RNA polymerase errors cause splicing defects and can be regulated by differential expression of RNA polymerase subunits.

Lucas B. Carey[1*]

[1] Department of Experimental and Health Sciences, Universitat Pompeu Fabra, Dr. Aiguader 88, 08003 Barcelona, Spain.

[*] Correspondence to: Lucas B. Carey (lucas.carey@upf.edu)

Errors during transcription may play an important role in determining cellular phenotypes: the RNA polymerase error rate is >4 orders of magnitude higher than that of DNA polymerase and errors are amplified >1000-fold due to translation. However, current methods to measure RNA polymerase fidelity are low-throughout, technically challenging, and organism specific. Here I show that changes in RNA polymerase fidelity can be measured using standard RNA sequencing protocols. I find that RNA polymerase is error-prone, and these errors can result in splicing defects. Furthermore, I find that differential expression of RNA polymerase subunits causes changes in RNA polymerase fidelity, and that coding sequences may have evolved to minimize the effect of these errors. These results suggest that errors caused by RNA polymerase may be a major source of stochastic variability at the level of single cells.

24 The information that determines protein sequence is stored in the genome
25 but that information must be transcribed by RNA polymerase and translated by the
26 ribosome before reaching its final form. DNA polymerase error rates have been well
27 characterized in a variety of species and environmental conditions, and are low, on
28 the order of one mutation per 10^8 - 10^{10} bases per generation[1-3]. In contrast, RNA
29 polymerase errors are uniquely positioned to generate phenotypic diversity. Error
30 rates are high (10^{-6}-10^{-5})[4-7], and each mRNA molecule is translated into 2,000 –
31 4,000 molecules of protein[8,9], resulting in amplification of any errors. Likewise,
32 because many RNAs are present in less than one molecule per cell in microbes[10,11]
33 and embryonic stem cells[12], an RNA with an error may be the only RNA for that
34 gene; all newly translated protein will contain this error. Despite the fact that
35 transient errors can result in altered phenotypes[13,14], the genetics and
36 environmental factors that affect RNA polymerase fidelity are poorly understood.
37 This is because current methods for measuring polymerase fidelity are technically
38 challenging[4], require specialized organism-specific genetic constructs[15], and can
39 only measure error rates at specific loci[16].
40
41 To overcome these obstacles I developed MORPhEUS (Measurement Of RNA
42 Polymerase Errors Using Sequencing), which enables measurement of differential
43 RNA polymerase fidelity using existing RNA-seq data **(Figure 1)**. The input is a set
44 of RNA-seq fastq files and a reference genome, and the output is the error rate at
45 each position in the genome. I find that RNA polymerase errors result in intron
46 retention and that cellular mRNA quality control may reduce the effective RNA

47 polymerase error rate. Moreover, our analyses suggest that the expression level of
48 the RPB9 Pol II subunit determines RNA polymerase fidelity *in-vivo*. Because it can
49 be run on any existing RNA-seq data, MORPhEUS enables the exploration of a
50 previously unexplored source of biological diversity in microbes and mammals.
51
52 Technical errors from reverse transcription and sequencing, and biological
53 errors from RNA polymerase look identical (single-nucleotide differences from the
54 reference genome). Therefore, a major challenge in identifying SNPs and in
55 measuring changes in polymerase fidelity is the reduction of technical errors[17-
56 19]**(Figure 1)**. First, I map full length (untrimmed) reads to the genome, and discard
57 reads with indels, more than two mismatches, reads that map to multiple locations
58 in the genome, and reads that do not map end-to-end along the full length of the
59 read. I next trim the ends of the mapped reads, as alignments are of lower quality
60 along the ends, and the mismatch rate is higher, especially at splice junctions. I also
61 discard any cycles within the run with abnormally high error rates, and bases with
62 low Illumina quality scores **(Figure 1 – figure supplement 1)**. Finally, using the
63 remaining bases, I count the number of matches and mismatches to the reference
64 genome at each position in the genome. I discard positions with identical
65 mismatches that are present more than once, as these are likely due to subclonal
66 DNA polymorphisms or sequences that Illumina miscalls in a systematic manner[20]
67 **(Figure 1 – figure supplement 2)**. The result is a set of mismatches, many of which
68 are technical errors, some of which are RNA polymerase errors. In order to
69 determine if RNA-seq mismatches are due to RNA polymerase errors it is necessary

70 to identify sequence locations in which RNA polymerase errors are expected to have
71 a measurable effect, or situations in which RNA polymerase fidelity is expected to
72 vary.
73
74 I reasoned that RNA polymerase errors that alter positions necessary for
75 splicing should result in intron retention, while sequencing errors should not affect
76 the final structure of the mRNA **(Figure 2a)**. However, mutations in the donor and
77 acceptor splice sites also result in decreased expression[36], and therefore are difficult
78 to measure using RNA-seq. I therefore used chromatin-associated and nuclear RNA
79 from Hela and Huh7 cells[37], and extracted all reads that span an exon-intron
80 junction for introns with canonical GT and AG splice sites, and measured the RNA-
81 seq mismatch rate at each position. I find that errors at the G and U in the 5' donor
82 site, and at the A in the acceptor site are significantly enriched relative to errors at
83 other positions **(Figure 2b)**, and to errors trinucleotides present in the splicing
84 motifs in the human genome **(Figure 2 – figure supplement 1)** suggesting that
85 RNA polymerase mismatches can result in changes in transcript isoforms. The
86 ability of RNA polymerase errors to significantly affect splicing has been proposed[22]
87 but never previously measured.
88
89 RPB9 is known to be involved in RNA polymerase fidelity *in vitro* and *in*
90 *vivo*[15,23]. I therefore reasoned that cell lines expressing low levels of RPB9 would
91 have higher RNA polymerase error rates. Consistent with this, I find that RPB9
92 expression varies 8-fold across the ENCODE cell lines, and this expression variation

