

1 LATENCY, EXPRESSION AND SPLICING DURING INFECTION WITH HIV
2 Scott Sherrill-Mix
3 A DISSERTATION
4 in
5 Genomics and Computational Biology
6 Presented to the Faculties of the University of Pennsylvania
7 in
8 Partial Fulfillment of the Requirements for the
9 Degree of Doctor of Philosophy
10 2015

11 Supervisor of Dissertation:

12 Frederic D. Bushman, Ph.D., Professor of Microbiology

13 Graduate Group Chairperson:

14 Li-San Wang, Ph.D., Associate Professor of Pathology and Laboratory Medicine

15 Dissertation Committee:

16 Nancy Zhang, Ph.D. Associate Professor of Statistics

17 Yoseph Barash, Ph.D., Assistant Professor of Genetics

18 Kristen Lynch, Ph.D., Professor of Biochemistry and Biophysics

19 Michael Malim, Ph.D., Professor of Infectious Diseases, King's College London

²⁰ LATENCY, EXPRESSION AND SPLICING DURING INFECTION WITH HIV

²¹ © COPYRIGHT

²² 2015

²³ Scott A. Sherrill-Mix

²⁴ This work is licensed under the

²⁵ Creative Commons Attribution

²⁶ NonCommercial-ShareAlike 3.0

²⁷ License

²⁸ To view a copy of this license, visit

²⁹ <http://creativecommons.org/licenses/by-nc-sa/3.0/>

Dedicated to William Maurer, Gayle Maurer & Michele Sherrill-Mix

ACKNOWLEDGEMENTS

32 I would like to thank

33 Rick Bushman

34 Christian Hoffmann wet lab

35 collaborators

³⁶ Bushman lab Rithun Mukherjee, Karen Ocwieja, Nirav Malani, Troy Brady, Young Hwang,
³⁷ Brendan Kelly, Kyle Bittinger, Rebecca Custers-Allen, Serena Dollive, Frances Male, Jacque
³⁸ Young, Rohini Sinha, Sam Minot, Aubrey Bailey, Christopher Nobles, Stephanie Grunberg,
³⁹ Vesa Turkki, Anatoly Dryga, Eric Sherman, Greg Peterfreund, Yinghua Wu, Alice Laughlin,
⁴⁰ Sesh Sundararaman, Alexandra Bryson, Christel Chehoud, Erik Clarke, Arwa Abbas

41 My committee—Nancy Zhang, Yoseph Barash, Kristen Lynch and Michael Malim—have
42 provided guidance and encouragement. Many faculty of GCB mentoring and teaching.
43 Hannah Chervitz, Tiffany Barlow, Mali Skotheim, Caitlin Greig and Laurie Zimmerman
44 for managing everything and helping manage the layers of bureaucracy. Funding from the
45 HIV Immune Networks Team (HINT) consortium P01 AI090935 and NRSA computational
46 genomics training grant T32 HG000046.

⁴⁷ Ram Myers and Mike James for great previous mentoring.

48 Xiaofen and Otto

49 . . .

ABSTRACT

51 LATENCY, EXPRESSION AND SPLICING DURING INFECTION WITH HIV

52 Scott Sherrill-Mix

53 Frederic D. Bushman, Ph.D.

54 Over 35 million people are living with human immunodeficiency virus (HIV-1). The
55 mechanisms causing integrated provirus to become latent, the diversity of spliced viral
56 transcripts and the cellular response to infection are not fully characterized and hinder the
57 eradication of HIV-1. We applied high-throughput sequencing to investigate the effects of
58 host chromatin on proviral latency and variation of expression and splicing in both the host
59 and virus during infection.

60 To evaluate the link between host chromatin and proviral latency, we compared genomic and
61 epigenetic features to HIV-1 integration site data for latent and active provirus from five cell
62 culture models. Latency was associated with chromosomal position within individual models.
63 However, no shared mechanisms of latency were observed between cell culture models. These
64 differences suggest that cell culture models may not completely reflect latency in patients.

65 We carried out two studies to explore mRNA populations during HIV infection. Single-
66 molecule amplification and sequencing revealed that the clinical isolate HIV_{89.6} produces at
67 least 109 different spliced mRNAs. Viral message populations differed between cell types,
68 between human donors and longitudinally during infection. We then sequenced mRNA
69 from control and HIV_{89.6}-infected primary human T cells. Over 17 percent of cellular genes
70 showed altered activity associated with infection. These gene expression patterns differed
71 from HIV infection in cell lines but paralleled infections in primary cells. Infection with
72 HIV_{89.6} increased intron retention in cellular genes and abundance of RNA from human
73 endogenous retroviruses. We also quantified the frequency and location of chimeric HIV-host
74 RNAs. These two studies together provided a detailed accounting of both HIV_{89.6} and host

75 expression and alternative splicing.

76 A more cost-effective method of detecting viral load would aid patients with poor access to
77 healthcare. We developed improved methods for assaying HIV-1 RNA using loop-mediated
78 isothermal amplification based on primers targeting regions of the HIV-1 genome conserved
79 across subtypes. Combined with lab-on-a-chip technology, these techniques allow quantitative
80 measurements of viral load in a point-of-care device targeted to resource-limited settings.

81 This work disclosed novel HIV-host interactions and developed techniques and knowledge
82 that will aid in the study and management of HIV-1 infection.

TABLE OF CONTENTS

84	ABSTRACT	v
85	TABLE OF CONTENTS.....	viii
86	LIST OF TABLES	ix
87	LIST OF ILLUSTRATIONS	x
88	CHAPTER 1 : Introduction	1
89	1.1 Impact of HIV	1
90	1.2 The HIV virus	1
91	1.3 Integration and latency	1
92	1.4 HIV splicing.....	1
93	1.5 Host cell interactions	1
94	1.6 HIV detection	1
95	1.7 Contribution summaries	2
96	CHAPTER 2 : HIV latency and integration site placement in five cell-based models	3
97	2.1 Abstract	3
98	2.2 Background	4
99	2.3 Methods	6
100	2.4 Results.....	10
101	2.5 Conclusions	21
102	2.6 Availability of supporting data	23
103	2.7 Author's contributions	23
104	2.8 Acknowledgements	24
105	CHAPTER 3 : Dynamic regulation of HIV-1 mRNA populations analyzed by single-molecule enrichment and long-read sequencing	25
107	3.1 Abstract	25
108	3.2 Introduction.....	26
109	3.3 Materials and methods.....	28
110	3.4 Results.....	35
111	3.5 Discussion	41
112	3.6 Acknowledgements	45
113	CHAPTER 4 : Gene activity in primary T cells infected with HIV _{89.6} : intron retention and induction of distinctive genomic repeats	46
115	4.1 Abstract	46
116	4.2 Background	47
117	4.3 Methods	48
118	4.4 Results.....	52

119	4.5	Discussion	70
120	4.6	Conclusions	74
121	4.7	Availability of supporting data	75
122	4.8	Acknowledgements	75
123	CHAPTER 5 : A reverse transcription loop-mediated isothermal amplification assay		
124	optimized to detect multiple HIV subtypes		76
125	5.1	Abstract	76
126	5.2	Introduction.....	77
127	5.3	Methods	78
128	5.4	Results.....	80
129	5.5	Testing different primer designs	84
130	5.6	Discussion	88
131	5.7	Acknowledgments.....	90
132	CHAPTER 6 : Conclusions and future directions.....		91
133	6.1	Latency and integration location	91
134	6.2	HIV-1 alternative splicing.....	92
135	6.3	Host expression during HIV infection	95
136	6.4	LAMP PCR and lab-on-a-chip	96
137	APPENDICES		100
138	A.1	Generalized linear models of changes in use of mutually exclusive HIV-1 splice	
139		acceptors	100
140	A.2	Reproducible report of HIV integration sites and latency analysis	106
141	BIBLIOGRAPHY		448

LIST OF TABLES

¹⁴³	TABLE 2.1 : Integrations from <i>in vitro</i> models of latency	9
¹⁴⁴	TABLE 2.2 : Genomic data available for comaprison to integration sites	11
¹⁴⁵	TABLE 4.1 : Samples and RNA-Seq sequencing coverage	53
¹⁴⁶	TABLE 4.2 : Data used for meta-analysis of expression changes in HIV	54

