intcomp: Benchmarking pipeline for integrative cancer gene prioritization algorithms based on gene expression and copy number data

Leo Lahti*and Martin Schäfer

November 15, 2011

1 Introduction

Several algorithms have been suggested to integrate gene expression and DNA copy number measurement to discover cancer-associated chromosomal regions, but quantitative comparison of these models has been missing. The *intcomp* R package provides a benchmarking pipeline for quantitative comparisons between the different implementations for integrative analysis of ge/cn data. This vignette provides installation instructions, practical examples and references to the algorithmic details of the *intcomp* benchmarking pipeline [1].

In the *intcomp* pipeline, the cancer gene detection performance of each algorithm is evaluated based on gene prioritization by using each method. Each method is used to order the gene list, and the resulting order is compared to golden standard lists of known cancer genes on simulated and real data sets. For details, see [1].

2 Installation

2.1 Installing the intcomp benchmarking pipeline

To install this package directly within R type:

```
> install.packages("intcomp", type = "source", repos = "http://R-Forge.R-project.org",
+     dependencies = TRUE)
```

In case of error messages, see below.

2.2 Dependencies

You may need to install dependencies before the intcomp package can be installed. The benchmarking pipeline depends on various external R packages. Install the dependencies from within R using the following commands:

*leo.lahti@iki.fi

```
> source("http://www.bioconductor.org/biocLite.R")
> biocLite(c("biomaRt", "DNAcopy", "DRI", "edira", "hgu133ahsentrezg.db",
+     "intCNGEan", "org.Hs.eg.db", "PMA", "SIM", "PREDA", "CGHcall",
+     "CNAmet"))
```

You may need to install the following packages manually: curl library[1] and the R packages XML[2], RCurl[3], edira[4], intCNGEan[5], org.Hs.eg.db[6], PREDA/SODEGIR[7], and CNAmet[8].

3 Benchmarking the comparison methods

The package contains a copy of the publicly available cancer data sets from [12] and [13] to benchmark the cancer gene detection algorithms on real experimental data, and two simulated data sets from previous publications [10, 5], called Ferrari and Schaefer data sets, respectively. The data sets have varying setups, depending on whether they include two-group comparisons or segmented/called copy number data. Showcases running the benchmarking pipeline on each data set are described below.

3.1 Hyman et al. (2002)

The Hyman et al. (2002) [12] breast cancer data set[9], and a golden standard list of known breast cancer genes from The Breast Cancer Gene Database [14] provide the first example data set for benchmarking the comparison algorithms. The cancer gene list was downloaded[10] and stored to the tgdb object. The gene symbols are converted into Entrez Gene IDs, the probes are matched between gene expression and copy number data, as detailed in the *read.hyman* function, and the known breast cancer genes from the TGDB golden standard list present in the ge/cn data are selected. Further details are detailed in the *read.hyman* and *the get.brca.genes* functions.

For Hyman, the original non-segmented data set from the publication is used (cn.seg = cn.raw) in the experiments (except with intCNGEan and CNAmet that require segmented and called data, respectively). To run the intcomp benchmarking pipeline on Hyman data set, use

```
> methods <- c("CNAmet", "edira")
> library(intcomp)
> data(hyman)
> library("org.Hs.eg.db")
```

[1] http://curl.haxx.se/download.html
[2] http://cran.r-project.org/web/packages/XML/index.html
[3] http://www.omegahat.org/RCurl/
[4] http://www.statistik.tu-dortmund.de/ schaefer/
[5] http://www.few.vu.nl/ wvanwie/software/intCNGEan/intCNGEan.html
[6] http://www.bioconductor.org/packages/release/data/annotation/html/org.Hs.eg.db.html
[7] http://www.bioconductor.org/packages/devel/bioc/html/PREDA.html
[8] http://csbi.ltdk.helsinki.fi/CNAmet/
[9] HymancdnaDataA.tab, HymancghDataA.tab and HymanAcc.mat obtained from http://www.ece.ucsb.edu/pubs/ieee/index.shtml accessed June 2, 2010.
[10] http://www.tumor-gene.org/cgi-bin/TGDB/tgdb_by_name.cgi accessed 5.6.2010; 'tgdb_by_name.cgi.html' and 'tgdb.txt'

```
> symbol2entrezid <- as.list(org.Hs.egALIAS2EG)
> hyman <- read.hyman(cdna, cgh, genenames, xx = symbol2entrezid)
> data(tgdb)
> cancerGenes <- get.brca.genes(rownames(hyman$ge$data), symbol2entrezid,
+     tgdb)
> res.hyman <- test.geneorder.pipeline(ge = hyman$ge, cn.raw = hyman$cn.raw,
+     cghCall = hyman$cghCall, cancerGenes = cancerGenes, methods = methods,
+     cn.default = "raw", references = "none")
> auc.ordered <- sort(unlist(res.hyman$auc))
```