93 is correlated with the RNA-seq error rate **(Figure 2c, Figure 2 – figure supplement**
94 **2)**. This suggests that low RPB9 expression may cause decreased polymerase
95 fidelity *in-vivo*.
96
97 In addition, export of mRNAs from the nucleus involves a quality-control
98 mechanism that checks if mRNAs are fully spliced and have properly formed 5' and
99 3' ends[24]. I hypothesized that mRNA export may involve a quality control that
100 removes mRNAs with errors. I used the ENCODE dataset in which nuclear and
101 cytoplasmic poly-A+ mRNAs I re sequenced, thus I can compare nuclear and
102 cytoplasmic fractions from the same cell line grown in the same conditions and
103 processed in the same manner. I find that the nuclear fraction has a higher RNA
104 polymerase error rate than does the cytoplasmic fraction **(Figure 2c,d)**, suggesting
105 that either that nuclear RNA-seq has a higher technical error rate or that the cell has
106 mechanisms for reducing the effective polymerase error rate by preventing the
107 export of mRNAs that contain errors.
108
109 Rpb9 and Dst1 are known to be involved in RNA polymerase fidelity *in-vitro*,
110 yet there is conflicting evidence as to the role of Dst1 *in-vivo*[6,15,23,25-27]. Part of these
111 conflicts may result from the fact that the only available assays for RNA polymerase
112 fidelity are special reporter strains that rely on DNA sequences known to increase
113 the frequency of RNA polymerase errors. While I found that RPB9 expression
114 correlates with RNA-seq error rates in mammalian cells, correlation is not
115 causation. Furthermore, differences in RNA levels do not necessitate differences in

116 stoichiometry among the subunits in active Pol II complexes. In order to determine
117 if differential expression of RPB9 or DST1 are causative for differences in RNA
118 polymerase fidelity in-vivo, I constructed two yeast strains in which I can alter the
119 expression of either RPB9 or DST1 using B-estradiol and a synthetic transcription
120 factor that has no effect on growth rate or the expression of any other genes[28,29]. I
121 grew these two strains (Z_3EV_{pr}-RPB9 and Z_3EV_{pr}-DST1) in different concentrations
122 of B-estradiol and performed RNA-seq. I find that cells expressing low levels of
123 RPB9 have high RNA polymerase error rates **(Figure 3a)**. Likewise, cells with low
124 DST1 have high error rates **(Figure 3a)**. The increase in errors rate is not a property
125 of cells defective for transcription elongation **(Figure 3 – figure supplement 1)**.
126 The increase in error rates due to mutations in Rpb9 and Dst1 have not been
127 robustly measured, however, there are some rough numbers. Here, the measured
128 increase in error rate is 13%, while the measured effect of Rpb9 deletion *in*-vitro is
129 5-fold[38] and in-vivo following reverse transcription is 30%[25]. If 2% of the observed
130 mismatches are due to RNA polymerase errors, a 5-fold increase in polymerase
131 error rate results in a 10% increase in measured mismatch frequency; this is
132 consistent with RNA polymerase fidelity of 10^{-6}-10^{-5} and overall RNA-seq error rates
133 of 10^{-4}. Note that in our assay cells still express low levels or RPB9, and we therefore
134 expect the increase in error rate to be lower, suggesting that RNA polymerase errors
135 constitute 5-10% of the measured mismatches. Our ability to genetically control the
136 expression of DST1 and RPB9, and measure changes in RNA-seq error rates is
137 consistent with MORPhEUS measuring RNA polymerase fidelity. In addition, genetic
138 reduction in RNA polymerase fidelity results in increased intron retention,

139 consistent with RNA polymerase errors causing reduced splicing efficiency **(Figure**
140 **3b)**.
141
142 A unique advantage of MORPhEUS is that it measures thousands of RNA
143 polymerase errors across the entire transcriptome in a single experiment, and thus
144 enables a complete characterization of the mutation spectrum and biases of RNA
145 polymerase. I asked how altered RPB9 and DST1 expression levels affect each type
146 of single nucleotide change. I find that, with decreasing polymerase fidelity,
147 transitions increase more than transversions, and that C->U errors are the most
148 common **(Figure 3c)**. This result, along with other sequencing based results[4], have
149 shown that DNA and RNA polymerase have broadly similar error profiles[2]; it will be
150 interesting to see if all polymerases share the same mutation spectra, and if this is
151 due to deamination of the template base, or is a structural property of the
152 polymerase itself. Interestingly, I find that coding sequences have evolved so that
153 errors are less likely to produce in-frame stop codons than out-of-frame stop
154 codons, suggesting that natural selection may act to minimize the effect of
155 polymerase errors **(Figure 4)**.
156
157 Here I have presented proof that relative changes in RNA polymerase error
158 rates can be measured using standard Illumina RNA-seq data. Consistent with
159 previous work *in-vivo* and *in-vitro*, I find that depletion of RPB9 or Dst1 results in
160 higher RNA polymerase error rates. Futhermore, I find that expression of RPB9
161 negatively correlates with RNA-seq error rates in human cell lines, suggesting that

162 differential expression of RPB9 may regulate RNA polymerase fidelity in-vivo in
163 humans. In addition, consistent with the errors detected by MORPhEUS being due to
164 RNA polymerase and not technical errors, in reads spanning an exon-intron
165 junction, the measured error rate is higher at the 5' donor splice site, suggesting that
166 RNA polymerase errors result in intron retention. Because it can be run on existing
167 RNA-seq data, I expect MORPhEUS to enable many future discoveries regarding both
168 the molecular determinants of RNA polymerase error rates, and the relationship
169 between RNA polymerase fidelity and phenotype.
170