LIST OF ILLUSTRATIONS

148	FIGURE 2.1 : Correlations of genomic features and latency	12
149	FIGURE 2.2 : Lasso regressions predicting latency	13
150	FIGURE 2.3 : Cellular expression and latency	15
151	FIGURE 2.4 : Strand orientation and latency	16
152	FIGURE 2.5 : Genes and latency	17
153	FIGURE 2.6 : Alphoid repeats and latency	18
154	FIGURE 2.7 : Acetylation and latency	20
155	FIGURE 2.8 : Shared expression status between near neighbors.....	20
156	FIGURE 3.1 : Mapping the splice donors and acceptors of HIV _{89.6}	27
157	FIGURE 3.2 : Spliced transcripts produced from HIV _{89.6}	34
158	FIGURE 3.3 : Novel transcripts utilizing acceptor A8c	39
159	FIGURE 3.4 : Temporal, cell type and donor variability in accumulation of HIV-1 messages.....	42
160		
161	FIGURE 4.1 : Comparisons among studies quantifying cellular gene expression after HIV infection	55
162		
163	FIGURE 4.2 : Comparisons of the effect of HIV infection on gene expression to studies comparing subsets of immune cells	57
164		
165	FIGURE 4.3 : Changes in the abundance of intronic regions with HIV infection ...	59
166		
167	FIGURE 4.4 : Repeat categories enriched upon infection with HIV	61
168		
169	FIGURE 4.5 : Characteristics of LTR12C sequences associated with induction upon infection with HIV _{89.6}	62
170		
171	FIGURE 4.6 : Estimating relative abundance of HIV _{89.6} message size classes using RNA-Seq data	64
172		
173	FIGURE 4.7 : Transcription and splicing of the HIV _{89.6} RNA	66
174		
175	FIGURE 4.8 : Chimeric RNA sequences containing both human and HIV sequences	70
176		
177		
178	FIGURE 5.1 : Amplification results for all RT-LAMP primer sets tested	82
174	FIGURE 5.2 : Subtype-agnostic RT-LAMP primers design	83
175		
176	FIGURE 5.3 : Performance of the AceIN-26 primer set with different starting RNA concentrations	86
177		
178	FIGURE 5.4 : Replicate tests of the ACeIN-26 primer set over six HIV subtypes..	87
178	FIGURE 6.1 : Ebola RT-LAMP primers design	98

CHAPTER 1: Introduction

180 **1.1 Impact of HIV**

181 **1.2 The HIV virus**

182 **1.3 Integration and latency**

183 **1.4 HIV splicing**

184 **1.5 Host cell interactions**

185 **1.6 HIV detection**

186 Immunoassays provide cheap immediate testing of HIV infection in patients. These tests are
187 based on the enzyme-linked immunosorbent assay (ELISA), using an enzyme linked to an
188 antibody to produce a detectable signal in the presence of antigen^{1–3}.

189 The isolation of HIV^{4–7} allowed the production of large quantities of virions. These virions
190 were bound to a substrate, sera from patients added and any patients antibodies sensitive
191 to HIV allowed to bind. Any unbound antibodies were washed away. Then a peroxidase
192 enzyme-labeled antibody targeted to human antibody bound was added, allowed to bind
193 and the unbound antibodies again washed away. Any HIV-targeted patient antibodies
194 would bind the antigen and be bound in turn by the peroxidase-labeled antibody and
195 the peroxidase would then change the color of media^{8,9}. These tests had a large false
196 positive rate and the standard procedure was to perform multiple ELISA tests follow by a
197 Western blot test^{10,11} but false positives were still prevalent¹². More conservative criteria
198 and cleaner lab procedures reduced false positives¹³. These assays have been developed
199 to fourth generation¹⁴ with more sensitive and specific detection of patient antibodies and
200 earlier detection using antibodies against the HIV capsid protein^{15,16}.

201 Slightly less specific but rapid immunoassays providing results in 30 minutes have been

202 developed to allow point-of-care testing with many fewer patients lost to follow up prior
203 to delivery of results^{17–19}. Rapid tests detecting HIV in oral fluids have been developed
204 and obviate the need for a blood draw^{20–22}. These rapid care tests allow self testing at
205 home^{23,24}.

206 Reverse transcription and PCR amplification offers another alternative^{25,26} but is not
207 currently cost effective for primary patient screening²⁷.

208 Tests allowing point-of-care qualitative HIV detection are now widespread but point-of-care
209 assays for viral load in a patient exist. In addition, existing laboratory-based tests are
210 relatively expensive and require specialized equipment making access difficult in resource-
211 limited settings^{28,29}. Without viral load measures, CD4⁺ T cell counts or clinical presentation
212 are used to infer the emergence of drug. These criteria are not specific or sensitive enough
213 without viral load measures so many patients are unnecessarily switched to second line
214 therapy^{30,31} or switched too late leading to accumulations of drug resistant mutations³². In
215 Chapter 5, we design loop-mediated isothermal amplification methods that can be used with
216 microfluidics to create a point-of-care assay of infection and viral load in resource-limited
217 settings.

218 1.7 Contribution summaries

219 Much of this work was performed as part of a large collaboration. It would not tell a
220 complete story in isolation. Therefore, I have preserved the chapters in published form in
221 and detailed my contribution to each project at the start of the chapter.

CHAPTER 2: HIV latency and integration site placement in five cell-based models

This chapter was originally published as:

S Sherrill-Mix, MK Lewinski, M Famiglietti, A Bosque, N Malani, KE Ocieja, CC Berry, D Looney, L Shan et al. 2013. HIV latency and integration site placement in five cell-based models. *Retrovirology*, 10:90. doi: 10.1186/1742-4690-10-90

I led the computational analysis, with assistance from CC Berry and N Malani. MK Lewinski, D Looney and J Guatelli analyzed integration sites using IonTorrent sequencing. M Famiglietti, A Bosque and V Planelles prepared DNA from latent and activated T cells using the Central Memory CD4 + model. L Shan, RF Siliciano, MJ Pace, LM Agosto, KE Ocwieja and U O'Doherty contributed data and suggestions. FD Bushman and I planned the overall study. I produced the figures. FD Bushman and I wrote the paper.

Additional files are available at [http://www.retrovirology.com/
content/10/1/90/additional](http://www.retrovirology.com/content/10/1/90/additional)

225 2.1 Abstract

Background: HIV infection can be treated effectively with antiretroviral agents, but the persistence of a latent reservoir of integrated proviruses prevents eradication of HIV from infected individuals. The chromosomal environment of integrated proviruses has been proposed to influence HIV latency, but the determinants of transcriptional repression have not been fully clarified, and it is unclear whether the same molecular mechanisms drive latency in different cell culture models.

Results: Here we compare data from five different *in vitro* models of latency based on primary human T cells or a T cell line. Cells were infected *in vitro* and separated into

234 fractions containing proviruses that were either expressed or silent/inducible, and integration
235 site populations sequenced from each. We compared the locations of 6,252 expressed
236 proviruses to those of 6,184 silent/inducible proviruses with respect to 140 forms of genomic
237 annotation, many analyzed over chromosomal intervals of multiple lengths. A regularized
238 logistic regression model linking proviral expression status to genomic features revealed no
239 predictors of latency that performed better than chance, though several genomic features
240 were significantly associated with proviral expression in individual models. Proviruses in the
241 same chromosomal region did tend to share the same expressed or silent/inducible status if
242 they were from the same cell culture model, but not if they were from different models.

243 Conclusions: The silent/inducible phenotype appears to be associated with chromosomal
244 position, but the molecular basis is not fully clarified and may differ among *in vitro* models
245 of latency.

246 2.2 Background

247 Highly active antiretroviral therapy (HAART) can suppress HIV-1 replication in infected
248 patients, but the ability of HIV to persist as an inducible reservoir of latent proviruses^{34–36}
249 obstructs eradication of the virus and functional cure³⁷. These latent proviruses are long
250 lived^{38,39} and relatively invisible to the immune system^{35,40}. The potential for even a single
251 virus to restart infection despite successful antiviral therapy means that it may be necessary
252 to eliminate all latent proviruses to eradicate HIV from an infected person.

253 After integration, a positive feedback loop of Tat transactivation appears to partition
254 proviral gene activity into either of two stable states^{41–43}—abundant Tat driving high
255 proviral expression or little Tat leading to quiescent latency. Similar to the positional effect
256 variegation observed in fruit fly chromosomal rearrangements^{44,45}, studies on cell clones
257 with single integrations show that differing integration sites can have large differences in
258 proviral expression^{46–48}. These data suggest that integration site location, along with the
259 cellular environment^{48–51}, influences the balance between latency and proviral expression.

260 Associations between latency and genomic features have also been reported in collections of
261 integration sites from cell culture models although the consistency of these effects across
262 model systems and their relationships to latency in patients remains uncertain. Lewinski
263 et al.⁵² reported that proviruses integrated in gene deserts, alphoid repeats and highly
264 expressed genes are more likely to have low expression. Shan et al.⁵³ reported an association
265 between latency and integration in the same transcriptional orientation as host genes. Pace
266 et al.⁵⁴ found that silent and expressed provirus integration sites differed in the abundance
267 and expression levels of nearby genes, GC content, CpG islands and alphoid repeats. In
268 model systems with defined integration sites, Lenasi et al.⁵⁵ reported decreased and Han
269 et al.⁵⁶ reported increased viral transcription when the provirus is downstream of a highly
270 expressed host gene.

271 Cell-based models of latency are important for many aspects of HIV research, including
272 screening small molecules that can reverse latency and potentially allow eradication^{57,58}.
273 Location-driven differences in expression are preserved even after demethylation and histone
274 deacetylase treatment⁴⁶, which suggests that integration location has the potential to
275 confound “shock and kill” anti-latency treatments^{59,60}. A greater understanding of the
276 effects of integration site location on latency could thus affect antiretroviral development.

277 To search for features of integration site associated with latency, we generated a set of
278 inducible and expressed integration sites using a primary central memory CD4⁺ T cell model
279 of latency^{61,62}, collected four previously reported integration site datasets and modeled
280 the effects of genomic features near the integration site on the expression status of these
281 proviruses. Although some genomic features associated with latency in individual models,
282 no feature was consistently associated with proviral expression across all five cell culture
283 models. However, closely neighboring proviruses within the same cellular model shared the
284 same latency status much more often than expected by chance suggesting that chromosomal
285 position of integration affects latency but that the mechanism remains unclear or differs
286 between cell culture models. Thus these data help inform the design of experiments in HIV

287 eradication research.

