Journal of Biotechnology 189 (2014) 154–156

Contents lists available at ScienceDirect

Journal of Biotechnology

journal homepage: www.elsevier.com/locate/jbiotec

Short communication

CEM-Designer: Design of custom expression microarrays in the post-ENCODE Era

Christian Arnold[a,c], Fabian Externbrink[a], Jörg Hackermüller[b,a,c,*], Kristin Reiche[b,a,c,**]

[a] Bioinformatics Group, Department for Computer Science, University of Leipzig, Leipzig, Germany
[b] Young Investigators Group, Bioinformatics and Transcriptomics, Department Proteomics, Helmholtz Centre for Environmental Research – UFZ, Leipzig, Germany
[c] RNomics Group, Department of Diagnostics, Fraunhofer Institute for Cell Therapy and Immunology – IZI, Leipzig, Germany

ARTICLE INFO

Article history:
Received 20 February 2014
Received in revised form 2 September 2014
Accepted 17 September 2014
Available online 28 September 2014

2014 MSC:
00-01
99-00

Keywords:
Transcriptomics
Pervasive transcription
Non-coding RNA
Custom expression microarray
Microarray design

ABSTRACT

Microarrays are widely used in gene expression studies, and custom expression microarrays are popular to monitor expression changes of a customer-defined set of genes. However, the complexity of transcriptomes uncovered recently make custom expression microarray design a non-trivial task. Pervasive transcription and alternative processing of transcripts generate a wealth of interweaved transcripts that requires well-considered probe design strategies and is largely neglected in existing approaches.

We developed the web server CEM-Designer that facilitates microarray platform independent design of custom expression microarrays for complex transcriptomes. CEM-Designer covers (i) the collection and generation of a set of unique target sequences from different sources and (ii) the selection of a set of sensitive and specific probes that optimally represents the target sequences. Probe design itself is left to third party software to ensure that probes meet provider-specific constraints.

CEM-Designer is available at http://designpipeline.bioinf.uni-leipzig.de.

1. Introduction

Microarrays are a powerful technology that are used ubiquitously in biomedical and biotechnological research. Custom expression microarrays (CEMs) are popular to monitor expression changes of a customer-defined set of RNA transcripts, called target sequences. Despite the emergence of newer technologies such as RNA-seq, microarrays still offer advantages and remain a useful and accurate tool (Malone and Oliver, 2011).

Several tools exist for designing oligonucleotide probes for a set of target sequences (Chou, 2010; Agilent; Shin et al., 2009; Wernersson, 2009). A high-quality probe must be sensitive, specific, and isothermal with the other probes (Stekel, 2003). However, in light of pervasive transcription in mammalian genomes (Clark et al., 2011), these tools only address the criteria for sensitivity and isothermality appropriately. The measures taken to ensure a specific probe are often too simplistic, for at least the following two reasons:

(i) Genomic loci often encode a variety of transcripts comprising several isoforms of protein- and non-coding RNAs (ncRNAs) (Wilusz et al., 2009), with a large number of overlapping transcripts (Djebali et al., 2012). Hence, selection of specific regions in target sequences in which a probe should be placed is necessary. If probes are designed independently for each of these transcripts, a probe may be located in exons that are shared between isoforms (transcripts) of the same (different) gene(s). Specificity of those probes is consequently low due to potential cross-hybridization with different RNAs transcribed from the same genomic locus. Probes should instead be placed in exons or splice-sites that are unique to a target sequence.

(ii) In humans, at least 75% of the genome is capable of transcription in a highly time-, tissue-, and developmental-specific manner (Cabili et al., 2011). An RNA sample may, hence, contain novel transcripts not yet contained in any public database.

[*] Corresponding author at: Young Investigators Group, Bioinformatics and Transcriptomics, Department Proteomics, Helmholtz Centre for Environmental Research – UFZ, Leipzig, Germany.

[**] Corresponding author at: Young Investigators Group, Bioinformatics and Transcriptomics, Department Proteomics, Helmholtz Centre for Environmental Research – UFZ, Leipzig, Germany. Tel.: +49 3412351011.

