REANNOTATION DES MAIZE OLIGONUCLEOTIDE ARRAYS

Felix Seifert, Heike Pospisil

Zusammenfassung

Die Microarray-Technologie hat sich zu einem etablierten Ansatz der Hochdurchsatz-Genexpressionsanalyse entwickelt. Das „maize oligonucleotide array“ (maizearray) ist eine der wenigen Microarray-Plattformen, welche für die genomweite Genexpressionsanalyse von Mais (*Zea mays* L.) erzeugt wurden. Die Sonden wurden basierend auf ESTs (expressed sequence tags) generiert. Mittlerweile ist die Genomsequenz von Mais verfügbar und ermöglicht eine genauere Annotation dieser Sonden. In dieser Arbeit wurden die Genompositionen aller Sonden und basierend darauf die zugrunde liegenden Gene sowie deren funktionelle Annotation bestimmt. Durch die Analyse konnten Redundanzen und nicht eindeutig bindende Sonden aufgedeckt und gleichzeitig die Zahl der Gene mit funktioneller Annotation verdoppelt werden. Unsere Reannotation wird funktionelle Analysen bereits existierender und zukünftiger Datensätze stark verbessern.

Abstract

The microarray technology has become an established approach for large-scale gene expression analysis. The maize oligonucleotide array (maizearray) is one of the few microarray platforms designed for genome-wide gene expression analysis in *Zea mays* L. The microarray probes were compiled based on expressed sequence tags (ESTs). Meanwhile, the maize genome sequence became available providing the possibility for an improved annotation of the microarray probe set. In this study we determined the genome positions of all maizearray probes to obtain current gene annotations and funktional annotations. These new data allow tracing redundancy of the probe set and interfering cross-hybridizations, and will largely improve the functional analysis of available and future datasets generated on this microarray platform.

EINFÜHRUNG

DNA-Microarrays wurden seit ihrer Vorstellung im Jahr 1995 (Shena *et al.* 1995) zu einer ausgereiften Methode zur Genexpressionsanalyse entwickelt. DNA-Microarrays werden durch Synthetisierung bzw. Drucken und Fixieren von Oligonukleotiden auf einer Trägeroberfläche erzeugt. Diese DNA-Elemente dienen als Sonden für die Hybridisierung mit fluoreszenzmarkierten cDNAs oder RNAs aus komplexen Transkriptproben (Phimister 1999).

Das „maize oligonucleotide array“ (maizearray) ist eine Microarrayplattform bestehend aus langen Oligonukleotiden (~70 nt) (Gardiner *et al.* 2005). Das initiale 57K Array umfasst 57.452 Sonden auf zwei Objektträgern, eine Überarbeitung führte zum 46K Array mit 43.536 Oligonukleotiden auf einem Objektträger. Die Sondensequenzen wurden basierend auf ESTs, TIGR Assembled Zea mays (AZM), repetitiven Elementen sowie Sequenzen aus Chloroplasten und Mitochondrien erzeugt. Zum Zeitpunkt der Erstellung dieses Arrays existierte die Genomsequenz von Mais noch nicht. Durch das Fehlen des Bezugs zu der Genomsequenz kann nicht ausgeschlossen werden, dass nicht alle Microarray-Sonden die Expression eines einzigen Gens repräsentieren und somit möglicherweise durch multiple Genkopien, alternative Spliceformen etc. beeinflusst werden.

Wir haben eine Neuannotation der 57K und der auf dieser basierenden 46K maizearray Plattform durch Lokalisation der Sonden auf dem Maisgenom durchgeführt. Basierend auf der Lokalisation kann für jede Sonde ermittelt werden, ob sie die Expression eines einzelnen oder mehrerer Gene/Loci widerspiegelt. Anhand der ermittelten Zielgene für die Sonden wurden funktionelle Informationen in Form von gene ontology (GO) Termen (The Gene Ontology Consortium 2000) mit Blast2GO ermittelt (Conesa & Götz 2008). Trotz der Einstellung der Produktion des maizearray erwarten wir, dass unsere Annotation einen hohen Wert für die Auswertung aktueller Arbeiten bzw. bestehender Datensätze hat.

