PREDICTING SNPS AND HAPLOTYPES FROM PUBLIC EST DATA

Jifeng Tang & Jack Leunissen

Background

- Sequence polymorphism = single-nucleotide polymorphisms (SNPs) and small insertions/deletions (indels)
- SNP = substitutions a/o insertions/deletions

For example:

```
5′ - CGATCTGAATGCAGCTGACTGTCATGCACGATCACACTCGTACGCT - 3′   allele 1

5′ - CGATCTGAATGCAGCTGACTGTCTTGCACGA-CACACTCGTACGCT - 3′   allele 2
```

A ↔ T substitution(transversion)
T ↔ - insertion/deletion(indel)

Background

- EST = expressed sequence tags
- cSNP or EST-SNP = SNP in coding region
- Merits
  - directly study expressed genes and map functional traits
  - non-synonymous SNP (nsSNP) are more likely to change protein function
  - abundance of public EST data
  - linkage disequilibrium analysis to better characterize associations between phenotype and genotype or haplotype

Background

- Programs / pipelines for SNP detection
  - phred/phrap/polyphred/consed (Picoult-Newberg, 1999)
  - phred/phrap/polybayes (Deantec, 2004)
  - phred/cap3/Jalview system (Somers, 2003)
  - AutoSNP (Barker, 2003)
    - no paralog identification, only cluster sizes [4,50]
  - SNiPpER (Kota, 2003)
    - no paralog identification, only cluster sizes [4,20]

Objective of the work

- Focus on identifying false positive SNPs
  - Identify sequencing errors
  - Detect paralogs
- Design a haplotype-based strategy to detect reliable SNPs and identify clusters with potential paralogs from EST sequences without trace or quality files, and without completed genome information

Haplotype definition

- A set of closely linked genetic markers present on one chromosome which tend to be inherited together (not easily separable by recombination)
- Rafalski (2002) showed that several closely linked SNPs can completely define haplotypes
- Schneider (2001) showed that variation in the expressed genes of *Beta vulgaris* was essentially confined to haplotypes

Haplotype model

```
>contig_32 EST:16 SNP:15
location info:  132 189 326 358 389 566 567 575 669 754 761 922 947 953 972
CK242805|ken|callus|Stu.4700 G  A  A  A  A  C  A  T  C  G  C  C  C  C  -
CK242806|ken|callus|Stu.4700 G  A  A  A  A  C  A  T  C  G  C
CK245425|ken|callus|Stu.4700 A  T  G  G  G  T  G  A  T  T  T  C  T  G  -
CK252198|ken|callus|Stu.4700 A  T  G  G  G  T  G  A  T  T  T  C  T  G  -
CK243684|ken|callus|Stu.4700 .  .  A  A  A  C  A  T  C  G  C  C  C  C  -
CK243685|ken|callus|Stu.4700 G  A  A  A  A  C  A  T  C  G  C  C  C  C  -
CK247648|ken|callus|Stu.4700 A  T  G  G  G  C  G  A  T  T  T  C  T  G  C
CK248794|ken|callus|Stu.4700 .  .  .  .  .  .  .  .  .  .  .  T
CK248221|ken|callus|Stu.4700 A  T  G  G  G  C  G  A  T  T  T  C  T  G  C
CK245638|ken|callus|Stu.4700 G  A  A  A  A  C  A  T  C  G  C  C  C  C  -
CK246194|ken|callus|Stu.4700 G  A  A  A  A  C  A  T  C  G  C  C  C  C  -
CK248793|ken|callus|Stu.4700 G  A  A  A  A  C  A  T  C  G  C  C  C  C
CK249476|ken|callus|Stu.4700 G  A  A  A  A  C  A  T  C  G  C  C  C  C
CK245639|ken|callus|Stu.4700 .  .  .  .  .  C  A  T  C  G  C  T  C  C  -
CK253729|ken|callus|Stu.4700 A  T  G  G  G  T  G  A  T  T  T
CK256382|ken|callus|Stu.4700 A  T  G  G  G  C  G  A  T  T  T
```

Haplotype model

```
>contig_32 EST:16 SNP:15
location info:  132 189 326 358 389 566 567 575 669 754 761 922 947 953 972
CK242805|ken|callus|Stu.4700 G   A   A   A   A   C   A   T   C   G   C   C   C   C   -
CK242806|ken|callus|Stu.4700 G   A   A   A   A   C   A   T   C   G   C
CK243684|ken|callus|Stu.4700 .   .   A   A   A   C   A   T   C   G   C   C   C   C   -
CK243685|ken|callus|Stu.4700 G   A   A   A   A   C   A   T   C   G   C   C   C   C   -
CK245638|ken|callus|Stu.4700 G   A   A   A   A   C   A   T   C   G   C   C   C   C   -
CK246194|ken|callus|Stu.4700 G   A   A   A   A   C   A   T   C   G   C   C   C   C   -
CK248793|ken|callus|Stu.4700 G   A   A   A   A   C   A   T   C   G   C   C   C   C
CK249476|ken|callus|Stu.4700 G   A   A   A   A   C   A   T   C   G   C   C   C   C
CK245639|ken|callus|Stu.4700 .   .   .   .   .   C   A   T   C   G   C   T   C   C   -

CK245425|ken|callus|Stu.4700 A   T   G   G   G   T   G   A   T   T   T   C   T   G   -
CK253729|ken|callus|Stu.4700 A   T   G   G   G   T   G   A   T   T   T
CK252198|ken|callus|Stu.4700 A   T   G   G   G   T   G   A   T   T   T   C   T   G   -

CK247648|ken|callus|Stu.4700 A   T   G   G   G   C   G   A   T   T   T   C   T   G   C
CK248221|ken|callus|Stu.4700 A   T   G   G   G   C   G   A   T   T   T   C   T   G   C
CK256382|ken|callus|Stu.4700 A   T   G   G   G   C   G   A   T   T   T

CK248794|ken|callus|Stu.4700 .   .   .   .   .   .   .   .   .   .   .   T
```

Haplotype No.1

No.2

No.3

Haplotype definition algorithm

- A haplotype is defined as a group of sequences within a cluster that have the same nucleotide at every polymorphic site
  - 1. defining the similarity of allelic variation on one polymorphic site between any EST and all current members of the haplotype
  - 2. defining the similarity of sequence and the haplotype depending on all its polymorphic sites

$$S_{ij} = \frac{\sum_{k=1}^{m} s_{ij}(k)}{\sum_{k=1}^{m} s_{ij}(k) + \sum_{k=1}^{m} d_{ij}(k)}$$

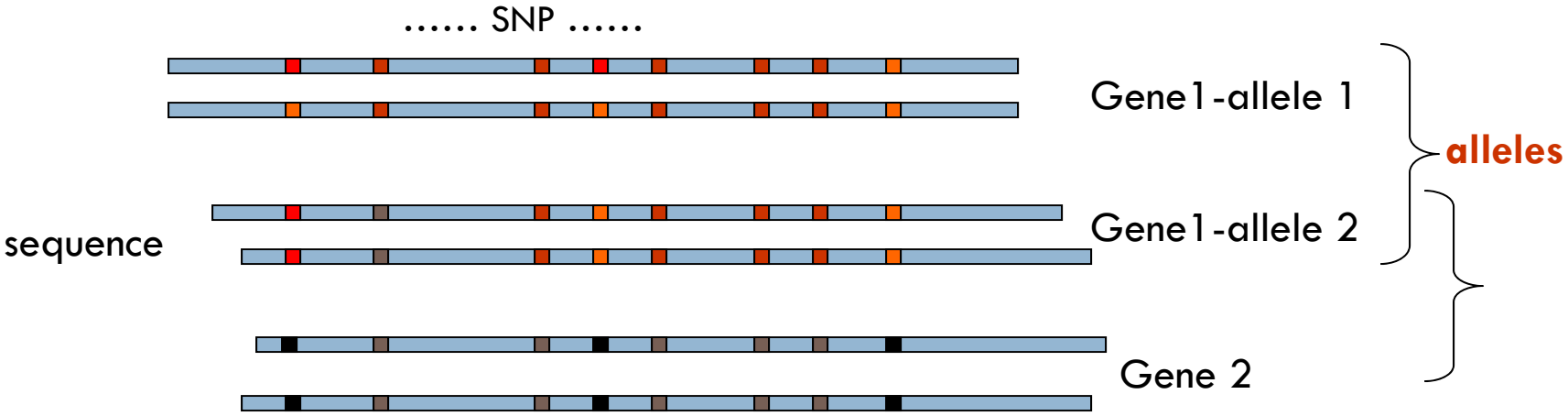
$$S_i = \frac{\sum_{j=1}^{n} S_{ij}}{\sum_{j=1}^{n} S_{ij} + \sum_{j=1}^{n} D_{ij}}$$

Paralogs definition

- Orthologs and paralogs are two types of homologous sequences
  - Orthology describes genes in different species that derive from a common ancestor
  - Paralogy describes homologous genes within a single species that diverged by gene duplication, where paralogs (may) evolve new functions, often related to the original one
- Paralogs are expected to contain more polymorphisms than allelic genes

Paralogs model

- Paralogs can be expected to contain more polymorphisms; this can be used to differentiate paralogs and alleles
  - Suppose gene2 is paralogous to gene1, but their sequences are quite similar, the model follows:



Paralogs identification algorithm

- Based on haplotypes, paralogs can be identified by calculating the standard deviation of variations among haplotypes in a cluster
  - Calculate the number of potential SNP defined in every haplotype:

$$snp_i \quad i \in [1, ahap]$$

ahap: the number of valid haplotypes

  - Normalize the number of SNPs per haplotype:

$$nrm\_snp_i = \frac{snp_i}{\frac{\sum_{i=1}^{ahap} snp_i}{ahap}} \quad \{i \mid i \in [1, ahap]\}$$

  - Calculate the standard deviation of the normalized number:

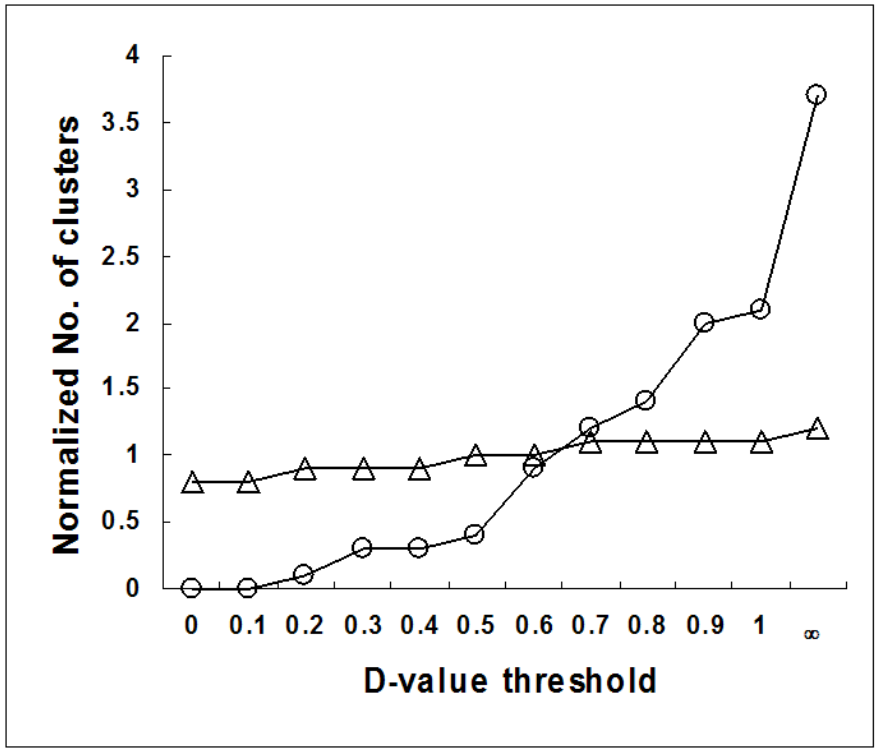
$$D = \sqrt{\frac{\sum_{i=1}^{ahap} (nrm\_snp_i - 1)^2}{ahap}}$$

- For larger D-values there is a higher probability that paralogs are contained in the cluster. But how to get the threshold of the D-value?