E-mail addresses: joerg.hackermueller@ufz.de (J. Hackermüller), kristin.reiche@ufz.de (K. Reiche).

http://dx.doi.org/10.1016/j.jbiotec.2014.09.012
0168-1656/© 2014 Elsevier B.V. All rights reserved.

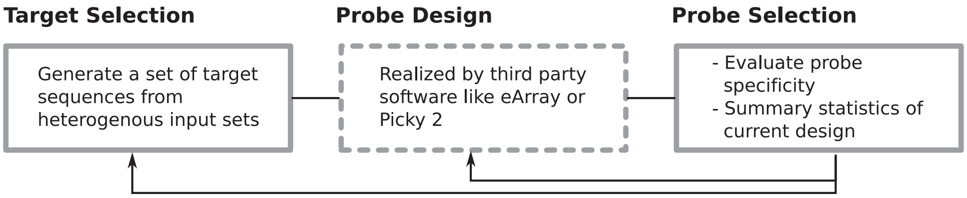
Target Selection
Probe Design
Probe Selection
Generate a set of target sequences from heterogenous input sets
Realized by third party software like eArray or Picky 2
- Evaluate probe specificity
- Summary statistics of current design

Fig. 1. The concept of the CEM-Designer web server specifically addressing target sequence selection and probe selection.

> However, current oligonucleotide probe designers return probes that are unique only with respect to currently annotated transcripts. Consequently, cross-hybridization of a probe must be tested against the complete transcriptome to not overestimate probe specificity, which is, however, in general not available. We, hence, propose to ensure probe uniqueness with respect to genome-wide alignments including alignments spanning introns.

We present the flexible web-server CEM-Designer that allows an adequate preparation of a set of target sequences and the identification of probes with better specificity than existing tools. Whenever appropriate, we compare an array design generated by CEM-Designer (GEO accession number GPL19136) with publicly available designs from Picky 2 (Chou, 2010) and Agilent.

2. Methods and results

Custom expression microarray design can be divided into three parts: target selection, probe design, and probe selection (Fig. 1). CEM-Designer addresses the first and last part more profoundly than available tools while leaving the probe design itself to third party software tools.

2.1. Target sequence selection

We interpret target sequence selection as the identification of subregions for probe placement prior to probe design to limit cross-hybridization with target sequences sharing the same genomic locus. Current probe designers address this only indirectly by assessing probe specificity according to all input target sequences after probe design (Chou, 2010; Wernersson, 2009; Agilent) but leave a profound preprocessing to the user. CEM-Designer provides a flexible and user-friendly handling of target sequences and accounts for the variety of known biotypes in complex transcriptomes (Harrow et al., 2012). It allows different input files, each of which may represent a particular target class (mRNAs, small or long ncRNAs). To maximize flexibility, processing parameters can be adjusted separately for each file.

2.1.1. Overlapping target sequences

Target sequences sharing the same genomic locus are often not separately detectable by state-of-the-art array designs (Supplemental Fig. S1A). CEM-Designer offers three choices for handling overlapping target sequences: (i) ignore overlaps, (ii) merge overlapping sequences to one combined sequence, or (iii) use exonic subregions that do not overlap with any other target sequence (Supplemental Fig. S1C). Options (i) and (ii) are favorable if genes should be represented on the microarray and a distinction into different isoforms is irrelevant. Option (iii) allows a distinct quantification of the expression of isoforms. Placing probes over unique splice-sites is currently not supported, as this would reduce the search space for optimal probes for a target sequence during the probe design step. A comparison with Picky 2 and eArray revealed an increased coverage of specific exons when CEM-Designer is used (Supplemental Table S1D).

2.1.2. Target sequence filters

Various user-adjustable sequence filters may be defined to discard target sequences or subregions thereof that are of insufficient length, redundant or negligible for the purpose of the study. For example, if only ncRNAs are of interest, overlapping exons of mRNAs can be excluded (Supplemental Fig. S1B).

2.1.3. Optimal coverage of sensitive probes

For a reasonable trade-off between selecting optimal positions for sensitive probes and optimal probe coverage of target sequences, CEM-Designer allows to partition target sequences into shorter subsequences. To ensure that probes may also be designed in the vicinity of the split positions, a specific overlap can be defined. In the probe design, users may select a maximal number of probes according to the length of the (sub)sequences.

