UNIVERSITÀ DEGLI STUDI DI TORINO

Area Ricerca e Relazioni Internazionali

SEZIONE RICERCA E FORMAZIONE AVANZATA
Via Bogino, 9 – 10123 Torino
Tel +39/(0)11/670.4373/670.4388/670.4371 – Fax 011-670.4380

DOTTORATO DI RICERCA IN

'COMPLEXITY IN POST-GENOMIC BIOLOGY'

CICLO: XXI

TITOLO DELLA TESI:

Identification and characterization of human replication origins on chromosome 19

Tesi presentata da: ***CESARONI MATTEO***

Tutor Interno: ***Prof. RAFFAELE CALOGERO***

Tutor esterno: ***D.ssa LUCILLA LUZI***

Coordinatore del ciclo: ***Prof. FEDERICO BUSSOLINO***

ANNI ACCADEMICI: 2005-2008

SETTORE SCIENTIFICO-DISCIPLINARE DI AFFERENZA: BIO-11

Table of contents

Abstract

Transmission of genetic information from one cell generation to the next requires the accurate duplication of the genome. Replication initiation is a well-conserved process determined in all eukaryotes by the binding of the pre-Replication Complex (ORC and MCM proteins) to replication origins. Despite the early success in the mapping of budding yeast replication origins, only few origins have been identified in mammals. We have developed a novel strategy to identify human replication origins based on ultracentrifugation in equilibrium density gradient of sheared crosslinked chromatin. Our results show that known replication origins are enriched in high-density fractions, containing naked DNA, and proteins of pre-replication complex are enriched in low-density fractions. In order to identify and characterize new replication origins we hybridized a custom tiled oligonucleotide array of the human chromosome 19 (NimbleGen Technology) with: i) origin-rich naked DNA, ii) DNA purified from ChIP assays using antibodies against proteins of the pre-replication complex and iii) double strand cDNAs to estimate the expression of genes involved in the replication process. To analyze all the data coming from these experiments we implemented a tool called CARPET (Collection of Automated Routine Programs for Easy Tiling). CARPET is a set of Perl, Python and R scripts, integrated on the Galaxy2 web-based platform, for the analysis of ChIP-chip and expression tiling data. The tool allows:

- quick experimental data entry;
- simple quality control of the data creating a simulation-image of the chip surface;
- easy identification and annotation of enriched ChIP-chip regions;

- detection of the absolute or relative transcriptional status of genes assessed by expression tiling experiments
- direct integration of ChIP-chip and expression data.

Since CARPET is integrated on Galaxy2, the results can be directly visualized in a genomic context within the UCSC genome browser as graph-based custom tracks. Both generated and uploaded data can be stored in sessions and easily shared with other users. The combined analysis of different hybridizations we performed allows us the identification and characterization of about 101 candidate replication origins. With an independent assay (nascent strand abundance assay) we have proved that 90% of these regions are newly identified replication origins.

Introduction

The origin of origins

Chromosomal DNA replication is one of the fundamental processes preceding cell division, but our understanding of how it occurs and how it is regulated is still far from complete. In bacteria, the regulation of DNA replication occurs at the level of the DNA sequence as well characterized DNA sequences identify the initiation sites of DNA replication. Whereas also in yeast the essential six-protein Origin Recognition Complex (ORC) has a specific activity for an origin consensus sequence, in metazoans (and human) ORC shows no sequence specificity and no origin consensus has been identified in their genomes.

The first replicon model was proposed by Jacob et al. in 1963 (Jacob and Brenner 1963) who hypothesized that specific cis-acting elements (DNA sequences) were activated upon their interaction with trans-acting elements (regulatory proteins) to initiate DNA replication. According to this model, then, the DNA sequence determines the initiation site of DNA replication, and the protein complex drives the triggering mechanism for the start of new DNA synthesis.

This initial model has been refined and widened, but the basic concept is still true, as already mentioned for prokaryotes and lower eukaryotes, but in mammals (particularly in the human genome) the situation is still far to be clear, because other factors equally contribute to determine if, when and where the DNA replication begins.

Replication Origins and Origin-Binding Proteins

Origins

The replication origin is defined as the initiation point of DNA synthesis (Bell and Dutta 2002), (Diffley 2004). It is surprising that, despite the importance of this process, cellular organisms show a great deal of variety in the mechanisms by which they ensure appropriate genome duplication (Robinson and Bell 2005). The simplest origin, called oriC, has been studied in bacteria (Escherichia coli) and it has been shown that typically a single origin exists per bacterial chromosome (Robinson and Bell 2005).

The oriC region (~250bp) is composed of multiple binding site (5) for DnaA protein, an AT-rich sequence and other binding sites for additional factors (Kaguni 2006). The binding of DnaA protein to the DNA mediates the local distortion of DNA, resulting in oligomerization of DnaA itself that melts the AT-rich region. The opened “bubble” of single-stranded DNA is then accessible to the helicase DnaB that, through the interaction with another protein (DnaC), acts as replicative helicase (Davey, Fang et al. 2002; Messer 2002) (Fig. 1).

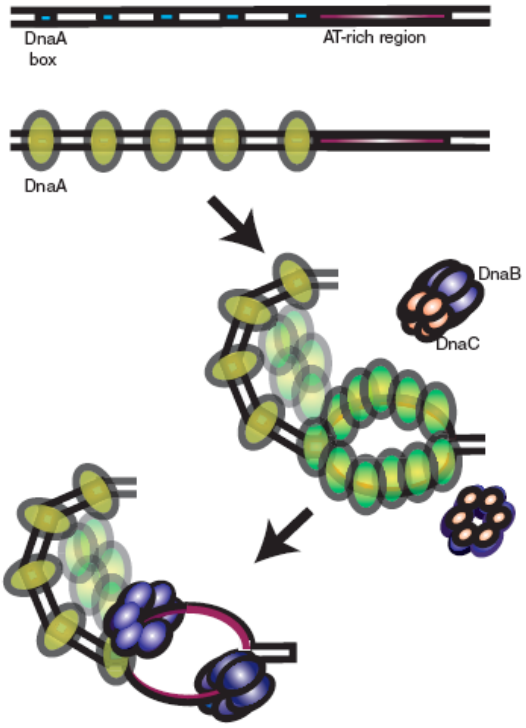


Fig. 1. Assembly of the DNA replication machinery on E.coli oriC. DnaA binds the DNA and after oligomerization melts the AT-rich region, making it accessible to DnaB/DnaC complex. After the dissociation of DnaC, DnaB starts to synthesize new DNA. (Adapted from (Robinson and Bell 2005)

To ensure complete replication of the eukaryotic genomic DNA, there are many different replication origins distributed along the genome (Tabancay and Forsburg 2006), The identification of initiation sites in eukaryotic cells is not obvious and it is experimentally challenging. One of the most studied and well characterized model is the ARS1 (Autonomously Replication Sequence) origin in Saccharomyces cerevisiae (Stinchcomb, Struhl et al. 1979). ARS1 is 125 bp long and it is constituted by 4 necessary elements: the ARS consensus element (ACS) that is highly conserved and essential, and other 3 different motifs known as B elements (B1, B2, B3) (Bell 2002; Bell and Dutta 2002). The ACS element recruits the Origin Recognition Complex (ORC) constituted by 6 closely associated proteins (Orc1-6) (Bell and Stillman 1992; Bell 2002). Before DNA synthesis starts, ORC complex recruits additional factors (Cdc6 and Cdt1) that are necessary and that have been shown to play a critical role in the loading process of the hexameric

minichromosomal maintenance (MCM) complex possessing the helicase activity (Bell and Dutta 2002; Robinson and Bell 2005) (Fig. 2).

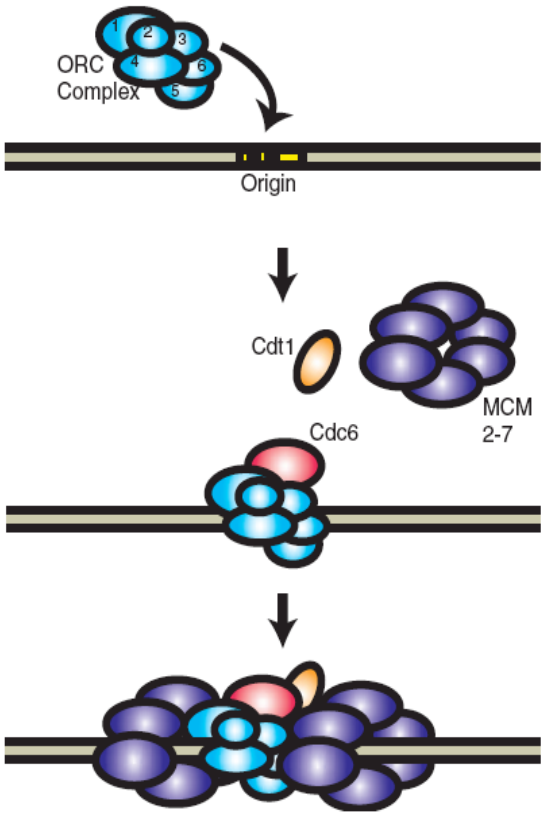


Fig. 2. Model for initiation of loading process in ARS1 origin. The origin is first bind by ORC complex, and then by Cdt1 and Cdc6, MCM complex is recruited (Adapted from (Robinson and Bell 2005)

Even if the basic features of origin activity are highly conserved also in higher eukaryotes, the mechanism of origin selection is still largely unknown. In fact, even if the proteins involved in the process are the same as in yeast, the genetic region containing replication origin is wider (1-6 Kb) compared to yeast region, and sequence does not contain a detectable consensus sequence.

In metazoans several replication origins have been isolated (Bielinsky and Gerbi 2001; Bell 2002; Cvetic and Walter 2005), but the binding of replication proteins to these origins is not sequence specific (Kong, Coleman et al. 2003; Vashee, Cvetic et al. 2003; Remus, Beall et al. 2004) and mutational analysis suggests that crucial

elements are dispersed and redundant (reviewed in (Tabancay and Forsburg 2006).

In Xenopous oocytes, exogenous DNA is efficiently replicated in a cell cycle dependent manner (Harland and Laskey 1980; Mechali and Kearsey 1984), suggesting that there must be sequence-independent mechanisms for DNA replication that can select the origins and determine their activity (Tabancay and Forsburg 2006). In addition, ORC seems to bind any DNA sequence (Cvetic and Walter 2005). This indicates that other factors like transcription, chromatin structure and nuclear localization could play a crucial role in determining specific start site for the DNA replication process (Gilbert 2004; Cvetic and Walter 2005). However, specific sequence elements could contribute to origin selection. Indeed when a DNA fragment containing a replication origin was moved to an ectopic location, it retained its function as a start site for DNA replication (Malott and Leffak 1999; Liu, Malott et al. 2003). Even if the specific sequence has some relevant function for the replication process, it is unclear what those sequences are doing.

Origin Recognition Complex (ORC)

Remarkably, despite the significant differences in origin sequence and structure, all eukaryotes use the same broad set of proteins to initiate DNA replication. The crucial complex that defines the origin is called Origin Recognition Complex, ORC, part of which binds to the origins throughout the cell cycle (Tabancay and Forsburg 2006). The assembly of the complete ORC, formed by six subunits, is the first step for the initiation of the events leading to origin firing (Li

and DePamphilis 2002). All ORC subunits bind the DNA with nanomolar affinity and bind preferentially AT-rich sequences. ORC binds ATP and has an ATPase activity. ATP binding site in Orc1 is required for ORC function (DePamphilis 2003). In human cells the Orc6 subunit has not been found to bind ORC complex (Ohta, Tatsumi et al. 2003), possibly because it may interact with other ORC subunits weakly. ORC2-5 complex has to be considered like a marker of potential replication origins, as its localization is not restricted only to active origins, but it was also found in unused origins; it may serve as a landing platform for different chromosomal proteins (Dillin and Rine 1997; Wyrick, Aparicio et al. 2001; Rusche, Kirchmaier et al. 2002; Tabancay and Forsburg 2006). ORC function is similar to the one played by DnaA in bacteria; it recruits additional proteins to the origin that form the pre-replicative complex (preRC), licensing the site for initiation of replication (Diffley 2004; Stillman 2005).

Prereplication Complex (preRC)

The prereplication complex (preRC) is assembled at the replication origins. It is formed by several multiprotein complexes that render the origins competent for initiation (Bell and Dutta 2002; Diffley 2004; Johnson and O'Donnell 2005). Together with ORC binding, also Cdc6 and Cdt1 proteins are required and critical for loading mini-chromosome maintenance (Mcm) proteins 2-7. Mcm2-7 exist as a hexameric complex that appears to have the helicase activity responsible for unwinding parental DNA strands (DePamphilis 2003). As DNA replication initiates in both directions, starting from the replication origin, at least two Mcm (2-7)

complexes are loaded at each ORC binding site (Edwards, Tutter et al. 2002; Harvey and Newport 2003).

Origins firing and Orc1 cycle

DNA must be replicated once and only once each time a cell divides. This is accomplished imposing two conditions: i) the preRC must be inactivated during S-phase, ii) new preRCs cannot be assembled until mitosis is complete. ORC1 subunit is the major responsible for this regulation: ORC1 binds ORC 2-5 complex in a cell cycle dependent manner. Supramolecular nuclear structures also play a role in the replication process. Indeed chromosomal DNA is organized into loops attached to the nuclear matrix. Each loop represents one individual replicon with the origin of replication localized within the loop and the ends of the replicon attached to the nuclear matrix. During late G1 phase, the replication origins are associated with the nuclear matrix and dissociated after initiation of replication in S phase (Anachkova, Djeliova et al. 2005). ORC1 is associated during metaphase with Cdk1/cyclinA in a hyper-phosphorylated form. As cell exits mitosis and enters in G1-phase, ORC1/Cdk1/cyclinA complex dissociates, ORC1 becomes active and it can bind chromatin and ORC2-5 complex (DePamphilis 2003). ORC2-5 complex is bound to potential origin sequences, while ORC1 selects active replication origins, where the replication machinery is assembled. The formation of ORC1-5 promotes binding of MCM complex to chromatin with the subsequent formation of preRC. Once the preRC is formed, ORC1 is degraded (by ubiquitination) and DNA synthesis starts (Ohta, Tatsumi et al. 2003) (Fig. 3). According to this model, not all the potential replication origins become active, ORC1 being the factor that

selects the initiation point (and then also the order and timing of origin firing). The cell cycle regulation of ORC1 (through the formation of the ORC1/Cdk1/cyclinA complex first and then its inactivation/degradation) ensures that origins fire only once per cell cycle (Ohta, Tatsumi et al. 2003).

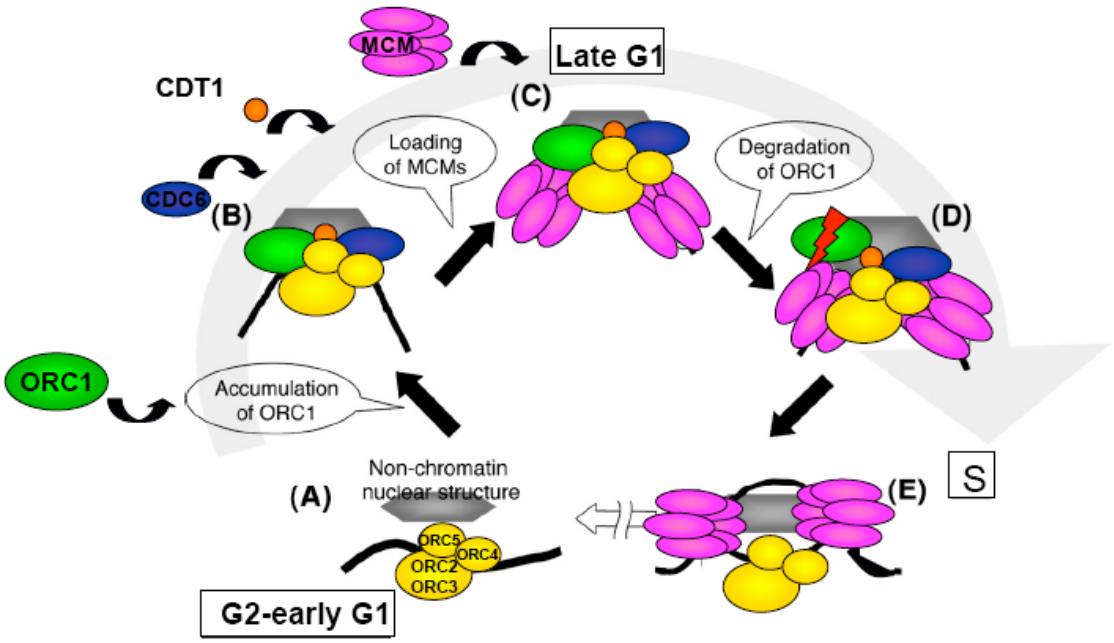


Fig. 3. Cell cycle dependent assembly of the human ORC1-5 complex. From G2 to early G1, ORC1 is present at very low level in nuclei. ORC2-5 subunits associate with DNA but not with non-chromatin nuclear structure (A). During the progression of the cell cycle ORC1 is accumulated in the nuclei and with CDC6 and CDT1 binds the ORC2-5 complex (B). Once it is formed, the complex is bound to non-chromatin nuclear structure (nuclear matrix), and MCM is loaded (C). After the assembly of this complex (pre-RC), ORC1 is degraded by proteasome (D). Sub sequentially, further initiation steps can then occur, including the assembly of the DNA synthesis apparatus (E). (adapted from Ohta, Tatsumi et al. 2003)

Assembling the active fork

Once the unwinding of DNA double strand starts, the DNA synthesis machinery, known as replisoma, is assembled and the DNA replication initiates (Bell and Dutta 2002; Garg and Burgers 2005; Johnson and O'Donnell 2005). A first DNA polymerase (alpha) associated with primase, initiates a short RNA primer and begins its extension. Then DNA polymerase alpha is replaced by a more processive polymerase enzyme, the DNA polymerase delta and also the PCNA protein, the “sliding clamp” processivity factor, is loaded. (Maga and Hubscher 2003). Leading strand synthesis is highly processive and efficient, following closely behind the helicase. Synthesis on the lagging strand, instead, requires multiple cycles of priming and extension as the underlying DNA is exposed. This synthesis occurs in the opposite direction as compared to the helicase progression (Tabancay and Forsburg 2006). The discontinuous fragments (Okazaki fragments) are assembled in a single DNA chain by DNA ligase (MacNeill 2001) (fig. 4). Synthesis of the leading and lagging strands remains coupled to ensure efficient DNA synthesis and to protect the underlying DNA template (Tabancay and Forsburg 2006).