Das Ziel von Hochdurchsatz-Experimenten wie Microarrays ist es, Schlüsselgene für untersuchte biologische Prozesse zu finden. Das Gene Ontology Consortium liefert ein einheitliches, dynamisches und kontrolliertes Vokabular, das durch GO-Terme zu Genfunktion, Lokalisation und biologischen Prozess für alle eukaryotischen Arten repräsentiert wird (The Gene Ontology Consortium 2000). In vielen Microarray-Experimenten wird eine Teilmenge von Genen mit ähnlicher Expression anhand ihrer GO-Terme auf eine Überrepräsentierung bestimmter Funktionen untersucht. Unsere funktionelle Annotationen ermöglichen GO-Anreicherungsanalysen für einen größeren Teil der Sonden und ermöglichen das Ausschließen von Sonden, welche mehrere Gene/Loci repräsentieren und Redundanz durch mehrere Sonden auf einem Gen.

MATERIAL UND METHODEN

Die Oligonukleotidsequenzen des 57K maizearray, welches alle Sonden des 46K Arrays enthält, wurden mittels BLASTn (standalone BLAST 2.2.26+,

e-value 0.0001, word-length 20) (Camacho *et al.* 2009) auf der Genomsequenz der Maislinie B73 in der Version RefGen v2 (ftp.maizesequence.org) lokalisiert. Alle Übereinstimmungen mit insgesamt weniger als drei Fehlpaarungen, Insertionen bzw. Deletionen (Indels) wurden mit Genannotationen (Exon, Intron) aus dem „working gene set“ (WGS) version 5a.59 bzw. repetitiven Elemente des TE Consortium (ZmB73_5a_MTEC+LTR_repeats.gff, beide Datensätze von ftp.maizesequence.org) abgeglichen und annotiert. Für alle Oligonukleotide mit zwei partiellen Sequenzübereinstimmungen innerhalb von 20.000 Basenpaaren auf dem selben Strang oder Fragmenten mit der Länge des Oligonukleotides minus der BLASTn-Wortlänge wurde die Oligonukleotidsequenz mit BLASTn gegen Mais cDNA Sequenzen des WGS Datensatzes Version 5a.59 mit den wie bereits zuvor genannten Parametern aligniert. Dieser Ansatz soll Sonden auf dem Genom lokalisieren, welche durch ein Intron unterbrochen sind. Die Annotationsprozedur ist in **Abbildung 1** dargestellt.

Die funktionelle Annotation des maizearray wurde mit Blast2GO Version 2.5.1 (Conesa & Götz 2008) durchgeführt. Dabei wurden für die ermittelten Gensequenzen homologe Genprodukte mittels BLASTx in der NCBI „non-redundant protein sequences (nr) database“ bestimmt, um anschließend die wahrscheinlichsten GO-Terme anhand der homologen Sequenzen zu ermitteln. Die Blast2GO-Analyse wurde mit Standardparametern und der Blast2GO PRO Datenbank b2g_apr12 für alle identifizierten Gene mit Ausnahme der repetitiven Sequenzen durchgeführt.

ERGEBNISSE

Die Lokalisierung der Oligonukleotide erzielte für 84.86 % aller Oligonukleotide mindestens eine entsprechende Genomposition. Durch die anschließende Annotation konnte für 70,82 % der Sonden des 57K Array bzw. für 73,98 % des 46K Array eine Annotation gefunden werden. Die Ergebnisse sind in **Tabelle 1** dargestellt [Seifert et al. 2012].

