OPEN ACCESS Freely available online

Message from ISCB

Getting Started in Tiling Microarray Analysis

X. Shirley Liu

Introduction

The availability of sequenced eukaryotic genomes and commercial oligonucleotide tiling microarrays has enabled many genomics applications. Different from expression microarrays, tiling microarrays have probes that cover the entire genome or contigs of the genome in an unbiased fashion. Currently three commercial sources provide tiling microarrays with different probe lengths and spacing, and array design characteristics. Affymetrix tiles 6 million 25-mer probes per array, which offers the lowest price per probe and the highest resolution (chromosomal distance between neighboring probe centers). Its arrays use one-color assays, so individual samples are hybridized to different arrays. NimbleGen can tile 385,000 50- to 75-mer probes, and Agilent can tile 244,000 60-mer probes per array. The latter two platforms, with longer oligonucleotide probes and two-color assays for which treatment and control samples are differentially labeled and put on the same array for competitive hybridization, have slightly better sensitivity. They are also flexible for custom array design, especially Agilent's multiplex arrays, which allow multiple samples to hybridize on different subareas of the same array. These tiling arrays offer diverse genomic applications, each with its own data analysis challenges.

ChIP-Chip

The most popular application for the tiling array platform is ChIP-chip, which maps the genome-wide binding locations of transcription factors and other DNA-binding proteins. In a ChIP-chip experiment, chromatin is crosslinked and fragmented to approximately 500 bp. An antibody to the protein of interest is used to precipitate the protein together with its interacting DNA (chromatin immunoprecipitation, or "ChIP"). The coprecipitated DNA is detected on a DNA microarray (the "chip") and mapped back to the genome [1,2]. In complex genomes, DNA-binding proteins often have thousands of binding sites throughout the genome, so genome tiling microarrays from Affymetrix [3], NimbleGen [4], and Agilent [5] can be used for unbiased binding site mapping.

For ChIP-chip on Affymetrix tiling microarrays, MAT (model-based analysis of tiling arrays) [6] is a very effective peak-finding algorithm. MAT standardizes probe behavior by its 25-mer probe sequence and genome copy number, and can work even without replicate ChIP or control samples. Occasionally Affymetrix genome tiling microarrays have blob-like image defects, which are visible when the array image is converted to a data .cel file. If users encounter array images with blob defects, they are advised to use a "microarray blob remover" [7] to detect and remove affected probes before running MAT. For NimbleGen tiling microarrays, TAMAL [8] is the best algorithm for locating binding sites, while MA2C [9] and TileScope [10] offer alternatives that are more user-friendly and flexible. For Agilent tiling arrays, the joint binding deconvolution [11] algorithm can detect ChIP-chip peaks, in addition providing finer peak spatial resolution than Agilent array tiling resolution.

After the ChIP-chip peaks are detected, biologists often want to find the sequence-specific binding motifs of their protein of interests. MEME [12] and Gibbs Motif Sampler [13] are the most popular tools for de novo motif discovery. As an alternative, biologists could use the cis-regulatory element annotation system [14] to annotate large-scale ChIP-chip data in human and mouse, such as retrieving ChIP-chip sequences, mapping nearby genes, plotting sequence conservation figures, and finding enriched known transcription factor motifs. For a more generalized genomics annotation pipeline, Galaxy (http://main.g2.bx.psu.edu/) offers more customized and interactive features to analyze additional sequenced genomes.

MeDIP-Chip and DNase-Chip

DNA methylation status often controls gene transcription status, and genome-wide DNA methylation sites can be mapped using methyl–DNA immunoprecipitation followed by microarray (MeDIP-chip). MeDIP-chip is similar to ChIP-chip in protocol, except that an antibody against 5-methyl-cytosine is used to directly precipitate methylated DNA [15,16]. Peak identification and annotation of MeDIP-chip experiments can be conducted with methods similar to ChIP-chip. The methylation level measured by MeDIP-chip should be calibrated by the GC content of the region, since poorly methylated CG-rich regions might still have a higher number of methyl-Cs to MeDIP than fully methylated CG-poor regions.

DNase-hypersensitive regions in the genome are often open chromatin harboring transcriptionally active or regulatory regions, which can be located using DNase-chip. Relying on the assumption that open chromatin is cleaved more often by DNase over a short distance, this experiment involves digesting chromatin with DNase I,

Editor: Olga Troyanskaya, Princeton University, United States of America

Citation: Liu XS (2007) Getting started in tiling microarray analysis. PLoS Comput Biol 3(10): e183. doi:10.1371/journal.pcbi.0030183

Copyright: © 2007 X. Shirley Liu. This is an open-access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

X. Shirley Liu is with the Department of Biostatistics and Computational Biology, Dana-Farber Cancer Institute, Harvard School of Public Health, Boston, Massachusetts, United States of America. E-mail: xsliu@jimmy.harvard.edu

isolating DNA fragments created by two DNase cleavages less than 1,200 bp apart, and hybridizing the DNA to tiling microarrays [17]. The resulting tiling array data can be analyzed with a regular ChIP-chip peak-finding algorithm, although window size needs to be adjusted based on the DNA fragment length distribution resulting from the level of DNase digestion.

