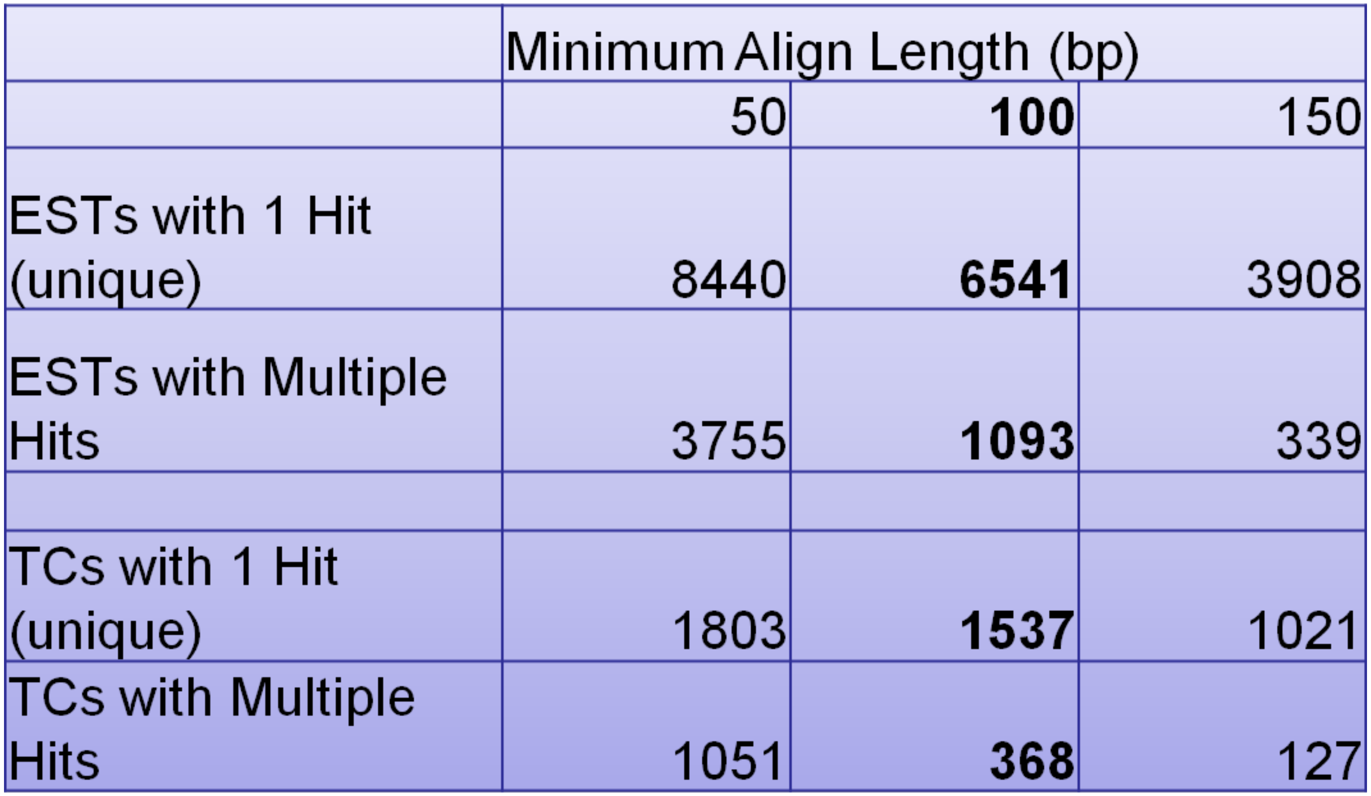
SNP Transitions:	
A => G	72,625
C => T	71,425
Total	144,050
SNP Transversions:	
G => T	21,466
A => C	22,268
A => T	29,471
C => G	16,080
Total	89,285

Number of SNP/InDels suitable for Infinium Genotyping Technology: 16,394 (~50 bp clear of flanking polymorphisms)

Key Consideration for Selecting SNP's for Infinium Design

- High ADT design scores >0.6
- Maximize the number of single bead assays
- Uniform distribution in sunflower genome
- SNP's should map to the EST sequences in the database
- Repetitive sequences & transposons elements should be eliminated
- SNP context sequences should not possess adjacent polymorphisms