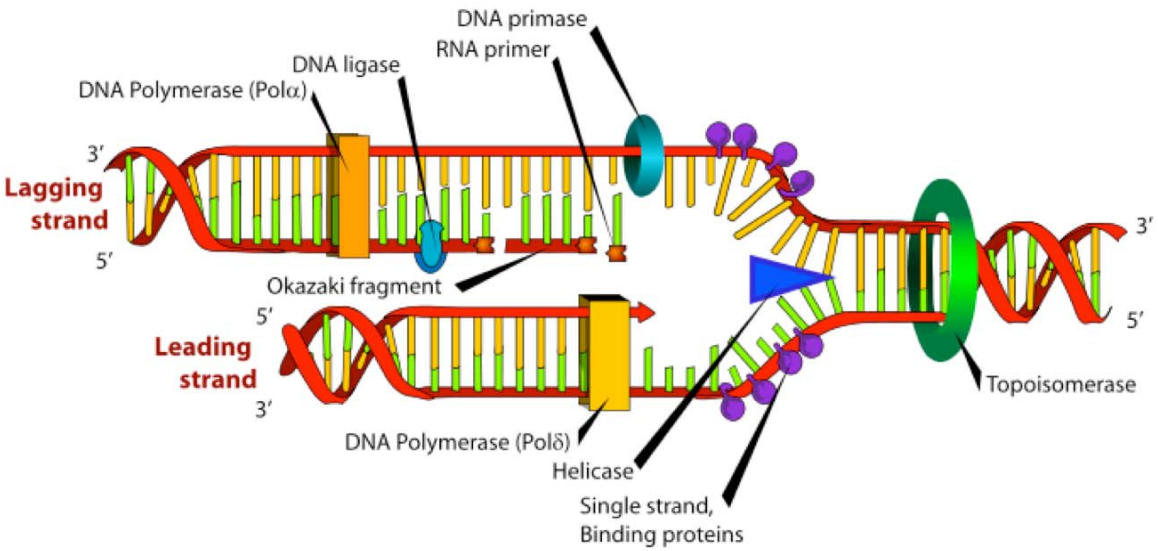


Fig. 4. Assembly active fork model. The helix unwinds and both strands replicate simultaneously as they unwind. Unwinding exposes single-stranded DNA, allowing binding of replication machinery. The leading strand replicates continuously from the 3' end, with the newest end of the forming strand facing into the replication fork. The lagging strand replicates by a series of fragments (Okazaki fragments) placed end-to-end, with the newest ends facing away from the fork; the Okazaki fragments are later joined together by DNA ligase.

Whole-genome era

With the completion of numerous genome sequences, a great deal of information became available and our understanding of several biological processes had the chance to drastically increase. Moreover scanning an entire genome sequence has been made feasible by the recent advances in microarray technologies. Advance in the synthesis of DNA microarray has increase the number of probes spotted on the array. The consequent is an increase in the

resolution and genomic sequence coverage. This new era of whole-genome study has revolutionized both mRNA expression analysis and has made easy performing functional genomics studies of DNA binding proteins. Chromatin immunoprecipitation (ChIP) followed by microarray-based detection of enriched DNA fragments, called ‘ChIP on chip’, is currently one of the most commonly used method for the identification of Transcription Factor binding sequences by a high-throughput approach (Bulyk 2006).

Tiling array technology

DNA microarray technology has revolutionized life-science research. Using arrays, researchers can examine the full complexity of a genome in a single experiment, allowing them to identify and study complex genetic regulatory networks and to begin to understand biology on a genome-wide scale. Arrays have been applied to studies in gene expression, genome mapping, SNP discrimination, transcription factor activity, toxicity, pathogen identification and detection, and many other applications. Genome tiling microarray construction is summarized in figure 5.

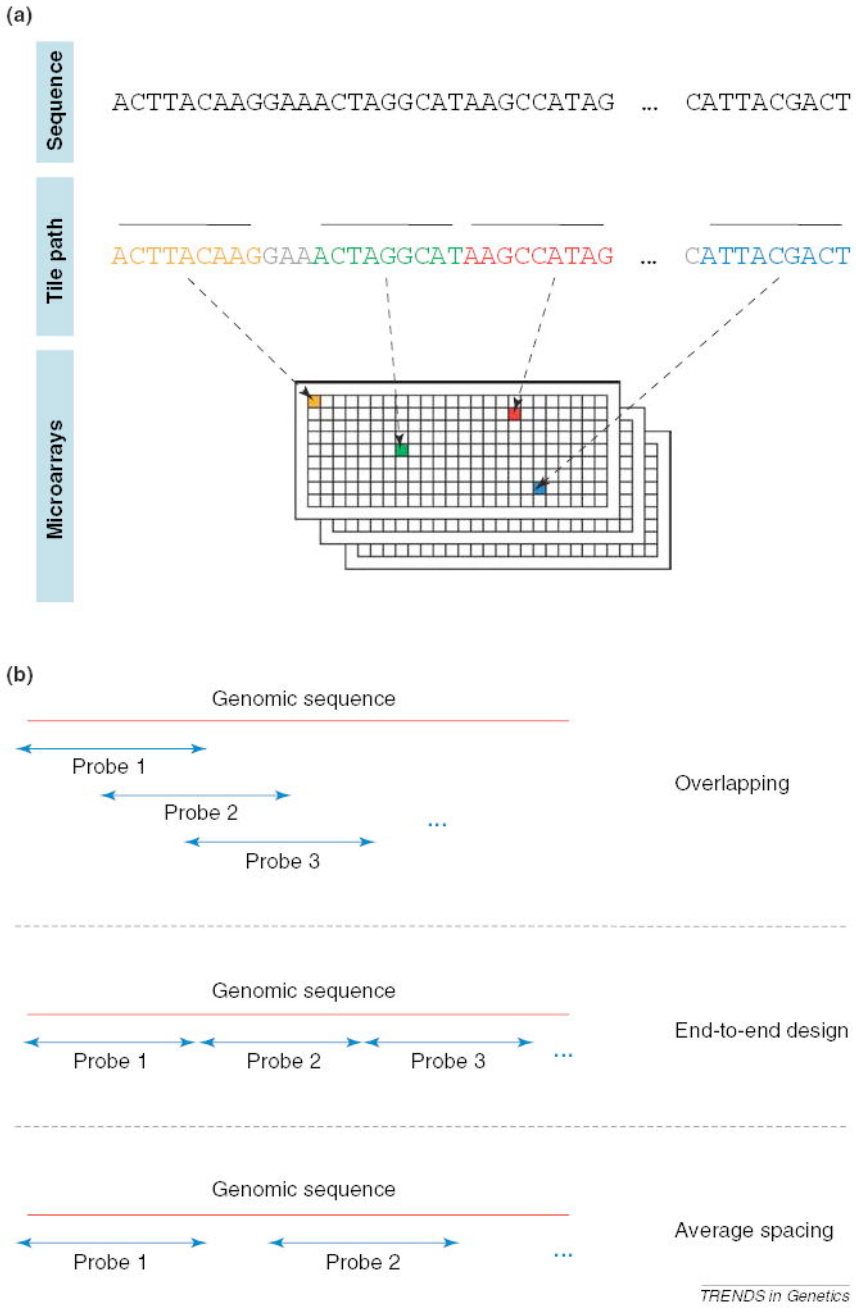


Fig. 5. Tiling microarray overview. (a) all the sequence is interrogated by different probes which are positioned randomly into the array. (b) Different tiling designs: overlapping, end-to-end or spaced (adapted from Royce, Rozowsky et al. 2005).

Each nucleic acid probe, immobilized on a glass slide, represents a little portion of a target genomic sequence. The probes can either overlap, lay end-to-end, or to

be spaced at a specific distance in the genomic context. The probe's sequence, spanning a genomic region, is called 'tiling', and the distance between the probes is called 'step' or 'resolution'. Tilling arrays may have either an isothermal design (matched melting temperatures and different probe) or an endothermal design (different melting temperatures and same probe length). The isothermal approach is the preferred design, because enables uniform probe performance, eliminating hybridization artifacts and/or bias and providing higher quality data . Probe lengths are adjusted (25mer - 60mer) to equalize the melting temperature (Tm) across the entire set. Thus, probes are optimized to perform equivalently at a given stringency in different genomic regions, including AT- and GC-rich regions. (Royce, Rozowsky et al. 2005).

Chromatin Immunoprecipitation (ChIP) on chip

ChIP on chip (or ChIP-chip) is a technique that combines ChIP and microarray technology and specifically, it allows the identification of binding sites of DNA-binding proteins on a genome-wide basis. Like conventional ChIP , this technique is based on freezing of transcription factors-DNA interactions *in vivo* through formaldehyde cross-linking. Following cross-linking, the DNA is broken into pieces 0.2-1 kb in length by sonication. At this point the immunoprecipitation with specific antibody against the protein of interest is performed resulting in the purification of protein-DNA complexes. After separation of DNA from the proteins, DNA can be isolated and quantified by PCR or used to identify genomic targets of transcription factor by DNA microarray hybridization (Ren, Robert et al. 2000). For ChIP-chip analyses, the immunoprecipitated DNA (referred as bound fraction) and the control

DNA (e.g. genomic DNA input) are fluorescently labeled using complementary dyes.

Both sets of labeled DNAs are hybridized to a microarray slide corresponding to defined genomic regions (e.g. the entire set of known promoter regions, or specific chromosomes, or the whole genome) (fig.5).

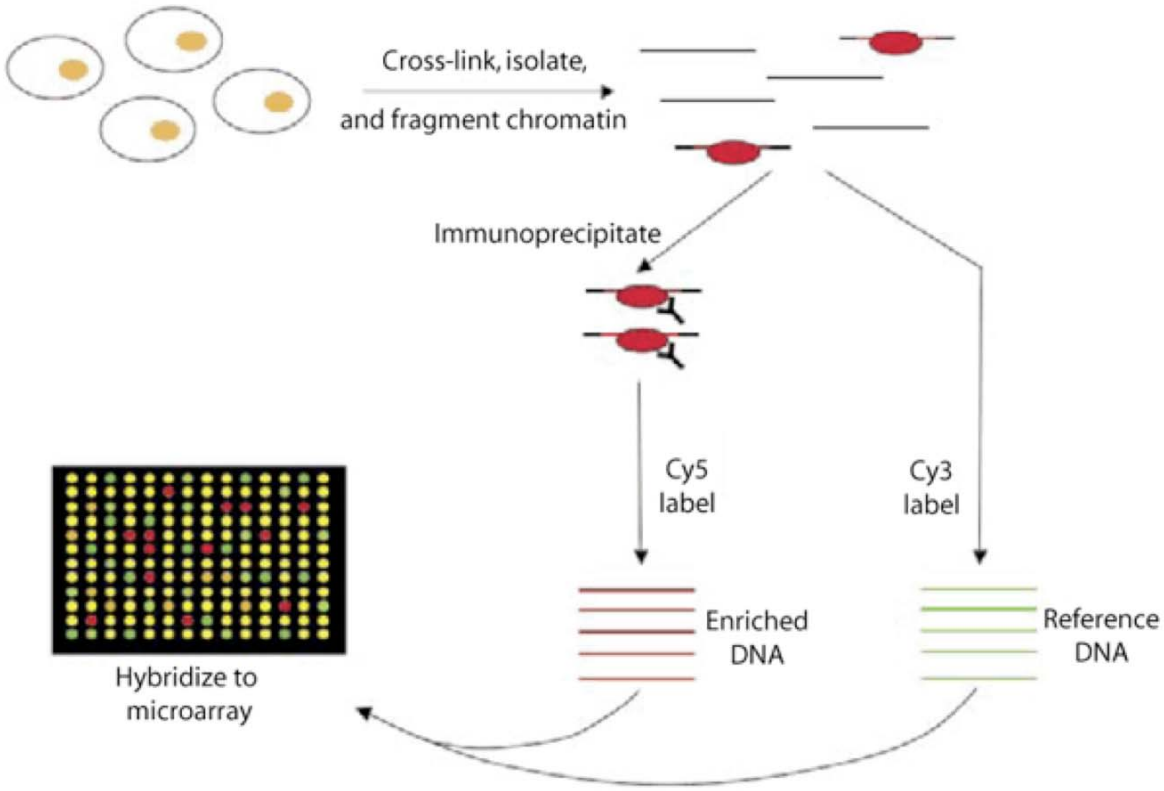


Fig.5. Schematic chip-chip procedure. After cross linking and sonication, antibody against a specific protein are used to immunoprecipitate the DNA. The DNA coming from immonoprecipitation is called enriched DNA and is spotted on the array with a reference DNA usually coming from the total DNA before immunoprecipitation. Enriched DNA and reference DNA are labeled with Cy5 and Cy3 color respectively, and the hybridization on the array is competitive (Hudson and Snyder 2006)

Bona fide protein binding sites are identified by the statistical analysis of the ratio between the ChIP DNA and the control DNA signals. The availability of high-density oligonucleotide arrays representing the entire non-repeated fraction of the human genome allowed the unbiased mapping of the interactions between DNA and the protein of interest. However, the comprehensive mapping of such interactions may not yet be practical so far in terms of costs required for a genome-wide analysis. Consistent with this, only few whole genome studies have exploited microarrays so far (Kim, Barrera et al. 2005; Carroll, Meyer et al. 2006; Kim, Abdullaev et al. 2007), and the vast majority of microarray-based screenings were conducted over 1% of the human genome under the ENCODE consortium (Birney, Stamatoyannopoulos et al. 2007).

ChIP-chip data analysis

After performing a ChIP experiment, the direct next step is to identify widely the target DNA sequences of the immunoprecipitated protein. Different types of array design are available for this analysis, and the most robust is the contiguous tiled DNA fragments that represent one entire genome, including the non–coding regions, alternatively just promoter regions or a specific chromosome can be cover from the stiling (Buck and Lieb 2004).

Data generated by a ChIP-chip experiment is similar to that of traditional microarrays, even though they differ in two main aspects. First, in microarray experiment each probe on the array measures the abundance of the corresponding RNA transcript. On the contrary, in a ChIP-chip experiment, each probe measures the abundance of a heterogeneous population of fragments

corresponding to the genomic DNA region of different length due to effects and degree of chromatin sonication. The final effect, looking at adjacent tiled probes around an enriched genomic position, is the production of a “peak” of signal centered on the binding site of the protein; a peak may span several probes, sometimes it may cover some kilobases. This so called “neighbor effect” is not an expected property of noise or of other spuriously high ratio measurements, and consequently it is a source of information and can be used for the analysis (fig.6) (Buck, Nobel et al. 2005).