Identifying paralogs – threshold of D

- Assumptions: all clusters with 4-20 members are without paralogous sequences; all clusters with at least 100 members will contain paralogous sequences

- The figure shows the relationship of the normalized number of the dataset containing allelic sequences (△) and the dataset containing paralogs (○) with the D-value threshold using the potato dataset



Identify reliable SNPs - 1

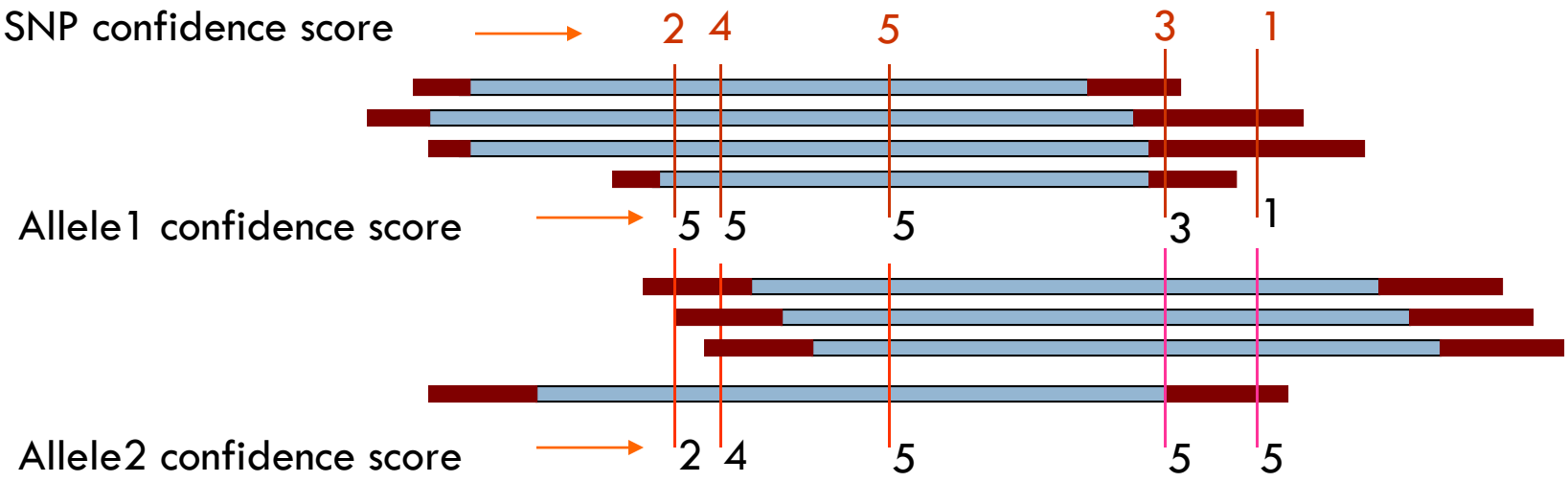
- A combination of two measures: major, minor allele haplotype score and confidence score based on sequence redundancy
  - Major allele haplotype score (*mahap*)

$$mahap = \sum_{i=1}^{ahap} mahap_i \left\{ mahap_i = 1 \,|\, \frac{wh \times ha_i + wl \times la_i}{hc_i} \geq Sij \right\}$$

  - Minor allele haplotype score (*mihap*)

$$mihap = \sum_{i=1}^{ahap} mihap_i \left\{ mihap_i = 1 \,|\, \frac{wh \times hb_i + wl \times lb_i}{hc_i} \geq Sij \right\}$$

Identify reliable SNPs - 2

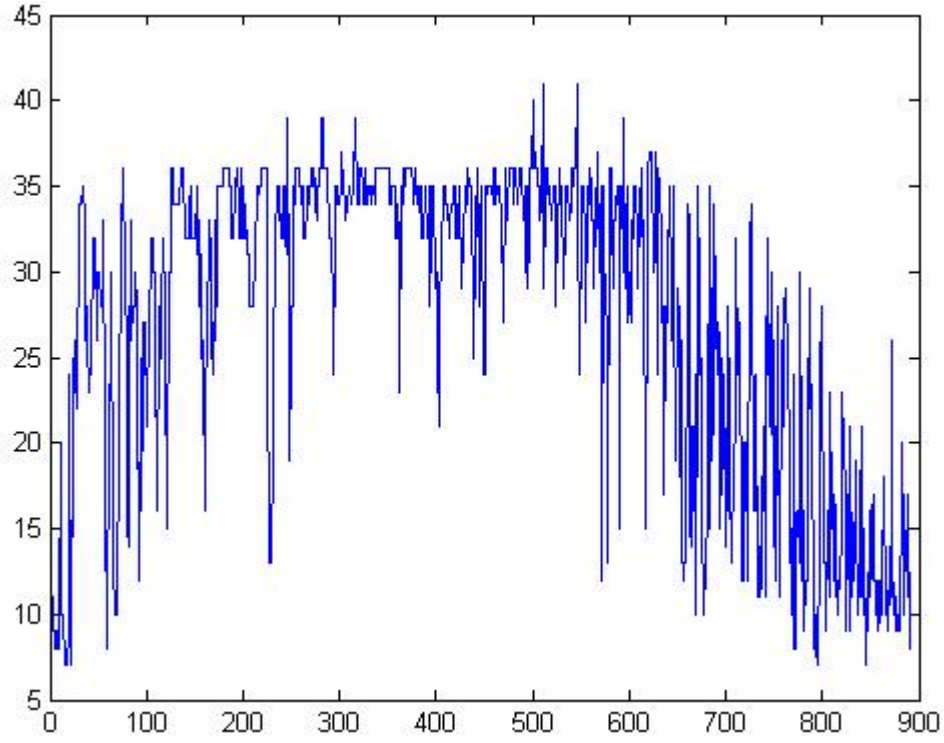


Confidence score is calculated for every putative SNP according to the number of occurrences of each allele in high and low quality regions

>BG592318|Kennebec|sprouting eyes from tubers

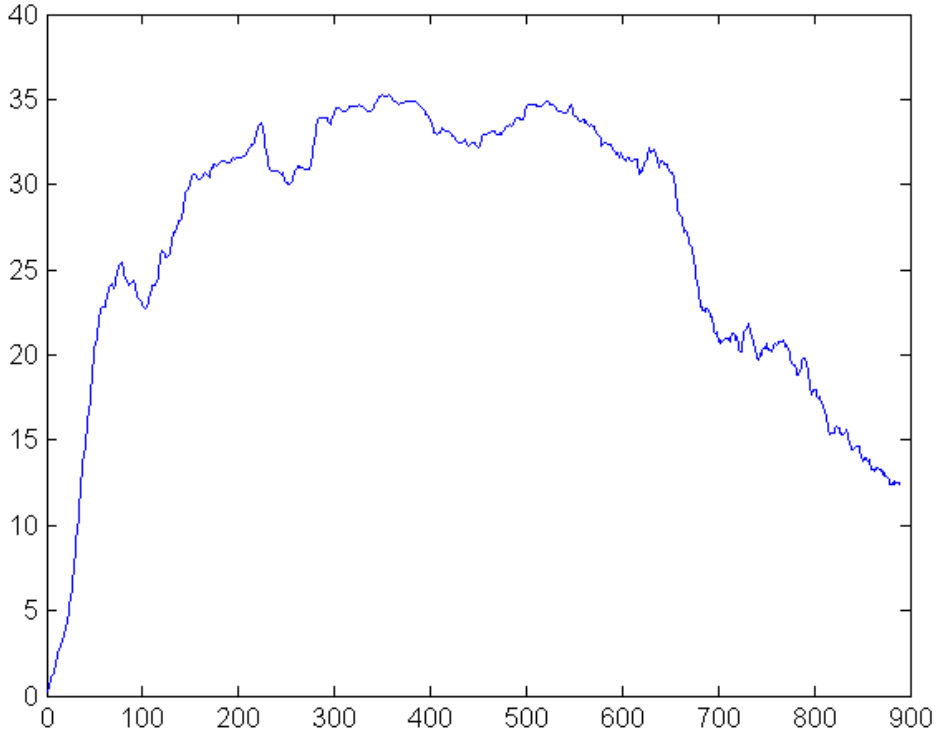
12 10 11 9 9 8 8 9 8 9 20 10 10 9 8 7 7 7 9 24 7 8 21 24 26 23 27 27 22 34 34 34 34 35 35 33 26
28 25 24 23 25 25 32 32 29 32 26 30 30 30 28 28 28 33 21 16 16 9 8 22 22 25 30 15 13 10 10
10 10 21 21 32 34 34 36 30 28 27 15 14 28 27 33 26 28 28 28 30 28 25 12 13 25 16 23 27 27 27
21 23 26 26 32 32 32 30 30 26 17 16 28 26 28 25 28 32 30 30 26 15 17 30 26 34 36 36 34 34 34
34 34 34 34 36 36 36 36 32 32 32 32 32 32 34 35 32 32 32 32 32 35 31 33 28 31 25 26 25 16 23
26 28 31 33 31 25 27 27 24 28 33 28 35 35 35 35 35 35 36 36 36 36 36 36 35 35 32 32 32 35 35
36 36 34 32 32 36 32 32 35 32 34 31 31 31 28 28 28 28 28 31 31 35 34 35 35 35 36 36 36 36 36
36 25 23 16 13 13 13 20 24 32 32 35 35 35 34 32 32 35 32 35 32 31 39 28 19 25 28 28 35 34 34
36 36 36 36 36 34 35 35 32 32 32 34 35 35 34 36 36 35 36 35 34 33 35 35 36 36 39 36 39 36 36
36 36 36 32 32 32 32 28 24 32 35 35 34 34 34 35 37 35 35 34 33 34 34 35 35 35 34 34 35 37 37
39 34 34 34 36 36 36 35 34 36 34 34 35 35 35 34 35 35 35 34 36 36 36 36 36 36 36 36 36 36 36
36 36 36 36 34 34 35 35 35 35 32 32 34 32 23 35 35 34 36 35 35 36 36 36 36 36 36 36 35 34 35
36 34 35 35 35 32 35 32 32 35 35 32 32 35 29 28 32 35 34 35 29 29 30 25 21 28 31 32 34 35 35
34 34 34 32 33 34 34 34 35 35 32 34 33 32 35 32 32 29 32 32 35 34 34 36 35 35 35 34 28 25 32
33 36 36 28 34 32 35 24 28 24 35 34 35 35 35 35 35 35 34 36 36 34 35 35 35 35 33 32 29 27 36
36 33 34 36 36 36 36 36 36 34 36 36 34 34 36 36 36 36 33 34 35 30 35 29 32 36 36 36 40 36 37
36 36 36 34 35 33 34 34 41 29 34 36 36 35 34 35 33 33 35 35 35 28 33 34 34 36 33 35 29 30 32
35 35 35 35 33 33 34 36 35 36 36 36 41 35 24 24 34 35 35 35 32 27 34 35 34 36 35 33 33 32 29
34 33 37 35 30 33 35 12 35 32 28 29 26 13 36 36 31 36 29 33 33 34 35 34 35 15 33 33 35 30 39
29 33 35 27 28 30 27 33 32 35 34 35 32 29 34 34 36 35 29 33 15 21 26 33 36 37 37 36 37 30 32
33 37 24 36 35 34 33 27 28 17 28 27 27 32 33 35 29 26 35 34 30 19 23 26 29 27 18 26 13 13 12
14 19 23 34 33 15 14 21 21 16 24 10 26 35 29 24 25 14 16 10 10 13 13 16 19 35 15 29 19 22 34
28 27 24 27 26 15 25 17 20 24 14 14 28 16 25 24 18 13 14 18 19 21 32 24 26 27 23 18 12 12 20
18 20 12 21 24 32 34 29 19 16 16 24 24 11 16 12 11 12 18 18 21 11 23 32 27 21 24 30 27 14 26
16 12 28 17 18 11 26 25 23 21 28 29 28 26 26 18 16 10 15 8 24 8 14 16 16 13 30 18 12 16 9 12
12 12 25 22 29 26 21 20 11 8 10 8 7 10 11 24 28 24 15 13 13 9 15 16 11 23 16 18 12 17 16 11
12 10 10 13 13 14 23 20 20 17 9 15 17 9 21 11 12 15 12 19 16 10 10 12 16 21 12 10 11 15 7 9 9
9 16 11 13 16 17 12 12 10 12 9 10 12 10 12 18 18 10 12 11 9 12 14 11 26 14 10 11 9 11 9 9 9
12 15 20 10 17 13 14 11 17 8

