Despite intense investigation, human replication origins and termini remain elusive. Existing data have shown strong discrepancies. Here we sequenced highly purified Okazaki fragments from two cell types and, for the first time, quantitated replication fork directionality and delineated initiation and termination zones genome-wide. Replication initiates stochastically, primarily within non-transcribed, broad (up to 150 kb) zones that often abut transcribed genes, and terminates dispersively between them. Replication fork progression is significantly co-oriented with the transcription. Initiation and termination zones are frequently contiguous, sometimes separated by regions of unidirectional replication. Initiation zones are enriched in open chromatin and enhancer marks, even when not flanked by genes, and often border ‘topologically associating domains’ (TADs). Initiation zones are enriched in origin recognition complex (ORC)-binding sites and better align to origins previously mapped using bubble-trap than λ-exonuclease. This novel panorama of replication reveals how chromatin and transcription modulate the initiation process to create cell-type-specific replication programs.
Genome-wide identification of mammalian replication origins has been attempted by various methods including trapping of replication-bubbles, purification of RNA-primed single DNA strands, immunoprecipitation of BrdU-labelled DNA or chromatin immunoprecipitation of origin recognition complex (ORC), followed by microarray hybridization or high-throughput sequencing. The number and position of origins thus detected present discrepancies (from 12,000 (ref. 4) to 250,000 (ref. 9) putative origins detected genome-wide). In addition, these methods cannot precisely quantify origin efficiency nor map termination.

An orthogonal method to map replication is based on the analysis of nucleotide compositional asymmetries of the two DNA strands. In most bacterial genomes, the GC and TA skews are strictly correlated with replication direction, suggesting that skew accumulation over evolutionary times. Thus, detection of mutational asymmetry of the leading and lagging strands led to the location of potential initiation and termination sites. The red and blue lines inside these structures indicate the Okazaki fragments synthesized during the progression of each fork. These Okazaki fragment populations are randomly sampled when extracted from asynchronously growing cell populations.

Skelton and Williams introduced a novel method that permits a high-resolution, quantitative analysis of replication fork initiation, progression and termination. They have been recently developed in Saccharomyces cerevisiae. The principle is to isolate and sequence Okazaki fragments to reveal the proportions of rightward- (R) and leftward- (L) moving forks throughout the genome. Zones of predominant initiation (termination) can then be detected as upward (downward) slopes in replication fork directionality (RFD = R – L) profiles (Fig. 1a–c). Since the very low amount of Okazaki fragments present in wild-type cells represents a challenge, these studies used conditional lethal mutants that massively increase Okazaki fragment abundance. The results were consistent with preexisting origin identifications and allowed to estimate origin efficiency. Furthermore, the results demonstrated that a large fraction of the genome can be replicated in either orientation and that termination occurs between origins at positions dictated by their relative firing times.

Here we present a new methodology to isolate and sequence Okazaki fragments (OK-Seq) that does not require overproducer mutant cells. This allowed us to determine RFD genome-wide in asynchronous cultures of two human cell lines, without any genetic modification. We infer the location of replication initiation and termination events at high resolution and paint a novel, comprehensive landscape of human genome replication in these two cell types.

We validate the replication model of N/U-domains. We identify in the two cell lines 9,386 and 5,684 (4,150 shared) broad initiation zones that typically support a single but randomly located initiation event. These zones are mostly non-transcribed, often but not always circumscribed by active genes, generally enriched in open chromatin and enhancer epigenetic marks, and often located at TAD borders. Initiation zones associated and not associated with active genes fire early and late in S phase, respectively. Replication terminates very dispersively between initiation zones, either within transcribed, early-replicating gene bodies or within large, late-replicating, silent DNA regions. A significant though incomplete co-orientation of replication with transcription is observed, due to preferential (but not exclusive) initiation of DNA replication upstream of active genes. The number of initiation zones detected in this work accounts for only a fraction of total initiation events. We, therefore, suggest that the human genome undergoes a superposition of efficient initiation at circumscribed ‘master’ zones identified here and more dispersed, less efficient initiation elsewhere.

**Results**

**Genome-wide profiling of RFD.** Our methodology is based on the replicative incorporation of the thymidine analogue B.

---

**Figure 1 | Principle of RFD mapping by OK-Seq.** (a) Schematic leftward (blue) and rightward (red) forks with Okazaki fragments. RFD is computed as the difference between the proportions of rightward- and leftward-moving forks in 1kb windows. (b) Expected RFD profiles for a fixed origin or a broad initiation zone. (c) Expected RFD profiles for a fixed terminus or a broad termination zone. The bubbles in a,b and X-shaped structures in c indicate the location of potential initiation and termination sites. The red and blue lines inside these structures indicate the Okazaki fragments synthesized during the progression of each fork. These Okazaki fragment populations are randomly sampled when extracted from asynchronously growing cell populations.
5-ethynyl-2'-deoxyuridine (EdU) followed by purification and sequencing of labelled Okazaki fragments (Fig. 2a and Methods section). Okazaki fragments purified from two cell types, HeLa (adenocarcinoma) and GM06990 (lymphoblastoid) revealed a totally new, full panorama of replication fork progression along the genome with excellent reproducibility (Fig. 2b,c, Supplementary Fig. 1). RFD profiles displayed series of alternating quasi-linear ascending, descending and flat segments (AS, DS, FS) of varying size and slope. RFD often reached values \( \leq 0.9 \) or \( \geq 0.9 \), indicating nearly complete purity of Okazaki fragments. A small-scale \(<300\,\text{kb}\) saw-tooth pattern was mostly observed in early-replicating regions (for example, 118–123 Mb, Fig. 2b,c). By contrast, most late-replicating regions displayed megabase-sized DS and/or FS of low absolute values of RFD (\( |\text{RFD}| \)) (for example, 115–118 Mb, Fig. 2b,c).