3.2 Pollack et al. (2002)

The Pollack et al. (2002) [13] data set[11] is also used in combination with the golden standard list from the TGDB (See Hyman data set). The gene identifiers in the Pollack data are converted into Entrez Gene IDs. To run the benchmarking tests on Pollack data set, use

```
> methods <- c("CNAmet", "edira")
> library(intcomp)
> data(pollack)
> pollack <- read.pollack(dat = CopyNoGeneDataset4719, clone2geneid = clone2geneid)
> library("org.Hs.eg.db")
> data(tgdb)
> cancerGenes <- get.brca.genes(rownames(pollack$ge$data), as.list(org.Hs.egALIAS2EG),
+     tgdb)
> res.pollack <- test.geneorder.pipeline(ge = pollack$ge, cn.raw = pollack$cn.raw,
+     cghCall = pollack$cghCall, cancerGenes = cancerGenes, methods = methods,
+     cn.default = "raw", references = "none")
> auc.ordered <- sort(unlist(res.pollack$auc))
```

3.3 Ferrari data set (2009)

The first simulated data set, where the exact ground truth is known, is provided by the simulation approach given in [10]:

```
> library(intcomp)
> ferrari <- test.simulation(GE, CN, method = "ferrari")
> res.ferrari <- test.geneorder.pipeline(ge = ferrari$ge, cn.raw = ferrari$cn.raw,
+     cn.seg = ferrari$cn.seg, cn.call = ferrari$cn.call, cghCall = ferrari$cn.cghCall,
+     cancerGenes = ferrari$cancerGenes, methods = methods)
> auc.ordered <- sort(unlist(res.ferrari$auc))
```

3.4 Schaefer data set (2009)

The second simulated data set is provided by the simulation approach given in [5] with added flexibility. The quantile grid to be simulated can be defined by the user, as well as the mixing weight, the number of different variances to be considered and the call probabilities.

[11] http://www.pnas.org/content/suppl/2002/09/23/162471999.DC1/4719CopyNoGeneDatsetLegend.html accessed June 2, 2010.

```
> methods <- c("CNAmet", "edira")
> library(intcomp)
> library(ediraAMLdata)
> data(AMLdata, package = "ediraAMLdata")
> schaefer <- test.simulation(GE, CN, method = "schaefer")
> res.schaefer <- test.geneorder.pipeline(ge = schaefer$ge, cn.raw = schaefer$cn.raw,
+     cghCall = schaefer$cn.cghCall, cancerGenes = schaefer$cancerGenes,
+     methods = methods, callprobs = schaefer$callprobs, cn.default = "raw")
> auc.ordered <- sort(unlist(res.schaefer$auc))
```

4 Notes on the benchmarking pipeline

The minimal input data for the test.geneorder.pipeline banchmarking function includes (i) gene expression data (ge), (ii) gene copy number data (cn.raw / cn.seg / cn.call / cghCall), (iii) a golden standard list of known cancer genes (cancerGenes), and (iv) the list of methods to compare (mehods).

The gene expression and copy number data sets are lists containing *data* and $info$ fields; the probes in gene expression and gene copy number need to be matched; *data* is a data matrix with gene expression (GE$data) or gene copy number (CN$data) data; $info$ field is a data frame containing additional information about genes: *loc* indicates the genomic location of the probes in base pairs (numeric); *chr* and *arm* are factors indicating the chromosome and chromosomal arm of the probe, respectively. The user can provide the copy number data as raw (cn.raw), segmented (cn.seg) or called (cn.call) version. Certain methods require specific versions of the copy number. For instance, the CNAmet requires called copy number data. The intCNGEan algorithm requires copy number as a cghCall object from the *CGHcall* R package. It is advisable to provide all four versions - cn.raw, cn.seg, cn.call and cghCall - in the input to the test.geneorder.pipeline function when possible. The cn.raw, cn.seg and cn.call should follow the data + info format explained above, and the cghCall contains the raw, segmented and called data in the cghCall format. Finally, if multiple versions of copy number data are available, the user can specify (through the cn.default argument) which version is coupled with gene expression data unless otherwise specified by particular methods. By default, the associations between gene expression and segmented copy number data (ge + cn.seg) are investigated.

