AWARD NUMBER: W81XWH-10-1-0463

TITLE: Origins of DNA Replication and Amplification in the Breast Cancer Genome

PRINCIPAL INVESTIGATOR:
Susan A. Gerbi, Ph.D.

CONTRACTING ORGANIZATION:
Brown University

Providence, RI 02912

REPORT DATE: September 2012

TYPE OF REPORT: Final

PREPARED FOR:

U.S. Army Medical Research and Materiel Command
Fort Detrick, MD 21702

DISTRIBUTION STATEMENT:

Approved for public release; distribution unlimited

The views, opinions and/or findings contained in this report are those of the author(s) and should not be construed as an official Department of the Army position, policy or decision unless so designated by other documentation.

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. REPORT DATE
1 September 2012

2. REPORT TYPE
Final

3. DATES COVERED
1 Sep 2010 – 31 Aug 2012

4. TITLE AND SUBTITLE

Origins of DNA Replication and Amplification in the Breast Cancer Genome

5a. CONTRACT NUMBER

5b. GRANT NUMBER
W81XWH-10-1-0463

5c. PROGRAM ELEMENT NUMBER

6. AUTHOR(S)

Susan A. Gerbi (P.I.); Alexander Brodsky (co-P.I.); Ben Raphael (co-P.I.)

Email: Susan_Gerbi@Brown.edu

5d. PROJECT NUMBER

5e. TASK NUMBER

5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
Brown University
Providence, RI 02912

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES)
U.S. Army Medical Research and Materiel Command
Fort Detrick, Maryland 21702-5012

10. SPONSOR/MONITOR'S ACRONYM(S)

11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION / AVAILABILITY STATEMENT
Approved for Public Release; Distribution Unlimited

13. SUPPLEMENTARY NOTES

14. ABSTRACT

The major goal of this IDEA Expansion Breast Cancer Research grant is to test whether a correlation existsbetween sites of DNA amplification and estrogen receptor (ER) binding in the breast cancer genome. Correlations would support our hypothesis that ER adjacent to replication origins may interact with the replication machinery to drive DNA amplification, a hallmark of many cancers. Our Specific Aims are: (1) Map replication origins in the MCF-7 breast cancer genome by genomic sequencing (2) Compare the replication origin maps between breast cancer (ER+, ER-) and normal breast cells (3) Correlate the origin map data with sites of DNA amplification and estrogen receptor binding (4) Pilot runs to map replication origins in ER+ human breast cancer tissue and sites of DNA amplification

15. SUBJECT TERMS

estrogen receptor, DNA amplification, replication origins

16. SECURITY CLASSIFICATION OF:			**17. LIMITATION OF ABSTRACT**	**18. NUMBER OF PAGES**	**19a. NAME OF RESPONSIBLE PERSON** USAMRMC
a. REPORT U	**b. ABSTRACT** U	**c. THIS PAGE** U	UU	43	**19b. TELEPHONE NUMBER** *(include area code)*

Table of Contents

Introduction

Gene amplification, reflecting genomic instability, is a hallmark of many cancers. It leads to aggressive growth of breast cancer cells. Therefore, it is desirable to prevent gene amplification. The problem is that the molecular basis for the trigger of gene amplification is unknown. Our current research in a model system suggests that the trigger may be a transcription factor, namely the receptor for the steroid hormone estrogen. The research proposed here describes some experiments to test this idea. Are there novel binding site(s) for the estrogen receptor (ER) near a replication origin that cause re-replication, resulting in amplification? As a first step to address this question, we need to map all origins of replication in the human (breast cancer) genome --- which is the **subject** of this DOD-funded grant. The **purpose and goal** of these experiments is to be able to see if correlations exist between origins that re-replicate (leading to DNA amplification) in breast cancer cells and sites of estrogen receptor binding. The **scope** of our research is summarized by the three Specific Aims: (1) Map replication origins in the human genome; (2) Comparison of replication origin maps between breast cancer (ER+, ER-) and normal breast cells; (3) Correlation of origin map data with sites of (a) DNA amplification and (b) estrogen receptor binding. The results from the proposed experiments will serve as the foundation for comparable experiments in surgically derived breast cancer tissue. These experiments are beyond the time frame of this grant, but we are already stockpiling tissue samples for these future experiments. Our proposed study could provide a new paradigm for hormonal induction of breast cancer via gene amplification, leading to new methods of diagnosis and treatment.

Body: Progress Report (year two)

As described in the DOD funded parent grant, to test our hypothesis we need to map origins of DNA replication in the genome and ask which of these coincide with sites of DNA amplification and with ER binding sites. In the parent grant we proposed to identify potential replication origins by mapping ORC binding sites. However, ORC can also bind to silent origins. We proposed to refine this strategy by using newly synthesized (nascent) DNA for direct mapping of all active replication origins in the genome. These data will then be compared to sites of DNA amplification and sites of ER binding to see if a correlation exists. In year one, as reported in last year's progress report, we accomplished the following:

Task (1) Map replication origins in the human genome.

Subtask (1a) (months 1-6) – preparation of short nascent strands from MCF-7 cells – completed.

We developed methodology using MCF-7 breast cancer cells to derive a genomic map of replication origins by Illumina sequencing of short nascent strands. Postdoc Michael Foulk in the Gerbi lab prepared nascent strands from MCF-7 cells as follows:

Nascent-strand-Seq Work Flow (Foulk method)

- Prepare genomic DNA from asynchronous MCF-7 cells with DNAzol (~30% in S-phase)
- Purify Replicative Intermediate (RI) DNA on BND-cellulose (100 ug input)
- Phosphorylate ends with T4-polynucleotide kinase
- Enrich for Nascent Strands by digesting with lambda-exonuclease (λ-exonuclease = Lexo)
- Size select for 1 kb-2 kb nascent strands on low melting point agarose gel to eliminate Okazaki fragments
- Test enrichment of the MYC origin of replication by Real-Time PCR

This was the method we used for the MCF-7 nascent strands used for sequencing on the Illumina GAIIX platform. Subsequently, the same DNA preparation was used for the first run on the Illumina

Hi-Seq platform. Real time PCR revealed that this nascent strand DNA preparation had 45-fold enrichment when tested for the Myc origin of DNA replication. Later nascent strand DNA preparations at the end of year one and in year two of DOD support followed the method of Cadoret et al. (2008; personal communication) that replaces the BND-cellulose column with a sucrose gradient; this method gave even higher levels of origin enrichment. The steps for the Cadoret method are summarized below:

Nascent-strand-Seq Work Flow (Cadoret method)

- Prepare genomic DNA from asynchronous MCF-7 cells with DNAzol (~30% in S-phase)
- Heat denature the DNA (95^0C for 5 min) and run on a 5-30% sucrose gradient; select the fractions with 300-2500 nt DNA.
- Phosphorylate ends with T4-polynucleotide kinase
- Enrich for Nascent Strands by digesting with lambda-exonuclease (λ-exonuclease = Lexo)
- Size select for 1 kb-2 kb nascent strands on low melting point agarose gel to eliminate Okazaki fragments
- Test enrichment of the MYC origin of replication by Real-Time PCR

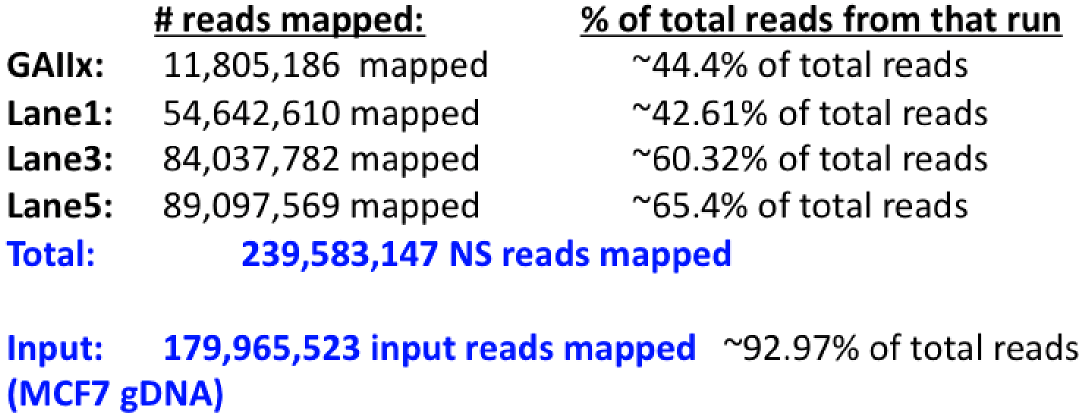
With either method, after enrichment of the MCF-7 nascent strand DNA, it was then subjected to DNA sequencing as follows:

- Subject nascent strands to fragmentation followed by making them double stranded with random primers and Klenow
- Standard library preparation for Illumina sequencing (200-500bp fragments)
- Sequence library on the Illumina GAIIx platform (pilot experiment) or High-Seq (subsequent experiments) using 42 bp single end reads
- Filter and align reads to the human genome (build hg18) and call peaks (using genomic input DNA as normalization control)

Several labs are now using the methodology we developed for nascent strand sequencing (NS-Seq) whose basis resides in the use of λ-exonuclease to enrich nascent strands coupled with size selection. Our NS-Seq protocol is based on our earlier report (Gerbi and Bielinsky, 1997; Bielinsky and Gerbi, 1998) that nascent DNA is resistant to lambda-exonuclease digestion because of the presence of a 5’ RNA primer. This allows the parental DNA to be digested while the nascent DNA is untouched. The nascent strands were size selected on gels for 1-2 kb, which gave greater origin enrichment than a 0.5-1 kb fraction that may have Okazaki fragment contamination. Using the c-Myc origin to assess for enrichment, the average of the several preparations used for sequencing had 45-fold or greater enrichment of nascent strands. Interestingly, in assessing this enrichment at the c-Myc origin, we discovered that the preferred origin resided in the second exon of the gene while it was previously determined to reside in the promoter of the gene (in HeLa cells: Tao et al., 2000). This observation was confirmed in our NS-Seq data, suggesting plasticity of origin usage at the c-Myc gene in different cell types.

Sub-task (1b) (months 7-8) – sequencing and analysis of results to map replication origins in the MCF-7 human breast cancer cell genome.

As reported last year, we switched from Helicos sequencing to Illumina sequencing. Last year we reported our first run of MCF-7 nascent strand DNA on the Illumina GAIIx machine. Since that time, in the current year we have used the more powerful Illumina Hi-Seq2000 machine as the platform for sequencing nascent strands. The new results confirm and extend our earlier data from the Illumina GAIIx machine. Of the three samples run this year on the Illumina Hi-Seq2000, the first sample was derived from the same material that had been used for the Illumina GAIIx machine. The other two samples were from different preparations (using the Cadoret sucrose gradient method – see above) of nascent strands, providing biological replicates.



	# reads mapped:	% of total reads from that run
GAIIx:	11,805,186 mapped	~44.4% of total reads
Lane1:	54,642,610 mapped	~42.61% of total reads
Lane3:	84,037,782 mapped	~60.32% of total reads
Lane5:	89,097,569 mapped	~65.4% of total reads
Total:	**239,583,147 NS reads mapped**	
Input: (MCF7 gDNA)	**179,965,523 input reads mapped**	~92.97% of total reads

We used BEDTools (Quinlan et al., 2010) and features of the genomic analysis of ChIP-Seq data (Euskirchen et al., 2007) for analysis of our data on DNA replication origin in the human genome. The reads of 42 bp listed in the table above were mapped to human genome build hg18 with Bowtie (Langmead et al, 2009; Langmead 2010). An input control was also sequenced and mapped to the genome to reduce the number of false positive results (This was not done previously for our pilot run on the Illumina GAIIx machine). The input control we have used is genomic DNA that has not been enriched for short nascent strands. Specifically, for MCF-7 genomic DNA (gDNA) we used estrogen starved G0 cells and for the MCF-10A gDNA (described below in Task 2) we just used asynchronous cells. In brief, total gDNA was isolated from the cells, sheared by sonication to a size range of about 100-600 bp and the entire prep was taken through the Illumina library preparation. In the end, the gDNA library was size selected on 2% agarose for 200-500 bp fragments which were then sequenced.

Once treatment and control reads are mapped to the genome, statistically significant peaks, which are areas where treatment reads pile-up high over the control reads, are called with a peak caller. The 239 million mappable reads (see table above) were combined into a single file. This file was used to call peaks against the input reads with MACS. As MACS was designed for ChIP-seq experiments, we spent considerable time optimizing the parameters for Nascent Strand sequencing (NS-seq).

Last year, using only one lane of sequencing from the Illumina GAIIx which produced ~11.8 million mappable reads (see table above), we called 54,100 peaks (53,914 peaks after removing chrY peaks, which are artifacts in the MCF-7 context). Since we did not have an input control for that pilot experiment, some of the peaks were probably false positives. Moreover, we did not reach saturation, which required more sequencing. Finally, to be rigorous we needed to include sequence reads from biological replicates. To reach saturation and to include biological replicates, in year two of DOD support, three lanes on the Illumina Hi-Seq2000 platform were used to sequence 3 biological replicates of short nascent strand preparations from MCF-7 cells. The Hi-Seq2000 data gave 54.6 million, 84 million, and 89 million mappable reads for the three different lanes (see table above). Including our previous 11.9 million reads, this gave 239.6 million mappable reads. We also obtained ~180 million mappable input control reads (MCF-7 gDNA).

In last year's progress report, we presented our analysis of the 53,914 peaks from the pilot GAIIx sequencing, including median and mean peak widths and inter-origin distances. We also presented the data demonstrating good congruence between our reads and some known origins, including c-Myc, DBF4, DHFR, β-Globin, RPE, as well as Lamin B2, and Glucose-6-Phosphate Dehydrogenase. There have been a few other reports of mapping origins in the human genome but, surprisingly, we found that the overlap in origins was less than 40% at best between the datasets from the various labs (Cadoret et al., 2008; Karnani et al., 2010.; Mesner et al. 2011; Martin et al. 2011, Valenzuela et al. 2011). Although some of these other reports only mapped origins to 1% of the HeLa genome (ENCODE project), the poor overlap of origins was especially surprising when comparing our data to that of Martin et al. (2011) who has also performed whole genome NS-Seq on MCF-7 cells. They used smaller nascent strands than us and we suspect that may have led to Okazaki fragment contamination in their samples. Indeed, their nascent strand enrichment was less than ours and not even reported in their paper nor were their results validated in their paper. We speculated that lack of saturation may be an explanation for poor congruence. However, now we suspect that there may be some technical explanations (described below). Very recently, a paper appeared by Besnard et al. (2012) that mapped origins in the human genome. Remarkably, they found more than twice as many origins in the human genome as discovered by us or Martin et al. (2011). This led us into experiments that are still ongoing to try and unscramble the confusion in the field on the number and location of replication origins in the human genome. We now suspect problems arising from the technique of nascent strand DNA preparation and also from analysis of the sequencing data as the explanation for the poor congruence between origin datasets from the different groups. Our experiments and analysis on this issue are described below. Once these experiments are completed, we anticipate writing two papers for publication: one on identification of the technical reasons for lack of congruence and our analysis of the true number of replication origins in the MCF-7 human breast cancer genome, and a second paper describing methods at the lab bench and for computational analysis for NS-Seq experiments. In addition, we also plan in the coming year to publish a validation of NS-Seq methods and analysis by applying it to genome-wide mapping of replication origins in the genome of the budding yeast, *Saccharomyces cerevisiae*, where the origins have already been mapped by a variety of other approaches.

Technical issues in NS-Seq:

It has been reported that there is a base composition skew in metazoan replication origins and that they are GC-rich (Cayrou et al. 2011). Moreover, the origin G-rich repeated elements (OGREs) predict G-quadruplex structures at origins (Cayrou et al., 2012). The recent paper by Besnard et al. (2012) claimed that a consensus G-quadruplex-forming DNA motif can predict the position of replication origins in the human genome. Although this might be correct, we realized that it also might be a technical artifact. Lambda exonuclease (Lexo) is used in NS-Seq protocols to enrich nascent strands by virtue of their resistant to Lexo degradation because of the 5' RNA primer that protects them. If G-quadruplex DNA is also resistant to Lexo digestion, then G-quadruplex DNA will also be present in the nascent strand enriched sample after Lexo digestion, thus contaminating the sample. Postdoc Michael Foulk in our lab has begun some experiments to explore this possibility, and his preliminary data (**Supporting Data Figure 1**) support the likely contamination of nascent strand DNA preparations by G-quadruplex DNA. He cut the plasmid pFRT.Myc with the restriction enzyme BglII to linearize the plasmid or with KpnI plus NotI to release a 1 kb fragment containing the G-quadruplex predicted to reside at the Myc origin. Samples were run on a gel with or without Lexo digestion. Moreover, in each case, one aliquot of each sample was native double-stranded DNA (unboiled) and another aliquot was boiled and placed on ice to favor intramolecular formation of the G-quadruplex. These latter samples that had been boiled revealed products that would be predicted if Lexo stopped digesting the DNA when it came to the predicted position of a G-quadruplex roadblock (2500 bp and 4600 bp fragments after Lexo digestion of the boiled sample of BglIII digested DNA; moreover, the 1 kb fragment released by KpnI and NotI is reduced to 850 bp and 140 bp after Lexo digestion of the boiled sample). In the next few weeks, we plan a few more experiments to confirm this preliminary finding. The correct way to combat this problem for NS-seq will then be to sequence genomic (non-replicating) DNA that is enriched for G-quadruplex (by boiling and quenching on ice followed by Lexo digestion) to map these structures in the MCF-7 genome to analyze if they are coincident or not with the replication origins already mapped from enriched nascent strand preparations. Another cautionary note is that Cayrou et al. (2011, 2012) boils her replicating DNA for 15 minutes before sucrose gradient centrifugation, and she uses pH 9.4 (rather than pH 8.8 recommended and used by us) for Lexo digestion. These harsh treatments could degrade RNA primers on the nascent DNA, making it even more likely that what she has sequenced is simply G-quadruplex DNA rather than nascent strands to identify replication origins. Similar concerns apply to the experiments of Martin et al. (2011) and Besnard et al. (2012).

Computational issues in NS-Seq analysis:

Most of the peak-caller software used for computational analysis of NS-Seq data was developed for analysis of ChIP-Seq data and has to be optimized for NS-Seq applications. In our case, we have used MACS and the optimization of the variables as done by graduate student John Urban in our lab is described below. The NS-Seq data of Martin et al. (2011) appears to have used a variant of MACS. In contrast, the NS-Seq data of Besnard et al. (2012) was analyzed with Sole-Search which they did not optimize and simply used the default parameters. John Urban in our lab has computed the effect of using MACS as compared to Sole-Search for analysis of our data, and the result is very striking, as described later in this progress report (see Task 2).

Optimizing the variables in MACS

(A) Redundant Reads - John has compared the effect of keeping just 1 (K1) or 3 copies (K3) of redundant reads and discarding the remaining redundant reads (usually thought to be an artifact such as from PCR). As shown in **Supporting Data Figure 2**, as expected, somewhat more reads are kept in K3 than K1 for both the experimental sample of the combined nascent strand reads as well as for the nonreplicating genomic DNA input used to correct for background. The K1 and K3 options were used for further analyses and little differences were found between them.

(B) Normalizing number of reads between the treatment (NS-Seq) and the control (input) –

MACS will either scale the counts toward the larger file or to the smaller file. For the current dataset, the larger of the two files is always the treatment (NS-seq) – see table above. This means that scaling to the large file adjusts the read counts in the input file and that scaling to the small file adjusts the read counts in the NS file. Henceforth, we will use the term “toLg” mean that the ‘scale to the large file’ option was used and we will use “toSm” mean that the ‘scale to the small file’ option was used. Another option for dealing with the disparity between the number of reads in the two files would be to adjust the number of reads in each file before submitting them to MACS such that the files have roughly the same number after MACS filters redundant reads out. John Urban has created a python script that calculates this adjustment and used it in a separate analysis where we called peaks using reads from only a single lane instead of combining all NS samples (discussed more below).

With just the two options discussed above, there are 4 possible sets of parameters. The parameter sets will be named such that it describes how many redundant reads were kept first, then whether the counts were scaled toward the large or small file: “K_to___”. For example, keeping just 1 read (K1) and scaling toward the small file (toSm) will be called K1toSm.

(C) P value to call peaks - MACS uses the Poisson distribution to call peaks. Briefly, the Poisson distribution deals with the probability of X events occurring in a given time/space interval given that the average number of events that occur is λ (lambda). For MACS, it is the probability of seeing X reads in a genomic interval given that the average number of reads in that interval is λ (lambda). The p-value, which is used to determine if the read count is statistically significant, is just the probability of seeing greater than or equal to X reads in a genomic interval given that the average number of reads in that interval is λ (lambda). P-values closer and closer to 0 are considered more and more significant (they are less so as they approach 1). MACS allows the user to pick a p-value cutoff, C, where read counts with p-values higher than C (closer to 1) are not called as peaks whereas those with p-values less than or equal to C are called as peaks. Unless otherwise stated, we used the p-value cutoff of 0.00001.