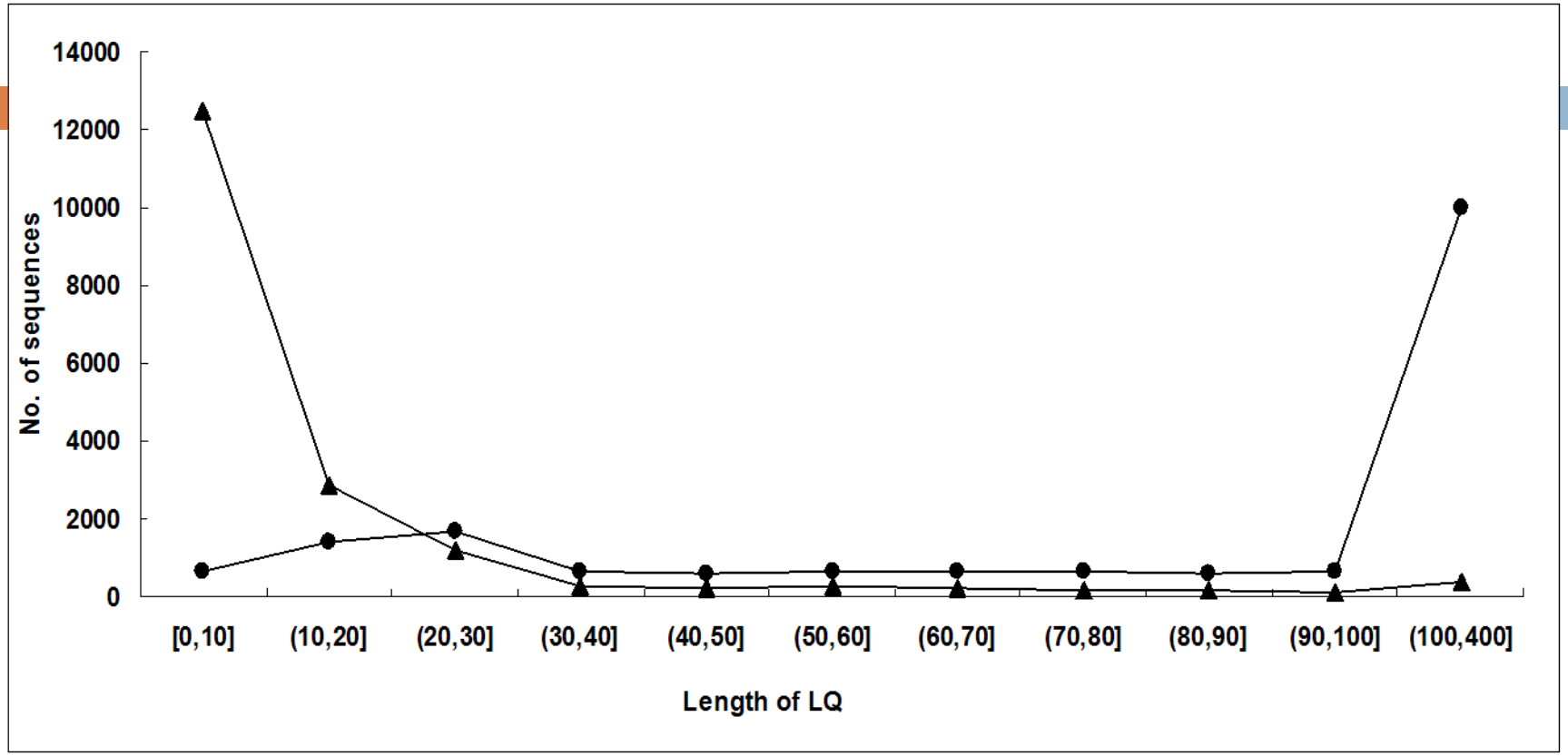
Distribution of quality scores



Raw values

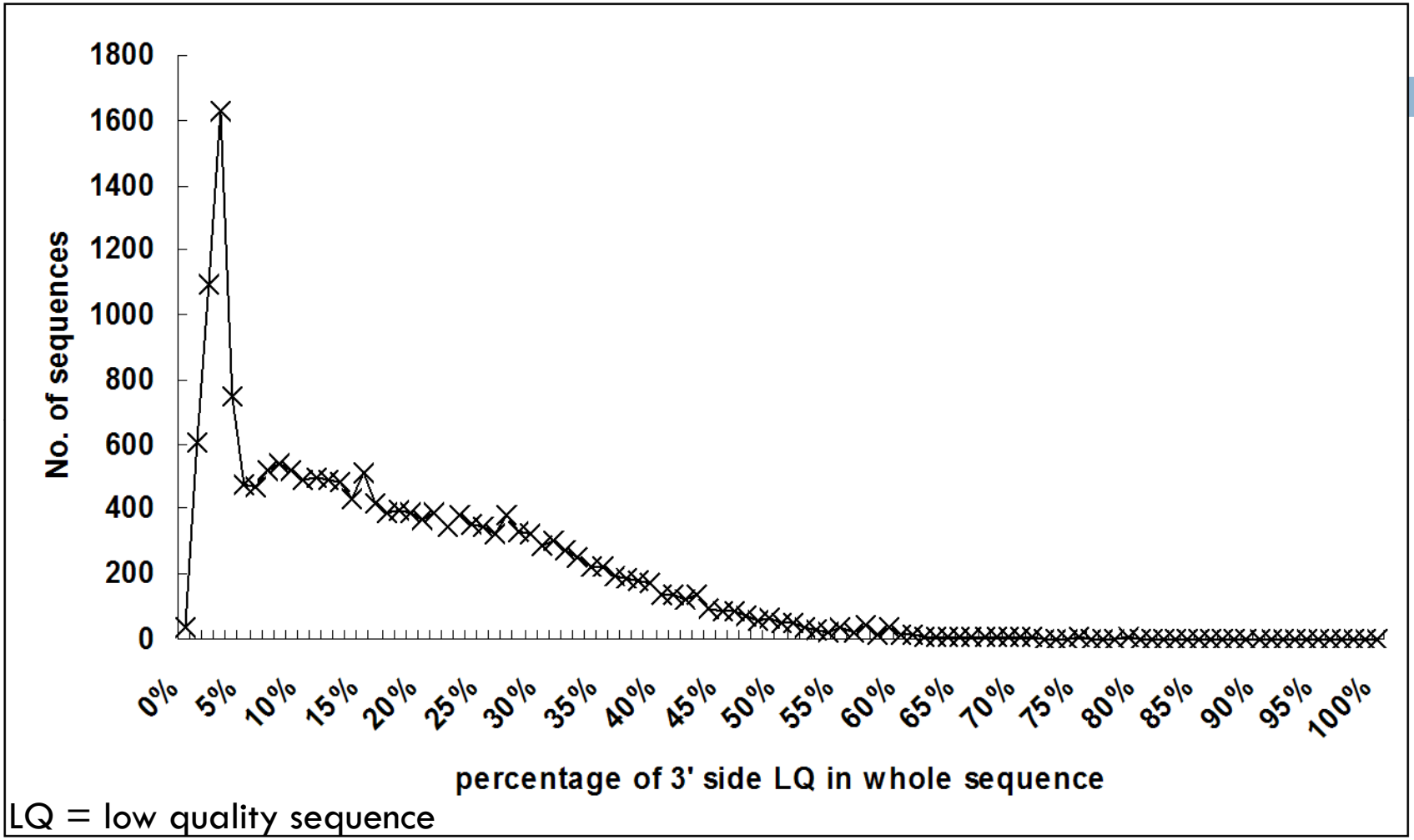
Smoothed values





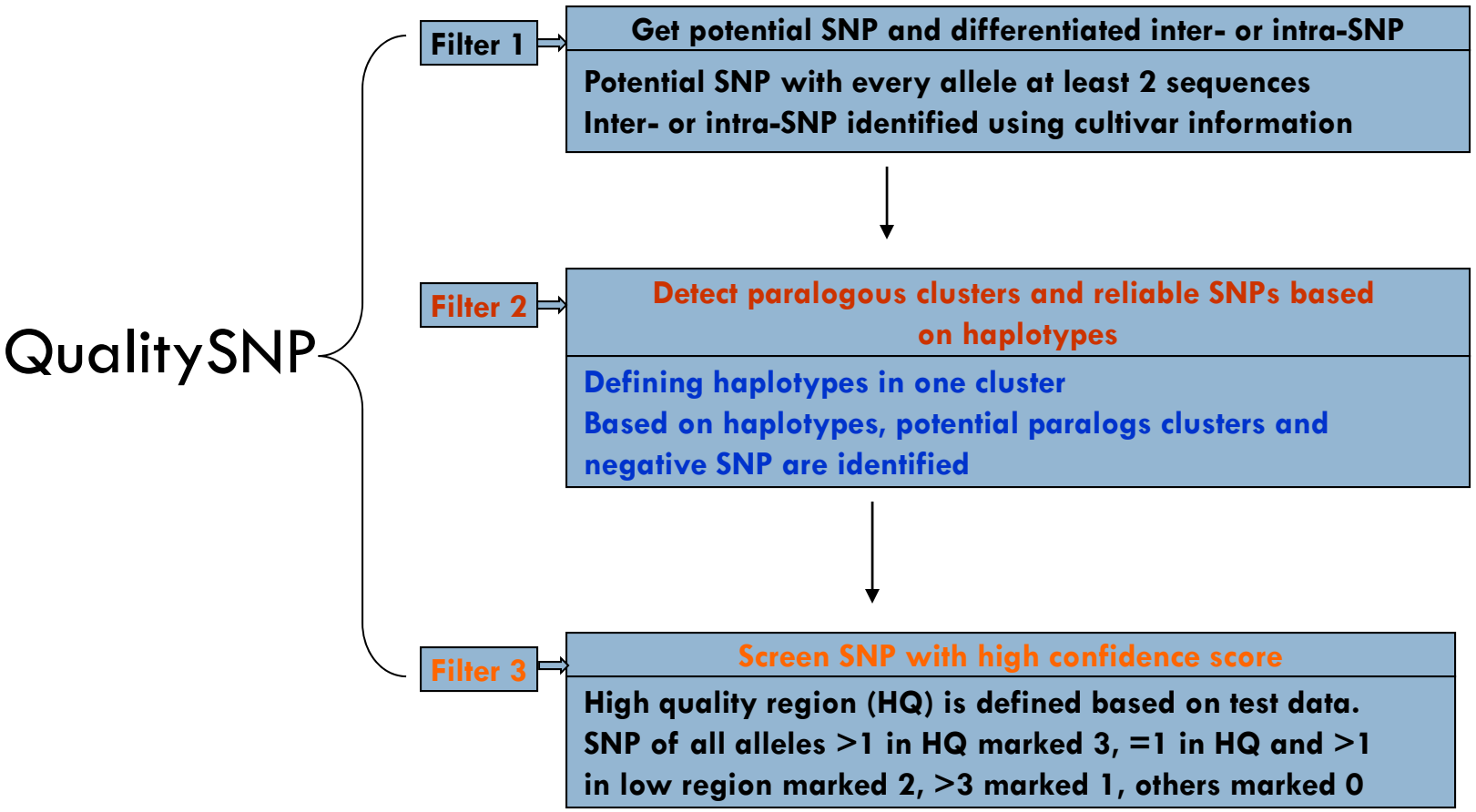
LQ = low quality sequence

The figure shows the number of sequences that have low quality scores in residue position intervals. It show that most sequences have LQ in the first 25 residues.

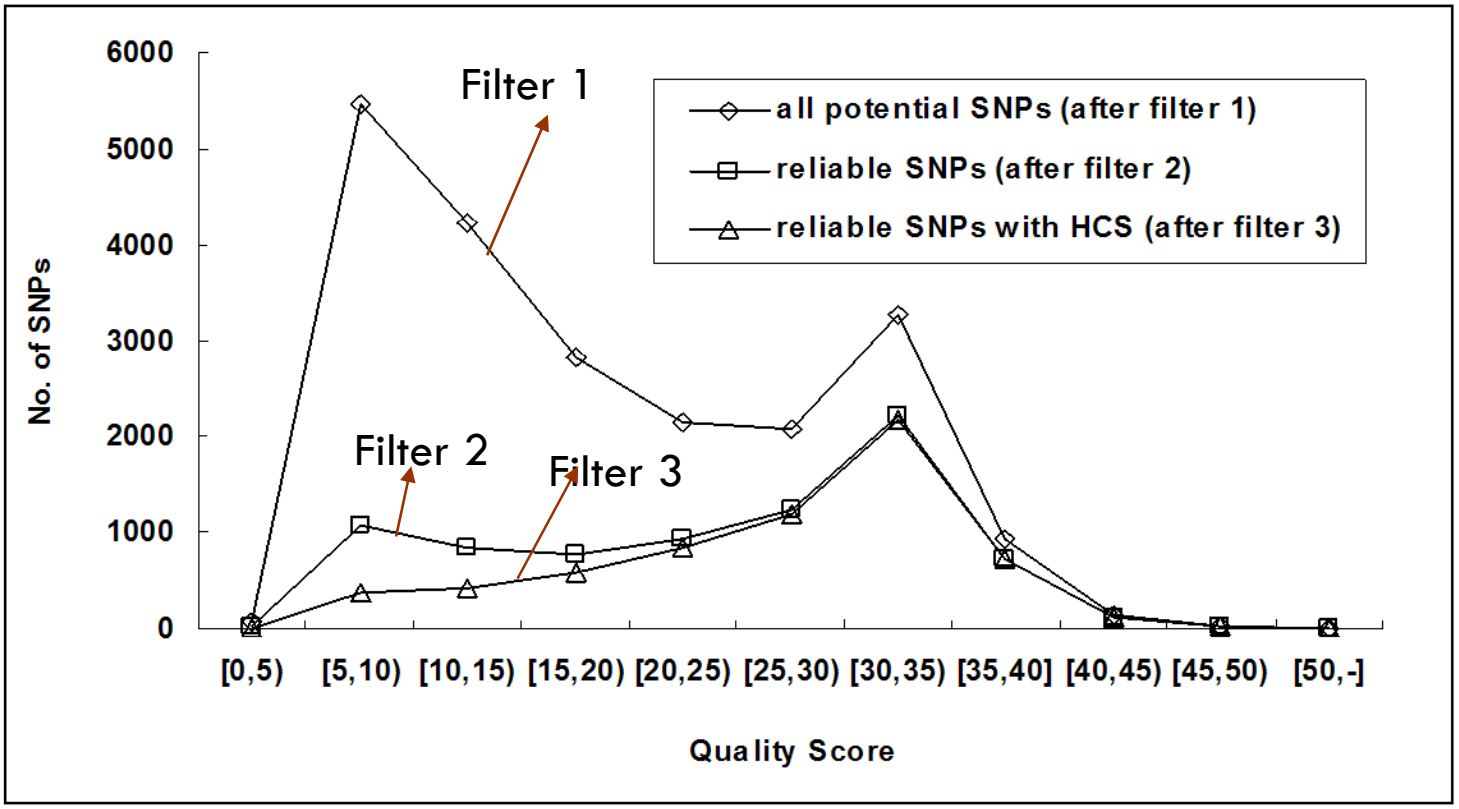


The figure shows the number of sequences that have low quality scores in the 3’ end of the sequence, as a percentage of the total length of the sequence.

Detect SNPs and haplotypes



Evaluation of QualitySNP



Validation of reliable SNPs with experimental data

From all predicted positive SNPs, 50 were selected randomly. 47 of these SNPs were verified experimentally as being true polymorphisms!

Evaluation of QualitySNP

QualitySNP compared to autoSNP (Batley *et al.* 2003)

Key to sequences:

```
A B73 gi|12967875|gb|BG264822.1|BG264822
B B73 gi|5499426|gb|AI855293.1|AI855293
C W23 gi|6828142|gb|AW331785.1|AW331785
D ohio_43 gi|4874469|gb|AI673989.1|AI673989
E ohio_43 gi|4887366|gb|AI677465.1|AI677465
F ohio_43 gi|5055938|gb|AI734825.1|AI734825
G gi|12970551|gb|BG267040.1|BG267040
H gi|14202672|gb|BG836349.1|BG836349
```

All SNPs correct

Summary of SNPs:

base	ABCDEFGH	SNP redundancy score	Cosegregation	Weighted cosegregation (%)
315	GGGCCCC.	3	7/9	68
380	AAAGGGG.	3	7/9	68
401	AAAGGGGA	4	7/9	78
455	AAACCCCA	4	7/9	78
514	CCCAAAAC	4	7/9	78
524	CCCTTTTC	4	7/9	78
541	TTCCCCCT	3	2/9	22
543	AACCCCCA	3	2/9	22
667	..AGGGGA	2	7/9	58

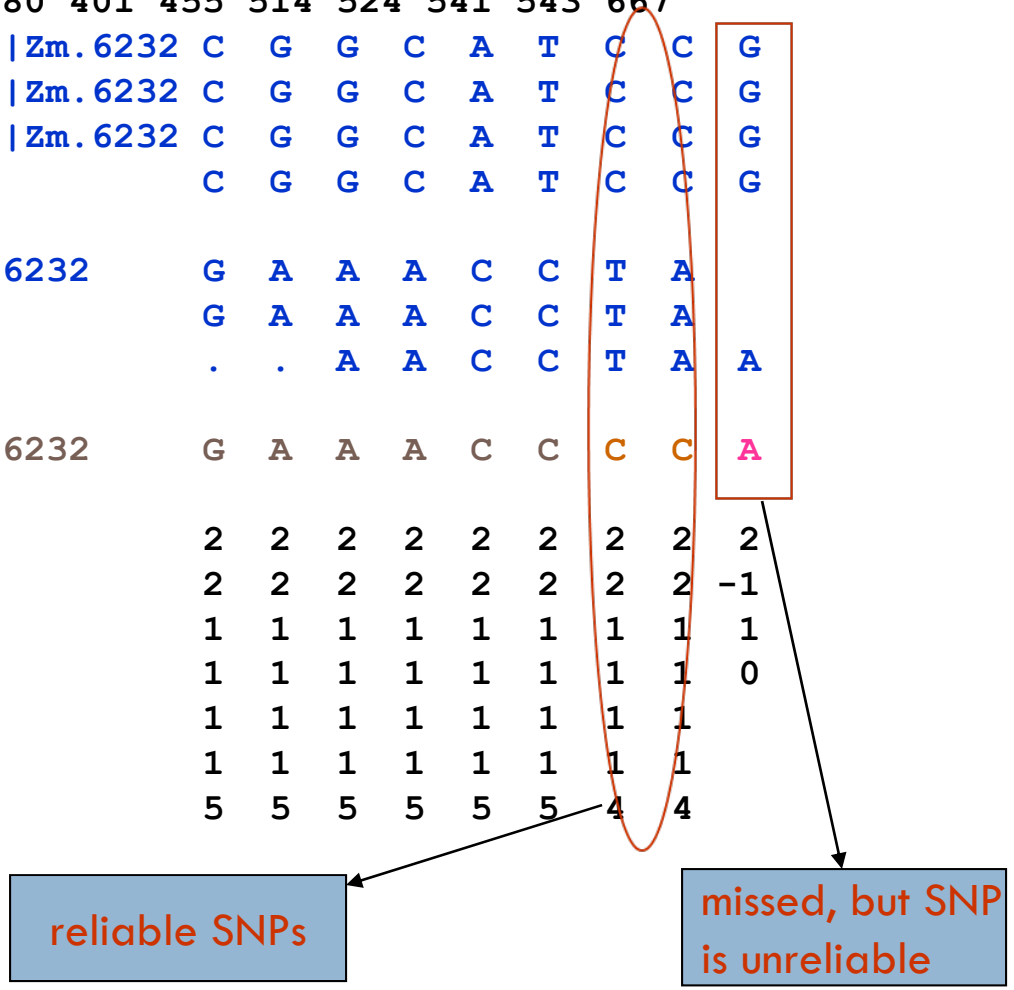
Missed SNPs

9 SNPs known, but Batley missed 2 SNPs

Batley,J., Barker,G., et al. (2003) Mining for Single Nucleotide Polymorphisms and Insertions/Deletions in Maize Expressed Sequence Tag Data. Plant Physiology,132,84-91

