A Complete Set of Nascent Transcription Rates for Yeast Genes

Vicent Pelechano[1¤], Sebastián Chávez[2], José E. Pérez-Ortín[1]*

1 Departamento de Bioquímica y Biología Molecular, Facultad de Ciencias Biológicas, Universitat de València, Burjassot, Spain, 2 Departamento de Genética, Facultad de Biología, Universidad de Sevilla, Sevilla, Spain

Abstract

The amount of mRNA in a cell is the result of two opposite reactions: transcription and mRNA degradation. These reactions are governed by kinetics laws, and the most regulated step for many genes is the transcription rate. The transcription rate, which is assumed to be exercised mainly at the RNA polymerase recruitment level, can be calculated using the RNA polymerase densities determined either by run-on or immunoprecipation using specific antibodies. The yeast *Saccharomyces cerevisiae* is the ideal model organism to generate a complete set of nascent transcription rates that will prove useful for many gene regulation studies. By combining genomic data from both the GRO (Genomic Run-on) and the RNA pol ChIP-on-chip methods we generated a new, more accurate nascent transcription rate dataset. By comparing this dataset with the indirect ones obtained from the mRNA stabilities and mRNA amount datasets, we are able to obtain biological information about posttranscriptional regulation processes and a genomic snapshot of the location of the active transcriptional machinery. We have obtained nascent transcription rates for 4,670 yeast genes. The median RNA polymerase II density in the genes is 0.078 molecules/kb, which corresponds to an average of 0.096 molecules/gene. Most genes have transcription rates of between 2 and 30 mRNAs/hour and less than 1% of yeast genes have >1 RNA polymerase molecule/gene. Histone and ribosomal protein genes are the highest transcribed groups of genes and other than these exceptions the transcription of genes is an infrequent phenomenon in a yeast cell.

Introduction

Gene transcription in eukaryotes is a complex process that starts with the recruitment of an RNA polymerase (RNA pol) complex to the gene promoter and is followed by a set of successive steps, such as initiation, elongation, splicing, termination, mRNA export, and degradation. Although it is well known that all of these steps are subject to strict regulation [1] the main objective of most regulatory studies is just the determination of the mRNA amount (RA) without being able to discriminate which steps are actually being regulated. RA can be easily measured by northern and RT-PCR techniques. Moreover, with the emergence of genomic techniques thousands of mRNAs can be simultaneously evaluated at the same time by DNA chip techniques [2] or by other more quantitative methods [3,4]. However, the RA is the result of two opposite reactions, transcription and mRNA degradation, that can be characterized by chemical kinetic rates (the transcription rate, or TR, and the degradation rate) [5]. The main regulatory step for the gene expression of many genes is the control of their TR, which is assumed to be exercised mainly at the RNA pol recruitment level. Thus, variation in the mRNA level is usually attributed to changes in RNA pol recruitment to the promoter,

and it is used to construct models in which transcription factors, nucleosome and histone modifications, among others, are the main players in the gene regulation game. However, as the regulation at the mRNA stability level is increasingly recognized to be important in gene regulation [5–8], the mRNA measurement can no longer be used as a direct estimation of gene transcription. Therefore, the existence of a complete set of TRs for a given organism would be of enormous interest for many researchers. TR can be mathematically calculated from RA and mRNA stability assuming steady-state conditions for gene expression [5]. In fact, the use of this kind of TR dataset has become very popular for yeast since Holstege *et al.* [9] provided a set of TR data as a supplementary material of that paper. Those data represent, however, the indirect calculation of the rate of appearance of mature mRNAs in the cytoplasm, taking into account all possible posttranscriptional processes of the mRNA, and do not represent the actual synthesis of new mRNAs by RNA pol in the genes (i.e. nascent TR).

We [10] and others [11,12] have developed genomic variants of the well-known run-on technique [13] to evaluate the nascent TR for most genes. In this technique (GRO, Genomic Run-on), elongating RNA pol molecules, that conserve the RNA, are forced

to incorporate radioactive UTP for a short length. The macroarray analysis of the *in vivo* labeled RNA measures the density of RNA polymerases in the analyzed genes that can be converted into TRs for all the yeast genes [10]. Like all experimental measurements, GRO is affected by an unavoidable precision error (random) and, potentially, by technical or biological biases (not random). Therefore, in order to improve the TR data obtained from GRO experiments, we have reduced the random error by increasing the number of biological repeats. Moreover, to decrease technical specific biases, we have used data from chromatin immunoprecipitation assay (ChIP) of RNA pol II inside the genes with specific antibodies (RNA Pol-ChIP-on-chip, RPCC) to detect and correct technical biases specifically associated to the GRO data and not present in the RPCC data. We also have incorporated the new estimations available for RA and stability and taken into account the dilution effect on the mRNA concentration due to the continuous increase of the total cellular volume during the exponential growth. All this has allowed us to obtain a reliable complete dataset for all the yeast gene nascent transcription rates for the first time in an eukaryote. We analyze this dataset and discover that histone genes are the most highly transcribed whereas most of yeast genes are scarcely transcribed. In fact, only 14% of them have an active RNA pol II molecules at a given moment and only a small proportion of RNA pol II molecules (6%) are actually productively transcribing the about 5900 ORF-containing *S. cerevisiae* genes.

Results and Discussion

Estimation of transcription rates using improved and corrected GRO values

When we first described the GRO method [10], we calculated the TRs for 5886 yeast genes, these being the vast majority of the RNA pol II genes. In that case, the values obtained for the TR from three individual samples of those cells exponentially growing in YPD medium in arbitrary units were converted into real ones (molecules/min) using the single published TR value of the *HIS3* gene [14]. That was the very first description of "nascent" TR dataset for an organism. The data we currently hold have been improved in five ways.

First, the current data are the averaged calculations from 24 individual samples exponentially growing in YPD to DO$_{600}$ of 0.45. Second, they have been corrected for a probe-length artifact occurring in GRO data (supplementary Figure S1). This artifact was discovered when comparing individual data between GRO and RPCC. Both showed a bias with regard to gene length [15]. The TR was seen to decrease slightly with the gene length in the RPCC, while a stronger tendency was observed in the GRO. We found that this strong dependence in length of GRO is due to a bias that increases the labeling towards the 3′ end of the transcript and we hypothesize that it is due to the downstream movement of RNA polymerases (see supplementary text S1 and supplemental Figure S1 for a detailed description). Consequently we have used the RPCC data, which do not present this bias, to correct it.

Third, we used an updated total amount of mRNA molecules per yeast cell that has been corrected from that previously used (15000 molecules from ref. [16]) to more recent and precise calculations obtained from massive parallel sequencing and competitive PCR transcriptome measurements: 26000 molecules from refs. [3,4]. Fourth, we used a more robust reference to convert TR into real units. We reasoned that because the *HIS3* TR was calculated mathematically by dividing the RA by mRNA stability [14], we could do a similar calculation for many more yeast genes if we used published data [3,4,17]. Therefore, we used

the 2000 genes with the highest (and, therefore, the most confident) indirect TR value (corresponding, thus, to an "increase of mature mRNA TR" dataset, see below) to plot them against the averaged and corrected nascent TR dataset in arbitrary units in order to convert them into real units (molecules/min).

Finally, we have taken into account the effect of the dilution of the mRNA concentration due to the cell division. The dilution has a negative effect on the mRNA concentration and it is usually overlooked on the kinetic analysis of the transcription. However, as the volume of the cellular population is continuously increasing this effect, as well as the mRNA degradation itself, has to be compensated by the TR. In the standard mRNA stability calculations this factor is not taken into account because the cells' growth is inhibited by the treatment used to stop the transcription [6,17]. For computing an indirect transcription rate it is necessary to use the mRNA amount (RA) and the rate of decrease of mRNA concentration. This decrease rate is due to both the mRNA degradation and to the dilution effect. The decrease in the RA due to the cellular dilution is directly proportional to the RA of each individual gene and inversely proportional to the generation time [RA*(ln2/generation time)]. To incorporate this factor we modified the indirect TR data (that were computed with non-growing cells) by adding this factor. Afterwards, we used the modified indirect TR dataset to convert the nascent TR data into real units. In this way, we have obtained an improved, highly confident, TR dataset for 4670 yeast genes in standard growing conditions (Table S1).

