Genomics and Bioinformatics

Examination - Week 14

December 17, 2013

Question 1 - Phylogenetic trees

Use the UPGMA algorithm to build the rooted tree T corresponding to the following distance matrix M:

M	a	b	c	d	e
a	0	4	10	10	14
b	4	0	10	10	14
c	10	10	0	2	14
d	10	10	2	0	14
e	14	14	14	14	0

Question 2 - Linear models

A microarray contains 6 probes (P1 to P6). The raw intensities for three experimental conditions A,B,C (three different arrays) have been recorded and reported in the table below. We want to compare them.

	P1	P2	P3	P4	P5	P6
A	166	121	166	270	39	121
B	49	49	90	126	18	90
C	10	14	24	14	10	3

1. Consider C as the control (untreated) sample, A as the treated sample. Write a linear model that describes the probe intensities as a function of the treatment A. Write it in matrix form, replacing all known quantities by their numeric value.

2. Here is the result of a similar linear regression with treatment B:

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)    12.50      11.37   1.099  0.29749
B              57.83      16.08   3.596  0.00488 **
```

- Would you say that treatment B has a significant effect (i.e. what is the probability to observe an even bigger difference in the future, given our data, under the hypothesis that B has no effect)?
- What is the expected increase in probe intensity when treatment B is applied?

3. Apply quantile normalization to all three samples in the data above. Explain the purpose of this operation.

Question 3 - Transcription

The following questions are based on the NET-seq paper discussed in week 10.

1. In figure 1b of the article, why is there a region with no (or very low) signal in the fragmented mature RNA?
2. What do you deduce from figure 2d?
3. Based on the data in figure 2d and knowing that RCO1 is required for histone H4 deacetylation:
    - What is the expected effect of RCO1 deletion on transcription?
    - Which figure demonstrates this effect on the genome-wide transcriptional levels? And how?

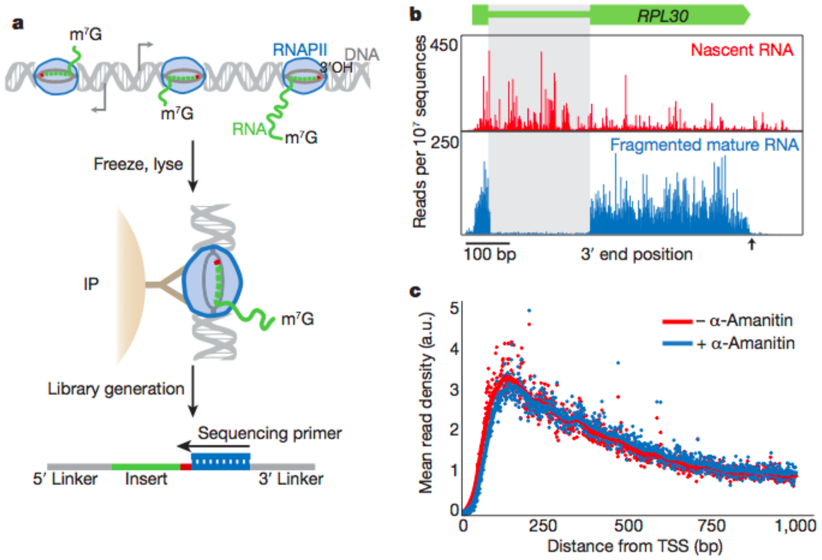