**OK-Seq validates replication model of N/U domains.** Previously, we detected in these cells, independently of skew analysis, 900–1,500 megabase-scaled domains characterized by parabolic (U-shaped) replication timing profiles and we predicted they would display N-shaped RFD profiles (Fig. 3a,b). This prediction results from the mathematical relationship between mean replication timing (MRT), fork speed (\( v \)) and RFD: \( \frac{\text{dMRT}}{\text{dx}} = \text{RFD}/v \) (refs 25,26,29) (the derivative of a parabolic profile is a straight line). In words, assuming \( v \) is constant, the steep slopes at the borders of the U predict a high [RFD] whereas the flat slope at the center of the U predicts a null RFD. In addition, U-domains frequently coincided with independently detected N-shaped domains of nucleotide compositional skew profile\(^{19,23,26,29}\) (Fig. 3c), proposed to reflect an N-shaped pattern of germline RFD\(^{19,24}\). OK-Seq fully verified the predicted 1–2 Mb N-shaped RFD gradients of replication timing U-domains and skew N-domains (Fig. 3d–f). These linear gradients of RFD likely result from cascades of origin firing propagating from domain borders to center\(^{15,23,26}\). The remarkable similarity of the RFD gradients deduced from skew\(^{19}\) or replication time\(^{25}\) or determined by OK-Seq provides a strong reciprocal validation of these data. Moreover, the superior resolution of OK-Seq reveals replication initiation and termination zones with much higher precision.

**Replication initiates stochastically within broad zones.** We used a hidden Markov model (HMM) to detect zones of predominant initiation as AS (see Methods section). The RFD values were computed within 15 kb sliding windows (stepped by 1 kb) to ensure reproducible AS detection (Fig. 4). In HeLa and GM06990, we detected 9,836 and 5,684 AS (mean size 30 kb, range 6–150 kb), covering 11.5 and 7.1% of the genome, respectively,
Figure 3 | OK-Seq corroborates the replicative organization of N/U-domains. (a) U-domains of replication timing (top) are predicted to show an N-shaped RFD profile (bottom). The steep slopes at U-domain borders predict a high RFD whereas the flat slope at the central, late-replicating valley predicts a null RFD. Blue and red arrows, expected proportions of Okazaki fragments from either strand across zones of predominant initiation (IZ) at borders and predominant termination (TZ) at center of U-domain. Dark line, expected RFD profile. The mathematical relationship dMRT/dx = RFD/v implies that RFD increases (AS) when the timing profile is convex (d²MRT/dx² > 0) and decreases (DS) when it is concave (d²MRT/dx² < 0; note that the time axis is oriented from top to bottom). (b) HeLa (orange) and GM06990 (teal blue) replication timing profiles showing four adjacent U-domains. (c) Nucleotide compositional skew profile (see Methods section) showing skew N-domains matching replication timing U-domains; (d) N-shaped RFD profiles determined by OK-Seq matching N-domains. (e) Heatmap of RFD along the 1,410 HeLa and 878 GM06990 U-domains, ordered by length and centred at zero abscissa. (f) Mean RFD values (± s.e.m.) along rescaled U-domains (all domains, empty circles; domains matching > 80% overlap) N-domains, full circles.

with 4,150 AS (5.2% of the genome) shared between the two cell lines (Fig. 5a, Supplementary Fig. 2a and Supplementary Table 1). The amplitude of the RFD shift (ARFD, see Methods section) across the segments ranged from 10 to 90%, indicating variable initiation efficiency (Supplementary Fig. 2b). As control, large (+)ARFD were detected around > 90% of the early replication timing peaks previously identified in HeLa and GM06990 (see Methods section), confirming them as bona fide initiation zones (Fig. 2b,c, red arrows, Supplementary Fig. 3).

The large size of most AS was striking, with only a few tens < 10 kb, suggesting that replication initiates mostly within broad zones and questioning whether site-specific origins exist in the human genome. First, we checked that the size estimate of AS was not biased by the presence of regions of lower read coverage at AS borders (Supplementary Fig. 4). Second, since HMM detection of AS required to compute RFD values within 15 kb sliding windows (Fig. 4), we checked whether this may overestimate the size of authentic site-specific origins. To begin with, we created in our profiles abrupt RFD shifts mimicking site-specific origins by deleting appropriate data points from AS. These were all detected by the HMM as 10–15 kb AS (Supplementary Fig. 5a,b). Next, we visually examined all detected AS < 15 kb from native HeLa and GM06990 RFD profiles (Supplementary Fig. 6 and Supplementary Table 2). Among the 1,297 examined AS from HeLa cells, only 13 clearly showed an abrupt RFD shift spanning at most 1 kb and only 53 spanned from 1 to 5 kb. Many others showed a smooth 5–15 kb AS. However, a few hundreds may be interpreted either as broad (5 to 15 kb) initiation zones with ample signal dispersion or as tandems of site-specific origins. Similar results were obtained for GM06990 cells (Supplementary Fig. 6 and Supplementary Table 2). We also investigated the ability of the HMM to detect isolated site-specific origins of low efficiencies. Detection required a ΔRFD > 20% (Supplementary Fig. 5c). Finally, we found that the extensive aneuploidy of the HeLa cancer cell line did not detectably affect the properties of initiation zones (Supplementary Fig. 7).

It has long been controversial whether replication initiates dispersively within broad zones or efficiently at narrowly localized sites in mammalian cells. Two-dimensional (2D)-gel and bubble trap studies identified almost exclusively broad initiation zones, where each probed restriction fragment showed a composite Y (single fork) + bubble pattern, with a complete bubble arc thicker at its lower end and thinner at its upper end. This typical bubble signal was clearly different from that generated by a bacterial or a viral site-specific origin. In reconstruction experiments, mixtures of bubbles generated from fixed origins located at different relative positions in the same fragment showed multiple, resolvable arcs even when origins were only 0.2–0.4 kb apart. Therefore, the continuity and thickness of mammalian bubble arcs could be only explained by the presence of bubbles generated at many positions (< 1 kb apart from each other). This conclusion was independently confirmed by studying the migration of intentionally broken bubble fragments, the size and origin of nascent strands from bubbles of increasing sizes, and the co-electrophoresis of replication intermediates of the same plasmid replicated in bacteria and in Xenopus egg extracts. The constant predominance of the Y signal implicated a low overall efficiency of initiation (~1% per kb). In contrast, SNS studies detected narrow peaks of initiation along the genome, suggesting highly efficient, site-specific origins. Even when multiple peaks could be clustered in zones, most remained separated by longer distances than the typical SNS size (1.5–2.0 kb), suggesting distinct, highly preferred sites within as well as outside the zones. The genome-wide overlap between the zones mapped by bubble-trap or obtained by merging multiple SNS peaks remained modest, even using generous merging (< 6 kb) and overlap (> 1 nt) criteria.