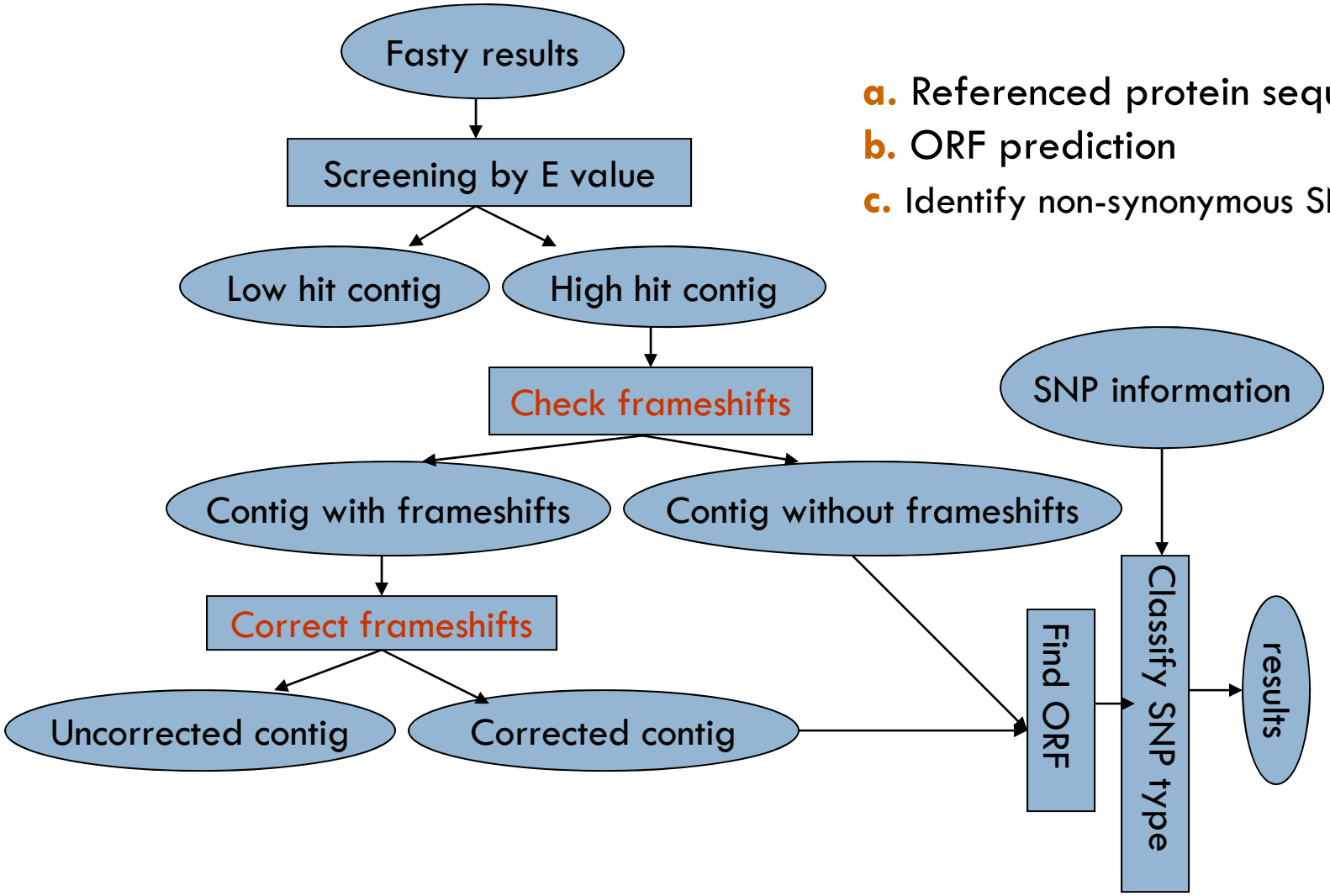
Evaluation of QualitySNP

```
>contig_1 EST:8 SNP:9
location info:                          315 380 401 455 514 524 541 543 667
AI673989|maize_ohio_43|At.23558|Zm.6232 C   G   G   C   A   T   C   C   G
AI677465|maize_ohio_43|At.23558|Zm.6232 C   G   G   C   A   T   C   C   G
AI734825|maize_ohio_43|At.23558|Zm.6232 C   G   G   C   A   T   C   C   G
BG267040|maize|At.23558|Zm.6232         C   G   G   C   A   T   C   C   G

BG264822|maize_B73|At.23558|Zm.6232     G   A   A   A   C   C   T   A
AI855293|maize_B73|                     G   A   A   A   C   C   T   A
BG836349|maize|At.23558|Zm.6232         .   .   A   A   C   C   T   A   A

AW331785|maize_W23|At.23558|Zm.6232     G   A   A   A   C   C   C   C   A

Inter-[2]&intra-[1]:                    2   2   2   2   2   2   2   2   2
SNP type:                               2   2   2   2   2   2   2   2  -1
major allele haplotype score:           1   1   1   1   1   1   1   1   1
minor allele haplotype score:           1   1   1   1   1   1   1   1   0
SNP pattern                             1   1   1   1   1   1   1   1
SNP block                               1   1   1   1   1   1   1   1
Confidence score:                       5   5   5   5   5   5   4   4
```



			QualitySNP(D<= 0.6)				autoSNP				their overlap		
Chromosome	UniGene	Size	Time(m)		Confirmed	unconfirmed	Time(m)		confirmed	unconfirmed	candidate SNPs	confirmed	unconfirmed
6	Hs.300701	3640	2	18	5 (27.8%)	13	150	6	0 (0%)	6	3	0(0%)	3
7	Hs.401316	1090	1	0	0 (0%)	0	3	4	0 (0%)	4	0	0(0%)	0
14	Hs.533717	1601	1	12	3 (25%)	9	26	166	1 (0.6%)	165	0	0(0%)	0
17	Hs.12956	622	1	10	2 (20%)	8	1	15	1 (6.7%)	14	9	1(11.11%)	8
19	Hs.515126	654	1	1	0 (0%)	1	2	44	0 (0%)	44	1	0(0%)	1
15	Hs.22543	847	1	10	1 (10%)	9	1	4	1 (25%)	3	1	1(100%)	0
2	Hs.468478	183	1	0	0 (0%)	0	1	0	0 (0%)	0	0	0(0%)	0
1	Hs.591503	200	1	6	2 (33.3%)	4	1	5	0 (0%)	5	3	0(0%)	3
6	Hs.567284	194	1	7	0 (0%)	7	1	8	0 (0%)	8	7	0(0%)	7
6	Hs.510172	282	1	1	0 (0%)	1	1	0	0 (0%)	0	0	0(0%)	0
17	Hs.406754	6453	2	49	25 (51%)	24	51	43	6 (14%))	37	14	5(35.71%)	9
14	Hs.510635	27193	4	535	198 (37%)	337	13	895	92 (10.3%)	803	143	86(60.14%)	57
7	Hs.61635	82	1	0	0 (0%)	0	1	0	0 (0%)	0	0	0(0%)	0
2	Hs.631881	355	1	5	0 (0%)	5	1	1	0 (0%)	1	0	0(0%)	0
8	Hs.104741	275	1	0	0 (0%)	0	1	0	0 (0%)	0	0	0(0%)	0
2	Hs.534639	1910	1	11	1 (9.1%)	10	6	9	0 (0%)	9	6	0(0%)	6
14	Hs.18069	1965	1	3	1 (33.3%)	2	1	1	0 (0%)	1	0	0(0%)	0
17	Hs.514220	6800	2	8	2 (25%)	6	267	13	0 (0%)	13	2	0(0%)	2
12	Hs.19192	397	1	1	0 (0%)	1	2	0	0 (0%)	0	0	0(0%)	0
Total		54743		677	240 (35.5%)	437		1214	101 (8.3%)	1113	189	93(49.21%)	96

Identify non-synonymous SNP



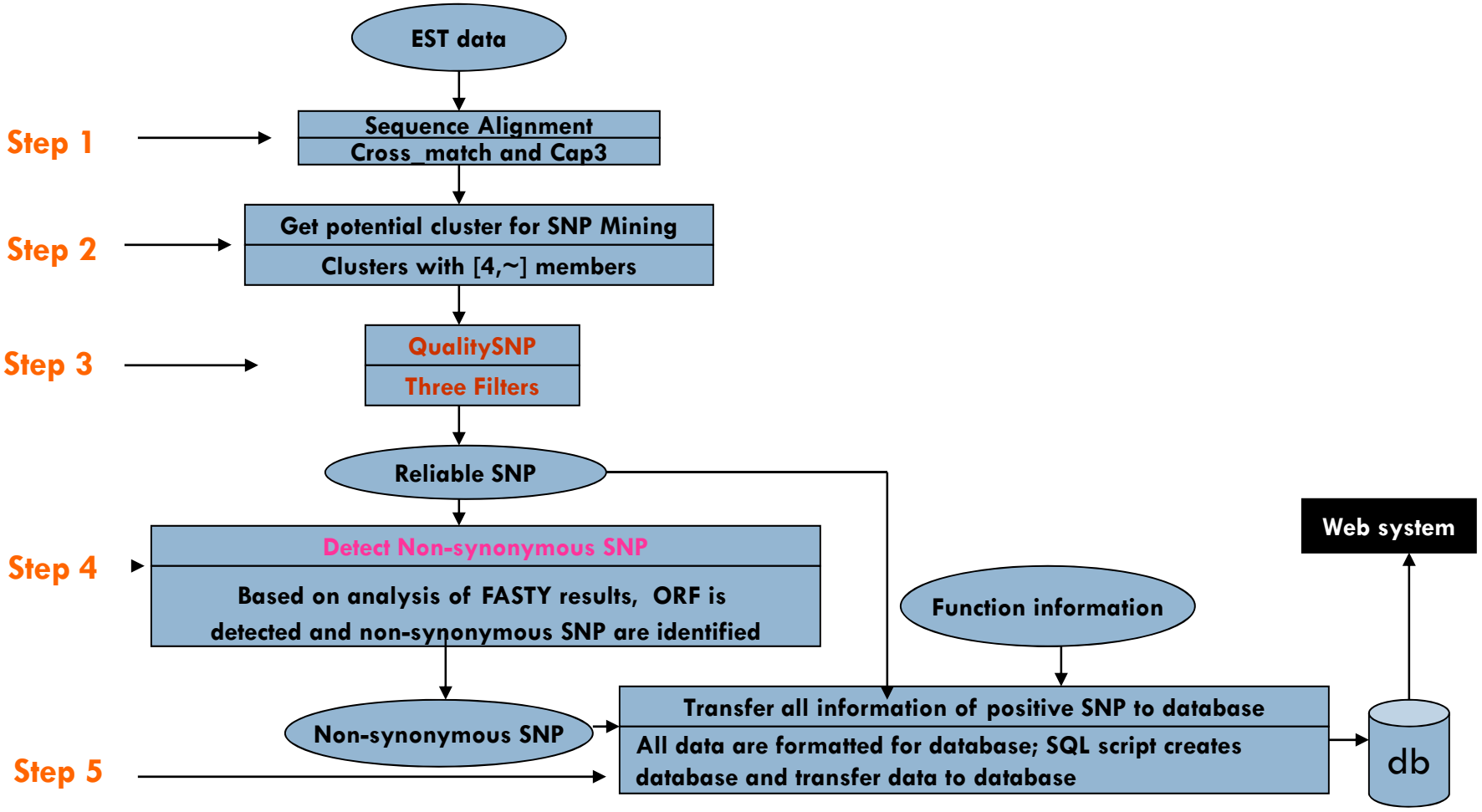
a. Referenced protein sequence

b. ORF prediction

c. Identify non-synonymous SNP

The QualitySNP pipeline

QualitySNP - A pipeline for mining SNP from EST data



Conclusions

- QualitySNP works at least as well as currently available methods, without the drawbacks of some of them, such as the necessity to provide a genomic sequence or sequence quality files. However, if quality files are available, this information can also be used by QualitySNP
- Using a haplotype-based strategy, QualitySNP not only predicts reliable SNPs but also identifies haplotypes, and thus can be used in EST-based genotyping
- The haplotype-based strategy can make full use of redundancy in sequences by reclustering them, not only to avoid influence of sequencing errors but also to remove poor quality sequences which might be single haplotypes
- QualitySNP identify paralogs and reliable SNPs on heterozygous diploid as well as polyploid species
- The method has been applied successfully on potato EST data from public sequence databases