Advantages of the nascent TR dataset

Until now many researchers have used an alternative TR dataset based on indirect calculations. In 1998 Holstege *et al.* [9] used an RA data set divided by "apparent half-life" data to obtain an apparent "transcriptional frequency", offered in the supplementary information at the authors' web site. This "apparent half-life" was obtained from the slope of decay curves that had been adjusted with just two experimental time points. Despite the uncertainty associated with indirect calculations and the use of poor half-life data, those data are still the most widely used in the literature as a source of yeast TR, probably because they positively correlate with the expected enrichment of the transcribed genes in some histone modifications or in active gene-associated proteins [18–20]. The advantage of this indirect approach to calculate TR is that it requires no experimental method and that it does not need to assume a constant RNA pol speed on genes to convert RNA pol II densities into TRs. However, it does present other possible drawbacks: 1) it relies on the hypothetical assumption of steady-state conditions, 2) the mathematical calculation of the data could increase the experimental error and, 3) the need to measure the mRNA disappearance during a prolonged time provokes a stressful situation that affects genes expression. Moreover, as the mRNA stabilities are computed in cells that are not actually growing, the dilution rate of the cellular content due to the cell growth, that is compensated by increasing the actual TR, is not taken into account. This indirect TR set, therefore, only accounts for the part of the TR that is devoted to compensate mRNA degradation. We have discarded the first potential drawback verifying the steady-state hypothesis during exponential growth in glucose [21]. However, the other error sources are unavoidable. As the indirect TR data are based on the measurement of mature mRNA disappearance, and thus it takes into account also all the posttranscriptional processes as well as the RNA degradation, the direct nascent TR data could be a better indicator of the effects of RNA pol II transcription on the *in vivo* gene chromatin structure features, such as histone modifications, transcription factor binding

or nucleosome positioning [18–20,22–27]. We have verified (see Figure 1) that our nascent TR dataset is better correlated with chromatin marks [25] directly related with transcription elongation, such as H3K36me3 that is deposited during RNA polymerase elongation [26] or with the presence of histone acetyltransferases related to elongation such as Gcn5p or Esa1p [27]. And it is, at least, as equally well correlated to chromatin modifications associated with active promoters [25] as the indirect TR data (see supplemental Figure S2). For these comparisons we have also used the indirect TR provided in [9] as well as a new set of indirect TR that we computed using more accurate and recent mRNA stability [17] and RA [3,4] (see details in [Methods] section). Moreover, we have repeated the same analyses using just the RA instead of TR. It can be seen that for transcription elongation markers indirect TR is slightly better correlated than RA and that the improvement caused by the use of nascent TR is much higher (Figure 1). With regard to non-elongating marks the results for all the datasets are much more similar (Figure S2). These comparisons demonstrate that the actual presence of active RNA pol II on genes, quantified by the nascent TR dataset, is the

best predictor of elongating-related chromatin features. This fact has been also confirmed by the direct comparison between the different TR datasets with RNA pol II occupancy (Figure 2): nascent TR dataset is also better correlated with the actual presence of RNA pol II, as determined by chip-on-chip analysis of the Rpb3 subunit [18].

It should be noted that all the nascent TR evaluations based on active RNA pol II density measure could differ from the increase of mature mRNA (Figure S3) owing to the fact that either some of the elongating polymerase molecules never reach the end of the gene or some of the nuclear pre-mRNAs are degraded [28]. This can differentially affect yeast genes, thus introducing a source of discrepancy with regard to those indirect TR calculations. In fact, it would be possible to use this discrepancy to study the different way in which the genes are expressed. To prove that our direct nascent TR data measures directly the transcription process, and thus differs from the indirect estimation of increase of mature mRNA concentration we compared both TR datasets with regard to the transcription features in which discrepancies are expected. First, our nascent
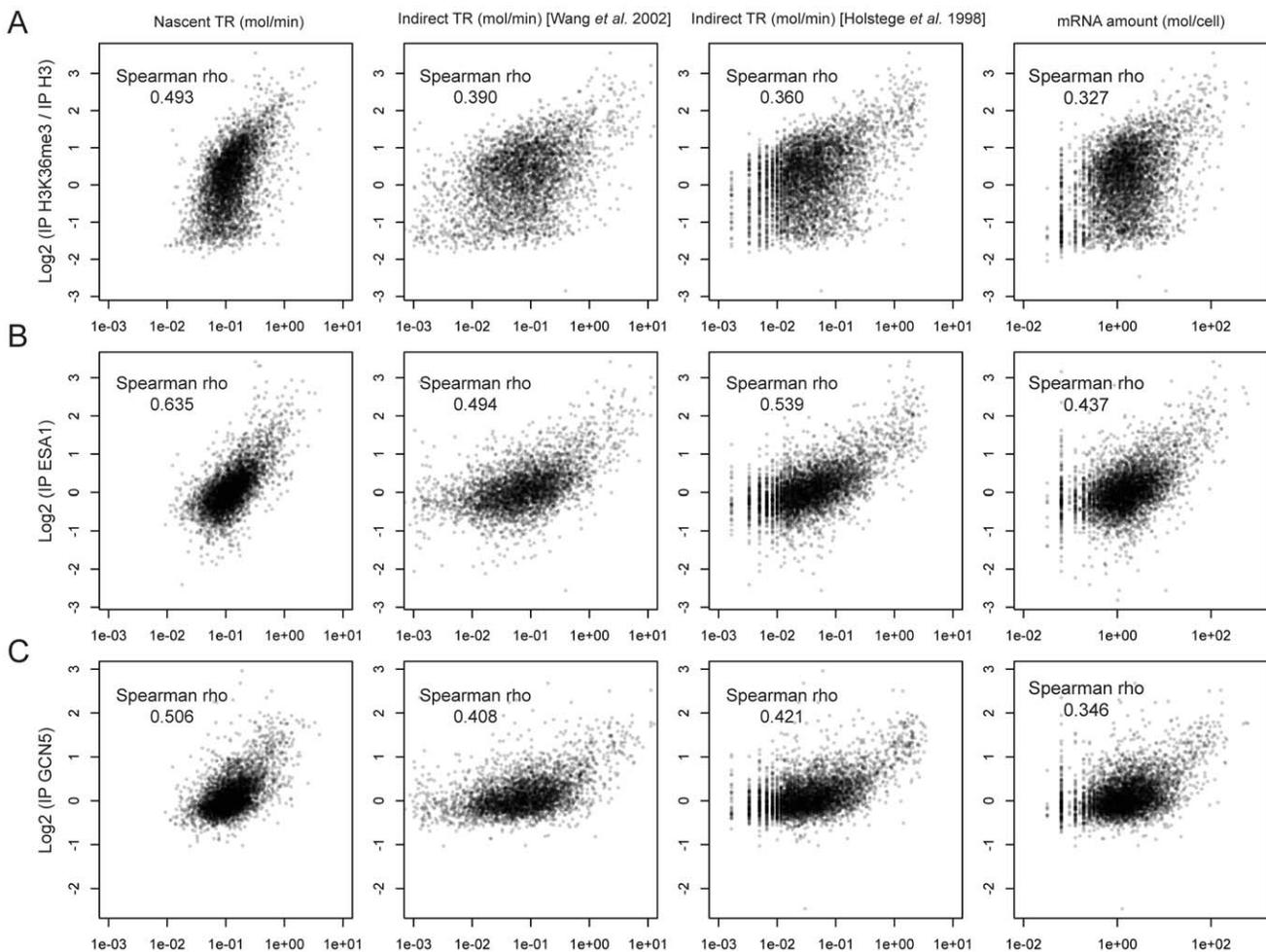
Figure 1. Comparison of the direct and indirect TR datasets with modifications related to elongation activity. Comparison of the new nascent TR dataset (this study) with a newly computed indirect TR datasets using the mRNA amount and the mRNA stability [17] that takes into account the effect of the dilution (see discussion), the indirect TR dataset from [9], and mRNA amount with different chromatin related parameters associated with active elongation as H3 trimethylation (H3K36me3, A) or presence of elongation related transcription factors the histone acetyltranferases Esa1p [27] (B) or Gcn5p (C) [25]. To avoid any bias depending on the wideness of the datasets Spearman rank correlation is used. Note that the correlations are always higher for the nascent TR dataset than for the indirect datasets, and that the indirect TR datasets show a better correlation than the mRNA amount data set.
doi:10.1371/journal.pone.0015442.g001
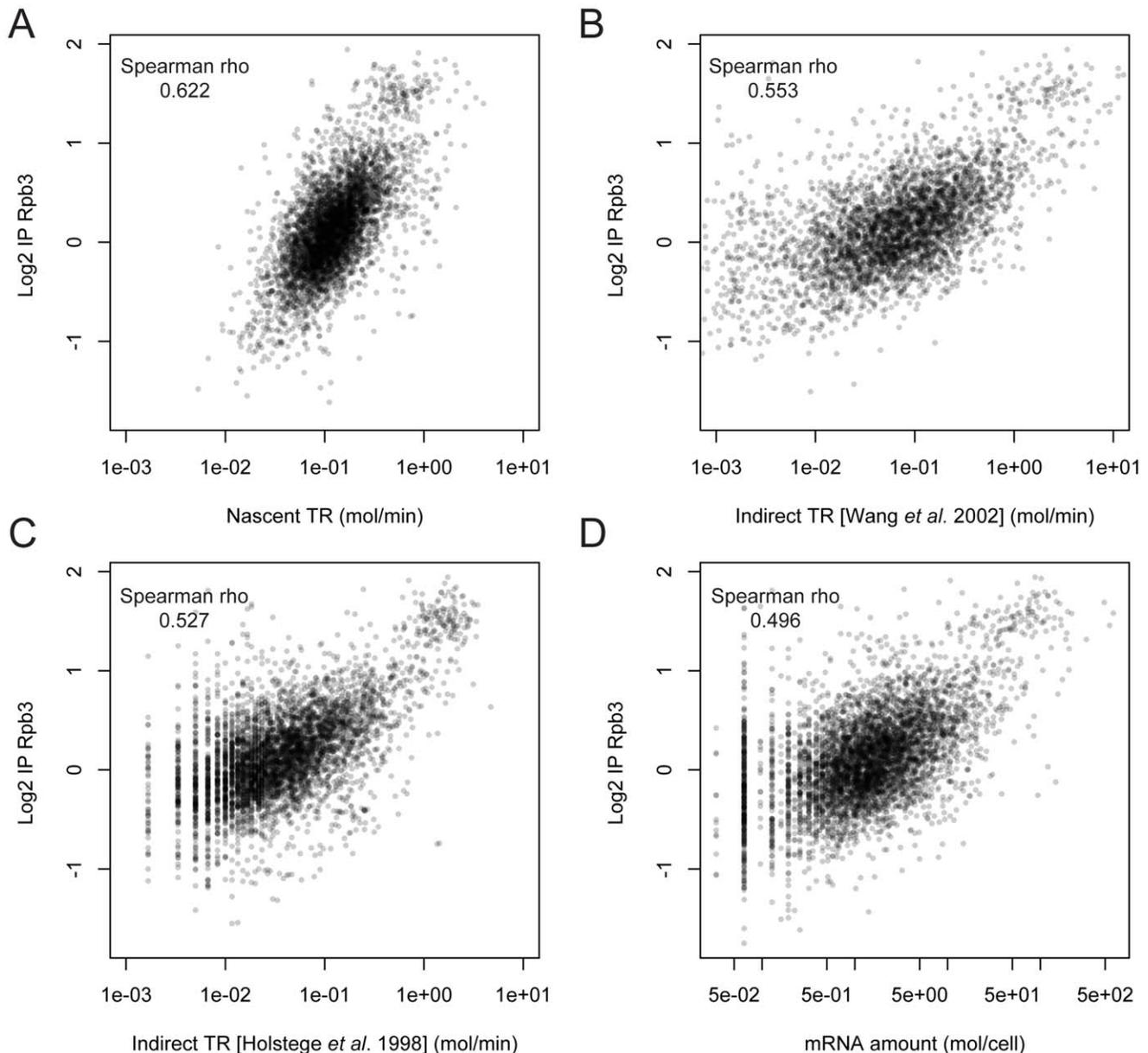
Figure 2. Comparison of the direct and indirect TR datasets with presence of RNA pol II. Comparison of the new direct TR dataset (this study) and indirect TR datasets computed using the mRNA amount and the mRNA stability [9,17] as in Fig. 1 with RNA pol II data (average IP along the coding region of Rpb3-TAP) [18]. To avoid any bias depending on the wideness of the datasets Spearman rank correlation is used. Note that the correlations are always higher for the nascent TR dataset (A) than for the indirect datasets (B, C), and that the indirect TR datasets show a better correlation than the mRNA amount data set (D).
doi:10.1371/journal.pone.0015442.g002