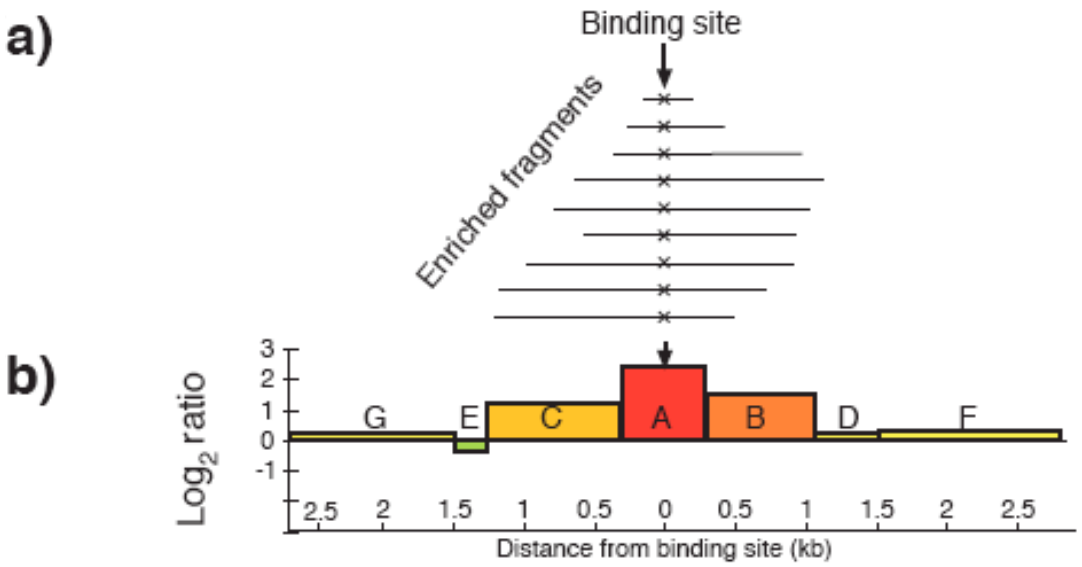


Fig.6. The neighbor effect and peaks formation. a) After ChIP experiment, purified DNA fragments bound by the protein of interest are of various lengths. b) Actual log2 ratios of different probes: probe A hybridizes with the higher number of matching fragments. That number is directly proportional with the signal intensity. Moving away from the binding site, the intensity of signal gradually decreases. (Adapted from Buck, Nobel et al. 2005)

The second difference in the interpretation of ChIP-chip and traditional gene expression data is that in expression experiments, the data are two-tailed distributed and approximately symmetric. It means that there is biological significance associated with both low and high ratio measurements, and measurements often occur with similar frequencies. On the contrary, the measurements coming from ChIP-chip experiment are a mixture of two different distributions. The most significant represents the population of the fragments corresponding to the ChIP enriched , and the second corresponds to the remaining population of genomic DNA that represent noise and/or background. The observed distribution of the log2ratios signals is therefore asymmetric around zero, with a distinct, positive oriented skew. The left-hand side of the distribution (the negative log ratios) is approximately Gaussian, but the positive log ratios show a heavier non-gaussian tail (fig.7). For the vast majority of ChIP-chip experiments, the genomic regions of biological interest will be confined to the positive side of the distribution, and the negative log ratios will arise solely from fragments that are considered to be background. Under the additional assumption that the distribution of unenriched fragments is symmetric about zero, we can estimate the distribution of background ratios using only the observed negative log ratios as a guide.

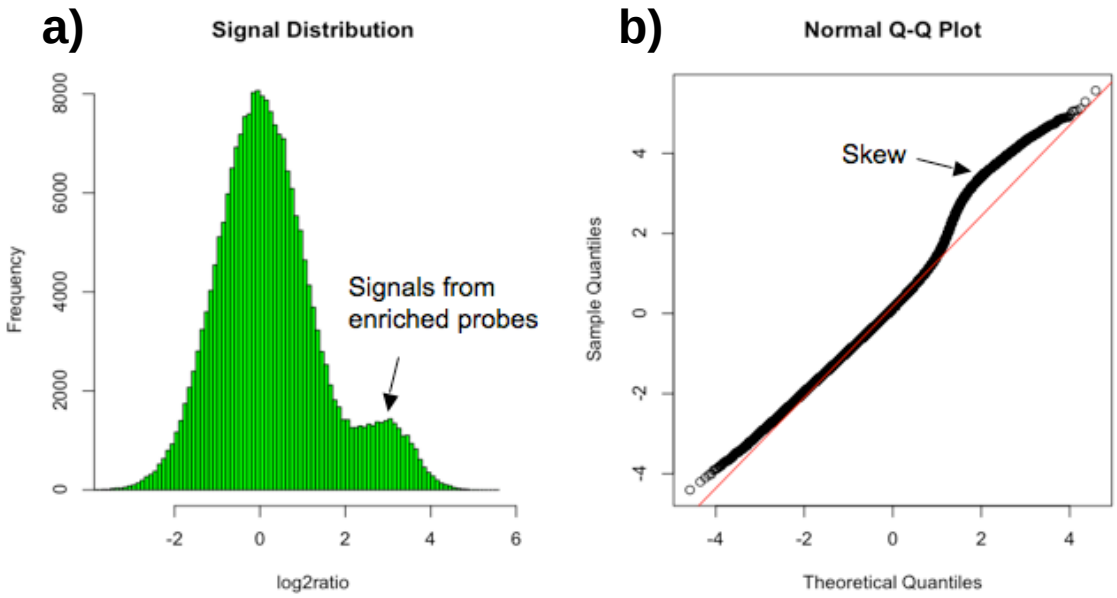


Fig.7. Raw signal distribution of the data coming from a ChIP-chip standard experiment. a) the distribution is skewed on the positive direction. Is possible to see two different distribution, one for the background signal and one for the enriched signal. b) A quantile-quantile plot (Q-Q plot) for the same experiment (black points) against a normal distribution (red line). On the upper part of the graph is visible the skew effect cause by the enriched probes.

Enriched regions (Peaks) identification

Starting from the two assumptions explicated above many different algorithms for the analysis of ChIP-chip data have been developed (Glynn, Megee et al. 2004; Scacheri, Crawford et al. 2006; Zhang, Rozowsky et al. 2007). Different approaches have been tested, but the main concept common to all the algorithms is that a single probe signal does not contain enough information. To solve this issue the smoothing of the signal is one of the most relevant procedures used until now to transform the data. Smoothing across genomically neighboring probes is often employed to ameliorate probe-specific variability in the data, that is, the effect that different probes measure the same target DNA amount with different

efficiency. This may be caused by different qualities of probe synthesis on the array, probe GC content, target cDNA secondary structure, cross-hybridization, and other reasons. Recently spike-in experiment has been performed and published and differences between platforms, algorithms and laboratories have been tested (Johnson, Li et al. 2008). Results suggests that the algorithm used is important, but also the chip design and the expertness of laboratory. The impact of algorithms choice diminishes when probes are tiled at a very high resolution (Johnson, Li et al. 2008).

Aim

The replication process not only is involved in the DNA duplication itself, but it also affects the expression and regulation of genes. It has been shown that early replication correlates most closely with an “open” chromatin structure that is permissive for transcription, rather than with transcription itself, and it has been proposed that the temporal program might be important for propagation of gene expression patterns into daughter cells. Cancer transformation may alter the regulation of origin activation resulting in altered origin usage between normal and transformed cells (Di Paola, Price et al. 2006). To elucidate how replication origins usage and timing are regulated in metazoans, it is necessary to identify and characterize DNA sequences that serve as replication origins. Despite the early successes in the identification of microbial eukaryotic origins (Newlon and Theis 1993; Gilbert 1998), the available methods for origin identification in mammalian genome and the main results so far generated by their application are sometimes controversial and have lead to the detailed characterization of really few origins (Ohta, Tatsumi et al. 2003).

Our laboratory has recently set up a novel strategy to isolate a population of DNA fragments enriched with human “replication origins ” from equilibrium density gradient and to ChIP ORC1 in the low density fraction of the gradient. The DNA purified from specific fractions of the gradient (low density fractions ChIPped with replication protein ORC1 and DNA naked high density fractions) has been then hybridized on high-density oligonucleotide custom arrays, created by NimbleGen

Systems and containing ~380,000 50mer tiled probes covering the whole human chromosome 19.

The aims of this research project are a) to use experimental data and computational analysis for the high-throughput mapping of human replication origins b) to characterize replication origins features and c) to find possible correlations between the position of newly identified origins and genome function and structure.

Materials and methods

Cell lines and culture conditions

U937-PR9 cells were maintained in RPMI 1640 medium supplemented with 10% fetal calf serum plus Penicillin/Streptomycin and Glutamine, in humidified atmosphere at 5% CO2.

CsCl equilibrium density gradient

Adapted from Schwartz et al.:300x106 U937-PR9 cells were cross-linked with 1% (final conc.) formaldehyde for 4 minutes at RT. The fixation was stopped by adding glycine 1.25 mM (final conc.). Cells were pelletted and washed twice with ice cold PBS, and harvested in SDS Buffer (50 mM Tris at pH 8.1, 0.5% SDS, 100 mM NaCl, 5 mMEDTA, and protease inhibitors). Lysate sonication is performed in sonication buffer (10mM HEPES, 1mM EDTA, 0,5mM EGTA, and protease inhibitors). Then, CsCl and N-laurylsarcosine were added to the lysate in centrifuge tubes and spun 140 hours at 34000 rpm. Fractions are collected and dialyzed with dialysis buffer (4% glycerol, 10mM Tris pH8, 1mM EDTA and 0.5mMEGTA. The cross-linking is reversed by incubation o/n at 65°C in of de-cross buffer (NaHCO3 0,1M, 1% SDS for DNA extraction supplemented with PK). The DNA was then phenol/chloroform extracted and ethanol-precipitated. Laemmli buffer was added to protein samples.

Immunoblotting

Proteins are loaded on standard SDS PAGE polyacrillamide gel, at different concentrations, and incubated with the primary antibodies reported in fig.1, and the correspondent secondary antibodies.

Chromatin immunoprecipitation (ChIP)

U937 PR9 cells were cultured as above. Formaldehyde was added to the culture medium to a final concentration of1% 8 min RT, and stopped by addition of glycine at a final concentration of 0.125 M, followed by an additional incubation for 5 min. Fixed cells were washed twice with TBS (20 mM Tris at pH 7.4, 150 mM NaCl) and harvested in SDS Buffer. Cells were pelleted by centrifugation, and suspended in 3 mL of IP Buffer (100 mMTris at pH8.6, 0.3% SDS, 1.7% Triton X-100, and 5 mM EDTA). Samples were disrupted by sonication with Branson 250 sonicator. The lysate was then diluted with IP buffer to the final volume defined on the number of IPs. For each immunoprecipitation, 1 mL of diluted lysate was precleared by addition of 30 µL of blocked protein A beads (50% slurry protein A-Sepharose, 0.5 mg/mLfatty acid-free BSA, and 0.2 mg/mL salmon sperm DNA in TE). Samples were immunoprecipitated overnight at 4 °C with specific antibodies on a rotating wheel. Immune complexes were recovered by adding 30 µL of blocked protein A beads and incubated for 2 h at 4 °C. Beads were washed 2 times in 1 mL of Mixed Micelle Buffer (20 mM Tris at pH 8.1, 150mMNaCl, 5mMEDTA, 5% w/v sucrose,1% Triton X-100, and 0.2% SDS), 2 times in Buffer 500 (50 mM HEPES at pH 7.5, 0.1% w/v deoxycholic acid, 1% Triton X-100, 500 mM NaCl, and 1 mM EDTA), 2 times in LiCl Detergent Wash Buffer (10 mM Tris at

pH 8.0, 0.5% deoxycholic acid, 0.5% NP-40, 250 mM LiCl, and 1 mM EDTA), and 1 time in TE (pH 7.5). Chromatin was de-crosslinked o/n at 65 °C in 250 ml of de-cross buffer. The beads were eluted and the DNA was phenol/chloroform extracted and ethanol-precipitated. Immunoprecipitated DNA from 10x106 cell equivalents was resuspended in 0.3 of TE1x.

NimbleGen chip hybridization

Purified DNA was hybridized on chromosome 19 chip, in accordance to what suggested by the NimbleGen.

Nascent strand abundance assay

Adapted from Giacca et al. Total phenol-chloroform extracted genomic DNA is resuspended in TNE buffer and loaded on sucrose gradient as reported in the paper. Fractions are then collected, and ethanol precipitated. The DNA purified is the assayed by qRT-PCR.

RNA extraction

RNA extraction was performed using Quiagen RNAeasy minikit following the protocol provided. A DNAse digestion step was added to assure the complete elimination of contaminant DNA.

cDNA synthesis

1 µg of RNA was retrotranscribed using random hexamers and oligo dT. RNA was denatured at 70° C for 15min and then placed immediately on ice. The first strand mix (0,5mM dNTP, 0,1mM DTT, Clontech first strand buffer, RNAse inhibitor and reverse transcriptase) was added. The mix was incubated at room temperature for 10 minutes and then at 42°C for 50 min. The enzyme was inactivated for 15 min at 70°C. Final reaction was diluted and used for real time PCR or hybridized on the NimbleGen chip.

MEME: Multiple Em for Motif Elicitation

MEME is a tool for discovering motifs in a group of related DNA or protein sequences (Bailey and Elkan 1994). A motif is a sequence pattern, possibly functional, that occurs repeatedly in a group of related protein or DNA sequences. MEME is an *ab initio* discovery program. MEME represents motifs as position-dependent letter-probability matrices which describe the probability of each possible letter at each position in the pattern. Individual MEME motifs do not contain gaps. Patterns with variable-length gaps are split by MEME into two or more separate motifs. MEME takes as input a group of DNA or protein sequences and outputs as many motifs as requested. MEME uses statistical modeling techniques to automatically choose the best width, number of occurrences, and description for each motif.

Clover: Cis-eLement OVERrepresentation

Clover is a program for searching functional sites, identified by known a positional weighted matrix (PWM) model, in a set of DNA sequences. If you give it a group of DNA sequences, that share a common function, it will compare them to a library of sequence motifs (e.g. transcription factor binding matrices), and identify which if any of the motifs are statistically overrepresented in the sequence set (Frith, Fu et al. 2004). Clover will compare each motif in turn to the sequence set, and calculate a "raw score" indicating how strongly the motif is present in the sequence set. Raw scores by themselves are hard to interpret, so Clover provides options to determine the statistical significance of the raw scores. Four ways of determining statistical significance are available. The first requires the use of one or more files of DNA sequences background. Each background file should contain sequences in FASTA format, with total length much greater than the target sequence set. For each background set, Clover will repeatedly extract random fragments matched by length to the target sequences, and calculate raw scores for these fragments. The proportion of times that the raw score of a fragment set exceeds or equals the raw score of the target set, e.g. 0.02, is called a P-value. The P-value indicates the probability that the motif's presence in the target set can be explained just by chance. For each motif, a separate P-value is calculated for each background file. The second way of determining statistical significance is to repeatedly shuffle the letters within each target sequence, and use these shuffled sequence sets as controls. P-values are calculated as above. The third way is to create random sequences with the same dinucleotide compositions as each target sequence. The fourth way is to shuffle the motif matrices, and obtain control raw

scores by comparing the shuffled motifs to the target sequences. When shuffling a motif, the counts of A, C, G and T within each position are not shuffled, but the positions are shuffled among one another.

Bioinformatic Analysis

With the introduction of tiling array technology, it has become feasible for scientists to interrogate entire genomes. Such high-resolution DNA microarrays represent a universal framework exploitable through several diverse experimental approaches to extract the full significance of whole-genome data (Mockler, Chan et al. 2005). Indeed, tiling arrays have been widely employed to detect genome-wide protein–DNA interactions through chip hybridization of chromatin immunoprecipitation experiments (ChIP-chip) (Wei, Wu et al. 2006; Birney, Stamatoyannopoulos et al. 2007; Guenther, Levine et al. 2007) and, more recently, to perform expression studies, confirming the potential of these techniques for the characterization of the whole transcriptome (Kapranov, Willingham et al. 2007). The integration of ChIP-chip and expression profiling data will be essential to decode the genetic and epigenetic networks that interlink DNA binding protein regulators, transcriptional events and chromatin state, both in physiological and pathological conditions. To support the analysis of such experimental approaches we have developed CARPET (Collection of Automated Routine Programs for Easy Tiling), a compilation of scripts integrated within the Galaxy2 platform (Blankenberg, Taylor et al. 2007), that helps biologists to analyze ChIP-chip and expression tiling data independently and, if necessary, to merge the results for a more comprehensive understanding of the significance of the experimental data.

CARPET

Introduction

CARPET is a set of Perl, Python and R scripts, integrated on the Galaxy2 web based platform (Blankenberg, Taylor et al. 2007), for the analysis of ChIP-chip and expression tiling data. CARPET allows rapid experimental data entry, simple quality control, easy identification and annotation of enriched ChIP-chip regions, detection of the absolute or relative transcriptional status of genes assessed by expression tiling experiments and, more importantly, it allows the integration of ChIP-chip and expression data. Results can be visualized instantly in a genomic context within the UCSC genome browser as graph-based custom tracks through Galaxy2. All generated and uploaded data can be stored within sessions and are easily shared with other users (fig. 8).