Using OK-Seq, we clearly identified no more than a few tens of isolated, > 20% efficient site-specific origins. In contrast, we identified hundreds of broad (> 15 kb) initiation zones. The quasi-linearity of the RFD shifts across most AS (Supplementary
values. This means that analysis of RFD profiles with an excessive proportion of AS not conserved between replicates, likely profiles strongly decreased for various sizes of than 0.05 (light blue). This proportion dropped down to 10% for l = 5 kb. For l = 5 kb, 40% of genome positions presented slopes differing by more than 0.05 (light blue). This proportion dropped down to 10% for l = 10 kb, and to 4% for l = 15 kb. This showed that the difference between the two profiles strongly decreased for l ≥ 15 kb. (c) AS of the rep 1 and rep 2 profiles were determined using the HMM procedure (see Methods section) for various sizes of l. The proportion of AS common to both replicate profiles (blue) or specific to one of the two profiles (red) was displayed as a function of l. The proportion of AS specific to only one replicate, that is, that would not be considered as bona fide AS, decreased from almost 50% for l = 5 kb to 20% for l = 15 kb and then decreased more slowly for larger l values. This means that analysis of RFD profiles with l < 15 kb would lead to an excessive proportion of AS not conserved between replicates, likely corresponding to too small signal/noise values.

Fig. 2c) suggested that each active initiation zone supported a single but randomly located initiation event, reminiscent of the Chinese hamster DHFR33,34 and mouse IgH35 initiation zones. This interpretation is in good agreement with 2D-gel34, DNA combing35,36 and bubble-trap4 studies. The progressive shift in Okazaki fragment strandedness observed over tens of kb at most human AS is in strong contrast to the abrupt shifts reported at S. cerevisiae origins27,28, which are defined by short specific sequences.

We caution that we would be unable to detect weak (ΔRFD < 20%) but site-specific origins, and that closely spaced tandems of site-specific origins would be detected as broad AS. If tandems of site-specific origins detected by SNS analysis were to account for most AS, however, we should detect a strong enrichment of SNS inside AS. As shown in a later section, bubble-trap origins were clearly more homogeneously enriched than SNS within AS. We, therefore, favour the alternative hypothesis that AS contain multiple inefficient sites spaced at < 1 kb intervals. The average slope of AS suggests an overall initiation rate of ~1% per kb (Supplementary Fig. 2b), consistent with 2D-gel studies. Although precise measurements of initiation rate fluctuations across individual AS will require further work, we expect that the dispersive nature of initiation observed across AS and bubbles would escape detection by SNS peak-searching algorithms. The lower overlap of SNS than bubbles with AS is further examined later in this manuscript, after characterization of AS location with respect to active genes.

Many replication initiation zones flank active genes. AS were predominantly intergenic and their ends bordered expressed genes more often than expected by chance (Fig. 5a-c); in HeLa cells, 19% of AS coincided (> 50%, 80% on average) with non-transcribed regions separating two active genes (type 1 AS), 27% were flanked (< 20 kb) by one active gene (type 2 AS) and 35% were not flanked (< 20 kb) by active genes (type 3 AS); many of the remaining 20% (type 4 AS) corresponded to type 1 or 2 AS with presumably mis-annotated transcripts (Fig. 5d, Supplementary Fig. 8). Similar results were observed for GM06990 cells (Supplementary Fig. 9d,e). Shared AS between the two cell lines were enriched in type 1 whereas cell-type-specific AS were enriched in type 3 (Supplementary Table 1). No particular positioning of inactive genes was detected for any AS type (Supplementary Fig. 10). Types 1 and 2 replicated almost exclusively in early S phase whereas type 3 replicated predominantly late (Supplementary Fig. 9d,e). Isolated genes expressed in only one cell type were flanked by AS only in that cell type (Fig. 5e,f). Initiation was routinely (70–90%) detected between active genes separated by > 20 kb, but less often within smaller intergenes; a preference for divergent over tandem over convergent gene pairs was observed, particularly in small intergenes (Fig. 5g). This preferential initiation upstream of active genes contributed to a significant co-orientation of replication with transcription (Fig. 5h), extending previous findings in skew N-domains37. These data show that many disperse initiation zones consist of large DNA segments precisely circumscribed by active genes. They are in agreement with early studies of a few model13,33,34,37 and with the recent key observation that transcription shapes a dramatic redistribution of the MCM2–7 complex, a core component of the replicative helicase marking potential origin sites, to exclude MCM2–7 from transcribed genes before S-phase entry in D. melanogaster38. However, this is probably not the single mechanism for specifying initiation zones. Using available global run-on sequencing (GRO-seq) and cap analysis of gene expression (CAGE) data39, we detected only a small amount of non-coding transcripts in the type 2 AS flanks that are devoid of active genes, and their level was not markedly higher than inside the AS; furthermore, we did not detect any significant level of non-coding transcripts in the non-genic flanks of type 3 AS (Supplementary Fig. 11). Therefore, initiation zones borders can also be delimited by non-transcriptional mechanisms.

Many replication termination zones overlap active genes. The HMM also detected 9,440 (HeLa) and 5,715 (GM06990) DS (8–2600 kb) covering 49.5% and 39.4% of the genome, respectively, with 5,000 (35% of the genome) shared between the two cell lines (Supplementary Figs 2a and 9f,g and Supplementary Table 1). In early S phase, DS predominantly overlapped transcribed genes, as expected from initiation predominance in intergenes; in late S phase, DS predominantly consisted of large non-expressed DNA regions (Fig. 5a,i, Supplementary Figs 1 and 9f,g). Many DS were directly contiguous to AS, implying that termination often occurs anywhere between adjacent AS.

Evidence for large unidirectional replication zones. The HMM detected 745 (HeLa) and 1,101 (GM06990) FS (20–1,200 kb).
Most of these (617 and 849, covering 4.1% and 7.3% of the genome, respectively, 235 shared segments) showed a strongly (>80%) favoured replication direction (RFD) > 0.6; Supplementary Fig. 2d; for an example see region 68.9–69.2 Mb of HeLa profile in Fig. 5f). These high-RFD FS bordered initiation zones, often started at an active transcription start site (TSS), and were enriched in expressed genes co-oriented with replication (Fig. 5j,k). These segments are likely devoid of origins and termini and are almost always replicated from their upstream initiation zone. They separate AS from DS so that the forks that traverse them terminate away, rather than at any distance, from their upstream initiation zone.