(D) Dynamic genomic interval for the read count - Not only does MACS use the Poisson distribution, it uses a "dynamic lambda", meaning that it does not necessarily use the same average read count for all genomic intervals. This is in contrast to a 'static lambda' such as only using genomic average, λ(genome) = [number of reads/genome size]*interval_size. Instead, MACS looks for local biases in the genomic interval it is currently looking in. It does this by calculating two more λ values: λ(small local window) and λ(large local window). The size of these two local window sizes can be tweaked. These parameters are called "slocal" and "llocal". The default slocal is 1000 bp. This is approximately twice the size one might expect ChIP-seq peaks to be. We expected ~1500 bp peaks so kept 'slocal' fixed at 3000 bp while tweaking only the llocal option. **Supporting Data Figure 3** shows a curve of the number of peaks called as a function of llocal. We wish to maximize the number of true peaks called. Note that all four conditions have a slight elevation in number of peaks at llocal = ~50 kb. This llocal value will continue to be most interesting in subsequent figures.

(E) False Discovery Rate to identify the true peaks – Supporting Data Figure 3 shows that there are 77,000-84,000 peaks in the MCF-7 genome from our combined NS-Seq data. How many of these are true peaks (bona fide replication origins) and how many are false positives? Using MACS, John Urban calculated the False Discovery Rate (FDR). As shown in **Supporting Data Figure 4**, the expected number of true origins based on the data stayed somewhat constant for all 4 conditions (with llocal=~30 kb) with a range from ~66,000 to ~70,000. Note that all conditions had a slight elevation of true peaks and a slight dip in false peaks around llocal=50 kb. As will be seen below, taken together this means that there is also a slight decrease in FDR at this llocal value.

All 4 parameter sets have a slight dip in FDR at llocal = ~50 kb (**Supporting Data Figure 5**). The "K1" sets have ~15% FDR after llocal = ~40-50 kb while the "K3" sets are slightly higher at ~16-17% FDR. When llocal = 50 kb, we have shown that the there is a slight elevation in the number of peaks, that the expected number of true positives is slightly elevated, that the expected number of false positives dips, and that as a result the FDR is most often lowest near llocal= 50 kb. This is why we chose to further explore the llocal = 50 kb sets for all conditions.

(F) Confirmation that the 50 kb llocal sets are appropriate for calling peaks. We wondered how different the llocal=50kb sets of a given condition (e.g. K1toLg) were from the other sets from the same condition when different llocal values were used. Note that for each condition the llocal=50 kb set contained the most peaks. Therefore, we asked whether or not the smaller sets were all proper subsets of the llocal=50 kb set – i.e. does the llocal=50 kb set contain every peak called from the smaller set in question? If not, we wanted to know how many peaks from the smaller set were not in the bigger set and whether that was lower than the expected number of false peaks in the smaller set. Each llocal value should have some unique peaks as a result of the differences in "dynamic lambdas" used. If the number of unique peaks exceeds the expected number of false peaks in a set, then that set would be considerably different from the llocal=50 kb set. This would be problematic because we would not have a way of knowing, which set was more representative of the truth. If the sets are reasonably similar, then choosing a set becomes more arbitrary and choosing the set that minimizes FDR while maximizing the number of peaks makes most sense. In such a case, we could move forward.

The peak sets are written out in BED files. Suffice it to say that each line in a BED file represents a peak and that the first three columns state the genomic coordinates of that peak by specifying the chromosome, the start position, and the end position respectively. BEDtools is a bioinformatics program that can manipulate BED files in numerous ways (Quinlan et al., 2010). We used BEDtools to compare two sets of peaks at a time to see how many peaks were shared in common between the sets. In **Supporting Data Figure 6**, we report how many peaks in the smaller set were NOT in the larger llocal=50 kb set. The conclusion is that the 50 kb sets are appropriate to use in peal calling.

We next wanted to make sure that the peak sets did not vary much between conditions. If the peak sets were much different from each other between K1 and K3, then once again determining which of among the sets more reflected the truth would not be straightforward. However, if the sets did not vary considerably between conditions, then it simply would be arbitrary in choosing a set. One would just pick one that minimized FDR while maximizing peak calls. **Supporting Data Figure 7** shows this analysis, which is analogous to Figure 6. This time the K3toLg set was biggest set so all smaller sets were compared to it. All sets were found to be reasonably similar. As the K1 sets had lower FDR, one of these was chosen as our final set. Scaling to small is supposed to have higher specificity and lower FDR. Nonetheless, we do not necessarily see this for the K1 sets. The K1toLg set actually seems to have a lower FDR.

(G) Number of peaks in each of these sets as a function of FDR%. This was analyzed primarily to access a given FDR set based on the present need. For example, one might prefer the 5% or 1% FDR sets when looking for a motif, but might prefer the entire set when calculating inter-origin distance. Moreover, we were able to see that the higher quality sets (e.g. 5% FDR) often vary much less between different conditions such as biological replicates (**Supporting Data Figure 8**).

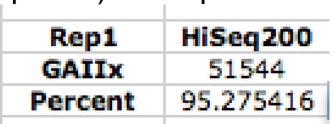
We also looked at how choosing different p-value cutoffs would affect number of peaks called as well as FDR (**Supporting Data Figure 9**). This was done on the K3toLg set before we decided on the K1toLg. Nonetheless, the trend should be similar in all 4 conditions we considered, as they were all similar. Note that our p-value cutoff was 0.00001 and that log10(0.00001) = -5. In other words, the most stringent p-value cutoff is leftmost and the p-value cutoffs get more and more lax as it goes right along the x-axis to log10(0.1) = -1. What is interesting about what we see here is that the expected number of true peaks remains relatively constant while the expected number of false peaks grows with less stringency solely contributing to the rising total number of peaks. This seems to indicate that even more stringent p-value cutoffs might keep the same number of expected true positives while reducing further the expected number of false positives until the total number of peaks and expected number of true peaks are approximately the same. We are yet to do this in large part because another peak caller we use, Sole-Search (discussed later), calculates FDR in a different way and gives completely different FDR estimates (e.g. 0.001% instead of 15%) while the regions with peaks remain relatively the same. Moreover, **Supporting Data Figure 10** of the expected numbers of false and true positives normalized by the total number of peaks shows FDR and “TDR” respectively as functions of pvalue cutoff. They seem to be leveling off in the left-direction indicating that the number of true peaks and total peaks will not actually converge, at least not until after most true and false positives are eliminated from consideration.

Biological and technical variation

We analyzed the biological variation in three different samples of MCF-7 nascent strand DNA. Having set the variables as described above, for each of the separate samples we used MACS to call peaks from the mappable reads after correction for redundant reads and subtraction of background from nonreplicating genomic DNA (**Supporting Data Figure 11**). The first DNA preparation (Rep1) was prepared by the BND-cellulose protocol, whereas the next two samples (Rep 2, Rep 3) were prepared with the sucrose gradient protocol. Therefore, we anticipated that Rep2 and Rep 3 would be more similar to each other, which indeed turned out to be the case, as now described. We calculated how many peaks from a given set (row) were represented by another given set (column) (**Supporting Data Figure 12**). This analysis is performed using BEDtools (Quinlan et al., 2010). This is followed by a table that instead shows the percent of the given set (row) that is represented in another set (column) (**Supporting Data Figure 13**). The diagonal from top-to-bottom, left-to-right in the first table(Figure 12) is in bold because this shows the total number of peaks as the row set and column set are the same at these intersections. In the second table (Figure 13) this is evident as the diagonal has 100% in all boxes.

It is clear that Rep2 and Rep3 are more like each other than they are like Rep1. We believe this is because Rep1 was subjected to the BND-cellulose approach discussed above while Rep2 and Rep3 were subject to the sucrose gradient approach. That Rep2 and Rep3 contribute ~173 million (173,135,267) mappable reads while Rep1 contributes just ~66.4 million (54,642,570 + ~11.8 million from GAIIx) speaks to why higher percentages of peaks from these files are represented in the Combined file and why higher percentages of peaks from the Combined file are represented in these files despite that they have less peaks than Rep1.

Next to look at the technical reproducibility, we compared the original GAIIx run reported last year with the HiSeq2000 Rep1 data treated in the same way as GAIIx. The HiSeq Rep1 data and the GAIIx data are derived from the same NS sample (the BND-cellulose approach). We did not have input data for the GAIIx analysis last year. Therefore, in a separate analysis on the Rep1 data, we treated it as if there was no input control. Below we show the comparison to find how many peaks in the GAIIx set (54,100 peaks) are represented in the larger HiSeq2000 set (117,446 peaks):



Rep1	**HiSeq200**
GAIIx	51544
Percent	95.275416

This shows that 51,544 peaks out of the 54,100 peaks in the GAIIx set (~95.3%) are represented in the HiSeq set. Therefore, this high throughput sequencing of a NS sample is highly technically reproducible. Thus, sequencing adds a small amount of technical variability compared to the variability between sample preparations, which is a mixture of biological variability and that introduced by the procedure and/or by using different nascent DNA isolation procedures.

Saturation

We wanted to know whether or not we have reached saturation, where saturation is defined as reaching an area of diminishing returns with more sequencing. To do this, the reads in the Combined file were shuffled to mix reads from all experiments together. Otherwise, the reads from each experiment are just stacked on top of each other in the Combined file – e.g. first the GAIIx reads, then the HiSeq2000 Rep1 reads, etc. Why shuffle them first? To certain extents, the analysis of lower read counts and number of peaks they give rise to has already been done. Rep1 had ~43.5 million reads, that were mappable and passed MACS filtering, to call peaks while Rep2 and Rep3 had 65 and 69 million. Rep1, 2, and 3 gave rise to ~80,000, ~55,000, and ~63,000 peaks respectively while the Combined set of mappable reads that passed filtering (~171 million) had ~79,000 peaks, suggesting that we have reached saturation (**Supporting Data Figure 14**). However, that the number of peaks seems to have leveled off also means that combining reads from biological replicates reduces the amount of spurious peaks called in a given replicate. For example, only ~65% of the peaks in Rep1 are represented in the Combined set (see analysis above). In other words, 35% of the peaks were not substantiated when biological replicate reads were added into the analysis. This is part of the power of combining reads from biological replicates. Therefore, it would be interesting to look at how many peaks are called when 10 million of the shuffled combined reads are used, 20 million, 30 million, and so on up to 230 million. Note that because (i) not all of the reads are mappable and (ii) there will be increasing rates of redundant read events with increasing number of reads, the true number of reads is equal to the number of mappable reads that pass the redundant reads filter in MACS set to keep only 1 read. **Supporting Data Figure 15** shows peaks when no input control is used. This simply means all possible peaks (true and false positives) are included in count. We are in the midst of performing this saturation analysis using input reads as well. It appears as though we have come close to saturation. Whereas the first 80 million reads gave rise to ~90,000 peaks, the second 80 million reads only gave rise to ~104,000 or just 14,000 more. Another 80 million would give rise to even less additional peaks and most would be very low enrichments that became statistically significant due to the large increase in sample size (~240 million reads).

Task (2) Comparison of replication origin maps between breast cancer (ER+, ER-) and normal breast cells. These results would indicate if replication origin usage changes between normal and breast cancer cells, and if it varies between ER positive and ER negative breast cancer cells.

We are nearing completion of subtask (2a) to map replication origins in an ER+ breast cancer cell line --- namely MCF-7 (see Task (1). Due to the additional experiments of NS-Seq method validation that were not in the original grant application, we requested a one year no cost extension that was approved by DOD. The experiments in subtask (2b) to map replication origins in ER- breast cancer cells (e.g., MDAMB231 cells, SKBR3 cells) have been deferred. However, in year two we have begun to carry out subtask (2c) to map replication origins in normal breast cells (MCF-10A) as in the timeline of the grant application. The similarities and differences in replication origin maps for ER+ and ER- breast cancer cell genomes and comparison to the origin map for the normal breast cell genome will address whether replication origins differ in different cell types, especially comparing breast cancer cells to normal breast cells, and comparing ER+ to ER- breast cancer cells.

MCF-10A

In year two we isolated nascent DNA as well as input genomic DNA from normal human breast cells (MCF-10A). We sequenced this material using the HiSeq2000 platform. **Supporting Data Figure 16** shows the results.

Data analysis with Sole-Search vs MACS

As indicated earlier in this progress report, a recent paper by the Lemaitre group (Besnard et al., 2012) obtained more than double as many peaks as us or Martin et al. (2011). Where we were calling nearly 80,000 peaks for MCF7 and nearly 68,000 peaks for MCF-10A, they were calling 200,000 to 250,000 for their cell lines. Though they used different human cell lines than us, there is no reason to expect this massive difference. A notable difference is that they used Sole-Search (with default parameters) rather than MACS as the peak caller. Below we describe our calculations to compare the effect of using the MACS vs. Sole-Search peak caller. To do this, we used our data with Sole-Search.

Sole-search has fewer parameters than MACS to tweak. Lemaitre's group kept all default parameters:

```
Permutation:5
Fragment:200
AlphaValue:0.0010
FDR:0.0001
PeakMergeDistance:0
```

We did roughly the same for our MCF-7 data:

```
Permutation:5
Fragment:350
AlphaValue:0.0010
FDR:0.0001
PeakMergeDistance:0
```

The difference is in bold. We used 350 because it accurately reflects our average fragment length. This parameter only eliminates peaks < 350 bp in length (or < 200 bp in Lemaitre's case).

Another difference is that we provided our own MCF-7 input control reads while the Lemaitre group used the generic genomic reference reads provided by the software. These reads are sampled from input controls from an array of cell lines. Ultimately, this should cause some problems because it does not account for the specific biases (such as amplifications) in the genome that the nascent strands were purified from. This difference will be explored further when we discuss the differences for our MCF-7 set when using our own input or using the generic input from the Sole-Search software. For now, we wanted to keep everything as similar to the MACS analysis as possible. This would tell us if the different methods of identifying peaks were responsible for the difference.

The results of using Sole-Search on our NS-Seq data for MCF-7 are striking:

MappedReads:	239566823
UniqueReads:	165908046
Number of peaks:	**280,368**
Average peak height:	52.88
Median peak height:	49
Highest peak:	229
Lowest peak:	23.2016210739615
Average peak width:	983.39

As shown in **Supporting Data Figure 17**, there were over 280 thousand peaks when using Sole-Search with the same parameters to the Lemaitre group. Moreover, if we change the Sole-Search FDR parameter from 0.0001 to 0.001 (less stringent) or 0.00001 (more stringent) we get 334,197 and 258,243 peaks from our MCF-7 data respectively.

We also used Sole-Search on our MCF-10A data (**Supporting Data Figure 18**). Specifically, for MCF-10A we used all of the same parameters as for MCF-7. As compared to the 67,812 peaks called by MACS, Sole-Search called 110,212 or 219,637 or 288,567 peaks for MCF-10A sets with FDR equal to 0.00001, 0.0001, and 0.001 respectively. Notice that the FDR reported for MACS was very high. Nonetheless, almost all of the MACS peaks are found within the Dole-Search set, as described below. The difference in numbers of peaks is mostly a consequence of having multiple smaller-width Sole-Search peaks in the same region as one larger-width peak from MACS (discussed in more detail below). Therefore, FDR between peak callers is not comparable. Each computes FDR differently and it is not apparent which approach is more appropriate or accurate.

It is seen for both cell lines that the different peak callers indeed give rise to different numbers of peaks when all of the data provided is the same. This is important because the number of peaks is used as a proxy for the number of origins of replication in the human genome. Moreover, the peak sets are used to estimate inter-origin distance, the average of which will be much smaller for the Sole-Search set, as well as for motif discovery and other downstream analyses. Finally, the peak locations are used to see what other genomic features they correlate with. These different outputs could lead to different conclusions – one output may give rise to a false correlation or break a true correlation.

We wanted to know if, despite the large difference in the number of peaks, if the peaks were in the same genomic regions. First, we tested how much peaks in the MACS set overlapped with peaks in the corresponding Sole-Search set and vice versa. High percentages of overlapping peaks would mean that, despite the large difference in number of peaks, that peaks were being called in the same regions. An overlap is counted if a peak in one set, overlaps at least 1 peak in the other set by at least 1 bp. We found that almost all of the MACS peaks were found in the Sole-Search set, but only about 60% of the Sole-Search peaks were found in the MACS set for MCF-7 (**Supporting Data Figure 19**) and for MCF-10A (**Supporting Data Figure 20**).

We next looked to see if each of the peak-callers called peaks in regions known to have replication initiation activity. If the sets from the 2 different peak callers are not similar as determined by the overlap test described above, then which might be more accurate (as determined by having peaks at know origin sites)? Conversely, if they are determined to be similar, do they both have peaks in these regions of known origins? To determine whether or not one, none, or both had peaks in regions known to have origin activity, we visualized the peaks from the MCF-7 and MCF-10A sets for both MACS and Sole-Search peak callers in the IGV browser and looked at 3 specific sites: the c-Myc locus, the HBB (beta globin) locus, and the RPE locus (**Supporting Data Figures 21, 22 and 23**, respectively. Both peak callers had peaks in all 3 of these places in both cell lines. This also visually shows that Sole-Search places many smaller-width peaks inside single, wider peaks called by MACS. It is clear that MACS has poorer resolution, but though the Sole-Search resolution seems to have finer resolution, it is unclear whether it parsed up the region too much, particularly at the c-Myc locus where some of the peaks called by Sole-Search do not coincide with experimental data. The top-most row always displays RefSeq genes (blue) at the given locus. It is then followed by MACS peaks (red) for MCF-7, Sole-Search peaks (green) for MCF-7, MACS peaks (red) for MCF10-A, and Sole-Search peaks (green) for MCF10-A, respectively. Above the rows showing genomic features (genes and NS peaks) is a representation of the chromosome and the width of the locus being viewed.

Finally, we looked at the density of peaks along the genome to see if the density rises and falls with each other. In other words, do the density curves of peaks along the genome visually correlate with each other? **Supporting Data Figures 24** shows the densities of peaks along the genome starting with chromosome 1 and going up through chromosome X. The density of RefSeq genes is first shown in blue. Next, in red, is the density of the MACS MCF-7 peaks followed by the density of MACS MCF-7 peaks when they are randomly shuffled across the genome. In green, the Sole-Search peaks for MCF-7 are shown followed by randomly shuffling them across the genome. Notice that the nascent strand peaks for both peak callers have similar profiles and that this profile is similar to the density of RefSeq genes. Moreover, this similarity is broken if the peaks are randomly shuffled. This suggests once again that the two different peak callers are detecting nascent strand signal from the same regions of the genome though they are parsing up these regions slightly differently leading to different numbers of peaks.

Visually, it appears that the peak densities of both peak callers go up and down with each other. For a more quantitative description of how the densities go up and down with each other, we tested for correlation using two tests: Pearson's r and Spearman's rho. The Pearson test for correlation assumes a linear relationship whereas the Spearman test does not. Instead, the Spearman ranks the scores in each set and then tests if the ranks of the different sets go up and down with each other. Both tests were used to test for correlation and found a moderately strong positive correlation between the peak densities of both peak callers (**Supporting Data Figures 25)**. Note that for zero peaks called by MACS, several can be called by Sole-Search, helping to explain the 40% additional peaks called by Sole-Search as compared to MACS (Supporting Data Figure 19). If one of the peak sets is randomly shuffled across the genome, the correlation is broken (**Supporting Data Figures 26)**.

<u>The effects of using the 'generic input' option in Sole-Search on our data</u>

As mentioned, the Lemaitre group (Besnard et al., 2012) did not use their own input control sequencing reads. Instead, they specified to Sole-Search to use generic reads that the creators of Sole-Search have amassed from various cell lines. The question becomes, "Does using the generic input option change our results in comparison to using our specific set of control reads when all other parameters are kept the same?" Sole-search called 280,368 peaks when our own specific input control reads for MCF-7 were provided, but only 194,815 peaks (~69.5% the size) when their generic cell line reads were used as a control (**Supporting Data Figure 27a**). This means that when using generic input, Sole-Search called 85,553 less peaks than when using the specific input, which indicates it has less sensitivity when generic input is used. The size of the peak set called with generic input is only ~69.5% the size of the specific input peak set (**Supporting Data Figure 27b**), but what percent of the specific input set is actually covered by the generic input set? The answer is only ~57% (**Supporting Data Figure 28**). What percent of the generic set is represented in the specific set? The answer is ~91.1% (Supporting Data Figure 28). Taken together, this shows that ~91.1% of the peaks called in the generic set agree with peaks in the specific set covering just 57% of the specific set. Relative to the specific set this implies that using the generic input leads to Sole-Search having a false negative rate of ~43% (i.e. the probability of not calling a peak that would be called with specific input is ~43%) (**Supporting Data Figure 29**). Moreover, that 91% of the generic set is represented in the specific set leaves ~8.9% unique to the generic set. Relative to the specific set, this means ~8.9% of the peaks are false peaks and that the minimum FDR, defined as the number of false peaks divided by the total number of peaks in the set (false/total), is ~8.9%. However, the FDR may be even higher. 91% of the generic set overlap 57% of the specific set. This means that there are instances when more than one generic peak overlaps the same specific peak. If one takes the number of peaks represented in the specific set and divides by the total number in the generic set, it gives an approximation to the percent of non-redundant peaks in the generic set that overlap a peak in the specific set, just 82%. That means the FDR relative to the specific set could be as high as 18%. Taken together, all of this could imply that even though the Lemaitre group report the FDR given by Sole-Search (0.0001), it may be that the FDR is far higher. This is in addition to having a false negative rate relative to the specific set that implies they could not have reached saturation of origins – just saturation at their level of sensitivity.