171 Acknowledgements

172 I thank members of the Carey lab and the computational genomics groups in the
173 PRBB for thoughtful discussions.
174

175 Materials and methods

176 Counting RNA polymerase errors in already aligned ENCODE data

177 Much existing RNA-seq data is available as bam files aligned to the human
178 genome. In order to bypass the most computationally expensive step of the pipeline,
179 I developed a method capable of using RNA-seq reads aligned with spliced aligners.
180 First, in order to avoid increased mismatch rates at splice junctions due to
181 alignment problems with both spliced and unspliced reads, I used samtools[30] and

182 awk to remove all alignments that don't align along the full length of the genome
183 (eg: for 76bp reads, only reads with a CIGAR flag of 76M). The remaining reads I re
184 trimmed (bamUtil , trimBam) to convert the first and last 10bp of each read to Ns
185 and set the quality strings to '!'. I then used samtools mpileup (-q30 –C50 –Q30) and
186 custom perl code to count the number of reads and number of errors at each
187 position in genome. Positions with too many errors (eg: more than one read of the
188 same non-reference base) I re not counted.
189

190 ## Measurement of error rates at splice junctions

191 I used the UCSC table browser[31] to download two bed files: hg19
192 EnsemblGenes introns with -10bp flanking from each side, and another file with the
193 introns and +10bp flanking on either side. I then used bedtools[32] (bedtools flank -b
194 20 -l 0 & bedtools flank -l 20 -b 0) to generate bed files with intervals that contain
195 the splicing donor and acceptor sites, respectively. In addition, I used bedtools
196 getfasta on the +10bp flanking bed file to keep only introns flanked by GT and AG
197 donor and acceptor sites. The final result is a pair of bam files with intervals
198 centered on the splicing donor or acceptor sites. I used this new bed file to count
199 error rates around each splice junction. The error rate at each position (eg: -10, -9, -
200 8, etc from the G at the 5' donor site) is the sum of all errors at that position, divided
201 by the sum of all reads. Positions are relative to the splicing feature, not to the
202 genome, as error rates at any single genomic position are dominated by sampling
203 bias. Per mono, di and tri-nucleotide background error rates I re calculated using
204 the same scripts, but without limiting mpileup to the splice junctions.

205

206 ## Strain construction and RNA sequencing for RPB9 and DST1 strains

207 The parental strain DBY12394[33] (GAL2+ s288c repaired HAP1, ura3Δ,
208 leu2Δ0::ACT1pr-Z3EV-NatMX) was transformed with a PCR product (KanMX-
209 Z3EVpr) to generate a genomically integrated inducible RPB9 (LCY143) or DST1
210 (LCY142). Correct transformants I re confirmed by PCR. To induce various levels of
211 expression, strains I re grown in YPD + 0,3,6,12 or 25nM β-estradiol (Sigma E4389)
212 for more than 12 hours to a final OD_{600} of 0.1 – 0.4. Cellular RNA was extracted using
213 the Epicenter MasterPure RNA Purification Kit, and Illumina sequencing libraries I
214 re prepared using the Truseq Stranded mRNA kit, and sequenced on a HiSeq2000
215 with at least 20,000,000 50bp sequencing reads per sample.
216
217 I used bwa[34] (-n 2, to permit no more than two mismatches in a read) to align the
218 yeast RNA-seq reads to the reference genome, and trimBam from bamUtil to mask
219 the first and last 10bp of each read. I used samtools mpileup[30] (-q 30 -d 100000 -
220 C50 –Q39) to count the number of reads and mismatches at each position in the
221 genome, discarding low confidence mapping, reads that map to multiple positions,
222 and low quality reads. Duplicate reads can be removed from the fastq file if the
223 coverage is low enough so that all unique read sequences are expected to come the
224 same RNA fragment; this is the case for low coverage paired-end reads with a
225 variable insert size, but not for very high coverage datasets or single-ended reads.
226

227 **pre-existing RNA-seq datasets.**

228 For the intron retention analysis in human cells, data were downloaded from
229 http://www.ncbi.nlm.nih.gov/bioproject/PRJNA253670 and reads mapped using
230 bwa. For RPB9 correlation, ENCODE[35] data (SRA PRJNA30709) from the Gingeras
231 lab at CSHL I re downloaded from NCBI SRA. Data for elc4 and spt4 were
232 downloaded from http://www.ncbi.nlm.nih.gov/bioproject/PRJNA167772 and
233 http://www.ncbi.nlm.nih.gov/bioproject/PRJNA148851, respectively.

234 Competing financial interests

235 The author declares no competing financial interests.

236 References

237 1. Lynch, M. The lower bound to the evolution of mutation rates. *Genome Biol*
238 *Evol* **3,** 1107–1118 (2011).
239 2. Zhu, Y. O., Siegal, M. L., Hall, D. W. & Petrov, D. A. Precise estimates of mutation
240 rate and spectrum in yeast. *PNAS* **111,** E2310–8 (2014).
241 3. Lang, G. I. & Murray, A. W. Estimating the per-base-pair mutation rate in the
242 yeast Saccharomyces cerevisiae. *Genetics* **178,** 67–82 (2008).
243 4. Gout, J.-F., Thomas, W. K., Smith, Z., Okamoto, K. & Lynch, M. Large-scale
244 detection of in vivo transcription errors. *PNAS* **110,** 18584–18589 (2013).
245 5. Lynch, M. Evolution of the mutation rate. *Trends Genet.* **26,** 345–352 (2010).
246 6. Shaw, R. J., Bonawitz, N. D. & Reines, D. Use of an in vivo reporter assay to test