288 **2.3 Methods**

289 **2.3.1 Integration sites**

290 Naive CD4⁺ T cells were purified by negative selection from peripheral blood mononuclear
291 cells. The cells were activated with anti-CD3 and anti-CD28 (+TGF-beta, anti-IL-12, and
292 anti-IL-4) to generate “non-polarized” cells (the in vitro equivalent of central memory T
293 cells). Five days after isolation, cells were infected with an NL4-3-based virus with GFP in
294 place of Nef and the LAI envelope (X4) provided in trans at a concentration of 500 ng of
295 p24 as measured by ELISA per million cells. Based on previous experience with this model,
296 this amount of p24 should produce an MOI of approximately 0.15. Cells were cultured
297 in the presence of IL-2. Two days post-infection, cells were sorted for GFP+; this active
298 population expresses GFP even when treated with flavopiridol, although for this study they
299 were not treated. The inducible population was the set of GFP negative cells from the initial
300 sort that, 9 days post-infection, were activated with anti-CD3 and anti-CD28 and sorted for
301 GFP production.

302 Genomic DNA from the inducible and expressed populations was digested with MseI, ligated
303 to an adapter, and amplified by ligation-mediated PCR essentially as in Wu et al.⁶³ and
304 Mitchell et al.⁶⁴ except that the nested PCR primers included sequence for the Ion Torrent
305 P1 adapter and adapter A sequence with a 5 base barcode sequence specific to the inducible
306 or expressed conditions. Amplicons were sequenced using an Ion Torrent Personal Genome
307 Machine (PGM) according to manufacturer’s instructions using an Ion 316 chip and the Ion
308 PGM 200 Sequencing kit (Life Technologies). The sequence reads were sorted into samples
309 by barcode. All reads were required to match the expected 5' sequence with a Levenshtein
310 edit distance less than 3 from the expected barcode, 5' primer and HIV long terminal repeat
311 (LTR). The 5' primer and HIV sequence, along with the 3' primer if present, were trimmed
312 from the read. Sequences with less than 24 bases remaining or containing any eight base

313 window with an average quality less than 15 were discarded. Duplicate reads and reads
314 forming an exact substring of a longer read were removed.

315 **2.3.2 Analysis**

316 All statistical analysis was performed in R 2.15.2⁶⁵. The analyses are described in a
317 reproducible report (Appendix A.2). The annotated integration site data necessary to
318 perform the analyses and the compilable code to generate this reproducible report are
319 provided as supplemental information³³. The new Central Memory CD4⁺ data set was
320 analyzed as in Berry et al.⁶⁶. The integration patterns appeared similar to previously
321 reported HIV integration site datasets⁶⁷.

322 **2.3.3 Previously published data**

323 We collected integration sites from three previously reported studies (Table 2.1), for a total
324 of four expressed versus silent/inducible pairs of samples. These studies used primary CD4⁺
325 T cells or Jurkat cells infected with HIV or HIV-derived constructs as cell culture models of
326 latency. Flow cytometry allowed cells expressing viral encoded proteins to be sorted from
327 non-expressing cells. In two of the studies, these non-expressing populations were stimulated
328 to ensure that the provirus could be aroused from latency. Specific differences in protocol
329 between the study sets are summarized below.

330 **Jurkat** Lewinski et al.⁵² infected Jurkat cells with a VSV-G pseudotyped, GFP-expressing
331 pEV731 HIV construct (LTR-Tat-IRES-GFP)⁴⁶ at an MOI of 0.1. The cells were
332 sorted into GFP+ and GFP- two to four days after infection. GFP+ cells were sorted
333 again two weeks after infection and cells that were again GFP+ were collected for
334 integration site sequencing. GFP- cells were sorted for GFP negativity twice more
335 then stimulated with TNF α . Cells that were GFP+ after stimulation were collected
336 for integration site sequencing. DNA was digested with MseI or a combination of NheI,
337 SpeI and XbaI, ligated to adapters for nested PCR, amplified and sequenced by Sanger
338 capillary electrophoresis.

339 **Bcl-2 transduced CD4⁺** Shan et al.⁵³ transduced CD4⁺ T cells with Bcl-2, costimulated
340 with bound anti-CD3 and soluble anti-CD28 antibodies, interleukin-2 and T cell growth
341 factor and then infected with X4-pseudotyped GFP-expressing NL4-3- δ 6-drEGFP
342 construct⁶⁸ at an MOI of less than 0.1. DNA was extracted, digested with PstI and
343 circularized⁶⁹. HIV-human junctions were amplified by reverse PCR and sequenced
344 using Sanger capillary electrophoresis.

345 **Active CD4⁺ & Resting CD4⁺** Pace et al.⁵⁴ spinoculated CD4⁺ T cells with HIV NL4-
346 3 at an MOI of 0.1. After 96 hours, the cells were stained for intracellular Gag CD25,
347 CD69 and HLA-DR and sorted into four subpopulations based on activation state and
348 Gag expression; activated Gag-, activated Gag+, resting Gag- and resting Gag+. The
349 ability of the viruses to reactivate was not tested although previous studies have shown
350 that the majority are likely inducible⁷⁰. Genomic DNA was extracted and digested
351 with restriction enzymes MseI and Tsp509 and ligated to adapters. Proviral LTR-host
352 genome junctions were sequenced by 454 pyrosequencing after nested PCR.

353 All datasets were processed using the hiReadsProcessor R package⁷¹. Adaptor trimmed
354 reads were aligned to UCSC freeze hg19 using BLAT⁷². Genomic alignments were scored
355 and required to start within the first three bases of a read with 98% identity. Alignments for
356 a given read with a BLAT score less than the maximum score for that read were discarded.
357 Reads giving rise to multiple best scoring genomic alignments were excluded, while reads
358 with a single best hit were dereplicated and converged if within 5bp of each other. The
359 Bcl-2 transduced CD4⁺ sample was sequenced from U3 in the 5' HIV LTR while the other
360 samples were sequenced from U5 in the 3' LTR. To account for the 5 base duplication of
361 host DNA caused by HIV integration, the chromosomal coordinates of the Bcl-2 transduced
362 CD4⁺ sample were adjusted by ± 4 bases.

363 To allow for alignment difficulties in the analysis of genomic repeats, reads with multiple
364 best scoring alignments, along with the single best hit reads used above, were included in
365 the repeat analyses. If any best scoring alignment for a read fell within a repeat, then that

Title	Cell type	Virus	Time of harvest after infection	Sequencing	Generation of expressed vs. silent/inducible	Citation	Silent/inducible unique sites	Expressed unique sites
Jurkat	Jurkat cells	HIV vector pEV731 (LTR-Tat-IRES-GFP)	2 weeks	Sanger	TNF α , GFP expression	Lewinski et al. ⁵²	463 inducible	643
Bcl-2 transduced CD4 $^{+}$	Primary CD4 $^{+}$ T cells (Bcl-2 transduced)	HIV NL4-3- δ 6-drEGFP (inactivated <i>gag</i> , <i>vif</i> , <i>vpr</i> , <i>vpu</i> , <i>nef</i> and <i>env</i> replaced by GFP)	3 days + 3-4 weeks + 3 days	Sanger	anti-CD3, anti-CD28 antibodies, GFP expression	Shan et al. ⁵³	446 inducible	273
Active CD4 $^{+}$	Primary active CD4 $^{+}$ T cells	HIV NL4-3	3 days	454	high vs. low Gag	Pace et al. ⁵⁴	1604 silent	1274
Resting CD4 $^{+}$	Primary resting CD4 $^{+}$ T cells	HIV NL4-3	3 days	454	high vs. low Gag	Pace et al. ⁵⁴	1942 silent	784
Central Memory CD4 $^{+}$	Primary central memory CD4 $^{+}$ T cells	HIV NL4-3 Δ Nef GFP	2 days/9 days	Ion-Torrent	anti-CD3, anti-CD28 antibodies, GFP expression	This paper	1729 inducible	3278

Table 2.1: HIV-1 integration datasets from *in vitro* models of latency where the proviruses were determined to be silent/inducible or expressed

³⁶⁶ read was considered to map to that repeat.

³⁶⁷ 2.3.4 Genomic features

³⁶⁸ A total of 140 whole genome features for CD4 $^{+}$ T-cells were gathered from data sources
³⁶⁹ indicated in Table 2.2. For features encoded as peaks or hotspots, the log of the distance of
³⁷⁰ each integration site to the nearest border was used for modeling. Integration sites from
³⁷¹ HIV 89.6 infection in primary CD4 $^{+}$ T cells⁷³ were used to count nearby integrations and
³⁷² determine a \pm 20bp position weight matrix for integration targets. Illumina RNA-Seq from
³⁷³ active CD4 $^{+}$ cells (Chapter 4) was used to estimate raw cellular expression and fragments
³⁷⁴ per kilobase of transcript per million mapped reads for genes as calculated by Cufflinks⁷⁴.
³⁷⁵ For sequence-based data like RNA-Seq and ChIP-Seq, the number of reads aligned within a
³⁷⁶ \pm 50, 500, 5,000 50,000 and 500,000 bp windows of each integration site were counted and
³⁷⁷ log transformed. In addition, chromatin state classifications derived from a hidden Markov

378 model based on histone marks and a few binding factors⁷⁵ were included as binary variables.