Most initiation sites are not associated with G4 or CGIs. Previous studies have reported an association of mammalian origins with TSSs and associated elements such as CpG islands (CGIs) and G-quadruplex (G4) elements. The degree of enrichment was variable, and concerns that it may arise from the failure of an-locunuclease to eliminate G-rich sequences in SNS-contaminating background DNA were raised. We found that CGIs were enriched at the borders of AS, especially AS shared by HeLa and GM06990, but not inside AS (Fig. 6a, Table 1). This reflected the frequent occurrence of CGIs around housekeeping gene TSSs. There was only a weak, if any, enrichment of G4s inside AS (Fig. 6a). Therefore, most...
Figure 6 | Comparison of initiation zones with previous origin mapping data. (a) Enrichment ratios (see Methods section) of CGIs (red) and G4s (black) around HeLa AS borders; pink box indicates position of half AS. (b) Enrichment ratios of bubbles (GM06990, blue), ORC1 (ref. 13) (HeLa, green) and SNS (HeLa, yellow) peaks around AS borders of the same cell type, for all peaks (left) or for peaks not overlapping CGIs (right). (c) Overlap (in Mb) of AS with bubbles, SNS and ORC1 peaks. (d) Exemplary profiles showing distribution of bubbles, CGIs, ORC1 and SNS peaks with respect to AS (GM06990, teal blue lines; HeLa, orange lines).

Table 1 | Number of common and cell-type-specific HeLa and GM06990 AS overlapping CGIs or not.

|                | Common   | HeLa specific | GM06990 specific |
|----------------|----------|---------------|------------------|
| All            | 4,186    | 5,650         | 1,573            |
| CGI –          | 1,710    | 3,416         | 1,039            |
| CGI +          | 2,476    | 2,234         | 534              |
| CGI%           | 59%      | 40%           | 34%              |

AS, ascending segment; CGI, CpG island.

Initiation sites within initiation zones are not associated with CGIs or G4s.

Initiation zones better align to bubbles than to SNS and ORC.

Two remarkable characteristics of AS were their sharp boundaries and their linearity, suggesting a fairly homogeneous internal rate of initiation falling off markedly at the borders. On average, short nascent strands (SNSs) and ORC1-binding sites were ~2 times more abundant within than outside AS from the same cell type (GM06990 for bubbles; HeLa for SNSs and ORC1 peaks; Fig. 6b–d), consistent with a higher efficiency of initiation within AS and a predominance of termination within DS. However, SNSs were not uniformly enriched along AS but peaked at their borders; these peaks disappeared after masking CGIs (Fig. 6b). ORC1-binding sites were also enriched at the borders but independently of CGIs (Fig. 6b) and of AS size (Supplementary Fig. 12). Similar results were obtained when TSSs from promoter regions that are both positive for ORC1 and enriched in SNS were analysed; the bimodal SNS enrichment pattern previously observed on both sides of such TSS/ORC1 sites was due to association of SNS with CGIs and not with ORC1 (Supplementary Fig. 13). Both 5′ and (to a lesser extent) 3′ gene ends contributed to ORC1 enrichment at AS borders (Supplementary Fig. 14). Globally, the density of ORC1 peaks was 2.4-fold higher in AS than in DS and FS. When the density of ORC1-binding sites was plotted along the segments delimited by consecutive AS, ORC1 enrichment peaked at the AS and declined in a distance-dependent manner in later-replicating sequences (Supplementary Fig. 15), consistent with the reported correlation of ORC1 density with replication timing13. Finally, trapped bubbles were enriched throughout the AS independently of CGIs (Fig. 6b,c), at all stages of S phase (Supplementary Fig. 16) and for all AS types (Supplementary Fig. 17).

The abundance of SNSs at CGIs suggested a high efficiency of initiation at these sites but this was neither reflected in bubble abundance nor by upshifts in RFD profiles. Only the trapped bubbles, and to a lesser extent the CGI-negative SNS, showed the homogeneous enrichment inside AS predicted by the linear slope of AS. Our data are consistent with the proposal that enrichment of SNS in CGI and other CG-rich sequences is due to an intrinsic bias of the λ-exonuclease technique1,3,40. Removing the CGI-associated SNS improved the homogeneity of SNS enrichment throughout AS but CGI-negative SNS were much less enriched than bubbles (Fig. 6b). The possibility that SNS peak detection algorithms overlooked a broad but uniform enrichment of SNS within AS was excluded by analysis of SNS raw data, which did not improve the enrichment inside AS (Supplementary Fig. 18). We, therefore, favour the hypothesis that bubbles are a more representative sample of human replication initiation sites than SNS. Together, the bubble-trap and OK-seq data favour the view that replication usually initiates within broad zones of multiple inefficient sites rather than at highly efficient specific sites.

The ORC1 signal was concentrated at AS borders rather than spread throughout AS as expected, if initiation events were to coincide strictly with ORC-binding sites. First, ORC has
other nuclear functions than replication initiation and some ORC-binding sites may not be origins. Second, metazoan ORC binds DNA \textit{in vitro} with no sequence specificity \cite{42-44} and the chromatin immunoprecipitation-sequencing (ChIP-seq) enrichment of ORC at specific chromatin sites is quite moderate in comparison to other chromatin-binding proteins \cite{45}. Indeed, the number of reads per ORC1 peak is relatively low in Dellino \textit{et al.} \cite{13}, which is the only reported genome-wide study of human ORC-binding sites. Clusters of ORC-binding sites might be difficult to detect if enrichment is too low compared with non-origin regions.