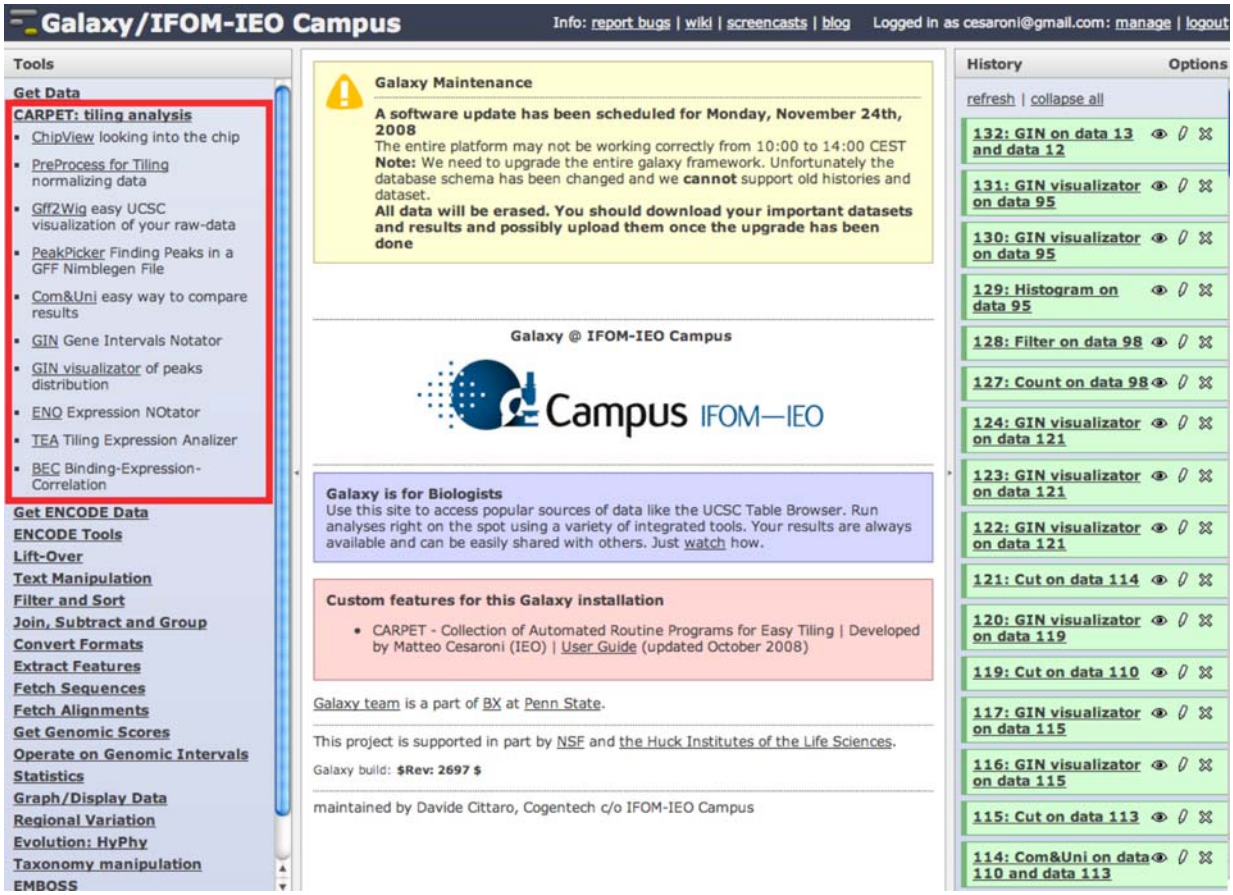


Fig. 8. Galaxy and CARPET website on the IFOM-IEO-Campus server. All the scripts available in the CARPET suite are highlighted with the red square.

Quality assessment by chip image visualization: ChipView

ChipView, allows you to create and visualize an image of the hybridized chip surface (fig. 10a), a feature not normally offered by NimbleGen. Chipview works with “Pair files” data provided by NimbleGen (see Appendix A for details) or “custom raw signal coordinates files”. With Chipview, the distribution of the signal over the chip can be easily inspected for the presence of artefacts or hybridization problems. Since Chipview considers only the chip raw signals and the corresponding x and y chip coordinates when creating the chip image, a standard

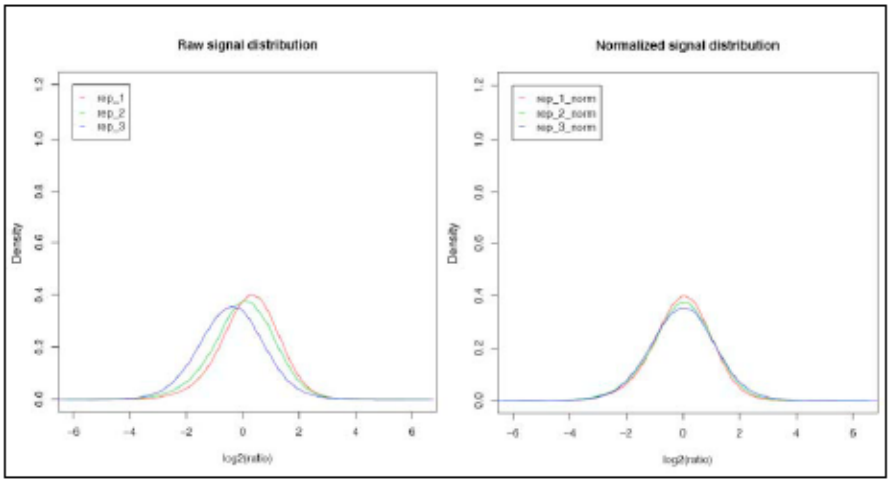
NimbleGen “Pair File” is not strictly required. ChipView can use all types of raw data tables, as long as the corresponding chip coordinates are also supplied. You will need to specify to the system which columns contain matching data.

Data normalization - PreProcess for Tiling: PPT

An important first step in chip data analysis is the normalization phase. PPT normalizes both ChIP-chip and expression tiling data (single or multi-replica experiments) using a program based, in part, on the Ringo Package (Toedling, Skylar et al. 2007). PPT also calculates and compares the correlation between replicates and finally creates a GFF file suitable for peak identification by PeakPicker or other user preferred methodologies/programs. First of all for each chip, the log2 of Cy5/Cy3 ratio is calculated (if not already provided). All the chips are then normalized, according to the type of normalization selected:

a) bi-weight: this procedure centers the probe log2(ratio)s around zero; scaling is performed by subtracting the bi-weight mean for the log2(ratio) values from each log2(ratio) value.
b) quantile: this procedure normalizes the distributions of the probe log2(ratio) of each chip with a quantile normalization.

The correlations between chip replicates can then be calculated. The program produces two outputs: a table file of the normalized data and a pdf file of graphs showing data distribution before and after the normalization process and the correlation between replicates (fig. 9).



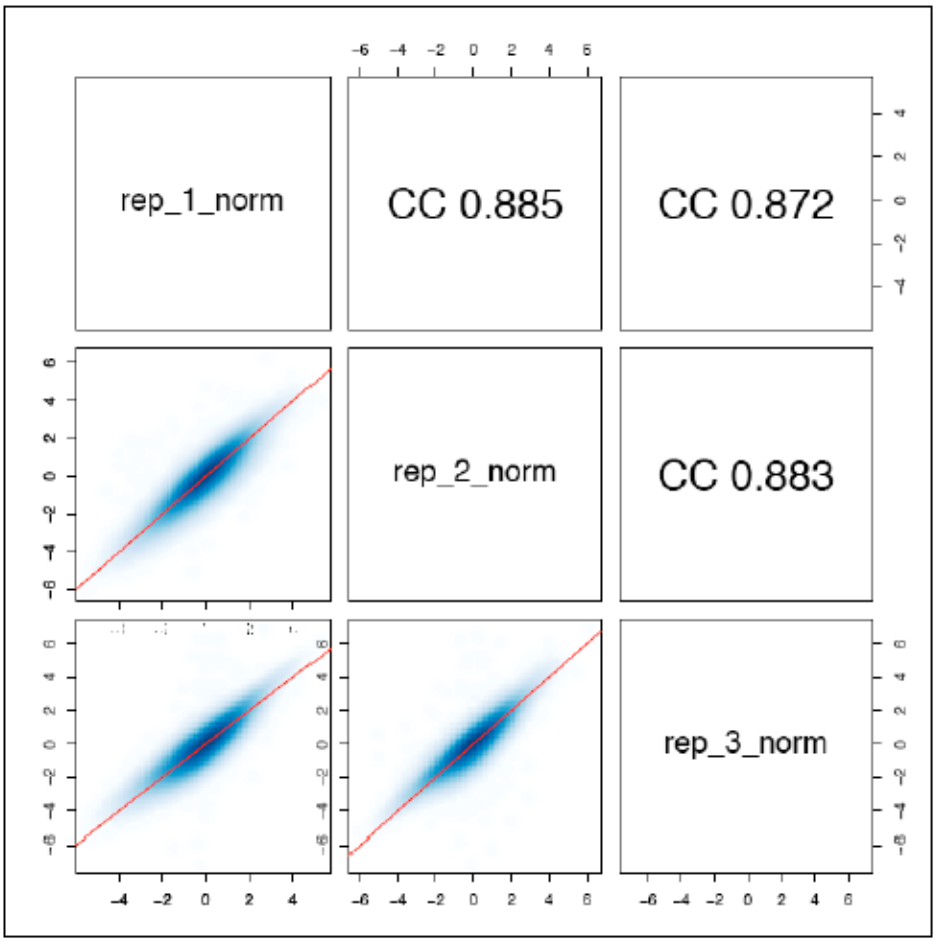


Fig. 9. PPT output. In the first part of the figure the distributions of the three chips (log2(ratio) signal) before and after normalization are plotted. In the second part of the graph each replica is plotted against each other and the correlation value are plotted.

Peak identification: PeakPicker

PeakPicker is a Perl script that is able to identify enriched regions (peaks) from a ChIP-chip experiment. The PeakPicker tool utilizes NimbleGen log2(ratio) files in GFF format (or properly reformatted GFF files obtained from other platforms) as the INPUT FILE and identifies regions of enriched signals (peaks), providing as an output a table in GFF format that contains the genomic peak coordinates and scores, alone or with statistical values.

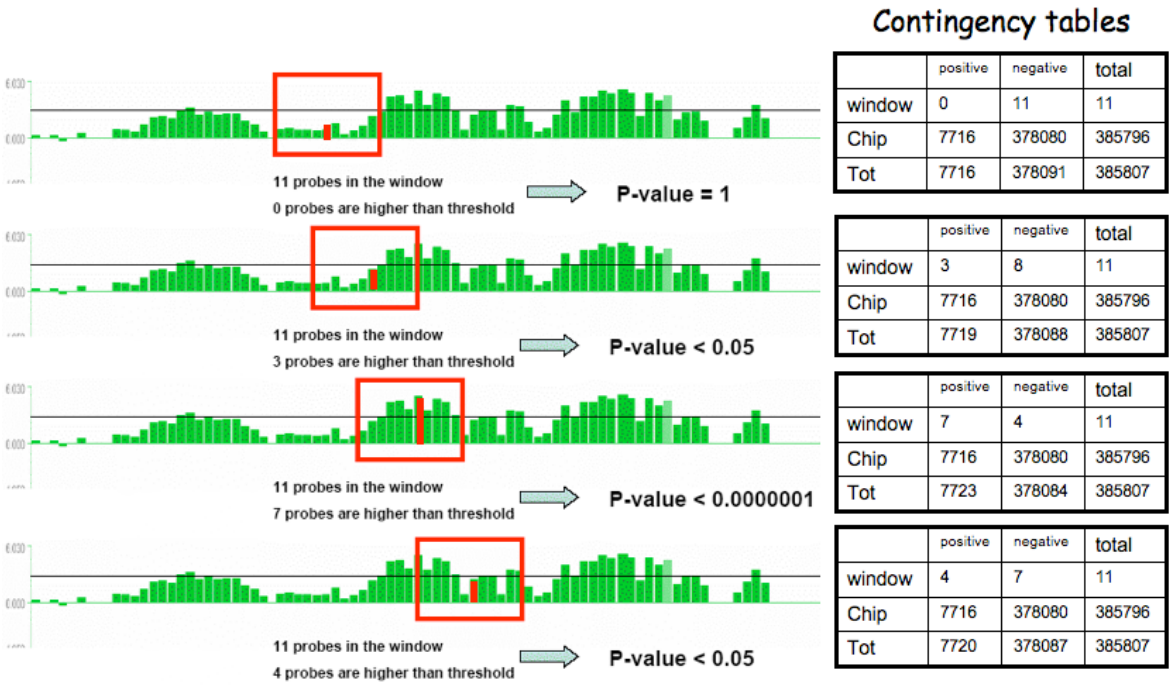
How does PeakPicker work?

A “peak” is defined as a region in the genome where multiple probes, with a log2(ratio) greater than a user-defined threshold, are located close to each another. PeakPicker makes two assumptions: i) data are enriched for signals in the positive direction ("one-tailed"); ii) a peak is represented by multiple probes located close to each other in the genome.

PeakPicker makes use of the sliding window statistical approach, essentially as described by Scacheri et al. (Scacheri, Crawford et al. 2006). A window moves along the array, centering on each probe in turn; in each window Chi squared is calculated (see formula)

$$\chi^2 = \sum \frac{(f(a) - f(e))^2}{f(e)}$$

by building a contingency table for each window/probe position (fig. 10). A p-value is then assigned.



	positive	negative	total
window	0	11	11
Chip	7716	378080	385796
Tot	7716	378091	385807

	positive	negative	total
window	3	8	11
Chip	7716	378080	385796
Tot	7719	378088	385807

	positive	negative	total
window	7	4	11
Chip	7716	378080	385796
Tot	7723	378084	385807

	positive	negative	total
window	4	7	11
Chip	7716	378080	385796
Tot	7720	378087	385807

Fig. 10. PeakPicker procedure. For each probe a contingency table is created and a Chi squared test performed. The calculated p-value takes in account the neighbor effect of the probes

Therefore, PeakPicker produces a new profile of your experiment (Fig. 11) derived from the "-log2(p-values)" associated to each probe. Profile peak margins are finally delineated by taking into account a user-defined p-value threshold. "-log2(p-values)" are used, instead of plain p-values, for this procedure since they take into account the so called “neighboring probes effect”, thereby, dramatically reducing the impact of the background signal (see Fig. 11).

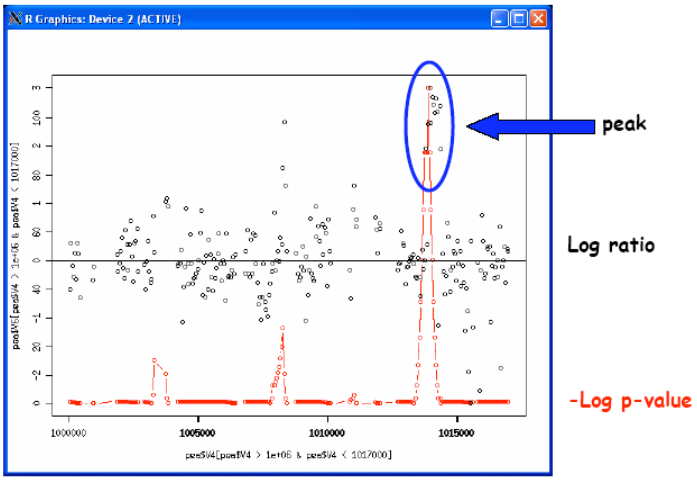


Fig. 11. Raw signal transformation. The red line represents the new signal coming from the p-value calculation. The background decreases drastically due the fact that each point take into account the information coming from the neighbor probes.

Alternatively or alongside the statistical p-value calculation for each peak, PeakPicker also estimates a peak score value that takes into account the length and the intensity of the raw log2(ratio) signals under the peak using this formula

$$S = \frac{\sum_1^n \log(ratio)}{n} + \sqrt{n}$$

The score value is derived from two main variables: peak height, the mean log2(ratio) value of the probes below the peak; peak length, an empirical correction factor derived from the square root of the total number of probes belonging to a peak.

PeakPicker allows the user to define a number of different parameters: the minimal number of probes that must exceed the defined threshold (fix this parameter in accordance with the expected peak length), the maximum distance permitted between probes, for probes to be considered as contiguous (fix this parameter according to your chip tiling design). The stringency of your analysis can be varied by setting different p-value thresholds. Neighbor enriched regions can also be joined together (fix this parameter according to the expected peak

spread). The analysis output is in the format of a GFF file that can be visualized simultaneously with your raw log2(ratio) data on the UCSC Genome Browser.

Peak comparison: Common & Unique (Com&Uni)

Several binding factors or histone modifications are often ChIPed within independent, but analogous, experiments, making cross-comparison of experiments essential for interpreting results correctly. The Com&Uni tool allows the user to compare two PeakPicker GFF output files, corresponding to two independent ChIP-chip experiments, in order to identify common and unique features. The program also permits the user to analyze peak flanking regions.

Peaks annotation: Genomic Interval Notator - (GIN)

Once you have identified and mapped enriched regions of binding for your ChIP-chip experiment, you can now determine the relationships between your data and gene loci. The GIN (Genomic Interval Notator) tool helps you with this task by annotating peak queries using user-defined annotation tables (e.g. RefSeq, UCSC genes, Ensembl Genes) and calculating the relative positions of peaks with respect to transcript associated features (e.g. promoter, exon, intron, intergenic).

How does GIN work

It uses two files: a GFF file with genomic intervals (i.e. the output file of PeakPicker) and any user-preferred transcript annotation table (e.g. RefSeq, UCSC genes) that can be easily downloaded from the UCSC Genome Browser database (see Appendix A of this manual for more information). GIN associates

genomic interval queries (i.e. your peaks) with the matching interrogated transcripts. The output for each interval includes the name and absolute chromosome coordinates of the assigned transcriptional units, as well as a call describing its relative position with respect to the transcribed unit (e.g. first exon, fourth intron, promoter) and the relative distance from the putative TSS (Transcription Start Site). Intervals that do not intersect any gene loci are annotated as “intergenic”. The user can arbitrarily define the length (in bps) of the putative promoter regions upstream of each TSS by setting the “Promoter definition” option. GIN has been programmed to produce a not-redundant annotation table, meaning that each peak will have a unique annotation. Since genes or gene features are often overlapping in genome sequences (e.g. antisense transcripts, bidirectional putative promoters) leading to an ambiguous annotation, the user can give priority to the annotation of genes or of (putative) promoter regions using the “Annotation priority” option. If the “promoter” option is chosen, GIN first tries to locate a peak within a promoter region. If more than one promoter is found, the peak is associated to the promoter of the closest transcriptional unit. If the “gene” option is selected, GIN first tries to locate a peak within an exon. A flowchart summarizing how GIN annotates peaks, depending on the priority specified, is shown in figure. 12.