379 All data from previous genomic freezes were converted to hg19 using liftover⁷⁶.

380 2.4 Results

381 The combination of integration site data newly reported here (set named “Central Memory
382 CD4⁺”) with previously published data (sets named “Jurkat”, “Bcl-2 transduced CD4⁺”,
383 “Active CD4⁺”, and “Resting CD4⁺”) provides a collection of 12,436 integration sites (Table
384 2.1) where the expression status of the provirus—silent/inducible or expressed—is known.

385 In three of the datasets, Jurkat, Central Memory CD4⁺ and Bcl-2 transduced CD4⁺, the
386 proviruses were sorted based on inducibility. In the Resting CD4⁺ and Active CD4⁺ datasets,
387 cells were sorted only based on proviral expression. Previous studies have shown that most
388 silent proviruses in this model system are inducible⁷⁰.

389 2.4.1 Global model

390 If a genomic feature and latency are monotonically related then we should be able to detect
391 this relationship using Spearman rank correlation. In addition if a feature has a consistent
392 effect across models we should see a consistent pattern in the direction of correlation. A
393 simple first look for correlation between genomic features (Table 2.2) and latency status
394 yielded inconsistent results among the five samples with no variables having a significant
395 Spearman rank correlation across all, or even four out of five, of the samples (Figure 2.1).

396 This suggests that there is not a consistent simple monotonic relationship between the
397 genomic variable and latency, or that any such correlations are modest and not detectable
398 across all studies given the available statistical power. We return to some of the stronger
399 trends below.

400 To investigate whether a combination of variables may affect latency, we fit a lasso-regularized
401 logistic regression, as implemented in the R package glmnet⁸⁵, to predict latency using
402 the genomic variables. The relationship between silent/inducible status and each genomic
403 variable was allowed to vary between models by including the interaction of genomic features

Group	Type	Source	Number	Types
T cell expression	RNA-Seq	Chapter 4	1	RNA
Jurkat expression	RNA-Seq	Encode ⁷⁷	1	wgEncodeHudsonalphaRnaSeq
Integration sites	Locations	Berry et al. ⁷³	1	sites
DNase sensitivity	DNA-Seq/peaks	Encode ⁷⁷	1	wgEncodeOpenChromDnase
Methylation	DNA-Seq	⁷⁸	1	Methyl
CpG	Locations	UCSC ⁷⁹	1	cpgIslandExt
Sequence-based	Continuous	—	4	% GC, HIV PWM score, distance to centrosome, chromosomal position
Repeats	Locations	UCSC ⁷⁹	16	DNA, LINE, Low_complexity, LTR, Other, RC, RNA, rRNA, Satellite, scRNA, Simple_repeat, SINE, snRNA, srpRNA, tRNA, alphoid
Histone features	ChIP-Seq/Peaks	Wang et al. ⁸⁰	18	H2AK5ac, H2AK9ac, H2BK120ac, H2BK12ac, H2BK20ac, H2BK5ac, H3K14ac, H3K18ac, H3K23ac, H3K27ac, H3K36ac, H3K4ac, H3K9ac, H4K12ac, H4K16ac, H4K5ac, H4K8ac, H4K91ac
Histone features	ChIP-Seq/Peaks	Barski et al. ⁸¹	23	CTCF, H2AZ, H2BK5me1, H3K27me1, H3K27me2, H3K27me3, H3K36me1, H3K36me3, H3K4me1, H3K4me2, H3K4me3, H3K79me1, H3K79me2, H3K79me3, H3K9me1, H3K9me2, H3K9me3, H3R2me1, H3R2me2, H4K20me1, H4K20me3, H4R3me2, PolII
Chromatin state	Binary	Ernst and Kellis ⁷⁵	51	state ₁ ,state ₂ ,...,state ₅₁
HATs and HDACs	ChIP-Seq	Wang et al. ⁸²	11	Resting-HDAC1, Resting-HDAC2, Resting-HDAC3, Resting-HDAC6, Resting-p300, Resting-CBP, Resting-MOF, Resting-PCAF, Resting-Tip60, Active-HDAC6, Active-Tip60
Nucleosome	ChIP-Seq	Schones et al. ⁸³	2	Resting-Nucleosomes, Active Nucleosomes
UCSC genes	Locations	Hsu et al. ⁸⁴	4	in gene, in gene (same strand), gene count, distance to nearest gene, in exon, in intron

Table 2.2: Genomic data available for comparison to HIV integration sites

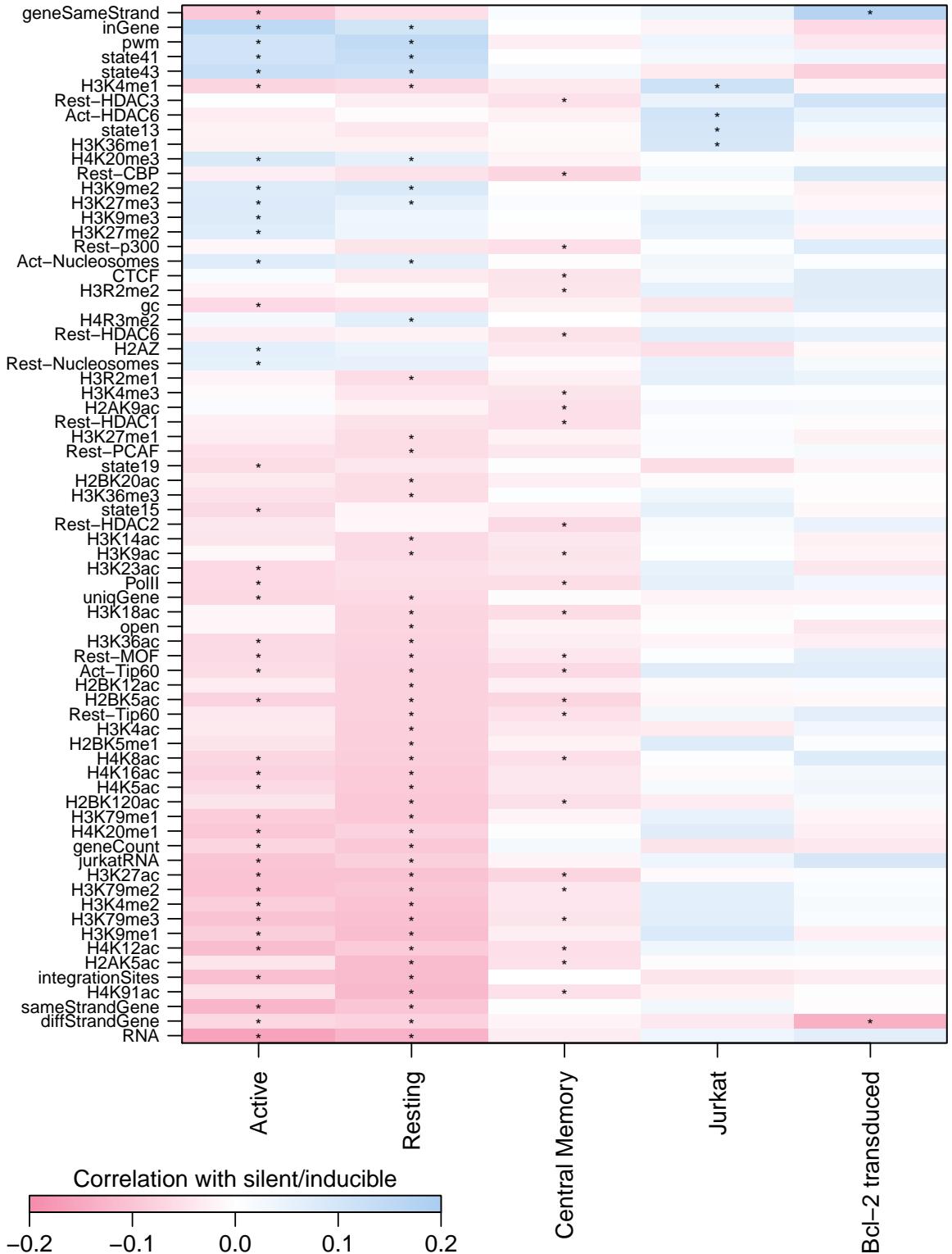


Figure 2.1: Spearman rank correlation between proviral expression status and genomic features. Only genomic features with at least one correlation with latency with a false discovery rate q -value < 0.01 (marked by asterisks) are shown.