In addition, the modest overlap observed between AS- and ORC-binding sites could result from initiation away from ORC. Multiple observations suggest that human MCM2–7 are broadly distributed throughout the nucleus and exhibit little colocalization with ORC \cite{46,47} and that initiation events can occur at MCM2–7 complexes spread over a broad zone around ORC \cite{48-51} in metazoan cells. Since MCM2–7 interact with only one side of ORC during loading \cite{52}, ORC bound at gene ends may load multiple copies of the MCM2–7 complex only on its non-genic side. Alternatively, MCM2–7 may be dispersively loaded on both sides of ORC \cite{48-50}, but erased from transcribed regions in G1 phase \cite{37,38,53-54}, ultimately leading to the formation of broad intergenic initiation zones. In strong support of this model, MCM2–7 loaded in late G1 in \textit{Drosophila} cells show a bimodal distribution of enrichment along transcribed and non-transcribed sequences, due to their specific displacement from transcribed genes and their inability to be re-established or to translocate into transcribed genes after the G1/S transition \cite{38}. This model, therefore, explains why the overlap of ORC1 with initiation sites is only partial and why many AS abut transcribed genes.

Initiation zones are enriched in specific chromatin marks. As expected from the bordering of many AS by expressed genes, epigenetic marks associated with active TSSs (H3K4me2 and 3, H3K9ac, H3K27ac) and gene bodies (H3K79me2 and H3K36me2) were enriched at the boundaries and flanks of AS, respectively, but not inside them (Fig. 7a–c and Supplementary Fig. 19 for HeLa, Supplementary Fig. 20 for GM06990). H4K20me1, previously associated with SNSs \cite{11}, was only weakly enriched at active TSSs and not inside AS (Fig. 7a–c). In contrast, DNAse I hypersensitive sites (HSSs), H3K4me1 and histone variant H2A.Z were more abundant within than outside AS, irrespective of the presence or absence of flanking transcribed regions.
Figure 8 | Model for replication of N-shaped RFD domains. (a) When initiation zones are close enough from each other such as in gene-rich early-replicating regions, the temporal and spatial stochasticity of origin firing may cause emanating forks to converge at different positions spanning the entire intervening segment. This results in a straight N-shaped RFD profile. (b) Flat segments of high |RFD| are generated when initiation zones are distant and efficient enough to prevent the fork emanating from one initiation zone to reach the next one before it can fire, resulting in a ‘crenellated’ N-shaped RFD profile. This model implies that high |RFD| FS are devoid of origins. Such segments are enriched in active genes co-oriented with replication fork progression. (c) Large, straight N-shaped RFD profiles can form between very distant initiation zones if cascades of intervening background origin firing initiate at their borders and propagate to their center at an increasing rate during S phase. This results in a progressive inversion of RFD along the segment (see text).

TAD borders are enriched in initiation zones. It has been recently reported that the boundaries of cell-type-specific replication timing domains often coincide with the boundaries of insulated compartments of chromatin interaction, identified by high-resolution chromatin conformation capture (Hi-C) and dubbed ‘topologically associated domains’ (TADs). When the directionality of chromatin interactions and RFD were mapped between the midpoints of consecutive AS, both exhibited similar N-shaped profiles with opposite extrema at domain borders and null values at centres (Fig. 7g for HeLa, Supplementary Fig. 22a,b for GM06990). Remarkably, the borders but not the centres of TADs were strongly associated with AS (Fig. 7h for HeLa, Supplementary Fig. 22c for GM06990). These findings imply that replication of TADs often starts at their borders and propagates to their center.

Initiation events outside AS. The number of detected AS was significantly lower than all current estimates of the number of initiation events per haploid genome. Thus, initiation likely occurs outside AS but without creating visible RFD upshifts. We, therefore, propose that the human genome presents a superposition of specific, ‘master’ initiation zones (AS) and a background of more dispersed, less efficient origins (Fig. 8), in agreement with the widespread occurrence of inefficient origins detected by other methods outside AS (Fig. 6c). Thus, the predominance of termination over initiation within DS, which arises from the net inward flux of forks emanating from adjacent
AS, is only relative. In small DS, background initiation may be negligible, and the linear inversion of RFD across DS would simply reflect the population average of forks emanating from adjacent AS and terminating at different sites (Fig. 8a,b). In longer DS, where background initiation becomes significant (Fig. 8c), the linear inversion of RFD can be quantitatively explained by a ‘cascade’ model for sequential activation of internal origins with increasing synchrony from domain borders to center15,25,26. In this model, background origins close to domain borders fire in a sequential wave from borders to center, so that center-oriented forks travel longer distances than border-oriented forks before their merging (high [RFD]). Closer to domain centres, background origins fire later and more synchronously so that forks moving in both directions travel more similar distances (low [RFD]). The master initiation zones are often adjacent to, but excluded from, transcribed sequences and are associated with open chromatin marks even when they are not flanked by active genes. The N-shaped pattern of chromatin interaction directionality observed between consecutive master origins at TAD borders may underlie the sequential activation of intervening background origins resulting in N-shaped RFD profiles.

A recent genome-wide, single-molecule analysis of chromosomal DNA replication in S. pombe62 concluded that early origins are focused but later origins are more dispersive, that replication propagates from the earliest origins by firing additional origins in adjacent regions and that the rate of origin firing strongly increases as S phase progresses. These results strikingly converge with the cascade model proposed for origin activation in the human genome15,25,26 based on previous work in human cells and other eukaryotes18,63–65. In S. pombe, however, the earliest initiation events do not always appear at the same locations on the chromosomes, in contrast to the ‘master’ initiation zones observed at the borders of N/U-domains and in this work.

In summary, the novel RFD maps produced here confirm and extend previous views of chromosomal domain organization inferred from replication timing, nucleotide compositional skew and chromatin conformation data. They provide new insight into replication initiation and termination zones genome-wide and help to clarify controversies between previous origin-mapping studies. Replication initiation and termination sites have been mapped previously by sequencing Okazaki fragments in S. cerevisiae, but this necessitated the introduction of conditional lethal mutations in ligase and checkpoint genes to massively increase Okazaki fragment abundance27,28. The methodology presented here, which does not require ligase inactivation or any other cell modification, can be applied to a wide variety of organisms and cell types to study how replication patterns change during differentiation, tumorigenesis and evolution.