Summary of Blast results

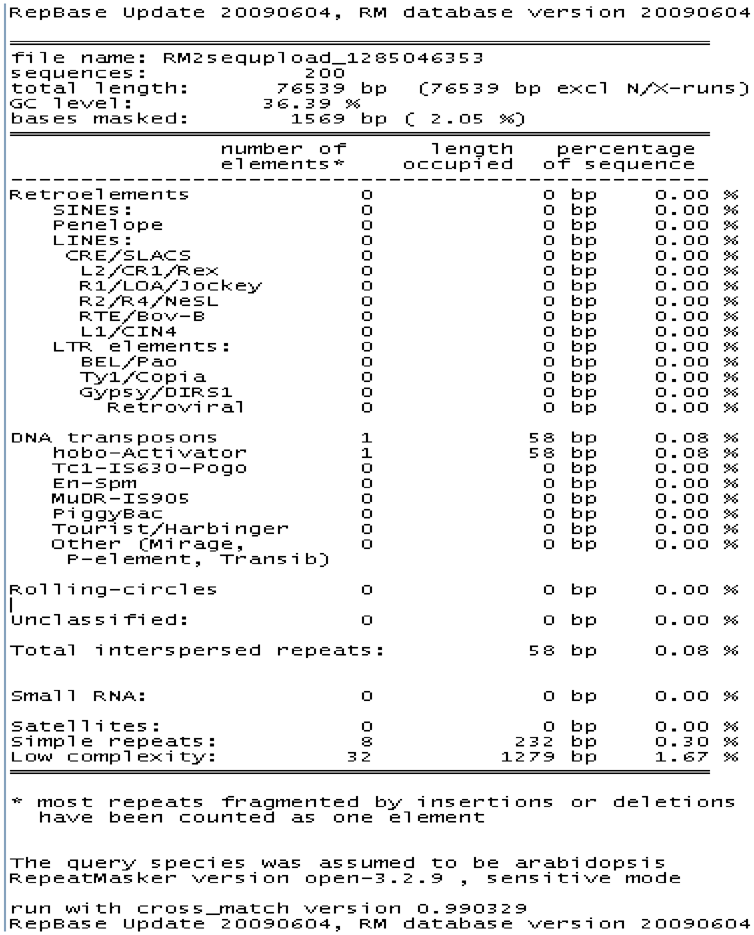


	Minimum Align Length (bp)		
	50	**100**	150
ESTs with 1 Hit (unique)	8440	**6541**	3908
ESTs with Multiple Hits	3755	**1093**	339
TCs with 1 Hit (unique)	1803	**1537**	1021
TCs with Multiple Hits	1051	**368**	127