transcription rates have not strand-specificity, comprising the transcription for both the sense and any hypothetical antisense or cryptic transcription. As most of the antisense and cryptic transcripts are degraded in the nucleus it is expected that the genes with known antisense transcription (CUTs or SUTs) [29] present a higher nascent TR over their indirect TR. In fact the relationship between nascent and indirect transcription is significant higher for those genes (Figure S4). This confirms that the nascent TR is able to detect the transcription process independently of the stability of the produced transcripts. Allowing thus to monitor directly the transcription process itself without taking into account any posttranscriptional process.

Second, Venters and Pugh [24] have recently proposed that some yeast genes (1070 genes called "Group 2") might be regulated at the level of elongation because they have enriched levels of RNA pol II molecules on their $5'$ end transcribed region. We analyzed the behavior of the 652 genes of this paused RNA pol II class present in our nascent TR dataset (Table S1 and Figure S5) and, as expected and according to their proposal, we found a significant excess of nascent TR over the indirect TR only in these genes (comparing the $\log_2$ lowess corrected values (see [Methods] section), t-test with sig. $4.71 \cdot 10^{-4}$). In opposition, the 561 genes of group 1 (characterized by accumulation of non-elongating polymerases in the promoter region) showed less nascent TR over

the indirect TR (sig. $8.24 \cdot 10^{-7}$). We do not know the reason of this but it could be explained if these genes have less cryptic transcription or a more efficient maturation and export process. Finally, there is no significant difference (sig. 0.93) between nascent and indirect TR for the 560 genes of group 3 that has an even or enriched to the 3′ distribution of polymerases.

Analyses of the nascent TR dataset

The distribution of the TR values (Figure 3) is similar to a log-normal as shown in most cases by the expression datasets, being the median TR about 0.12 mRNA molecules/min (equivalent to 7 mRNAs/hour). Furthermore, 90% of the genes have TRs between 2.33 and 29.7 mRNAs/hour and, if we assume that the 4670 genes for which we have data are representative of the 5796 non-dubious ORFs, the total transcription for RNA pol II in a yeast cell growing in standard conditions is about 60200 mRNAs/h. By assuming the known datum that RNA polymerase molecules transcribe at 25 nt/s [30], then the median RNA pol II density inside the genes is 0.078 molecules/kb.

Gene functional groups tend to have similar TRs (Fig. 3B first graph), which reflects that coordinated transcription is one of the main steps in coordinating gene expression. However, some of GO categories, e.g. the hexose catabolism genes, present a wider range of TRs. This can reflect that the coordination of the biological function is less demanding at the transcriptional level, that there is strong control in a later expression step or, simply, that the GO category is less strictly defined. For most groups, the average position within the whole population in RA and TR is similar (Fig. 3B, compare first and third graphs). An exception to this is the group of histone genes. Their relatively high RA value is the result of the highest average TR level that is compensated by quite unstable mRNAs (Fig. 3B fourth graph). This group follows an expression strategy with very high TR and very low mRNA stability [7] that allows for fast changes in RA [5]. Histone genes are, thus, among the highest transcribed genes with a median TR of almost 56 mRNAs/hour. Moreover, because histone genes are only transcribed during the S phase, which represents only 27% of the cell cycle in exponentially growing yeast [31], the TR for actual active histone genes on an asynchronic cell population will be 206 mRNAs/hour that corresponds to a RNA pol II density of 2.29 molecules/kb. This is still 3 times lower than the RNA pol I density on rDNA active genes [32] and 4 times lower than the theoretical maximal density calculated just by using the physical size of the elongating RNA pol II and its maximal rate of initiation [33]. On the other hand, it can be seen also in Fig. 3 that sporulation is among the groups with the lowest TRs, coinciding with a very low RNA pol II density, as determined by Venters and Pugh [24]. Other group with very low median TR is amino acid transport (Figure 3B and Supplementary Figure S6). All these results are logical for cells exponentially growing in complete YPD medium.