<u>Nucleotide composition of the peaks</u>

To begin analyzing the sequences of our NS peaks for both MACS and Sole-Search, we aligned all of the peaks by either aligning the peak summits (nucleotide of highest coverage) and/or the peak centers (position in middle of peak length). From the given alignment focal point, we looked at the first 2000 nucleotides (nt) in both directions. Therefore, we looked at the 4000 nt centered around the summit/center for all peaks to find the nucleotide proportions at each position. This provides insight into whether or not there are any skews in the nt distribution in the peaks deviating from the random background distribution. To show the random background distribution, the peaks and/or peak summits were also randomly shuffled around the genome and treated the same as discussed above.

When this analysis was done for MCF-7, the following distributions were seen:

(1) MACS MCF7 NS peak summits (**Supporting Data Figure 30**)
(2) MACS MCF7 NS peak centers (**Supporting Data Figure 31**)
(3) MACS MCF7 shuffled NS peak summits (**Supporting Data Figure 32**)
(4) Sole-Search MCF7 NS peak centers (**Supporting Data Figure 33**)
(5) Sole-Search MCF7 shuffled NS peak centers (**Supporting Data Figure 34**)

From the random distribution obtained from shuffling peaks, it is clear that the background proportions for A and T are ~29% each and for G and C are ~21% each regardless of position. In other words, at random, the nt distribution is position-independent. However, when centered at summits or peak centers, the distribution becomes position-dependent. It is clear that within 500-1000 nt from the peak center in both directions there is a non-random nt distribution that increases in GC content as it approaches the center.

When this analysis was done for MCF-10A, the following distributions were seen:

(1) MACS MCF-10A NS peak summits (**Supporting Data Figure 35**)
(2) MACS MCF-10A peak centers (**Supporting Data Figure 36**)
(3) Sole-Search MCF-10A peak centers (**Supporting Data Figure 37**)

The background distribution stays the same as shown above. The same trend of higher GC content as one approaches the summit/center is also seen in the MCF-10A data.

To explore whether the nt skew found in the summit/center of peaks was meaningful or whether it was an artifact, we also looked at peaks that were not derived from Nascent Strand enrichment, using genomic DNA (gDNA) as a control. We used MACS to call peaks from just the control input reads. We did this originally to approximate areas of amplifications. The peaks were called using the static lambda (genome background average read count for a given interval). A dynamic lambda would not work to call amplifications because it would use the local average read count to see whether there was significant enrichment over background in a given area. However, an amplicon is not significantly enriched over itself. It is significantly enriched over the genomic average. These amplicon peaks were assessed the same as described above for their nt composition centered around the summits (nt of highest coverage). It is clear, that the same GC bias arises around the summit of these peaks, which have not come from our NS isolation protocol. The results shown in **Supporting Data Figures 38 and 39** argues that the task of centering peaks around their summits/centers in combination with biases common to all Illumina preparations (PCR bias, sequencing bias) at least partially explains this distribution. In other words, it argues that this GC skew by summits and centers that has been reported by others is an artifact.

As another control for the GC skew, we looked at the “negative peaks” MACS called for our MCF-7 and MCF-10A datasets. Negative peaks arise from the FDR approximation process of MACS. First, it calls all peaks in the treatment file using the input file as background. Regions in the treatment file enriched over the same regions in the input file are called as peaks. To approximate how many of the peaks called are ‘false peaks’, areas enriched over background by chance, MACS swaps the roles of the treatment and input files. Now it looks for regions in the input file that are significantly enriched over the same regions in the treatment file using all of the same parameters. MACS considers these to be false peaks by definition. It then uses this number as a proxy for the number of false peaks that were called in the treatment file to estimate the FDR of the entire set:
FDR = #false/#total. It also provides all of the locations for these “negative peaks”. Negative peaks are interesting for this analysis because they arise in regions that are seriously deprived or depleted

of reads from the Nascent Strand sequencing relative to the number of reads at those regions in the input control. This argues that these are regions that lack replication origins. Moreover, the negative peak summits in these regions may arise at non-random locations as reads are non-uniformly distributed across the genome, most likely due to PCR or sequencing biases. When the negative peaks are centered at the summits, we see similar nucleotide skews away from the random distribution for both MCF-7 (**Supporting Data Figures 40 and41**) and for MCF-10A (**Supporting Data Figures 42 and 43**). This argues that the center/summits of peak regions are, in general, non-randomly enriched at areas within those regions of higher GC content. Therefore, though our peaks may indeed localize at or near origins of replication, where the peak summits and/or centers is not necessarily meaningful. Aligning the peaks by either gives rise to an artifact. This means another approach needs to be taken to align the peaks such as a Hidden Markov Model approach that will align the peaks by state paths modeling nt compositions or a motif.

We next examined our MACS MCF7 set for any base composition bias for the “best” peaks as determined by p-value scores (**Supporting Data Figures 44 and45**). Interestingly, they show slightly different distribution patterns than all the peaks combined. However, both still have a GC rise in the middle. The top 114 peaks show a lot of variation from one position to the next, but have a general trend of AT content that is higher than the genomic average. That is interesting as origins of replication from all bacteria, bacteriophages, animal viruses, and unicellular eukaryotes studied thus far have high AT content, presumably to facilitate unwinding of the DNA.

Task (3) Correlation of origin map data w ith sites of (a) DNA amplification and (b) estrogen receptor binding. These data will support or refute the hypothesis that ER may bind next to the replication machinery and induce DNA amplification.

These results will support or refute the hypothesis that ER may bind next to the replication machinery and induce DNA amplification. This analysis was originally scheduled for year two. However, because of the unanticipated delay necessitated by our forging new ground to refine technical and computational issues for NS-Seq, task (3) will be deferred to year 3 (no cost extension). We will compare the origin map data to data that already exists on sites of DNA amplification (to identify amplification origins) as well as confirm and expand these data using our own data on the number of reads from sequencing bulk genomic DNA from the various cell lines we are using. This information will, in turn, be compared to existing data on sites of ER binding. It may prove necessary to undertake some ChIP (chromatin immunoprecipitation) experiments for validation of ER binding, though not proposed in the original grant application. These data will indicate if a correlation exists between ER binding and origins that re-replicate (amplify), thereby testing our hypothesis.

Key Research Accomplishments

- Development of the method of Nascent Strand-Seq (NS-Seq) to map replication origins in the genome - We have developed and refined this method. We plan to also apply it to the yeast genome for validation of the method.

- Application of NS-Seq to map replication origins in the MCF-7 breast cancer genome - We have obtained results of NS-Seq to map replication origins in the MCF-7 genome. Our initial results with the Illumina GAIIx platform have now been extended to three samples run on the Illumina HiSeq2000 machine, providing both biological and technical replicates.

- Validation of the NS-Seq results by finding known replication origins in our data set - We have validated NS-Seq on known origins, including Myc, DBF4, DHFR, β-Globin, RPE, as well as Lamin B2, and Glucose-6-Phosphate Dehydrogenase.

- Comparison of our data to the data sets of other labs to map replication origins in the human genome. The data sets from Cadoret et al., 2008; Karnani et al., 2010., and Mesner et al. 2011 were based on using ENCODE (1% of the human genome) for HeLa cells, so finding only a small amount of overlap could be due to their use of a different cell line than that used by our lab. Moreover, even when comparing the results between these three data sets, there was not complete agreement, suggesting lack of saturation of the data. The Martin et al. (2011) data set used MCF-7 cells and was for the full genome, but did not give full overlap with our data. They did not show any data for validation of their results, and we suspect that they had contamination from Okazaki fragments as they selected small nascent strand DNA. We have spent considerable effort to analyze the effects of different peak callers (MACS and Sole-Search), and our results demonstrate that the larger number of peaks called by the Lemaitre group (Besnard et al., 2012) reflects their use of the Sole-Search peak caller.

- Analysis of base composition at the peak summits or centers. Our computational analysis reveals that the apparent GC skew at the peaks seen by other groups appears to be a computational artifact. In fact, the “best” peaks in our data set have some AT bias instead, which is more consistent with the base composition of replication origins from bacteria, viruses and yeast.

- NS-Seq has been performed on normal human breast cells (MCF-10A) to allow comparison to the replication origins in human breast cancer cells (MCF-7).

Reportable Outcomes

Our results have been presented at the DOD Era of Hope meeting and the Cold Spring Harbor DNA Replication meeting as reported in last year’s progress report and will be written up for publication soon.

Conclusion

We are forging new ground to refine the method for nascent strand sequencing (NS-Seq) and to discover which computational method 9that we are developing) is best to use for analysis of the results. Our data thus far has revealed serious issues that may flaw the results published so far by other groups to map replication origins in the human genome. Therefore, when completed, our study will correct the conclusions drawn by others, as well as serve as the framework to test the hypothesis if there is a correlation between amplification origins and sites of estrogen receptor binding.

Personnel Paid From This Grant

PI and co-PIs:

Susan Gerbi	Professor of Biology (PI)
Alexander Brodsky	Assistant Professor of Medical Science (co-PI)
Benjamin Raphael	Associate Professor of Computer Science (co-PI)

Lab Personnel:

Yutaka Yamamoto	Research Associate
Michael Foulk	Research Associate
Jacob Bliss	Research Assistant
Souriya Vang	Research Assistant

References Cited

E Besnard, A babbled, L Lapasset, O Milhavet, H Parrienello, C Dantec, JM Marin, JM Lemaitre (2012). Unravelling cell type-specific and reprogrammable human replication origin signatures associated with G-quadruplex consensus motifs. Nature Struct. Mol. Biol. 19: 837-844.

AK Bielinsky AK and Gerbi SA (1998). Discrete start sites for DNA synthesis in the yeast ARS1 origin. Science 279:95-98.

JC Cadoret JC, Meisch F, Hassan-Zadeh V, Luyten I, Guillet C, Duret L, Quesneville H and Prioleau MN (2008). Genome-wide studies highlight indirect links between human replication origins and gene regulation. Proc. Nat. Acad. Sci. 105: 15837-15842.

C Cayrou, P Coulombe, A Vigneron, S Stanojcic, O Ganier, I Peiffer, E Rivals, A Puy, S Laurent-Chabalier, R Desprat, M Mechali (2011). Genome-scale analysis of metazoan replication origins reveals their organization in specific but flexible sites defined by conserved features. Genome Res. 21: 1438-1449.

C Cayrou, P Coulombe, A Puy, S Rialle, N Kaplan, E Segal, M Mechali (2012). New insights into replication origin characteristics in metazoans. Cell Cycle 11: 658-667.

Euskirchen GM, Rozowsky JS, Wei CL, Lee WH, Zhang ZD, Hartman S, Emanuelsson O, Stolc V, Weissman S, Gerstein MB, Ruan Y, Snyder M. (2007) Mapping of transcription factor binding regions in mammalian cells by ChIP: comparison of array- and sequencing-based technologies. Genome Res. 17: 898-909.

Feng J, Liu T and Zhang Y. (2011). Using MACS to identify peaks from ChIP-Seq data. Curr Protoc Bioinformatics, Chapter 2:Unit 2.14.

Gerbi SA and Bielinsky AK (1997). Replication initiation point mapping. Methods 13 (3): 271-280.

Karnani N, Taylor CM, Malhotra A, Dutta A. (2010) Genomic study of replication initiation in human chromosomes reveals the influence of transcription regulation and chromatin structure on origin selection. Mol Biol Cell. 21: 393-404.

Langmead B, Trapnell C, Pop M and Salzberg SL. (2009). Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biol.10(3):R25.

Langmead B. (2010). Aligning short sequencing reads with Bowtie. Curr Protoc Bioinformatics. Chapter 11:Unit 11.7

Martin MM, Ryan M, Kim R, Zakas AL, Fu H, Lin CM, Reinhold WC, Davis SR, Bilke S, Liu H, Doroshow JH, Reimers MA, Valenzuela MS, Pommier Y, Meltzer PS and Aladjem MI (2011). Genome-wide depletion of replication initiation events in highly transcribed regions. Genome Res. (Sept. 22, 2011 Epub).

Mesner LD, Valsakumar V, Karnani N, Dutta A, Hamlin JL and Bekiranov S. (2011). Bubble-chip analysis of human origin distributions demonstrates on a genomic scale significant clustering into zones and significant association with transcription. Genome Res 21:377-389.

Neve RM, Chin K, Fridlyand J, Yeh J, Baehner FL, Fevr T, Clark L, Bayani N, Coppe JP, Tong F, Speed T, Spellman PT, DeVries S, Lapuk A, Wang NJ, Kuo WL, Stilwell JL, Pinkel D, Albertson DG, Waldman FM, McCormick F, Dickson RB, Johnson MD, Lippman M, Ethier S, Gazdar A and Gray JW (2006). A collection of breast cancer cell lines for the study of functionally distinct cancer subtypes. *Cancer Cell* 10: 515-527.

Quinlan AR, Hall IM. (2010) BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics. 26: 841-2.

Robinson JT, Thorvaldsdóttir H, Winckler W, Guttman M, Lander ES, Getz and Mesirov JP (2011).Integrative Genomics Viewer. Nature Biotech 29: 24–26.

Tao L, Dong Z, Leffak M, Zannis-Hadjopoulos M and Price G (2000). Major DNA replication initiation sites in the c-myc locus in human cells. J. Cell Biochem 78:442-457

Valenzuela MS, Chen Y, Davis S, Yang F, Walker RL, Bilke S, Lueders J, Martin MM, Aladjem MI, Massion PP and Meltzer PS. (2011). Preferential localization of human origins of DNA replication at the 5'-ends of expressed genes and at evolutionarily conserved DNA sequences. PLoS One. 2011;6(5):e17308.

Zhang Y, Liu T, Meyer CA, Eeckhoute J, Johnson DS, Bernstein BE, Nusbaum C, Myers RM, Brown M, Li W and Liu XS. (2008). Model-based analysis of ChIP-Seq (MACS). Genome Biol.9(9):R137.

DOD Progress Report 2012 Supporting Data (Figures)

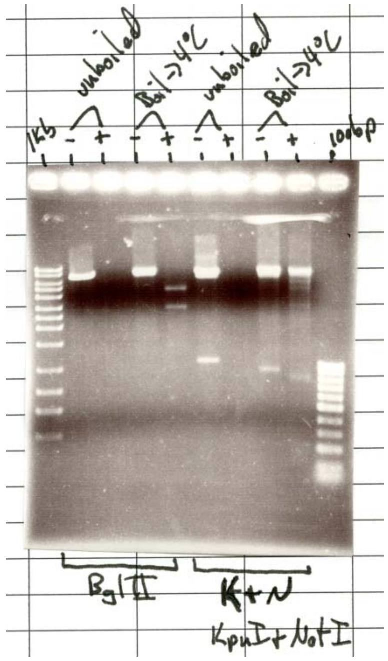


Figure 1. G-quadruplex is resistant to Lexo digestion (see text). Plus or minus lanes were digested or not digested with Lexo. Unboiled DNA shows the pFRT.myc plasmid DNA in the minus lane but in the plus lane where it was fully digested with Lexo no DNA remains. In contrast, Lexo resistant fragments are seen in the plus lanes of the boiled samples (see text). The far left and far right lanes are size markers.

File	Starting # reads	Keep 1 read leaves:	Keep 3 reads leaves:
NS-Comb-Align	239,583,014	171,804,663	212,189,231
Input	179,965,523	165182107	176,806,690

Figure 2. Number of redundant reads retained by MACS.
First, MACS has options on how to deal with “redundant reads”. Redundant reads are reads that map to the same genomic position on the same strand. It is thought that this is an artifact, usually a consequence of the PCR amplification steps in next-generation sequencing preparations. MACS allows one to keep all occurrences, though this is advised against (since it is likely an artifact), as well as two other options: keep only 1 of the reads at that site (discard all others) or use the binomial distribution to determine how many reads could reasonably pile up at the same site on the same strand. In our case, the binomial option allows 3 reads to be kept at a site. We explored these latter two options. Henceforth, we will use the term “K1” to refer to the option that keeps only 1 read and “K3” to be the binomial option that allows 3 reads to be kept in our data. Note that all but 1 or 3 reads at a redundant site are discarded and this reduces the number of reads in the treatment and control files used to call peaks. The table presented here in Figure 2 shows how many are left in each file for each option.

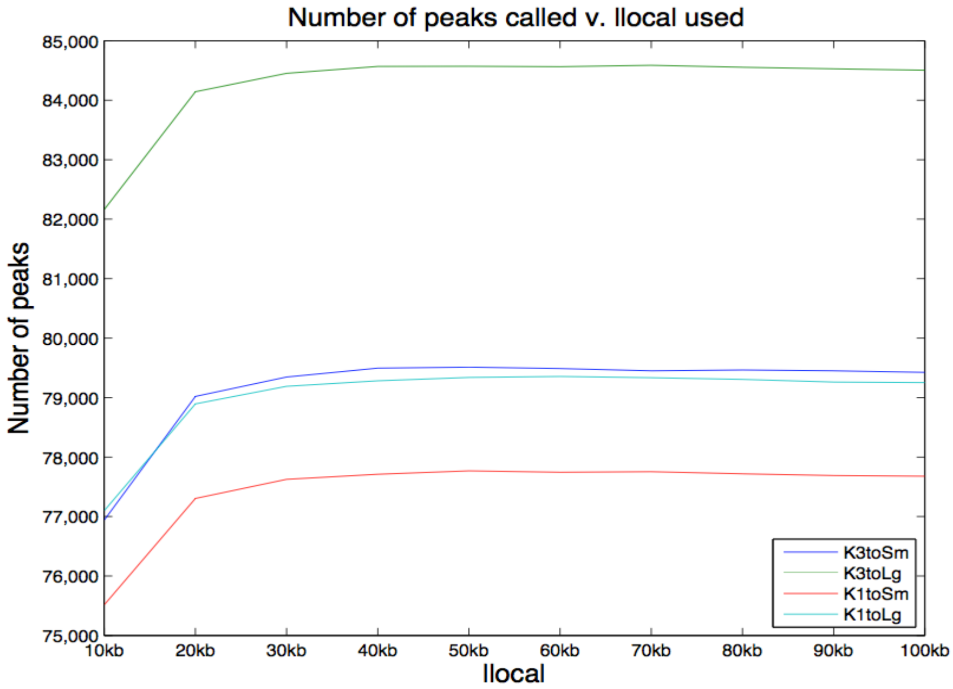
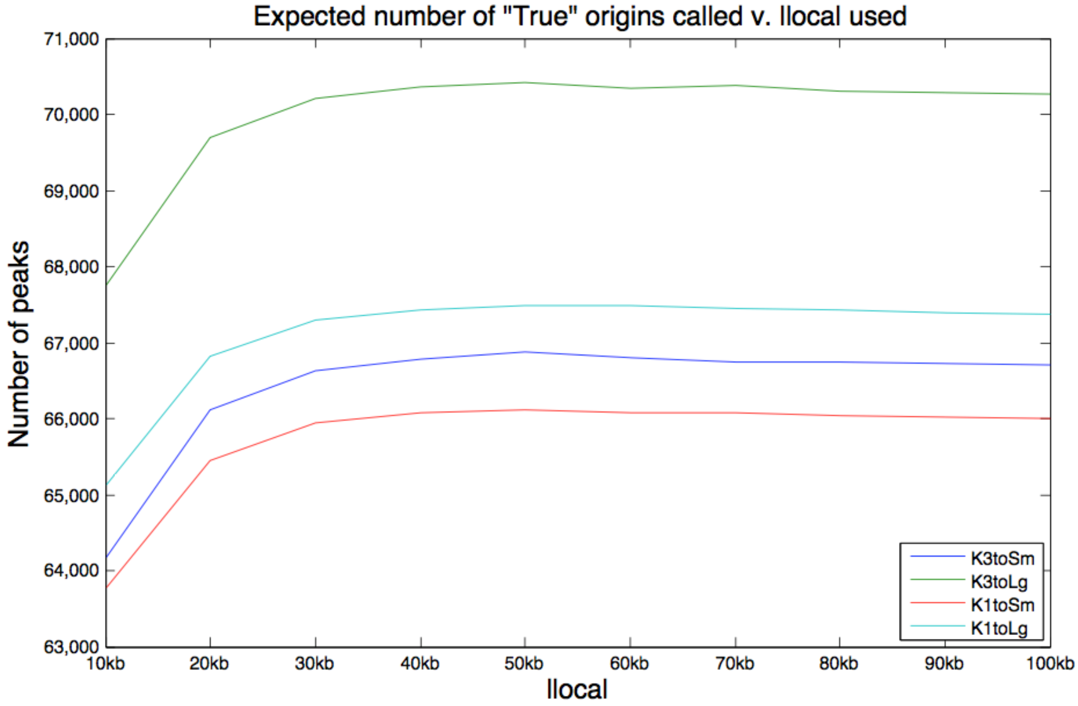


Figure 3. Effect of varying the llocal value on the number of peaks called. Note that all conditions have a slight elevation in number of peaks at llocal = ~50kb.