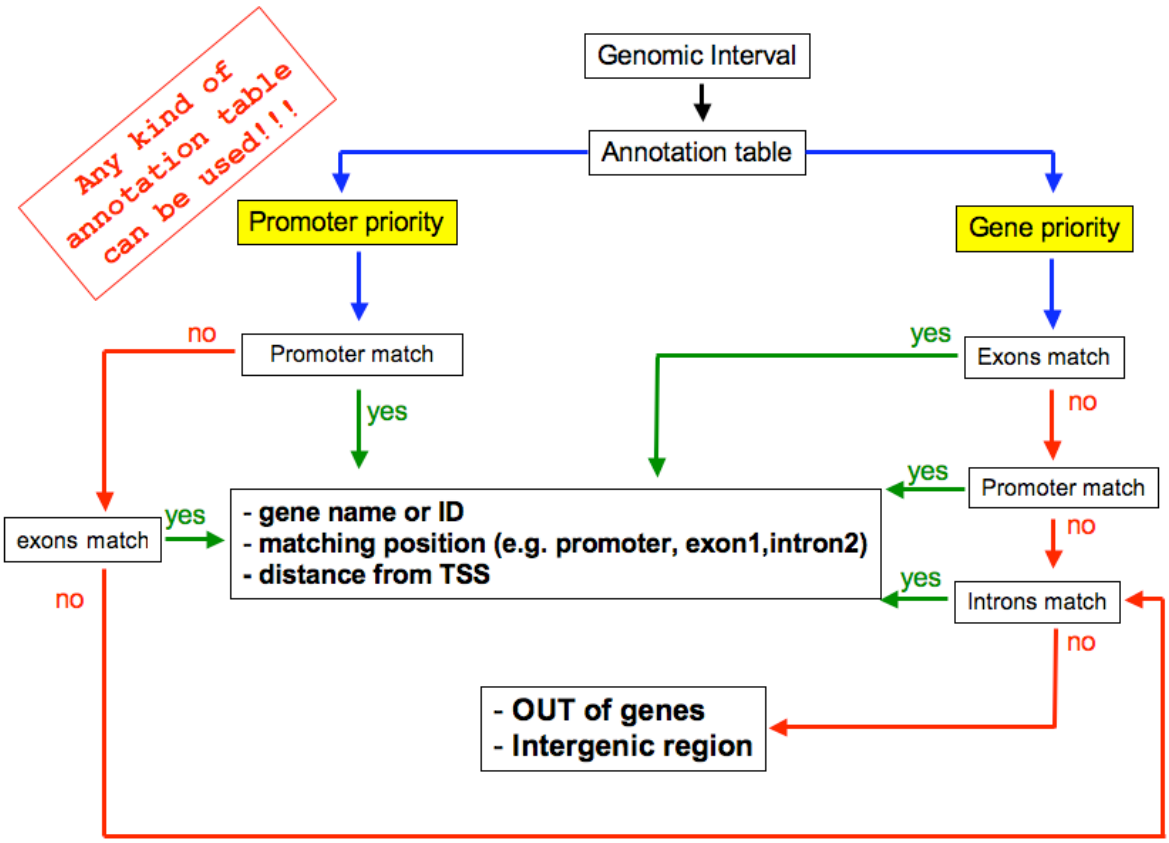


Fig. 12. GIN flowchart. The procedure for the annotation of peaks is shown here. Any kind of annotation table can be used.

Peak characterization: GIN visualizator

Now that you have an annotated list of the positions of your ChIP-chiped binding peaks relative to gene transcripts, you can visualize this information using the GIN visualizator tool. For this purpose, the GIN output file can be directly submitted to GIN visualizator to portray the distribution of peak intervals around the TSS, as shown in figure 13. Knowledge of the location of binding elements with respect to the TSS of the associated transcriptional unit is often helpful when characterizing regulatory DNA binding proteins tested by ChIP.

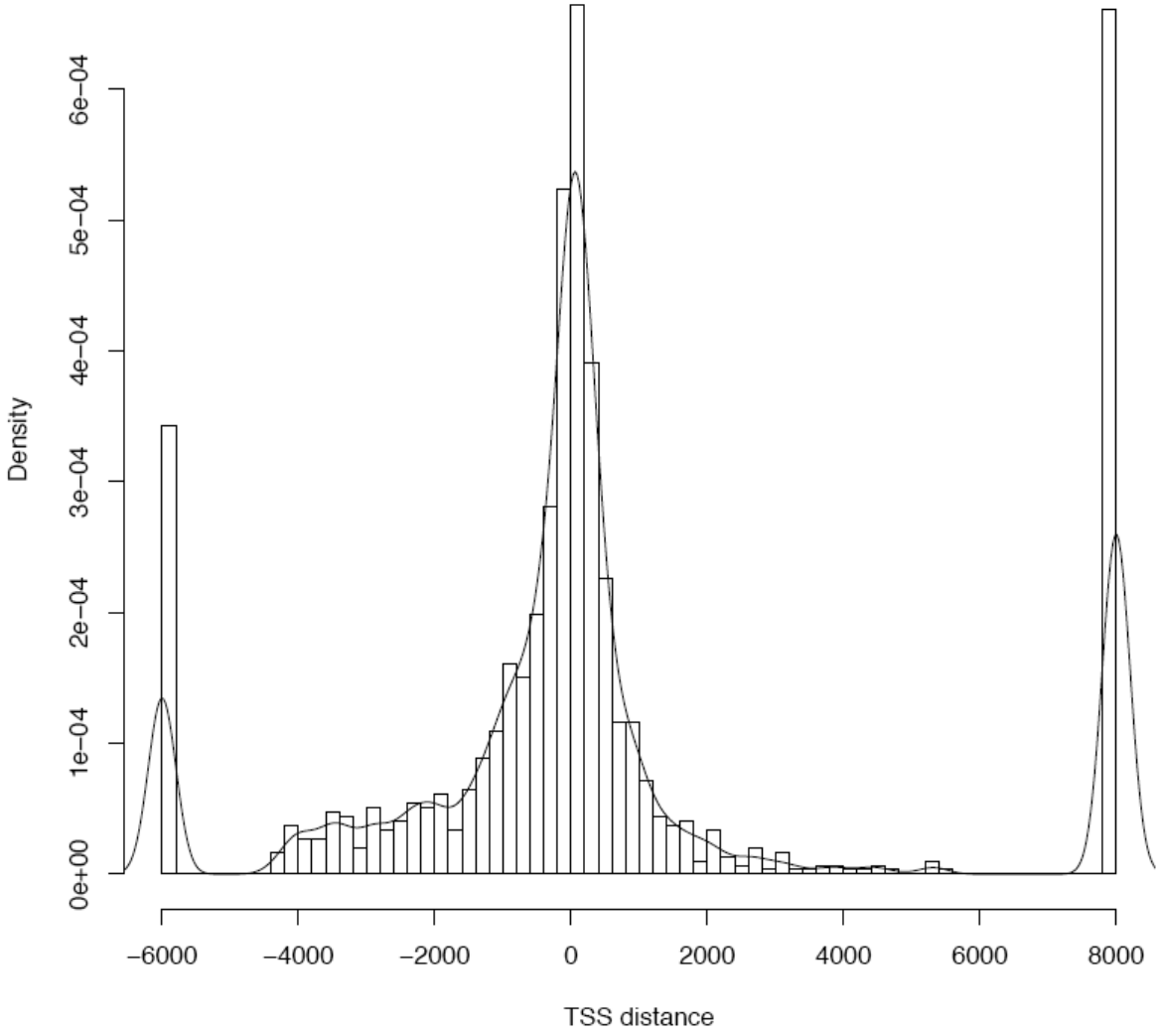


Fig. 13. Output of GIN visualizator.

Expression chip annotation: Expression Notator (ENO)

The first step in analyzing an expression tiling experiment (namely a cDNA hybridation on a tiling platform), is to assign chip probes to their corresponding gene, in particular, to the corresponding exon. The Expression Notator (ENO) tool annotates each probe on the chip using a user-defined transcript annotation table (e.g. RefSeq, UCSC genes) downloaded from the UCSC Genome Browser database. Since this annotation step relies on chip design, it is necessary to

perform this step only once for each transcript annotation table that you want to use. If a probe matches with more than one transcript (e.g. two overlapping antisense transcripts or two different splicing isoforms) the program will take into account the same probe for both transcripts. In figure 14, a scheme of different annotation features is reported.

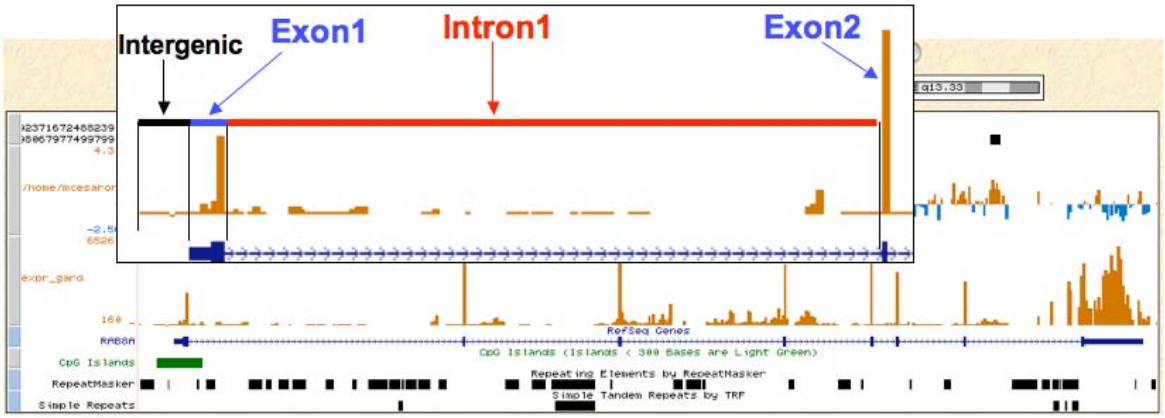


Fig. 14. Annotation procedure

Analysis of Tiling expression data: Tiling Expression Analyzer (TEA)

The TEA tool (Tiling Expression Analyzer) performs two different tasks, depending on the number of experiments uploaded. For simple expression estimation, TEA (starting from your ENO annotation file) calculates an expression value based on the mean and/or the median of the probe signals associated with the exons of a particular transcriptional unit. In comparison experiments, TEA analyzes the signal distribution for each gene under different conditions (e.g. untreated vs. treated) and calculates the fold-change and statistical p-values. The user may also choose to operate a False Discovery Rate (FDR) correction

(Benjamini and Hochberg, 1995). Many filters can also be applied to the data, e.g. on the raw signals, fold-change and p-values. TEA utilizes NimbleGen expression files in GFF format and the annotated table produced by ENO as INPUT FILES to generate a table of the expression value for the transcripts studied. When comparing two different conditions, TEA computes, from the two distinct expression level evaluations, the fold-change in expression between the two conditions and a statistical p-value. In this analysis, the expression tiling GFF file should contain raw signals (NOT the log2(ratio)), as usually provided by NimbleGen; alternatively the transformed log2 of the raw signal, obtained after normalizing with the “PreProcess for Tilling” tool, can be use as the input.

How does TEA work?

For each gene, the program builds the signal distribution of the probes that match the exons. In a simple expression experiment, the mean or the median signal distribution for each gene is reported in the TEA output. In a comparison experiment the signal distribution over the gene exons is compared between the two conditions (1 and 2) and a t-test is performed (fig. 15); in addition the user can introduce a FDR correction (Benjamini and Hochberg, 1995).

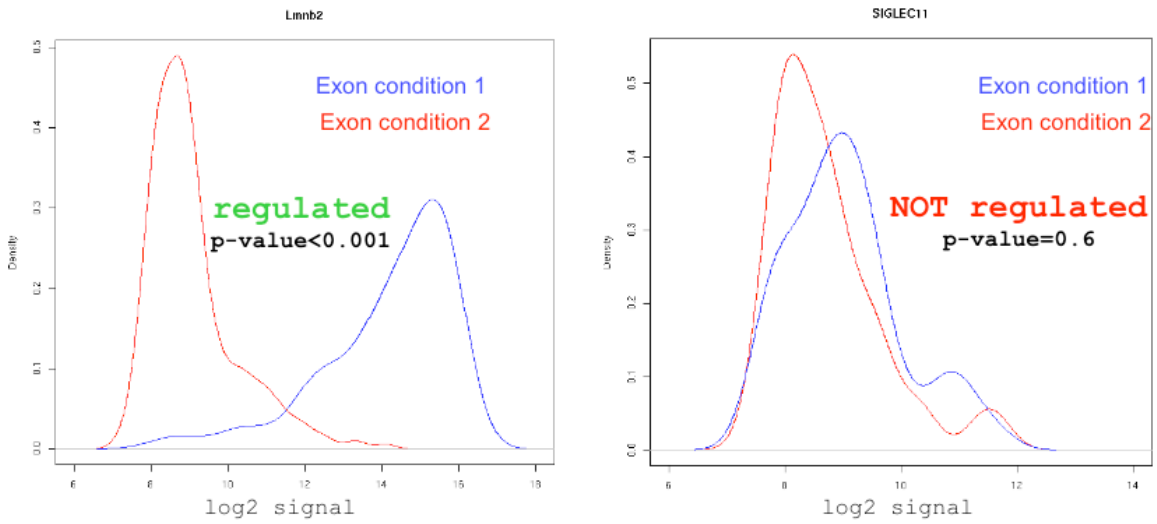


Fig.15. Distributions of the signals. The graphs represent the signal of all the probes belonging to the exon of the same gene in two different conditions. For gene A the difference between the two distribution is statically significant, but for the second one no.

Binding-Expression Correlation (BEC)

Merging expression and ChIP-chip results can be helpful when formulating hypotheses regarding, for example, the mechanistic implications of the binding of a transcription factor (TF) near to, or within, putative target gene loci. Outputs of ChIP-chip results from PeakPicker and expression data from TEA can be rapidly compared using BEC. For each gene, BEC gives the number of peaks that match the strict transcriptional unit or the user-defined putative promoter region around the TSS. Correlations between gene regulation or expression and TF binding can, therefore, immediately be evaluated. BEC integrates the results of expression analyses and ChIP-chip analyses. For each transcript, the number of peaks within the promoter region and/or the gene body is calculated.

Results

Isolation of origin-rich DNA

It has been shown that before DNA synthesis starts, very large protein complexes are assembled on replication origins, which in turn interact with the nuclear matrix, giving rise to the so-called replication factories (Anachkova, Djeliova et al. 2005; Jackson 2005). Furthermore, Schwartz et al. (Schwartz, Kahn et al. 2005), demonstrated that in CsCl equilibrium density gradients, the buoyant density of chromatin varies dramatically among different regions. In equilibrium density gradients, performed on shared cross-linked chromatin, it is possible to distinguish between bulk cross-linked chromatin fragments (buoyant density of 1,42-1,39 g/cm3), free DNA (≅1,69 g/cm3) and cross-linked proteins (≅1,25 g/cm3). These two assumption let us hypothesize that replication origins might have peculiar physical properties.

In order to demonstrate this hypothesis, we performed a CsCl density gradient on shared cross-linked chromatin of U937-PR9 cells (and of HeLa cells, where most of known replication origins have been mapped). After gradient fractionation we purified DNA and proteins from each fraction and assayed their distribution by gel electrophoresis (fig. 16). Ethidium bromide staining of purified DNA (fig. 16a upper panel) shows that the most of DNA is found in the central fractions of the gradient, where, as reported, it is fractionated the bulk of chromatin. The distribution of the DNA of two known replication origins (MCM4, TOP1) along the

gradient was then assayed by qRT-PCR (fig. 16a lower panel). Normalizing the data with the total DNA content of each fraction, we observed that the MCM4 and the TOP1 origins show a modest, but significant enrichment in low-density fractions, and a strong enrichment in the high-density fractions of the gradient. Thus, high-density fractions of CsCl gradient contain naked DNA strongly enriched in replication origins. Low-density fractions are particularly enriched in proteins and histone distribution seems to follow the total protein profile (fig. 16b).

