Error-pooling-based statistical methods for identifying novel temporal replication profiles of human chromosomes observed by DNA tiling arrays

Taesung Park¹, Youngchul Kim¹, Stefan Bekiranov² and Jae K. Lee³,*

¹Department of Statistics, Seoul National University, Korea, ²Department of Biochemistry and Molecular Genetics, University of Virginia School of Medicine, USA and ³Division of Biostatistics and Epidemiology, University of Virginia School of Medicine, USA

Received February 6, 2007; Revised March 20, 2007; Accepted February 16, 2007

ABSTRACT

Statistical analysis on tiling array data is extremely challenging due to the astronomically large number of sequence probes, high noise levels of individual probes and limited number of replicates in these data. To overcome these difficulties, we first developed statistical error estimation and weighted ANOVA modeling approaches to high-density tiling array data, especially the former based on an advanced error-pooling method to accurately obtain heterogeneous technical error of small-sample tiling array data. Based on these approaches, we analyzed the high-density tiling array data of the temporal replication patterns during cell-cycle S phase of synchronized HeLa cells on human chromosomes 21 and 22. We found many novel temporal replication patterns, identifying about 26% of over 1 million tiling array sequence probes with significant differential replication during the four 2-h time periods of S phase. Among these differentially replicated probes, 126,941 sequence probes were matched to 417 known genes. The majority of these genes were found to be replicated within one or two consecutive time periods, while the others were replicated at two non-consecutive time periods. Also, coding regions found to be more differentially replicated in particular time periods than noncoding regions in the gene-poor chromosome 21 (25% differentially replicated among genic probes versus 18.6% among intergenic probes), while such a phenomenon was less prominent in gene-rich chromosome 22. A rigorous statistical testing for local proximity of differentially replicated genic and intergenic probes was performed to identify significant stretches of differentially replicated sequence regions. From this analysis, we found that adjacent genes were frequently replicated at different time periods, potentially implying the existence of quite dense replication origins. Evaluating the conditional probability significance of identified gene ontology terms on chromosomes 21 and 22, we detected some over-represented molecular functions and biological processes among these differentially replicated genes, such as the ones relevant to hydrolase, transferase and receptor-binding activities. Some of these results were confirmed showing >70% consistency with cDNA microarray data that were independently generated in parallel with the tiling arrays. Thus, our improved analysis approaches specifically designed for high-density tiling array data enabled us to reliably and sensitively identify many novel temporal replication patterns on human chromosomes.

INTRODUCTION

Various mechanisms of human chromosome replication are still unclear, including whether the molecular structure and biological function of genes are correlated with replication timing on chromosomes. A better understanding of the replication process of human chromosomes may be achieved by obtaining a detailed knowledge of their time of replication, and many recent studies have addressed replication timing of human genome based on such a strategy (1,2). Recently, DNA tiling microarrays have been used to assay patterns of DNA replication at different stages of S phase on human chromosomes 21 and 22 (3). However, statistical analysis on tiling array data introduces new challenges beyond the standard analysis approaches to the widely used RNA expression profiling microarrays due to noisy and heterogeneous errors in...
tiling array probes compared to gene expression arrays. This is likely due to the minimal probe selection process, which results in a wide range of probe sensitivities and specificities on the arrays. Furthermore, experiments that use tiling arrays are typically performed with a limited number of replicates—three in the above-mentioned replication study. Consequently, many classical statistical methods that rely on a relatively large sample size and homogeneous variance for their maximal performance are severely underpowered and biased when applied to this kind of data. Indeed, recent studies introduced several approaches to tiling data analysis, including hidden Markov models (4), G-TRANS (5) and empirical Bayes model (TileMap) (6); however, their statistical inference is based on individual error estimates that may not be accurate for an extremely large number of tiling array probes with a small number of replicates.

Based on a non-parametric test on sliding windows of tiling array probes, Jeon et al. (3) showed many interesting findings on the replication timing of human chromosomes, reporting that ~60% of interrogated tiling array probes were evenly replicated across the four time periods. Here, significantly improving the analysis accuracy and fidelity by several novel statistical approaches, we reanalyze this data set to identify sequence probes and genes that are replicated at specific times in S phase. Specifically, overcoming the aforementioned difficulties, we apply an improved error estimation approach to small-sample tiling array data using a recent error-pooling method called local pooled error (LPE) (7). We also use several novel statistical methods that are well suited for analyzing tiling array data: a weighted ANOVA modeling for simultaneously identifying differentially replicated probes across the four time periods of the replication in cell-cycle S-phase, significant stretch analysis for testing the local proximity of differentially replicated genic and intergenic probes and conditional probability evaluation on gene ontology terms for discovering over-represented functions and molecular mechanisms of differentially replicated genes.