247 for transcriptional and translational fidelity in yeast. *J. Biol. Chem.* **277,**
248 24420–24426 (2002).
249 7. de Mercoyrol, L., Corda, Y., Job, C. & Job, D. Accuracy of wheat-germ RNA
250 polymerase II. General enzymatic properties and effect of template
251 conformational transition from right-handed B-DNA to left-handed Z-DNA.
252 *Eur. J. Biochem.* **206,** 49–58 (1992).
253 8. Schwanhäusser, B. *et al.* Global quantification of mammalian gene expression
254 control. *Nature* **473,** 337–342 (2011).
255 9. Futcher, B., Latter, G. I., Monardo, P., McLaughlin, C. S. & Garrels, J. I. A
256 sampling of the yeast proteome. *Mol. Cell. Biol.* **19,** 7357–7368 (1999).
257 10. Fuhrmann, C. N., Halme, D. G., O'Sullivan, P. S. & Lindstaedt, B. A complete set
258 of nascent transcription rates for yeast genes. *PLoS ONE* **5,** e15442–249
259 (2010).
260 11. Hereford, L. M. & Rosbash, M. Number and distribution of polyadenylated
261 RNA sequences in yeast. *Cell* **10,** 453–462 (1977).
262 12. Islam, S. *et al.* Characterization of the single-cell transcriptional landscape by
263 highly multiplex RNA-seq. *Genome Res.* **21,** 1160–1167 (2011).
264 13. Gordon, A. J. E., Satory, D., Halliday, J. A. & Herman, C. Heritable change caused
265 by transient transcription errors. *PLoS Genet.* **9,** e1003595–e1003595 (2013).
266 14. Gordon, A. J., Satory, D., Halliday, J. A. & Herman, C. Lost in transcription:
267 transient errors in information transfer. *Curr. Opin. Microbiol.* **24C,** 80–87
268 (2015).
269 15. Irvin, J. D. *et al.* A genetic assay for transcription errors reveals multilayer

270 control of RNA polymerase II fidelity. *PLoS Genet.* **10,** e1004532 (2014).
271 16. Imashimizu, M., Oshima, T., Lubkowska, L. & Kashlev, M. Direct assessment of
272 transcription fidelity by high-resolution RNA sequencing. *Nucleic Acids Res.*
273 **41,** 9090–9104 (2013).
274 17. Kleinman, C. L. & Majewski, J. Comment on "Widespread RNA and DNA
275 sequence differences in the human transcriptome". *Science* **335,** 1302–author
276 reply 1302 (2012).
277 18. Pickrell, J. K., Gilad, Y. & Pritchard, J. K. Comment on "Widespread RNA and
278 DNA sequence differences in the human transcriptome". *Science* **335,** 1302–
279 author reply 1302 (2012).
280 19. Li, M. *et al.* Widespread RNA and DNA sequence differences in the human
281 transcriptome. *Science* **333,** 53–58 (2011).
282 20. Meacham, F. *et al.* Identification and correction of systematic error in high-
283 throughput sequence data. *BMC Bioinformatics* **12,** 451 (2011).
284 21. Tilgner, H. *et al.* Deep sequencing of subcellular RNA fractions shows splicing
285 to be predominantly co-transcriptional in the human genome but inefficient
286 for lncRNAs. *Genome Res.* **22,** 1616–1625 (2012).
287 22. Fox-Walsh, K. L. & Hertel, K. J. Splice-site pairing is an intrinsically high fidelity
288 process. *Proc. Natl. Acad. Sci. U.S.A.* **106,** 1766–1771 (2009).
289 23. Knippa, K. & Peterson, D. O. Fidelity of RNA polymerase II transcription: Role
290 of RPB9 in error detection and proofreading. *Biochemistry* **52,** 7807–7817
291 (2013).
292 24. Lykke-Andersen, J. mRNA quality control: Marking the message for life or

293 death. *Curr. Biol.* **11,** R88–91 (2001).

294 25. Nesser, N. K., Peterson, D. O. & Hawley, D. K. RNA polymerase II subunit Rpb9
295 is important for transcriptional fidelity in vivo. *Proc. Natl. Acad. Sci. U.S.A.* **103,**
296 3268–3273 (2006).

297 26. Walmacq, C. *et al.* Rpb9 Subunit Controls Transcription Fidelity by Delaying
298 NTP Sequestration in RNA Polymerase II. *J. Biol. Chem.* **284,** 19601–19612
299 (2009).

300 27. Kireeva, M. L. *et al.* Transient reversal of RNA polymerase II active site closing
301 controls fidelity of transcription elongation. *Mol. Cell* **30,** 557–566 (2008).

302 28. McIsaac, R. S., Gibney, P. A., Chandran, S. S., Benjamin, K. R. & Botstein, D.
303 Synthetic biology tools for programming gene expression without nutritional
304 perturbations in Saccharomyces cerevisiae. *Nucleic Acids Res.* **42,** e48–e48
305 (2014).

306 29. McIsaac, R. S., Oakes, B. L., Botstein, D. & Noyes, M. B. Rapid synthesis and
307 screening of chemically activated transcription factors with GFP-based
308 reporters. *Journal of visualized experiments : JoVE* e51153–e51153 (2013).

309 30. Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics*
310 **25,** 2078–2079 (2009).

