

Design of Oligonucleotide Microarrays for Vertebrate Chromatin Immunoprecipitation: From Promoters to Whole Genomes

George W. Bell¹, Bingbing Yuan², Tong Ihn Lee³, and Fran Lewitter⁴

Keywords: microarray design, chromatin immunoprecipitation, transcription factors

1 Abstract.

The binding of transcriptional regulators to DNA can induce or repress gene expression in a cell-specific manner in different physiological conditions. To help understand this mechanism on a genome-wide scale, we developed a series of microarrays for chromatin immunoprecipitation (ChIP chips) for human and mouse.

To assay DNA binding that occurs near to transcriptional start sites (TSSs), we first created gene-centric ChIP chips to reflect activity in transcriptional units' proximal promoters. To select TSS data of high confidence for human and mouse, we merged mapping data from at least three gene sets (such as RefSeq, Ensembl, and MGC), selecting TSSs ($n = 18k$ for human; $16k$ for mouse) consistent across at least two databases. Promoters (-8 kb to $+2$ kb relative to the TSS) were extracted from the genome after processing with RepeatMasker¹. To assay transcriptional regulators, including those that bind far from known transcriptional units, we also designed a much larger chipset based on the entire RepeatMasked genome.

These genome-scale oligonucleotide ChIP chips required the selection of sequences that can optimally represent a region of the genome of at least 300 nt. To help minimize false positive signals, representative probes were designed close enough that adjacent probes could measure any binding event. On the other hand, probes were selected far enough apart so a minimal number could represent large regions or all of the genome. Finally, to ensure that we could identify binding events near the edge of repetitive genomic regions, we selected probes near the distal ends of each non-repetitive region.

2 Results.

Human and mouse gene-centric chipsets were designed with 4×10^5 features, including control features. Features included positive controls representing regions known to bind well-characterized

¹ Bioinformatics and Research Computing, Whitehead Institute for Biomedical Research, Nine Cambridge Center, Cambridge, MA. 02142, USA. E-mail: gbell@wi.mit.edu

² Bioinformatics and Research Computing, Whitehead Institute for Biomedical Research, Nine Cambridge Center, Cambridge, MA. 02142, USA. E-mail: yuan@wi.mit.edu

³ Whitehead Institute for Biomedical Research, Nine Cambridge Center, Cambridge, MA. 02142, USA. E-mail: tlee@wi.mit.edu

⁴ Bioinformatics and Research Computing, Whitehead Institute for Biomedical Research, Nine Cambridge Center, Cambridge, MA. 02142, USA. E-mail: lewitter@wi.mit.edu

transcriptional regulators; negative controls representing “gene deserts”, euchromatin regions distant from any annotated transcribed region; matched intensity controls for all chips; and negative controls comprising plant-specific DNA. A human whole genome chipset was designed with 4×10^6 features, including control features such as matched intensity controls and plant-specific DNA.

Using a series of 3 filters of decreasing stringency, we were able to minimize gaps during probe selection and represent virtually all non-repetitive genome regions. Our custom algorithm selected oligonucleotide probes at a more optimal spacing than ArrayOligoSelector² (AOS), thus creating a more representative and more efficient microarray design. Our design methods are also being applied to the zebrafish genome.

3 Methods.

After extracting the non-repetitive regions of the genome we wished to represent, potential probes (60 nt) from each genomic region were assayed for uniqueness, complexity, secondary structure, and GC content using AOS, locally optimized for high performance. BLAT³ was used both to assay for uniqueness and to map each selected probe to the genome. Using the four AOS parameters above, up to 3 filters of decreasing stringency were applied. After removing poorly scoring oligonucleotides by filtering, sequences were chosen at optimal spacing (300 nt) so that any binding event would be captured by at least two adjacent sequences.

4 Discussion.

Our designs are based on two major assumptions. First, repetitive regions of the genome are unlikely to contain key sites of transcriptional regulation. Whereas this may lead to some missing of regulatory regions, it effectively halves the size of the genome sequence represented by the ChIP chips. Second, the gene-centric designs assume that most transcriptional regulation occurs -8 kb to +2 kb relative to the TSS. Based on well-curated data from the literature such as TRANSFAC⁴, more than 90% of all binding events occur within 1 kb of the TSS. Although these data are surely biased due to the difficulty of carefully assaying large genomic regions, assaying the 10 kb around a TSS is an efficient way to study many genes at once. Results from whole genome ChIP chip experiments will help determine distal regions where transcriptional regulators may also be active.

Results of several ChIP chip experiments, which verified our design principles, will also be presented.

5 References.

- [2] Bozdech Z, Zhu J, Joachimiak MP, Cohen FE, Pulliam B, DeRisi JL. 2003. Expression profiling of the schizont and trophozoite stages of *Plasmodium falciparum* with a long-oligonucleotide microarray. *Genome Biology* 4(2):R9.
- [3] Kent WJ. 2002. BLAT--the BLAST-like alignment tool. *Genome Research* Apr;12(4):656-64.
- [4] Matys V, Fricke E, Geffers R, Gossling E, Haubrock M, Hehl R, Hornischer K, Karas D, Kel AE, Kel-Margoulis OV, Kloos DU, Land S, Lewicki-Potapov B, Michael H, Munch R, Reuter I, Rotert S, Saxel H, Scheer M, Thiele S, Wingender E. 2003. TRANSFAC: transcriptional regulation, from patterns to profiles. *Nucleic Acids Research* 31(1):374-8.
- [1] Smit, AFA, Hubley, R & Green, P. 1996-2004. *RepeatMasker Open-3.0*. <<http://www.repeatmasker.org>>.