METHODS

High-density genome tiling and cDNA microarray data

The DNA tiling array data that comprise 1 020 653 probe pairs interrogating the repeat-masked sequences of chromosomes 21 and 22 (Affymetrix, Santa Clara, CA), which were originally reported in Jeon et al. (3), are reanalyzed in this study. These tiling-array probes are ordered from the centromere to the end of the long arm of each chromosome, with average probe spacing 35 bp and unbiased selection between coding (exon) and noncoding (intron) regions. Replication products that were obtained from cell-cycle-synchronized HeLa cells by thymidine-aphidicolin block were hybridized to tiling arrays at different stages of S phase with four consecutive time periods of 2-h intervals: 0–2, 2–4, 4–6 and 6–8 h. Overall, HeLa cells released from a G1-S block showed that DNA content initiated to increase by 2 h, and the whole content was doubled after 8–10 h. More details regarding the experimental procedures can be found in Jeon et al. (3).

We also obtained the array data from the cDNA microarray experiment that was performed in parallel with the above tiling array experiment. In brief, a human cDNA array platform containing 1589 cDNA clones spotted in duplicate across the entire human genome, including 21 cDNA clones for chromosome 21 and 20 for chromosome 22 that met internal quality control, i.e. consistency between duplicated clones, was also used to examine the replication expression patterns during the above four time periods and 8–10 h of S phase. The sample from each time period was labeled with Cy5 and the pooled sample from 9 to 10 h with Cy3. For our analysis, the log-ratio intensities between the Cy3 and Cy5 samples were available for the five time periods.

Tiling data preprocessing and normalization based on perfect match probes

In the analysis of Affymetrix oligonucleotide microarrays, it has been reported that the intensity evaluation can be more reliable if based only on the perfect match than on the difference of perfect match (PM) and mismatch (MM) (8). We also found that the error variability from tiling arrays can be significantly reduced by using this strategy. Thus, we performed all the subsequent analyses based on PM probe intensities. Prior to the main analysis, all replicate arrays were normalized to a baseline array by interquartile and lowess normalization in order to make the baseline distributions of different tiling arrays comparable across all replicate arrays (9,10).

Sliding window analysis

The temporal replication behavior of neighboring probe pairs would be expected to be similar. However, there are several possible reasons why some neighboring probes have different replication enrichment patterns. First, a specific probe will tend to have a different binding affinity from its neighboring probes, so that it may show quite different enrichment patterns from its neighboring ones due to cross-hybridization. Second, at various potential break points in replication, e.g. a transition point from an exon region to the next intron region and, if it exists, a replication starting point, some adjacent probes may also exhibit dramatically different replication patterns. While the different replication patterns from the latter case will be consistently shown before and after such a break point, the heterogeneous expression patterns from the former case would be more sporadic and result in much noisier patterns than those without such inconsistent probes. Therefore, in order to minimize such noisy individual probe effects, we use a robust averaging method based on a sliding window with a fixed window width. We use a bandwidth of 10 kb (and 700 bp in Supplementary Result), which approximately corresponds to 10 min of replication. This bandwidth was found to be short and sensitive, yet reliable in detecting consistent temporal replication changes; some relevant issues and alternative approaches to this bandwidth selection and sliding window methods are discussed later. Our statistical testing is
then performed on these sliding windows to search for chromosome sequences that are differentially replicated in time during S phase of cell cycle.

Weighted ANOVA using LPE variance

One of the fundamental difficulties in analyzing the tiling array data is the small sample size; in this data set, triplicates were obtained for each time period. This results in extremely underpowered statistical testing framework for reliably identifying differentially expressed probes out of millions of tiling array probes. In order to overcome this limitation, we adopt the recent error-pooling technique called LPE, which pools the error information in each local intensity range, and consequently shrinks each error variance estimate toward the mean of hundreds of other probe intensity values in the local intensity region (7). This LPE method is quite well suited for analyzing tiling array data, given the cost of these experiments and limited starting sample. The error variability is found to be dependent on (and generally a function of) each probe’s mean intensity value, which can be accurately estimated by a large number of probes with similar intensity values. Therefore, as in Jain et al. (7), we estimate the baseline error distribution for each of the four time periods across the entire range of tiling array intensities. Based on these baseline error distributions for the variance estimates, we use the weighted ANOVA model for identifying the genomic positions that are differentially enriched over a given 2-h interval of labeling compared to other intervals. This weighted ANOVA approach is designed to interrogate tiling array data in two ways: (i) the errors of the same probe can be different among the four periods and/or (ii) those of multiple probes in the same sliding window can also be heterogeneous even within the same time period. In order to control for the multiple comparison error rate in our weighted ANOVA analysis, we evaluate the false discovery rate (FDR) to identify the probes with temporal differential expression patterns, as proposed in (11).