Our results differ, in some aspects, from a previous study based in "mature mRNA appearance" by Struhl [33]. In that study the most transcribed genes in exponential growth in glucose were the ribosomal protein (RP) genes. However, in our exponential growth YPD nascent TR dataset RP genes show a median TR of 32.7 mRNAs/hour (Figures 3 and 4), six times lower than the histone genes. In fact, summing up all RP genes their transcription represents 8.5% of total RNA pol II transcription, which is much lower than previous estimations (50%) based on indirect calculations [34]. This is, in part, due to the fact that they exceed an amount of non-active RNA polymerases that reduces their effective TR (see ref. [35] for discussion). It was also discussed that GAL genes and stress-responsive genes have the highest TRs,

obviously in growth conditions other than the exponential growth in YPD [33]. However, using our previously published data, we have computed the TR level of GAL genes in exponential growth in galactose [10] and of stress-responsive genes after oxidative [36] or osmotic stress [37] and we have found that GAL genes in YPGal medium have a median TR of about 10.5 molecules/hr and that stress-responsive genes after stress reach a median TR of 124 molecules/hr, less than the one of the histone genes in exponential growth in YPD (Figure 4A and Supplementary Table S2).

Respective contributions of TR for compensating mRNA degradations rates and dilution rate

If we assume a steady-state situation for the exponential growth in YPD (demonstrated in [21]) the productive transcription rate should compensate the disappearance of mRNA occurring by mRNA degradation and by the dilution caused by continuous growth. Therefore, knowing the growth rate, it is possible to estimate the part of the TR devoted to compensate the mRNA degradation and the part devoted to compensate the dilution. In our data, obtained from exponentially growing cells in YPD with a doubling time of 113 minutes, although the dilution effect on the decrease of the mRNA concentration is much lesser than the mRNA decrease due to the mRNA degradation (only 8% on average), its contribution is not negligible and not identical for every gene. In fact, for some genes the dilution is the main factor for the mRNA disappearance and most of the transcriptional activity is devoted to compensate it (Supplementary Table S3). As the decrease due to the dilution is directly proportional to the mRNA amount for every gene this effect is more important for "housekeeping" genes that maintain a high RA and are relatively stable. The GO categories related with translation and biosynthetic processes are enriched in the group of genes in which TR devoted to compensate the dilution effect is more important (Supplementary Fig. S7). On the other hand, in the mRNAs expressed at low level and with low stability most of its transcription is devoted to compensate the mRNA degradation, and the effect of the dilution is almost negligible. This group is enriched in GO categories related with replication, transcription and mitosis (see Supplementary Figure S7).

Transcription and chromatin

It is known that nucleosome stability and distribution are affected during transcription elongation [38], and it has been demonstrated that active genes could be partially depleted of H2A–H2B dimers and H3–H4 tetramers [39]. In fact, the active copies of rDNA gene are totally depleted of nucleosomes because of the steric hindrance caused by high density of RNA pol I molecules [40]. Our RNA pol II density calculations (see Fig. 3) predict that even in the case of the highest transcribed RNA pol II genes, theoretically, nucleosomes have the time and room required to re-form, at least, partially. That is, a RNA pol II complex will transcribe in histone genes during their maximum transcription (206 mRNAs/hour) every 17.5 seconds through the region occupied by a nucleosome. With a RNA pol II speed of 25 bp/s it takes around 6 seconds to pass through 147 bp. Thus the nucleosomes could be associated with the DNA for 65% of the time. This will cause a nucleosome density more than three times higher than that shown for fully activated GAL genes [39]. In that case, however, the authors demonstrated that there is a significant reduction in histone-DNA contact when the genes are activated [39], and this has been genomically extended to other active genes [41]. Therefore, it can been concluded that the effect of RNA pol II passage on chromatin extends for longer periods than the actual time required to read the nucleosome-wrapped DNA sequence
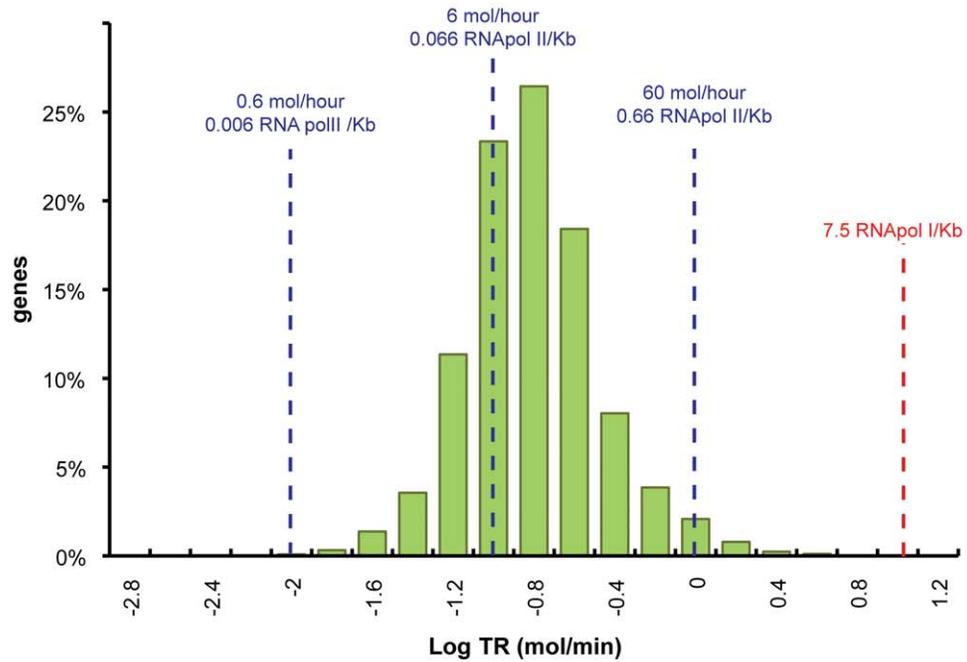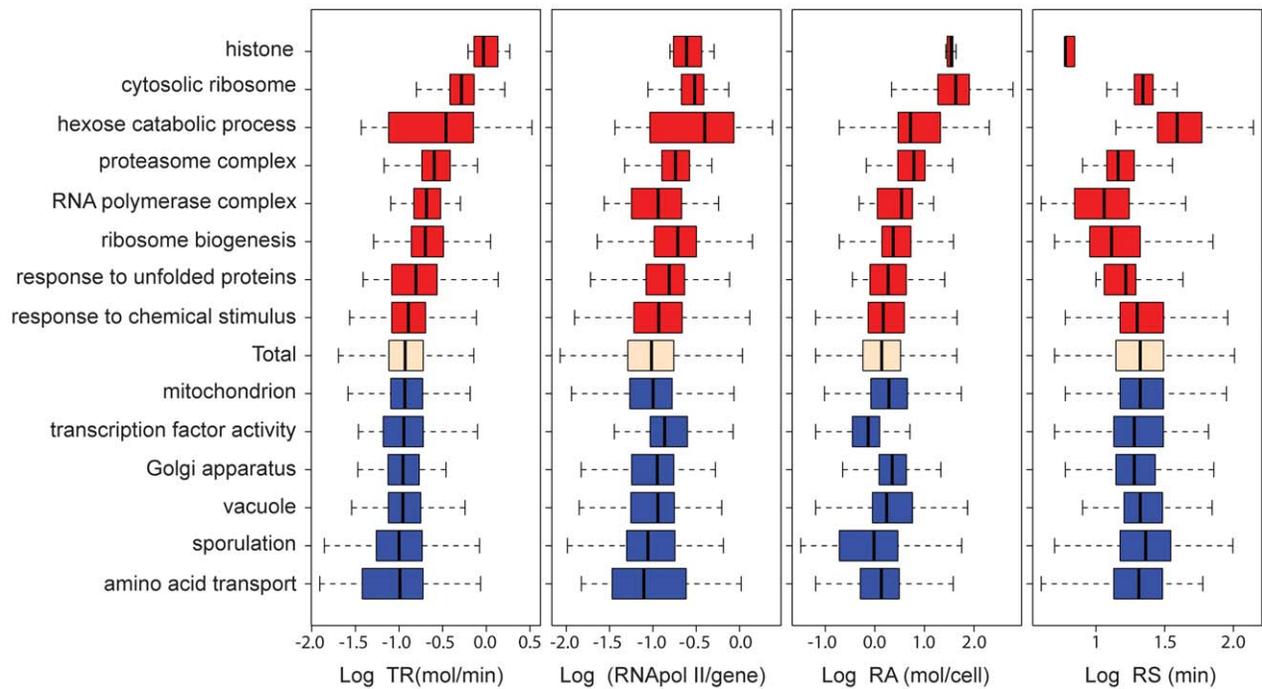
Figure 3. Nascent TR data set. A) Histogram of the yeast nascent productive TR data set in real units (molecules per hour, mol/hour) and RNA pol II densities (RNA pol II molecules per kb, mol/kb) for all the genes. The RNA pol I density [32] is shown as a reference with a dashed red line. The natural values of RNA pol II density and TR corresponding to −2, −1 and 0 units in log scale are shown as dashed blue lines. B) Comparison of nascent TR, RNA pol II molecules per gene, RNA amount (RA, using RNA-Seq data corrected with GATC-PCR, see [Methods]) and mRNA stability from ref. [17] for the different selected groups of genes (GO categories). The box plots represent the median and quartiles of the data. The whiskers show the maximum and minimum of the data set (excluding the outliers which lie beyond 1.5 times the inter quartile range).
doi:10.1371/journal.pone.0015442.g003

