Cite this poster as:

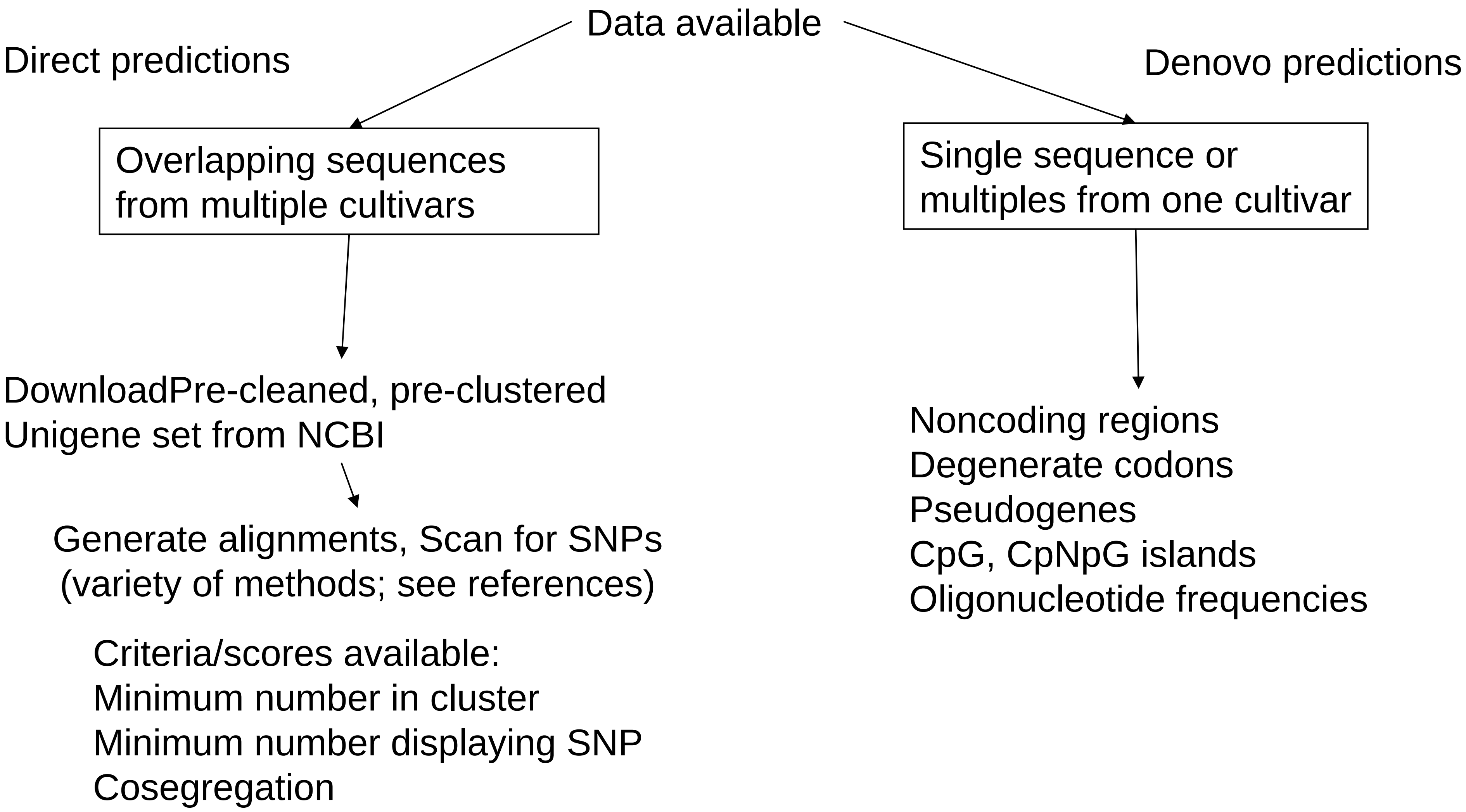
Baldo, A.M. and J. Labate. 2003. Polymorphism prediction. ISMB 2003, Brisbane, Australia.

