Identification of non reliable probes on customized Affymetrix Mouse430_2 platform

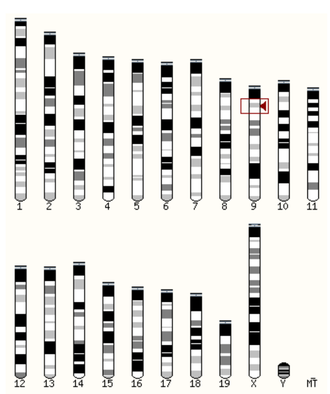
Noura Chelbat[1], Adetayo Kassim[2] Ulrich Bodenhofer[1], W.Talloen[3]
Sepp Hochreiter[1],, Ziv Shkedy[2],
[1]Institute of Bioinformatics, Johannes Kepler University, Linz, Austria
[2]Center for Statistics, Hasselt University, Diepenbeek, Belgium
[3]Johnson & Johnson Pharmaceutical Research & Development, a division of Janssen Pharmaceutica n.v., Beerse, Belgium

Motivation

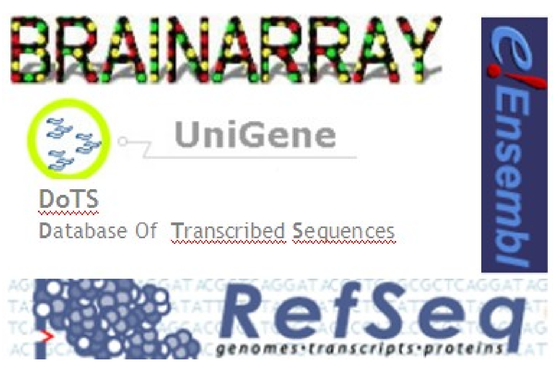
It is well known that Affymetrix Chip Definition Files (CDFs), contain wrongly annotated probes that lead to misinterpretation of the experiments results. Several methods to identify such groups of probes have been lately developed. One of them is founded on customized CDFs where probes map uniquely to genes contained in the EntrezGene database based on the latest genome and transcriptome released information.

Keywords: Gene Filtering, FARMS (Factor Analysis for Robust Microarray Summarization), I/NI calls (Informative/Non-Informative calls), SPC (Single Probe Constribution),LCMM (Latent Class Mixed Model),CDFs (Chip Definition Files)

Customized CDFs → Probesets are redefined to ensure each probe hits only one genomic location and all probes within the same probeset mapped to the same target transcript



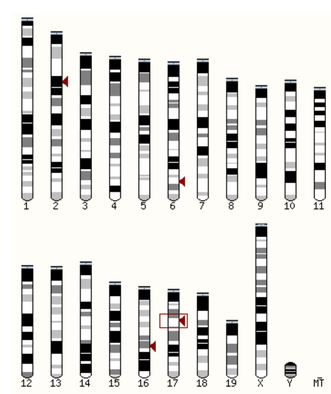
Blasting of the probeset 68743_at to the mouse genome



Mouse4302_Mm_ENTREZG_12

Pair	X	Y	Seq	Genome Location
1	311	881	TAATAATTTGAATGTAACCTTGATT	Mm 17 - 34133098
2	799	335	CATGACCCTCATTCTCTAGCGTGAA	Mm 17 - 34133416
3	285	869	TCATGACCCTCATTCTCTAGCGTGA	Mm 17 - 34133417
4	797	969	TTCATGACCCTCATTCTCTAGCGTG	Mm 17 - 34133418
5	437	603	GTTCATGACCCTCATTCTCTAGCGT	Mm 17 - 34133419
6	161	711	GGTTCATGACCCTCATTCTCTAGCG	Mm 17 - 34133420
7	162	711	GGTTCATGACCCTCATTCTCTAGCG	Mm 17 - 34133420
8	539	151	AGCTGCAATAGTCACTGGAGCTGTG	Mm 17 - 34134020
9	540	151	AGCTGCAATAGTCACTGGAGCTGTG	Mm 17 - 34134020
10	279	999	TTGGAGCTGCAATAGTCACTGGAGC	Mm 17 - 34134024
11	280	999	TTGGAGCTGCAATAGTCACTGGAGC	Mm 17 - 34134024
12	450	403	CCTTGGAGCTGCAATAGTCACTGGA	Mm 17 - 34134026
13	670	649	GTCCTTGGAGCTGCAATAGTCACTG	Mm 17 - 34134028
14	671	649	GTCCTTGGAGCTGCAATAGTCACTG	Mm 17 - 34134028
15	770	993	TTGTCCTTGGAGCTGCAATAGTCAC	Mm 17 - 34134030
16	771	993	TTGTCCTTGGAGCTGCAATAGTCAC	Mm 17 - 34134030

Probe content and genomic location in probeset 14972_at according to custom CDF version 12.1.0, entrezg



Example of blasting the Non reliable probe sequence "TAATAATTTGAATGTAACCTTGATT" to the mouse genome