**Methods**

**EdU labelling.** Asynchronously growing 2–3 x 10⁶ HeLa (MRL2 strain, a gift from O. Bensaude, IBENS) or 10⁴ GM06990 cells (Corriell) were labelled by adding 20 µM EdU (Jenabioscience) to growth media for 2 min. Adherent HeLa cells were washed with ice-cold 1 x phosphate-buffered saline (PBS) and harvested by scrapping with a cell lifter. Non-adherent GM06990 cell suspensions were chilled in ice-cold bath water for 10 min and collected by centrifugation at 400 g for 10 min at 4 °C. As an alternative to EdU, 5-ethyl-2’-deoxyctydine (EdC) was also used with identical results.

**Genomic DNA preparation.** Genomic DNA was extracted by standard proteinase Kiphenol extraction followed by ethanol precipitation66, dissolved in 4 ml of 10 mM Tris-HCl, pH 8.0, 1 mM EDTA and stored at 4 °C.

**Okazaki fragment isolation and biotinylation.** Genomic DNA was denatured at 94 °C for 10 min, chilled on ice and size-fractionated on 5–30% (w/v) linear sucrose gradients in 10 mM Tris-HCl pH 8.0, 1 mM EDTA, 100 mM NaCl in a Beckman Ultracentrifuge SW28 rotor for 17 h at 26,000 g at 20 °C (ref. 67). Fractions containing < 200 b-strands (typically upper 8 ml) were concentrated on Amicon Ultra 10k centrifuging filter (Millipore), buffer exchanged twice with 5 ml of water and concentrated to 390 µl. Click reaction was performed in the dark in a final volume of 500 µl by sequential addition of 10 mM Tris-HCl pH 8.0, 2 mM CuSO₄, 1 mM biotin-TEG-Azide, 10 mM sodium ascorbate, incubation for 45 min at room temperature, followed by precipitation with ethanol and resuspension in 90 µl water. DNA was hydrolysed for 30 min at 37 °C by adding 10 µl of 2.5 M NaOH, followed with neutralization with 20 µl of 2.5 M acetic acid and purification on Micro Bio-Spin Columns with Bio-Gel P-30 (Bio-Rad). DNA 5’ ends were phosphorylated with 30 U of T4 Polynucleotide kinase (Fermentas) for 30 min at 37 °C, followed by enzyme inactivation for 10 min at 75 °C, ethanol precipitation and resuspension in 20 µl of water.

**Sequencing adapter ligation and Okazaki fragment capture.** Double-stranded adapters with single-stranded random hexanucleotide overlaps were generated by annealing equimolar amounts of complementary oligonucleotides in 50 mM NaCl in a thermal cycler by heating at 95 °C for 5 s and cooling down to 4 °C at 0.1 °C s⁻¹. The sequences of oligonucleotides pairs were as follows: A₁-F: 5’-CAACYCTTTCCCTACACGACGCCTCCTC-3’; A₁-R: 5’-NNNNNGGAAAAAGCTGGTATAAGGAAGTT-3’; A₂-F: 5’-[Phos]-AGATCGGAAAGGACACACTGCTGTAACCTGATCA [ddC]-3’; A₂-R: 5’-TGACTGGAAGTCAGAGTTGCTCTCCGATCTNNNNNNN [ddC]-3’. The A₁-F and A₂-R sequences were identical to the two strands of the Illumina TruSeq adapter except that A1 was truncated by five bases to reduce unwanted annealing with A2 during library preparation69. 5’-phosphorylated single-stranded DNA strands were annealed to 80 pM of each adapter in a thermal cycler by heating at 65 °C for 10 min and cooling down to 16 °C for 5 min. Ligation was performed at 16 °C overnight in 40 µl of ligase buffer added with 500 pmol biotin, 5% PEG8000 and four Weiss Units of T4 DNA Ligase (Fermentas). Biotinylated fragments were captured with 200 µg of Dynabeads MyOne Streptavidin T1 according to the manufacturer’s protocol (Life Biosciences). Unligated adapters were removed by washing five times with 5 mM Tris-HCl pH 7.5, 0.5 mM EDTA, 1 M NaCl, 0.05% Tween 20. DNA was denatured at 100 °C for 10 min, chilled on ice and size-fractionated on 5–30% (w/v) linear sucrose gradients in 10 mM Tris-HCl pH 8.0, 1 mM EDTA, 0.05% Tween 20 and once with water. Adapter annealing and ligation were then repeated using 20 µl of bead suspension and 40 µl of each adapter. Beads were washed as above and resuspended in 20 µl of 10 mM Tris-HCl pH 8.0.

**Library amplification and sequencing.** Libraries were amplified in 12 cycles of PCR using 10 µl of bead suspension as a template, Phusion polymerase (NEB) and the following primers. The forward primer was PE PCR Primer 1.0 truncated by 5’-ATGATAGCGGACCGACACGATCAGTCCTT TCCCTACACAGGCTCCTCC-3’ and the reverse primer was TruSeq Index primer: 5’-CAACYCTTTCCCTACACGACGCCTCCTC-3’ (‘indicates Illumina index). PCR products were purified on Qiagen MinElute columns and fragments from 150 to 400 bp were size-selected from 4% small fragment agarose gel (Eurogentec) using Qiagen gel extraction kit. Generation of clusters was performed on Illumina cBot following manufacturer's protocols except that a sequencing primer was replaced by 5’-CA CACTCTTCTCCCTACACAGGCTCCTCC-3’ (ref. 68). Sequencing was performed on HiSeq 1000 sequencer (single read 51 pb) according to standard procedures.

**Sequencing data.** All Illumina sequencing runs were performed at IMGIF high-throughput sequencing platform. Sequence reads were identified and demultiplexed using the standard Illumina software suite and adaptor sequences were removed by Cutadapt (version 1.2.1). Reads >10 nt were aligned to the human reference genome (hg19) using the BWA (version 0.7.5a) software with default parameters. We considered only uniquely mapping reads and removed PCR duplicates reads using Picard (http://picard.sourceforge.net).