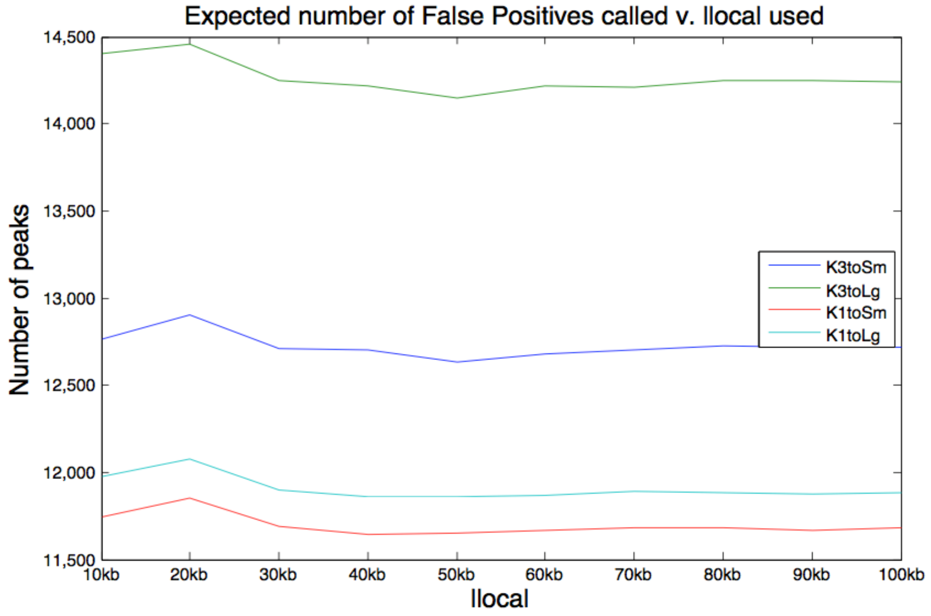


Figure 4. False Discovery Rate (FDR) of the peaks called.
Note that the expected number of true origins based on the data stayed somewhat constant for all 4 conditions after llocal=~30kb with a range from ~66,000 to ~70,000. Moreover, note that all conditions had a slight elevation of true peaks and a slight dip in false peaks around llocal=50kb. As will be seen below, taken together this means that there is also a slight decrease in FDR at this llocal value.

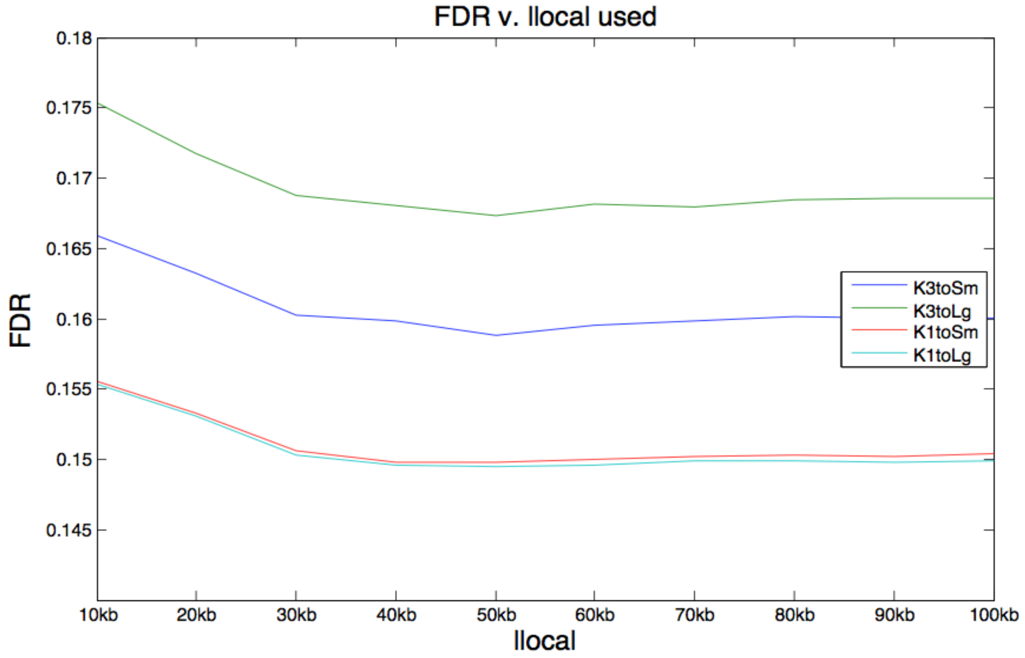


Figure 5. Effect of varying llocal on the False Discovery Rate.

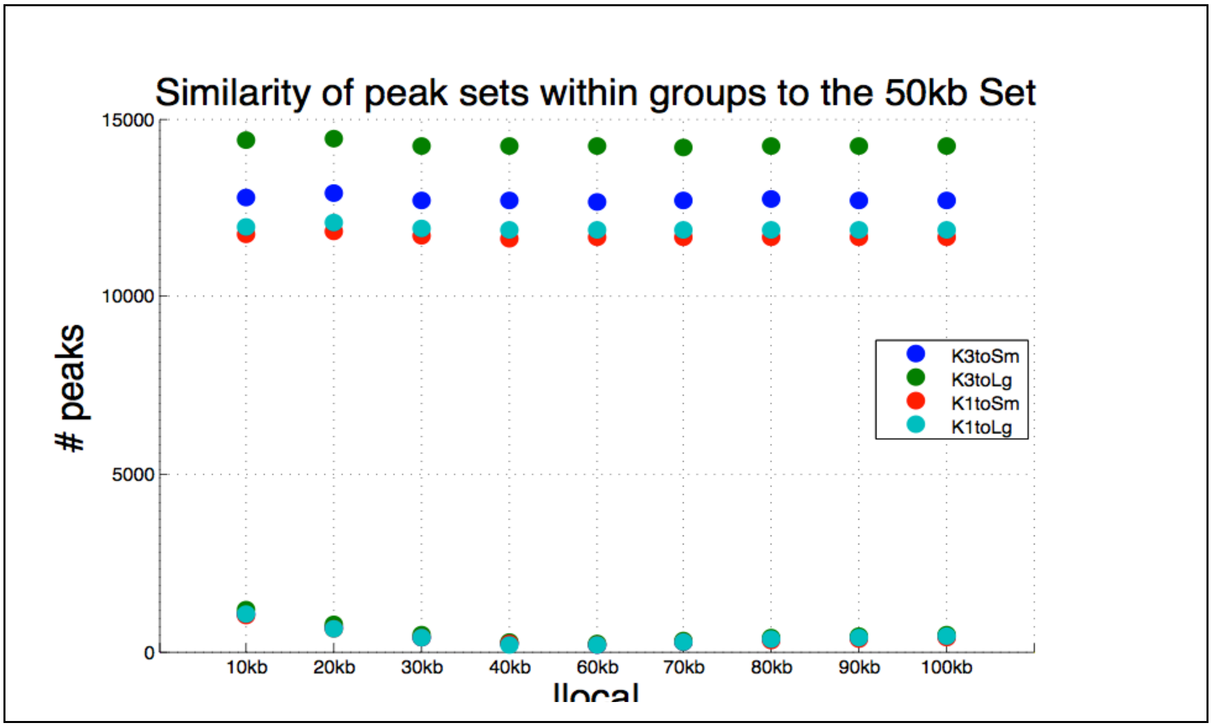


Figure 6. Similarity of peak sets within groups to the 50 kb set. The top set of data points represent the expected number of false peaks in each condition for the given llocal value. The bottom set of data points represent how many peaks were in the smaller set of given llocal value that were NOT in the 50 kb set. The latter number is always a tiny fraction of total peaks (total peaks all >66,000; not shown here, see above figures). Moreover, it is always a small percentage of the number of peaks expected to be false suggesting that the discrepancies could be explained by differences in false peaks alone. If the number was greater than the number of expected false, then one would have to conclude that true peaks definitely differed between sets. Though some true peaks may differ here, even if all, that number is small. Therefore, the 50 kb sets were considered fine representatives of each condition.

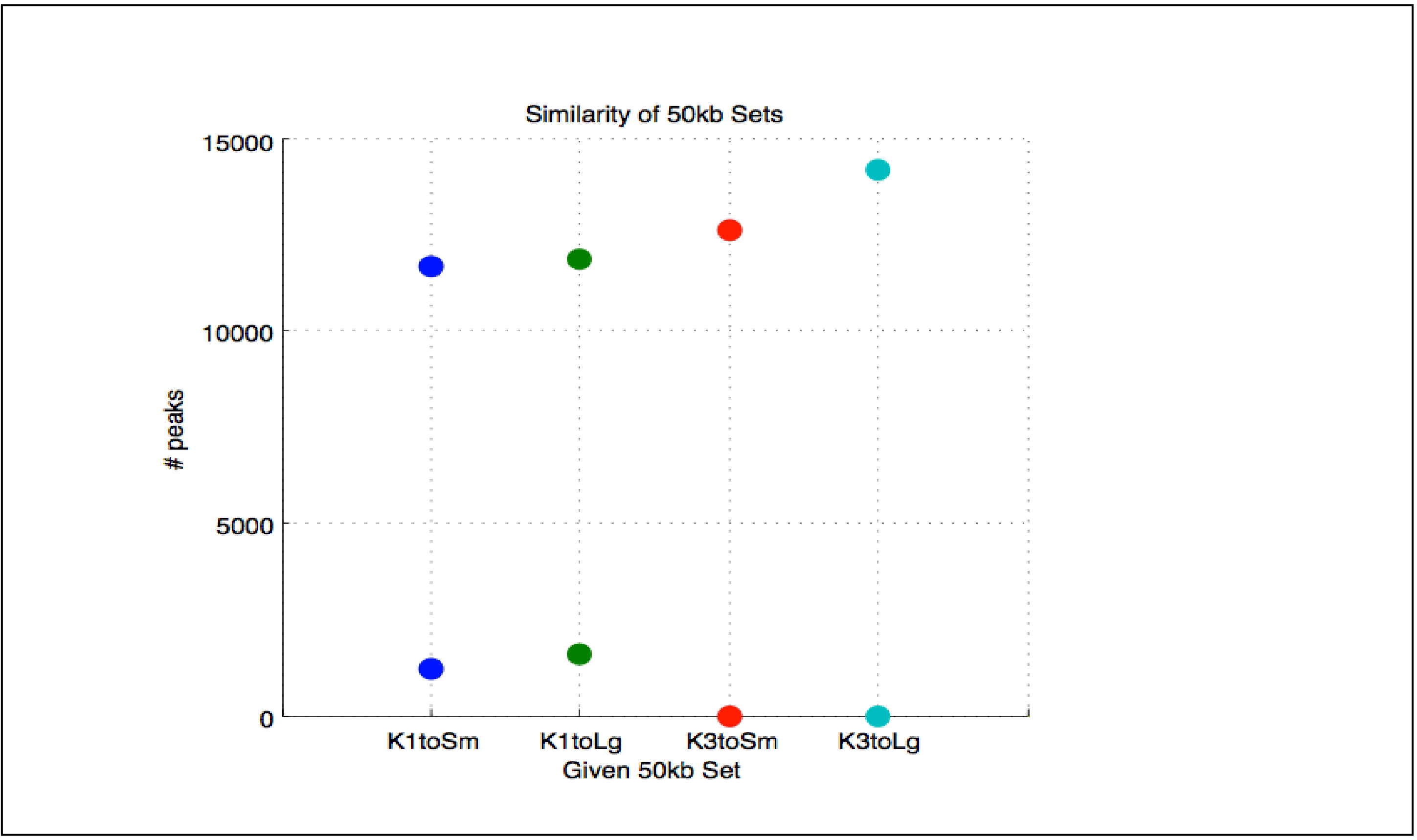


Figure 7. Similarity of the number of peaks regardless of the 50 kb set used. The top row of data points is expected number of false positives. The bottom row of data points are the number of peaks in the smaller set NOT in the largest K3toLg set. Both K1 sets had between 1000 and 2000 peaks that were not in the K3toLg set – numbers much less than the expected number of false peaks and far less than the total number of peaks. The K3toSm set is a proper subset of K3toLg. This is not surprising as all 'toSm' sets considered are proper subsets of their corresponding 'toLg' sets. All sets are therefore considered to be reasonably similar. As the K1 sets had lower FDR, one of these was chosen as our final set. Scaling to small is supposed to have higher specificity and lower FDR. Nonetheless, we do not necessarily see this for the K1 sets. The K1toLg set actually seems to have a lower FDR.

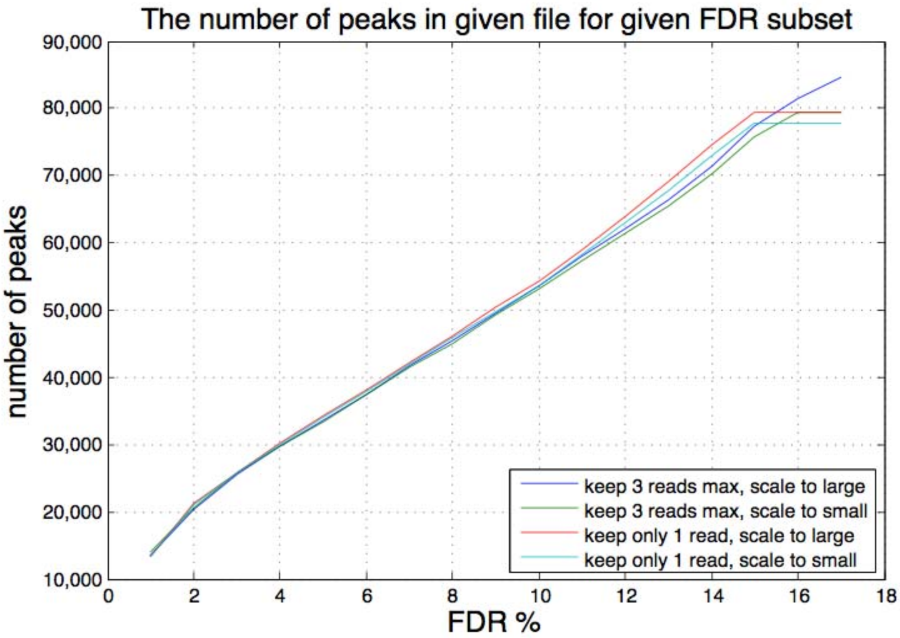


Figure 8. Effect of varying the False Discovery Rate on the number of peaks.

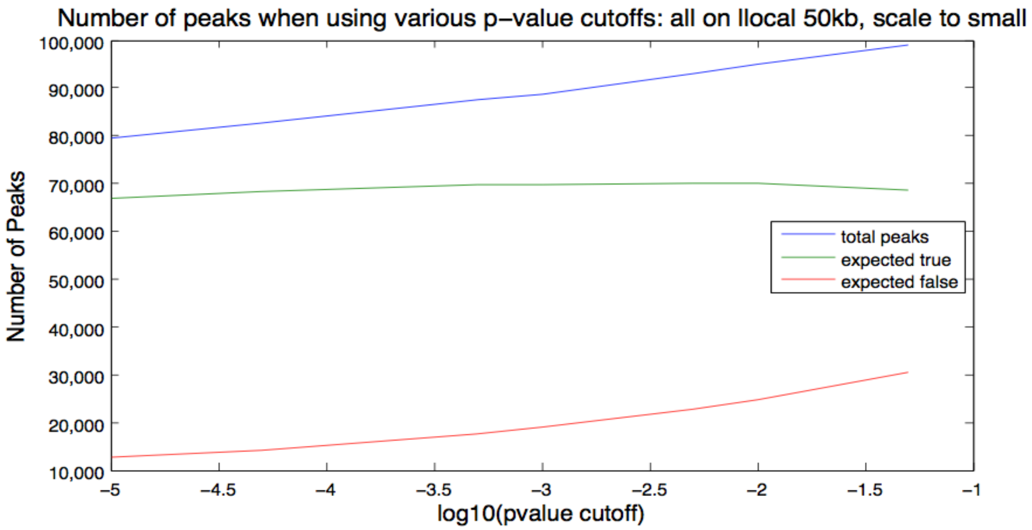


Figure 9. Effect of various p-value cutoffs on the number of peaks.

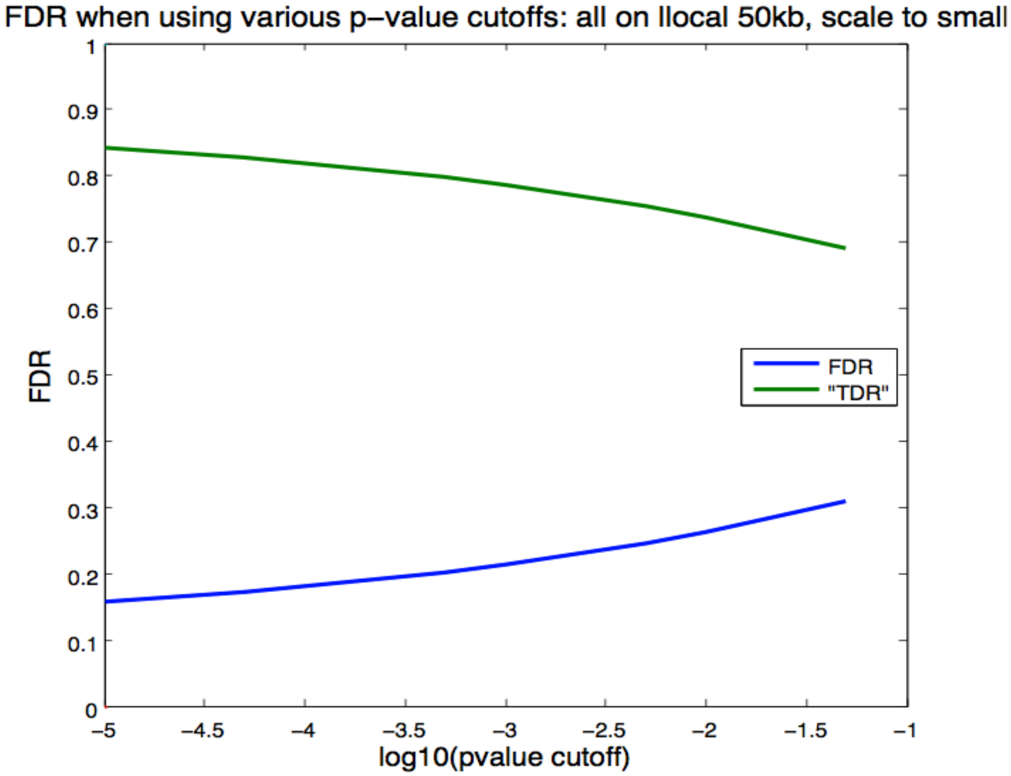
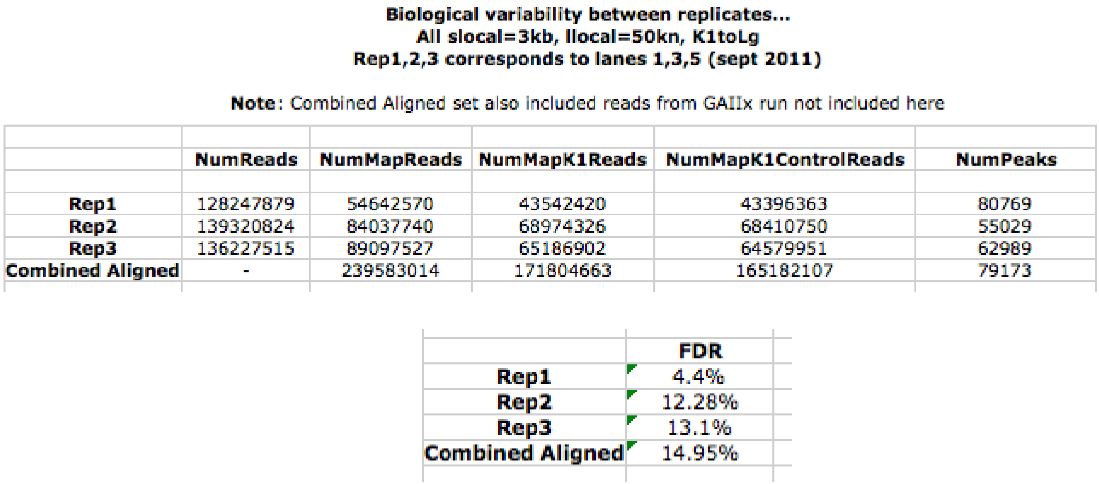


Figure 10. Effect of varying the p-value cutoff on the False Discovery Rate.