Prozedur der Neuannotierung

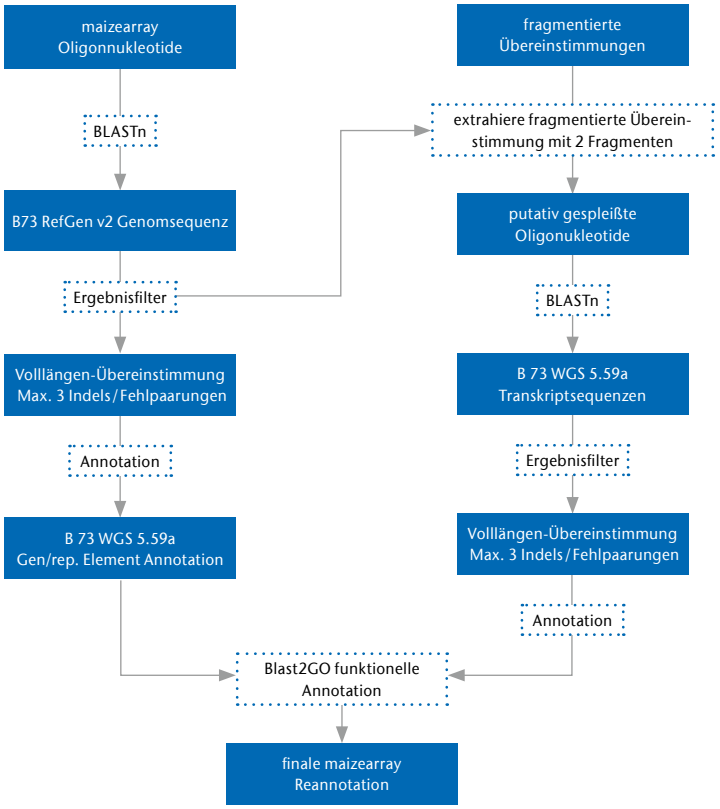

Abb. 1) Die maizearray Oligonukleotidsequenzen wurden mit BLASTn auf dem B73 Refgen v2 Maisgenome lokalisiert. Übereinstimmungen über die volle Länge wurde anhand der B73 WGS 5.59a Gen/repetitive Elemente-Annotation für beide DNA-Stränge annotiert. Fragmentierte Oligonukleotidübereinstimmungen wurden mit BLASTn auf die B73 WGS 5.59a Transkriptsequenzen gemappt und bei vollständiger Übereinstimmung über das Transkript annotiert. Die funktionelle Annotation erfolgt über Blast2GO.

Die Blast2GO-Annotation aller Oligonukleotide, welche einem oder mehreren Genen (Exon, Intron) zugeordnet werden konnten, resultierte in 47.562 annotierten Genen, welche von 32.745 (57,00 %) aller Sonden repräsentiert werden. Insgesamt konnten den annotierten Genen 238.700 GO-Terme zugewiesen werden. Somit entfallen 5,92 GO-Terme auf jedes annotierte Oligonukleotid [Seifert et al. 2012].

DISKUSSION

Die Lokalisierung der Oligonukleotide auf der B73 Genomsequenz resultierte mit fast 85 % in einer hohen Abdeckung des Genoms. Insgesamt konnte ein großer Anteil der Array-Sonden auf dem Mais-Referenzgenom lokalisiert werden. Dies zeigt, dass die Array-Plattform trotz Ermangelung der Genomsequenz bei dem Entwurf des

	57K maizearray	46K maizearray
unannotiert	16.760 (29.18%)	11.326 (26.02%)
anti-sense / Intron	3.811 (6.63%)	2.349 (5.39%)
mehrere Gene (inkl. repetitive Elemente)	8.473 (14.74%)	6.492 (14.91%)
einzelnes Gen, >1 Transkript	11.167 (19.44%)	9.715 (22.32%)
einzelnes Gen, 1 Transkript	17.241 (30.01%)	13.654 (31.36%)

Tab. 1) Annotationsergebnisse für Oligonukleotide beider maizearray Versionen

Microarrays einen hohen Informationsgehalt in den Expressionsdaten bietet.