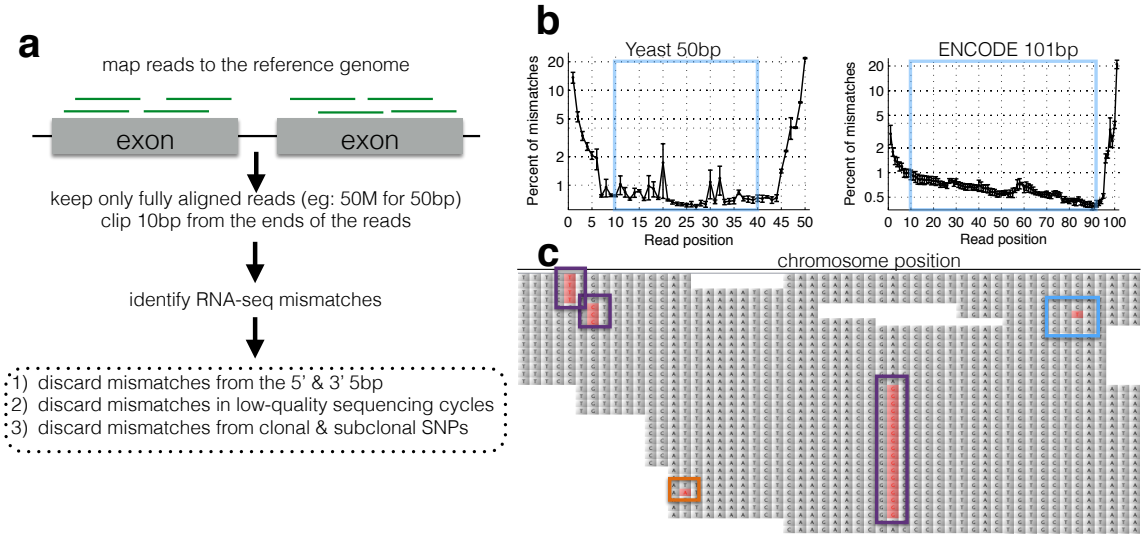
311 31. Karolchik, D. *et al.* The UCSC Table Browser data retrieval tool. *Nucleic Acids*
312 *Res.* **32,** D493–6 (2004).

313 32. Quinlan, A. R. & Hall, I. M. BEDTools: a flexible suite of utilities for comparing
314 genomic features. *Bioinformatics* **26,** 841–842 (2010).

315 33. McIsaac, R. S. *et al.* Synthetic gene expression perturbation systems with

316 rapid, tunable, single-gene specificity in yeast. *Nucleic Acids Res.* **41,** e57–e57
317 (2013).
318 34. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-
319 Wheeler transform. *Bioinformatics* **25,** 1754–1760 (2009).
320 35. ENCODE Project Consortium. An integrated encyclopedia of DNA elements in
321 the human genome. *Nature* **489,** 57–74 (2012).
322 36 Jung, H. et al. Intron retention is a widespread mechanism of tumor-
323 suppressor inactivation. Nat. Genet. (2015).
324 37. Dhir, A. *et al.* Microprocessor mediates transcriptional termination of long
325 noncoding RNA transcripts hosting microRNAs. Nat. Struct. Mol. Biol. 22, 319–
326 327 (2015).
327 38. Walmacq, C. et al. Rpb9 subunit controls transcription fidelity by delaying
328 NTP sequestration in RNA polymerase II. J. Biol. Chem. 284, 19601–19612
329 (2009).
330
331
332

333 ## Figures



334
335 **Figure 1. A computational framework to measure relative changes in RNA**

336 **polymerase fidelity. (a)** Pipeline to identify potential RNA polymerase errors in

337 RNA-seq data. High quality full-length RNA-seq reads are mapped to the reference

338 genome or transcriptome using bwa, and only reads that map completely with two

339 or fewer mismatches are kept. **(b)** Then 10bp from the front and 10bp from the end

340 of the read are discarded as these regions have high error rates and are prone to

341 poor quality local alignments. **(c)** Errors that occur multiple times (purple boxes)

342 are discarded, as these are likely due to sub-clonal DNA mutations or sequences that

343 sequence poorly on the HiSeq. Unique errors in the middle of reads (cyan box) are

344 kept and counted.

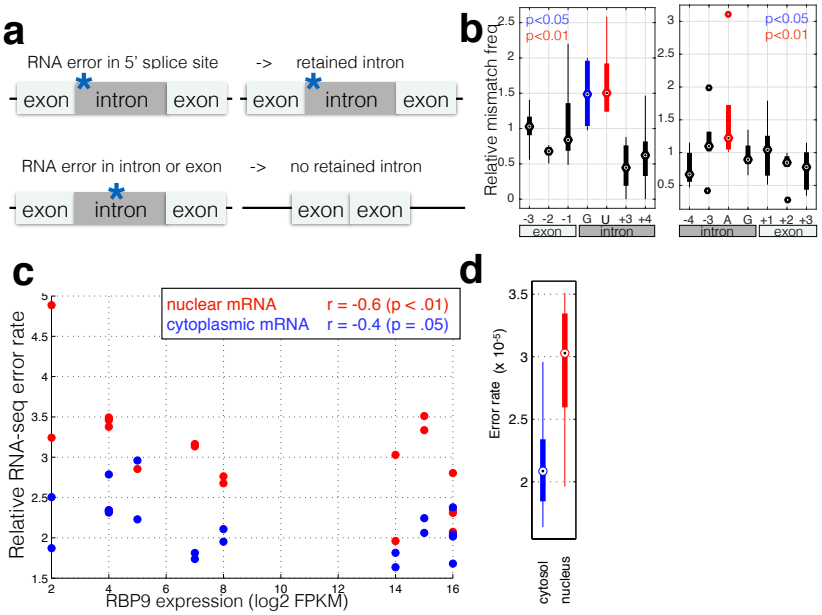
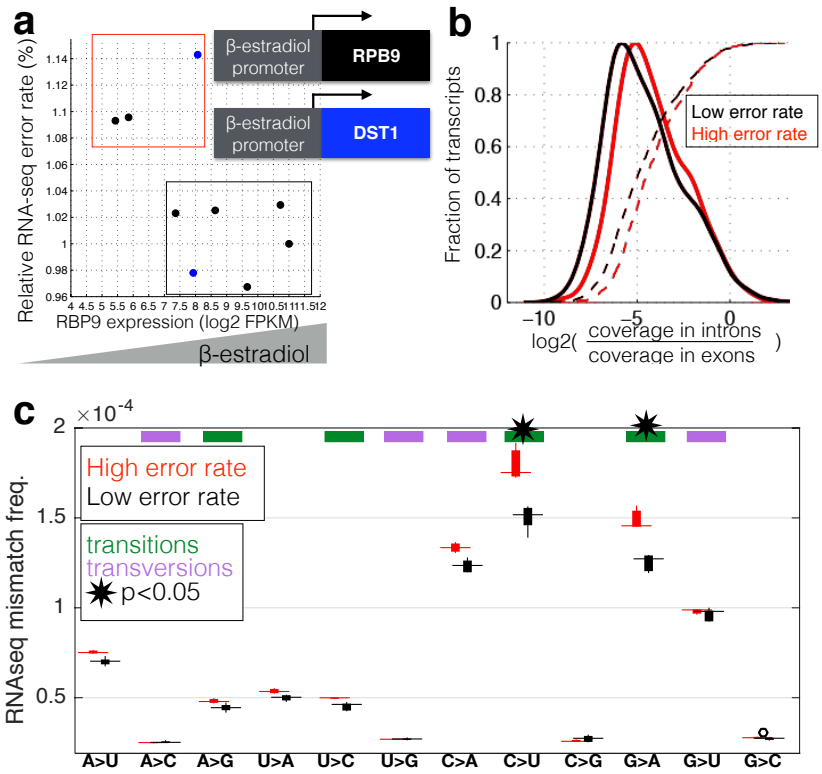