NCBI

DFCI

Evaluating the SNP Sequences for the presence of Repetitive elements

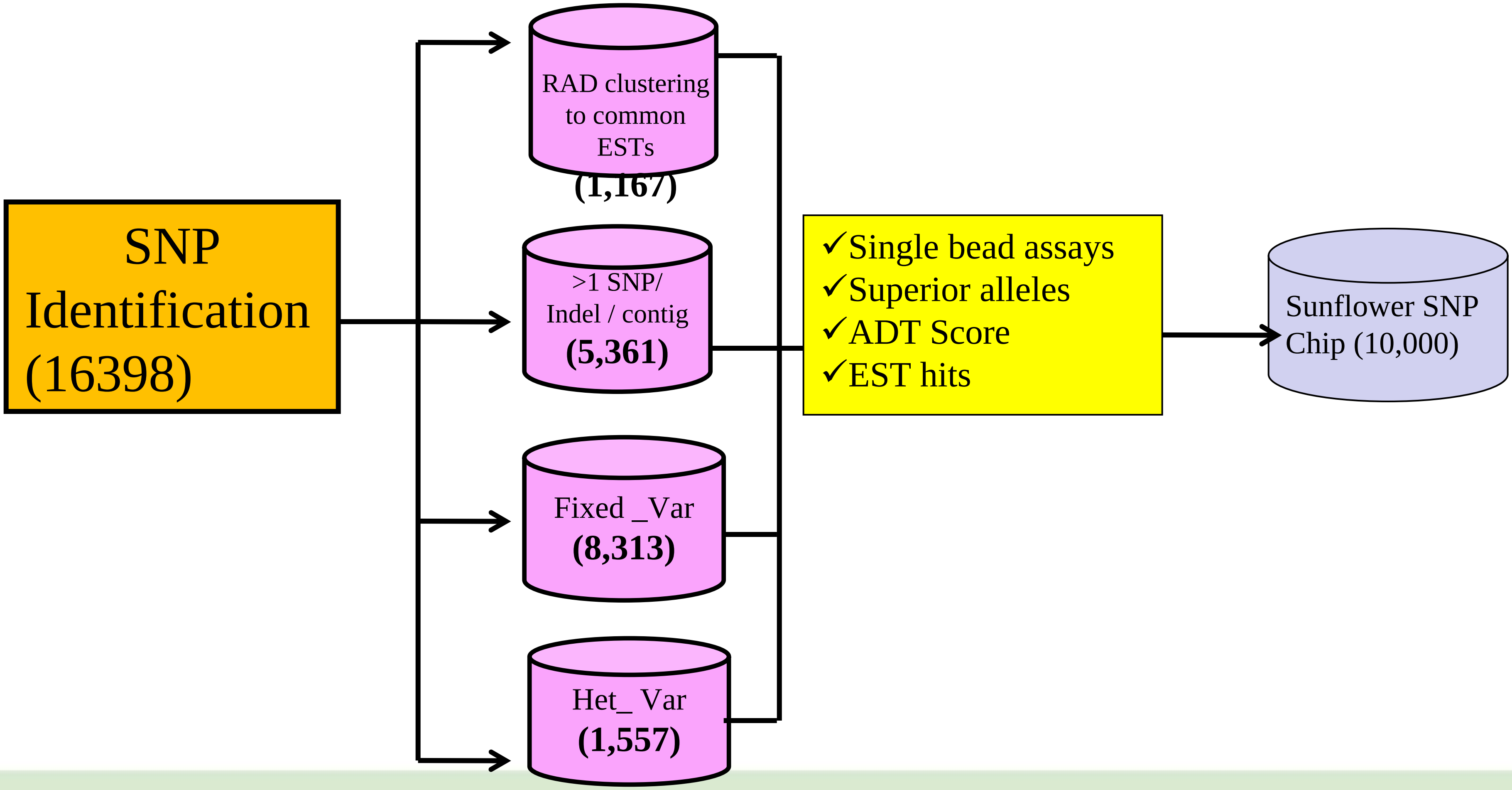


```
RepBase Update 20090604, RM database version 20090604

==================================================
file name: RM2sequpload_1285046353
sequences:           200
total length:      76539 bp  (76539 bp excl N/X-runs)
GC level:         36.39 %
bases masked:       1569 bp ( 2.05 %)
==================================================
               number of      length   percentage
               elements*    occupied  of sequence
--------------------------------------------------
Retroelements            0          0 bp    0.00 %
   SINEs:                0          0 bp    0.00 %
   Penelope              0          0 bp    0.00 %
   LINEs:                0          0 bp    0.00 %
    CRE/SLACS            0          0 bp    0.00 %
     L2/CR1/Rex          0          0 bp    0.00 %
     R1/LOA/Jockey       0          0 bp    0.00 %
     R2/R4/NeSL          0          0 bp    0.00 %
     RTE/Bov-B           0          0 bp    0.00 %
     L1/CIN4             0          0 bp    0.00 %
   LTR elements:         0          0 bp    0.00 %
     BEL/Pao             0          0 bp    0.00 %
     Ty1/Copia           0          0 bp    0.00 %
     Gypsy/DIRS1         0          0 bp    0.00 %
       Retroviral        0          0 bp    0.00 %

DNA transposons          1         58 bp    0.08 %
   hobo-Activator        1         58 bp    0.08 %
   Tc1-IS630-Pogo        0          0 bp    0.00 %
   En-Spm                0          0 bp    0.00 %
   MuDR-IS905            0          0 bp    0.00 %
   PiggyBac              0          0 bp    0.00 %
   Tourist/Harbinger     0          0 bp    0.00 %
   Other (Mirage,        0          0 bp    0.00 %
    P-element, Transib)

Rolling-circles          0          0 bp    0.00 %

Unclassified:            0          0 bp    0.00 %

Total interspersed repeats:        58 bp    0.08 %


Small RNA:               0          0 bp    0.00 %

Satellites:              0          0 bp    0.00 %
Simple repeats:          8        232 bp    0.30 %
Low complexity:         32       1279 bp    1.67 %
==================================================

* most repeats fragmented by insertions or deletions
  have been counted as one element


The query species was assumed to be arabidopsis
RepeatMasker version open-3.2.9 , sensitive mode

run with cross_match version 0.990329
RepBase Update 20090604, RM database version 20090604
```

Approximately, 2% of nucleotides were masked with in the RAD sunflower assemblies using the Arabidopsis repeat database

Similarly, Panicoid, Triticale & Rice dbs yielded same results

Categorization of SNP’s selected for the final chip synthesis



SNP Identification (16398)

RAD clustering to common ESTs (1,167)

>1 SNP/ Indel / contig (5,361)

