oligoMask: A Framework for Assessing and Removing the Effect of Genetic Variants on Microarray Probes

by Daniel Bottomly, Beth Wilmot and Shannon K. McWeeney

Abstract As expression microarrays are typically designed relative to a reference genome, any individual genetic variant that overlaps a probe's genomic position can possibly cause a reduction in hybridization due to the probe no longer being a perfect match to a given sample's mRNA at that locus. If the samples or groups used in a microarray study differ in terms of genetic variants, the results of the microarray experiment can be negatively impacted. The **oligoMask** package is an R/SQLite framework which can utilize publicly available genetic variants and works in conjunction with the **oligo** package to read in the expression data and remove microarray probes which are likely to impact a given microarray experiment prior to analysis. Tools are provided for creating an SQLite database containing the probe and variant annotations and for performing the commonly used RMA preprocessing procedure for Affymetrix microarrays. The **oligoMask** package is freely available at https://github.com/dbottomly/oligoMask.

Introduction

It has been observed that for mRNA microarrays from a given sample, genetic differences of that sample relative to the probe sequences can affect hybridization to short oligonucleotide probes resulting in false positives or negatives depending on the experimental and array design (Walter et al., 2007; Alberts et al., 2007). Several approaches currently exist to identify and flag/remove probes that have hybridization artifacts due to genetic variants. Removal of probes based on pre-defined genetic variant databases is one such approach (Benovoy et al., 2008). Software (Kumari et al., 2007) and databases (Duan et al., 2008) allowing the interrogation of the relationships between microarray probes and single nucleotide variants have been described. The R package **CustomCDF** (Dai et al., 2005) is an example of this approach that removes probes from an environment formed from the Affymetrix chip description file (CDF) prior to analysis. One potential limitation of the CDF filtering approach is that for some of the more recent arrays platforms such as the Affymetrix Gene or Exon arrays the use of such environments has been superseded (e.g. the SQLite databases in the **oligo** (Carvalho and Irizarry, 2010) package or ROOT scheme files as in the **xps** (Stratowa et al., 2013) package).

In addition, the actual expression data itself can be interrogated to identify and mask out variants. R packages exist to effectively deal with a two group comparison between several strains or species through procedures based mainly on the expression data such as **maskBAD** (Dannemann et al., 2012) and **SNEP** (Fujisawa et al., 2009). However, models based on two (genetic) groups have limited utility when analyzing more complicated experimental designs such as those found in expression-based analyses using more genetically diverse mouse lines such as Diversity Outbred (Svenson et al., 2012), Collaborative Cross (Collaborative Cross Consortium, 2012) or other Heterogenous Stock (Chia et al., 2005) mice.

In order to facilitate eQTL mapping and other expression analyses in complex mouse crosses we devised an R package **oligoMask** based on the use of high quality publicly available genetic variant databases to screen microarray probes identifying probes impacted by variants. The key to this is the relatively recent availability of genome-wide variant databases in the variant call format (VCF) such as those from the Sanger Mouse Genomes Project (Keane et al., 2011) and the 1000 Genomes for humans (1000 Genomes Project Consortium, 2012) as well as the ability to query and parse these files via the **VariantAnnotation** (Obenchain et al., 2014) package. The **oligoMask** package is designed to work in conjunction with the **oligo** Bioconductor package to facilitate removal of aberrant probe expression prior to the commonly used robust multi-array average (RMA) (Irizarry et al., 2003) pre-processing procedure for Affymetrix arrays. Our package works by removing potentially impacted probes from the overall expression matrix prior to the call to the RMA processing functions. This can be done before the background correction step or after the normalization step. The annotation for these impacted probes are most easily derived from VCF files with the parsed data stored in an SQLite database. This database can be optionally wrapped in an R package with appropriate metadata to facilitate sharing and reproducibility. High-level S4 classes and methods provide a convenient interface with **oligoMask** and **oligo**. In addition, users can define new database schemas, add custom data as well as create their own functionality. Below we give an overview as well as demonstrate using publicly available data the steps involved for the use of **oligoMask**.

Example data

The example data presented in this article and in the vignette was downloaded from the gene expression omnibus (GEO) with accession number GSE33822 (Sun et al., 2012). For demonstration purposes we use a subset (n=8) of the dataset including only those samples derived from whole brain which received the vehicle treatment and that were run on version 1 of the Mouse Gene ST array. In our example, we are looking for expression differences between the NOD/ShiLtJ (NOD) inbred strain and the C57BL/6J (B6) inbred strain, the genome of which serves as the mouse reference genome. As we can expect the microarray probe sequences to be heavily biased towards the reference genome, looking for differential expression between these two strains may be problematic as differences in expression may be due to hybridization artifacts or true gene expression differences. First we create a NOD-specific database and then filter out those probes that are impacted by at least one variant in the NOD strain but not in the B6 strain and then carry out the differential expression analysis as per standard statistical workflows.