Figure 2. RNA polymerase errors cause intron retention and error rates are correlated with RPB9 expression. (a) RNA polymerse errors at the splice junction should result in intron retention, as DNA mutations at the 5' donor site are known to cause intron retention. **(b)** Shown are the RNA-seq mismatch rates at each position relative to the 5' donor splice site, for sequencing reads that span an exon-intron junction. Mismatch rates from chromatin associated and nuclear RNAs are higher at the 5' and 3' splice sites, suggesting that RNA polymerase errors at this site result in intron retention. **(c)** For all ENCODE cell lines, RPB9 expression was determined from whole-cell RNA-seq data, and the RNA-seq error rate was measured separately for the cytoplasmic and nuclear fractions. **(d)** The RNA-seq error rate is higher (paired t-test, p=0.0019) in the nuclear than the cytoplasmic fraction, suggesting that quality control mechanism may block nuclear export of low quality mRNAs.

358

359



360

361 **Figure 3. RNA polymerase error rate is determined by the expression level of**

362 **RPB9 and DST1. (a)** RNA-seq error rates I re measured for two strains (Z_3EVpr-

363 RPB9, black points , Z_3EVpr-DST1, blue points) grown at different concentrations of

364 β-estradiol. The points show the relationship between RPB9 expression levels

365 (determined by RNA-seq) and RNA-seq error rates. The blue points show RPB9

366 expression levels for the Z_3EVpr-DST1 strain, in which DST1 expression ranges from

367 16 FPKM at 0nM β-estradiol to 120 FPKM native expression to 756 FPKM at 25nM

368 β-estradiol. Low induction of both DST1 or RPB9 results in high RNA-seq error rates

369 (red box), while wild-type and higher induction levels result low RNA-seq error

370 rates (black box). **(b)** Across all genes, the intron retention rate is higher in

371 conditions with low RNA polymerase fidelity (t-test between high and low error rate

372 samples, p=0.029), consistent with the hypothesis that RNA polymerase errors

373 result in splicing defects. **(c)** The error rate for each of the 12 single base changes

374 are shown for induction experiments that gave high (red) or low (black) RNA-seq

375 error rates. Transitions (G<->A , C<->U) are marked with green boxes and

376 transversions (A<->C , G<->U) with purple

377

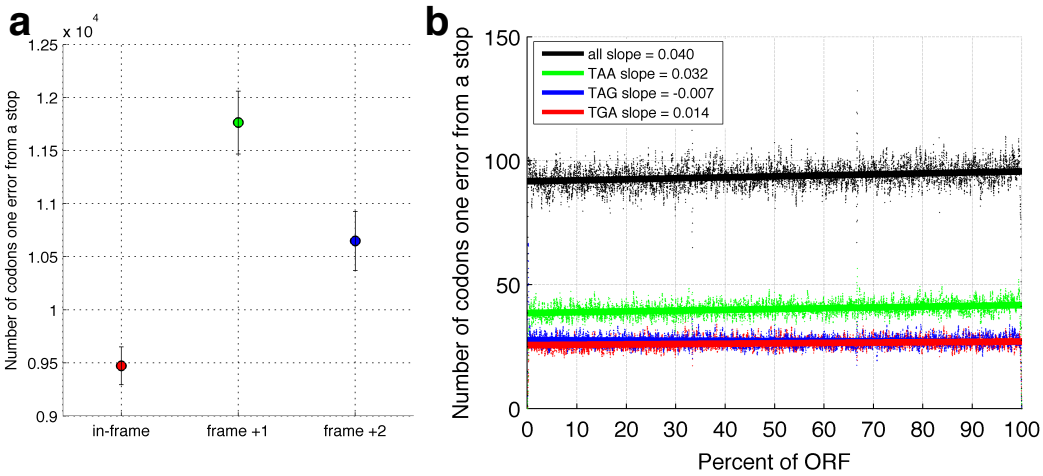
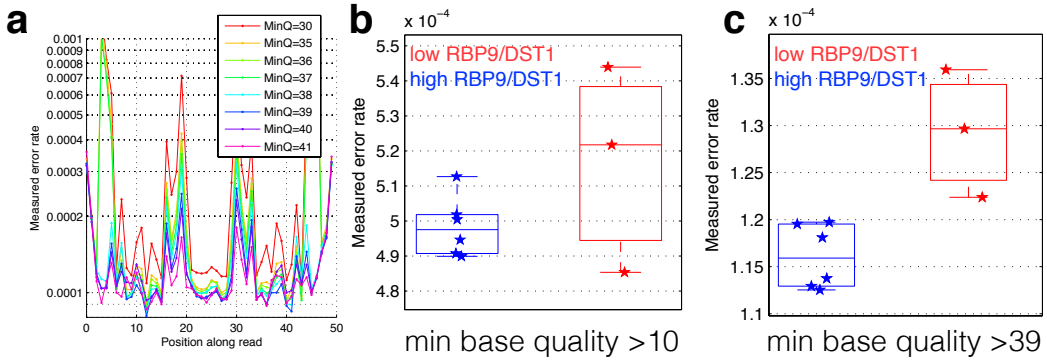