Identification of regions of differentially replicated probes

In order to identify the significant region of differentially replicated probes, we calculated the statistical significance for all stretches containing, e.g. 50 consecutive differentially replicated probes for their significance of ‘probe proximity’ compared to all spotted probes. For a stretch of \( n \) consecutive intergenic (or genic) differentially replicated probes along a chromosome, we want to assess its significance if the probes in the stretch have very close positions on the chromosome. Define \( X_i \) to be the normalized rank of probe \( i \) on the chromosome (i.e. rank of probe \( i \)/number of all spotted probes on the chromosome) and assume \( X_i \sim \text{Uniform} (0,1) \). Define \( Y = X_{(b)} - X_{(1)} \), where \( X_{(i)} \)'s are order statistics, as the normalized rank distance of this stretch. Note that all normalized rank distances are reduced to the order statistics of uniform distributions. Thus, we assess the ‘tightness’ by the rank distance of the probes on the two ends of the stretch of \( n \) probes relative to the positions of all the spotted probes on the chromosome. A direct calculation of this high-polynomial probability is rather prohibitive, so that a sampling-based significance analysis is performed as follows. For a given chromosome, let \( N_i \) and \( N \) be the number of differentially replicated and total number of tiled probes, respectively. Also, let \( X_i \) be the position of the \( i \)th probe on the chromosome, \( i = 1, 2, \ldots, N \), and let \( X_{(i)} \) be the \( i \)th ranked normalize position of differentially replicated probes. For a fixed-width window of size \( k \),

1. Calculate the observed range \( Y_{(1)} \), \ldots, \( Y_{(N-k+1)} \), \( Y_{(i)} = X_{(i+k)} - X_{(i)} \).
2. For \( b = 1, \ldots, B \) times,
   a. Sample random numbers \( U_{b1}, U_{b2}, \ldots, U_{bN} \) from Uniform (0,1)
   b. Obtain ordered scores \( U_{b(1)}, U_{b(2)}, \ldots, U_{b(N)} \)
   c. Compute the range \( Y_{(b)} = U_{b(i+k)} - U_{b(i)}, \ \ i = 1, 2, \ldots, N, k+1 \)
   d. Calculate \( P \)-value by counting statistics
      \[
      P_i = \frac{\sum_{b=1}^{M} I(Y_{(b)} \leq Y_{(i)})}{B}, \ \ i = 1, 2, \ldots, N - k + 1
      \]

Gene ontology analysis for differentially replicated genes

We further investigated the genes identified with temporal differential expression patterns for their functional roles and biological mechanisms. We collected and summarized the gene ontology terms for these genes’ biological process, molecular function and cellular components (12). We used GOSTat, software, which is based on Fisher’s exact test between the expected and observed frequencies of certain GO terms, to discover biological functions and molecular mechanisms that are statistically significantly over-represented among these genes (12). In this analysis, accounting for a restricted set of GO term categories on chromosomes 21 and 22, we derived the conditional probabilities \( P[B|A] = P[AB]/P[A] \) by separately obtaining \( P[A] \) and \( P[AB] \), where \( A = \{ \text{genes on chromosomes 21 or 22} \} \) and \( B = \{ \text{differentially replicated genes} \} \).

RESULTS

All the replicated tiling arrays in each time period were pooled and normalized based on the AM transformation on the PM probe intensities as described in the Methods section (Supplementary Figure 1). After normalization, the correlation coefficients among three replicate tiling arrays ranged from 0.82 to 0.93, compared to 0.65–0.87 before normalization. Using these normalized array data, LPE baseline error distributions were estimated for the four temporal phases of the tiling array experiment (Figure 1). These LPE baseline error distributions showed a non-increasing relationship with the intensity, and the four time periods exhibited quite different magnitudes of baseline error; note that these
LPE estimates were constant for $A < 13$, which were thresholded to avoid artificially low error variability in the $AM$ transformation plots. Considering this heterogeneity among the baseline error distributions of different intensity ranges and different time periods, we used weighted ANOVA modeling on 10-kb moving windows of tiling array probes, based on the LPE estimates as detailed in the Methods section. The window length of 10 kb is used because the average replication fork speed in human is $\sim 1$ kb/min, so that each window ($\sim 10$ min interval) would most likely fall in one of the four 2-h time periods. The results from a much shorter window length (700 bp) are summarized in Supplementary Results for comparison.

We were interested in discovering probes with significant differences among the four time periods of replication and tightly controlling numerous false positives from millions of tiling array probes. Differentially replicated probes were identified by controlling the false discovery rate (FDR) 5% level of the weight ANOVA F-statistics (13). The coordinate of each 10-kb sliding window represents the first probe location of each 10-kb window for simplicity; this may introduce a slight bias of location, but we found that it did not impact gene identification much at this fine resolution. We first investigated the relationship between the chromosome location of probes and replication time by examining the positions of differentially replicated probes and the frequencies of these probes. Figure 2 shows the histogram of the number of differentially replicated probes in coding and nucoding regions for both chromosomes 21 and 22, averaged over 500-kb intervals. The start and end of chromosomes seemed to have a much higher concentration of differential replicated probes for both chromosomes, and the number of these probes is larger near the centromere and telomere of q-arms than the remaining chromosome positions.