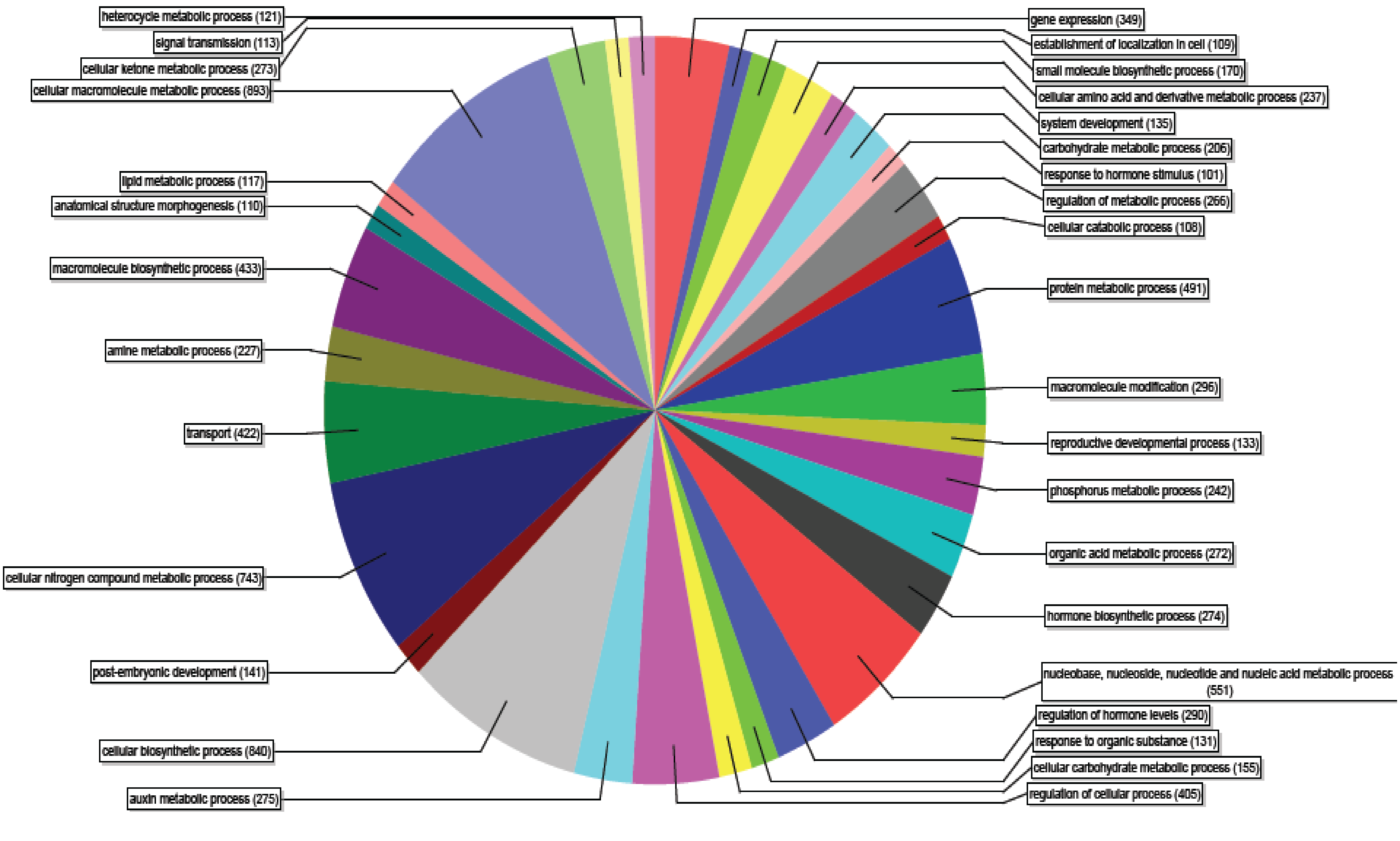
Fixed _Var (8,313)

Het_ Var (1,557)

- ✓Single bead assays
- ✓Superior alleles
- ✓ADT Score
- ✓EST hits

Sunflower SNP Chip (10,000)

Associating SNPs to various functional groups



Conclusion

—NGS & high throughput genotyping technologies can now provide

- ✓ Abundant
- ✓ Robust
- ✓ Cost effective molecular markers

—Marker application in sunflower breeding will ensure accurate and rapid trait selection enabling breeders to quickly release new sunflower hybrids into market

—DNA-based diagnostic methods can be used as quality assurance tools to produce a premium sunflower seed and growers can demand premium prices for their superior genetics

—Collaboration among all the relevant stakeholders is essential to meet this overall goal