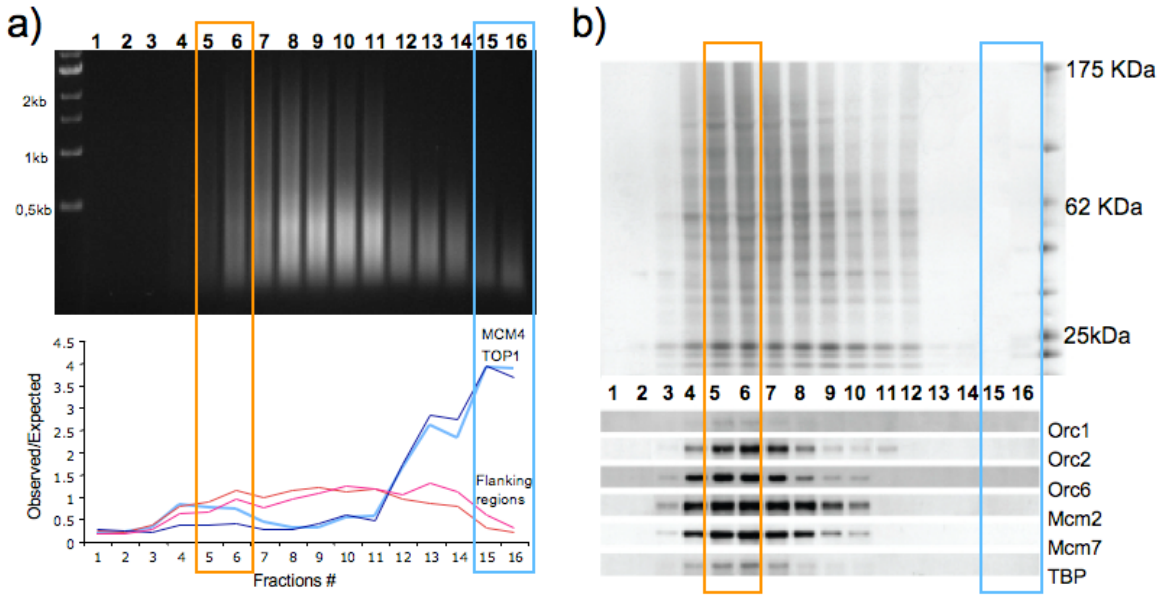


Fig. 16. Replication origins do not segregate randomly in a CsCl equilibrium gradient. a) Ethidium bromide staining of the DNA extracted from each single fraction of the gradient shows that the high amount of DNA segregates in the central part of the gradient. In the lower panel qRT-PCR of two different know replication origins shows a high enrichment of the amplicons located on the origin (blue lines) in the high-density fraction of the gradient and a small but significant enrichment in the low density fraction (marked with the orange rectangle). b) Comassie staining of proteins extracted from the same gradient and in the lower panel western blot of the same proteins with the antibodies written on the right.

The distribution of different subunits of the pre-RC (fig. 16b lower panel), assayed by western blot, shows that all the ORC and MCM proteins “peak” in low-density fractions. Also TBP, a protein of the basal transcription machinery, shows the same distribution, suggesting that both replication and transcription complexes co-segregate at low-density fractions. The combined analysis of protein and DNA distribution suggested us the use of low-density fractions chromatin, containing replication proteins, in ChIP assays using an antibody directed against Orc1, that is known to interact with actively firing origins (Ohta, Tatsumi et al. 2003).

Identification of new replication origins

NimbleGen custom tiled array

In order to map new replication origins, DNA populations obtained from high-density fraction number 16 (FRA 16) and anti-Orc1 ChIP of low-density fraction, were labeled and used as probes to hybridize a DNA tiled chip. Chip used for this hybridization is a custom tiled chip of the entire chromosome 19 produced by NimbleGen with an isothermal design. The resolution is about 50 bps and the number of probes spotted on the array is 385.796 with an average length of 47 nucleotides. The coverage is 28% of the entire chromosome, due to masking of the repeats. In figure 17 is shown the spatial distribution of the probes along the chromosome 19. The presence of a great amount of the repeats near left part of the centromeric region leads to an incomplete coverage of the chromosome in this specific region (Fig. 17).

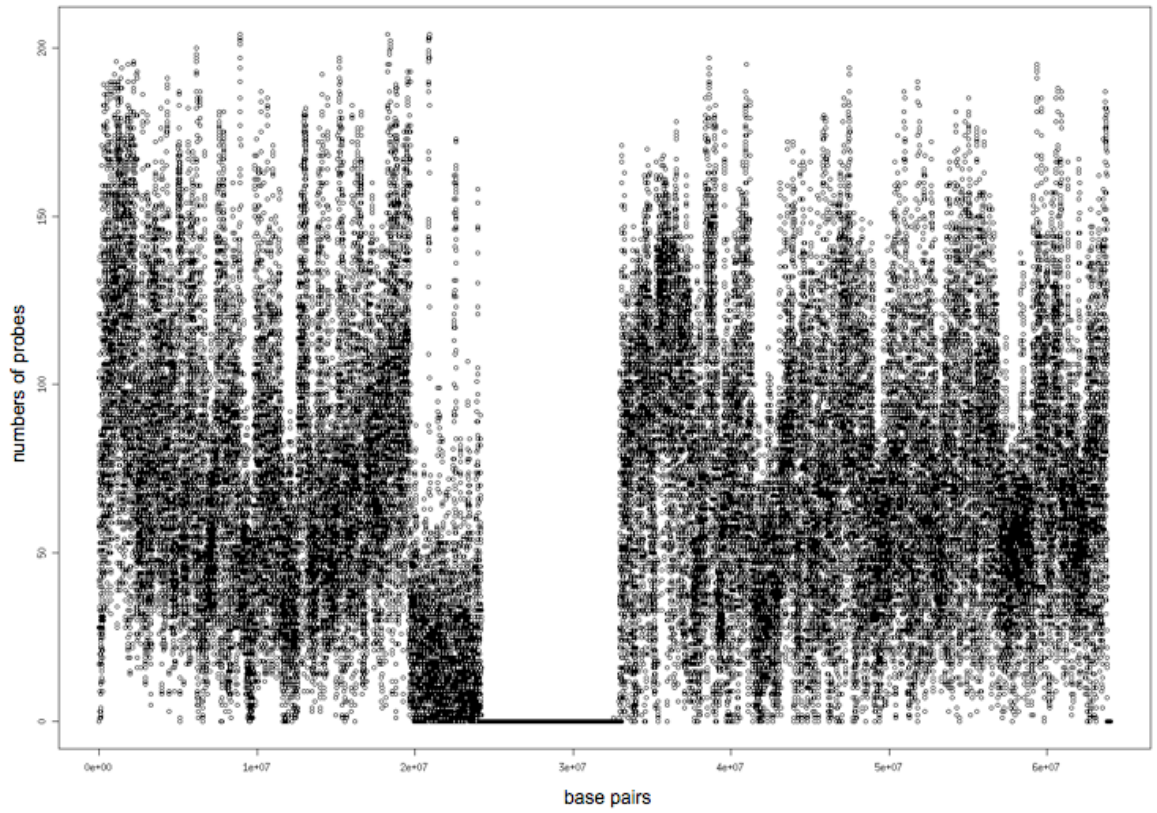


Fig. 17. Probe density distribution along chromosome 19. Each point represents the number of probes counted in a defined window of 10 Kb. Region without signal represents the centromeric region. Close by the centromer (on the left) a region with a low-resolution cause by the high number of repeats.

Raw Data Analysis and Normalization

The two chips used for the hybridization have been analyzed using CARPET pipeline (Cesaroni, Cittaro et al. 2008). An image of the two chips has been generated using ChipView, in order to verify the goodness of the hybridization. The two chips showed a high-quality hybridization without the presence of artifact (fig. 18a). Subsequently, normalization procedure has been applied on the data

using PPT (PreProcess for Tiling). Both the datasets have been normalized by Bi-Weight mean method to scale the distribution around 0 (fig. 18b).

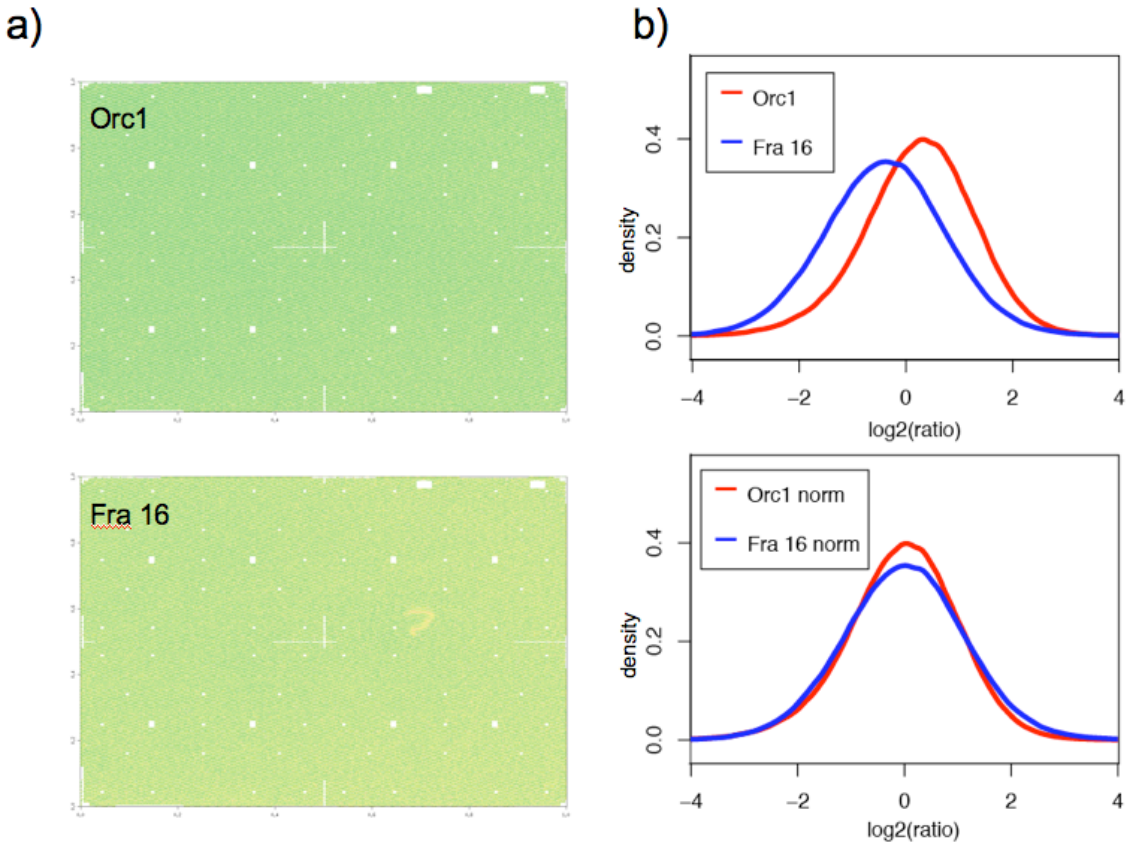


Fig. 18. Raw data analysis and normalization. a) Images of the two chips have been generated by ChipView implemented in CARPET (Cesaroni, Cittaro et al. 2008). The due images don't show hybridization problems and great artifacts are not present. b) Raw signals distribution of the two chips before and after Bi-weight normalization procedure. The normalization has been performed using PPT implemented in CARPET.

Enriched regions identification

After an in-depth raw data check and performing normalization procedure, a statistical analysis to identify enriched regions (also termed “peaks”) has been carried out. We used PeakPicker algorithm implemented in CARPET. We setup four parameters to define a peak:

a) sliding window of 500 bps

b) threshold of 95 percentile on the raw signal

c) log2(p-value) higher than 7

d) at least 3 neighbor probes that pass the b) and c) thresholds.

We estimate these parameters by a permutations procedure, trying to have less the 5% of false positive. We found 1059 peaks for Orc1 chip and 1178 for FRA 16 chip. A region with binding of Orc1 and enrichment in naked DNA (FRA 16) was defined as replication origin. We checked for common regions between the two dataset, using the Com&Uni program implemented in CARPET and found 217 regions in common: we considered these regions to represent putative replication origins (fig. 19).

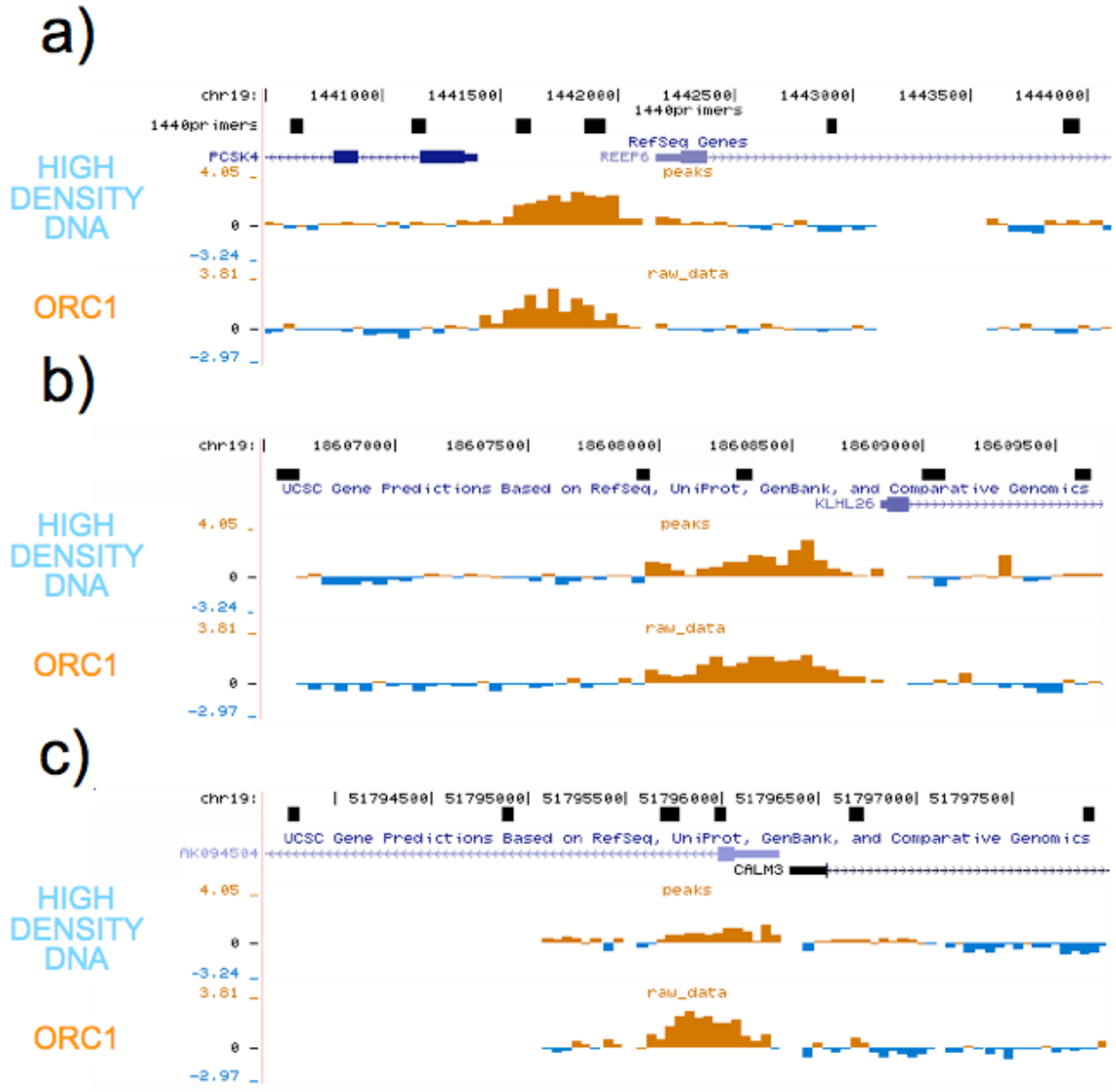


Fig. 19. Three examples of putative replication origins.

Validations

After the identification of the putative replication origins the subsequently step was to validate the results. The validation has been done on two levels: validation of Orc1 binding and validation of the firing activity of the origins. We tested 66 Orc1 binging regions with ChIP assays on chromatin from low-density fractions of the gradient. We found that 51 regions were positive, while 15 were negative, so we had a validation rate of 77% (fig. 20a). To confirm the firing activity of the

validated Orc1 positive regions we performed Nascent Strand Abundance (NSA) assay (Giacca, Pelizon et al. 1997). With this assay is possible to isolate, through a neutral sucrose gradient, short neo-synthesized DNA fragment (1-2kb) from the total genomic DNA. We then tested by qRT-PCR the enrichment of the region, where we map Orc1 binding, versus a couple of flanking regions. If the ratio scored for the tested new origin was higher than the observed for the MCM4/PRKDC, our positive control, we confirmed the mapping of a new active origin. We analyzed 28 Orc1 positive binding sites and we found that 21 were positive for NSA and 7 not with a validation rate of 78% (fig. 20b). The summary of the results is reported in table 1. Probably Orc1 is involved in other cellular process.

Orc1 binding	Validation rate=77%	
	Tested=66	Positive=51 Negative=15
Nascent strand assay	Validation rate=78%	
	Tested=28	Positive=21 Negative=7

Table 1. Summary of validation results for the 217 putative replication origins

We finally tested by ChIP experiments on total chromatin the presence of other proteins of the pre-RC (Orc2 and Mcm7) at the newly identified replication origins (fig 20c). Although at the moment we have tested only few regions, all of them confirm the assembly of all the pre-RC on our candidate replication origins.