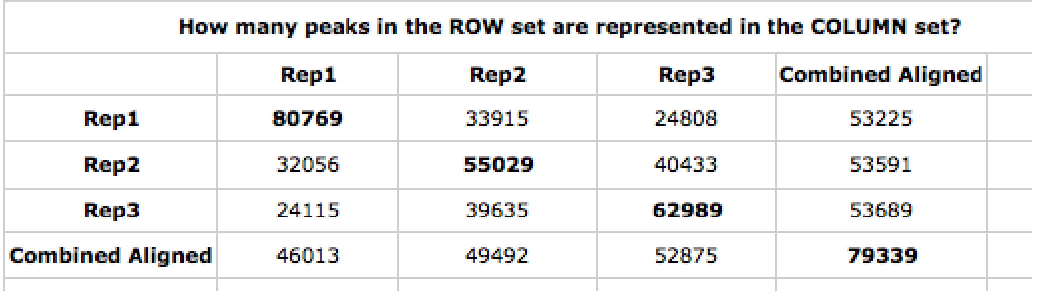
Biological variability between replicates...
All slocal=3kb, llocal=50kn, K1toLg
Rep1,2,3 corresponds to lanes 1,3,5 (sept 2011)

Note: Combined Aligned set also included reads from GAIIx run not included here

	NumReads	NumMapReads	NumMapK1Reads	NumMapK1ControlReads	NumPeaks
Rep1	128247879	54642570	43542420	43396363	80769
Rep2	139320824	84037740	68974326	68410750	55029
Rep3	136227515	89097527	65186902	64579951	62989
Combined Aligned	-	239583014	171804663	165182107	79173

	FDR
Rep1	4.4%
Rep2	12.28%
Rep3	13.1%
Combined Aligned	14.95%

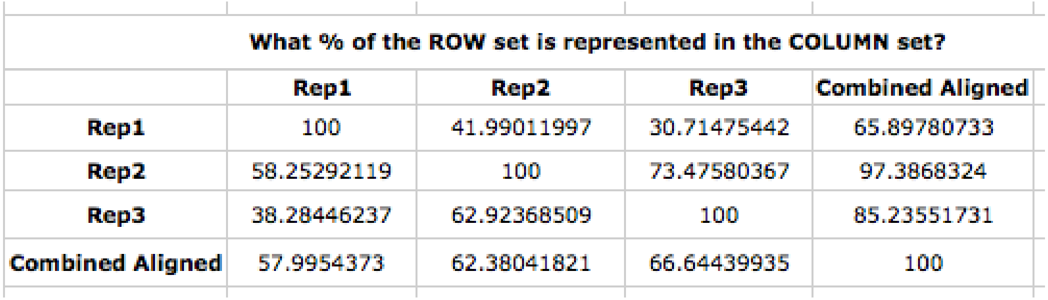
Figure 11. Peaks called from three different samples of MCF-7 nascent DNA.



How many peaks in the ROW set are represented in the COLUMN set?

	Rep1	Rep2	Rep3	Combined Aligned
Rep1	**80769**	33915	24808	53225
Rep2	32056	**55029**	40433	53591
Rep3	24115	39635	**62989**	53689
Combined Aligned	46013	49492	52875	**79339**

Figure 12. Shared peaks by the different samples of MCF-7 nascent DNA.



What % of the ROW set is represented in the COLUMN set?

	Rep1	Rep2	Rep3	Combined Aligned
Rep1	100	41.99011997	30.71475442	65.89780733
Rep2	58.25292119	100	73.47580367	97.3868324
Rep3	38.28446237	62.92368509	100	85.23551731
Combined Aligned	57.9954373	62.38041821	66.64439935	100

Figure 13. Percent peak overlap in different samples of MCF-7 nascent DNA.

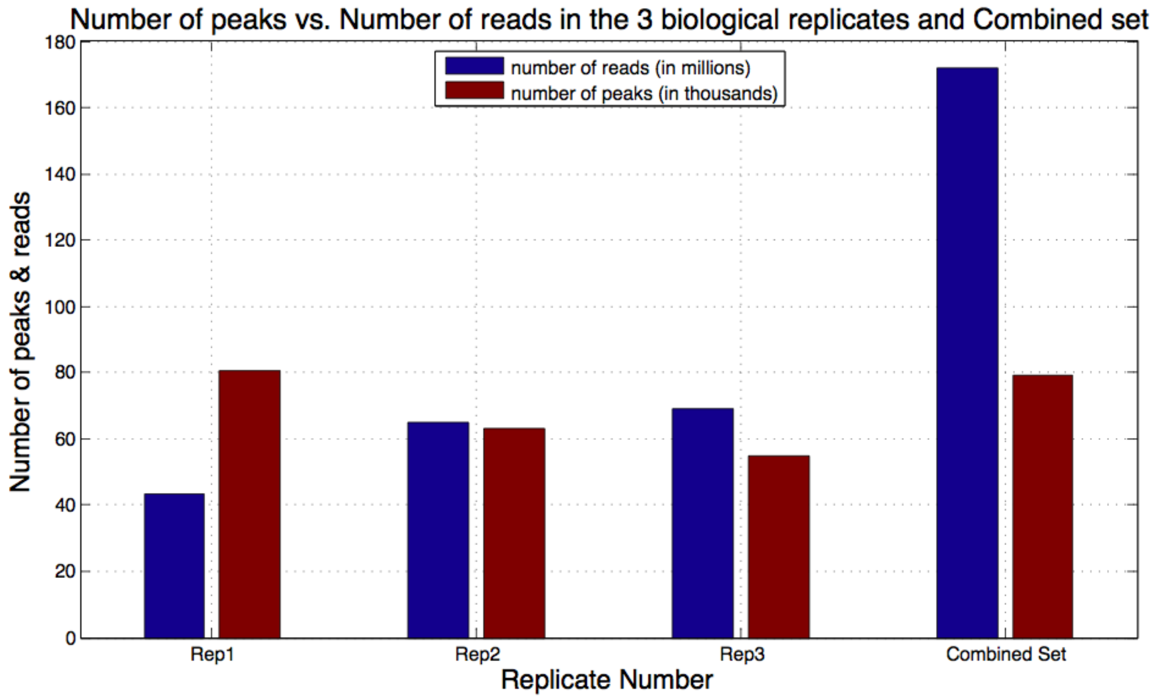


Figure 14. Number of peaks vs number of reads to suggest saturation.

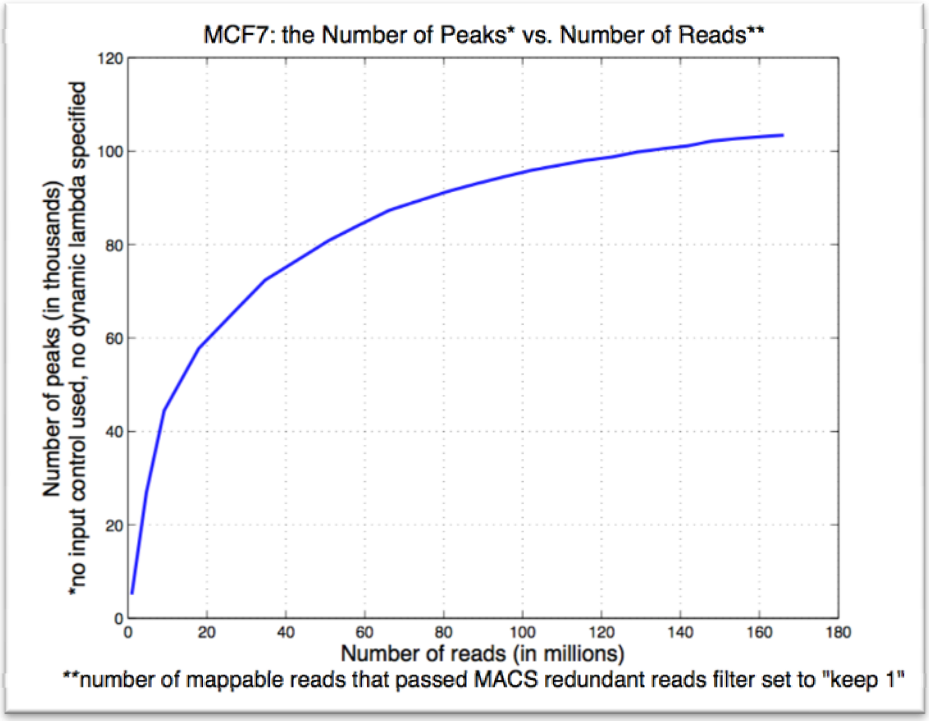


Figure 15. Saturation curve of number of peaks vs number of reads.

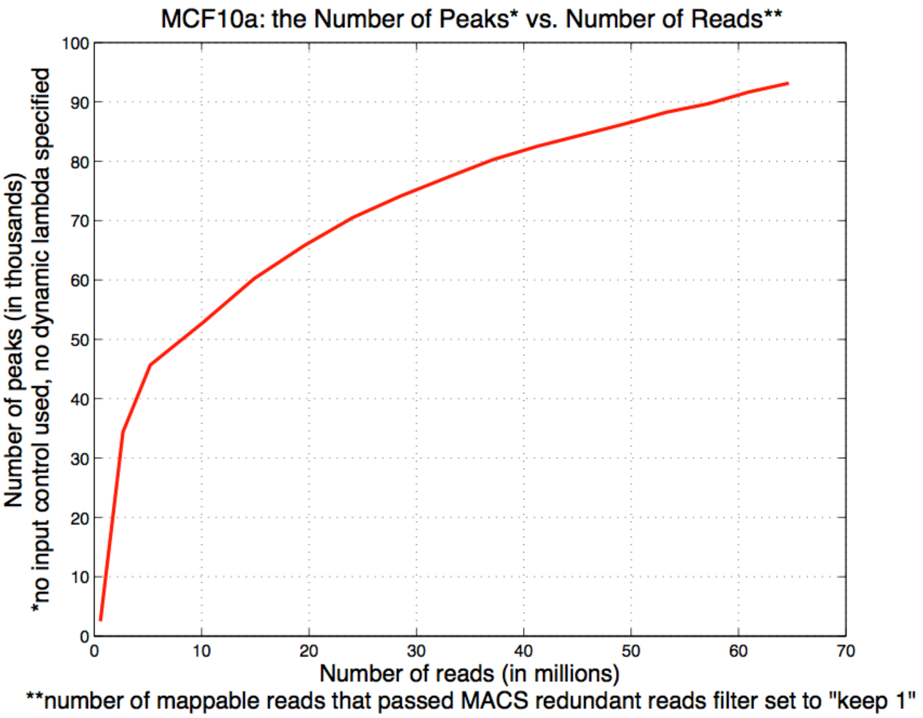


Figure 16. NS-Seq on MCF-10A replicating DNA approaches saturation.

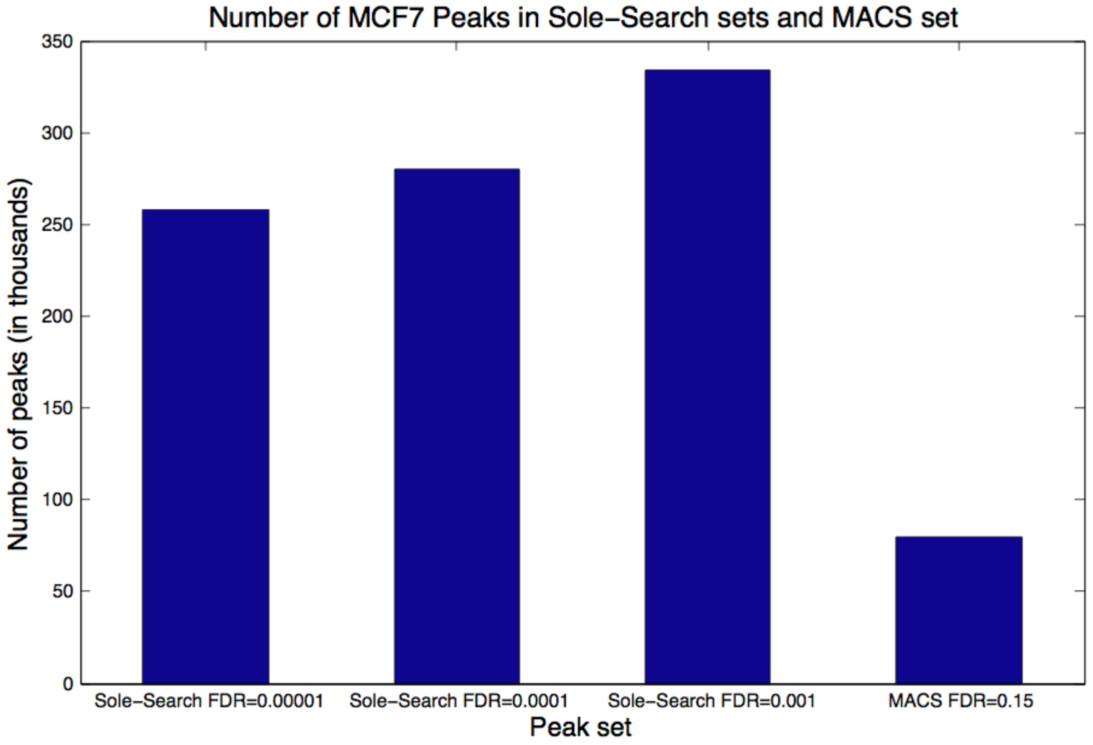


Figure 17. More MCF-7 peaks with Sole-Search than with MACS.

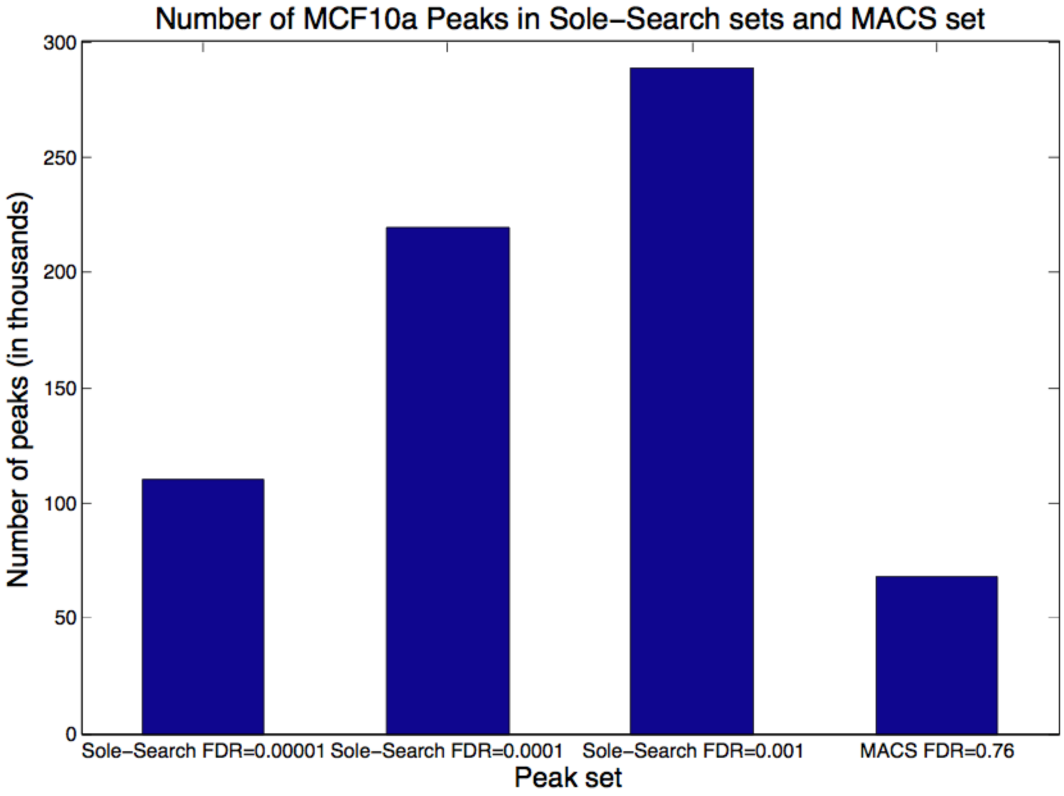


Figure 18. More MCF-10A peaks with Sole-Search than with MACS.

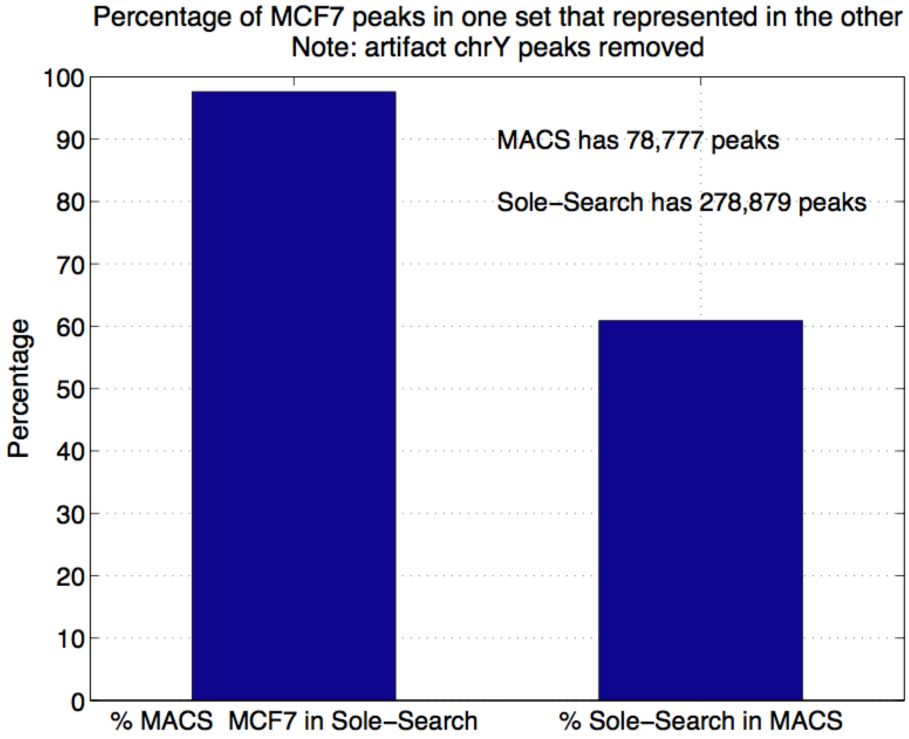


Figure 19. Percentage of MCF-7 peaks shared by MACS and Sole-Search.

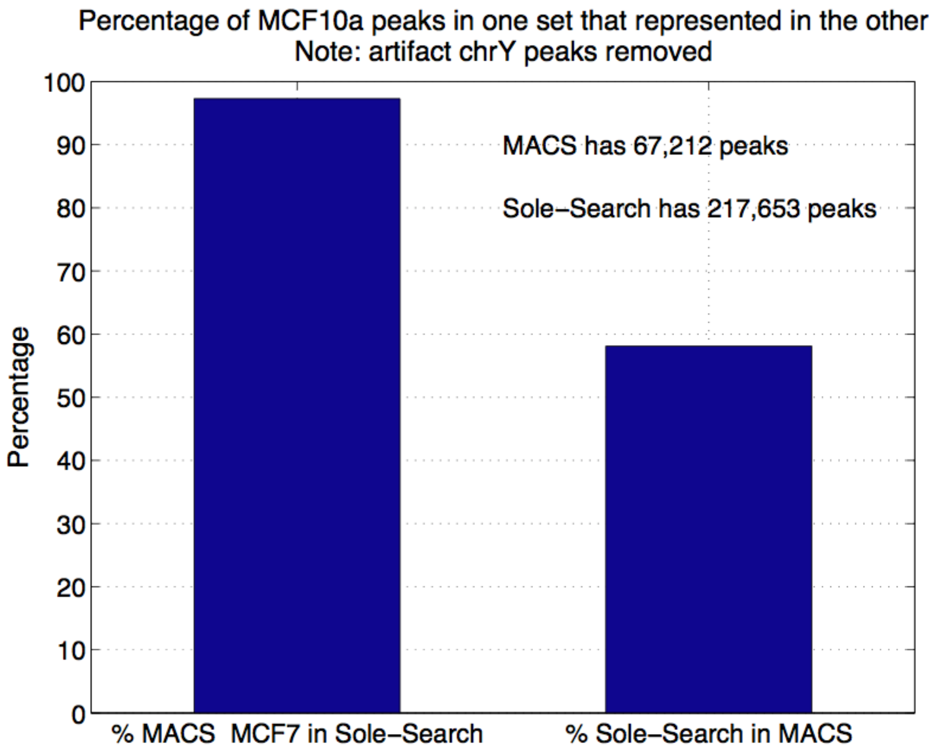


Figure 20. Percentage of MCF-10A peaks shared by MACS and Sole-Search.