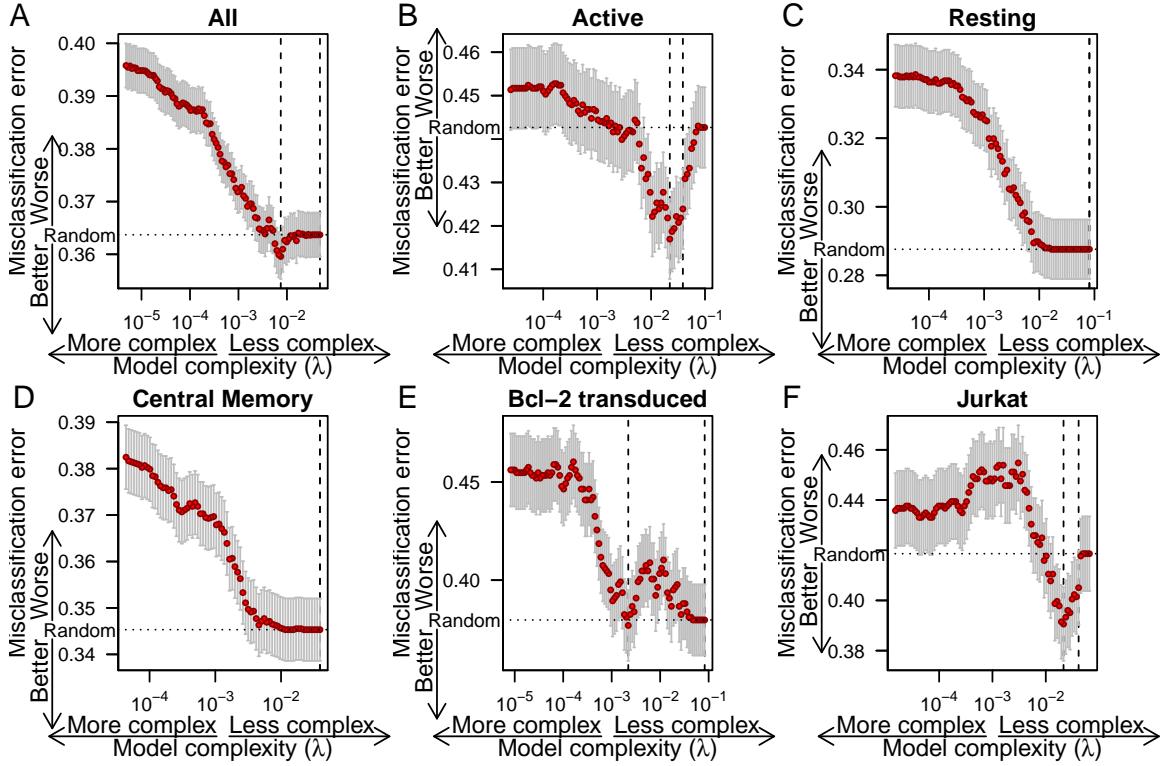


Figure 2.2: Misclassification error from cross validation for lasso regressions of silent/inducible status on genomic features as a function of λ , the regularization coefficient for the lasso regression, for all cell culture models combined and each individual cell culture model. The number of variables included and size of coefficients in the model increases to the left. Whiskers show the standard error of mean misclassification error. Dashed vertical lines indicate the minimum misclassification error and the simplest model within one standard error. Dotted horizontal line indicates the misclassification error expected from random guessing.

404 with dummy variables indicating cellular model. The λ smoothing parameter of the lasso
 405 regression was optimized by finding the λ with lowest classification error in 480-fold cross
 406 validation and finding the simplest model with misclassification error within one standard
 407 error.

408 The proportion of silent/inducible sites varied between the samples. To avoid the model
 409 overfitting on this source of variation, an indicator variable for each sample was included in
 410 the base model. The base model with no genomic variables was selected as the best model by
 411 cross validation (Figure 2.2A). This suggest that there is not a consistent linear relationship
 412 between an additive combination of genomic variables and latency across all models.

413 When each dataset was fit individually with leave-one-out cross validation, improvements in
414 cross-validated misclassification error were only observed in the Active CD4⁺ (5.8% decrease
415 in misclassification error, standard error: 2.1) and Jurkat (6.7% decrease in misclassification
416 error, standard error: 3.5) samples (Figure 2.2B-F). There was no overlap in variables
417 selected for the Active CD4⁺ and Jurkat samples.

418 Finding little global association between latency and genomic features, we investigated
419 whether predictors of latency reported previously by single studies were consistently associ-
420 ated with latency across studies.

421 **2.4.2 Cellular transcription**

422 Model systems with defined integration sites show upstream transcription can interfere with
423 viral transcription⁸⁶ and that cellular transcription in the same orientation may interfere
424 with viral transcription⁵⁵ or increase viral transcription⁵⁶ and in opposite orientations may
425 decrease transcription⁵⁶. In integration site studies, integration outside genes appears to
426 increase latency⁵² but high transcription of nearby host cell genes may cause increased
427 latency^{52,53}. In addition, Tat or other viral proteins may affect cellular transcription^{87,88}.

428 To look at transcription and latency, we ran a logistic regression of silent/inducible status
429 on a quartic function of RNA expression, as determined by RNA-Seq reads within 5,000
430 bases in Jurkat cells for the Jurkat sample or CD4⁺ T cells for the remaining samples,
431 interacted with indicator variables encoding cell culture model. There appears to be little
432 agreement between samples (Figure 2.3). The Resting CD4⁺ and Active CD4⁺ datasets
433 show an enrichment in silent proviruses in regions with low gene expression. The other three
434 studies show the opposite or no relationship for low expression regions. The two samples
435 showing increased silence in areas of low expression (Resting CD4⁺ and Active CD4⁺) are
436 from a study that did not check whether inactive viruses could be activated. One possible
437 explanation is that regions with low gene transcription may harbor proviruses that are not
438 easily activated, though some other discrepancy between *in vitro* systems could also explain

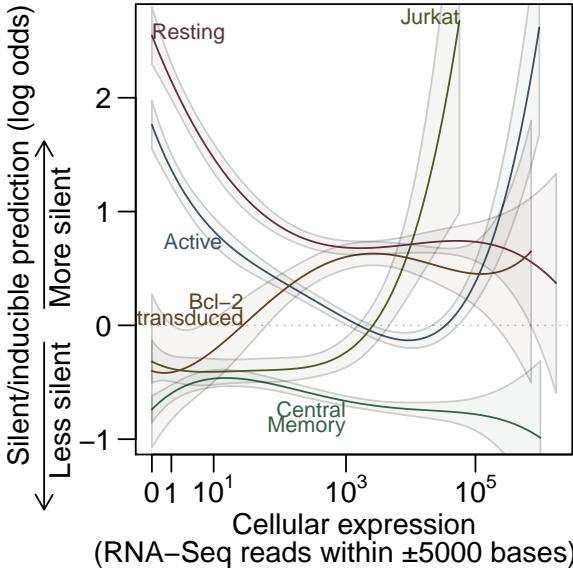


Figure 2.3: Predictions from a logistic regression of silent/inducible status on cellular RNA expression. High y-axis values are predicted to be silent/inducible. Dashed line shows where equal odds of silent/inducible and expressed are predicted. Solid lines show predictions from the regression for each sample and shaded regions indicate one standard error from the modeled predictions.

439 the difference. Both the Jurkat and Active CD4⁺ samples appear to increase in latency with
 440 increasing expression while the remaining three studies did not show a strong trend.

441 2.4.3 Orientation bias

442 Shan et al.⁵³ reported that inducible proviruses were oriented in the same strand as the
 443 host cell genes into which they had integrated more often than chance. This orientation bias
 444 was still reproduced after our reprocessing of the Bcl-2 transduced CD4⁺ sample from Shan
 445 et al.⁵³. However, the proportion of provirus oriented in the same strand as host genes did
 446 not differ significantly from 50% in the other samples (Figure 2.4). Perhaps orientation bias
 447 and transcriptional interference are especially sensitive to parameters of the model system.

448 2.4.4 Gene deserts

449 Lewinski et al.⁵² reported increased latency in gene deserts. In the collected data, integration
 450 outside known genes was associated with latency (Fisher's exact test, $p < 10^{-6}$). This
 451 seemed to largely be driven by the Active CD4⁺ and Resting CD4⁺ samples with significant
 452 association found individually in only those two samples (both $p < 10^{-8}$) and no significant
 453 association observed in the other three samples (Figure 2.5A). Looking only at integration
 454 sites outside genes, silent sites in the Resting CD4⁺ sample had a mean distance to the

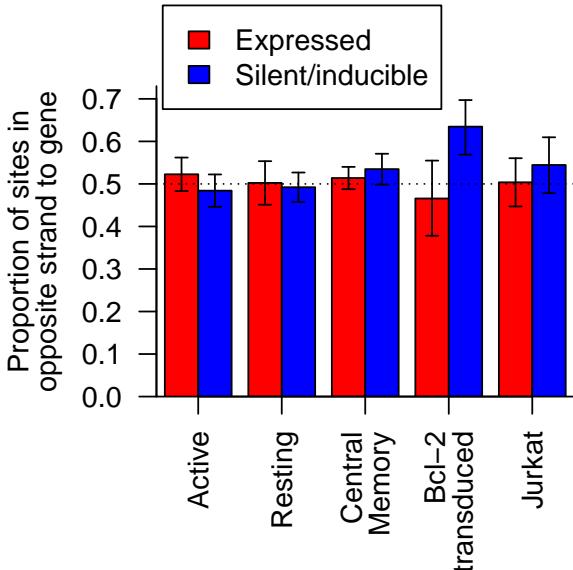


Figure 2.4: The proportion of provirus integrated in the opposite strand compared to cellular genes in silent/inducible (blue) and expressed (red) samples. Error bars show the 95% Clopper-Pearson binomial confidence interval.

455 nearest gene 2.5 times greater than that of expressed sites (95% CI: 2.2–6.2 \times , $p < 10^{-6}$,
 456 Welch two sample t-test on log transformed distance) (Figure 2.5B). The Active CD4 $^{+}$
 457 sample had a small difference that did not survive Bonferroni correction.