SNPs results from potato EST data

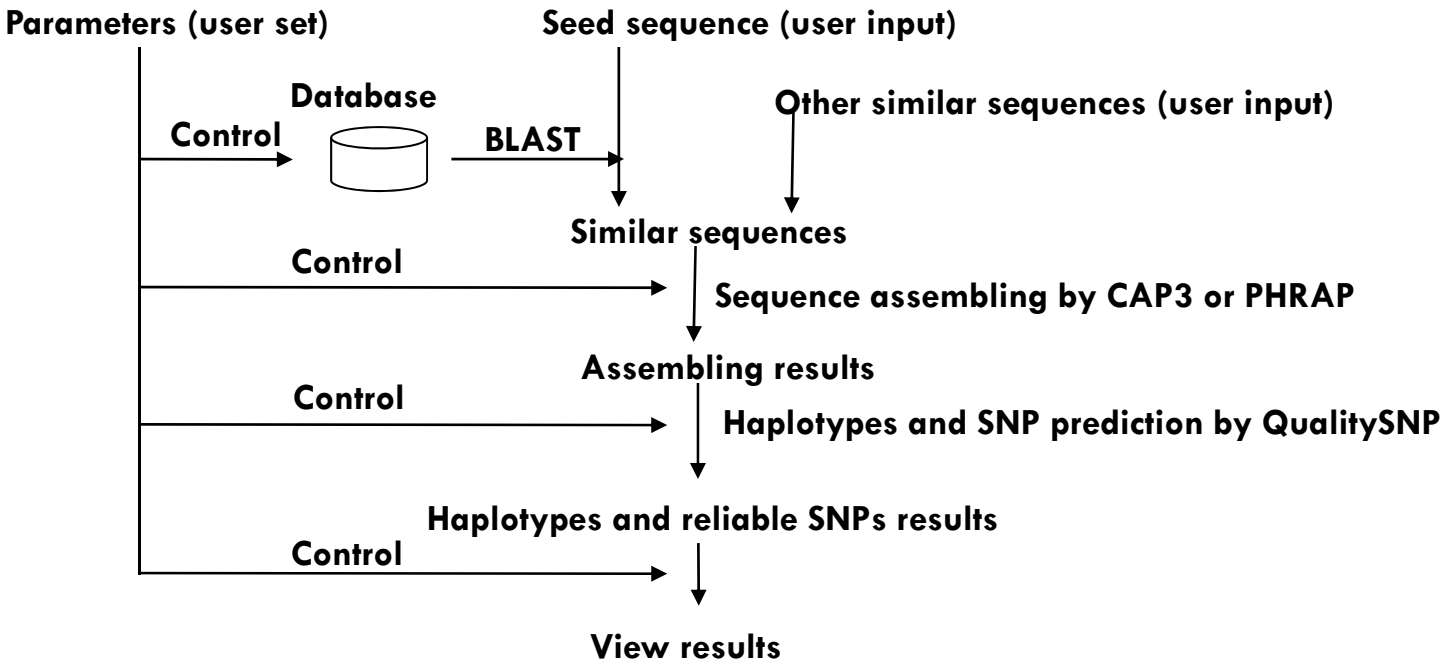
Title	Kennebec EST
total EST	83565
total contigs	10670
total contigs with SNP	3081
potential SNP statistic analysis	
total potential SNPs including tri-SNP	31815
bp/SNP	118.1
bp/indel	790.1

Reliable SNPs with confidence score more than 1(2651 clusters without potential paralogs clusters Under D-value less than 0.6)	
reliable SNP	16772
bp/SNP	224.0
bp/indel	2070
Transition (AG,CT)	9853
Transversion (AT, AC, CG,TG)	5057
Indel	1815
tri-SNP	47
tr/tv	1.95
reliable SNP/potential SNP	0.67

nsSNP analysis (without potential paralogs clusters)	
total contigs	2651
hit contigs	2576
lowhits(fasty)	75
high hit	2576
frameshifts(fasty)	506
contig with ORF	2065
corrected frameshifts contig(fasty)	102
total contig with ORF	2167
contig with uncorrected frameshifts	409
total bi-SNP	14188
Indel	1523
SNP without Indel 3' UTR	475
SNP without Indel 5' UTR	1836
SNP without Indel in UTR	0.16 (2311/14188)
Indel	0.11 (1523/14188)
bi-SNP in coding region	0.73 (10354/14188)
nsSNP coding region	0.34 (3536/10354)

HaploSNPer allele and SNP discovery

- A flexible web-based tool for detecting alleles and SNPs in user specified input sequences from diploid and polyploid species



HaploSNPer

allele and SNP detection tool

User Input

Seed sequence (example)

Browse..

Other similar sequences (example)

Browse..

Select a database (data info)

--- Select species data set ---

Select a sequence alignment program

CAP3

Help

- Quick start
- Online manual
- Download manual (PDF)

Related publications

- QualitySNP (BMC bioinformatics)

Pre-processing of sequences

Remove vectors

☐ Use Cross_match to remove vectors

Mask repeats

☐ Use RepeatMasker to mask repeats

Parameters

BLAST / CAP3 and PHRAP

E-value 1e-60

Similarity for CAP3 and PHRAP 95 percent (must be > 65)

Haplotype construction

Similarity per polymorphic site 0.75 [0..1]

Similarity over all polymorphic sites 0.80 [0..1]

Low quality region

LQ length from 5'side 30 nucleotides

LQ length from 3'side 20 percent of the whole sequence

The weight value of LQ region 0.5 [0..1]

SNP detection

Minimum cluster size 4 sequences

Minimum redundancy of each allele 2 sequences

Minimum confidence score 2 [1..5]

Output format

Output format

○ Only cluster containing the seed sequence

◉ All clusters related to the seed sequence

Submit your job to HaploSNPer

Your e-mail address (optional):

(Output returned by email instead of waiting) Submit Reset

Last Modified March 23, 2009

WAGENINGEN UR
For quality of life

HaploSNPer - results

Your input parameters

You have chosen the **Human** EST database.

E-value for BLAST: **1e-60**
Minimum number of sequences forCluster: **4**
Minimum number of sequences for each allele for every potential SNP: **2**
Low quality region from 5' side : **30 nucleotides**
Low quality region from 3' side : **20% of the whole length of sequence**
Weight value of low quality region: **0.5**
You have chosen CAP3 for sequence alignment and similarity for CAP3: **95%**
Similarity per polymorphic site: **0.75**
Similarity over all polymorphic sites: **0.80**
Minimum Confidence Score: **2**

Summary Information

Seed sequence:**myseedNM_003087.1|SNCG** is found in **Cluster2**
List all clusters related to your input sequences

Cluster	Number of ESTs	Number of Potential SNPs
2	**111**	**60**
4	**4**	**0**
7	**11**	**2**
1	**2**	**0**
3	**3**	**0**
5	**2**	**0**
6	**3**	**0**

Summary of SNPs and haplotypes information

Cluster	Number of potential SNPs in reliable haplotypes	Number of reliable SNPs	D-value	Number of haplotypes	Number of single haplotypes
2	**55**	**14**	**0.5435**	**17**	**11**
7	**2**	**1**	**0.1768**	**9**	**7**

Statistic information of potential and reliable SNPs

SNP	C/T	A/G	transition	A/T	A/C	C/G	T/G	transversion	Indel
Potential SNPs	**12**	**3**	**15**	**5**	**12**	**6**	**6**	**29**	**12**
reliable SNPs	**7**	**1**	**8**	**1**	**4**	**2**	**0**	**7**	**0**

Potential SNPs: variations are defined by Minimum size of each allele, and they include bi-allelic,tri-allelic,tetra-allelic, penta-allelic SNPs
Reliable SNPs: reliable SNPs are identified by QualitySNP from bi-allelic SNPs.
Haplotype: a group of sequences within a cluster that represent the same allele of a gene
Reliable haplotype: a haplotype containing at least 2 sequences
Single haplotypes: a haplotype containing only 1 sequence
D-value: it is calculated and used to identify paralogs by QualitySNP
For further information on parameters and output please consult the manual and the QualitySNP reference.
Download results