Nucleosome Localization

A nucleosome, which consists of ~146 bp of DNA wrapped around eight histone proteins, forms the fundamental structural unit of eukaryotic chromatin. Since nucleosomes limit DNA accessibility to regulatory factors, it is important to map positioned nucleosomes or nucleosome-free regions in the genome. Nucleosome mapping experiments involve digesting the chromatin with micrococcal nuclease to remove the linker DNA between neighboring nucleosomes, and isolating the remaining nucleosomal DNA to be labeled and hybridized to a tiling microarray. The controls for such experiments are often naked genomic DNA (without chromatin structure) cleaved with hydroxyl radicals or micrococcal nuclease to the same size distribution. Unlike ChIP-chip, the occupancy difference between positioned nucleosomes and linker regions is often less than 10-fold, and positioned nucleosomes occupy only about 100–200 bp of DNA. This requires the tiling microarray to have both high sensitivity and high resolution. Long oligonucleotide microarrays tiled at 5–20 bp resolution are often custom-made to cover selected genomic regions (e.g., promoters or a few megabases on a chromosome) for this application.

In a nucleosome mapping study conducted in yeast Chromosome III [18], a hidden Markov model was applied. The model defines a stretch of probes with low signals as linkers, six to eight probes that span approximately 146 bp with high signals as well-positioned nucleosomes, and more than eight probes with intermediately high signals as delocalized nucleosomes. A Viterbi algorithm is used to infer the most likely partition of probes along the chromosome into the different nucleosomal states. In a similar study conducted in human promoters [19], wavelet transformation was first used to remove noise from the probe signal, which eliminated the high frequency and low coefficient signals. Laplacian Gaussian edge detection was applied to the smoothed probe signal curve to detect peaks and troughs (zero first derivatives) with a reasonable width as positioned nucleosomes and linker regions, respectively.

ArrayCGH and Copy Number Variation

In an array-based comparative genome hybridization (arrayCGH) experiment, DNA from normal and diseased individuals are differentially hybridized to microarrays to identify copy number variations in the genome that are potential biomarkers or causal genes of disease [20]. Early microarrays used in arrayCGH studies have long (e.g., BAC clones) and/or sparse probes to cover the genome. Recently, tiling microarrays have been used to improve the copy number variation detection sensitivity and resolution [21]. One method proposes a structural change model to use dynamic programming to segment the genome into a number of regions with different copy numbers; within each region the probe signals (thus genome copy number) are similar [22]. However, selecting the number of regions could be difficult for big genomes with complex copy number variations. Hidden Markov model is also a plausible approach to infer the hidden copy number based on observed probe values. One complication that all arrayCGH applications need to reconcile with is that sample impurities (e.g., patient DNA degradation or heterogeneous tumor DNA) sometimes give rise to noisy signals or non-integer copy numbers.

Transcriptome Mapping

Hybridizing RNA samples to tiling microarrays is gaining popularity for detecting novel transcripts in sequenced genomes. Early studies often called positive probes based on a probe signal cutoff [23], then defined stretches of genomic regions with a significant number of positive probes as transfrags (transcribed fragments). One study on yeast 4-bp resolution tiling arrays adopted a structural change model similar to that used in arrayCGH [24]. In a more recent study profiling multiple *Drosophila* embryogenesis stages on genome tiling microarrays, a Kruskal-Wallis test (a nonparametric analog of one-way ANOVA) was used to detect a stretch of probes giving differential expression among conditions [25]. In addition, the study checked neighboring transfrags with correlated expression in different conditions to find novel 5′, 3′, or internal exons of previously annotated genes. With more transcriptome conditions profiled at better tiling resolution, more advanced algorithms can be developed to refine transfrag borders and detect differential expression, alternative splicing, and antisense transcripts.

Prospective

All commercial tiling microarray companies strive to put more probes on the array at reduced cost. This trend seems to follow the Moore's Law observed in the semiconductor industry, which dictates that chips double their density at half the cost every 18 months. A few years from now might see tiling microarrays covering the whole mammalian genome at single-base resolution that cost only a few thousand dollars. Tiling arrays will have much wider applications, and researchers might use them for different experiments and informatically select a subset of the probes for analysis. At the same time, high-throughput sequencing technologies such as 454, Illumina Solexa, and ABI SOLiD are making fast progress as well. If enough coverage can be achieved at a cost similar to tiling microarrays, they might give more sensitive and unbiased results. These technologies each entail different challenges and opportunities for computational biologists to develop efficient analysis algorithms. The competition between the different technology companies will inevitably benefit researchers regardless of the winner. Therefore, we look forward to a very exciting decade of genomics advances ahead. ■

Funding. XSL was supported by US National Institutes of Health grant 1R01 HG004069–01.

Competing interests. The author has declared that no competing interests exist.