458 Lewinski et al.⁵² also reported decreased latency near CpG islands and reasoned this was
 459 tied to the increased latency in gene deserts. In the Resting CD4 $^{+}$ sample, silent sites were
 460 on average further from CpG islands than expressed sites (Bonferroni corrected Welch's two
 461 sample T test, $p = 0.006$), but there was no significant relationship between silent/inducible
 462 status and log distance to CpG island after Bonferroni correction if the integration site's
 463 location inside or outside of a gene was accounted for first (analysis of deviance).

464 2.4.5 Alphoid repeats

465 Alphoid repeats are repetitive DNA sequences found largely in the heterochromatin of cen-
 466 tromeres⁸⁹. Integration near heterochromatic alphoid repeats has been reported to associate
 467 with latency^{47,52,54}. Looking only at uniquely mapping sites, there was no statistically
 468 significant association between latency and location inside an alphoid repeat in pooled or
 469 individual samples (Fisher's exact test).

470 Since alphoid repeats are both problematic to assemble in genomes and difficult to map

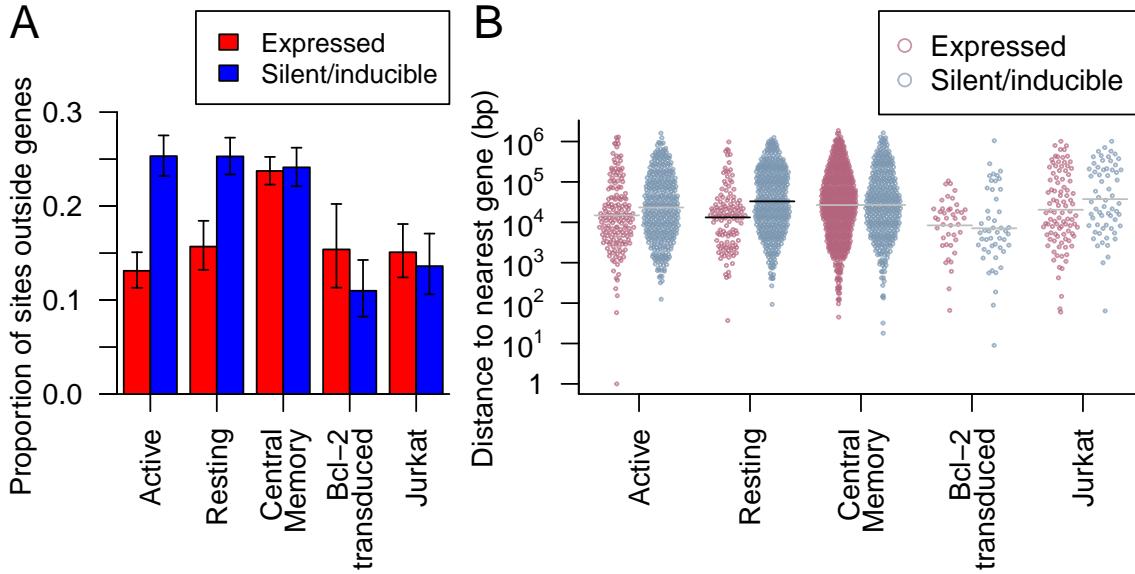


Figure 2.5: (A) The proportion of provirus integrated outside genes in silent/inducible (blue) and expressed (red) samples. Error bars show the 95% Clopper-Pearson binomial confidence interval. (B) The nearest distance to any gene for integration sites (points) outside genes in the five samples. Points are spread in proportion to kernel density estimates. Horizontal lines indicate sample means where there was a significant difference in means between silent/inducible and expressed provirus (black) or no significant difference (grey).

471 onto, we reasoned that some alphoid hits might be lost or miscounted in the filtering
 472 procedures of the standard workup. To counteract this, we treated each sequence read as an
 473 independent observation of a proviral integration and included sequence reads with more
 474 than one best scoring alignment. For multiply aligned reads, we considered the read to have
 475 been inside an alphoid repeat if any of its best scoring alignments fell within a repeat. We
 476 found 74 reads with potential alphoid mappings. Integration inside alphoid repeats was
 477 significantly associated with the expression status of a provirus in the Resting CD4⁺, Jurkat
 478 and Central Memory CD4⁺ datasets (Bonferroni corrected Fisher's exact test, all $p < 0.05$)
 479 and approached significance in the Active CD4⁺ dataset ($p = 0.053$) (Figure 2.6). The Bcl-2
 480 transduced CD4⁺ data did not contain any integration sites in alphoid repeats, probably due
 481 to 1) the relatively low number of integration sites in the dataset and 2) to the requirement
 482 for cleavage at two Pst1 restriction sites, which are not found in the consensus sequence of
 483 alphoid repeats⁹⁰. Of the 1340 repeat types in the RepeatMasker database⁹⁰, only alphoid
 484 repeats achieved a significant association with proviral expression in more than two datasets.

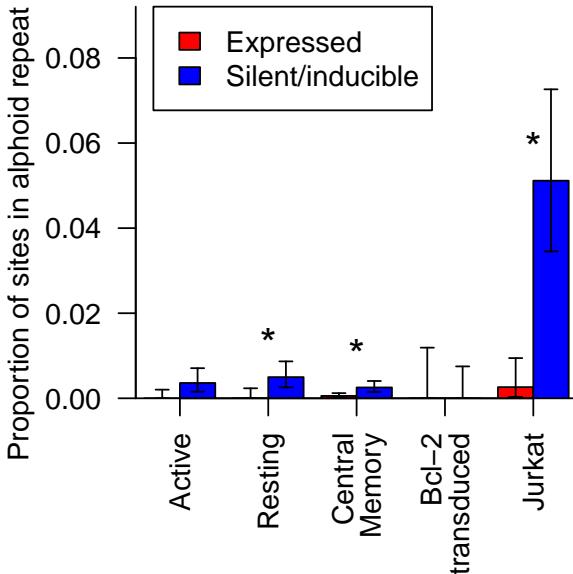


Figure 2.6: The proportion of integration sites with matches in alphoid repeats in silent/inducible (blue) and expressed (red) cells in five samples. Error bars show the 95% Clopper-Pearson binomial confidence interval. Asterisks indicate significant associations between integrations within an alphoid repeat and proviral expression status (Bonferroni corrected Fisher’s exact test $p < 0.05$).

485 2.4.6 Acetylation

486 Histone marks or chromatin remodeling, especially involving the key “Nuc-1” histone near
 487 the transcription start site in the viral LTR, appear to affect viral expression^{48,91,92}. Based
 488 on this effect, histone deacetylase inhibitors have been developed as potential HIV treatments
 489 and show some promise in disrupting latency⁶⁰. In these genome-wide datasets, we do
 490 not have information on the state of individual LTR nucleosomes. However, repressive
 491 chromatin does seem to spread to nearby locations if not blocked by insulators^{44,45} and
 492 the state of neighboring chromatin could affect proviral transcription independently of
 493 provirus-associated histones.

494 We found that the number of ChIP-seq reads near an integration site from several histone
 495 acetylation marks (Figure 2.1) were associated with efficient expression in the Active CD4⁺,
 496 Resting CD4⁺ and Central Memory CD4⁺ samples. H4K12ac had the strongest association
 497 (Bonferroni corrected Fisher’s method combination of Spearman’s ρ , $p < 10^{-25}$) with
 498 silence/latency (Figure 2.7A).

499 Although the appearance of several significantly associated acetylation marks might suggest
 500 acetylation exerts a considerable effect on the expression of a provirus, there are strong

correlations among these marks, so their effects may not be independent. To account for the correlations between these variables, we performed a principal component analysis (PCA) to convert the correlated acetylation marks into a series of uncorrelated principal components that capture much of the variance within a few components. Here, the first principal component explained 59% of the variance and the first ten components 84%. Several of these principal components again displayed significant associations with latency in the Active CD4⁺, Resting CD4⁺ and Central Memory CD4⁺ samples but no significant correlations in the Bcl-2 transduced CD4⁺ or Jurkat samples (Figure 2.7B). A logistic regression of expression status on the first ten principal components and sample did not reduce misclassification error from a base model including only sample in 480-fold cross validation (base model misclassification error: 36.4%, PCA model: 36.5%). This suggests that acetylation of neighboring chromatin does not exert strong effects on latency in all samples.

2.4.7 Clustering

We reasoned that if there was a strong relationship between latency and chromosomal position, then integration sites that are near one another on the same chromosome should share the same expression status more often than expected by chance. To test this, we compared how often pairs of proviruses shared the same expression status in relation to the distance between the two sites (Figure 2.8). Pairs of sites with little distance between integration locations did share the same expression status more often than expected by chance (e.g. neighbors closer than 100bp, Fisher exact test $p = 0.0002$). Breaking out the data to separate between sample and within sample pairings showed that this matching was limited to neighbors within the same experimental model (Figure 2.8), emphasizing that chromosomal environment does appear to influence latency, but the factors involved differ among experimental models of latency.