[39], which confirms previous suggestions that, in highly transcribed RNA pol II genes, nucleosome organization and positioning are extensively altered [42] with the help of chromatin remodeling complexes [38], but without an important contribution

of RNA polymerase steric hindrance. Because our results show that RNA pol II molecules are not so closely spaced, even in the highest transcribed genes, to sterically prevent DNA-octamer re-association behind the transcribing first RNA pol II molecule, the model that

A)

median nacent TR (mol/hour)	Glucose (this study)	Galactose [a]	Oxidative stress [b]	Osmotic stress [c]
Total	7.02	3.24	6.84	10.5
Histone [d]	55.68	23.34	30.12	91.02
Ribosomal Proteins	32.7	10.5	18.3	9.24
GAL genes	3.72	9.06	4.02	5.28
HSP genes	40.74	9.78	124.08	102.48

Precentage of RNA pol II engaged in productive transcription

	100%	n.a. [e]	73%	144%
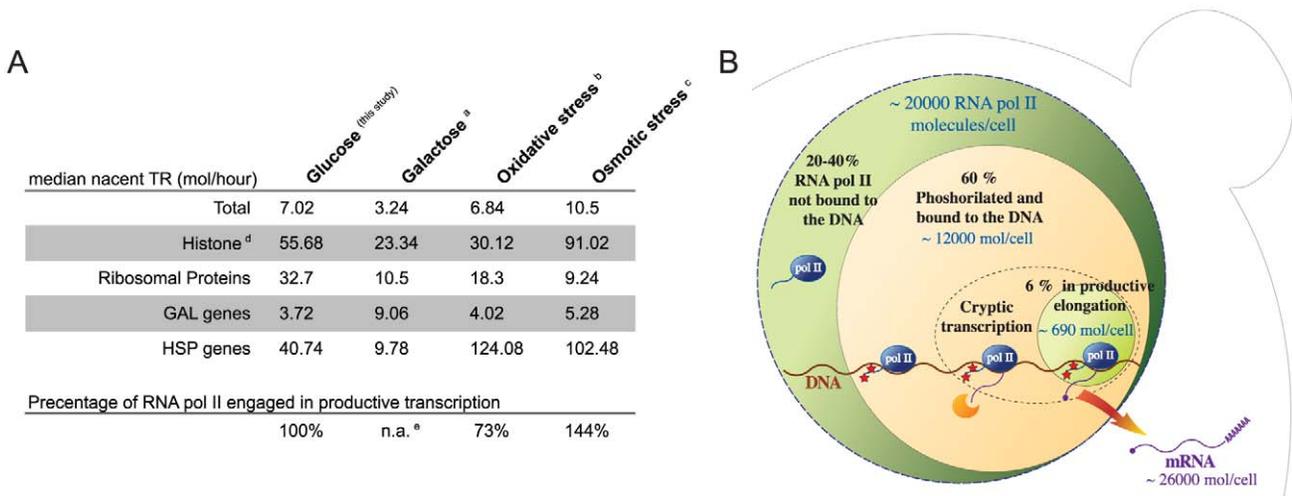
Figure 4. A snapshot of the yeast transcriptome. A) Selected data from the yeast nascent TR dataset. It is shown the median values for the TR for some selected groups of genes in exponential growth in YPD (glucose, this study) as well as in different conditions using previously published GRO datasets normalized with the newly computed nascent data: [a] cells growing exponentially in YPGal, (galactose: time 5 from [10]); [b] 16 min after the oxidative stress due to the addition of 0.1 mM t-BOOH [36] [c] 10 min in 0.4 M NaCl [37]. It is also shown the percentage of RNA pol II engaged in productive transcription (total transcription activity) relative to the exponential growth in glucose. [e] In the case of yeast cells growing in galactose this percentage is not significant because it is not possible to assume that the total amount of RNA pol II has not changed after several cell cycles using galactose as carbon source. On the other hand, we assume that the total amount of RNA pol II has not change during the short stresses. B) Schema of the yeast nucleus showing the computed distribution of transcriptional machinery during the exponential growing in rich medium. The size of the respective circles reflects the amount of RNA pol II molecules in each category.
doi:10.1371/journal.pone.0015442.g004

better fits the observations is that the next RNA pol II molecule could approach a partially remodeled nucleosome and displace it easily because of its lowered stability [38].

A possible error involved in the RNA pol II density calculated from the TR data is that the average speed for RNA pol II molecules is, perhaps, higher than 20–30 pb/s [43]. In fact, it has been recently shown that in human genes the average speed for many genes is about 63 pb/s [44]. Although, it has to be considered that this data was calculated at 37°C instead of the 28–30°C commonly used for yeast experiments. If the RNA pol II speed were higher the calculated densities for RNA pol II molecules would be even lower. Alternatively, if the number of mRNA molecules per cell were higher than 26000 as has also been suggested [43] the densities of the RNA pol II molecules would be higher. Therefore, we keep our calculations as they are in Table S1.

A snapshot of the yeast cell transcription

It was said that transcription occurs multiple times in most yeast genes within a single cell cycle [33]. This opinion sees the glass half full. However, although our current results confirm this view, we see the glass half empty because most genes are productively transcribed just once or a few times per cell cycle and that most of the transcriptional machinery is not active at the same time. In fact we can say that productive transcription on a yeast gene is a rare phenomenon in an actively growing cell. In any case, it is necessary to stress that the conversion of the initial nascent TR values in arbitrary units into absolute TR is based in the comparison with the absolute values from the indirect dataset (see [Methods] section) that normalizes the total amount of transcription events only to those that finally produce a mature mRNA in the cytoplasm. This means that only the "productive" events of transcription are considered. Part of the transcribing RNA pol II molecules that are detected in GRO never will produce an mRNA because of getting arrested during transcription or because the mRNA will not be exported to the cytoplasm.

With the current technology it is not possible to measure the fraction of "unproductive events" but if we take as indicator of the cryptic transcription the relative abundance of transcripts outside the coding regions, it does not seem likely that it represents a high percentage of the total RA [45]. However, it has to be taken into account that all these genome wide studies measure only the mature mRNA that has not been degraded and can be detected by microarray or RNAseq technology. To discard this we have computed the total signal due to CUTs (cryptic unstable transcripts) in a *rrp6* mutant from the experiment described in [29]. We found that cryptic transcripts in *rrp6* only represents around the 5% of the total RA signal. Thus, it can be assumed that the fraction of cryptic transcription should not be very high.

According to productive events the median elongating RNA pol II density of 0.078 molecules/kb corresponds to an average of 0.096 molecules/gene. The statistical distribution shows that less that 1% of yeast genes have >1 molecule of elongating RNA pol II/gene. Therefore, we calculate that about 690 RNA pol II molecules are actively transcribing in a snapshot of an average cell in exponential growth in YPD medium. This figure can be increased if we consider that part of the transcription events on real genes will not result in mature transcripts in the cytoplasm. For instance, we have preliminary evidence (Montón and Pérez-Ortín, unpublished) that almost half of the RNA pol II molecules that initiate elongation do not reach the end of the gene *YLR454w* when using the experimental protocol described by Mason and Struhl [28]. This fact matches the proposed model of "transcription factories" [46] which imply that, at least for short genes, two RNA pol II molecules are not simultaneously reading a gene. It has been shown by single molecule analysis that a yeast gene has a median distribution of 2 RNA pol II molecules on it [43]. The gene analyzed (*MDN1*) is, however, 14 kb long. Therefore, it is not contradictory with our conclusion. On the other hand, because our data are population averages it is very likely that some individual cells have higher densities in some genes.