Figure 4. In-frame stop codons are less likely to be created by polymerase errors. For all genes in yeast, I calculated the number of codons which are one polymerase error from a stop codon. **(a)** Fewer in-frame codons can be turned into a stop codon by a single nucleotide change, compared to out-of-frame codons. **(b)** Codons that are one error away from generating an in-frame stop codon are more likely to be found at the ends of ORFs, compared to the beginning of the ORF.

387



388

389 **Figure 1 – figure supplement 1. Cycle-specific error rates and better**
390 **differentiation of genetically determined error rates using base quality value**
391 **cutoffs.** Six yeast RNA-cDNA libraries were sequenced on the same lane in a HiSeq.
392 **(a)** The average mismatch rate (across the six cDNA libraries) to the reference
393 genome at each position was determined using different minimum base-quality
394 thresholds using GATK ErrorRatePerCycle. Independent of the quality threshold,
395 cycles at the ends, as well as some cycles in the middle, have high error rates. **(b)**
396 The measured error rate for each sample using a minimum base quality of 10. **(c)**
397 The measured error rate for each sample using a minimum base quality of 39.

398

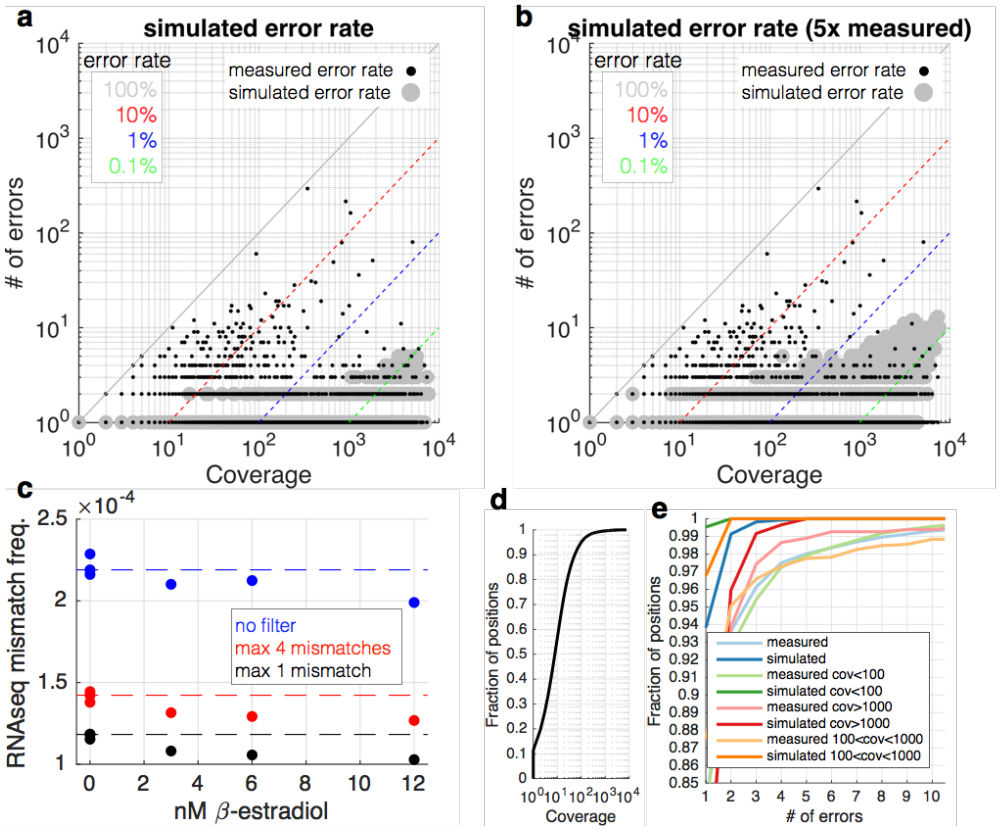


Figure 1 – figure supplement 2. RNA-seq data are enriched for mismatches to the reference genome that occur far more often than expected.