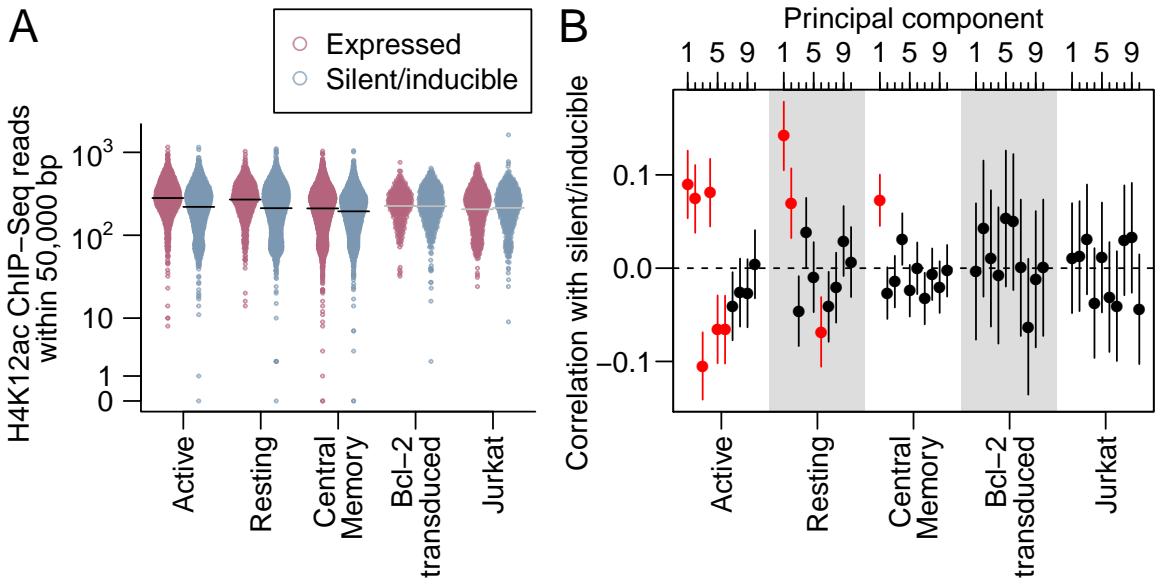


Figure 2.7: (A) The number of ChIP-seq reads for H4K12ac, the histone mark with the lowest Fisher's method p -value for correlation with latency, within 50,000 bases across the five samples. Integration sites (points) are spread in proportion to kernel density estimates. Horizontal lines indicate sample means where there was a significant difference (black) in means between silent/inducible and expressed provirus or no significant difference (grey). (B) The correlation (points) and its 95% confidence interval (vertical lines) between principal components of acetylation and silent/inducible status for each of the five samples. Red indicates correlations with a Bonferroni-corrected p -value < 0.05 .

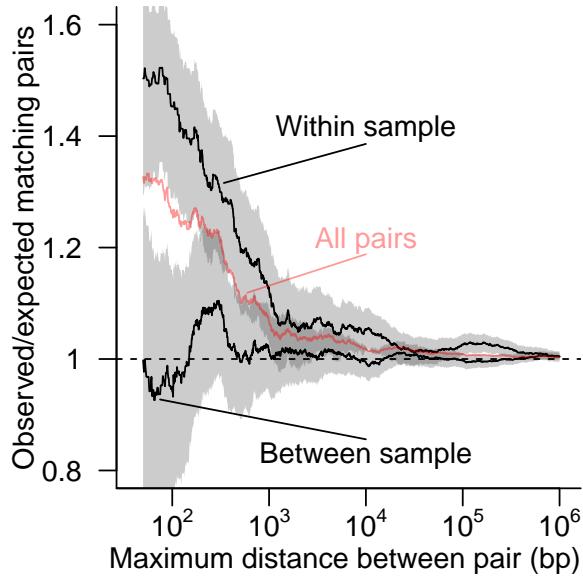


Figure 2.8: The ratio of the number of pairs of proviruses with matching expression status to the number of matches expected by random pairings given the frequency of silent/inducible proviruses. All possible pairs of proviruses integrated within a given distance of each other on the same chromosome (red line) were separated into two sets; one with both proviruses from within the same cell culture model and one with proviruses paired between two different cell culture models (black lines). The shaded region shows the 95% Clopper-Pearson binomial confidence interval for within and between sample pairings. The dashed horizontal line shows the ratio of 1 expected if there is no association between the expression status of neighboring proviruses.

526 **2.5 Conclusions**

527 Here we compared the latency status of HIV-1 proviruses in five model systems with the
528 genomic features surrounding their integration sites. Surprisingly, no relationships between
529 genomic features near the integration location and latency achieved significance in all models.
530 Proviruses from the same cellular model integrated in nearby positions did share the same
531 latency status much more often than predicted by chance, indicating the existence of local
532 features influencing latency, but these were not consistent among models. This suggests that
533 whatever features are affecting latency are highly local and model-specific, and that we may
534 not have access to all relevant chromosomal features e.g. 93–96.

535 In addition to differences in experimental conditions, methodological issues have the potential
536 to obscure patterns. Examples include multiply infected cells, inactivated viruses and
537 inaccurate assessment of HIV gene activity—each of these are discussed below.

538 A latent provirus integrated into the same cell as an expressed provirus will be erroneously
539 sorted as expressed, potentially confounding analysis. A low multiplicity of infection (MOI)
540 will help to avoid this problem, but there is still the potential for a significant proportion of
541 the cells studied to contain multiple integrations. This problem arises because although cells
542 with multiple integrations form a small proportion of total cells, most of the total are cells
543 lacking an integrated provirus and thus are excluded by experimental design. For example,
544 assuming integrations are Poisson distributed with an MOI of 0.1 (1 integration per 10 cells),
545 90.5% of cells will not contain a provirus, 9% of cells will contain one proviral integration
546 and 0.5% of cells will contain multiple integrations. The cells without an integration are
547 not amplified by HIV-targeted PCR leaving only 9.5% of the total cells. Of these cells
548 actually under study, 4.9% will contain multiple integrations. Thus the signal from expressed
549 proviruses may be muted by the presence of latent proviruses in the expressed population.

550 The replication cycle of HIV is error prone, and a significant proportion of virions contain
551 mutated genomes⁹⁷. In studies that do not check for inducibility, mutant proviruses

552 integrated in regions of the genome otherwise favorable to proviral expression can be sorted
553 into the latent pool due to mutational inactivation. This problem of inactivated provirus
554 is worse when latent provirus are rare and exacerbated further when looking at latency in
555 the cells of HIV patients due to selective enrichment of inactivated proviruses incapable
556 of spreading infection³⁵. Here, the effects of mutation are minimized in the datasets that
557 required inducible viral expression (Jurkat, Bcl-2 transduced CD4⁺, Central Memory CD4⁺)
558 but may be a confounder in the two datasets that were sorted based on lack of viral expression
559 only (Active CD4⁺, Resting CD4⁺).

560 Inaccurate staining or leaky markers may also result in misclassification of proviruses. False
561 positives and false negatives will result in incorrectly sorted latent and expressed integrations.
562 For example, if 5% of cells not containing Gag are labeled as Gag+ and there are an equal
563 amount of latent and expressed integration sites, then 4.8% of integrations labeled expressed
564 will actually be latent. If a category is rare, false staining has even greater potential to cause
565 error. For example, if only 5% of sites are latent and a Gag stain has a false negative rate
566 of 5%, then we would expect 48.7% of sites classified as latent to actually be mislabeled
567 expressed integrations.

568 Attempts to induce latent proviruses in patients have so far focused on using histone
569 deacetylase inhibitors, raising interest in associations with histone acetylation in these data.
570 An important caveat in results from these genome-wide data is that histone modification
571 near the integrated provirus may not be representative of modification within the provirus
572 at the key “Nuc-1” nucleosome of the transcription start site⁹², though local correlations
573 in chromatin states are well established from studies of position effect variegation^{44,45}. We
574 found that some histone acetylation marks were significantly associated with viral expression
575 in some but not all samples (Figures 2.1, 2.7). This lack of association may be due to a
576 lack of power in these studies, but the confidence intervals suggest that any correlations
577 between acetylations and latency are unlikely to be strong. These weak correlations raise
578 the possibility that there are populations of latent proviruses that are not associated with

579 acetylation and may not be inducible by histone deacetylase inhibitors.

580 This study highlights that the choice of model system can have a large effect on measurements
581 of latency. Further studies are needed to determine which *in vitro* models best reflect latency
582 *in vivo*. Different cell models may report genuinely different mechanisms of latency. While
583 we did see some relationship between histone acetylation and latency, paralleling a recent
584 clinical trial of SAHA⁶⁰, associations with histone acetylation did not explain a large fraction
585 of the difference between latent and expresssed proviruses in any of the five models. One
586 possible explanation is that there may be multiple mechanisms that maintain proviruses in a
587 latent state. To be successful, shock-and-kill treatments must induce and destroy all latent
588 proviruses to eliminate HIV from an infected individual, raising the question of whether
589 multiple simultaneous inducing treatments will be necessary.

590 **2.6 Availability of supporting data**

591 Sequence reads from the Central Memory CD4⁺ sample reported here, the Resting CD4⁺
592 and Active CD4⁺ data reported by Pace et al.⁵⁴, the Bcl-2 transduced CD4⁺ data reported
593 by Shan et al.⁵³ and reprocessed data originally reported by Lewinski et al.⁵² are available
594 at the Sequence Read Archive under accession number SRP028573.