5 Comparison methods

The following *implementations* are available in the *intcomp* benchmarking pipeline: *CNAmet* [2, 3], variants of *DRI* [4], *edira* [5], *intcngean* [6], *OrtizEstevez* [7], *pint* [8], variants of *SIM* [11], *PMA* [9], *PREDA/SODEGIR* [10, 15]. The list of available methods in the pipeline is retrieved with:

```
> library(intcomp)
```

```
> list.methods()
```

```
 [1] "edira"        "DRI.cp"       "DRI.cs"       "DRI.ct"       "SIM.full"
 [6] "SIM.window"   "CNAmet"       "intcngean"    "PMA.raw"      "pint"
[11] "OrtizEstevez" "PREDA"
```

6 Benchmarking results

The prioritized cancer gene list provided by each method is compared to the golden standard list of known cancer genes; the result contains running times of the algorithms and the AUC values from ROC analysis. The AUC values provide quantitative estimates of model performance in cancer gene detection and provide the basis for the comparisons as reported in [1].

6.1 Version details

The following package versions were used to produce this vignette:

```
> sessionInfo()

R version 2.13.0 (2011-04-13)
Platform: x86_64-unknown-linux-gnu (64-bit)

locale:
 [1] LC_CTYPE=en_US.UTF-8       LC_NUMERIC=C
 [3] LC_TIME=en_US.UTF-8        LC_COLLATE=en_US.UTF-8
 [5] LC_MONETARY=C              LC_MESSAGES=en_US.UTF-8
 [7] LC_PAPER=fi_FI.UTF-8       LC_NAME=C
 [9] LC_ADDRESS=C               LC_TELEPHONE=C
[11] LC_MEASUREMENT=en_US.UTF-8 LC_IDENTIFICATION=C

attached base packages:
[1] compiler  stats     graphics  grDevices utils     datasets  methods
[8] base

other attached packages:
 [1] intcomp_0.3.27       intCNGEan_0.53       PREDA_0.99.1
 [4] annotate_1.30.0      multtest_2.8.0       lokern_1.1-2
 [7] sfsmisc_1.0-16       ediraAMLdata_1.0.4   CNAmet_1.1
[10] CGHcall_2.12.0       CGHbase_1.10.0       marray_1.30.0
[13] limma_3.8.2          SIM_1.20.0           quantreg_4.71
[16] SparseM_0.89         PMA_1.0.8            plyr_1.5.2
[19] pint_1.5.34          dmt_0.8.06           MASS_7.3-12
[22] Matrix_0.999375-50   lattice_0.19-23      mvtnorm_0.9-9991
[25] org.Hs.eg.db_2.5.0   RSQLite_0.9-4        DBI_0.2-5
[28] AnnotationDbi_1.14.1 edira_1.1.3          DRI_1.1
[31] cghFLasso_0.2-1      impute_1.26.0        DNAcopy_1.26.0
[34] biomaRt_2.8.1        affy_1.30.0          Biobase_2.12.1

loaded via a namespace (and not attached):
 [1] affyio_1.20.0          globaltest_5.6.1      grid_2.13.0
 [4] preprocessCore_1.14.0 quantsmooth_1.18.0    RCurl_1.6-6
 [7] splines_2.13.0         survival_2.36-5       XML_3.4-3
[10] xtable_1.5-6
```

Acknowledgements

This work has been supported by EuGESMA COST Action BM0801: European Genetic and Epigenetic Study on AML and MDS. We would also like to thank Fransesco Ferrari for providing simulated data for the study.