Workflow

Creating a variant database

The first step in the use of **oligoMask** is the creation of an SQLite database containing the probe annotation (including alignments to a given genome), variant annotation and the overlap,if any, between probes and variants. In a general sense, the probes sequences are first realigned to the given genome using the **BSgenome** and **Biostrings** Bioconductor packages (Pages, 2013; Pages et al., 2013) with the probe location and mappability of the probes being recorded. The locations of the uniquely mapping probes are then used to compute overlap with the variants in the specified VCF file using import and overlap functionality in the **VariantAnnotation** package. The locations of the variants in the genome, type of variant and the individual/population it was observed in is also recorded along with the overlap between probe alignments and variants. A convenience function for database creation is provided (`create.sanger.mouse.vcf.db`) for use with the case of variants derived from VCF files from the Sanger Mouse Genomes Project and variants of the Affymetrix Mouse Gene ST arrays. The **oligoMask** Vignette demonstrates in the section 'Data preparation' how the NOD variant database package (**om.NOD.mogene.1.0.st**) can be created using this function.

Additional array platforms and variant genotype file types can be supported through a modification of `create.sanger.mouse.vcf.db` as well as specifying the database schema as a `"TableSchemaList"` object as returned in the pre-defined `SangerTableSchemaList` function. The `"TableSchemaList"` S4 class serves a similar role as an object-relational mapping approach (ORM) in other languages and allows the R code to interact with a given database in a general way.

Masking procedure

The masking procedure first requires the installation of the **oligo** package along with the appropriate platform design databases that can be downloaded from Bioconductor. In our use case of Affymetrix Gene ST arrays, the CEL files are first read in using the `read.celfiles` function of **oligo** resulting in a `"GeneFeatureSet"` object. Next, the **oligoMask** database package is loaded, the parameters for the masking procedure are defined and finally the RMA summarization is performed as is shown below starting from the `"GeneFeatureSet"` object distributed with **oligoMaskData**.

```
library(oligoMask)
library(oligoMaskData)
library(om.NOD.mogene.1.0.st)
library(pd.mogene.1.0.st.v1)
library(limma)
data(oligoMaskData)

var.parms <- VariantMaskParams(om.NOD.mogene.1.0.st, geno.filter = FALSE,
  rm.unmap = FALSE, rm.mult = FALSE)

sun.gfs.mask <- maskRMA(oligoMaskData, target = "core", apply.mask = TRUE,
  mask.params = var.parms)
```

The result of these commands is a summarized `"GeneFeatureSet"` object with all probes overlapping variants from the NOD inbred strain of mouse removed prior to the background correction step of RMA. Users can control several aspects of the masking procedure through creation of a parameter object. For instance users can additionally remove probes that map to multiple locations as well as those that do not map at all to the reference genome by supplying TRUE to rm.multi and/or rm.unmap. Similarly, masking can be performed using only those variants that passed quality filters encoded in the VCF file by setting geno.filter to TRUE.

The maskRMA method carries out the RMA procedure and provides a similar interface to the rma method from **oligo**. In addition it requires specification of a "VariantMaskParams" object and whether the masking procedure should be performed before the background correction function or after background correction and normalization but before summarization by setting the mask.type argument to before.rma or before.summary respectively.

Assessment of masking procedure

As a demonstration of **oligoMask** next we perform a basic linear-model based differential expression analysis with the Sun *et al.* 2012 data comparing results with and without the NOD mask applied. Below we illustrate the basic approach using the masked data.

```
sun.exprs.mask <- exprs(sun.gfs.mask)
phen.dta <- data.frame(t(sapply(strsplit(colnames(sun.exprs.mask), "_"), c))[, 1:3])
names(phen.dta) <- c("tissue", "strain", "exposure")
use.mod <- model.matrix(~strain, data = phen.dta)
fit <- lmFit(sun.exprs.mask, use.mod)
fit <- eBayes(fit)
sun.exprs.mask.res <- decideTests(fit)
```

We then repeat this procedure but this time setting apply.mask = FALSE in maskRMA to provide the baseline standard RMA values for the comparison.

```
sun.gfs.unmask <- maskRMA(oligoMaskData, target = "core", apply.mask = FALSE,
         mask.params = var.parms)

sun.exprs.unmask <- exprs(sun.gfs.unmask)
um.phen.dta <-
    data.frame(t(sapply(strsplit(colnames(sun.exprs.unmask), "_"), c))[, 1:3])
names(um.phen.dta) <- c("tissue" , "strain" , "exposure")
um.mod <- model.matrix(~strain , data = um.phen.dta)
um.fit <- lmFit(sun.exprs.unmask , um.mod)
um.fit <- eBayes(um.fit)
sun.exprs.unmask.res <- decideTests(um.fit)
```