Abstract

The distribution of single nucleotide polymorphisms (SNPs) has been demonstrated as nonrandom in the genomes of animals and plants in a number of recent studies. While genetic variation is traditionally expected in noncoding regions, additional genetic features such as CpG islands, particular codons, pseudogenes, and oligonucleotide composition have been correlated with the presence of SNPs. We investigate whether this information might be useful in the context of predicting regions more likely to contain genetic markers in agricultural crops.

When multiple overlapping sequences are available in the form of expressed sequence tags (ESTs), a more direct approach for SNP prediction is available. We outline here our current approach and plans for incorporating both approaches in the future.

Approaches



Data available

Direct predictions

Overlapping sequences
from multiple cultivars

DownloadPre-cleaned, pre-clustered
Unigene set from NCBI

Generate alignments, Scan for SNPs
(variety of methods; see references)

Criteria/scores available:
Minimum number in cluster
Minimum number displaying SNP
Cosegregation

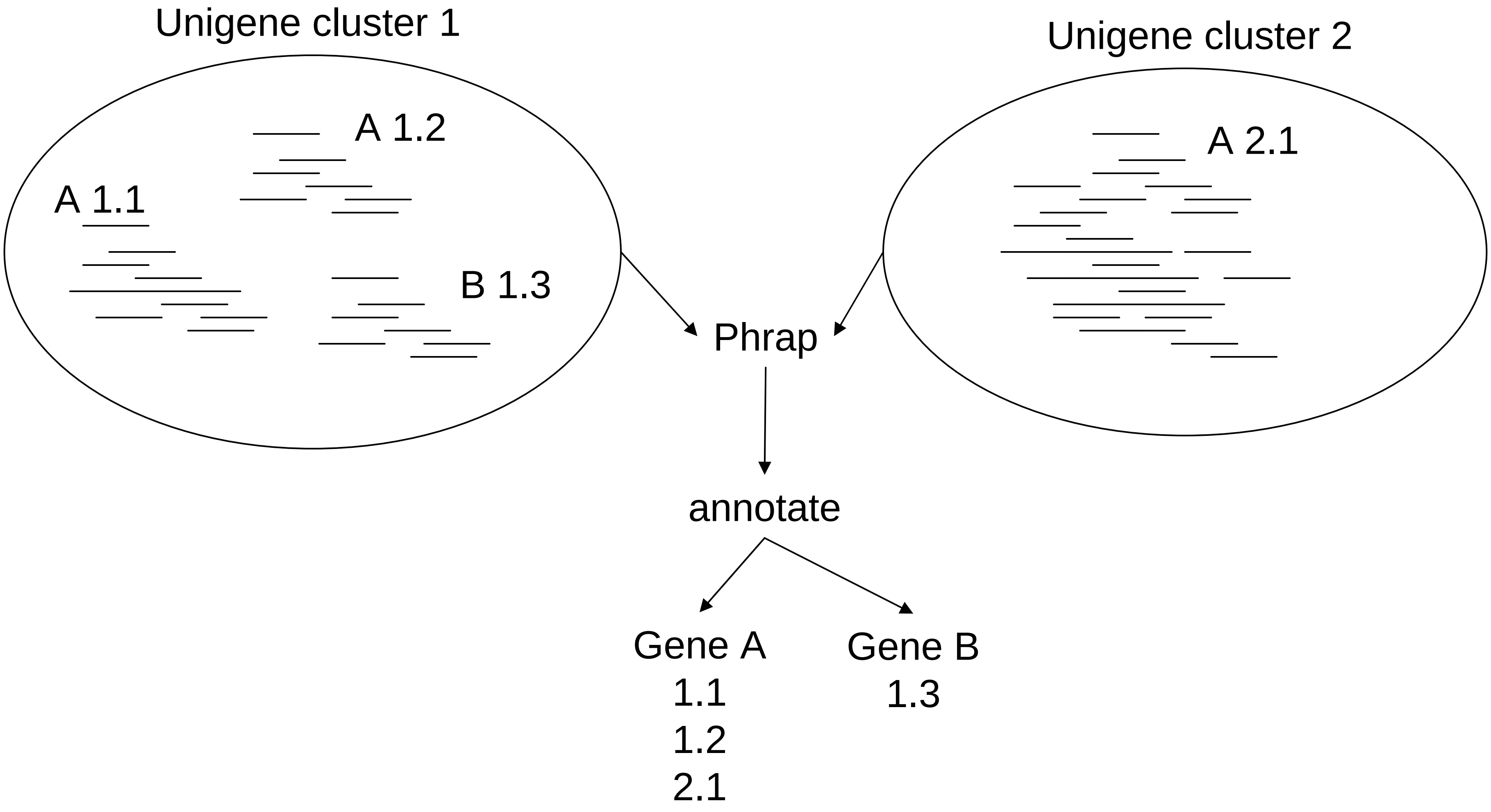
Denovo predictions

Single sequence or
multiples from one cultivar

Noncoding regions
Degenerate codons
Pseudogenes
CpG, CpNpG islands
Oligonucleotide frequencies

What is a “cluster”?

Unigenes do not unambiguously cluster with Phrap, and are a moving target as data are continuously deposited



Unigene cluster 1

A 1.2

A 1.1

B 1.3

Unigene cluster 2

A 2.1

Phrap

annotate

Gene A
1.1
1.2
2.1

Gene B
1.3

How much data is necessary to establish atypical nucleotide distributions?

Number of SNPs	p value 5’ deviation (A +1.43%, C +4.91%, G -1.70%, T -4.62%)*	p value 3’ deviation (A -4.44%, C -1.59%, G +5.05%, T +0.99%)*
200	0.320	0.247
500	0.033	0.016
1000	0.001	0.000
2.6 million (actual number reported, from unconfirmed SNPs)	0.000	0.000

* Zhao et. al., 2002 reported human background nucleotide frequencies:
A 29.55%, C 20.44%, G 20.46%, T 29.54% (chromosomal GC content 38.26%-48.33%)

How much confirmed SNP data is available in plants?

Crop	Predicted	Confirmed	Reference
Arabidopsis		37,344	Jander, et. al. 2002 http://www.arabidopsis.org/Cereon
Maize	14,832	264	Batley, et. al. 2003
Rice	2,800	213	Nasu, et. al. 2002
Soybean		234	Zhu, et. al. 2003 http://ncbi.nih.gov/SNP
Human (for reference)	4,145,633	512,247	Zhao, et. al. 2002 http://ncbi.nih.gov/SNP

Direct Method Conclusions

Current estimates of prediction method accuracy are as high as 80% (various authors, pers. comm.) Some populations of organisms exhibit a higher degree of polymorphism than others. A comparision of available methods using the same data sets is needed.

Not all methods incorporate all scoring and acceptance criteria. Valuable criteria (some of which are currently unavailable) include:

- Cosegregation (as a check for accuracy as well as estimation of linkage disequilibrium among SNPs and/or population substructure)
- Cross-library validation
- Intra-variety/population validation
- Sequence quality prediction in the absence of trace files
- Mapping SNPs onto predicted coding regions, reporting/scoring synonymous/nonsynonymous changes
- Processing validation data for accuracy reporting

Denovo Conclusions

There is a need for SNP distribution data in plants.

At least 1000 samples are necessary to distinguish the kinds of neighboring nucleotide frequency differences that have been observed in humans. A similar study has not been conducted in plants, but is feasable with current *Arabidopsis* data.

A sliding window of fewer than 1000 nucleotides would also be insufficient to distinguish regions of similar atypical distributions. Di and trinucleotide distributions might facilitate smaller windows.

It may be possible to create an overlapping scoring scheme based on observations of plant SNP distribution among: CpG and CpNpG islands, degenerate codons, noncoding regions, etc.