HaploSNPer - results

	consensus sequence	G*TCACCTCCGGGGTGGTGCGCAAGGAGGACTTGA*GGCCA*TCTGCCCCCC*AACAGGAGGGTGAGGCATCCAAAGAGAAAGAGGAAGTGGCAG*AGGAGG
	SNP sign	......*..*.....*.....*.....*.....
ID	sequence name	410.....420.....430.....440.....450.....460.....470.....480.....490.....500.....51
3	BX474511	G*TCACCTCCGGGGTGGTGCGCAAGGAGGACTTGA*GGCCA*TCTGCCCCCC*AACAGGAGGGTGAGGCATCCAAAGAGAAAGAGGAAG
3	CD616264	G*TCACCTCCGGGGTGGTGCGCAAGGAGGACTT
3	CD616265	G*TCACCTCCGGGGTGGTGCGCAAGGAGGACTT
3	DA115597	G*TCACCTCCGGGGTGGTGCGCAAGGAGGACTTGA*GGCCA*TCTGCCCCCC*AACAGGAGGGTGAGGCATCCAAAGAGAAAGAGGAAGTGGCAG*AGGA
3	DA122942	G*TCACCTCCGGGGTGGTGCGCAAGGAGGACTTGA*GGCCA*TCTGCCCCCC*AACAGGAGGGTGAGGCATCCAAAGAGAAAGAGGAAGTGGCAG*AGGA
3	DA190029	G*TCACCTCCGGGGTGGTGCGCAAGG
3	DN994208	G*TCACCTCCGGGGTGGTGCGCAAGGAGGACTTGA*GGCCA*TCTGCCCCCC*AACAGGAGGGTGAGGCATCCAAAGAGAAAGAGGAAGTGGCAG*AGGA
3	DT2200	Homo sapiens|Whole brain Human adult whole brain, large insert, pCMV|unknown GGGAGGACTTGA*GGCCA*TCTGCCCCCC*AACAGGAGGGTGAGGCATCCAAAGAGAAAGAGGAAGTGGC
3	DT220934	G*TCACCTCCGGGGTGGTGCGCAAGGAGGACTTGA*GGCCA*TCTGCCCCCC*AACAGGAGGGTGAGGCATCCAAAGAGAAAGAGGAAGTGGCAG*AGGA
4	BG707764	G*TCACCTCCGGGGTGGTGCGCAAGGAGGACTTGA*GGCCA*TCTGCCCCCC*AACAGGAGGGTGTGGCATCCAAAGAGAAAGAGGAAGTGGCAG*AGGA
4	BI597796	G*TCACCTCCGGGGTGGTGCGCAAGGAGGACTTGA*GGCCA*TCTGCCCCCC*AACAGGAGGGTGTGGCATCCAAAGAGAAAGAGGAAGTGGCAG*AGGA
4	DB504413	G*TCACCTCCGGGGTGGTGCGCAAGGAGGACTTGA*GGCCA*TCTGCCCCCC*AACAGGAGGGTGTGGCATCCAAAGAGAAAGAGGAAGTGGCAG*AGGA
4	DB575656	G*TCACCTCCGGGGTGGTGCGCAAGGAGGACTTGA*GGCCA*TCTGCCCCCC*AACAGGAGGGTGTGGCATCCAAAGAGAAAGAGGAAGTGGCAG*AGGA
5	BQ006197	.....A
5	CA421283	G*TCACCTCCGGGGTGGTGCGCAAGGAGGACTTGA*GGCCA*TTTGCCCCCC*AACAGGAGGGTGAGGCATCCAAAGAGAAAGAGGAAGTGGCAG*AGGA
6	BI603159	G*TCACCTCAGGGGTGGTGCGCAAGGAGGACTTGA*GGCCA*TCTGCCCCCC*AACAGGAGGGTGTGGCATCCAAAGAGAAAGAGGAAGTGGCAG*AGGA
6	BI603171	G*TCACCTCAGGGGTGGTGCGCAAGGAGGACTTGA*GGCCA*TCTGCCCCCC*AACAGGAGGGTGTGGCATCCAAAGAGAAAGAGGAAGTGGCAG*AGGA

HaploSNPer - results

Reliable SNPs and Haplotypes information

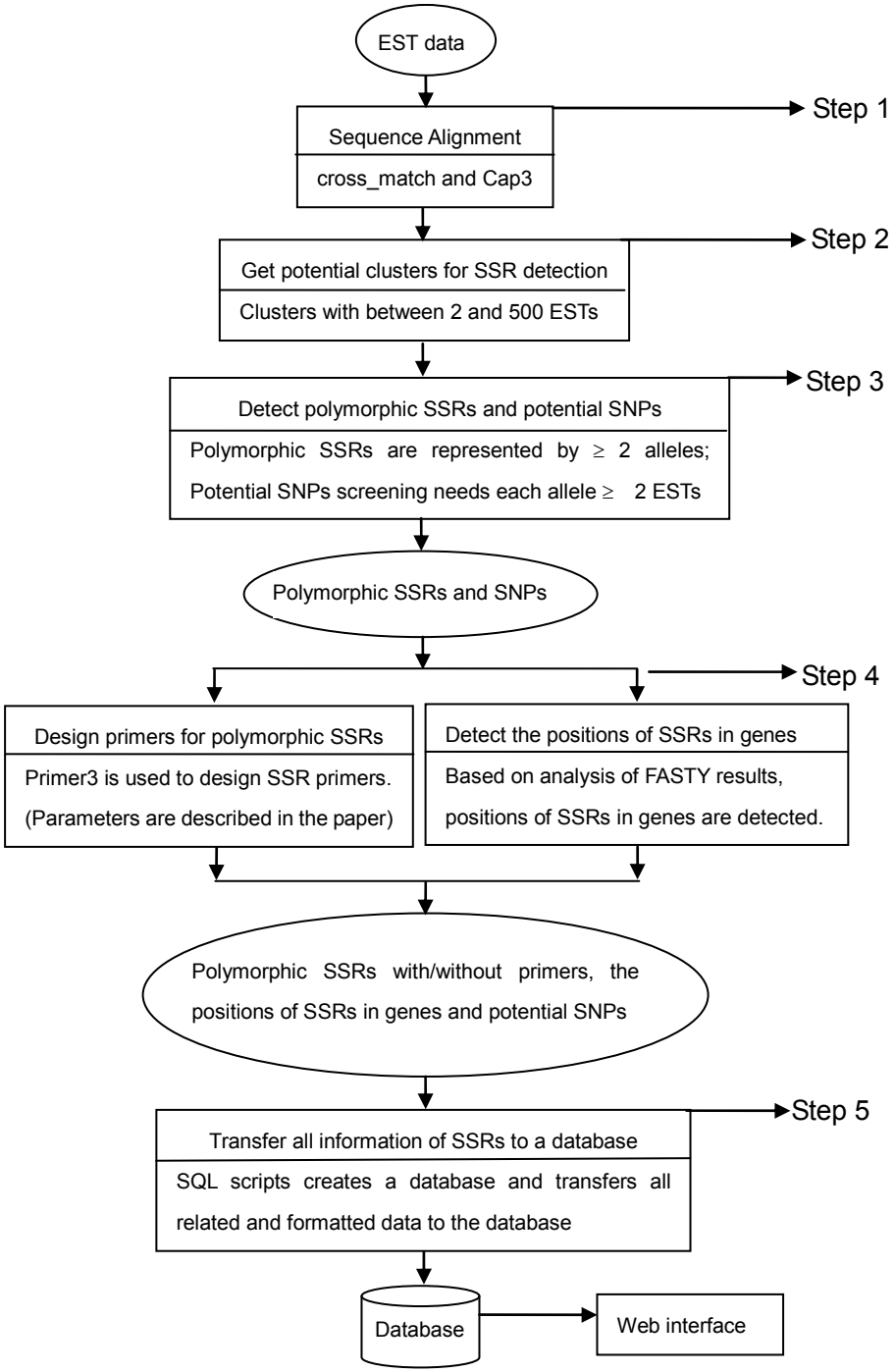
Haplotype ID	Sequence name	SNP location	88	118	265	335	419	475	577	580	584	664	696	699	714	744
1	**myseedNM_003087.1**		.	A	C	G	C	T	C	C	C	C	C	C	C	C
1	BP381244		A	A	C	G	C	T	C	C	C	C	C			
1	BQ901053		A	A	C	G	C	T	C	C	C	C	C	C	C	C
1	BU157619		A	A	C	G	C	T	C	C	C	C	C	C	C	C
1	BU179779		A	A	C	G	C	T	C	C	C	C	C	C	C	C
1	BX090816	Homo sapiens|ovary Soares ovary tumor NbHOT adult, 36 years|unknown					C	T	C	C	C	C	C	C	C	C
	Other 39 sequences in haplotype 1 were not shown in here															
2	BI836596		A	A	C	C	C	A	C	C	C	C	C	C	C	C
2	BM665098		.	.	.	C	C	A	C	C	C	C	C	C	C	C
2	BP212912		A	A	C	C	C	A	C	C	C					
2	CA443299		.	.	C	C	C	A	C	C	C	C	C	C	C	C
2	CB107161		A	A	C	C	C	A								
2	AL712443		A	A	C	C	C	A								
	Other 19 sequences in haplotype 2 were not shown in here															
3	BG328738		A	C	C	C	C	A	C	C	C	C	C	C	C	C
3	BI757131		A	C	C	C	C	A	C	C	C	C	C	C	C	
3	BM921124		A	C	C	C	C	A	C	C	C	C	C	C	C	C
3	BQ221776		A	C	C	C	C	A	C	C	C	C	C	C	C	C
3	BQ439430		A	C	C	C	C	A	C	C	C	C	C	C	C	C
3	BQ882072		A	C	C	C	C	A	C	C	C	C	C	C	C	C
	Other 16 sequences in haplotype 3 were not shown in here															
4	BG707764		G	A	C	G	C	T	G	C	C	C	C	C	C	
4	BI597796		G	A	C	G	C	T	G	C	C	C	C	C	C	C
4	DB504413		G	A	C	G	C	T	G	C	C					
4	DB575656		.	.	.	G	C	T	G	C	C	C	C	C	C	C
5	BQ006197		.	.	.	.	.	.	C	T	C	T	T	T	T	T
5	CA421283		.	.	.	.	C	A	C	T	C	T	T	T	T	T
6	BI603159		A	A	T	G	A	T	C	C	A	C	C	C	C	C
6	BI603171		A	A	T	G	A	T	C	C	A	C	C	C	C	C

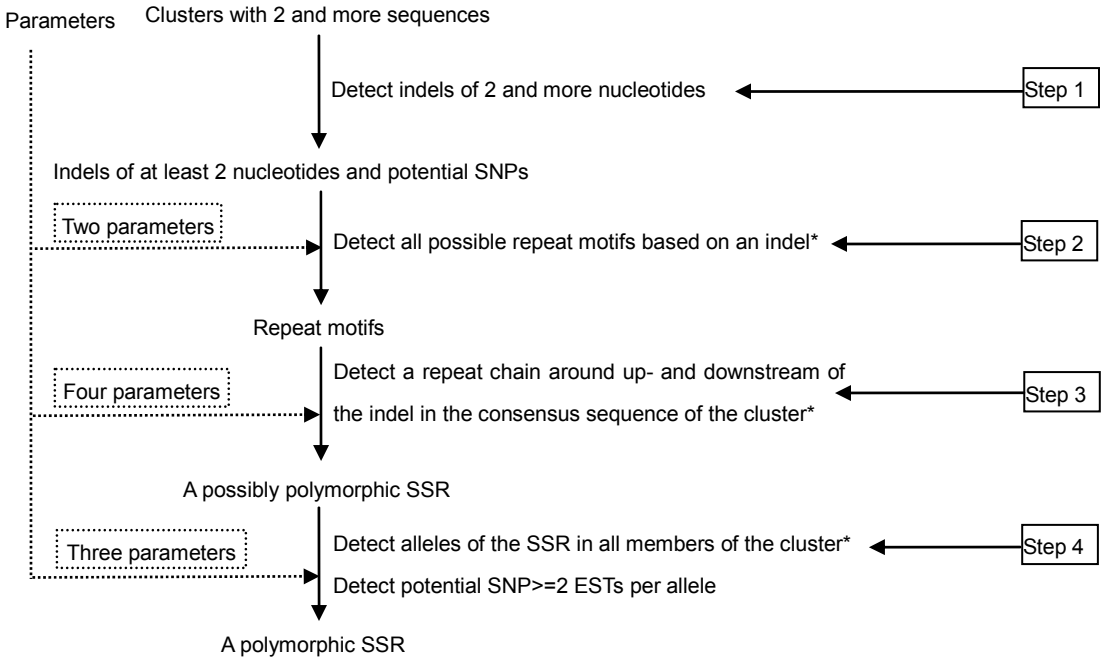
Conclusions

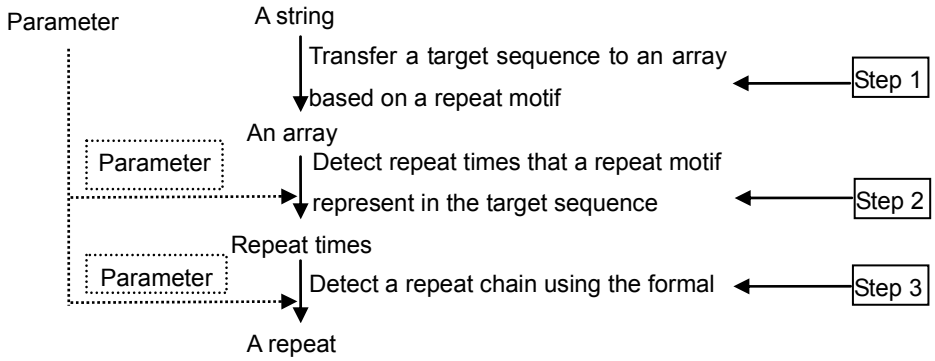
- QualitySNP works at least as well as currently available methods, without the drawbacks of some of them, such as the necessity to provide a genomic sequence or sequence quality files. However, if quality files are available, this information can also be used by QualitySNP
- Using a haplotype-based strategy, QualitySNP not only predicts reliable SNPs but also identifies haplotypes, and thus can be used in EST-based genotyping
- The haplotype-based strategy can make full use of redundancy in sequences by reclustering them, not only to avoid influence of sequencing errors but also to remove poor quality sequences which might be single haplotypes
- QualitySNP identify paralogs and reliable SNPs on heterozygous diploid as well as polyploid species
- The method has been applied successfully on potato EST data from public sequence databases (Illumina GoldenGate)