Figure 1 | NET-seq visualizes active transcription via capture of 3′ RNA termini. a, Schematic diagram of NET-seq protocol. A yeast culture is flash frozen and cryogenically lysed. Nascent RNA is co-purified via an immunoprecipitation (IP) of the RNAPII elongation complex. Conversion of RNA into DNA results in a DNA library with the RNA as an insert between DNA sequencing linkers. The sequencing primer is positioned such that the 3′ end of the insert is sequenced. m^{7}G refers to the 7-methylguanosine cap structure at the 5′ end of nascent transcripts. **b**, The 3′ end of each sequence is mapped to the yeast genome and the number of reads at each nucleotide is plotted at the *RPL30* locus for nascent RNA and lightly fragmented mature RNA. Note that for the nascent transcripts, the introns (grey box) and regions after the polyadenylation site (black arrow) are readily detected. **c**, Metagene analysis for well-expressed genes ($n = 471$, >1.5 reads per bp in both conditions) of the mean read density (arbitrary units, a.u.) in the presence and absence of transcription inhibitor, α-amanitin. TSS, transcription start site.

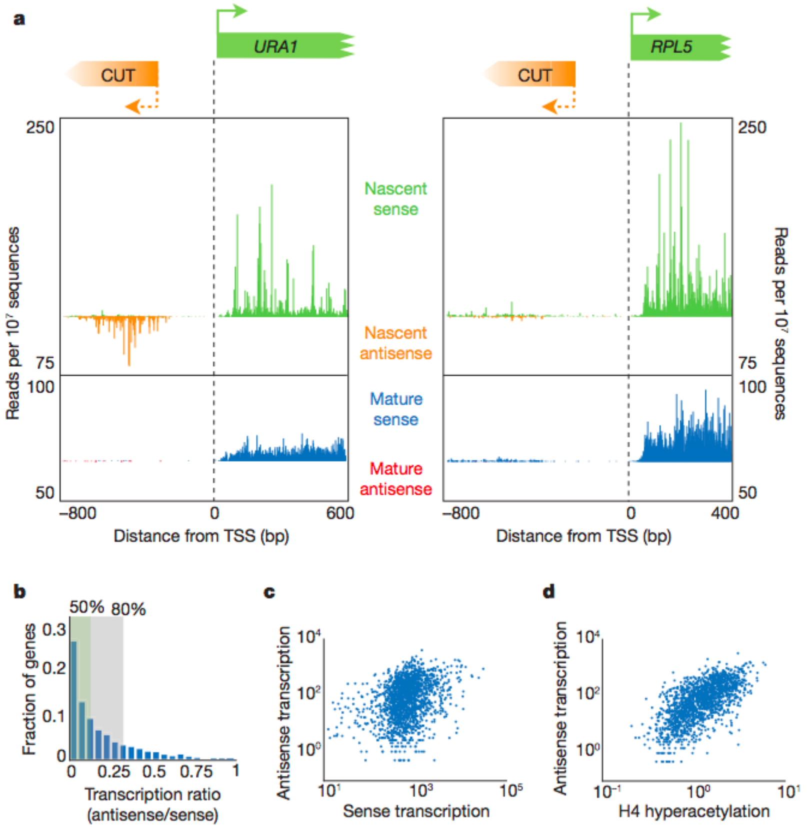

Figure 2 | Observation of divergent transcripts reveals strong directionality at most promoters. **a,** Nascent and mature transcripts initiating from *URA1* and *RPL5* promoters in the sense and antisense directions. Note that there are cryptic unstable transcripts (CUTs) in the antisense direction for *URA1* but not *RPL5*. **b,** A histogram of the transcription ratio (antisense/sense transcription levels) for 1,875 genes. The green and grey boxes indicate the subset of genes with a ratio of less than 1:8 and less than 1:3, respectively. **c,** Antisense transcription levels are plotted versus sense transcription for each tandem gene (Spearman correlation coefficient, $r_s = 0.34$). **d,** The level of antisense transcription for each promoter is plotted versus the local enrichment for H4 hyperacetylation using available data[24] ($r_s = 0.65$).

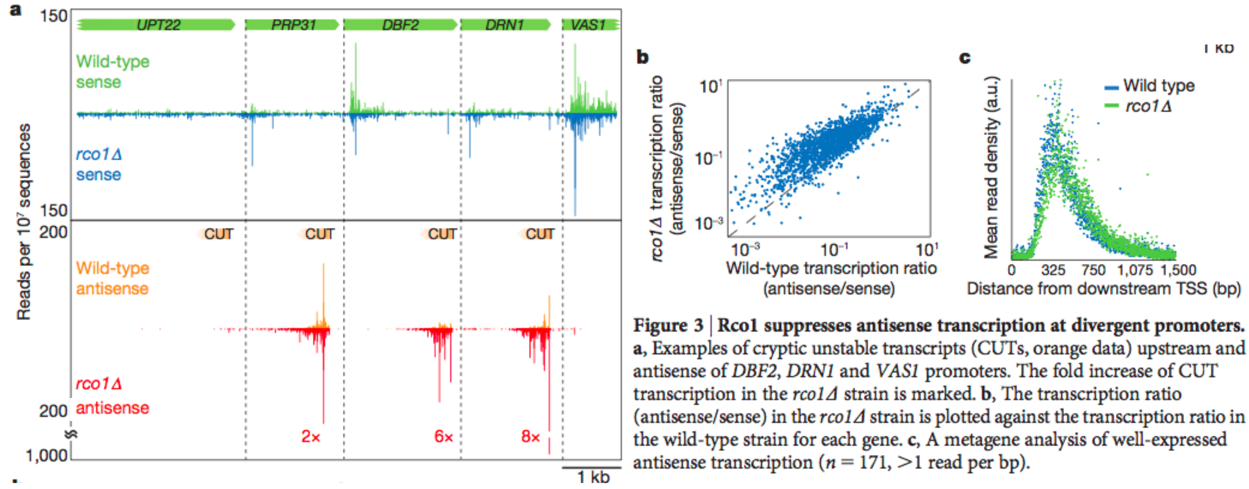