According to the data discussed by Struhl [33] there are about 20000 RNA pol II molecules per cell (30000 according [47]) and 60% of those are hyperphosphorylated. On the other hand, the chromatin-associated RNA pol II molecules have been calculated to be between 60–80% using FRAP (fluorescence recovery after photobleaching) [48]. Therefore, at least 12000 RNA pol II molecules would be hyperphosphorylated in Ser2 or/and Ser5 of the carboxy terminal domain (CTD) and associated to genes but only 690–1380 (6–12% of them, depending on the proportion of mRNA nuclear degradation) would be actually transcribing the ORF-containing genes. This result coincides with the calculations made by Struhl [33] who estimated that 90% of the RNA pol II initiation events represent "transcriptional noise". The rest of the hyperphosphorylated RNA pol II molecules would be either backtracked onto ORF-containing genes (see below) or transcribing onto other genome locations from either cryptic promoters which transcripts are not stabilized in a *rrp6* mutant, or on still unknown genes (see Fig. 4b).

On the other hand, we think that the excess of phosphorylated RNA pol II molecules without transcriptional activity associated to ORF-containing genes over the active ones could be a general feature of yeast genes. In a previous paper we showed that some functional categories, such as RP genes had an excess of backtracked RNA pol II molecules [35]. This result was deduced from the comparison between RPCC with Ser5-phosphorylated-specific antibody, that detects all RNA pol II molecules having this post-translational modification onto their CTD tail, and GRO, that detects only transcriptionally active ones. However, in that study we were only able to determine that RP genes were above the general behavior but we cannot quantify the absolute number of RNA pol II molecules that are unable to elongate. This data suggests that, at least, part of the vast excess of phosphorylated molecules (about 11000) that constitute the "transcriptional noise" could be also arrested non-active competent RNA pol II molecules [49]. Another corollary of these calculations is that the RNA pol II density maps that have been published using variants of chip-on-chip technology [42,50] represent, probably, a compendium of active and non-active (arrested) molecules more than a transcribing map of the yeast genome and that it is necessary to couple those data with measurements of the polymerases activity along the genes [51] to understand the dynamics of the transcription process.

Conclusion

Our data show that most yeast genes are not being productively transcribed in a snapshot of the cell. A large proportion of the transcription is concentrated in a few genes, in fact, 25% of the total transcription is due only to 5% of the genes with the highest TR. This coincides with previous calculations that indicated that many mRNAs present in a cell are inherited from the mother [52]. Assuming that the 4670 genes, from which we obtained TR data, are representative of the 5796 non dubious ORF-containing genes, it is possible to calculate approximately 1000 productive transcription events (leading to the production of new mature mRNAs) occurring each minute in a yeast cell. This means that 113500 events take place in a typical cell cycle. Given that the current data for mRNA molecules per yeast cell is thought to be 26000 [3], the turnover of the whole set of mRNAs would be more than four times per cell cycle.

Materials and Methods

Respective contributions of mRNA degradation and dilution to the disappearance of mRNA during cell growth

Taking into account the mRNA amount for each gene and the cellular generation time, it is possible to compute the mRNA decrease inside the cell due just to the cell doubling, and thus the TR devoted to compensate this effect. In our hands the BY4741 strain has a doubling time of 113 min in exponential phase in YPD at 28°C. We computed the decrease in the mRNA concentration due to the dilution for each gene as RA*(ln2/doubling time). For this we computed a new mRNA amount dataset derived from the RNA-Seq data [4], which had been normalized to real units (mRNAs/cell) using the 2000 most abundant mRNAs from the GATC-PCR data [3].

Generation of a nascent transcription rate dataset

An averaged GRO dataset was generated using the ArrayStat software and 8 different GRO experiments of exponentially growing cells done in triplicate (for a total of 24 independent biological samples) with a minimum Pearson correlation among them of 0.7. Only those genes with at least 5 valid measures throughout the 8 different experiments were accepted. The GRO analyses were done as previously described [10]. Briefly, the signals from ^{33}P-UTP labeled run-on samples were normalized using the signals from the same filters hybridized with ^{33}P-dCTP random primer labeled genomic DNA samples.

In order to remove the gene length dependent bias (see Text S1), two different lowess smoothing analyses were carried out for those genes with lengths above or below 3 kb. Specifically, the z-score values on the logarithmic scale were computed for both GRO and RPCC data. The datasets were changed to $\log_2$ scale and to each individual value the population mean was subtracted. Then, this difference was divided by the standard deviation. Afterwards, the z-score values on the logarithmic scale were represented against the ORF length (Fig. S1). Two independent lowess smoothers were performed for the ORF above or below 3 kb using a smoother span of 20% of the data. The GRO data were normalized subtracting to each value the difference between the GRO and RPCC lowess smoothers in order to correct any probe length bias. Finally, the GRO standardization was reversed (multiplying each value by the original standard deviation and adding the population mean).

To transform the TR in arbitrary units into real ones (molecules/min), a general steady-state for the gene expression was assumed [21]. First, a indirect TR dataset was computed using genomic mRNA half-life data [17] (in minutes) and the newly computed mRNA amount dataset described above. However, this indirect TR dataset was computed using non-growing cells [17]. Thus, in order to incorporate the dilution effect (decrease in the RA due to the cellular dilution) we added to each gene the TR needed to compensate the dilution ([RA*(ln2/generation time)], as computed before). Finally, the nascent TR data, in arbitrary units, previously obtained in this study were normalized to real units (molecules/min) using the 2000 genes with the highest indirect dilution-corrected TR.

Correlations of the TR datasets with some chIP datasets

To test the improvement of the new nascent TR dataset and it relationship with modifications associated to active transcription we compared the new data with some alternative transcription datasets. These datasets include a newly computed indirect TR dataset using genomic mRNA half-life data [17] and new estimations of RA [3,4] (see above), the original indirect TR computed by Holstege *et al.* [9] and the new estimation of mRNA amount. We compared those datasets with data from [25], and computed the Spearman correlation.

Comparison between nascent TR and indirect TR datasets

In order to compare the nascent TR dataset with the indirect one, we used a Log_2(TR/indTR) dataset in which the general bias due to the different transcription datasets amplitudes was corrected. Specifically, we plot Log_2(TR/indTR) against Log_2(TR + indTR) and corrected the bias related with the TR using a lowess smoothing with a smoother span of 20% of the data.

Accession numbers

All the experiments were done in triplicate. Genomic data are stored in the Valencia Yeast (VYdBase; http://vydbase.uv.es/) and the GEO databases. GEO series accession numbers are GSE14060 for RPCC hybridizations, and GSE11521 and GSE13096 for GRO data.

Supporting Information

Text S1   Supplementary analysis and references. (DOC)

Table S1   Whole dataset of the nascent TR obtained in this work and other transcription parameters. It is presented the nascent (TR) and the indirect transcription rates (TR indirect) computed from [17]. Both data are shown before and after the correction used to compensate the dilution effect. (XLS)

Table S2   Nascent TR datasets from to other physiological conditions. We used previously published GRO datasets to expand the change from cells growing in glucose to galactose [10] oxidative stress [36] and osmotic stress [37]. (XLS)

Table S3   Percentage of TR devoted to compensate the dilution effect due to the cell growth. The TR necessary to compensate the dilution is computed using the RA and a generation time of 113 minutes. As the TR necessary to compensate the dilution is computed independently to the nascent TR some values show percentage greater than 100% or are negative. In those cases we have arbitrary substituted the values to either 100% or 0%, meaning that the TR devoted to compensate the dilution is much larger (100%) or negligible (0%) in respect to the TR devoted to compensate the degradation. (XLS)

Figure S1   Transcription data biases due to the ORF and probe length. A) Comparison between RPCC (red), GRO (blue, data from [10]) and random-priming cDNA data (green, data from [10]). B) Confirmation of the 3′ labeling bias using an oligo-dT cDNA labeling (from ref. [S2], purple) and random-priming cDNA data as a control (green). All the curves represent the smoothness of the data using the averages values for a sliding window of 100 genes. The 3 kb ORF length (vertical red line) marks a change in the probe length design for the arrays from the complete ORF for genes shorter than 3 kb to the 3′ last 1 kb in longer genes [S3]. C) Lowess correction (yellow) of the probe length dependent bias for the GRO data (blue) using RPCC data (red). All the values are presented as z-score standardized arbitrary units. (PDF)