Die offizielle Annotation enthielt 43.381 Gen-assoziierte Oligonukleotide und wies für nur 16.549 dieser Oligonukleotide insgesamt 113.584 GO-Terme auf (6,86 GO-Terme pro Sonde). Durch unsere Neuannotation wurde die Anzahl GO-annotierter, genassoziierter Oligonukleotide um fast den Faktor zwei erhöht, während die Anzahl der Terme pro Gen geringfügig gesunken ist.

Die Oligonukleotide, welche nicht lokalisiert werden konnten, stammen möglicherweise von ESTs anderer Maislinien bzw. Transgenen. Die Lokalisation eines Oligonukleotids an mehreren Loci weist auf Sequenzen hin, die durch Transposition oder Genomduplikation vervielfältigt wurden. Oligonukleotide, die auf dem Gegenstrang lokalisiert sind, entsprechen vermutlich falsch orientierten ESTs bzw. natürlichen anti-sense Transkripten (NATs) (Jin *et al.* 2008). Die Sonden, welche anti-sense zu repetitiven Elementen gefunden wurden, entsprechen wahrscheinlich Zwischenprodukten des RNAi-Mechanismus (Ito 2012).

Die durchgeführte Neuannotation des maizearray über eine Lokalisierung der Sonden auf dem B73 Maisgenom und Funktion der Gene basierend auf GO-Termen mittels Blast2GO erzielte eine hohe Abdeckung von annotierten Oligonukleotiden. Die neue Annotation ermöglicht den Ausschluss von Sonden, welche nicht einem einzigen Gen zugeordnet werden können bzw. mehrere Oligonukleotide identische Gene repräsentieren. Diese hinzugewonnenen Informationen erlauben spezifischere Auswertungen von Experimenten, die auf der Basis dieses Microarrays erzeugt wurden.

LITERATUR

Camacho, Ch., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K., Madden, T.L. (2009): BLAST+: architecture and applications. BMC Bioinformatics 10:421

Conesa, A., Götz, S. 2008: Blast2GO (2008): A Comprehensive Suite for Functional Analysis in Plant Genomics. Int. J. Plant Genomics 1-13

Gardiner, J.M., Buell, C.R., Elumalai, R., Galbraith, D.W., Henderson, D.A., Iniguez, A.L., Kaeppler, S.M., Kim, J.J., Liu, J., Smith, A., Zheng, L., Chandler, V.L. (2005): Design, Production, and Utilization of Long Oligonucleotide Microarrays for Expression Analysis in Maize. Maydica 50: 425-435

Ito, H. (2012): Small RNAs and transposon silencing in plants. Dev Growth Differ 54: 100-107

Jin, H., Vacic, V., Girke, Th., Lonardi, St., Zhu J.-K. (2008): Small RNAs and the regulation of cis-natural anti-sense transcripts in Arabidopsis. BMC Molecular Biol. 9: 6

Phimister, D. (1999): Going global. Nat. Genet. 21: 1

Schena, M., Shalon, D., Davis, R.W., Brown, P.O. (1995): Quantitative monitoring of gene expression patterns with a complementary DNA microarray. Science 270. 467-470

The Gene Ontology Consortium (2000): Gene ontology: tool for the unification of biology. Nat. Genet. 25(1): 25-29

Seifert, F., Thiemann, A., Pospisil, H., Scholten, S. (2012): Re-annotation of the maize oligonucleotide array. Maydica 57.1: 49-55

AUTOREN

Felix Seifert, Dipl.-Biol.,
TH Wildau, Bioinformatik / High Performance Computing in Life Sciences; Universität Hamburg, Biozentrum, Entwicklungsbiologie und Biotechnologie

Prof. Dr. rer. nat. Heike Pospisil,
TH Wildau, Bioinformatik / High Performance Computing in Life Sciences