Prediction of regulatory targets of alternative isoforms of the epidermal growth factor receptor in a glioblastoma cell line

Weinholdt Claus[1], Wichmann Henri[2], Kotrba Johanna[3], Ardell David H.[4], Kappler Matthias[2], Eckert Alexander W.[2], Vordermark Dirk[5], and Grosse Ivo[1,6]

[1] Institute of Computer Science, Martin-Luther-University Halle-Wittenberg, Germany; [2] Department of Oral and Maxillofacial Plastic Surgery, Martin Luther University Halle-Wittenberg, Germany; [3] Institute for Molecular and Clinical Immunology, Otto von Guericke University, Germany; [4] Molecular Cell Biology, School of Natural Sciences, University of California, Merced, USA; [5] Department of Radiotherapy, Martin Luther University Halle-Wittenberg, Germany; [6] German Centre for Integrative Biodiversity Research (iDiv) Halle-Jena-Leipzig, Germany;

Abstract

The epidermal growth factor receptor (EGFR) is a major regulator of proliferation in tumour cells. There are at least four isoforms of EGFR in humans, named I through IV, where isoform I is the full-length membrane protein, while isoforms II-IV are shorter protein isoforms. All isoforms are capable of binding the ligand epidermal growth factor (EGF), but the function and biomarker potential of the three soluble EGFR isoforms II-IV is not well investigated.

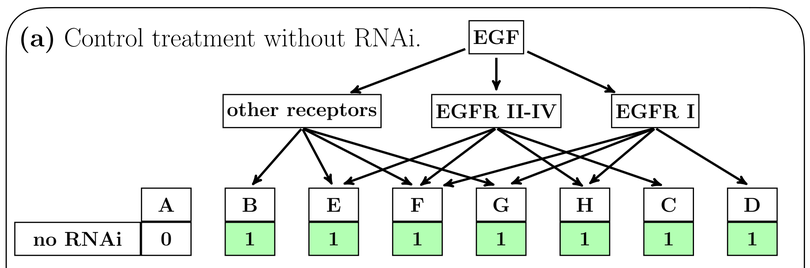
There is no siRNA with a region specific to isoforms II-IV, so it is impossible to knock-down only isoforms II-IV and not isoform I. Hence, we designed two nested RNAi experiments, where we selectively knocked down isoform I in one experiment and all four isoforms in the other, and we measured expression data from glioblastoma cell line SF767 in the corresponding samples to predict target genes regulated by isoforms II-IV but not by isoform I nor other receptors.

Finally, we developed a two-step bioinformatics approach based on the Bayesian Information Criterion for analysing expression data from such type of nested experimental design. The two-step Bayesian Gene Selection Criterion (BGSC) approach first defines conceptual groups of genes with qualitatively and quantitatively different responses to EGF and RNAi treatment and second classifies each gene into one of the four reduced gene groups based on the approximated posterior probability

Step 1.a - Conceptual gene groups

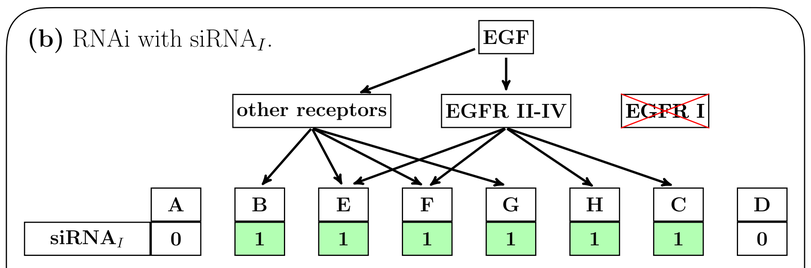
- Gene groups (A-H) contains all possible theoretical regulations of a gene after EGF addition in combination with the three RNAi treatments
- Differentially expressed genes markt by 1
- No differentially expressed genes markt by 0

(a) Control treatment without RNAi.



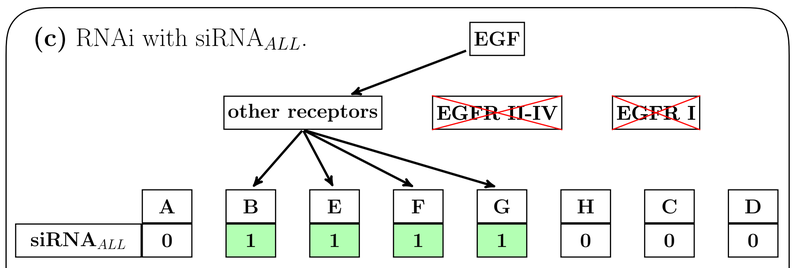
- Control experiment without RNAi treatment
  - no down-regulation of EGFR isoforms

(b) RNAi with siRNA$_I$.



- RNAi treatment with siRNA$_I$
  - down-regulation of EGFR isoform I (Red cross)

(c) RNAi with siRNA$_{ALL}$.



- RNAi treatment with siRNA$_{ALL}$
  - down-regulation of EGFR isoforms I and II-IV (Red cross)

Step 1.b - Reduction of the conceptual gene groups

	Differential expression by EGF							
group	A	B	E	F	G	H	C	D
no RNAi	0	1	1	1	1	1	1	1
siRNA$_I$	0	1	1	1	1	1	1	0
siRNA$_{ALL}$	0	1	1	1	1	0	0	0
simplified group	a	b				c		d

The eight gene groups A-H can be reduced to four the groups *a-d*:

- Genes of group a are never regulated by EGF.
- Genes of group b are regulated by EGF through other receptors and EGFR isoforms.
- Genes of group c are regulated by EGFR isoforms II-IV and not by other receptors.
- Genes of group d are regulated by EGFR isoform I and not EGFR isoforms II-IV or other receptors.