We summarized the numbers of all detected probes, including exon and intron probes, and the corresponding matched genes for varying FDR rates in Table 1A. For example, with FDR $= 0.05$, on chromosome 21 we found 113,841 probes that were differentially replicated during S phase (Table 1B). Among these sequences, 50,929 sequence probes were in exon regions, corresponding to 154 genes (out of 336 genes on chromosome 21) that were differentially replicated consistently with multiple probes representing these genes; the full list of these genes is provided in Supplementary Table 1; the remaining sequences were located within intergenic regions. For chromosome 22, 157,889 sequence probes were detected, among which 81,887 probes were matched to 256 genes (out of 688 genes on chromosome 22). The lengths and the numbers of tiling array probes of chromosomes 21 and 22 are similar, but chromosome 22 yielded almost two times more differentially replicated genes during S phase than
chromosome 21, which was somewhat expected since it contains twice as many known genes.

In order to find out more specific patterns of the probes for these differentially replicated genes, we closely examined four genes, randomly selecting each one from those replicated during one of the four time periods: HASF2BP (0–2 h), COL6A2 (2–4 h), PCNT2 (4–6 h) and ANKRD21 (6–8 h). Figure 3A shows that HASF2BP has the highest peak at the early replication time where each line in this figure shows a probe representing this gene. Most of these probes peak at 0–2 h. Figure 3B–D shows the replication patterns of the probes for the other three genes: COL6A2 showing the peak at time 2–4 h, PCNT2 at time 4–6 h and ANKRD21 at time 6–8 h. For convenience, we classify these genes as early (0–2 h), middle (2–4 h or 4–6 h) or late (6–8 h) replicated genes. As shown, even though their intensities differ, the sequence probes that represent the same gene exhibit quite similar replication

Figure 2. Frequency of differentially replicated probes on each 500-kb interval of chromosomes 21 and 22. The start and end parts of chromosomes have much higher concentration of differential replication for both chromosomes, and the number of these probes is larger near the centromere and telomere of q-arms than the remaining chromosome positions. Frequencies of differentially replicated probes in coding and noncoding regions on chromosomes 21 and 22.
patterns across the four time periods. We also examined the replication timing on the whole region of a particular gene BAGE3 (Figure 4A–F). In this figure, the whole sequence of BAGE3 was divided into six different consecutive regions, which showed homogeneous enrichment patterns of replication timing in each region. It shows that sequence probes from the same region tend to have quite similar replication patterns, while such patterns may slightly vary at different chromosome locations. For example, most regions show a consistent up-down-up pattern, but sequences in the middle of the gene (Figure 4D) show a constantly increasing pattern.

In Figure 5, the peak replication times (y-axis) of all differentially replicated genes were displayed by averaging over each gene’s multiple probes, together with the frequency of exon probes for comparison. We first found that even though some genes are nearby, their replication times can be quite different. Even the regions concentrated with high frequencies of exons exhibited similar patterns, showing quite different replication times between adjacent genes. These observations imply that...
The number of differentially replicated genes was determined by using interphase FISH (14). We classified these genes into three categories: (1) genes replicated in two or three consecutive time periods; (2) genes replicated in two different time periods and less replicated in 6–8 h compared to 0–2 h; and (3) genes replicated at two disjoint time periods, e.g. 0–2 h and 6–8 h. Thus, a large proportion of differentially replicated genes (44.6%) were replicated at two non-consecutive time periods, implying that such high-frequency regions were not concentrated at particular time periods, which again suggests the existence of highly dense replication origins.

We finally investigated whether the temporal replication was relevant to each gene's functions and molecular mechanisms. In order to assess over-represented functional categories of genes, we obtained Gene Ontology information of biological processes, molecular functions and cellular components for the 410 genes provided in Supplementary Table 1 that displayed the same replication pattern across each selected gene. We analyzed these genes using GOstat for evaluating statistical significance of overrepresented functional and molecular mechanisms (http://www.stat.wehi.edu.au) and conditional probability evaluation of such terms on chromosomes 21 and 22. GOstat simply derives the statistical significance between expected and observed functional categories based on the Fisher's exact test. Table 2 shows the list of these over-represented GO terms on chromosomes 21 and 22. Among the over-represented functional categories, lipid transport, glutathione biosynthesis and cyanate metabolism were significantly over-represented at FDR <0.05.

We further investigated the peak replication time points among adjacent neighboring genic and intergenic probes by averaging the frequencies at each 200-spotted probe window (Supplementary Figure S2). The question was whether many prolonged sequence regions of chromosomes were replicated (regardless of coding and noncoding regions) within the same time period. Thus, as detailed in the Methods section, we evaluated the statistical significance for proximity of 50 consecutive differentially replicated probes (or stretches) at each time period, based on the standardized ranks of these probes compared to the locations of all the probes. From this analysis, we found that statistically significant stretches (with FDR <0.05) were often observed at similar locations between coding and noncoding regions and among different time periods, implying that such high-frequency regions were not concentrated at particular time periods, which again suggests the existence of highly dense replication origins.

We identified three types of differentially replicated genes: (1) genes replicated in two or three consecutive time periods; (2) genes replicated at two non-consecutive time periods, e.g. 0–2 h and 6–8 h; and (3) genes replicated at two disjoint time periods, e.g. 0–2 h and 6–8 h. Thus, a large proportion of differentially replicated genes (44.6%) were replicated at two non-consecutive time periods, implying that such high-frequency regions were not concentrated at particular time periods, which again suggests the existence of highly dense replication origins.