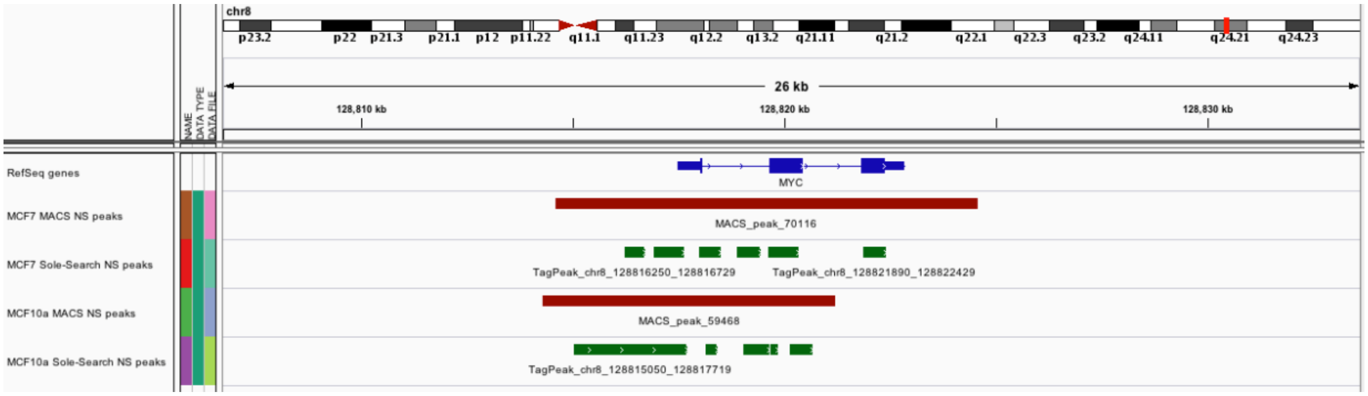


Figure 21. Peaks in the Myc locus called with MACS and with Sole-Search.

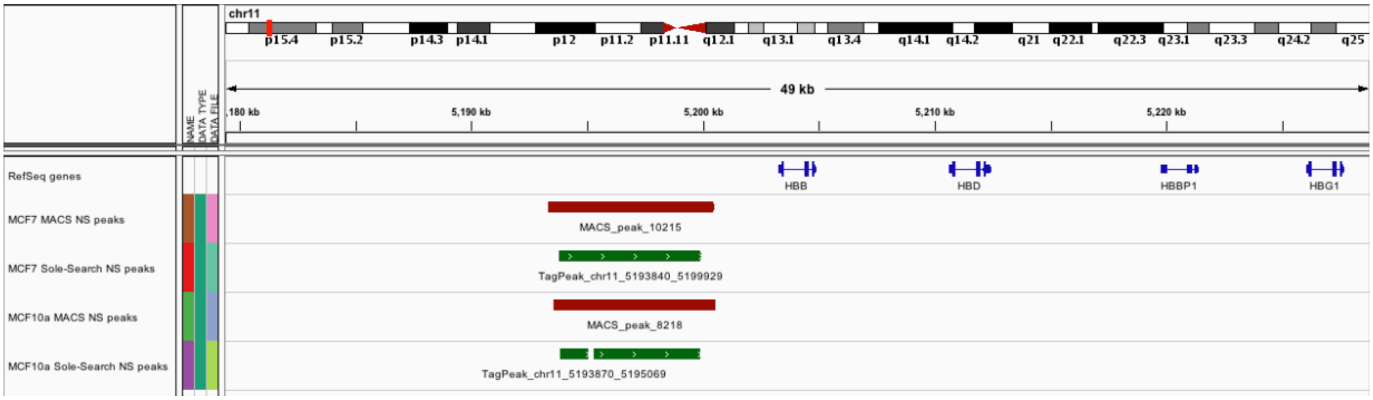


Figure 22. Peaks in the HBB locus called with MACS and with Sole-Search.

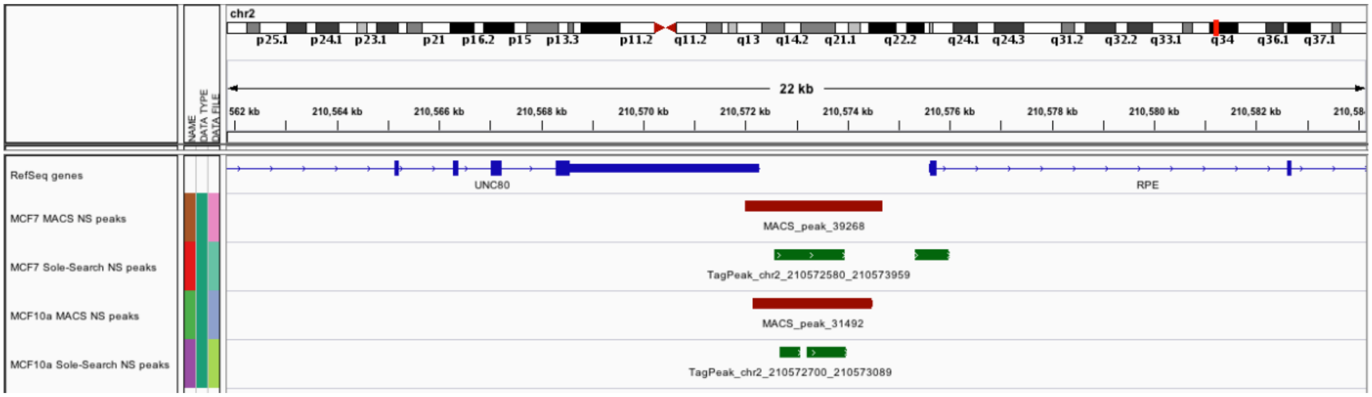

chr2
p25.1 p24.1 p23.1 p21 p16.2 p15 p13.3 p11.2 q11.2 q13 q14.2 q21.1 q22.2 q24.1 q24.3 q31.2 q32.2 q33.1 q34 q36.1 q37.1
NAME
DATA TYPE
DATA FILE
22 kb
562 kb
210,564 kb
210,566 kb
210,568 kb
210,570 kb
210,572 kb
210,574 kb
210,576 kb
210,578 kb
210,580 kb
210,582 kb
210,58
RefSeq genes
UNC80
RPE
MCF7 MACS NS peaks
MACS_peak_39268
MCF7 Sole-Search NS peaks
TagPeak_chr2_210572580_210573959
MCF10a MACS NS peaks
MACS_peak_31492
MCF10a Sole-Search NS peaks
TagPeak_chr2_210572700_210573089

Figure 23. Peaks in the RPE locus called with MACS and with Sole-Search.

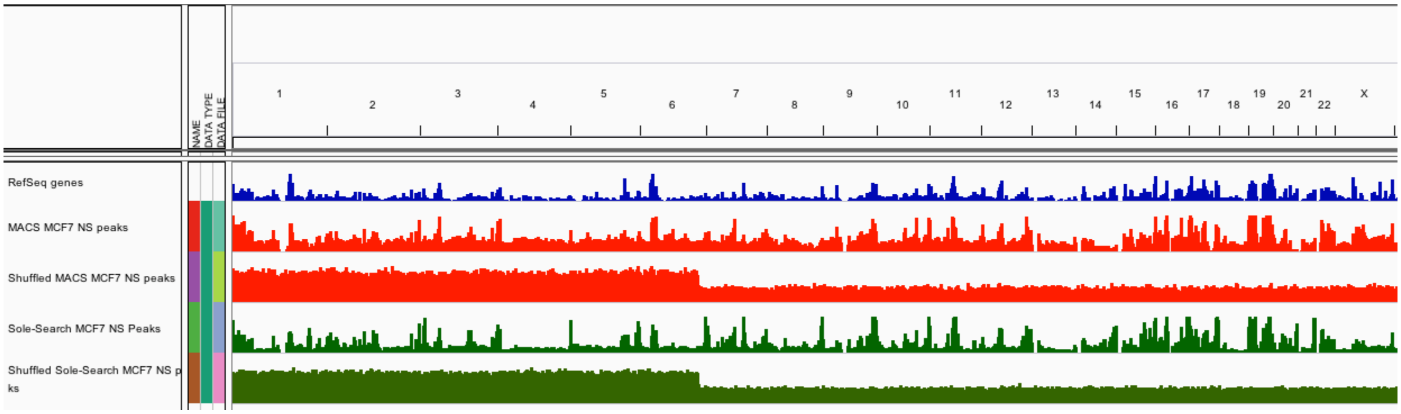

NAME
DATA TYPE
DATA FILE
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 X
RefSeq genes
MACS MCF7 NS peaks
Shuffled MACS MCF7 NS peaks
Sole-Search MCF7 NS Peaks
Shuffled Sole-Search MCF7 NS p
ks

Figure 24. Density of peaks in the human genome called by MACS and by Sole-Search.

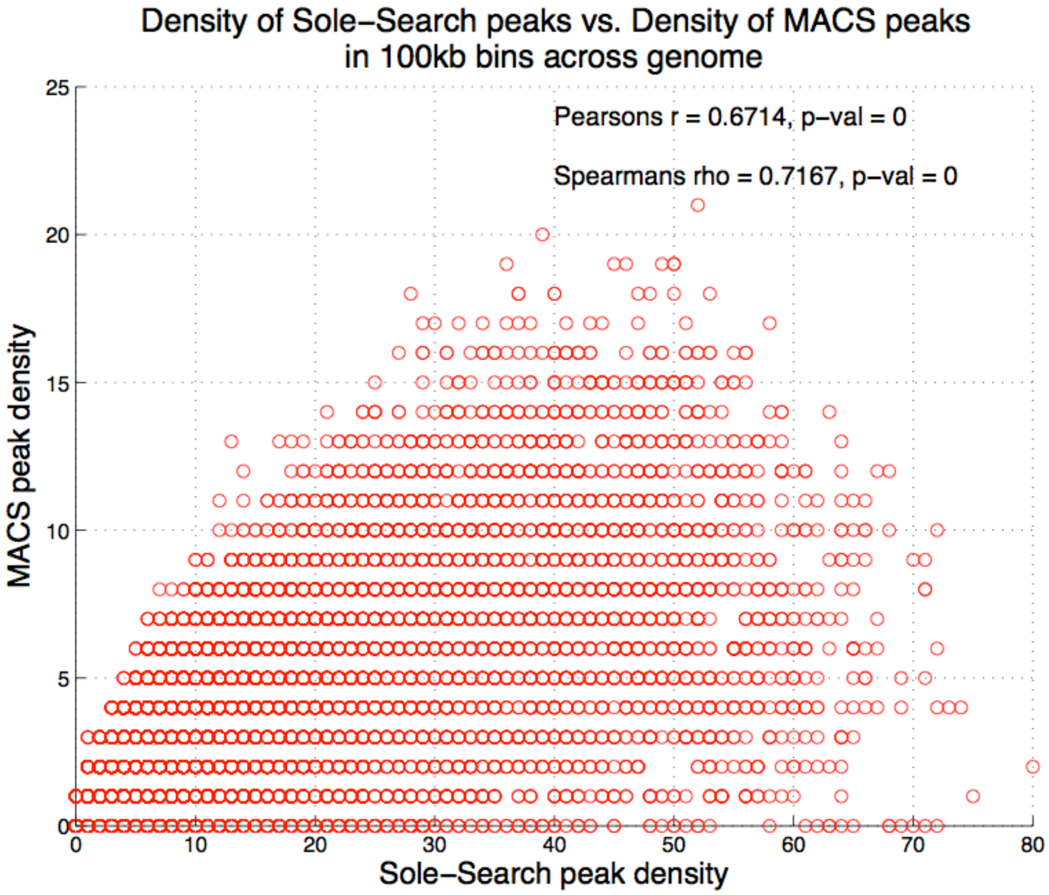


Figure 25. Density of Sole-Seach vs MACS peaks in 100 kb bins across the genome.

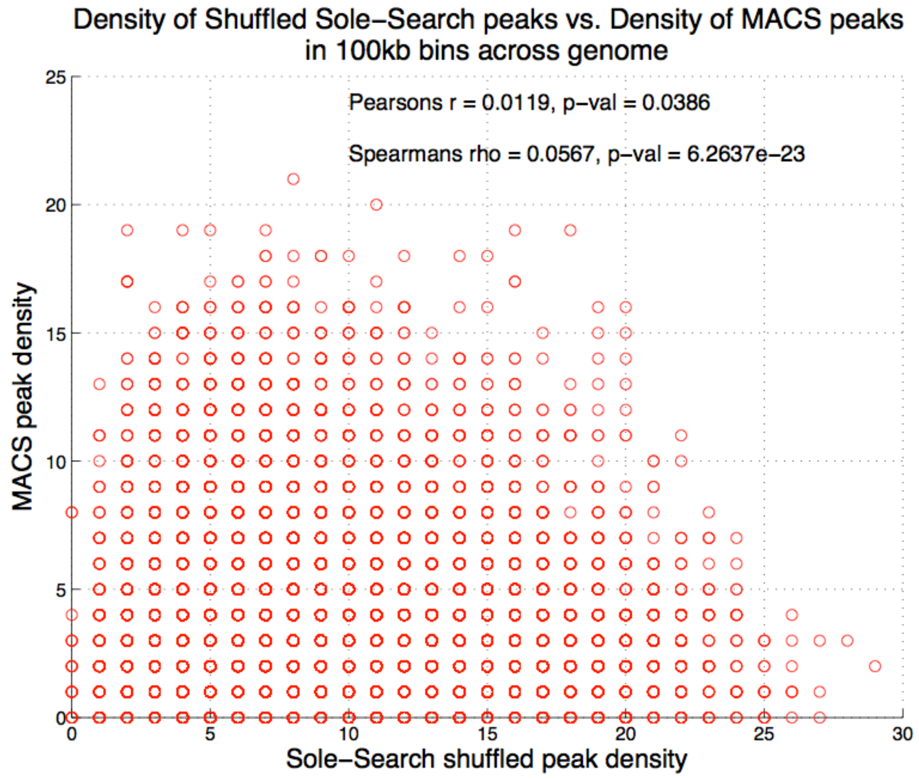


Figure 26. Density of Sole-Seach vs MACS shuffled peaks in 100 kb bins in the genome.

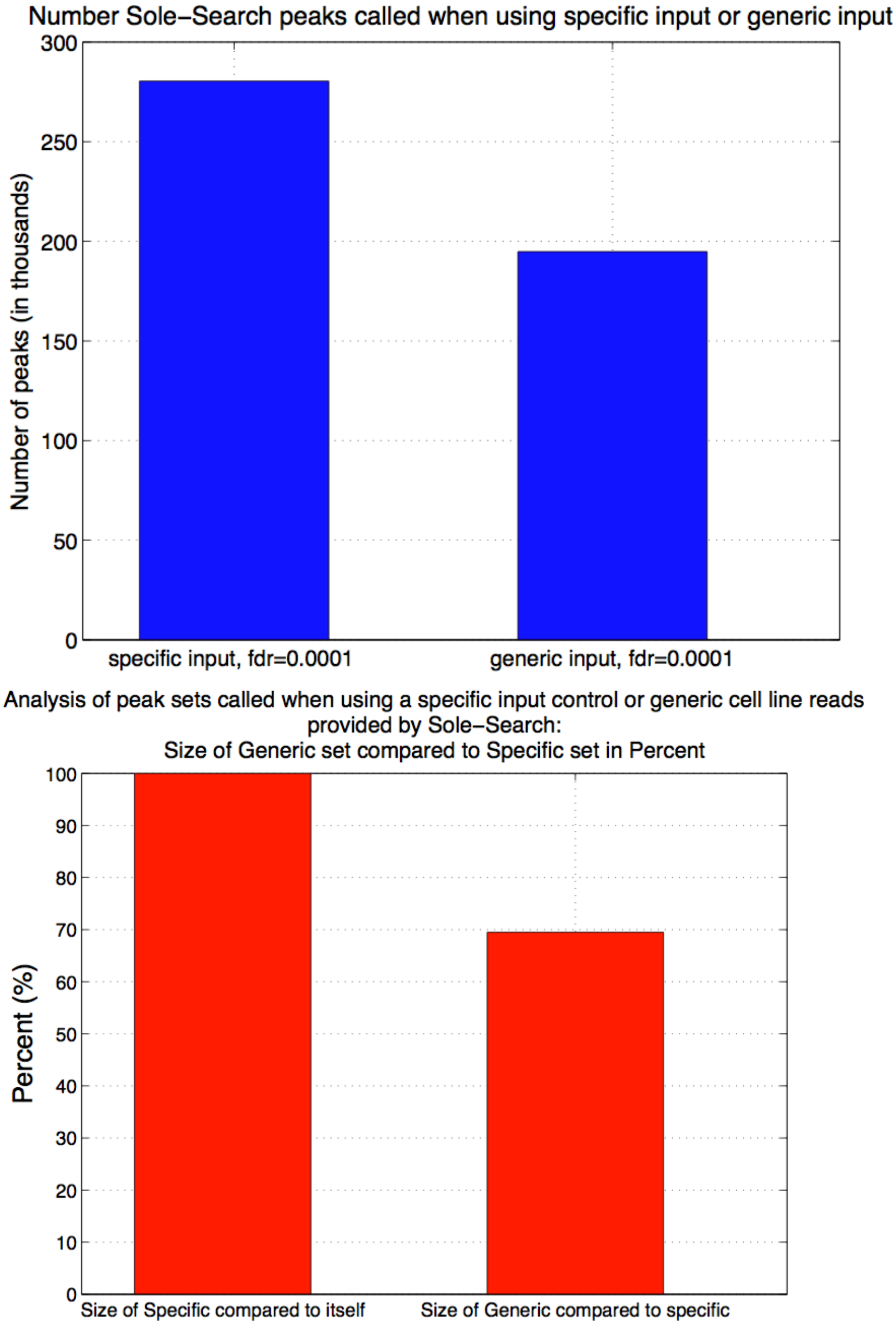


Figure 27. (a) Number or **(b)** percent of Sole-Search peaks using specific or generic input control.

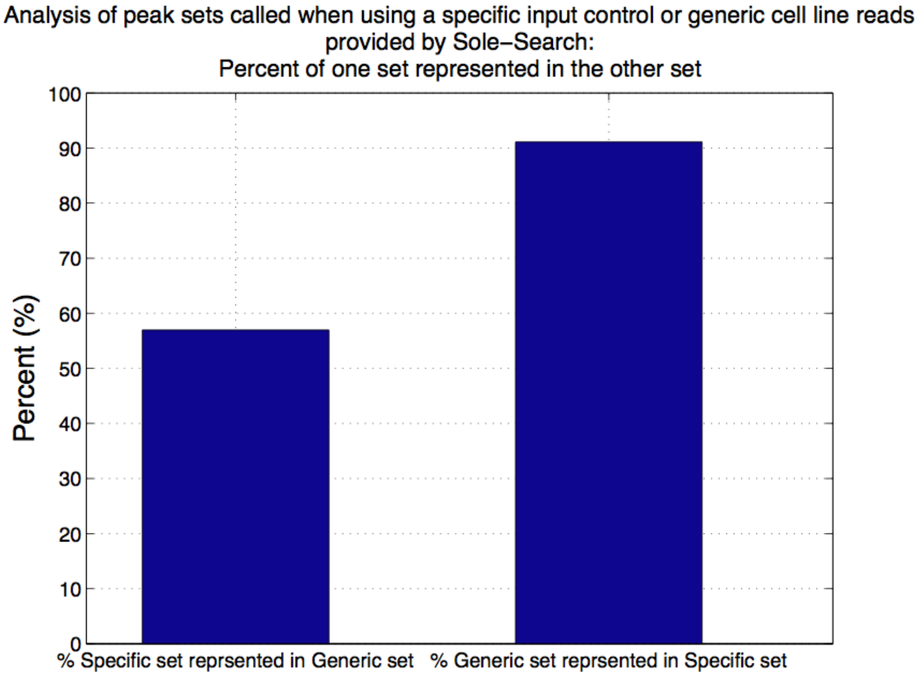


Figure 28. Percent specific set in generic set and vice versa when using a specific input or generic input control with Sole-Search.