POLYSSR DETECTION

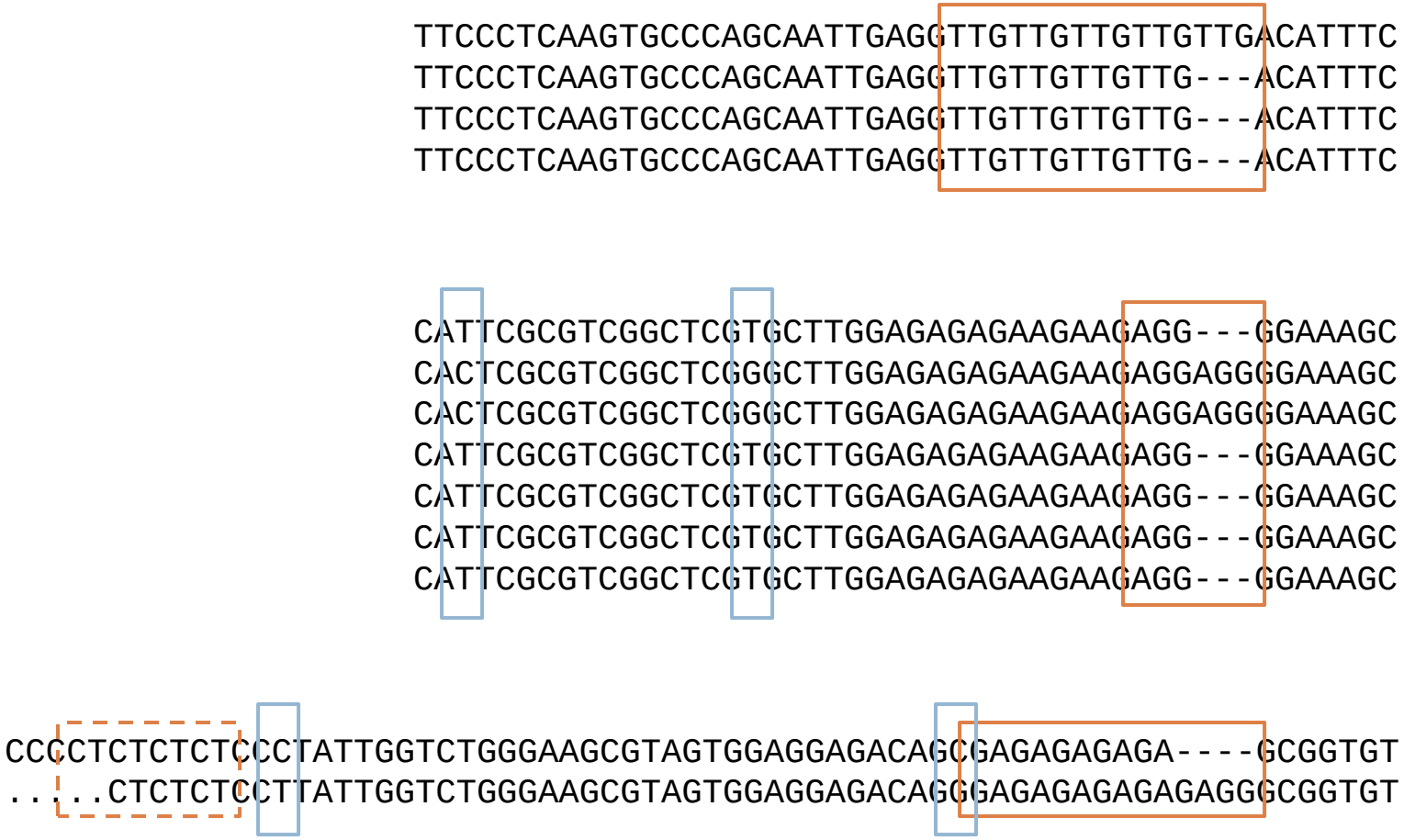
Detection of polymorphic SSRs


EST data
Step 1
Sequence Alignment
cross_match and Cap3
Step 2
Get potential clusters for SSR detection
Clusters with between 2 and 500 ESTs
Step 3
Detect polymorphic SSRs and potential SNPs
Polymorphic SSRs are represented by ≥ 2 alleles;
Potential SNPs screening needs each allele ≥ 2 ESTs
Polymorphic SSRs and SNPs
Step 4
Design primers for polymorphic SSRs
Primer3 is used to design SSR primers.
(Parameters are described in the paper)
Detect the positions of SSRs in genes
Based on analysis of FASTY results,
positions of SSRs in genes are detected.
Polymorphic SSRs with/without primers, the positions of SSRs in genes and potential SNPs
Step 5
Transfer all information of SSRs to a database
SQL scripts creates a database and transfers all related and formatted data to the database
Database
Web interface


Parameters
Clusters with 2 and more sequences
Detect indels of 2 and more nucleotides
Step 1
Indels of at least 2 nucleotides and potential SNPs
Two parameters
Detect all possible repeat motifs based on an indel*
Step 2
Repeat motifs
Four parameters
Detect a repeat chain around up- and downstream of
the indel in the consensus sequence of the cluster*
Step 3
A possibly polymorphic SSR
Three parameters
Detect alleles of the SSR in all members of the cluster*
Step 4
Detect potential SNP>=2 ESTs per allele
A polymorphic SSR


Parameter
A string
Transfer a target sequence to an array
based on a repeat motif
Step 1
An array
Parameter
Detect repeat times that a repeat motif
represent in the target sequence
Step 2
Repeat times
Parameter
Detect a repeat chain using the formal
Step 3
A repeat

Examples


TTCCCTCAAGTGCCCAGCAATTGAGGTTGTTGTTGTTGTTGACATTTC
TTCCCTCAAGTGCCCAGCAATTGAGGTTGTTGTTGTTG---ACATTTC
TTCCCTCAAGTGCCCAGCAATTGAGGTTGTTGTTGTTG---ACATTTC
TTCCCTCAAGTGCCCAGCAATTGAGGTTGTTGTTGTTG---ACATTTC
CATTCGCGTCGGCTCGTGCTTGGAGAGAGAAGAAGAGG---GGAAAGC
CACTCGCGTCGGCTCGGGCTTGGAGAGAGAAGAAGAGGAGGGGAAAGC
CACTCGCGTCGGCTCGGGCTTGGAGAGAGAAGAAGAGGAGGGGAAAGC
CATTCGCGTCGGCTCGTGCTTGGAGAGAGAAGAAGAGG---GGAAAGC
CATTCGCGTCGGCTCGTGCTTGGAGAGAGAAGAAGAGG---GGAAAGC
CATTCGCGTCGGCTCGTGCTTGGAGAGAGAAGAAGAGG---GGAAAGC
CATTCGCGTCGGCTCGTGCTTGGAGAGAGAAGAAGAGG---GGAAAGC
CCGCTCTCTCTCCCTATTGGTCTGGGAAGCGTAGTGGAGGAGACAGCGAGAGAGAGA----GCGGTGT
.....CTCTCTCCTTATTGGTCTGGGAAGCGTAGTGGAGGAGACAGGGAGAGAGAGAGAGGGCGGTGT

Polymorphic SSRs mining for EST data

Home
Search db

PolySSR retrieval system - example database

SSR type: all types (* required)

SSR position: ----------

Reference species: potato (EMBL 88) (* required)

Search by text:

Advanced filter for: ----------

Output settings :

- only poly-SSR with primers
- both poly-SSR with or without primers

results are available for downloading

Search Reset

Searching parameters

Brassica(including all brassica species) has been chosen!
Search all-polymorphic SSRs

Output setting:
Polymorphic SSRs both with or without primers will be shown

Searching Results

997 polymorphic SSRs, of which **937** with primers are found in the database
997 SSRs with and without primers are listed in the below table

SSR position: hit: a SSR locates in coding region with available reference protein; coding: a SSR is in coding region without reference protein; 3UTR: a SSR locates in 3' UTR; 5UTR: a SSR locates in 5' UTR. tstart: a SSR is around the translation start sites; tstop: a SSR is around the translation stop.

allele info consists of allele 1:repeat times:sequences in the allele - allele 2:repeat times:sequences in the allele.....
species within alleles : allele 1(species1,species2 ...) - allele 2(species1,species2 ...).....
primers:at most five pairs of primers are listed with product size,the start position of left primer,left primer,the temperature of left primer, the start position of right primer,right primer,the temperature of right primer

Cluster ID	SSR motif	SSR start position	SSR stop position	SSR position	SSR type	Minimum repeat times	No.of alleles	minimum sequence per allele	allele info	species within alleles	product size	t l s
29	ACGTCC	62	82	hit	6	3	2	1	1:4:3-2:3:1	1(Brassica oleracea var. alboglabra_A12Dhd,Brassica napus,Brassica oleracea var. alboglabra_A12DHd)-	129	3
71	ATG	446	459	hit	3	4	2	1	1:4:1-2:5:2	1(Brassica oleracea var. alboglabra_A12Dhd)-2(Brassica napus)	355	2
119	CT	113	122	5utr	2	4	2	1	1:4:1-2:5:1	1(Brassica oleracea var. alboglabra_A12Dhd)-2(Brassica napus)	339	4
212	CTAC	971	986	3utr	4	5	2	3	1:4:3-2:3:4	1(Brassica rapa subsp. pekinensis,Brassica oleracea var. alboglabra_A12DHd)-2(Brassica napus_samoura	396	6
231	AC	285	300	3utr	2	5	2	1	1:8:1-2:6:2	1(Brassica oleracea var. alboglabra_A12Dhd)-2(Brassica	301	8

Acknowledgement

- Jifeng Tang
- Ben Vosman
- Roeland Voorrips
- Gerard van der Linden

URL = http://www.bioinformatics.nl/tools/