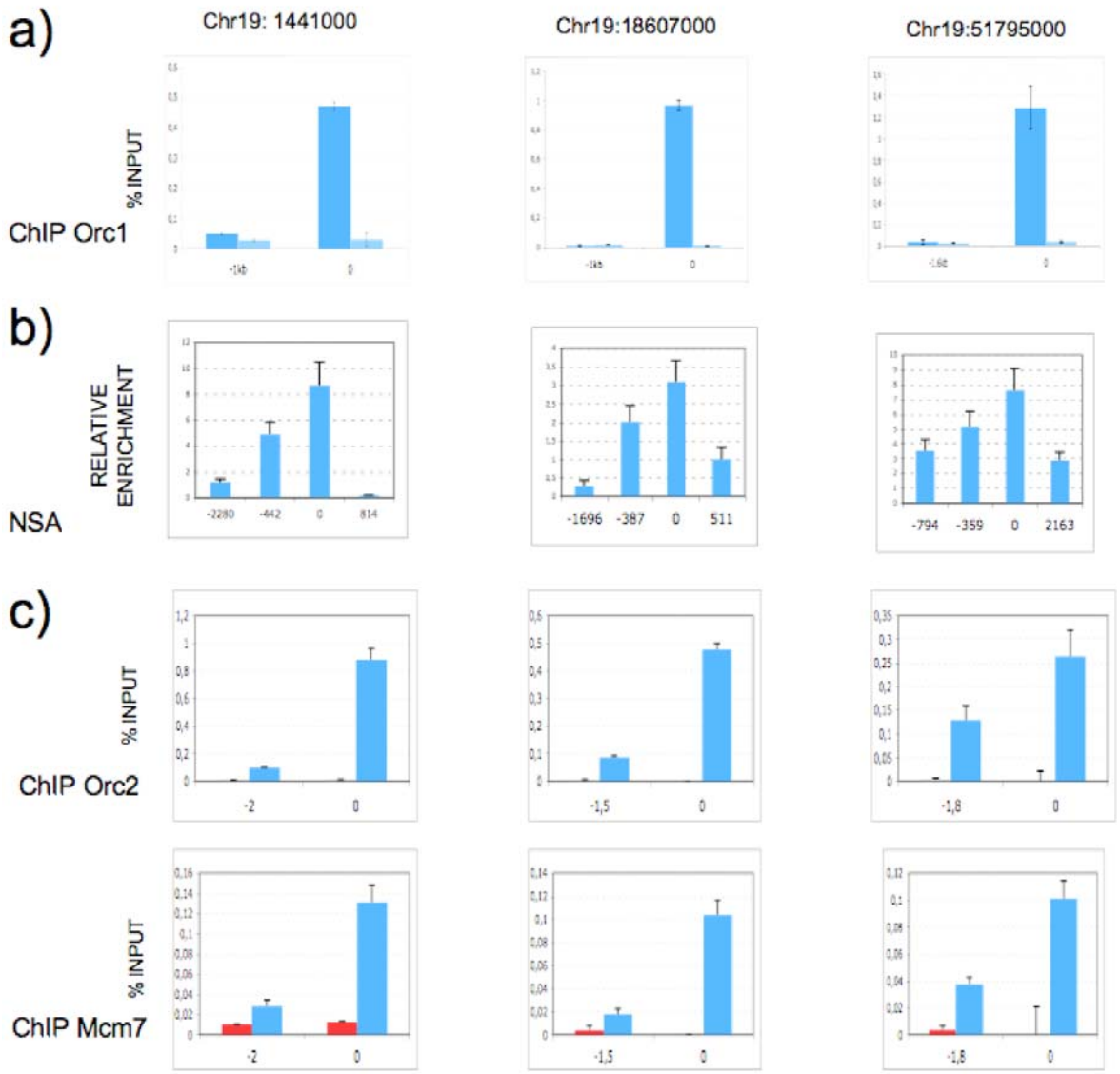


Fig. 20. Origin validation panel. a) ChIP Orc1 for three different replication origins. b) NSA validation. c) ChIP Orc2 and Mcm7.

Sequence analysis

To understand if some information was stored in the sequence and to better understand if there was any different between validated and not-validated replication origins, we performed a sequence analysis of the 28 regions tested by

NSA. We divided the dataset in two subsets: NSA positive regions and NSA negative regions. We approached the analysis in two different ways. We used MEME to find out if a specific motif was present in our positive dataset, and Clover to find out which specific transcription factor motifs were enriched in positive dataset against negative dataset. We ran MEME with the default parameters and it figured out a specific sequence with a CCAAT box (fig. 20). This motif was then compared with the transcription factor binding database Jaspar and it resulted very similar to the Nuclear Transcription Factor Y (NF-Y) binding site (p<0.0001). The CCAAT box has been already shown to be the binding of site for NF-Y (Borghini, Vargiolu et al. 2006).

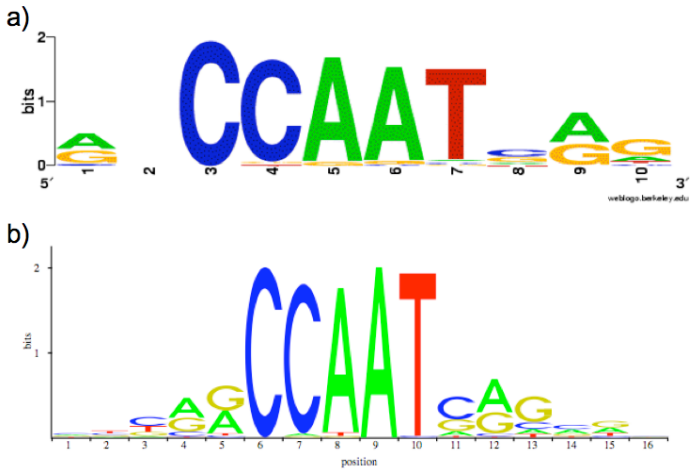


Fig. 20. Sequence found by MEME analysis (a) compared with the sequence of the NF-Y binding site (b). The matrix b has been used for the Clover analysis and was enriched in positive NSA dataset compared with negative NSA dataset.

Clover has been run on the positive and negative NSA dataset. We use two different backgrounds: all the promoter regions of the RefSeq database and the entire chromosome 20. We set the threshold to 6 for the raw score and to 0.01 for the p-value. Notably from the Clover analysis one of the most enriched transcription factor in the positive dataset was just NF-Y with a p-value ~= 0 (Data not shown).

NF-Y binging to the replication origins

The presence of the CCAAT box inside most of the newly identified replication origins and the identification of NF-y matrix as one of the most over-represented matrix in our dataset, suggests a possible biological role for NF-Y transcription factor in specifying active replication origins. We first confirmed, by ChIP on total chromatin, NF-Y binding to all the validated replication origins, and then we hybridized labeled ChIPped DNA with anti-NF-Y antibody onto the human chromosome 19 array. As well as the other chips, NF-Y has been checked for the quality of the hybridization and for the distribution of the intensities. The background resulted to be very low cause by good antibody specificity that performed a very specific enrichment. Therefore we decided to decrease the signal threshold to 90 percentile. The consequentially statistical analysis (Cesaroni, Cittaro et al. 2008) revealed 2,387 “peaks”, among which only 101 overlapped with the 217 putative replication origins dataset. 40 out of 101 putative replication origins was already tested for Orc1 binding and 19 for nascent strand. Interestingly, the validation rate of these “new” 101 putative replication origins raised from 78% of 217 to 95% of 101 (table 2).

Orc1 binding	Validation rate=90%	
	Tested=40	Positive=36 Negative=4
Nascent strand assay	Validation rate=95%	
	Tested=19	Positive=18 Negative=1

Table 2. Validation rate of the peaks common between Orc1, Fra 16 and NF-Y

Previous NF-Y ChIP-on-chip data and bioinformatics analysis showed that 60% of all the promoter regions are bound by this protein and contain a CCAAT box (Testa, Donati et al. 2005).

In order to understand which was the relation between CCAAT containing replication origins and genes, we annotated putative replication origin positions considering genes TSSs, exons and introns. To perform this analysis we used GIN (Cesaroni, Cittaro et al. 2008). We downloaded the RefSeq table hg18 from UCSC Genome Browser containing 1637 different genes and we set the promoter regions as -2000 bp from Transcription Starting Site (TSS). The annotation analysis shown that 77% of our replication origins are located near gene promoters (-2000bp < TSS > +2000bp), figure 22, as already found for most if not all the known replication origins, in accordance with what reported by Cadoret et. Al (Cadoret, Meisch et al. 2008), that mapped 283 replication origins overlapping with gene regulatory elements.

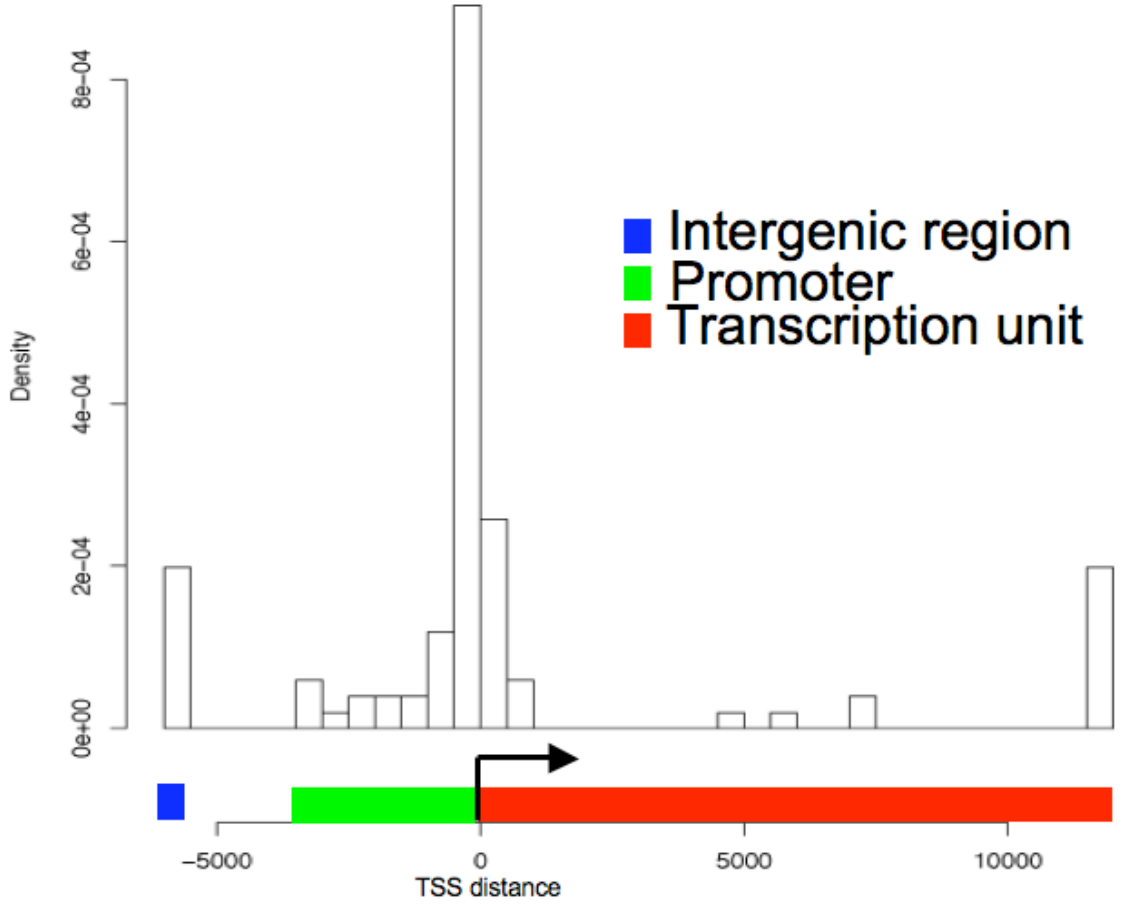


Fig. 22. Distribution of 101 putative replication origins around TSS. More than 70% of the peaks we found are located in the proximity of Transcription Starting Site of the genes. Only a few percentage of the peaks are located in intergenic regions.

Gene expression chip

Considering the annotation results, we decided to further investigate the possible role of gene regulation and transcription in origin selection. At this purpose we performed an expression tiling array, on the Chr19 chip. The probe for this hybridization was a cDNA retro transcribed from total mRNA of U937-PR9 cells.

To analyze this kind of chip we create ENO to annotated probes and TEA to calculate expression (Cesaroni, Cittaro et al. 2008). We use the RefSeq table downloadable from UCSC genome browser to annotate the probes by ENO. Then applying TEA we found that 761 out of 1637 genes were called expressed with a p-value < 0.01. 77 putative replication origins out of 101 has been found within the promoter region and 68 out of 77 genes are called expressed by TEA. Notably the average expression level of the 77 genes located close to new replication origins was higher than the average level of all the chromosome 19 genes (fig. 23).

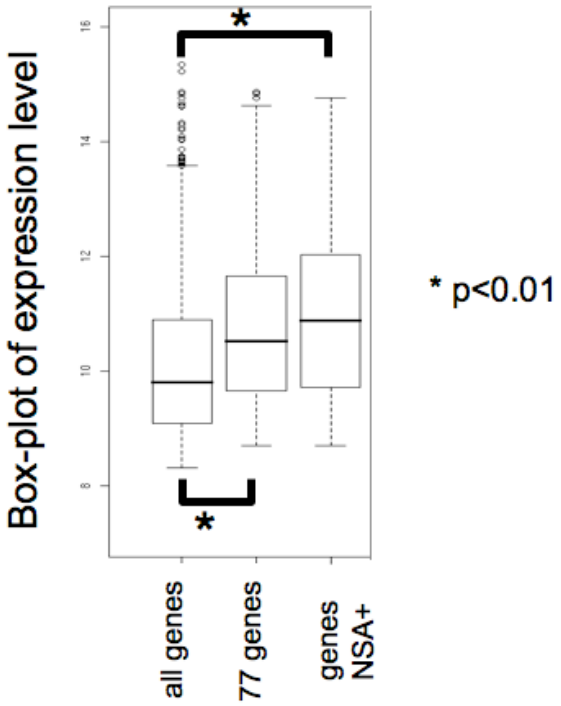


Fig. 23. Box plot of the different classes of the expression chip. First box-plot represents the distribution of the intensity of all the genes presented on the chip. The intensity of each gene has been calculated as the average of all the probe signal matching exons. Second box-plot represents the intensity distribution of 77 genes close to putative replication origins. Third box-plot represents the distribution of only the genes with a validated replication origin on the promoter. The differences between the first and the second and between the first and the third are both statistically significant.

The difference in the expression level of the two distributions has been tested by T-test with a p-value <0.01, mining that these two groups belong to two different populations. Plotting only the validated origins close to genes, the difference between the two groups increases and the statistical significance as well. This result has been further confirmed, by the binding of Polymerase II (Pol II) and histone H3 lysine 79 dimethylation (H3K79me2) detected on the promoter and within these genes (fig 24), that taken together could be considered good markers of gene expression (Guenther, Levine et al. 2007).

Fig. 24. Replication origins are located close to promoters of expressed genes. Three examples of validated replication origin posed into the promoter region of genes. PolII and H3K79me2 results are reported expressed as %input.

Chromatin structure

To test if expression level, Pol II binding, H3K79me2 and origin choice correlate with an open chromatin structure, we performed chromosome walking of H3 binding by ChIP and qRT-PCR. We observed that in all the regions analyzed, the replication origin mapped, Orc1 and high-density fraction peaks do coincide with a drop in the binding level of H3 (fig. 24), suggesting the presence of a more relaxed chromatin structure.

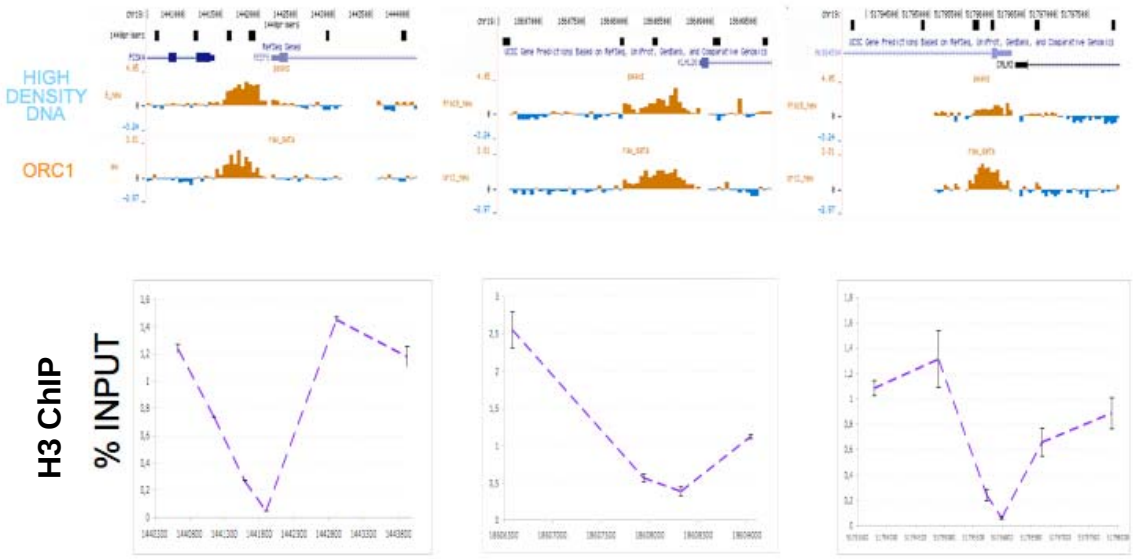


Fig. 24. H3 content is lower at replication origins. H3-ChIP profile on replication origins shows a decrease in H3 content, corresponding to Orc1 and high density fraction peaks. In the same positions we observed high levels of NFY binding.