Finally we produce a summary venn diagram of the results.

```
comb.mat <- cbind(sun.exprs.unmask.res[ , "strainNOD", drop = FALSE], Masked = 0)
comb.mat[rownames(sun.exprs.mask.res), "Masked"] <-
  sun.exprs.mask.res[, "strainNOD"]
colnames(comb.mat)[1] <- "UnMasked"
vennDiagram(vennCounts(comb.mat))
```

The Venn diagram provides a visual comparison between the masked and unmasked results, allowing the user to assess impact of variants on expression (Figure 1). An executable version of this code is provided in the accompanying vignette accessed by:

```
Stangle(system.file("doc/oligoMask.Rnw" , package = "oligoMask"))
```

Conclusion

oligoMask is a flexible R/SQLite framework for pre-processing and QA/QC of hybridization based expression data. Not only can it remove the effect of spurious probe intensities due to genetic variants but it can additionally correct for design artifacts (probes mapping to multiple places or not mapping at all). It utilizes SQLite and works in conjunction with the **oligo** Bioconductor package. Currently it supports the Affymetrix Mouse Gene ST array though support could be easily added for other array types or species as described above. Approaches for removal of probes based off of the variant and mapping information in the SQLite database are already implemented. More sophisticated algorithms could be built on top of the database to provide masking additionally based off of the

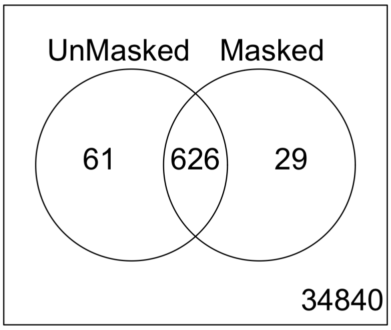

Figure 1: Concordance of differentially expressed genes between the masked and unmasked versions of the analysis

expression data itself or inferred haplotypes. We are currently working on allowing the masking to be further controlled by enforcing positional constraints on the variants relative to the probes as well as enabling masking to be based solely on the mapping information. Source code is freely available to all users from `https://github.com/dbottomly/oligoMask`. Note that **om.NOD.mogene.1.0.st** and **oligoMaskData** are available as part of release 0.99.08 on github (`https://github.com/dbottomly/oligoMask/releases`).

Acknowledgements

We thank the two anonymous reviewers for their comments. This work was supported in part by grants from the National Institute of Allergy and Infectious Disease (1U19AI100625), the National Center for Advancing Translational Sciences (5UL1RR024140) and the National Cancer Institute (5P30CA069533-13).

Bibliography

1000 Genomes Project Consortium. An Integrated Map of Genetic Variation From 1,092 Human Genomes. *Nature*, 491:1, 2012. [p159]

R. Alberts, P. Terpstra, Y. Li, R. Breitling, J.-P. Nap, and R. C. Jansen. Sequence Polymorphisms Cause Many False Cis eQTLs. *PloS ONE*, 2(7):e622, 2007. [p159]

D. Benovoy, T. Kwan, and J. Majewski. Effect of Polymorphisms Within Probe-Target Sequences on Olignonucleotide Microarray Experiments. *Nucleic Acids Research*, 36(13):4417–4423, 2008. doi: 10.1093/nar/gkn409. URL `http://nar.oxfordjournals.org/content/36/13/4417.abstract`. [p159]

B. S. Carvalho and R. A. Irizarry. A Framework for Oligonucleotide Microarray Preprocessing. *Bioinformatics*, 2010. ISSN 1367-4803. doi: http://dx.doi.org/10.1093/bioinformatics/btq431. [p159]

R. Chia, F. Achilli, M. F. Festing, and E. M. Fisher. The Origins and Uses of Mouse Outbred Stocks. *Nature Genetics*, 37(11):1181–1186, 2005. [p159]

Collaborative Cross Consortium. The Genome Architecture of the Collaborative Cross Mouse Genetic Reference Population. *Genetics*, 190:389–401, 2012. [p159]

M. Dai, P. Wang, A. D. Boyd, G. Kostov, B. Athey, E. G. Jones, W. E. Bunney, R. M. Myers, T. P. Speed, H. Akil, et al. Evolving Gene/Transcript Definitions Significantly Alter the Interpretation of GeneChip Data. *Nucleic Acids Research*, 33(20):e175–e175, 2005. [p159]

M. Dannemann, M. Lachmann, and A. Lorenc. 'maskBAD' - A Package to Detect and Remove Affymetrix Probes With Binding Affinity Differences. *BMC Bioinformatics*, 13(1):56, 2012. ISSN 1471-2105. doi: 10.1186/1471-2105-13-56. URL `http://www.biomedcentral.com/1471-2105/13/56`. [p159]