**Replication fork directionality.** RFD was computed for each 1 kb window as follows:

\[ RFD = (C - W) / (C + W) \]  

where C and W correspond to the number of reads mapped on Crick and Watson strand, that reveal, respectively, the proportions of rightward- and leftward-moving forks within each window. Since the total amount of replication on both strands shall be constant across the genome, we derive the difference between two strands by the total read count to account for variation in read depth due to copy number, sequence bias and so on. RFD ranges from −1 (100% leftward moving forks) to +1 (100% rightward moving forks) and 0 means equal proportions of leftward- and rightward-moving forks. Repeat experiments produced RFD that strongly corroborate each other (Supplementary Fig. 1, for HeLa cells, R = 0.92 (Pearson), P < 10⁻¹⁵ (t-test) and for GM06990 cells, R = 0.93, P < 10⁻¹⁵). Similar correlations were observed between EdU and EdC profiles.
Segmentation of RFD profiles. A four-state HMM was used to detect within the RFD profiles the AS, DS and FS segments representing regions of predominant initiation ('Up' state), predominant termination ('Down' state) and constant RFD ('Flat1' and 'Flat2' states; see Supplementary Fig. 9a for a graphic representation). The RFD values were computed within 15 kb sliding windows (stepped by 1 kb across the autosomes). The HMM used the ARFD values between adjacent windows (that is, ARFD_{(n+1)} = ARFD_{(n)} / 2 for window n). Windows with <30 reads on one strand were masked. The ARFD values were divided in five quantiles and the HMM package of R (http://www.r-project.org/) was used to perform the HMM prediction with probabilities of transition and emission quantiles and the HMM package of R (http://www.r-project.org/) was used. The RFD values were computed within 15 kb sliding windows (stepped by 1 kb). To compare the enrichments, the density profiles were normalized by the average density within the examined region (Fig. 6a,b, Supplementary Figs 12–16). In Supplementary Fig. 13a–c, raw HeLa SNS-seq data was used to compute enrichments around TSS in adjacent 100 nt windows.

Histone and open chromatin marks. The data of signal enrichment for histone marks, including H2A.Z, H3K4me1, H3K4me2, H3K9ac, H3K36me3, H3K27ac, H3K27me3, H3K36me3, H3K97me2 and H4K20me1 were obtained from GM12878 (closely related to GM06990) cells were downloaded from the Board Histone track at the UCSC genome browser. The data for nucleosome positioning (MNase-seq) in GM12878 cells were downloaded from the Stanf NUosome track at the UCSC genome browser. The DNAse I-seq data (UW DNase I HS data track) and the DNase I hypersensitivity uniform peaks of HeLa and GM069090 cells were downloaded from ENCODE.

Hi-C chromatin interaction data. The Hi-C data of HeLa69 and GM06990 (ref. 60) cells were retrieved and the Directionality Index (DI) was computed as described:

\[
DI = \frac{(B - A)}{[\frac{(A - E)^2}{E} + \frac{(B - E)^2}{E}]} 
\]

where \(A\) and \(B\) are the number of reads that map from a given 40 kb bin to the upstream and downstream 2 Mb, respectively, and \(E\) the expected number of reads under the null hypothesis \((E = \frac{(A + B)}{2})\). The coordinates of the TADs of H1 human embryonic stem cells and IMR90 fibroblasts were retrieved from ref. 61.

References
1. Hyrien, O. Peaks cloaked in the mist: the landscape of mammalian replication origins. J. Cell Biol. 208, 147–160 (2015).
2. Gilbert, D. M. Evaluating genome-scale approaches to eukaryotic DNA replication. Nat. Rev. Genet. 11, 673–684 (2010).
3. Urban, J. M., Foulk, M. S., Casella, C. & Gerbi, S. A. The hunt for replication origins. Genome Res. 23, 1774–1788 (2013).
4. Mesner, L. D. et al. Bubble-seq analysis of the human genome reveals distinct chromatin-mediated mechanisms for regulating early- and late-firing origins. Genome Res. 23, 30 (2013).
5. Cadoret, J. C. et al. Genome-wide studies highlight indirect links between human replication origins and gene regulation. Proc. Natl Acad. Sci. USA 105, 15837–15842 (2008).
6. Sequeira-Mendes, J. et al. Transcription initiation activity sets replication origin efficiency in mammalian cells. PLoS Genet. 5, e1000446 (2009).
7. Cayrou, C. et al. Genome-scale analysis of metazoan replication origins reveals their organization in specific but flexible sites defined by conserved features. Genome Res. 21, 1438–1449 (2011).
8. Besnard, E. et al. Unraveling cell-type-specific and reprogrammable human replication origin signatures associated with G-quadruplex consensus motifs. Nat. Struct. Mol. Biol. 19, 837–844 (2012).
9. Martin, M. M. et al. Genome-wide depletion of replication initiation events in highly transcribed regions. Genome Res. 21, 1822–1832 (2011).
10. Picard, F. et al. The spatiotemporal program of DNA replication is associated with specific combinations of chromatin marks in human cells. PLoS Genet. 10, e1004282 (2014).
11. Karnani, N., Taylor, C. M., Malhotra, A. & Dutta, A. Genomic study of replication initiation in human chromosomes reveals the influence of transcription regulation and chromatin structure on origin selection. Mol. Biol. Cell 21, 393–404 (2010).
12. Dellino, G. I. et al. Genome-wide mapping of human DNA-replication origins: levels of transcription at ORC1 sites regulate origin selection and replication timing. Genome Res. 23, 1–11 (2013).
13. Lobry, J. R. Asymmetric substitutions patterns in the two DNA strands of bacteria. Mol. Biol. Evol. 13, 660–665 (1996).
14. Hyrien, O., et al. Simple bacterial and archaeal replications to replication N/U-domains. J. Mol. Biol. 425, 4673–4689 (2013).
15. Grigoriev, A. Analyzing genomes with cumulative skew diagrams. Nucleic Acids Res. 36, 2286–2290 (2008).
16. Brodie Of Brodie, E. B. et al. From DNA sequence analysis to modeling replication in the human genome. Phys. Rev. Lett. 94, 248103 (2005).
17. Touchon, M. et al. Replication-associated strand asymmetries in mammalian genomes: Toward detection of replication origins. Proc. Natl Acad. Sci. USA 28, 28 (2005).
18. Hlavat, M. et al. Human gene organization driven by the coordination of replication and transcription. Genome Res. 17, 1278–1285 (2007).
19. Woodfine, K. et al. Replication timing of human chromosome 6. Cell Cycle 4, 172–176 (2005).
21. Chen, C. L. et al. Impact of replication timing on non-CpG and CpG substitution rates in mammalian genomes. *Genome Res.* 20, 447–457 (2010).