Discussion

Mammalian genomes are thought to contain several thousand replication origins, but the lack of a generally applicable and functional screening assay has slowed down their identification and the definition of their characteristic features. In this thesis we propose a novel high-throughput approach for identifying new replication origins. Our strategy consists in gathering a compilation of potential human DNA origins achieved by ORC1-binding DNA fragments enriched by chromatin immunoprecipitation (ChIP) and in their hybridization on a tiling custom array. The key point of our procedure is that, before the immunoprecipitation, we introduce a biochemical fractionation phase which results in a significant DNA-ORC1 binding enrichment: this fractionation step allowed us to both ChIP ORC1 from the specific protein enriched low-density fraction and to isolate naked DNA from the high-density fraction 16. Both these two DNA populations have been then hybridized on a NimbleGen tiling custom chip of the whole chromosome 19. Genomic binding regions shared by the two experiments results in the identification of new putative replication origins.

With this approach we have initially mapped and characterized 217 new putative human replication origins, on the basis of their biophysical properties. On a subsample of 18 putative origins we have further shown by conventional chromatin immunoprecipitation that, together with Orc1, both Orc2 and Mcm7 proteins were recruited on those replication origins, indicating the presence of the entire pre-RC complex on these sites. Then, we have established, by Nascent Strand Assay that those regions are actively firing. Looking for consensus sequence that would

be recognized by the origin recognition complex (ORC), we found that all the 21 validated regions shared the presence of a CCAAT box consensus sequence, considered the binding site for the NF-Y transcription factor. By standard ChIP and ChIP-on-chip experiments we have next confirmed that finally all the 101 identified replication origins indeed bound NF-Y-A.

The annotation of all the 101 replication origins revealed that the most of them were positioned near gene transcription start site, in fact 77 replication origins mapped within -2000 bp +1000 bp from TSS. Combining information coming from ChIP of Pol II and H3K79me2 with tiling chip expression analysis we confirmed that 68 out of 77 genes are actively transcribed. Finally, H3 ChIP revealed a decrease in histone content exactly where we mapped replication firing and NF-Y binding.

The identification of the CCAAT box and the binding of NF-Y let us to hypothesize that NF-Y transcription factor may play an important role in replication origin determination and selection. The finding that our new identified replication origins are localized close to gene TSSs is in accordance with, and expand, previous indications attained from the few fully characterized human replication origins, that was shown to map near gene promoters (e.g. MCM4/PRKDC, TOP1). This is in agreement also with what recently published by Cadoret et al (Cadoret, Meisch et al. 2008), that map replication origins within regulatory elements. Characteristics such the high expression level of neighboring gene transcripts, the co-occurrence of ORC1 and the transcription factor (NF-Y) binding and the observed drop in the histone H3, all strongly suggest that starting of DNA replication could be favored by the presence of an “open” chromatin state. The finding that the transcription factor NF-Y is bound to all our newly identified replication origins is an important

new discovery; NF-Y could work in replication initiation as an “organizer” of the chromatin structure, as it does in transcription, at the promoter site (Ceribelli, Dolfini et al. 2008). In addition, NF-Y, through its histone-like fold domain, could substitute for core histones, as it does in other genomic regions (Gatta and Mantovani 2008), also at replication origin sites.

Beside the demanding wet-lab activity, the huge amount of data, produced all along the project, required an equivalent bioinformatics effort. To answer to this critical need, we developed a custom and easy interface for the analysis, for comparing and for managing the results. To address this necessity we created CARPET, a collection of tools integrated on the Galaxy2 web-based platform (Blankenberg, Taylor et al. 2007). That package was recently made public available through the web server of the IFOM CAMPUS institute and published. CARPET provides a very powerful, user-friendly, and comprehensive set of tools for ChIP-chip and expression tiling analysis. With CARPET we produced not only specific instruments for the biologists, but also a pipeline to follow in order to analyze and integrate different type of genome-wide data. No bioinformatics skills are required to use CARPET, all the scripts are accessible through a friendly interface perfectly integrated in GALAXY2 environment and a detailed manual is available. In CARPET all the main steps of a complete analysis are present: i) visualization of the chip surface, ii) different strategies of normalization, iii) peaks identification, iv) expression estimation, v) comparison and integration of the results. Other main assets are that i) it provides a collection of coordinated programs directly accessible through the web on our site; ii) no knowledge is needed of programming languages, such as R or Perl, iii) the integration of CARPET with the Galaxy2 environment makes data storage and sharing very

easy and allows the direct graphic visualization of results as custom tracks in the UCSC Genome Browser, iv) it facilitates the comparison of results obtained through different experimental approaches. The development of CARPET results finally functional and crucial for the achievement and integration of the replication origins identification and characterization.

Few previous attempts to map DNA replication origins were previously done (Lucas, Palakodeti et al. 2007; Cadoret, Meisch et al. 2008), and in all the cases they end up with the identification of a limited number of replication origins. We have set up a new method to map replication origins that, in a near future, could be easily applied to all the genome, possibly with the cheapest variant of the high-throughput sequencing. This it will surely contribute to better understand which are all the different aspects that contribute to drive and control DNA replication origins in a certain position of the genome.

The genome wide recognition and characterization of all replication origins, possibly in different cellular systems, will be instrumental not only to understand the role of DNA replication in physiological states, but also to deeply comprehend which are the pathological implications of altered replication processes.

References

Anachkova, B., V. Djeliova, et al. (2005). "Nuclear matrix support of DNA replication." J Cell Biochem **96**(5): 951-61.

Bailey, T. L. and C. Elkan (1994). "Fitting a mixture model by expectation maximization to discover motifs in biopolymers." Proc Int Conf Intell Syst Mol Biol **2**: 28-36.

Bell, S. P. (2002). "The origin recognition complex: from simple origins to complex functions." Genes Dev **16**(6): 659-72.

Bell, S. P. and A. Dutta (2002). "DNA replication in eukaryotic cells." Annu Rev Biochem **71**: 333-74.

Bell, S. P. and B. Stillman (1992). "ATP-dependent recognition of eukaryotic origins of DNA replication by a multiprotein complex." Nature **357**(6374): 128-34.

Bielinsky, A. K. and S. A. Gerbi (2001). "Where it all starts: eukaryotic origins of DNA replication." J Cell Sci **114**(Pt 4): 643-51.

Birney, E., J. A. Stamatoyannopoulos, et al. (2007). "Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project." Nature **447**(7146): 799-816.

Blankenberg, D., J. Taylor, et al. (2007). "A framework for collaborative analysis of ENCODE data: making large-scale analyses biologist-friendly." Genome Res **17**(6): 960-4.

Borghini, S., M. Vargiolu, et al. (2006). "Nuclear factor Y drives basal transcription of the human TLX3, a gene overexpressed in T-cell acute lymphocytic leukemia." Mol Cancer Res **4**(9): 635-43.

Buck, M. J. and J. D. Lieb (2004). "ChIP-chip: considerations for the design, analysis, and application of genome-wide chromatin immunoprecipitation experiments." Genomics **83**(3): 349-60.

Buck, M. J., A. B. Nobel, et al. (2005). "ChIPOTle: a user-friendly tool for the analysis of ChIP-chip data." Genome Biol **6**(11): R97.

Bulyk, M. L. (2006). "DNA microarray technologies for measuring protein-DNA interactions." Curr Opin Biotechnol **17**(4): 422-30.

Cadoret, J. C., F. Meisch, et al. (2008). "Genome-wide studies highlight indirect links between human replication origins and gene regulation." Proc Natl Acad Sci U S A **105**(41): 15837-42.

Carroll, J. S., C. A. Meyer, et al. (2006). "Genome-wide analysis of estrogen receptor binding sites." Nat Genet **38**(11): 1289-97.

Ceribelli, M., D. Dolfini, et al. (2008). "The histone-like NF-Y is a bifunctional transcription factor." Mol Cell Biol **28**(6): 2047-58.

Cesaroni, M., D. Cittaro, et al. (2008). "CARPET: a web-based package for the analysis of ChIP-chip and expression tiling data." Bioinformatics.

Cvetic, C. and J. C. Walter (2005). "Eukaryotic origins of DNA replication: could you please be more specific?" Semin Cell Dev Biol **16**(3): 343-53.

Davey, M. J., L. Fang, et al. (2002). "The DnaC helicase loader is a dual ATP/ADP switch protein." EMBO J **21**(12): 3148-59.

DePamphilis, M. L. (2003). "The 'ORC cycle': a novel pathway for regulating eukaryotic DNA replication." Gene **310**: 1-15.

Di Paola, D., G. B. Price, et al. (2006). "Differentially active origins of DNA replication in tumor versus normal cells." Cancer Res **66**(10): 5094-103.

Diffley, J. F. (2004). "Regulation of early events in chromosome replication." Curr Biol **14**(18): R778-86.

Dillin, A. and J. Rine (1997). "Separable functions of ORC5 in replication initiation and silencing in Saccharomyces cerevisiae." Genetics **147**(3): 1053-62.

Edwards, M. C., A. V. Tutter, et al. (2002). "MCM2-7 complexes bind chromatin in a distributed pattern surrounding the origin recognition complex in Xenopus egg extracts." J Biol Chem **277**(36): 33049-57.

Frith, M. C., Y. Fu, et al. (2004). "Detection of functional DNA motifs via statistical over-representation." Nucleic Acids Res **32**(4): 1372-81.

Garg, P. and P. M. Burgers (2005). "DNA polymerases that propagate the eukaryotic DNA replication fork." Crit Rev Biochem Mol Biol **40**(2): 115-28.

Gatta, R. and R. Mantovani (2008). "NF-Y substitutes H2A-H2B on active cell-cycle promoters: recruitment of CoREST-KDM1 and fine-tuning of H3 methylations." Nucleic Acids Res.

Giacca, M., C. Pelizon, et al. (1997). "Mapping replication origins by quantifying relative abundance of nascent DNA strands using competitive polymerase chain reaction." Methods **13**(3): 301-12.

Gilbert, D. M. (1998). "Replication origins in yeast versus metazoa: separation of the haves and the have nots." Curr Opin Genet Dev **8**(2): 194-9.

Gilbert, D. M. (2004). "In search of the holy replicator." Nat Rev Mol Cell Biol **5**(10): 848-55.

Glynn, E. F., P. C. Megee, et al. (2004). "Genome-wide mapping of the cohesin complex in the yeast Saccharomyces cerevisiae." PLoS Biol **2**(9): E259.

Guenther, M. G., S. S. Levine, et al. (2007). "A chromatin landmark and transcription initiation at most promoters in human cells." Cell **130**(1): 77-88.

Harland, R. M. and R. A. Laskey (1980). "Regulated replication of DNA microinjected into eggs of Xenopus laevis." Cell **21**(3): 761-71.

Harvey, K. J. and J. Newport (2003). "Metazoan origin selection: origin recognition complex chromatin binding is regulated by CDC6 recruitment and ATP hydrolysis." J Biol Chem **278**(49): 48524-8.

Hudson, M. E. and M. Snyder (2006). "High-throughput methods of regulatory element discovery." Biotechniques **41**(6): 673, 675, 677 passim.

Jackson, D. A. (2005). "The amazing complexity of transcription factories." Brief Funct Genomic Proteomic **4**(2): 143-57.

Jacob, F. and S. Brenner (1963). "[On the regulation of DNA synthesis in bacteria: the hypothesis of the replicon.]." C R Hebd Seances Acad Sci **256**: 298-300.

Johnson, A. and M. O'Donnell (2005). "Cellular DNA replicases: components and dynamics at the replication fork." Annu Rev Biochem **74**: 283-315.

Johnson, D. S., W. Li, et al. (2008). "Systematic evaluation of variability in ChIP-chip experiments using predefined DNA targets." Genome Res **18**(3): 393-403.

Kaguni, J. M. (2006). "DnaA: controlling the initiation of bacterial DNA replication and more." Annu Rev Microbiol **60**: 351-75.

Kapranov, P., A. T. Willingham, et al. (2007). "Genome-wide transcription and the implications for genomic organization." Nat Rev Genet **8**(6): 413-23.

Kim, T. H., Z. K. Abdullaev, et al. (2007). "Analysis of the vertebrate insulator protein CTCF-binding sites in the human genome." Cell **128**(6): 1231-45.

Kim, T. H., L. O. Barrera, et al. (2005). "A high-resolution map of active promoters in the human genome." Nature **436**(7052): 876-80.

Kong, D., T. R. Coleman, et al. (2003). "Xenopus origin recognition complex (ORC) initiates DNA replication preferentially at sequences targeted by Schizosaccharomyces pombe ORC." EMBO J **22**(13): 3441-50.

Li, C. J. and M. L. DePamphilis (2002). "Mammalian Orc1 protein is selectively released from chromatin and ubiquitinated during the S-to-M transition in the cell division cycle." Mol Cell Biol **22**(1): 105-16.

Liu, G., M. Malott, et al. (2003). "Multiple functional elements comprise a Mammalian chromosomal replicator." Mol Cell Biol **23**(5): 1832-42.

Lucas, I., A. Palakodeti, et al. (2007). "High-throughput mapping of origins of replication in human cells." EMBO Rep **8**(8): 770-7.

MacNeill, S. A. (2001). "DNA replication: partners in the Okazaki two-step." Curr Biol **11**(20): R842-4.

Maga, G. and U. Hubscher (2003). "Proliferating cell nuclear antigen (PCNA): a dancer with many partners." J Cell Sci **116**(Pt 15): 3051-60.

Malott, M. and M. Leffak (1999). "Activity of the c-myc replicator at an ectopic chromosomal location." Mol Cell Biol **19**(8): 5685-95.

Mechali, M. and S. Kearsey (1984). "Lack of specific sequence requirement for DNA replication in Xenopus eggs compared with high sequence specificity in yeast." Cell **38**(1): 55-64.

Messer, W. (2002). "The bacterial replication initiator DnaA. DnaA and oriC, the bacterial mode to initiate DNA replication." FEMS Microbiol Rev **26**(4): 355-74.

Mockler, T. C., S. Chan, et al. (2005). "Applications of DNA tiling arrays for whole-genome analysis." Genomics **85**(1): 1-15.

Newlon, C. S. and J. F. Theis (1993). "The structure and function of yeast ARS elements." Curr Opin Genet Dev **3**(5): 752-8.

Ohta, S., Y. Tatsumi, et al. (2003). "The ORC1 cycle in human cells: II. Dynamic changes in the human ORC complex during the cell cycle." J Biol Chem **278**(42): 41535-40.

Remus, D., E. L. Beall, et al. (2004). "DNA topology, not DNA sequence, is a critical determinant for Drosophila ORC-DNA binding." EMBO J **23**(4): 897-907.

Ren, B., F. Robert, et al. (2000). "Genome-wide location and function of DNA binding proteins." Science **290**(5500): 2306-9.

Robinson, N. P. and S. D. Bell (2005). "Origins of DNA replication in the three domains of life." FEBS J **272**(15): 3757-66.

Royce, T. E., J. S. Rozowsky, et al. (2005). "Issues in the analysis of oligonucleotide tiling microarrays for transcript mapping." Trends Genet **21**(8): 466-75.

Rusche, L. N., A. L. Kirchmaier, et al. (2002). "Ordered nucleation and spreading of silenced chromatin in Saccharomyces cerevisiae." Mol Biol Cell **13**(7): 2207-22.

Scacheri, P. C., G. E. Crawford, et al. (2006). "Statistics for ChIP-chip and DNase hypersensitivity experiments on NimbleGen arrays." Methods Enzymol **411**: 270-82.

Schwartz, Y. B., T. G. Kahn, et al. (2005). "Characteristic low density and shear sensitivity of cross-linked chromatin containing polycomb complexes." Mol Cell Biol **25**(1): 432-9.

Stillman, B. (2005). "Origin recognition and the chromosome cycle." FEBS Lett **579**(4): 877-84.

Stinchcomb, D. T., K. Struhl, et al. (1979). "Isolation and characterisation of a yeast chromosomal replicator." Nature **282**(5734): 39-43.

Tabancay, A. P., Jr. and S. L. Forsburg (2006). "Eukaryotic DNA replication in a chromatin context." Curr Top Dev Biol **76**: 129-84.

Testa, A., G. Donati, et al. (2005). "Chromatin immunoprecipitation (ChIP) on chip experiments uncover a widespread distribution of NF-Y binding CCAAT sites outside of core promoters." J Biol Chem **280**(14): 13606-15.

Toedling, J., O. Skylar, et al. (2007). "Ringo--an R/Bioconductor package for analyzing ChIP-chip readouts." BMC Bioinformatics **8**: 221.

Vashee, S., C. Cvetic, et al. (2003). "Sequence-independent DNA binding and replication initiation by the human origin recognition complex." Genes Dev **17**(15): 1894-908.

Wei, C. L., Q. Wu, et al. (2006). "A global map of p53 transcription-factor binding sites in the human genome." Cell **124**(1): 207-19.

Wyrick, J. J., J. G. Aparicio, et al. (2001). "Genome-wide distribution of ORC and MCM proteins in S. cerevisiae: high-resolution mapping of replication origins." Science **294**(5550): 2357-60.

Zhang, Z. D., J. Rozowsky, et al. (2007). "Tilescope: online analysis pipeline for high-density tiling microarray data." Genome Biol **8**(5): R81.