References

Balasubramanian, S., Harrison, P., Hegyi, H., Bertone, P.,Luscombe, N., Echols, N., McGarvey, P., Zhang, Z.L., Gerstein, M. 2002. SNPs on human chromosomes 21 and 22 – analysis in terms of protein features and pseudogenes. Pharmacogenomics 3(3):393-402

Barker, G., Batley, J., O'Sullivan, H., Edwards, K.J., Edwards, D. 2003. Redundancy based detection of sequence polymorphisms in expressed sequence tag data using autoSNP. Bioinformatics 19(3):421-422

Batley, J., Barker, G., O'Sullivan, H., Edwards, K.J., Edwards, D. 2003. Mining for Single Nucleotide Polymorphinsms and Insertions/Deletions in Maize Expressed Sequence Tag Data. Plant Physiology 132:84-91

Estivill, X., Cheung, J., Angel Pujana, M., Nakabayashi, K., Scherer, S.W., Tsui, L.C. 2002. Chromosomal regions containing high-density and ambiguously mapped putative single nucleotide polymorphisms (SNPs) correlate with segmental duplications in the human genome. Hum. Mol. Genet. 11(17):1987-1995

Green, P. http://www.phrap.org

Huntley, D. 2003. SEAN SNP Prediction and Display Programs http://zebrafish.doc.ic.ac.uk/Sean

Jander, G; Norris, SR; Rounsley, SD; Bush, DF; Levin, IM; Last, RL 2002. Arabidopsis Map-Based Cloning in the Post-Genome Era Plant Physiol. 129:440-450

Kuittinen, H., Salguero, D., Aguade, M. 2002. Parallel Patterns of Sequence Variation Within and Between Populations at Three Loci of Arabidopsis thaliana. Mol. Biol. Evol. 19(11):2030-2034

Lazo, G.R., Lui, N., You, F., Hummel, D., Chao, S., and Anderson, O.D. 2003. Charting Contig-Component Relationships Within The Triticeae. 23rd Stadler Genetics Symposium, Genome Exploitation: Data Mining, J.P. Gustafson, ed., March 31-April 2, 2003, Columbia, MO. Kluwer Academic/Plenum Publishers, New York (in press).

Lercher, M.J., Hurst, L.D. 2002. Human SNP variability and mutation rate are higher in regions of high recombination. TIG 18(7):337-340

Majewski, J., Ott, J. 2003. Amino acid substitutions in the human genome: evolutionary implications of single nucleotide polymorphisms. Gene 305:167-173

Morton, B.R., Oberholzer, V.M., Clegg, M.T. 1997. The Influende of Specific Neighboring Bases on Substitution Bias in Noncoding Regions of the Plant Chloroplast Genome. J. Mol. Evol. 45:227-231

Nasu, S., Suzuki, J., Ohta, R., Hasegawa, K., Yui, R., Kitazawa, N., Monna, L., Minobe, Y. 2002. Search for and Analysis of Single Nucleotide Polymorphisms (SNPs) in Rice (Oryza sativa, Oryza rufipogon) and Establishment of SNP Markers. DNA Research 9:163-171

Rafalski, A. 2002. Applications of single nucleotide polymorphisms in crop genetics. Curr. Opin. Plant Bio. 5:94-100

Tomso, D.J., Bell, D.A. 2003. Sequence Context at Human Single Nucleotide Polymorphisms: Overrepresentation of CpG Dinucleotide at Polymorphic Sites and Suppression of Variation in CpG Islands. JMB 237:303-308

Zhao, Z.M., Boerwinkle, E. 2002. Neighboring-Nucleotide Effects on Single Nucleotide Polymorphisms: A Study of 2.6 Million Polymorphisms Across the Human Genome. Genome Research 12:1679-1686

Zhu, Y.L., Song, Q.J., Hyten, D.L., VanTassell, C.P., Matukumalli, L.K., Grimm, D. R., Hyatt, S. M., Fickus, E.W., Young, N.D., Cregan, P.B. 2003. Single-Nucleotide Polymorphisms in Soybean Genetics 163:1123-1134