We finally investigated whether the temporal replication was relevant to each gene's functions and molecular mechanisms. In order to assess over-represented functional categories of genes, we obtained Gene Ontology information of biological processes, molecular functions and cellular components for the 410 genes provided in Supplementary Table 1 that displayed the same replication pattern across each selected gene. We analyzed these genes using GOstat for evaluating statistical significance of overrepresented functional and molecular mechanisms (http://www.stat.wehi.edu.au) and conditional probability evaluation of such terms on chromosomes 21 and 22. GOstat simply derives the statistical significance between expected and observed functional categories based on the Fisher's exact test. Table 2 shows the list of these over-represented GO terms on chromosomes 21 and 22. Among the over-represented functional categories, lipid transport, glutathione biosynthesis and cyanate metabolism were significantly over-represented at FDR <0.05.

We further investigated the peak replication time points among adjacent neighboring genic and intergenic probes by averaging the frequencies at each 200-spotted probe window (Supplementary Figure S2). The question was whether many prolonged sequence regions of chromosomes were replicated (regardless of coding and noncoding regions) within the same time period. Thus, as detailed in the Methods section, we evaluated the statistical significance for proximity of 50 consecutive differentially replicated probes (or stretches) at each time period, based on the standardized ranks of these probes compared to the locations of all the probes. From this analysis, we found that statistically significant stretches (with FDR <0.05) were often observed at similar locations between coding and noncoding regions and among different time periods, implying that such high-frequency regions were not concentrated at particular time periods, which again suggests the existence of highly dense replication origins.

We finally investigated whether the temporal replication was relevant to each gene's functions and molecular mechanisms. In order to assess over-represented functional categories of genes, we obtained Gene Ontology information of biological processes, molecular functions and cellular components for the 410 genes provided in Supplementary Table 1 that displayed the same replication pattern across each selected gene. We analyzed these genes using GOstat for evaluating statistical significance of overrepresented functional and molecular mechanisms (http://www.stat.wehi.edu.au) and conditional probability evaluation of such terms on chromosomes 21 and 22. GOstat simply derives the statistical significance between expected and observed functional categories based on the Fisher's exact test. Table 2 shows the list of these over-represented GO terms on chromosomes 21 and 22. Among the over-represented functional categories, lipid transport, glutathione biosynthesis and cyanate metabolism were significantly over-represented at FDR <0.05.

We further investigated the peak replication time points among adjacent neighboring genic and intergenic probes by averaging the frequencies at each 200-spotted probe window (Supplementary Figure S2). The question was whether many prolonged sequence regions of
organic solvent but not, or sparingly, in an aqueous solvent. Glutathione biosynthesis means that the formation from simpler components of glutathione, the tripeptide glutamylcysteinylglycine, which acts as a coenzyme for some enzymes and as an antioxidant in the protection of sulfhydryl groups in enzymes and other proteins. Cyanate metabolism is the chemical reaction involving cyanate, NCO–, the anion of cyanic acid. One-carbon-compound catabolism is the chemical reaction involving compounds containing a single carbon atom. Table 2B shows significantly over-represented GO terms in molecular function (after accounting for a restricted set of categories on chromosomes 21 and 22), such as hydrolase activity, transferase activity and

Figure 3. Replication patterns for four genes: HASF2BP, COL6A2, PCNT2 and ANKRD21, with differential replication in time. These genes are called as early (0–2 h), middle (2–6 h) or late (6–8 h) replicated genes. For example, Figure 3A shows that HASF2BP has the highest peak at early replication time, where each line in this figure represents a sequence probe for this gene.
receptor binding. Hydrolase activity in cyclic amidines is catalysis of the hydrolysis of any non-peptide carbon–nitrogen bond in a cyclic amide. Transferase is the systematic name for any enzyme of EC class 2 and a catalysis of the transfer of a group, e.g. a methyl group, glycosyl group, acyl group, phosphorus-containing or other groups, from one compound (generally regarded as the donor) to another (generally regarded as the acceptor). We found that there was no significantly over-represented GO term in the cellular component.

The results on the tiling array data obtained by our error-pooling analysis approach would benefit from further biological validation. However, due to the limited availability of biological materials, it was prohibitive to confirm the observations made from our tiling array data analysis by conventional methods such as quantitative RT-PCR (qPCR). Instead, we utilized cDNA microarray data described earlier, which were independently, but simultaneously, obtained for the four time periods of S phase along with the tiling array data. There were 41 cDNA clones (21 for chromosome 21 and 20 for chromosome 22, respectively, among the 1589 clones that were spotted twice on the cDNA array across the entire human genome) that met statistical quality control criterion, i.e. internal consistency between the duplicated clones on the cDNA arrays. This cDNA array experiment would thus be conceptually and effectively equivalent to performing independent qPCR confirmation experiments for the tiling array probes representing the corresponding 41 genes. Of these 41 clones, we found that 29 clones (15 for chromosome 21 and 14 for chr 22) showed concordant (i.e. exactly the same or adjacent) replication times with their corresponding tiling array genes (Figure 6; P-value = 0.011 against the random chance of such concordant replication times; from the binomial test with success probability ½). In fact, 19 cDNA clones showed the exact same replication times between the tiling and cDNA microarrays (P-value = 0.003 from the binomial test with success probability ¼). Note that the replication times from the cDNA arrays are directly inferred by examining each individual cDNA clone’s replication patterns on the cDNA array data, not using any of our proposed analysis approaches to the tiling array data. Thus, while the noise level may be considerable in both the cDNA and tiling array