Figure 3 | Rco1 suppresses antisense transcription at divergent promoters. **a,** Examples of cryptic unstable transcripts (CUTs, orange data) upstream and antisense of *DBF2*, *DRN1* and *VAS1* promoters. The fold increase of CUT transcription in the *rco1Δ* strain is marked. **b,** The transcription ratio (antisense/sense) in the *rco1Δ* strain is plotted against the transcription ratio in the wild-type strain for each gene. **c,** A metagene analysis of well-expressed antisense transcription ($n = 171$, >1 read per bp).

Question 4 - DNA Binding

A transcription factor's DNA-binding domain (of length 4) is described by a Position-Weight Matrix (PWM). In units of nucleotide frequencies, the matrix is:

$$M = \begin{array}{c} \\ 1 \\ 2 \\ 3 \\ 4 \end{array} \begin{array}{c} \begin{array}{cccc} \text{A} & \text{C} & \text{G} & \text{T} \end{array} \\ \begin{pmatrix} 0.082594539 & 0.610295685 & 0.224515236 & 0.082594539 \\ 0.610360368 & 0.370202277 & 0.009718678 & 0.009718678 \\ 0.009718678 & 0.009718678 & 0.370202277 & 0.610360368 \\ 0.082594539 & 0.224515236 & 0.610295685 & 0.082594539 \end{pmatrix} \end{array} .$$

We usually work with the logarithm of this matrix, which has the same units as the free energy and, after subtracting an arbitrary constant to make the numbers simpler, the PWM is as follows:

$$W = \begin{array}{c|cccc} & \text{A} & \text{C} & \text{G} & \text{T} \\ \hline 1 & -1 & 1 & 0 & -1 \\ 2 & 1 & 0.5 & -3.14 & -3.14 \\ 3 & -3.14 & -3.14 & 0.5 & 1 \\ 4 & -1 & 0 & 1 & -1 \end{array} .$$

1. What is the consensus sequence for this motif?

2. What are the two second best sequences?

3. Compute the best motif score on the following sequence: `ATAGCCTAG`

4. Using the limit of low protein concentration we know that the occupancy of a sequence by the transcription factor is proportional to $\exp(W)$ (we set $\beta = 1$) and assuming that the accessible part of the genome consists of 1000 times the motif `CTGG` and 1 motif `CATG`, which proportion of transcription factors will bind to `CATG`? Here are some numerical values that can be useful:

x	e^x
-3	0.05
-4	0.02
-4.64	0.01
-5.00	0.007
-6.14	0.002
-6.64	0.001
-7.64	0.0005
-8.28	0.0002