2.2. Oligonucleotide probe design

As CEM manufacturers like Agilent provide their own models for probe design to optimize melting temperatures for their specific array technology, we do not include a probe design tool. However, preprocessed target sequences are provided in standard file formats that can be directly used as input for subsequent probe design.

2.3. Oligonucleotide probe selection

After oligonucleotide probe design, CEM-Designer evaluates probe specificity, generates summary statistics, and prepares subsequent CEM designs, if necessary.

2.3.1. Probe specificity

Cross-hybridization to other RNA transcripts than the target transcript is assessed by aligning a probe to the full genome, including intron spanning alignments, using BLAT (Kent, 2002). To assure high probe specificity, non-uniquely mapping probes are discarded. The strategy to assess probe uniqueness with respect to genome-wide alignments is outperforming existing tools, as they test probe specificity with respect to known transcripts only (Supplemental Table S1A). The latter strategy is, for example, neglecting transcriptionally active pseudogenes (Frankish and Harrow, 2014; Johnsson et al., 2013) because probes designed to represent a parent gene may map to an expressed pseudogene due to high sequence similarity. Indeed, pseudogenes showed a large fraction of non-unique probes for probe sets from Picky 2 and eArray (Supplemental Table S1B).

2.3.2. Preparation of subsequent designs

CEM-Designer calculates summary statistics such as the number of uniquely and non-uniquely mapping probes, and the distribution of the target sequences probe coverage. It further lists target sequences for which fewer probes than anticipated were designed. To maximize target coverage, they can be used as input for probe redesign or reapplication of CEM-Designer with altered parameters (Fig. 1). This scheme can in principle be repeated until no new uniquely mapping probes can be designed.

2.4. Assessing expression changes based on probes generated by CEM-Designer

Biological meaningful expression variation was successfully captured by probes generated by CEM-Designer and spotted on an Agilent custom microarray (GEO accession number GPL19136). The CEM was used to investigate expression changes of long ncRNAs and mRNAs in human primary foreskin fibroblasts synchronized in G0 and G1 phases of mitotic cell cycle (Hackermüller et al., 2014). A principal component analysis of all custom probes passing unspecific filtering resulted in two well separated clusters accounting for 76% overall expression variation in the first principle component and reflecting biological differences between cells in G0 and G1 (Supplemental Fig. S2).

3. Discussion and conclusion

Mammalian transcriptomes are pervasively transcribed, which makes a suitable selection and preprocessing of target sequences prior to probe design as well as the evaluation of probe specificity for expression microarrays non-trivial. As all three issues have a profound impact on the outcome of microarray expression studies, we designed CEM-Designer to explicitly address transcriptome complexity. Cross-hybridization effects to other target sequences, even to yet unannotated transcripts that may nevertheless be contained in the RNA sample, are minimized. CEM-Designer combines several input sources into a single array design but allows the selection of distinct parameters for each source, thereby providing high flexibility for the selection and preprocessing of unique target sequences. Quality tests after probe design provide summary statistics and identify target sequences with low coverage of probes. The concept as implemented in CEM-Designer was successfully transferred to studies investigating the expression of long ncRNAs and mRNAs (Hackermüller et al., 2014; Reiche et al., 2014).

Funding

This work was supported by the Initiative and Networking Fund of the Helmholtz Association (VH-NG738), the Bundesministerium für Bildung und Forschung (BMBF), and the Sächsische Aufbaubank (SAB), grant # 14494. Funders had no role in design of the software and decision to publish, or preparation of the manuscript.

Appendix A. Supplementary Data

Supplementary data associated with this article can be found, in the online version, at http://dx.doi.org/10.1016/j.jbiotec.2014.09.012.

References

Malone, J.H., Oliver, B., 2011. Microarrays, deep sequencing and the true measure of the transcriptome. BMC Biol. 9 (1), 34.

