Cold Spring Harbor Protocols
cshprotocols.cshlp.org

doi:10.1101/pdb.prot4786
Cold Spring Harb Protoc 2007. 2007: pdb.prot4786-

Protocol

SNP Mining from Maize 454 EST Sequences

W. Brad Barbazuk[1,4], **Scott Emrich**[2] **and Patrick S. Schnable**[3]

Author Affiliations

[4]Corresponding author (bbarbazuk@danforthcenter.org)

INTRODUCTION

In this protocol, 454 expressed sequence tags (ESTs) are generated by sequencing shoot apical meristem (SAM) cDNA from maize inbred lines on the 454 Life Sciences GS-20 sequencing system. The computational tool PolyBayes (Marth et al. 1999) is then used to identify single-nucleotide polymorphisms (SNPs). PolyBayes has been used successfully to identify SNPs in many different systems, including maize, and is particularly recommended for identifying SNPs in 454 sequences.

RELATED INFORMATION

For related protocols on tissue preparation, RNA extraction, and amplification, refer to Maize Tissue Preparation and Extraction of RNA from Target Cells for Genotyping and T7-Based RNA Amplification for Genotyping from Maize Shoot Apical Meristem. The use of PolyBayes to identify SNPs in maize is described in Useche et al. (2001).

MATERIALS

Reagents

Maize shoot apical meristem (SAM) cDNA from inbred lines B73 and Mo17

Prepare the cDNA as described in Maize Tissue Preparation and Extraction of RNA from Target Cells for Genotyping and T7-Based RNA Amplification for Genotyping from Maize Shoot Apical Meristem.

Equipment

454 Life Sciences Genome Sequencer 20 (GS 20)

454 Life Sciences has a sequencing service center that will provide sequences from cDNA and genomic DNA samples. Inquiry with the company regarding requirements for cDNA quantity and quality is recommended.

BLAST (Altschul et al. 1990)

Cross_match (P. Green, unpubl.)

PolyBayes (http://genome.wustl.edu/tools/software/polybayes.cgi)

METHOD

1. Generate 454 ESTs by sequencing SAM cDNA from the maize inbred lines B73 and Mo17 on the 454 Life Sciences GS-20 sequencing system.

2. Assign 454 ESTs to maize genomic anchor sequences using BLAST. Identify the highest-scoring alignment between each 454 EST and the collection of genomic sequences (1e−8 minimum *E*-value).

In place of genomic DNA, assembled ESTs can be used as an anchor. The main requirement is that the anchor sequences be of high quality because they are driving the multiple sequence alignment (MSA). Although "best hit" criteria are used during EST-to-anchor

assignment, poor alignments or alignments between paralogs will be caught either during formation of MSAs by cross_match (see below) or by the internal paralog filter implemented within PolyBayes. The genome of B73 maize is currently being sequenced and this will provide an excellent collection of anchor sequences.

3. Run cross_match on each anchor sequence and its associated 454 ESTs to create an anchored MSA. The following cross_match parameters are recommended:

-discrep_lists

-tags

-masklevel 5

-gap_init -1

-gap_ext -1.

Low initiation (-gap_init) and gap extension (-gap_ext) are used to increase alignment tolerance between the short 454 ESTs and genomic anchors. Substitute higher values for gap_init and gap_ext if the anchored MSAs are unspliced (i.e., ESTs aligned to an EST anchor, or genomic sequence aligned to a genomic sequence anchor).

4. Run PolyBayes on the MSA. Recommended PolyBayes parameters for maize are:

-maskAmbiguousMatches

-nofilterParalogs

-priorParalog 0.03

-thresholdNative 0.75

-screenSnps

-considerAnchor

-noconsiderTemplateConsensus

-prescreenSnps

-priorPoly 0.01

-thresholdSnp 0.5.

It is necessary to include sequence quality files for the anchor sequence and the sequences aligned to it (member sequences). If these are unavailable or unreliable, set default quality values with:

-anchorBaseQualityDefault

-memberBaseQualityDefault.

Because cross_match aligns each sequence individually to the anchor during MSA construction, and PolyBayes assesses base quality on an individual basis, the use of a stringent default rather than the base quality information provided by 454 Life Sciences is expected to increase the accuracy of polymorphism detection.

5. Perform post-processing by reading the PolyBayes output files and deciding on appropriate rules to distinguish putative SNPs from false positives.

See Discussion.

DISCUSSION

In maize SNP-mining experiments conducted by the authors, both Mo17 and B73 454 ESTs were available, and the B73 maize MAGI assemblies were used as alignment anchors. Because Mo17 and B73 are inbreds, they should be

monoallelic at every base position, with relatively rare exceptions caused by nearly identical paralogs (NIPs). Hence, putative SNPs were filtered using rules designed to substantially decrease the rate of false positives. These rules were:

- Polymorphic sites require a minimum of 2X representation in the Mo17 454 ESTs.
- All Mo17 base calls at sites that are polymorphic between Mo17 454 ESTs and the B73 MAGI anchors are expected to be identical. This ensures monoallelism within the Mo17 454 ESTs.
- When B73 454 EST sequences also align across polymorphic sites that pass the first two rules, all of the B73 454 ESTs and the MAGI3.1 anchor base calls must agree. This avoids polymorphisms resulting from incorrect MAGI base calls or NIPs within B73.

REFERENCES

Altschul S.F., Gish W., Miller W., Myers E.W., Lipman D.J. (1990) Basic local alignment search tool. *J. Mol. Biol.* **215**:403–410. CrossRef Medline

Marth G.T., Korf I., Yandell M.D., Yeh R.T., Gu Z., Zakeri H., Stitziel N.O., Hillier L., Kwok P.Y., Gish W.R. (1999) A general approach to single-nucleotide polymorphism discovery. *Nat. Genet.* **23**:452–456. CrossRef Medline

Useche F.J., Gao G., Harafey M., Rafalski A. (2001) High-throughput identification, database storage and analysis of SNPs in EST sequences. *Genome Inform.* **12**:194–203. Medline

Related Protocols

Protocol
T7-Based RNA Amplification for Genotyping from Maize Shoot Apical Meristem
Kazuhiro Ohtsu and Patrick S. Schnable
Cold Spring Harb Protoc; 2007; doi:10.1101/pdb.prot4785
Abstract Full Text

Protocol
Maize Tissue Preparation and Extraction of RNA from Target Cells for Genotyping
Kazuhiro Ohtsu and Patrick S. Schnable
Cold Spring Harb Protoc; 2007; doi:10.1101/pdb.prot4784
Abstract Full Text