Figure 4. Replication patterns of gene BAGE3 divided into six different consecutive regions with tightly homogeneous expression patterns for the replication times. Sequence probes of BAGE3 from the same gene seem to have homogeneous replication patterns with minor variation on their physical positions.

A. 10,004,271–10,045,549 bp  B. 10,045,550–10,045,943 bp  C. 10,045,944–10,049,035 bp
D. 10,049,036–10,049,332 bp  E. 10,049,333–10,062,755 bp  F. 10062756–10,063,007 bp
Figure 5. Replication timing and exon density of differentially replicated probes on (A) chromosome 21 and (B) chromosome 22 for the entire time period of S phase. Replication period (y-axis) averaging over multiple probes of each of the genes with differential temporal expression along the chromosomal position (x-axis) was shown compared to the frequency of exon probes at the same position. Even though some genes are nearby, their replication times seem to be quite different, and the regions concentrated with exon exhibited little or no difference compared to other regions.
Table 2. Overrepresented GO terms of the genes with temporal differential replication on chromosome 21 and 22 from 700-bp window analysis

A. Over-represented biological process of differentially replicated genes

| Best go                        | Gene symbols | Number of annotated gene/total gene | FDR     |
|-------------------------------|--------------|-------------------------------------|---------|
| Lipid transport               | ABCG1 APOL4 APOL6 APOL2 APOL5 OSBP2 APOL1 | 8/84                              | 0.0611  |
| Glutathione biosynthesis      | GGT1 GGTLA1 GGT2 | 3/8                                  | 0.0611  |
| Cyanate metabolism            | TST MPST     | 2/2                                  | 0.0611  |
| Cyanate catabolism            | TST MPST     | 2/2                                  | 0.0611  |
| One-carbon-compound catabolism| TST MPST     | 2/2                                  | 0.0611  |

B. Over-represented molecular functions of differentially replicated genes

| Best go                        | Gene symbols | Number of annotated gene/total gene | Gostat FDR |
|-------------------------------|--------------|-------------------------------------|------------|
| Hydrolase activity, acting on carbon-nitrogen (but not peptide) bonds, in cyclic amidines | APOBEC3G ARP10 APOBEC3F Q5IFJ4 Q8NFD1 | 7/41 | 0.00604 |
| Structural constituent of eye lens | APOBEC3C APOBEC3A | 5/19 | 0.00618 |
| Gamma-glutamyl transferase activity | CRYA ACRYB3 CRYB4 CRYB1 CRYBB2 | 5/22 | 0.00889 |
| Glucocorticoid receptor binding | YWHAH NRIP1 | 2/2 | 0.0317 |
| Oncostatin-M receptor binding  | LIF OSM      | 2/2 | 0.0317 |
| Transferase activity, transferring amino-acyl groups as acceptor | GGT1 GGTLA1 Q6ISH0 Q5NV76 GGT2 | 5/33 | 0.0317 |
| Oxidoreductase activity, acting on superoxide radicals as acceptor | SOD1 KIAA0179 D21S2056E | 3/9 | 0.0317 |
| Superoxide dismutase activity  | SOD1 KIAA0179 D21S2056E | 3/9 | 0.0317 |
| Hydrolase activity, acting on carbon-nitrogen (but not peptide) bonds | APOBEC3G Q8NFD1 ARP10 APOBEC3F Q5IFJ4 UPB1 | 9/132 | 0.0361 |
| Protein carrier activity       | APOBEC3C HDAC10 APOBEC3A | | |
| Carbonyl reductase (NADPH) activity | Q6ICM2 SEC14L2 SEC14L3 SEC14L4 | 4/24 | 0.0361 |
| Thiosulfate sulfuroxtransferase activity | TST MPST | 2/3 | 0.0361 |
| Hematopoietin/interferon-class (D200-domain) cytokine receptor activity | IFNA1 CSF2RB IL2RB ENSP00000343289 IL17R | 7/85 | 0.038 |
Figure 6. Concordant replication timing between tiling and cDNA array data. The majority of matched pairs of tiling array probe and cDNA clone showed concordant replication times: 29 (15 on chromosome 21 and 14 on chromosome 22) of 41 pairs with exact or adjacent time periods and 19 pairs (10 on chromosome 21 and 9 on chromosome 22) had the exact same replication times.
experiments, the tiling array replication times obtained by our error-pooling moving-window approach were well matched and effectively confirmed by those of the independent cDNA array study.