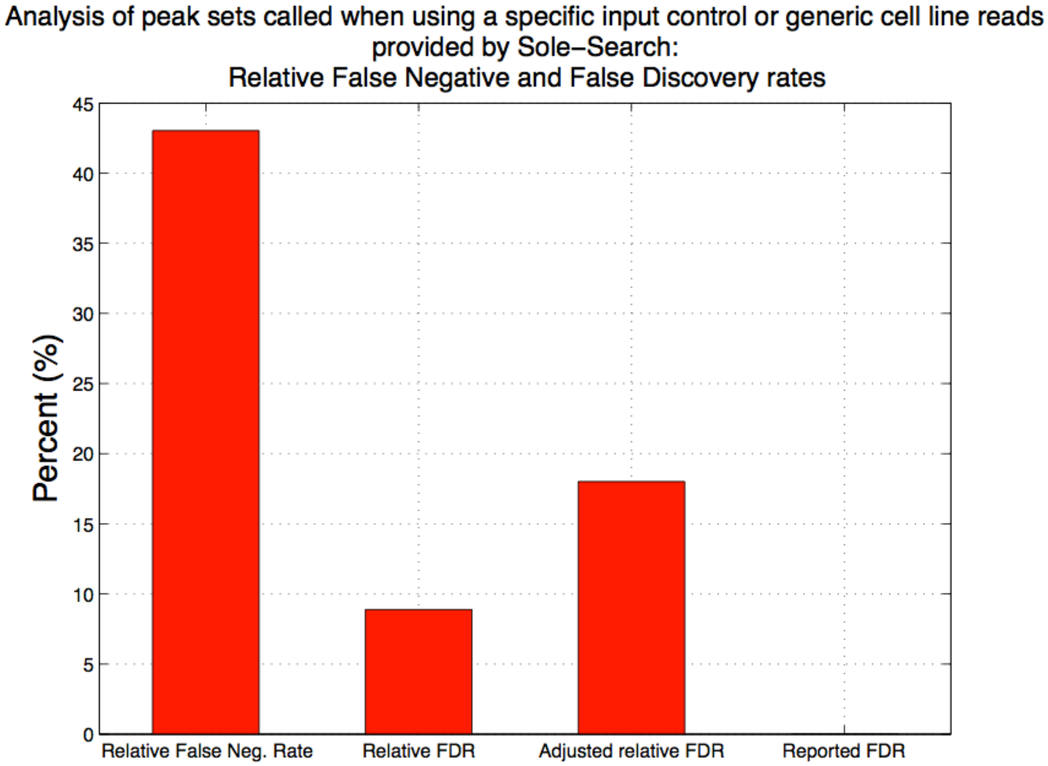


Figure 29. Relative False Negative and False Discovery Rates when using a specific input or generic input control with Sole-Search.

