Bioinformation by Biomedical Informatics Publishing Group

open access

www.bioinformation.net

Database

SNPinProbe_1.0: A database for filtering out probes in the Affymetrix GeneChip® Human Exon 1.0 ST array potentially affected by SNPs

Shiwei Duan[1], Wei Zhang[1], Wasim Kamel Bleibel[1], Nancy Jean Cox[1, 2] and M. Eileen Dolan[1, 3, 4, *]

[1]Section of Hematology/Oncology, Department of Medicine; [2]Department of Human Genetics; [3]Committee on Clinical Pharmacology and Pharmacogenomics; [4]Cancer Research Center, The University of Chicago, IL 60637, USA;
M. Eileen Dolan* - E-mail: edolan@medicine.bsd.uchicago.edu; Phone: 773 702 4441; Fax: 773 702 0963; * Corresponding Author

received June 18, 2008; revised July 20, 2008; accepted July 23, 2008; published August 01, 2008

Abstract:
The Affymetrix GeneChip® Human Exon 1.0 ST array (exon array) is designed to measure both gene-level and exon-level expression in human samples. This exon array contains ~1.4 million probesets consisting of ~5.4 million probes and profiles over 17,000 well-annotated gene transcripts in the human genome. As with all expression arrays, the exon array is vulnerable to SNPs within probes, because these SNPs can affect the hybridization of the probes and thus produce misleading expression values. In some cases, this could result in dramatic fluctuations of the exon-level expression. For this reason, we performed a genome-wide search for SNPs within regions that hybridize to probes by evaluating approximately 18 million SNPs in dbSNP (Build 129) and about 5.4 million probes in the exon array. We identified 597,068 probes within 350,382 probe sets that hybridized to regions containing SNPs. These affected probes and/or probesets can be filtered in the data processing procedure thus controlling for potential false expression phenotypes when using this exon array.

Keywords: database; probes; SNP; Affymetrix GeneChip® human exon 1.0 ST array; human genome

Availability: http://cid-fb2a64e541add2be.skydrive.live.com/browse.aspx/Affy%7C_HuEx%7C_1.0ST?uc=2.

Background:
Using high-throughput gene expression microarrays, thousands of genes are now able to be profiled in a single analysis. The Affymetrix GeneChip® Human Exon 1.0 ST array has been designed to detect novel exons, spliced exons or sub-exons of a gene in human samples **[1]**. The exon array uses over 5.4 million probes representing about 1.4 million probesets that are designed based on the genomic regions of known genes and regions that may harbor hypothetical genes. Compared with other arrays including the Affymetrix Genome Human Focus® array, U95® and U133® series array, the probes on the exon array are designed to cover the whole gene region instead of the 3′-untranslated regions **[1]**. Additionally, gene structures are represented by the probe sets with each probe set on the exon array consisting of up to 4 perfect match probes transcribed to a region of the exon. This is quite different from previous Affymetrix gene expression arrays that contain a set of perfect match and mismatch set of oligonucleotides tiled onto the microarray that account for nonspecific hybridization **[1, 2]**. However, studies have shown that SNPs within probes can affect hybridization of the 3′ expression arrays **[3]** as well as the exon arrays **[4-6]**. Given that there are 5.4 million probes on this human exon array, there are more probes hybridizing to regions containing SNPs and the effect can be dramatic when evaluating exon level expression. SNPs found in the probe-covered regions were shown to affect the hybridization efficiency of some probes and this can cause false relationships between the SNP genotypes and gene expression levels that are represented by the probes **[4-6]**. Furthermore, the hybridization difference of certain probes among individuals may not actually reflect the actual expression differences of the probe-representing regions but be due to the genotype differences of the common SNPs inside the hybridized sequences of the probes **[3-6]**. Quality control should include the identification of the probes containing SNPs in order to filter out the affected probes prior to expression analysis, thereby controlling the confounding effects that can be caused by these SNPs **[5, 7, 8]**.

Methodology:

Dataset

The dataset **[9]** contains the probes affected by the SNPs in their hybridization regions based on the dbSNP database (version 129, genome build 36, April, 2008) **[10]**.