**DISCUSSION**

The statistical analysis of tiling array data is challenging due to the extremely large number of individual probe sequences, much higher noise level (than standard expression arrays) and the limited number of replicates. Unfortunately, many classical statistical methods assume homogeneity of variance and/or a relatively large sample size for their maximal performance, so that their statistical inference is severely underpowered and biased for tiling data analysis. The error variances of tiling array probes often greatly differ among different probes, yet are quite dependent on the underlying mean intensities of individual probes. An error-pooling method was thus critical and important for accurately estimating technical error variability of tiling array data with a small number of replicates across the entire intensity range. Taking this into account, the proposed LPE approach dramatically improved the accuracy of tiling array error estimates, which resulted in considerably higher statistical power in tiling array data analysis (7). We also introduced a weighted ANOVA analysis on 10-kb sliding windows based on the LPE baseline error distributions. Our sliding-window approach further stabilized and reduced highly variable individual probe effects for our error estimation.

Based on our improved analysis approach, we found many novel observations in replication timing on human chromosomes with a high resolution. We first identified 26% probes as being differently replicated with a tight statistical cutoff criterion (5% FDR), which is somewhat smaller than the original study reported (~30–35%). Among these differentially replicated probes, 47% probes were from coding regions, which then matched 410 known genes. We found that a much higher proportion of generic probes (25.0%) were differentially replicated than intergenic probes (18.6%) on chromosome 21, suggesting a gene-centric view of replication timing as opposed to one based on physical location along a chromosome. More specifically, we observed that early replicating regions tended to be associated with gene-dense regions and late replicating sequences were found in relatively gene-poor loci similar to Jeon et al. Assuming that (1) chromatin tends to be open in gene-dense regions and late replicating sequences were found in relatively gene-poor loci similar to Jeon et al. Assuming that (1) chromatin tends to be open in gene-dense regions and in a compacted state (i.e. heterochromatin) in gene-poor regions upon entry of S phase and (2) the replication machinery cannot access DNA in heterochromatic regions, many have formulated a chromatin-centric model of replication dynamics. Finally, we note that these two views are not necessarily mutually exclusive, given that nucleosome depletion tends to occur just 5' of actively transcribed genes and that typical replication fork rates are ~1 kb/min. It is somewhat surprising that we found that replication timing seems to be relevant to certain molecular functions and biological processes. Furthermore, even though the number of mutually validated genes between the tiling and cDNA microarrays was relatively small (and majority of these validated genes were sparsely distributed), we found that some of these were very closely located, e.g. no other gene exists between two consecutive genes for the four areas highlighted in Supplementary Table S3, yet showed quite different replication times. This small number of mutually confirming, switch-over cases cannot yet prove the existence of a large number of replication break points across the entire human genome. However, considering that >70% of tiling array genes' replication times obtained by our proposed approach were consistent with and effectively confirmed by the cDNA array data (29/41), we believe that our observations on the frequent changes of replication times between adjacent genes should apply to the rest of the human genome.

We used a sliding window strategy to reduce the noise level of individual sequence probes. As pointed by (5), the determination of a suitable window size is constrained by two factors: the spacing of the probes along the chromosomes and the characteristic size of the functional elements being assayed (e.g. exon for an mRNA assay) on these chromosomes. One additional fact to be mindful of when using a sliding-window-based approach is that although the false-positive rate was reduced in identifying the sites of transcription, the signal from the probe pairs was smoothed, making strict determination of the transcription boundaries problematic (5). We experimented with a few different strategies, including varying the window width and using disjoint windows. The results were generally consistent if the window width exceeded a certain minimum. For example, we analyzed the data with a much shorter window size, 700 bp, in order to identify finer replication patterns and check the consistency of the results. This analysis resulted in a much smaller number of significantly differentially replicated probes and corresponding genes—6524 probes and 178 genes (64 genes in chromosome 21 and 114 genes in chromosome 22; see Supplementary Results). This was expected since longer windows result in greater power to detect long stretches of significant replication differences. Nevertheless, the main conclusions were quite consistent with those from the 10-kb window analysis. Twice as many genes were identified on chromosome 22 than chromosome 21; significant exon and intron probes were quite evenly distributed across the four time periods, as shown in Figures 2–5 for the 10-kb window case.

A rigorous sampling-based method was applied to evaluate statistical significance for the stretch proximity of 50 consecutive differentially replicated probes at each time period based on 10,000 random samples. From this analysis, we found that these stretches of replication are more or less evenly distributed during the four time periods and not concentrated in any local region at a particular time period. This appears to contradict the existence of replication break points, with a considerable length between two adjacent break points. This may suggest that such break points may exist at much finer scales than has been assayed. We believe these results provide a novel insight into the replication of genes, which awaits further confirmation.
The use of GO analysis may provide useful functional insights regarding the replication timing of particular genes during the cell cycle. We detected some over-represented ontologies for differentially replicated genes by using GOstat after accounting for the biased representation of such GO terms on chromosomes 21 and 22. Genes with certain molecular functions such as hydrolase activity, transferase activity and receptor binding were found to show differential temporal expression patterns and tend to replicate at particular time periods. In the biological process category, lipid transport, glutathione biosynthesis and cyanate metabolism were over-represented with marginal significance. On the contrary, no significantly over-represented cellular component was found.