S. Duan, W. Zhang, W. K. Bleibel, N. J. Cox, and M. E. Dolan. SNPinProbe_1. 0: A Database for Filtering Out Probes in the Affymetrix GeneChip® Human Exon 1.0 ST Array Potentially Affected by SNPs. *Bioinformation*, 2(10):469, 2008. [p159]

H. Fujisawa, Y. Horiuchi, Y. Harushima, T. Takada, S. Eguchi, T. Mochizuki, T. Sakaguchi, T. Shiroishi, and N. Kurata. SNEP: Simultaneous Detection of Nucleotide and Expression Polymorphisms Using Affymetrix GeneChip. *BMC Bioinformatics*, 10(1):131, 2009. ISSN 1471-2105. doi: 10.1186/1471-2105-10-131. URL http://www.biomedcentral.com/1471-2105/10/131. [p159]

R. A. Irizarry, B. Hobbs, F. Collin, Y. D. Beazer-Barclay, K. J. Antonellis, U. Scherf, and T. P. Speed. Exploration, Normalization, and Summaries of High Density Oligonucleotide Array Probe Level Data. *Biostatistics*, 4(2):249–264, 2003. [p159]

T. M. Keane, L. Goodstadt, P. Danecek, M. A. White, K. Wong, B. Yalcin, A. Heger, A. Agam, G. Slater, M. Goodson, et al. Mouse Genomic Variation and Its Effect on Phenotypes and Gene Regulation. *Nature*, 477(7364):289–294, 2011. [p159]

S. Kumari, L. Verma, and J. Weller. AffyMAPSDetector: A Software Tool to Characterize Affymetrix GeneChip® Expression Arrays With Respect to SNPs. *BMC Bioinformatics*, 8(1):276, 2007. ISSN 1471-2105. doi: 10.1186/1471-2105-8-276. URL http://www.biomedcentral.com/1471-2105/8/276. [p159]

V. Obenchain, M. Lawrence, V. Carey, S. Gogarten, P. Shannon, and M. Morgan. VariantAnnotation: A Bioconductor Package for Exploration and Annotation of Genetic Variants. *Bioinformatics*, 2014. doi: 10.1093/bioinformatics/btu168. URL http://bioinformatics.oxfordjournals.org/content/early/2014/03/28/bioinformatics.btu168.abstract. [p159]

H. Pages. *BSgenome: Infrastructure for Biostrings-Based Genome Data Packages*, 2013. R package version 1.30.0. [p160]

H. Pages, P. Aboyoun, R. Gentleman, and S. DebRoy. *Biostrings: String Objects Representing Biological Sequences, and Matching Algorithms*, 2013. R package version 2.30.0. [p160]

C. Stratowa, Vienna, and Austria. *xps: Processing and Analysis of Affymetrix Oligonucleotide Arrays including Exon Arrays, Whole Genome Arrays and Plate Arrays*, 2013. [p159]

W. Sun, S. Lee, V. Zhabotynsky, F. Zou, F. A. Wright, J. J. Crowley, Z. Yun, R. J. Buus, D. R. Miller, J. Wang, et al. Transcriptome Atlases of Mouse Brain Reveals Differential Expression Across Brain Regions and Genetic Backgrounds. *G3: Genes| Genomes| Genetics*, 2(2):203–211, 2012. [p160]

K. L. Svenson, D. M. Gatti, W. Valdar, C. E. Welsh, R. Cheng, E. J. Chesler, A. A. Palmer, L. McMillan, and G. A. Churchill. High-Resolution Genetic Mapping Using the Mouse Diversity Outbred Population. *Genetics*, 190(2):437–447, 2012. [p159]

N. A. Walter, S. K. McWeeney, S. T. Peters, J. K. Belknap, R. Hitzemann, and K. J. Buck. SNPs Matter: Impact on Detection of Differential Expression. *Nature Methods*, 4(9):679–680, 2007. [p159]

Daniel Bottomly
Oregon Clinical and Translational Research Institute
Oregon Health and Science University
3181 SW Sam Jackson Park Rd.
Portland, Oregon 97239
USA bottomly@ohsu.edu

Beth Wilmot
Oregon Clinical and Translational Research Institute
Division of Bioinformatics and Computational Biology, DMICE
3181 SW Sam Jackson Park Rd.
Portland, Oregon 97239
USA wilmotb@ohsu.edu

Shannon K. McWeeney
Oregon Clinical and Translational Research Institute
Knight Cancer Institute
Division of Bioinformatics and Computational Biology, DMICE
Division of Biostatistics, PHPM
3181 SW Sam Jackson Park Rd.
Portland, Oregon 97239
USA mcweeney@ohsu.edu