Development

The genomic positions (build 36) of over 18 million SNPs were retrieved in the dbSNP database (version 129). The sequences of over 5.4 million probes and over 1.4 million

probe sets were downloaded at the Affymetrix website **[11]**. Since the probesets are given with the genomic regions (build 36), while the probes are still annotated with the old genomic regions (build 34), a local BLAT **[12]** between probes and their probesets were performed to update the probe covered genomic regions. Then, a genome-wide search process was performed between ~18 million SNPs and over 5.4 million probes to identify the probes affected by the probesets.

Database content

This database **[9]** provides 597,068 probes within 350,382 probesets affected by the known SNPs in dbSNP (version 129).

Database usage

The user can download the list of affected probes and probesets **[9]**, and then apply the list to filter out the affected probes using the program provided by the Affymetrix Power Tools (1.8.6) (Figure 1). This software is a free tool with the functionality to filter out a known set of probes. Removal of affected probes can be accomplished by using their highly experimental workflow through using the *apt-probeset-summarize* function together with the *--kill-list* function **[13]**. Resulting probeset intensities will be summarized solely on those probes not affected by SNPs. The generated expression data will be good for routine expression analysis.

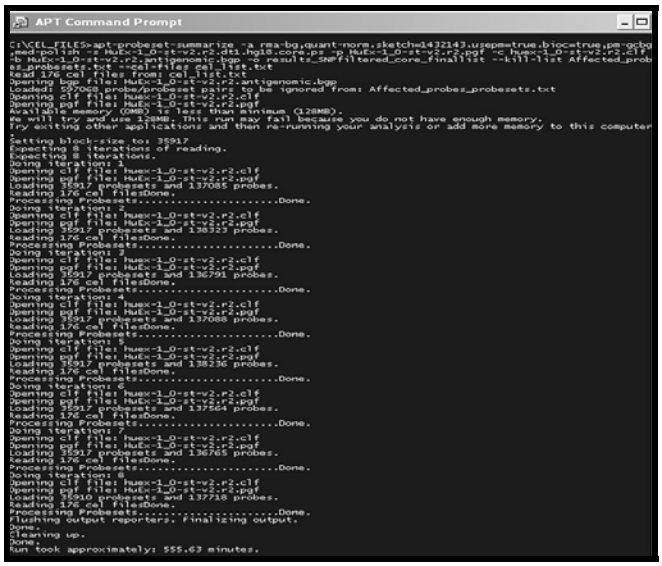

Figure 1: The process of filtering out the affected probes by SNPs inside.

Caveats:

There are 111,685 probes (2% of the total probes) that failed in the BLAT process possibly due to the fact that they are the background controls. We also include them in the database **[9]**.

Acknowledgment:

This Pharmacogenetics of Anticancer Agents Research (PAAR) Group http://pharmacogenetics.org) study was supported by the NIH/NIGMS grant U01GM61393.

References:

[01] http://www.affymetrix.com/support/technical/whitepapers/exon_probeset_trans_clust_whitepaper.pdf

[02] http://www.affymetrix.com/support/technical/technotes/25mer_technote.pdf

[03] E. Sliwerska, *et al.*, *Biol Psychiatry*, 61: 13 (2007) [PMID: 16690034]

[04] R. Alberts, *et al.*, *PLoS ONE*, 2: e622 (2007) [PMID: 17637838]

[05] W. Zhang, *et al.*, *Am J Hum Genet.*, 82: 631 (2008) [PMID: 18313023]

[06] A. Sequeira, *et al.*, *Mol Psychiatry*, 13: 363 (2008) [PMID: 18347597]

[07] W. Zhang and M. E. Dolan, *Bioinformation,* 2: 238 (2008) [PMID: 18317571]

[08] S. Duan, *et al.*, *Am J Hum Genet.*, 82: 1101 (2008) [PMID: 18439551]

[09] http://cid-fb2a64e541add2be.skydrive.live.com/browse.aspx/Affy|_HuEx|_1.0ST?uc=2

[10] S. T. Sherry, *et al., Nucleic Acids Res.*, 29: 308 (2001) [PMID: 11125122]

[11] http://www.affymetrix.com

[12] W. J. Kent, *Genome Res.*, 12: 656 (2002) [PMID: 11932250]

[13] http://www.affymetrix.com/support/developer/powertools/changelog/VIGNETTE-expression-mask-probes.html

Edited by P. Kangueane

Citation: Duan *et al.*, Bioinformation 2(10): 469-470 (2008)