References

1. Bernstein BE, Humphrey EL, Liu CL, Schreiber SL (2004) The use of chromatin immunoprecipitation assays in genome-wide analyses of histone modifications. Methods Enzymol 376: 349–360.
2. Hanlon SE, Lieb JD (2004) Progress and challenges in profiling the dynamics of chromatin and transcription factor binding with DNA microarrays. Curr Opin Genet Dev 14: 697–705.
3. Carroll JS, Meyer CA, Song J, Li W, Geistlinger TR, et al. (2006) Genome-wide analysis of estrogen receptor binding sites. Nat Genet 38: 1289–1297.
4. Kim TH, Barrera LO, Zheng M, Qu C, Singer MA, et al. (2005) A high-resolution map of active promoters in the human genome. Nature 436: 876–880.
5. Lee TI, Jenner RG, Boyer LA, Guenther MG, Levine SS, et al. (2006) Control of developmental regulators by Polycomb in human embryonic stem cells. Cell 125: 301–313.
6. Johnson WE, Li W, Meyer CA, Gottardo R, Carroll JS, et al. (2006) Model-based analysis of tiling-arrays for ChIP–chip. Proc Natl Acad Sci U S A 103: 12457–12462.
7. Song JS, Maghsoudi K, Li W, Fox E, Quackenbush J, et al. (2007) Microarray blob-defect removal improves array analysis. Bioinformatics 23: 966–971.
8. Bieda M, Xu X, Singer MA, Green R, Farnham PJ (2006) Unbiased location analysis of E2F1-binding sites suggests a widespread role for E2F1 in the human genome. Genome Res 16: 595–605.
9. Song JS, Johnson WE, Zhu X, Zhang X, Liu Y, et al. (2007) Model-based analysis of 2-color arrays (MA2C). Genome Biol. In press.
10. Zhang ZD, Rozowsky J, Lam HY, Du J, Snyder M, et al. (2007) Tilescope: Online analysis pipeline for high-density tiling microarray data. Genome Biol 8: R81.
11. Qi Y, Rolfe A, MacIsaac KD, Gerber GK, Pokholok D, et al. (2006) High-resolution computational models of genome binding events. Nat Biotechnol 24: 963–970.
12. Bailey TL, Williams N, Misleh C, Li WW (2006) MEME: Discovering and analyzing DNA and protein sequence motifs. Nucleic Acids Res 34: W369–W373.
13. Thompson W, Rouchka EC, Lawrence CE (2003) Gibbs Recursive Sampler: Finding transcription factor binding sites. Nucleic Acids Res 31: 3580–3585.
14. Ji X, Li W, Song J, Wei L, Liu XS (2006) CEAS: Cis-regulatory element annotation system. Nucleic Acids Res 34: W551–W554.
15. Zhang X, Yazaki J, Sundaresan A, Cokus S, Chan SW, et al. (2006) Genome-wide high-resolution mapping and functional analysis of DNA methylation in arabidopsis. Cell 126: 1189–1201.
16. Weber M, Hellmann I, Stadler MB, Ramos L, Paabo S, et al. (2007) Distribution, silencing potential and evolutionary impact of promoter DNA methylation in the human genome. Nat Genet 39: 457–466.
17. Sabo PJ, Kuehn MS, Thurman R, Johnson BE, Johnson EM, et al. (2006) Genome-scale mapping of DNase I sensitivity in vivo using tiling DNA microarrays. Nat Methods 3: 511–518.
18. Yuan GC, Liu YJ, Dion MF, Slack MD, Wu LF, et al. (2005) Genome-scale identification of nucleosome positions in *S. cerevisiae*. Science 309: 626–630.
19. Ozsolak F, Song JS, Liu XS, Fisher DE (2007) High-throughput mapping of the chromatin structure of human promoters. Nat Biotechnol 25: 244–248.
20. Sebat J, Lakshmi B, Troge J, Alexander J, Young J, et al. (2004) Large-scale copy number polymorphism in the human genome. Science 305: 525–528.
21. Urban AE, Korbel JO, Selzer R, Richmond T, Hacker A, et al. (2006) High-resolution mapping of DNA copy alterations in human chromosome 22 using high-density tiling oligonucleotide arrays. Proc Natl Acad Sci U S A 103: 4534–4539.
22. Daruwala RS, Rudra A, Ostrer H, Lucito R, Wigler M, et al. (2004) A versatile statistical analysis algorithm to detect genome copy number variation. Proc Natl Acad Sci U S A 101: 16292–16297.
23. Bertone P, Stolc V, Royce TE, Rozowsky JS, Urban AE, et al. (2004) Global identification of human transcribed sequences with genome tiling arrays. Science 306: 2242–2246.
24. David L, Huber W, Granovskaia M, Toedling J, Palm CJ, et al. (2006) A high-resolution map of transcription in the yeast genome. Proc Natl Acad Sci U S A 103: 5320–5325.
25. Manak JR, Dike S, Sementchenko V, Kapranov P, Biemar F, et al. (2006) Biological function of unannotated transcription during the early development of *Drosophila melanogaster*. Nat Genet 38: 1151–1158.