Figure S2   Comparison of the direct and indirect TR datasets with chromatin modifications related to active promoters. Comparison of the new direct TR dataset (this study) and indirect TR datasets computed using the mRNA amount and the mRNA stability [9,17] as in Fig. 1 with chromatin modification associated with active promoters as H3K9 acetylation (A) or H3K14 acetylation (B) and H3K4 trimethylation [25]. To avoid any bias depending on the wideness of the datasets Spearman rank correlation is used. Note that no clear improvement in the correlation indexes exists when comparing nascent and indirect datasets. For simplicity, we have used the average values for the probes covering the coding region as indicator of the chromatin modification or protein binding, or when using RA dataset. (PDF)

Figure S3   Comparison between nascent and indirect transcription rates. Comparison of the new direct TR dataset (this study) and indirect TR datasets computed using the mRNA amount and the mRNA stability [9,17]. Note than the correlation is higher when comparing the two indirect TR datasets (C) than when comparing direct and indirect datasets (A, B). Spearman rank correlation is showed. (PDF)

Figure S4   Differences between nascent and indirect transcription rates according to the overlap with non-coding transcripts. The box plot shows the relationship Log_2(TR/indTR) after lowess correction (see [Methods] section) for ORFs overlapping with CUTs (red), SUTs (blue) or without any overlapping transcript (bisque) according to [29]. The box plots show the median and quartiles of the data. The whiskers show the maximum and minimum of the data set (excluding the outliers which lie beyond 1.5 times the inter quartile range). The significance values represent the p value for the t test of difference of the mean. (PDF)

Figure S5   Differences between nascent and indirect transcription rates according to RNA pol II density across the genes. The box plot shows the relationship Log_2(TR/indTR) after lowess correction (see [Methods] section) for genes classified according to their distribution of RNA pol II along the gene. It is shown cluster 1 (red, genes with accumulation of non-elongating polymerases in the promoter region), cluster 2 (blue, genes with enriched levels of RNA pol II molecules on their 5′ end transcribed region) and cluster 3 (bisque, genes with an even or enriched to the 3′ distribution of polymerases) in reference to the rest of the analyzed genes (white boxes). See Venters and Pugh [24] for cluster details. The box plots show the median and quartiles of the data. The whiskers show the maximum and minimum of the data set (excluding the outliers which lie beyond 1.5 times the inter quartile range). The significance values represent the p value for the t test of difference of the mean. (PDF)

Figure S6   Functional analysis of the TR using the FatiScan tool. (Babelomics software, [S3]). GOs in red represent the groups associated with high TRs (either overrepresented groups in genes with high TRs (left) or underrepresented groups in genes with low TRs (right)). GOs in blue represent the groups associated with low TRs (either overrepresented groups in genes with low TRs (right) or underrepresented groups in genes with high TRs (left)). Only GOs with a statistically significant difference from the population is shown (multitest corrected p-value <0.001). The three parts represent independent searches for GOs in "biological process", "molecular function" and "cellular component", in that order. (PDF)

Figure S7   Functional analysis of the percentage of TR devoted to compensate the dilution. (Babelomics software, [S3]). GOs in red represent the groups associated with genes that devote a high percentage of its TR to compensate the dilution due to the cellular growth, whereas in blue are represented those genes in which the contribution of the dilution to the mRNA disappearance is smaller. Only GOs with a statistically significant difference from the population are shown (multitest corrected p-value <0.001). The three parts represent independent searches for GOs in "biological process", "molecular function" and "cellular component", in that order. (PDF)

Acknowledgments

Author Contributions

Conceived and designed the experiments: VP SC JEP-O . Performed the experiments: VP. Analyzed the data: VP SC JEP-O . Wrote the paper: JEP-O .

References

1. Benjamin Lewin (2007) Genes IX. Massachusetts: Jones & Bartlett Publishers.

2. Pérez-Ortín JE, García-Martínez J, Alberola TM (2002) DNA chips for yeast biotechnology. The case of wine yeasts. J Biotechnol 98: 227–241.

3. Miura F, Kawaguchi N, Yoshida M, Uematsu C, Kito K, et al. (2008) Absolute quantification of the budding yeast transcriptome by means of competitive PCR between genomic and complementary DNAs. BMC Genomics 9: 574.

4. Nagalakshmi U, Wang Z, Waern K, Shou C, Raha D, et al. (2008) The transcriptional landscape of the yeast genome defined by RNA sequencing. Science 320: 1344–1349.

5. Pérez-Ortín JE, Alepuz PM, Moreno J (2007) Genomics and gene transcription kinetics in yeast. Trends Genet 23: 250–257.

6. Grigull J, Mnaimneh S, Pootoolal J, Robinson MD, Hughes TR (2004) Genome-wide analysis of mRNA stability using transcription inhibitors and microarrays reveals posttranscriptional control of ribosome biogenesis factors. Mol Cell Biol 24: 5534–5547.

7. García-Martínez J, González-Candelas F, Pérez-Ortín JE (2007) Common gene expression strategies revealed by genome-wide analysis in yeast. Genome Biol 8: R222.

8. Mata J, Marguerat S, Bahler J (2005) Post-transcriptional control of gene expression: a genome-wide perspective. Trends Biochem Sci 30: 506–514.

9. Holstege FC, Jennings EG, Wyrick JJ, Lee TI, Hengartner CJ, et al. (1998) Dissecting the regulatory circuitry of a eukaryotic genome. Cell 95: 717–728.

10. García-Martínez J, Aranda A, Pérez-Ortín JE (2004) Genomic run-on evaluates transcription rates for all yeast genes and identifies gene regulatory mechanisms. Mol Cell 15: 303–313.

11. Core LJ, Waterfall JJ, Lis JT (2008) Nascent RNA sequencing reveals widespread pausing and divergent initiation at human promoters. Science 322: 1845–1848.

12. Fan J, Yang X, Wang W, Wood WH, III, Becker KG, et al. (2002) Global analysis of stress-regulated mRNA turnover by using cDNA arrays. Proc Natl Acad Sci U S A 99: 10611–10616.

13. Hirayoshi K, Lis JT (1999) Nuclear run-on assays: assessing transcription by measuring density of engaged RNA polymerases. Methods Enzymol 304: 351–362.

14. Iyer V, Struhl K (1996) Absolute mRNA levels and transcriptional initiation rates in Saccharomyces cerevisiae. Proc Natl Acad Sci U S A 93: 5208–5212.

15. García-Martínez J, Pelechano V, Pérez-Ortín JE (2010) Genomic methods for evaluating transcription rates in yeast. In: Yeast Genetic Networks New York; Humana press inc. In press.

16. Hereford LM, Rosbash M (1977) Number and distribution of polyadenylated RNA sequences in yeast. Cell 10: 453–462.

17. Wang Y, Liu CL, Storey JD, Tibshirani RJ, Herschlag D (2002) Precision and functional specificity in mRNA decay. Proc Natl Acad Sci U S A 99: 5860–5865.

18. Jasiak AJ, Hartmann H, Karakasili E, Kalocsay M, Flatley A, et al. (2008) Genome-associated RNA polymerase II includes the dissociable Rpb4/7 subcomplex. J Biol Chem 283: 26423–26427.

19. Marin A, Gallardo M, Kato Y, Shirahige K, Gutierrez G, et al. (2003) Relationship between G+C content, ORF-length and mRNA concentration in Saccharomyces cerevisiae. Yeast 20: 703–711.

20. Rosaleny LE, Ruiz-García AB, García-Martinez J, Pérez-Ortín JE, Tordera V (2007) The Sas3p and Gcn5p histone acetyltransferases are recruited to similar genes. Genome Biol 8: R119.

21. Pelechano V, Pérez-Ortín JE (2010) There is a steady-state transcriptome in exponentially growing yeast cells. Yeast 7: 413–422.

22. Komarnitsky P, Cho EJ, Buratowski S (2000) Different phosphorylated forms of RNA polymerase II and associated mRNA processing factors during transcription. Genes Dev 14: 2452–2460.

23. Liu CL, Kaplan T, Kim M, Buratowski S, Schreiber SL, et al. (2005) Single-nucleosome mapping of histone modifications in S. cerevisiae. PLoS Biol 3: e328.

24. Venters BJ, Pugh BF (2009) A canonical promoter organization of the transcription machinery and its regulators in the Saccharomyces genome. Genome Res 19: 360–371.