Head arrows: multiple alignments and hits

Rectangular box: where the probeset maps

Method

Factor Analysis Based Method → I/NI and SPC

- → Informative non informative call and Single Probe Contribution [1]
- → Probe sets by where the majority of the probes are consistent in terms of intensity [1]
- → Filtering score: Computed value of signal to noise ratio for each probe as individual donation to its probeset
- → "non reliable-bad" probes those group of probes that fail to detect a signal confirmed by other probes

$$Var(z|x) = \sigma^2 = (\lambda\lambda^T)\Psi_{ij}^{-1}$$

$$SPC = (\lambda\lambda^T)\Psi_{ij}^{-1} / \sum (\lambda\lambda^T)\Psi_{ij}^{-1}$$

Highly correlated probes ↑↑λ and ↓↓Ψ

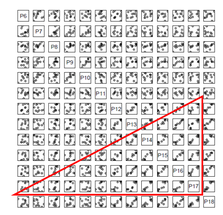
Probabilistic method → LCMM and ICC

- → Latent Class Mixed Model and Intra-Cluster Correlation
- → Probesets by where probes are grouped and array-array variability differing between such groups [2]
- → Filtering score: Intracluster correlation quantify average correlation between any pairs of observation in a probeset
- → "Informative Probeset" acording to ICC cut off

$$LCMM = \log_2(PM_{ij}) = \mu_j + Z_{jg}b_{ij} + \varepsilon_{ij}$$

$$ICC = \rho_g = \sigma^2_{bg} / \sigma^2_{bg} + \sigma^2_{\varepsilon}$$

Limitation Class to which probes belong to is Unknown



Results

The discrimination between non reliable-bad probes from reliable-good probes was computed through two alternative methods obtaining accuracies of wrongly performing probes in the range of 60–70% for both approaches

Materials

Experiments/Datasets: real-life data [3]

Genotype	description	Time and samples
Wild type Slc17A5 +	Total RNA Brain Native Sialin protein	18 day old mice/6x
Knockout Slc17A5 –	Total RNA Brain Mutated Sialin protein	18 day oldmice /6x

Platform and annotation file: customized CDFs

Affymetrix GeneChip®	Customized CDF	Samples	Probesets	Probes
MG-U74Av2	mouse4302mmgentrez	12	16 395	240 917

Reliable-Non Reliable Probes Definitions

	INI/SPC	LCMM/ICC
R+	>5xE-02	>5xE-01
R-	<5xE-03	<5xE-01

Training model selection

From Factor Analysis based method on SPC filteered probes the best predictors are selected [4]

K-mer	3	4	5
CV (10x)	61.0	62	60.0

Informative Probesets	2 318
Informative Probes	34144
SPC probes	20 298

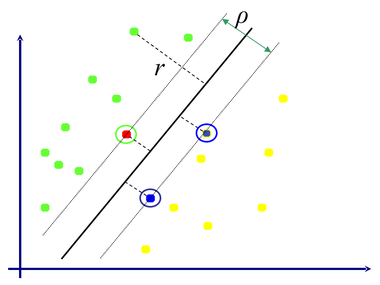
1. Training model on SPC
2. Explicit K-mer representation generation on IP
3. Generalization model based on Position Independent Kernel

Supervised Approach

Class labeling for binary classification task

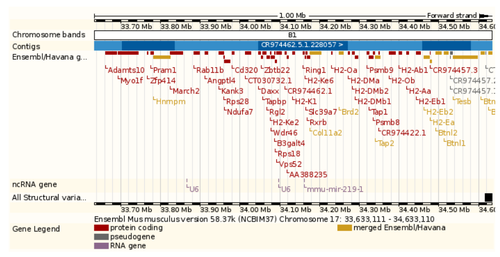
+1 Predicted Reliable probes

-1 Predicted Non Reliable probes



Runs	Accuracy
SPC_s3	64.5
SPC_s4	70.0
SPC_s5	61
ICC based	70.5

PIK counts the occurrences of k-length subsequences



The genomic location of probes in probeset 14972_at

Conclusions

- → Biological relevance on alternative splicing and multiple alignments were found
- → We demonstrated that even though wrongly annotated probes are removed from the curated CDFs still some probes on the arrays show different responses to a signal even if they are supposed to detect the same signal
- → "Outliers" will lead to noisy measurements and should be identified and removed
- → Improvements needed in customized annotation files for the better post processing and impact on the biological analysis

References

1. W.Talloen, D.-A.Clevert, S. Hochreiter, D. Amaratunga, L. Bijnens, S. Kass and H.W.H.Göhlmann: **I/NI-calls for the exclusion of non-informative genes: a highly effective filtering tool for microarray data.** Bioinformatics 2007; 23: 2897 – 2902
2. Nandini Raghavan, An M. I. M. De Bondt, Willem Talloen, Dieder Moechars, Hinrich W. H. Göhlmann and Dhammika Amaratunga: **The high-level similarity of some disparate gene expression measures.** Bioinformatics 2007; 22: 3032-3038
3. Noura Chelbat, Ulrich Bodenhofer, Sepp Hochreiter: **Filtering and Identifying non-reliable probes in Affymetrix GeneChip® platforms.** Poster at ISMB/ECCB, Stockholm, Sweden , June 27 - July 2 2009