(a) At each coverage (x-axis), a point is shown if there is any positions in the genome with the observed number of errors (y-axis) (small black dots). The diagonal lines show mismatch frequencies of 100%, 10%, 1% and 0.1% — any point falling on these lines has the given mismatch frequency. With large grey circles are shown simulated data in which the same coverage as the yeast RNA-seq data are used, but with a mismatch frequency identical to the measured overall mismatch frequency of the yeast data. Locations in the graph in which a black point occurs but there is no grey point are locations in which there are more mismatches than expected by change. Note that at a coverage of less than 100, we expect to see no

411 mismatches more than twice, and 0.5% of positions with 2 observances of identical
412 mismatches. **(b)** Identical to (a) but with the simulated mismatch frequency 5x the
413 observed. **(c)** Shown are measured mismatch frequencies for the yeast RPB9 and
414 DST1 induction data at different B-estradiol concentrations, at different filters for
415 the maximal allowed number of observed identical mismatches. The dashed lines
416 show the average mismatch frequency for the 0nM condition. For all filters, low B-
417 estradiol conditions have higher RNA-seq mismatch frequencies. **(d)** The coverage
418 of the yeast RNA-seq data; ~95% of the genome is covered by less than 100 reads.
419 **(e)** Shown are the fraction of positions in the genome (y-axis) with X errors (x-axis)
420 for the yeast RNAseq data (cyan) and simulated data (blue). Also shown are the
421 same data for positions of the genome with different coverage. For positions
422 covered by less than 1000 reads (95% of the genome) the expectation is 0 or 1
423 sequence mismatch (blue and orange lines). Positions with far greater numbers of
424 mismatches are likely due to sub-clonal mutations and technical bias.
425

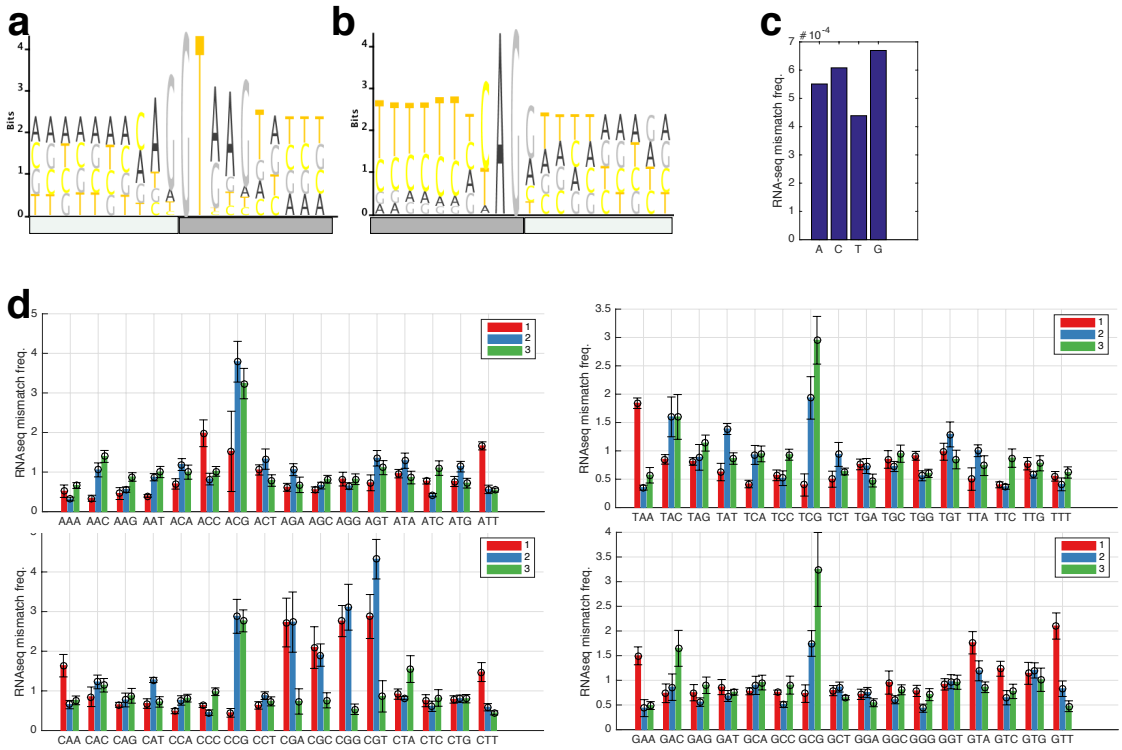


Figure 2 – figure supplement 1. RNA-seq mismatch rates for all trinuculeotides in chromatin associated and nuclear RNAs. (a,b) The 5' and 3 splicing motifs in the human genome. **(c)** The RNA-seq mismatch frequencies for all single nucleotides. **(d)** The RNA-seq mismatch rate to the reference genome for each trinucleotide, normalized to the average mismatch rate across all trinucleotides. For each trinucleotide, red shows the mismatch frequency at the first base, blue at the second, and green at the third. Error bars are standard deviation across all samples.

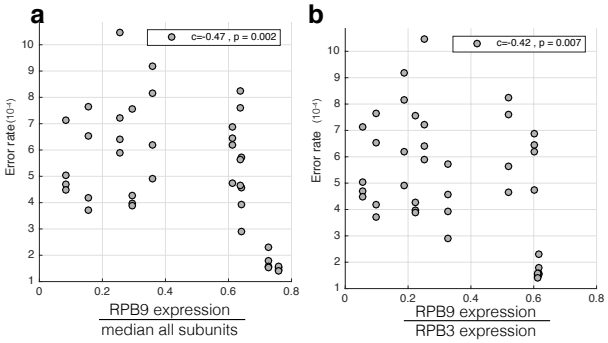


Figure 2 – figure supplement 2. RBP9 expression negatively correlates with RNA-seq mismatch rates. The mismatch frequency is shown across all cells lines. **(a)** RPB9 mRNA expression is normalized by the median expression level of all subunits. **(b)** RPB9 mRNA expression is normalized by RBP3 (POLR2C) expression.

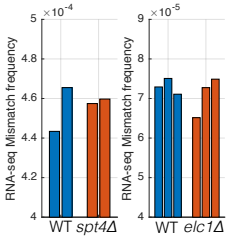


Figure 3 – figure supplement 1. Mutations that affect transcription elongation do not affect measured RNA-seq mismatch frequencies. Two separate experiments were performed with wild-type controls and mutants involved in transcription elongation. Individual bars show the RNA-seq mismatch frequency of biological replicates.