25. Pokholok DK, Harbison CT, Levine S, Cole M, Hannett NM, et al. (2005) Genome-wide map of nucleosome acetylation and methylation in yeast. Cell 122: 517–527.

26. Rando OJ, Chang HY (2009) Genome-wide views of chromatin structure. Annu Rev Biochem 78: 245–271.

27. Ginsburg DS, Govind CK, Hinnebusch AG (2009) NuA4 lysine acetyltransferase Esa1 is targeted to coding regions and stimulates transcription elongation with Gcn5. Mol Cell Biol 29: 6473–6487.

28. Mason PB, Struhl K (2005) Distinction and relationship between elongation rate and processivity of RNA polymerase II in vivo. Mol Cell 17: 831–840.

29. Xu Z, Wei W, Gagneur J, Perocchi F, Clauder-Munster S, et al. (2009) Bidirectional promoters generate pervasive transcription in yeast. Nature 457: 1033–1037.

30. Edwards AM, Kane CM, Young RA, Kornberg RD (1991) Two dissociable subunits of yeast RNA polymerase II stimulate the initiation of transcription at a promoter in vitro. J Biol Chem 266: 71–75.

31. Niemisto A, Nykter M, Aho T, Jalovaara H, Marjanen K, et al. (2007) Computational methods for estimation of cell cycle phase distributions of yeast cells. EURASIP J Bioinform Syst Biol 2007: 46150.

32. French SL, Osheim YN, Cioci F, Nomura M, Beyer AL (2003) In exponentially growing Saccharomyces cerevisiae cells, rRNA synthesis is determined by the summed RNA polymerase I loading rate rather than by the number of active genes. Mol Cell Biol 23: 1558–1568.

33. Struhl K (2007) Transcriptional noise and the fidelity of initiation by RNA polymerase II. Nat Struct Mol Biol 14: 103–105.

34. Warner JR (1999) The economics of ribosome biosynthesis in yeast. Trends Biochem Sci 24: 437–440.

35. Pelechano V, Jimeno-González S, Rodríguez-Gil A, García-Martínez J, Pérez-Ortín JE, et al. (2009) Regulon-specific control of transcription elongation across the yeast genome. PLoS Genet 5: e1000614.

36. Molina-Navarro MM, Castells-Roca L, Belli G, García-Martínez J, Marín-Navarro J, et al. (2008) Comprehensive transcriptional analysis of the oxidative response in yeast. J Biol Chem 283: 17908–17918.

37. Romero-Santacreu L, Moreno J, Pérez-Ortín JE, Alepuz P (2009) Specific and global regulation of mRNA stability during osmotic stress in Saccharomyces cerevisiae. RNA 15: 1110–1120.

38. Kulaeva OI, Gaykalova DA, Studitsky VM (2007) Transcription through chromatin by RNA polymerase II: histone displacement and exchange. Mutat Res 618: 116–129.

39. Schwabish MA, Struhl K (2004) Evidence for eviction and rapid deposition of histones upon transcriptional elongation by RNA polymerase II. Mol Cell Biol 24: 10111–10117.

40. Lucchini R, Sogo JM (1998) The dynamic structure of ribosomal RNA gene chromatin. In: Paule MR, ed. Transcription of ribosomal RNA genes by eukaryotic RNA polymerase I. Austin: Springer-Verlag and RG Landes Co. pp 255–276.

41. Lee CK, Shibata Y, Rao B, Strahl BD, Lieb JD (2004) Evidence for nucleosome depletion at active regulatory regions genome-wide. Nat Genet 36: 900–905.

42. Koerber RT, Rhee HS, Jiang C, Pugh BF (2009) Interaction of transcriptional regulators with specific nucleosomes across the Saccharomyces genome. Mol Cell 35: 889–902.

43. Zenklusen D, Larson DR, Singer RH (2008) Single-RNA counting reveals alternative modes of gene expression in yeast. Nat Struct Mol Biol 15: 1263–1271.

44. Singh J, Padgett RA (2009) Rates of in situ transcription and splicing in large human genes. Nat Struct Mol Biol 16: 1128–1133.

45. van BH, Nislow C, Blencowe BJ, Hughes TR (2010) Most ''dark matter'' transcripts are associated with known genes. PLoS Biol 8: e1000371.

46. Carter DR, Eskiw C, Cook PR (2008) Transcription factories. Biochem Soc Trans 36: 585–589.

47. Borggrefe T, Davis R, Bareket-Samish A, Kornberg RD (2001) Quantitation of the RNA polymerase II transcription machinery in yeast. J Biol Chem 276: 47150–47153.

48. Sprouse RO, Karpova TS, Mueller F, Dasgupta A, McNally JG (2008) Regulation of TATA-binding protein dynamics in living yeast cells. Proc Natl Acad Sci U S A 105: 13304–13308.

49. Sigurdsson S, Dirac-Svejstrup AB, Svejstrup JQ (2010) Evidence that transcript cleavage is essential for RNA polymerase II transcription and cell viability. Mol Cell 38: 202–210.

50. Steinmetz EJ, Warren CL, Kuehner JN, Panbehi B, Ansari AZ (2006) Genome-wide distribution of yeast RNA polymerase II and its control by Sen1 helicase. Mol Cell 24: 735–746.

51. Rodríguez-Gil A, García-Martínez J, Pelechano V, Munoz-Centeno MD, Geli V, et al. (2010) The distribution of active RNA polymerase II along the transcribed region is gene-specific and controlled by elongation factors. Nucleic Acids Res 38: 4651–64.

52. Bon M, McGowan SJ, Cook PR (2006) Many expressed genes in bacteria and yeast are transcribed only once per cell cycle. FASEB J 20: 1721–1723.

Supplementary analysis

Detection and correction of the specific probe length data in GRO

When plotting the GRO and RPCC data against the gene length a negative correlation can be detected between the gene length and the TR. Although the fact that shorter genes tend to be more expressed has already been described [3, 10, 19] the slope of this bias is higher in the GRO dataset (blue) than in the RPCC (red) or a mRNA labeling using random priming (green) (supplemental Fig S1). Moreover, the GRO curve shows a change in tendency for those genes whose ORF was longer than 3 kb. The average line increases to levels higher than 0 with genes reaching levels of about 1 kb in length. The macroarrays used for both RPCC and GRO present for those genes longer than 3 kb a probe that covers only the 1 kb downstream of the ORF, whereas the rest of the genes (<3 kb) are represented by the whole of the ORF [S1]. As this technical effect is only observed in the GRO data, this allows to determine that the differences in the slope against the length of the gene is in fact a technical bias specific of the GRO technique

As the GRO method labeled the elongating mRNA by extending it 200-300 nt downstream [13], it is expected to bias the label to the 3' part of the genes. This would lead to an increased average label in the genes as their length decreases and the relative influence of their 3' part increases. It also predicts a sudden increase for the genes represented by probes covering only their last 1 kb of the ORF. What this reasoning predicts is that other labeling protocols with a bias to the 3' end of the mRNA will also present a similar average curve. To confirm this hypothesis we made a similar analysis using the data of mRNA labeled using an oligo-dT primer for the cDNA synthesis. This kind of labeling which mainly labels the 3' end of the mRNA (purple curve in supplemental Fig S1B), shows a similar profile to that of the GRO curve, while the curve obtained by random priming labeling of cDNA (green curve) is similar to that of the RPCC. With this reasoning, we may conclude that there is a general bias in the GRO that increases the calculated TR inversely with the ORF (i.e., probe) length. Additionally, there is an artifact for ORFs that is longer than 3 kb due to the use of 1 kb probes from their 3' end. Thus, this gene/prove dependent effect can be corrected by using the RPCC data for lowess normalization of the GRO data.

We analyzed the GRO and RPCC data for other biases related with G+C content, transcription rate or gene length. However we were not able to identify other technical bias that we could unequivocally assign to one of the two methods.

Supplementary references

S1 Alberola TM, García-Martínez J, Antúnez O, Viladevall L, Barceló A et al. (2004) A new set of DNA macrochips for the yeast Saccharomyces cerevisiae: features and uses. Int Microbiol 7: 199-206.

S2 Pelechano V, Pérez-Ortín JE (2008) The transcriptional inhibitor thiolutin blocks mRNA degradation in yeast. Yeast 25: 85-92.

S3 Al-Shahrour F, Carbonell J, Mínguez P, Goetz S, Conesa A et al. (2008) Babelomics: advanced functional profiling of transcriptomics, proteomics and genomics experiments. Nucleic Acids Res 36: W341-W346.