Chou, H.-H., 2010. Shared probe design and existing microarray reanalysis using PICKY. BMC Bioinform. 11, 196, http://dx.doi.org/10.1186/1471-2105-11-196.

Agilent, eArray, https://earray.chem.agilent.com/earray/

Shin, S.-Y., Lee, I.-H., Cho, Y.-M., Yang, K.-A., Zhang, B.-T., 2009. EvoOligo: oligonucleotide probe design with multiobjective evolutionary algorithms. IEEE Trans. Syst. Man Cybern. B: Cybern. 39 (6), 1606–1616, http://dx.doi.org/10.1109/TSMCB.2009.2023078.

Wernersson, R., 2009. Probe design for expression arrays using OligoWiz. Methods Mol. Biol. 529, 23–36, http://dx.doi.org/10.1007/978-1-59745-538-1_2.

Stekel, D., 2003. Microarray Bioinformatics. Cambridge University Press, http://dx.doi.org/10.1017/CBO9780511615535.

Clark, M.B., Amaral, P.P., Schlesinger, F.J., Dinger, M.E., Taft, R.J., Rinn, J.L., Ponting, C.P., Stadler, P.F., et al., 2011. The reality of pervasive transcription. PLoS Biol. 9 (7), e1000625, http://dx.doi.org/10.1371/journal.pbio.1000625 (discussion e1001102).

Wilusz, J.E., Sunwoo, H., Spector, D.L., 2009. Long noncoding RNAs: functional surprises from the RNA world. Genes Dev. 23 (13), 1494–1504, http://dx.doi.org/10.1101/gad.1800909.

Djebali, S., Davis, C.A., Merkel, A., Dobin, A., Lassmann, T., Mortazavi, A., Tanzer, A., Lagarde, J., Lin, W., Schlesinger, F., et al., 2012. Landscape of transcription in human cells. Nature 489 (7414), 101–108.

Cabili, M.N., Trapnell, C., Goff, L., Koziol, M., Tazon-Vega, B., Regev, A., Rinn, J.L., 2011. Integrative annotation of human large intergenic noncoding RNAs reveals global properties and specific subclasses. Genes Dev. 25 (18), 1915–1927, http://dx.doi.org/10.1101/gad.17446611.

Harrow, J., Frankish, A., Gonzalez, J.M., Tapanari, E., Diekhans, M., Kokocinski, F., Aken, B.L., Barrell, D., Zadissa, A., et al., 2012. GENCODE: the reference human genome annotation for The ENCODE Project. Genome Res. 22 (9), 1760–1774., http://dx.doi.org/10.1101/gr.135350.111.

Kent, W.J., 2002. BLAT-the BLAST-like alignment tool. Genome Res. 12 (4), 656–664, http://dx.doi.org/10.1101/gr.229202.

Frankish, A., Harrow, J., 2014. GENCODE pseudogenes. Methods Mol. Biol. 1167, 129–155.

Johnsson, P., Ackley, A., Vidarsdottir, L., Lui, W.O., Corcoran, M., Grander, D., Morris, K.V., 2013. A pseudogene long-noncoding-RNA network regulates PTEN transcription and translation in human cells. Nat. Struct. Mol. Biol. 20 (4), 440–446.

Hackermüller, J., Reiche, K., Otto, C., Hösler, N., Blumert, C., Brocke-Heidrich, K., Böhlig, L., Nitsche, A., Kasack, K., Ahnert, P., Krupp, W., Engeland, K., Stadler, P.F., Horn, F., 2014. Cell cycle, oncogenic and tumor suppressor pathways regulate numerous long and macro non-protein coding RNAs. Genome Biol. 15 (3), R48, http://dx.doi.org/10.1186/gb-2014-15-3-r48.

Reiche, K., Kasack, K., Schreiber, S., Lüders, T., Due, E.U., Naume, B., Riis, M., Kristensen, V.N., Horn, F., Børresen-Dale, A.L., Hackermüller, J., Baumbusch, L.O., 2014. Long Non-Coding RNAs Differentially Expressed between Normal versus Primary Breast Tumor Tissues Disclose Converse Changes to Breast Cancer-Related Protein-Coding Genes. PLoS ONE 9, 9–e106076.