22. Hansen, R. S. et al. Sequencing newly replicated DNA reveals widespread plasticity in human replication timing. *Proc. Natl. Acad. Sci. USA* 107, 139–144 (2010).

23. Audit, B. et al. DNA replication timing data corroborate in silico human replication origin predictions. *Phys. Rev. Lett.* 99, 248102 (2007).

24. Chen, C. L. et al. Replication-associated mutational asymmetry in the human genome. *Mol. Biol. Cell.* 28, 2327–2337 (2011).

25. Baker, A. et al. Replication fork polarity gradients revealed by megabase-sized U-shaped replication timing domains in human cell lines. *PLoS Comput. Biol.* 8, e1002443 (2012).

26. Guilbaud, G. et al. Evidence for sequential and increasing activation of replication origins along replication timing gradients in the human genome. *Mol. Biol. Cell.* 7, e1003232 (2011).

27. Smith, D. J. & Whitehouse, I. Intrinsic coupling of lagging-strand synthesis to chromatin assembly. *Nature* 483, 434–438 (2012).

28. McGuffee, S. R., Smith, D. J. & Whitehouse, I. Quantitative, genome-wide analysis of eukaryotic replication initiation and termination. *Mol. Cell.* 50, 123–135 (2013).

29. Audit, B. et al. Multiscale analysis of genome-wide replication timing profiles using a wavelet-based signal-processing algorithm. *Nat. Protoc.* 8, 98–110 (2013).

30. Schwartzman, J. B., Martínez-Robles, M. L. & Hernandez, P. The migration behaviour of DNA replicative intermediates containing an internal bubble analyzed by two-dimensional agarose gel electrophoresis. *Nucleic Acids Res.* 21, 5474–5479 (1993).

31. Kalejta, R. F., Lin, H. B., Dijkwel, P. A. & Hamlin, J. L. Characterizing replication intermediates in the amplified CHO dihydrofolate reductase domain by two novel gel electrophoretic techniques. *Mol. Cell. Biol.* 16, 4923–4931 (1996).

32. Santamaria, D. et al. Bi-directional replication and random termination. *Nucleic Acids Res.* 28, 2099–2107 (2000).

33. Dijkwel, P. A. & Hamlin, J. L. Initiation of DNA replication in the dihydrofolate reductase locus is confined to the early S phase in CHO cells synchronized with the plant amino acid mimose. *Mol. Cell. Biol.* 12, 3715–3722 (1992).

34. Wang, S., Dijkwel, P. A. & Hamlin, J. L. Laggard-strand, early-labelling, and two-dimensional gel assays suggest multiple potential initiation sites in the Chinese hamster dihydrofolate reductase origin. *Mol. Cell. Biol.* 18, 39–50 (1998).

35. Demczuk, A. et al. Regulation of DNA replication within the immunoglobulin heavy-chain locus during B cell commitment. *PLoS Biol.* 6, e1001360 (2008).

36. Lebofsky, R., Heilig, R., Sonnleitner, M., Weissbenz, I. & Bensimon, A. DNA replication origin interference increases the spacing between initiation events in human cells. *EMBO J.* 17, 4199–4207 (1998).

37. Hyrien, O., Marie, C. & Méchal, M. Transition in specification of embryonic metazoan DNA replication origins. *Science* 270, 994–997 (1999).

38. Powell, S. K. et al. Dynamic loading and redistribution of the Mcm2-7 helicase complex through the cell cycle. *EMBO J.* 34, 531–543 (2015).

39. Andersson, R. et al. Nuclear stability and transcriptional directionality separate functionally distinct RNA species. *Nat. Commun.* 5, 5336 (2014).

40. Foulk, M. S., Urban, J. M., Casella, C. & Gerbi, S. A. Characterizing and controlling intrinsic biases of lambda exonuclease in nascent strand sequencing by the human origin recognition complex. *Cell* 129, 399–409 (2007).

41. Hyrien, O., Marie, C. & Méchal, M. Transition in specification of embryonic metazoan DNA replication origins. *Science* 270, 994–997 (1999).

42. Powell, S. K. et al. Dynamic loading and redistribution of the Mcm2-7 helicase complex through the cell cycle. *EMBO J.* 34, 531–543 (2015).

43. Andersson, R. et al. Nuclear stability and transcriptional directionality separate functionally distinct RNA species. *Nat. Commun.* 5, 5336 (2014).

44. Foulk, M. S., Urban, J. M., Casella, C. & Gerbi, S. A. Characterizing and controlling intrinsic biases of lambda exonuclease in nascent strand sequencing by the human origin recognition complex. *Cell* 129, 399–409 (2007).

45. Schepers, A. & Papior, P. Why are we where we are? Understanding replication origins and initiation sites in eukaryotes using ChiP-approaches. *Chromosome Res.* 18, 63–77 (2010).

46. Krude, T., Musahl, C., Laskey, R. A. & Knippers, R. Human minichromosome maintenance proteins and human replication origin predictions. *EMBO J.* 23, 191–201 (2004).

47. Scheper, A. & Papior, P. Why are we where we are? Understanding replication origins and initiation sites in eukaryotes using ChiP-approaches. *Chromosome Res.* 18, 63–77 (2010).

48. Edwards, M. C. et al. MCM2-7 complexes bind chromatin in a distributed pattern surrounding the origin recognition complex in Xenopus egg extracts. *J. Biol. Chem.* 277, 33049–33057 (2002).

49. Harvey, K. J. & Newport, J. CpG methylation of DNA restricts prereplication complex assembly in Xenopus egg extracts. *Mol. Cell. Biol.* 23, 6769–6779 (2003).
Competing financial interests: The authors declare no competing financial interests.

Reprints and permission information is available online at http://npg.nature.com/reprintsandpermissions/

How to cite this article: Petryk, N. et al. Replication landscape of the human genome. Nat. Commun. 7:10208 doi: 10.1038/ncomms10208 (2016).

This work is licensed under a Creative Commons Attribution 4.0 International License. The images or other third party material in this article are included in the article’s Creative Commons license, unless indicated otherwise in the credit line; if the material is not included under the Creative Commons license, users will need to obtain permission from the license holder to reproduce the material. To view a copy of this license, visit http://creativecommons.org/licenses/by/4.0/