In this study, we did not directly compare the performance of our approach with other approaches for the following reasons. First, the results from a simulation study, which may be one of the standard ways of comparison, may heavily depend on its simulation setting, but no realistic setting was seemingly appropriate for comparing different approaches objectively on the tiling array data, especially since our error-pooling approach, which assumes different probes with similar mean intensity values to have a similar technical error variance, is conceptually different from other ‘within-probe-error’-based approaches. Second, the improvement of LPE error estimation was demonstrated in (7), in which a dramatically higher statistical power of LPE estimation was shown than that of other within-probe (or within-gene) approaches when the number of replicated arrays was less than five. Also, evaluated by FDR, which is a recent statistical significance concept that simultaneously controls false positives and false negatives in a large-screening microarray data analysis, we found that our approach tightly controlled both false-positive and false-negative errors; for example, as shown in Table 1, our weighted ANOVA approach provided many significant probes with FDR <0.05. Finally, as shown in the results, the probes significantly identified by our approach were quite biologically consistent, e.g. many adjacent probes and the probes of the same gene tend to share the same statistical significance level and replication patterns. From these observations, we believe that our approach has significantly improved tiling array data analysis.

There still exist several remaining issues regarding our tiling array data analysis on cell cycle replication. For example, the current normalization across all the tiling arrays was performed based on the assumption that the interquartile range of each array is the same. While this may be a conservative assumption in terms of temporal differential replication discovery, this may not be true if one particular time period is more active in replication than the others; this should be more carefully evaluated biologically in order to avoid a biased identification of differential replication patterns. We also did not correct for the large differences in the hybridization affinity of tiling array probes. Despite these shortcomings, we believe that our current result demonstrates that a series of improved statistical analysis methods can yield novel insights at effectively higher resolution from genomic tiling array data.

ACKNOWLEDGEMENTS

We greatly thank Dr Anindya Dutta for providing many helpful comments and feedback on the analysis of the tiling array data of human chromosome replication and on the preparation of this manuscript and Dr Yesu Jeon for providing us the cDNA microarray data. This study is supported in part by NIH Grant 1R01HL081690 for J.K.L. and by the National Research Laboratory Program of Korea Science and Engineering Foundation (M10500000126) for T.P. Funding to pay the Open Access publication charge was provided by Funding to pay the Open Access publication charges for this article was provided by the grant for the project of the National Institutes of Health.

Conflict of interest statement. None declared.

REFERENCES

1. Woodfine,K. et al (2005) Replication timing of human chromosome 6. Cell Cycle, 4, 172–176.
2. White,E.J. et al (2004) DNA replication-timing analysis of human chromosome 22 at high resolution and different developmental states. Proc. Natl. Acad. Sci. USA, 101, 17771–17776.
3. Jeon,Y. et al (2005) Temporal profile of replication of human chromosomes. Proc. Natl. Acad. Sci. USA, 102, 6419–6424.
4. Li,L. et al (2005) Tiling microarray analysis of rice chromosome 10 to identify the transcripitome and relate its expression to chromosome architecture. Genome Biol., 6, 52.
5. Kampa,D. et al (2004) Novel RNAs identified from an in-depth analysis of the transcriptome of human chromosomes 21 and 22. Genome Res., 14, 331–342.
6. Ji,H. and Wong,W.H. (2005) TileMap: create chromosomes map of tiling array hybridizations. Bioinformatics, 21, 3629–3636.
7. Jain,N. et al (2003) Local-pooled-error test for identifying differentially expressed genes with a small number of replicated microarrays. Bioinformatics, 19, 1945–1951.
8. Irizarry,R.A. et al (2003) Summaries of affymetrix genechip probe level data. Nucleic Acids Res., 31, e15.
9. Dean,N. and Raftery,A.E. (2005) Normal uniform mixture differential gene expression detection for cDNA microarrays. BMC Bioinformatics, 6, 173.
10. Cho,H. and Lee,J.K. (2004) Bayesian hierarchical error model for analysis of gene expression data. Bioinformatics, 20, 2016–2025.
11. Storey,J.D. and Tibshirani,R. (2003) Statistical significance for genomewide studies. Proc. Natl. Acad. Sci. USA, 100, 9440–9445.
12. Beissbarth,T. and Speed,T.P. (2004) GOstat: find statistically overrepresented Gene Ontologies within a group of genes. Bioinformatics, 20, 1464–1465.
13. Storey,J.D. and Tibshirani,R. (2003) Statistical methods for identifying differentially expressed genes in DNA microarrays. Methods Mol. Biol., 234, 149–157.
14. Simon,I. et al (1999) Asynchronous replication of imprinted genes is established in the gametes and maintained during development. Nature, 401, 929–932.