References

[1] L. Lahti, M. Schäfer, H.-U. Klein, S. Bicciato, and M. Dugas, "Cancer gene prioritization by multi-platform data integration: a comparative review. upcoming,"

[2] S. Hautaniemi, M. Ringnér, P. Kauraniemi, R. Autio, H. Edgren, O. Yli-Harja, J. Astola, A. Kallioniemi, and O. Kallioniemi, "A strategy for identifying putative causes of gene expression variation in human cancers," *Journal of the Franklin Institute*, vol. 341, no. 1-2, pp. 77–88, 2004.

[3] R. Louhimo and S. Hautaniemi, "CNAmet: an R package for integrating copy number, methylation and expression data," *Bioinformatics*, vol. 27, no. 6, pp. 887–888, 2011.

[4] K. Salari, R. Tibshirani, and J. R. Pollack, "DR-Integrator: a new analytic tool for integrating DNA copy number and gene expression data.," *Bioinformatics*, vol. 26, no. 3, pp. 414–6, 2010.

[5] M. Schäfer, H. Schwender, S. Merk, C. Haferlach, K. Ickstadt, and M. Dugas, "Integrated analysis of copy number alterations and gene expression: a bivariate assessment of equally directed abnormalities," *Bioinformatics*, vol. 25, no. 24, pp. 3228–3235, 2009.

[6] W. N. van Wieringen and M. A. van de Wiel, "Nonparametric testing for DNA copy number induced differential mRNA gene expression," *Biometrics*, vol. 65, pp. 19–29, 2009.

[7] M. Ortiz-Estevez, J. De Las Rivas, C. Fontanillo, and A. Rubio, "Segmentation of genomic and transcriptomic microarrays data reveals major correlation between DNA copy number aberrations and gene-loci expression," *Genomics*, vol. 97, pp. 86–93, 2011.

[8] L. Lahti, S. Myllykangas, S. Knuutila, and S. Kaski, "Dependency detection with similarity constraints," in *Proceedings MLSP'09 IEEE International Workshop on Machine Learning for Signal Processing XIX*, (Piscataway, NJ), pp. 89–94, IEEE Signal Processing Society, September 2-4 2009. Implementation available in pint package of R/BioConductor http://www.bioconductor.org/packages/release/bioc/html/pint.html.

[9] D. M. Witten, R. Tibshirani, and T. Hastie, "A penalized matrix decomposition, with applications to sparse principal components and canonical correlation analysis," *Biostatistics*, vol. 10, no. 3, pp. 515–534, 2009.

[10] S. Bicciato, R. Spinelli, M. Zampieri, E. Mangano, F. Ferrari, L. Beltrame, I. Cifola, C. Peano, A. Solari, and C. Battaglia, "A computational procedure to identify significant overlap of differentially expressed and genomic

imbalanced regions in cancer datasets," *Nucleic Acids Research*, vol. 37, pp. 5057–5070, 2009.

[11] R. X. Menezes, M. Boetzer, M. Sieswerda, G.-J. B. van Ommen, and J. M. Boer, "Integrated analysis of DNA copy number and gene expression microarray data using gene sets.," *BMC bioinformatics*, vol. 10, no. 1, p. 203, 2009.

[12] E. Hyman, P. Kauraniemi, S. Hautaniemi, M. Wolf, S. Mousses, E. Rozenblum, M. Ringner, G. Sauter, O. Monni, A. Elkahloun, O.-P. Kallioniemi, and A. Kallioniemi, "Impact of DNA Amplification on Gene Expression Patterns in Breast Cancer," *Cancer Res.*, vol. 62, pp. 6240–6245, Nov. 2002.

[13] J. R. Pollack, T. Sø rlie, C. M. Perou, C. A. Rees, S. S. Jeffrey, P. E. Lonning, R. Tibshirani, D. Botstein, A.-L. Bø rresen Dale, and P. O. Brown, "Microarray analysis reveals a major direct role of DNA copy number alteration in the transcriptional program of human breast tumors.," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 99, pp. 12963–8, Oct. 2002.

[14] R. Baasiri, S. Glasser, D. Steffen, and D. Wheeler, "The Breast Cancer Gene Database: a collaborative information resource," *Oncogene*, vol. 18, no. 56, pp. 7958–7965, 1999.

[15] Ferrari F, Solari A, Battaglia C, Bicciato S. PREDA: an R-package to identify regional variations in genomic data. *Bioinformatics* online July 7, 2011.