Goal: Prediction of putative target genes regulated by EGFR isoforms II-IV and not by other receptors (group c)

Step 2 - Statistical model of the data in the reduced gene groups

1. Assumption: expression data are log-normally distributed
2. Assumption: genes of group ***a*** are not differentially expressed
   - Likelihood:
$$p(x \mid a, \theta_a) = \prod_{n=1}^{N} \mathcal{N}(x_n \mid \mu_{a,0}, \sigma_a)$$
3. Assumption: genes of group $z \in \{b, c, d\}$ are differentially expressed by EGF-stimulation
   - Likelihood:
$$p(x \mid z, \theta_z) = \prod_{n=1}^{N} \mathcal{N}(x_n \mid \mu_{z,g_n}, \sigma_z)$$

group a	no EGF	EGF
no RNAi	$g_1 = 0$	$g_2 = 0$
siRNA$_I$	$g_3 = 0$	$g_4 = 0$
siRNA$_{ALL}$	$g_5 = 0$	$g_6 = 0$

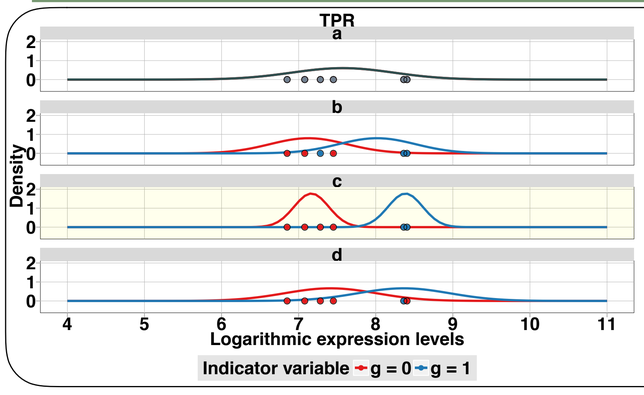
group b	no EGF	EGF
no RNAi	$g_1 = 0$	$g_2 = 1$
siRNA$_I$	$g_3 = 0$	$g_4 = 1$
siRNA$_{ALL}$	$g_5 = 0$	$g_6 = 1$

group c	no EGF	EGF
no RNAi	$g_1 = 0$	$g_2 = 1$
siRNA$_I$	$g_3 = 0$	$g_4 = 1$
siRNA$_{ALL}$	$g_5 = 0$	$g_6 = 0$

group d	no EGF	EGF
no RNAi	$g_1 = 0$	$g_2 = 1$
siRNA$_I$	$g_3 = 0$	$g_4 = 0$
siRNA$_{ALL}$	$g_5 = 0$	$g_6 = 0$

Expression pattern

Likelihood example for gene *TPR*



- Probability density of group ***a*** (grey curve)
- Probability densities of group $z \in \{b, c, d\}$ (red & blue curves)
- log2-expression values marked with points (coloured by indicator variable g)

Model of group ***c*** fits best the expression profile of *TPR* — best separation between the two estimated means (μ_{c0}, μ_{c1}) and the smallest estimated pooled variance (σ_c)

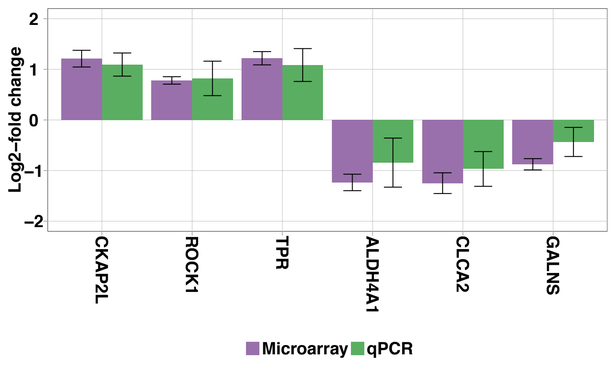
Approximate posterior probability

1. Approximation of marginal likelihoods p(x | k) for group k ∈ {*a, b, c, d*} using the Bayesian Information Criterion
2. Compute approximate posterior probability p(k | x) using prior probability p(a)=0.7 and p(b)=p(c)=p(d)=0.1
3. Assign each gene to that group k with highest approximate posterior probability p(k | x)

Results

Group	*a*	*b*	*c*	*d*
Assigned genes	8, 449	3, 822	3,143	1, 328

Validation by qPCR



- Person correlation coefficient of log2-fold changes of qPCR and microarray data = 0.99

Conclusions

- Developed Bayesian Gene Selection Criterion approach for predicting putative target genes of EGFR isoforms II-IV
- Predicted 3,143 putative target genes for group ***c***
- Validated six of these genes *CKAP2L, ROCK1,TPR, ALDH4A1, CLCA2*, and *GALNS* by triplicated qPCR experiments
- Validated target genes positively associated with migration, metastasis, and invasion of tumour cells and negatively associated with cell proliferation
- Datasets and R-scripts for reproducing all studies available at github.com/GrosseLab/BGSC

We thank Ralf Eggeling, Ioana Lemnian, Martin Porsch, and Teemu Roos for valuable discussions and the German Research Foundation, the German Federal Ministry of Education and Research, and the funding program Open Access Publishing by the DFG for financial support.