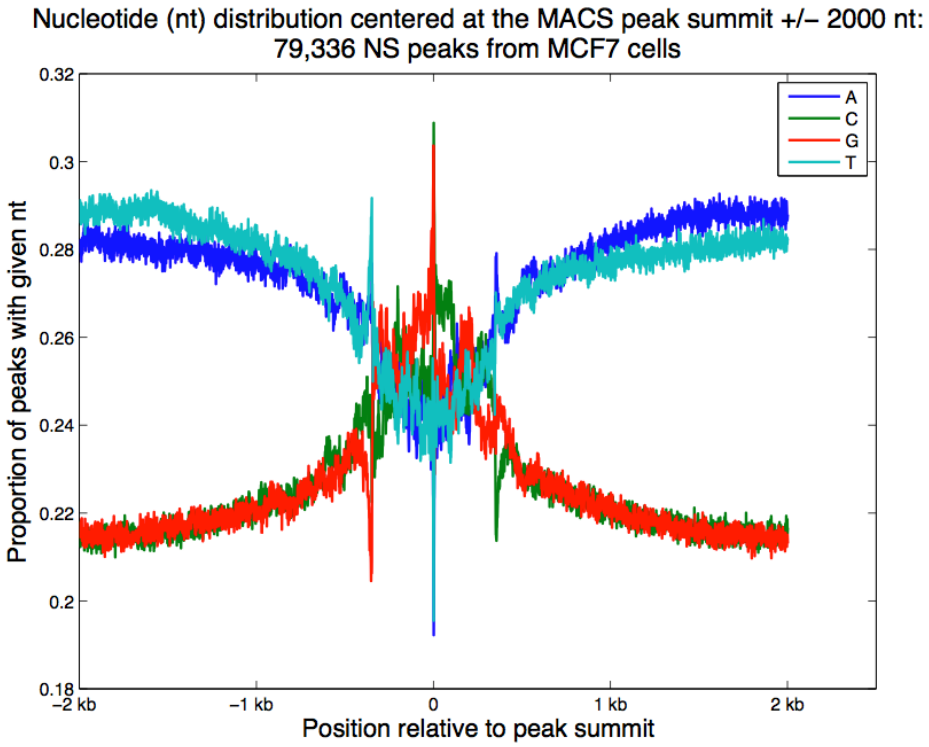


Figure 30. MACS MCF7 NS peak summits

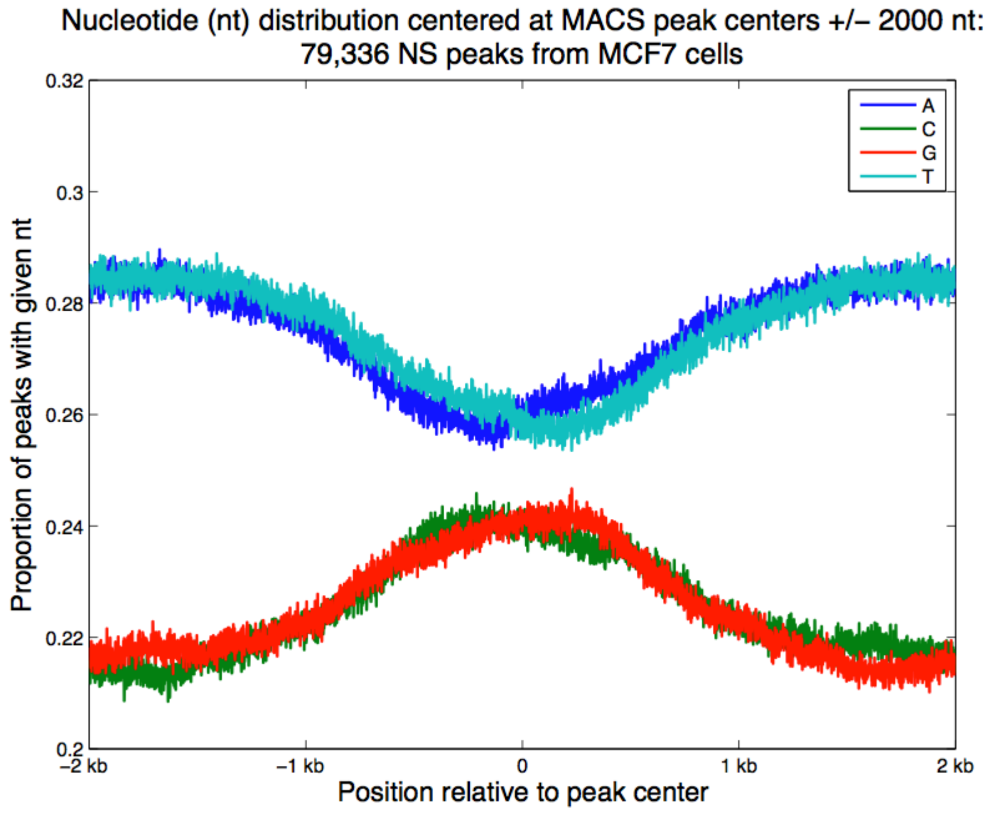


Figure 31. MACS MCF7 NS peak centers

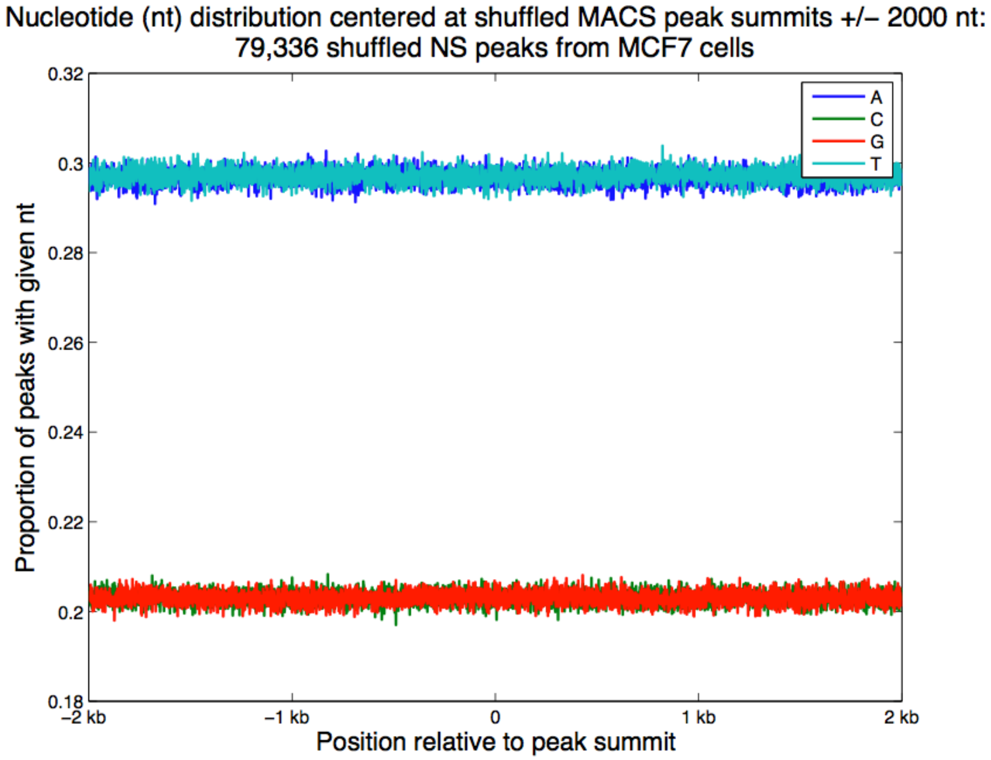


Figure 32. MACS MCF7 shuffled NS peak summits

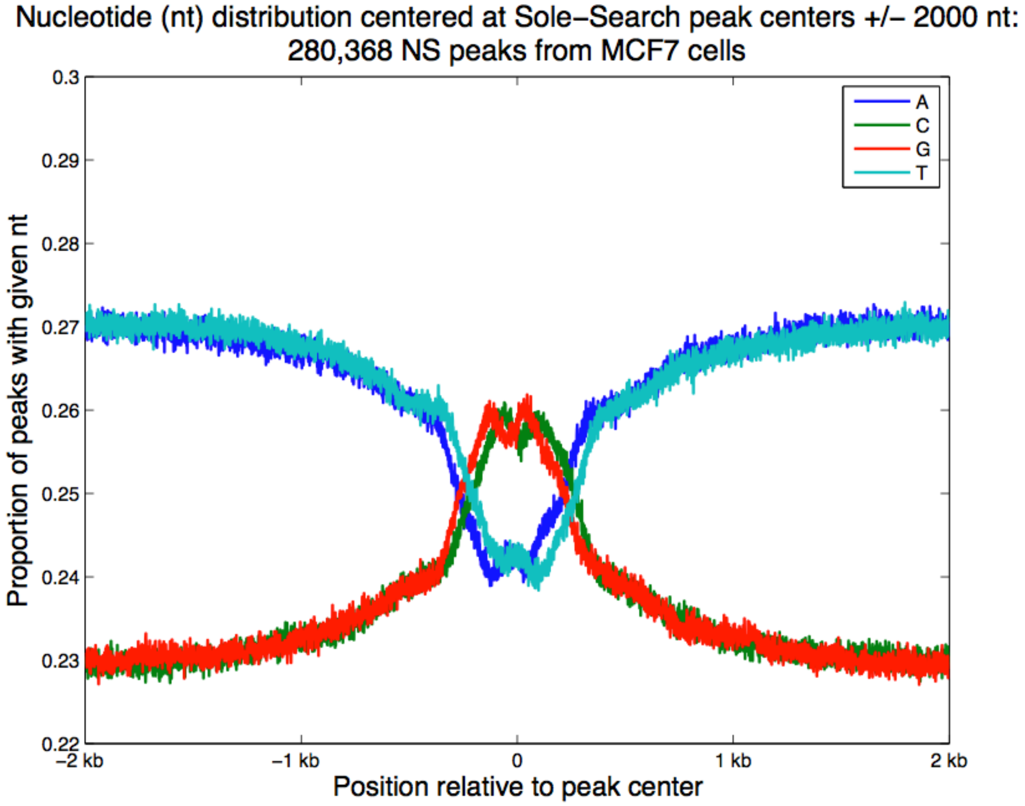


Figure 33. Sole-Search MCF7 NS peak centers

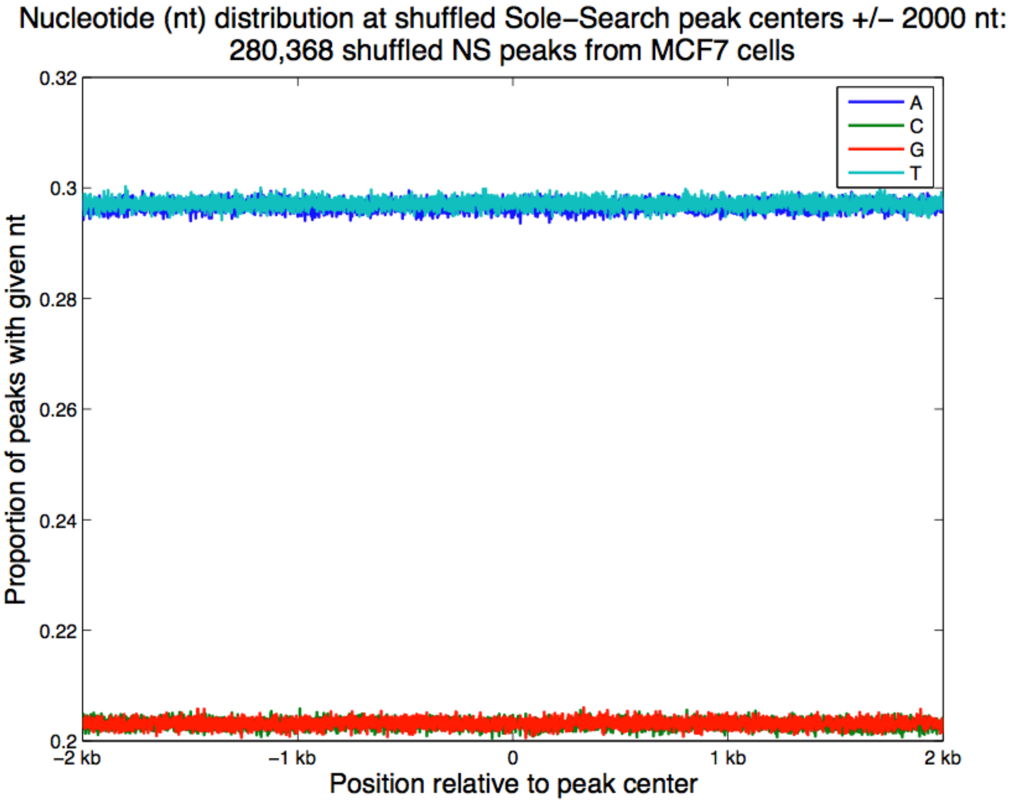


Figure 34. Sole-Search MCF7 shuffled NS peak centers

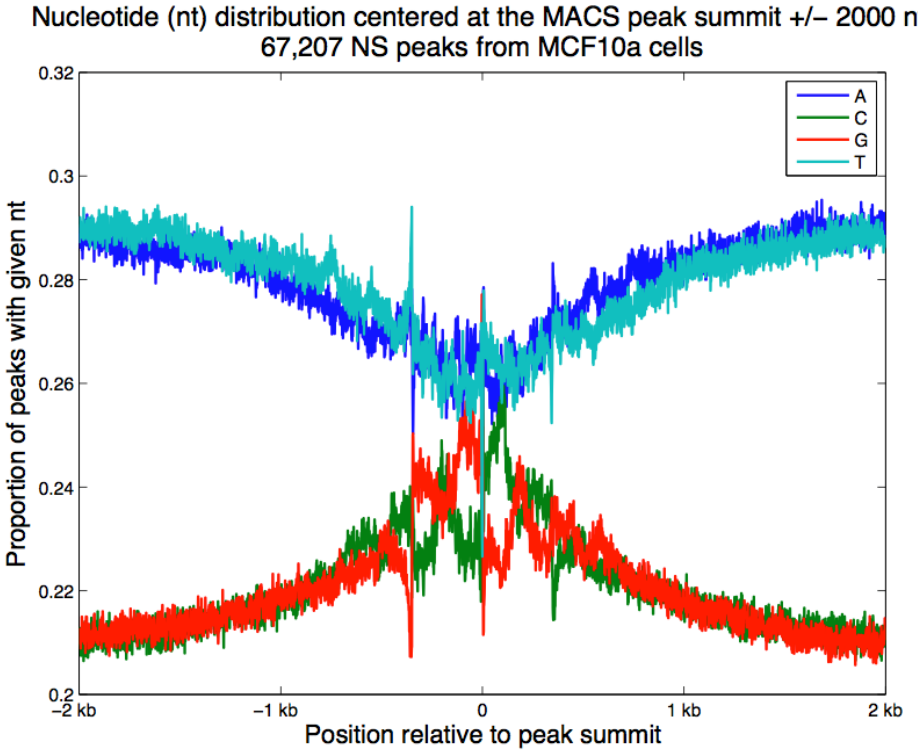


Figure 35. MACS MCF-10A NS peak summits

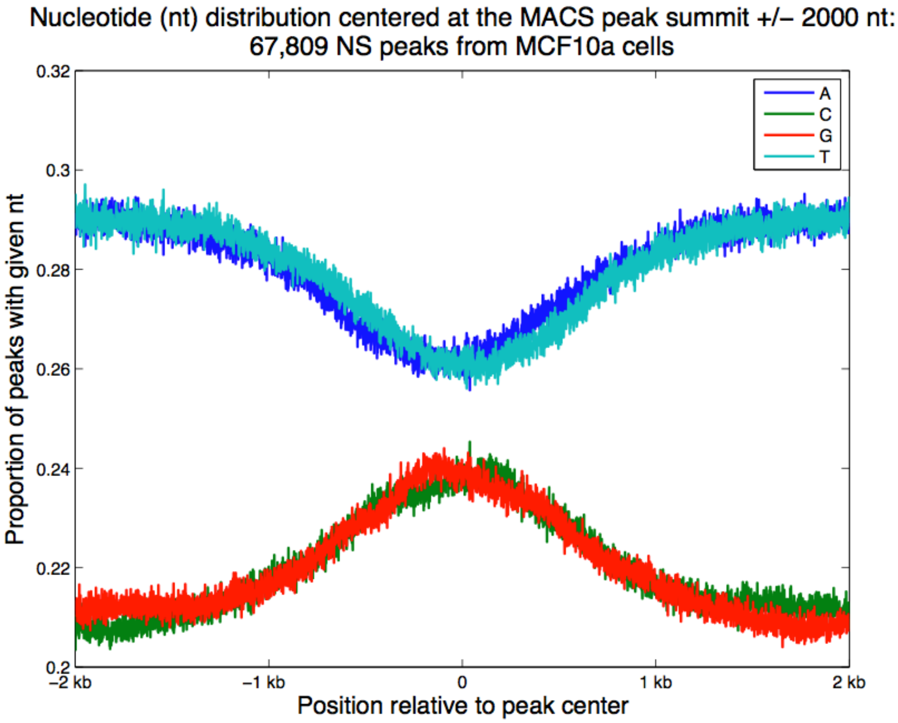


Figure 36. MACS MCF-10A peak centers

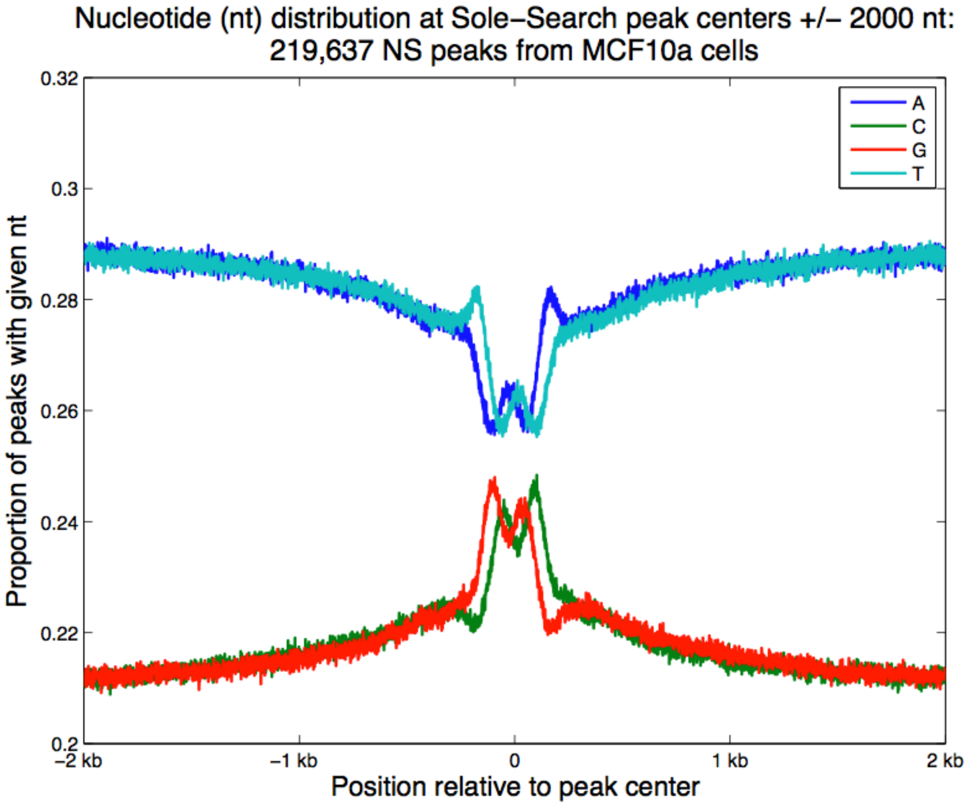


Figure 37. Sole-Search MCF-10A peak centers

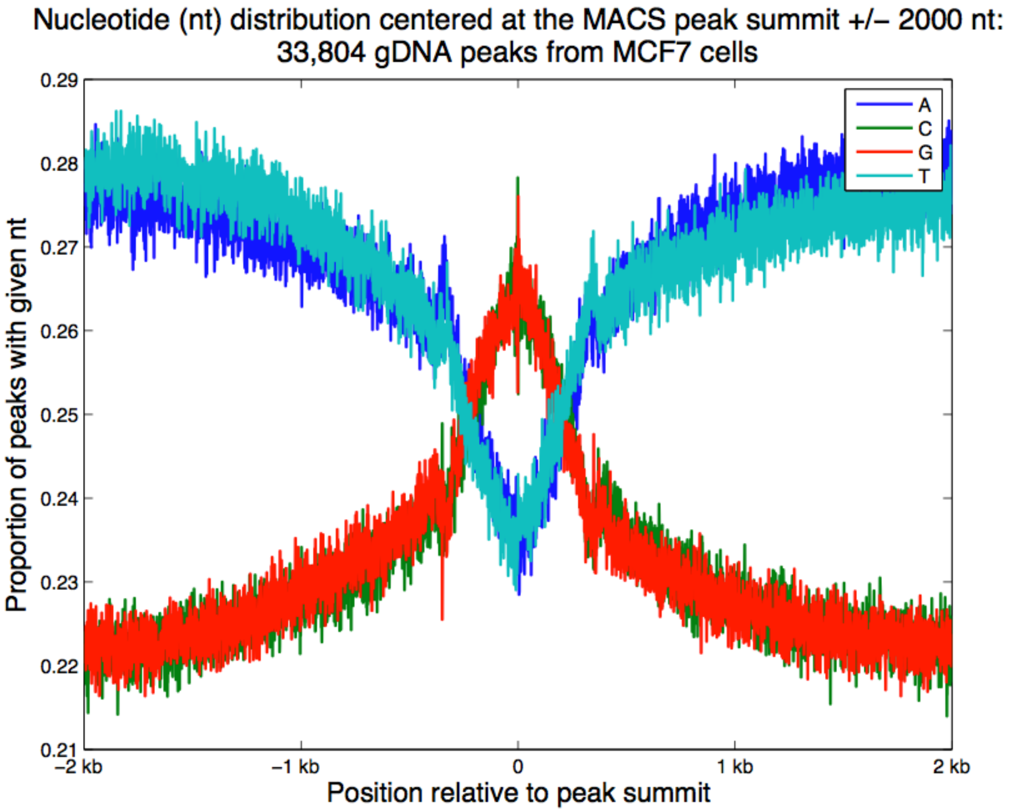


Figure 38. GC skew at peak summits from nonreplicating genomic DNA input

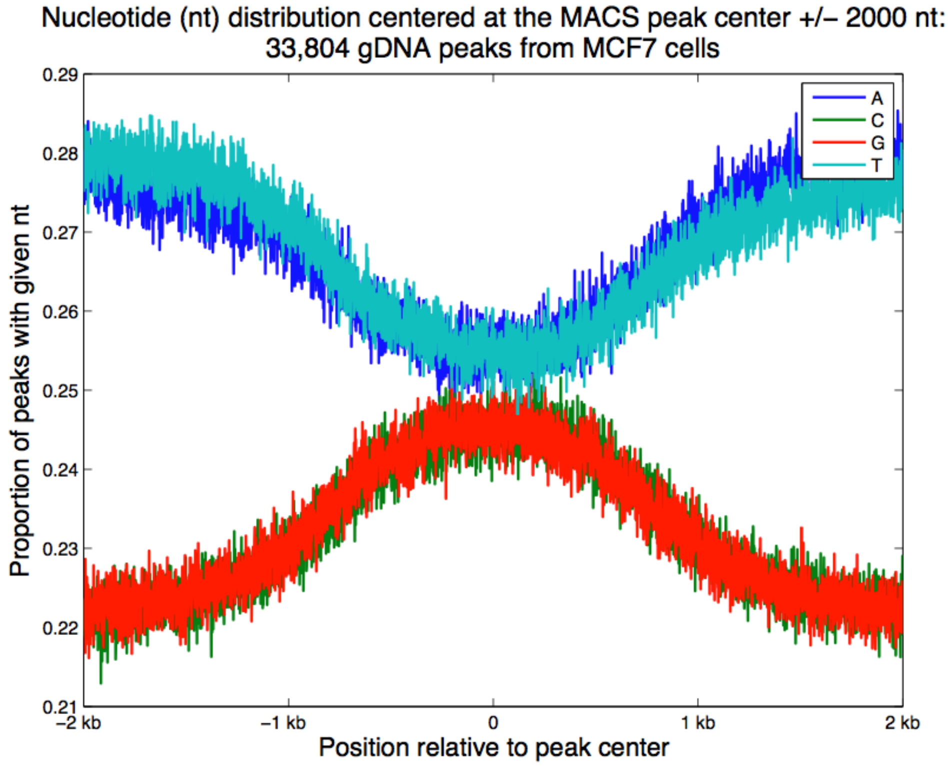


Figure 39. GC skew at peak centers from nonreplicating genomic DNA input

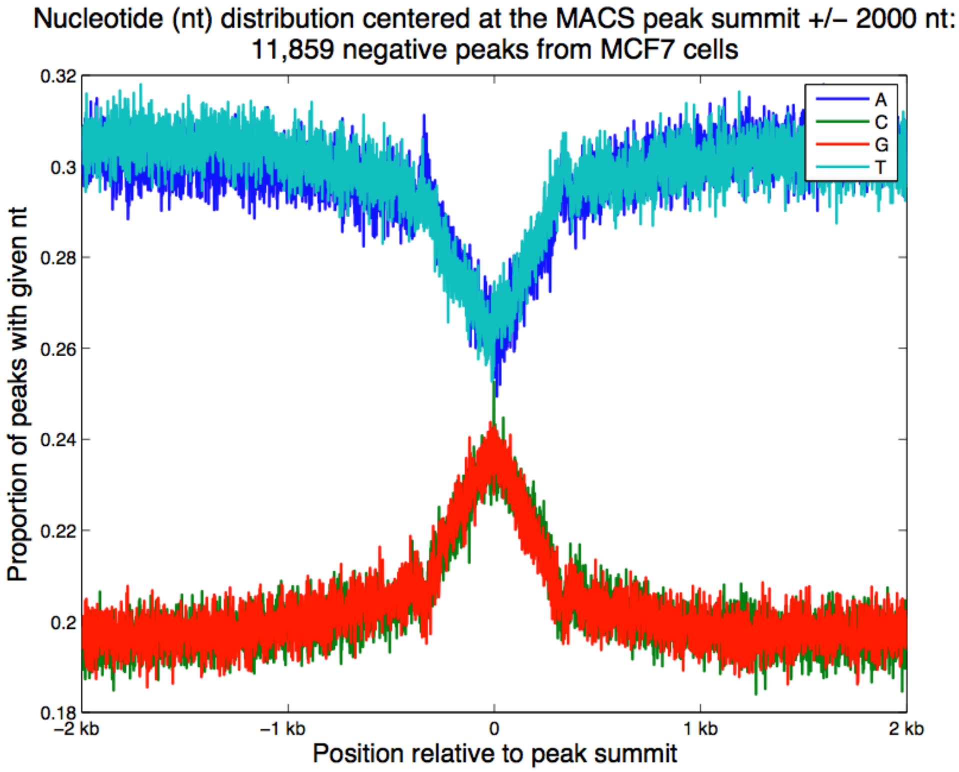


Figure 40. GC skew at peak summits from MCF-7 negative peaks

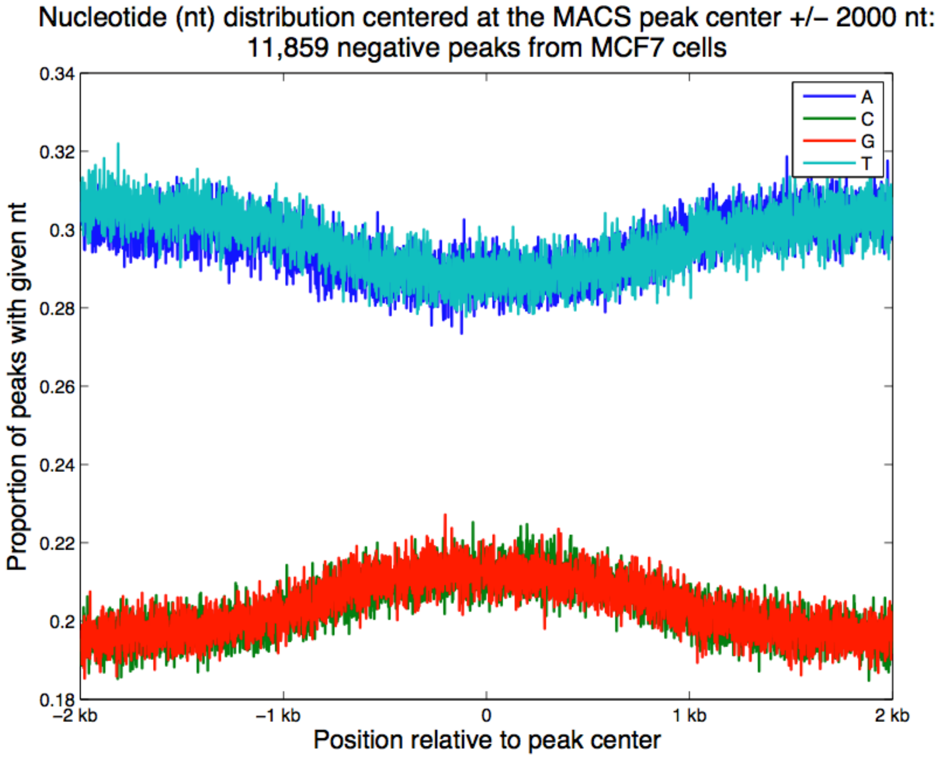


Figure 41. GC skew at peak centers from MCF-7 negative peaks

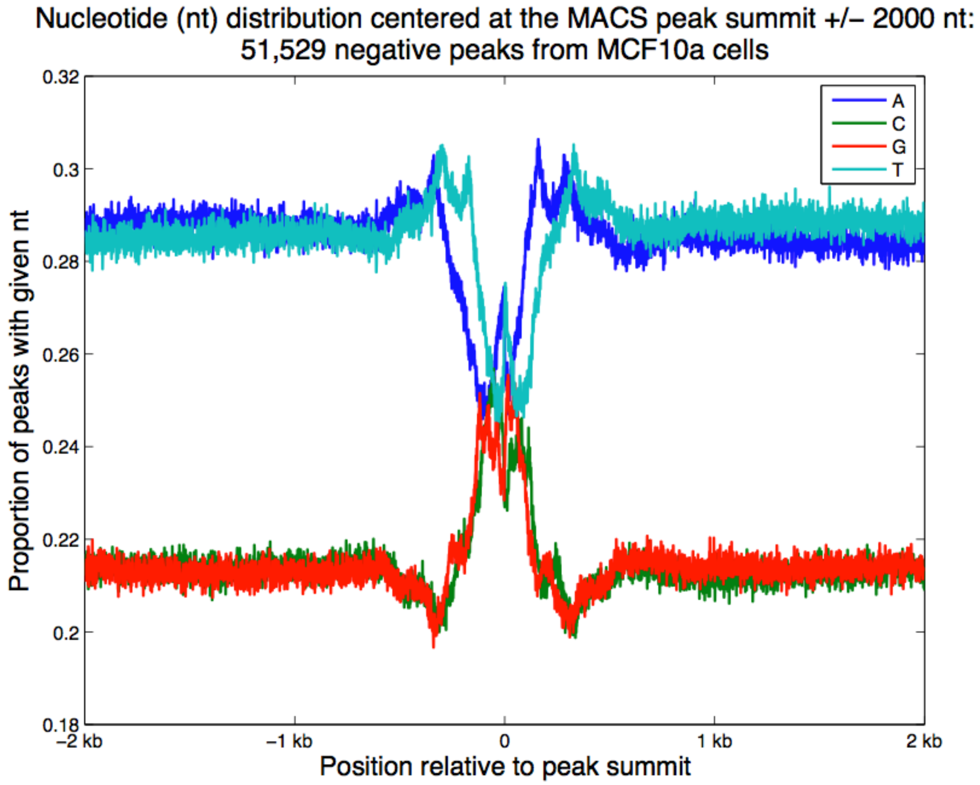


Figure 42. GC skew at peak summits from MCF-10A negative peaks

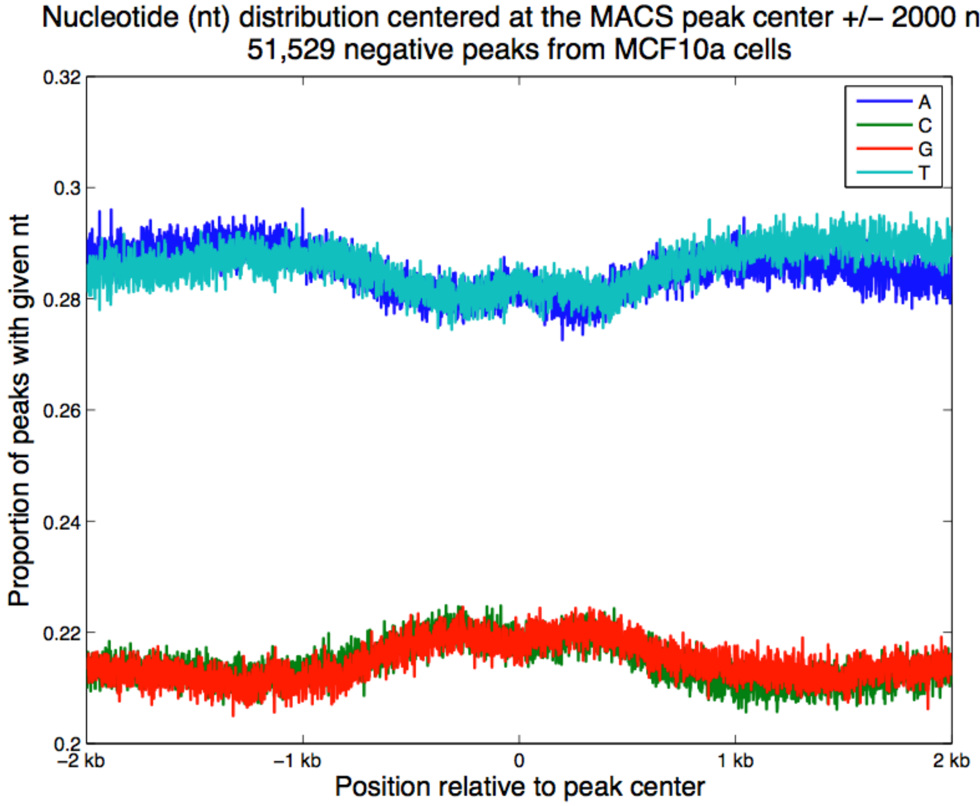


Figure 43. GC skew at peak centers from MCF-10A negative peaks

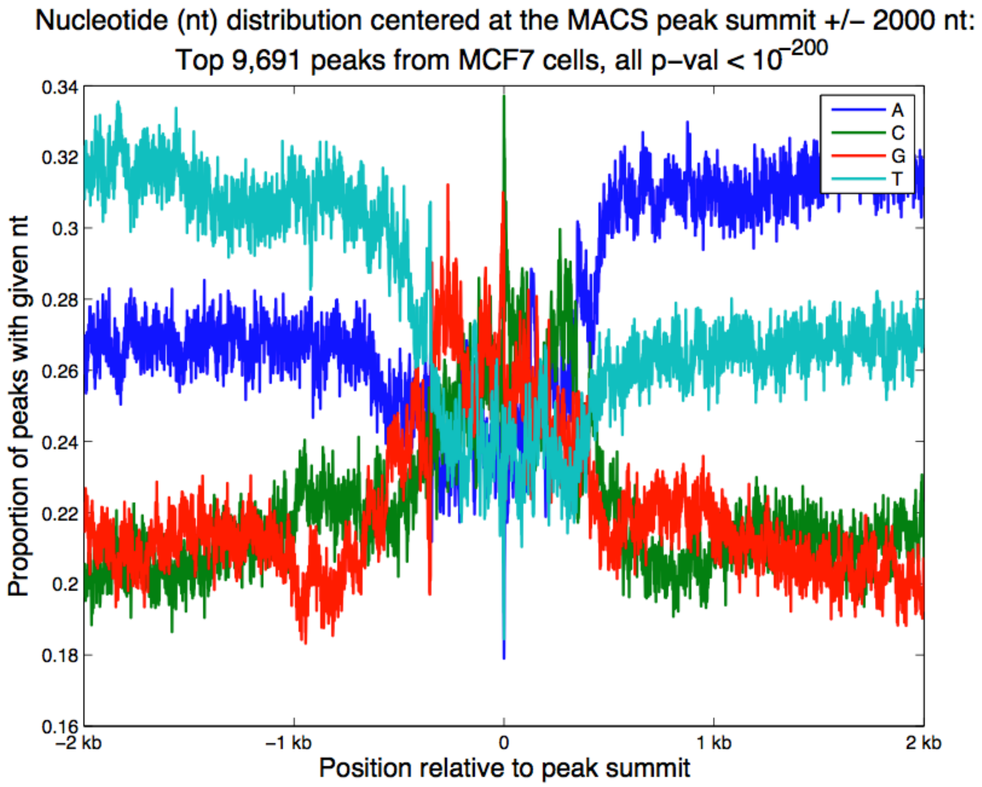


Figure 44. Base composition at peak summit of top 9691 MCF-7 Nascent Strand peaks.

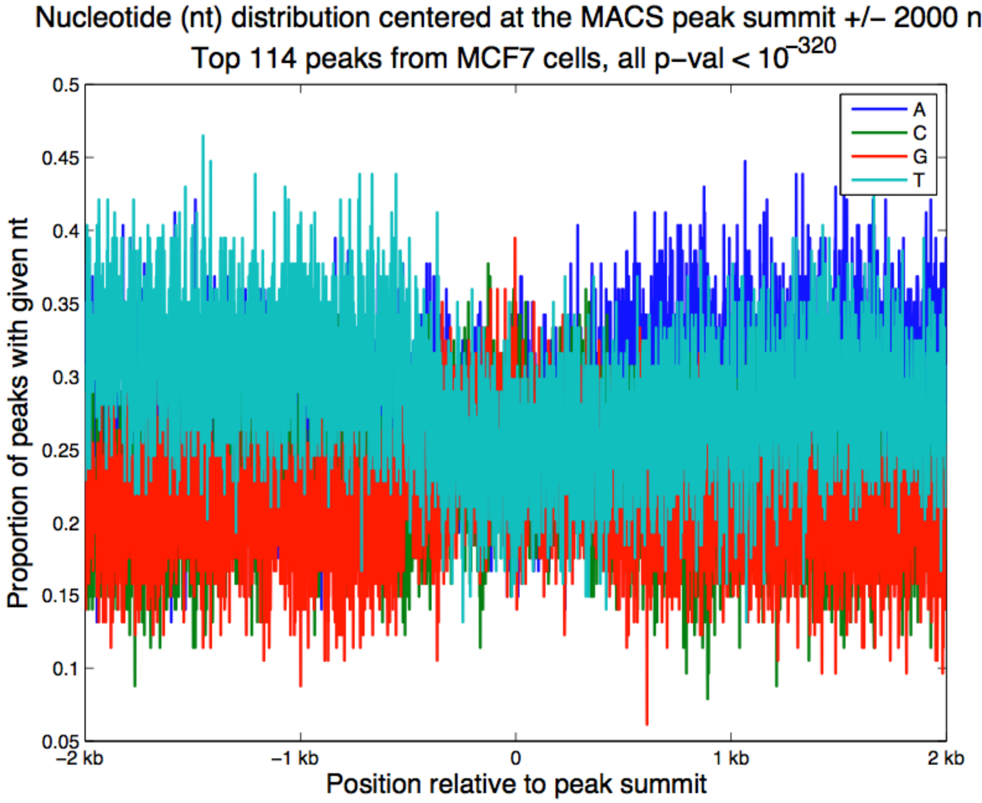


Figure 45. Base composition at peak summit of best 114 MCF-7 Nascent Strand peaks.