595 **2.7 Author's contributions**

596 SS-M led the computational analysis, with assistance from CCB and NM. MKL, DL and JG
597 analyzed integration sites using IonTorrent sequencing. MF, AB and VP prepared DNA
598 from latent and activated T cells using the Central Memory CD4⁺ model. LS, RFS, MJP,
599 LMA and UO'D contributed data and suggestions. SS-M, KEO and FDB planned the
600 overall study, and SS-M and FDB wrote the paper. All authors read and approved the final
601 manuscript.

602 **2.8 Acknowledgements**

603 We would like to thank Werner Witke for assistance with IonTorrent sequencing. This
604 work was supported in part by NIH grants R01 AI 052845-11 to FDB, R21AI 096993 and
605 K02AI078766 to UO'D, 5T32HG000046 to SS-M, AI087508 to VP and R01AI038201 to
606 JG, the Penn Genome Frontiers Institute, the University of Pennsylvania Center for AIDS
607 Research (CFAR) P30 AI 045008 and the University of California, San Diego, CFAR P30
608 AI036214.

609 **CHAPTER 3: Dynamic regulation of HIV-1 mRNA populations**
610 **analyzed by single-molecule enrichment and long-read**
611 **sequencing**

This chapter was originally published as:

KE Ocwieja, S Sherrill-Mix, R Mukherjee, R Custers-
Allen, P David, M Brown, S Wang, DR Link, J Olson
et al. 2012. Dynamic regulation of HIV-1 mRNA popu-
lations analyzed by single-molecule enrichment and long-
read sequencing. *Nucleic Acids Res*, 40:10345–10355. doi:
10.1093/nar/gks753

612 FD Bushman, K Travers, DR Link, E Schadt, KE Ocwieja and R Mukher-
jee conceived and designed the experiment. KE Ocwieja and R Custers-
Allen carried out sample preparation and experimental validation. P
David and J Olson performed single-molecule amplification. K Travers
and S Wang performed sequencing. KE Ocwieja, M Brown and I analyzed
the data. KE Ocwieja and I produced the figures. KE Ocwieja, FD
Bushman and I wrote the manuscript.

Supplementary data are available at [http://nar.oxfordjournals.org/
content/40/20/10345/suppl/DC1](http://nar.oxfordjournals.org/content/40/20/10345/suppl/DC1)

613 **3.1 Abstract**

614 Alternative RNA splicing greatly expands the repertoire of proteins encoded by genomes.
615 Next-generation sequencing (NGS) is attractive for studying alternative splicing because
616 of the efficiency and low cost per base, but short reads typical of NGS only report mRNA
617 fragments containing one or few splice junctions. Here, we used single-molecule amplification
618 and long-read sequencing to study the HIV-1 provirus, which is only 9700 bp in length, but
619 encodes nine major proteins via alternative splicing. Our data showed that the clinical isolate
620 HIV_{89.6} produces at least 109 different spliced RNAs, including a previously unappreciated
621 ~1 kb class of messages, two of which encode new proteins. HIV-1 message populations

622 differed between cell types, longitudinally during infection, and among T cells from different
623 human donors. These findings open a new window on a little studied aspect of HIV-1
624 replication, suggest therapeutic opportunities and provide advanced tools for the study of
625 alternative splicing.

626 3.2 Introduction

627 Alternative splicing greatly expands the information content of genomes by producing
628 multiple mRNAs from individual transcription units. Approximately 95% of human genes
629 with multiple exons encode RNA transcripts that are alternatively spliced, and mutations
630 that affect alternative splicing are associated with diseases ranging from cystic fibrosis to
631 chronic lymphoproliferative leukemia^{99–103}. Work to decipher an RNA ‘splicing code’ has
632 revealed that multiple interactions between trans-acting factors and RNA elements determine
633 splicing patterns, though regulation is little understood for most genes¹⁰⁴.

634 The integrated HIV-1 provirus is ~9700 bp in length and has a single transcription start
635 site, but according to the published literature yields at least 47 different mRNAs encoding
636 9 proteins or polyproteins, making HIV an attractive model for studies of alternative
637 splicing¹⁰⁵. HIV mRNAs fall into three classes: the unspliced RNA genome, which encodes
638 Gag/Gag-Pol; partially spliced transcripts, ~4 kb in length, encoding Vif, Vpr, a one-exon
639 version of Tat, and Env/Vpu; and completely spliced mRNAs of roughly 2 kb encoding
640 Tat, Rev and Nef (Figure 3.1A). Additional rare ‘cryptic’ splice donors (5' splice sites)
641 and acceptors (3' splice sites) contribute even more mRNAs^{106–111}. A complex array of
642 positive and negative cis-acting elements surrounding each splice site regulates the relative
643 abundance of the HIV-1 mRNAs, and disrupting the balance of message ratios impairs viral
644 replication in several models^{112–119}. Studies have suggested strain-specific splicing patterns
645 may exist^{105,120,121}. However, detailed studies of complete message populations have not
646 been reported for clinical isolates of HIV-1.

647 Several groups have demonstrated tissue- and differentiation-specific splicing of cellular

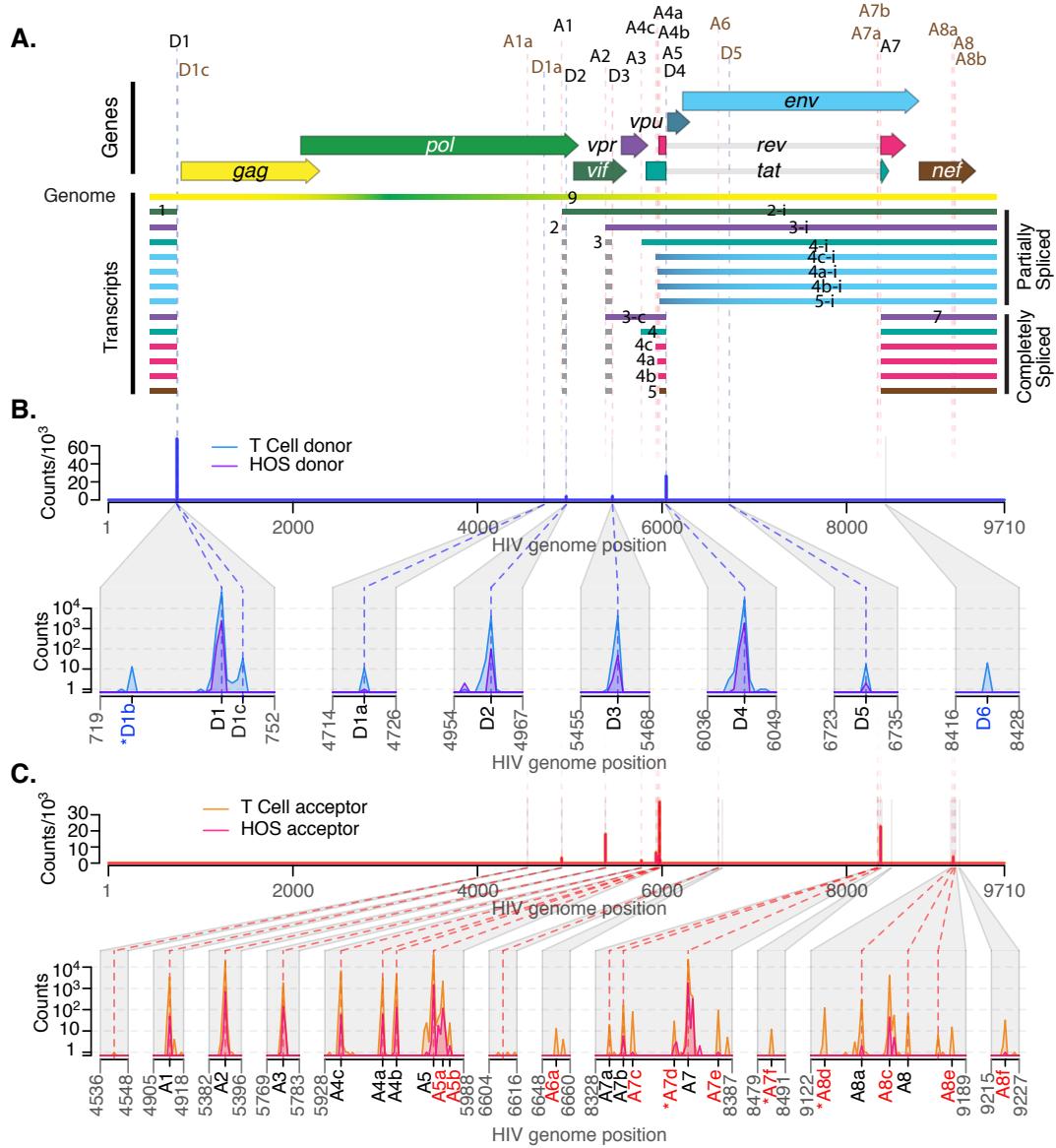


Figure 3.1: Mapping the splice donors and acceptors of HIV_{89.6}. PacBio sequence reads of HIV_{89.6} cDNA from infected HOS-CD4-CCR5 (HOS) and CD4⁺ T cells were aligned to the HIV_{89.6} genome shown in (A). Exons of the conserved HIV-1 transcripts are colored according to the encoded gene. Conserved (black) and published cryptic (brown) splice donors ('D') and acceptors ('A') are shown. Gaps in HIV-1 sequence alignments with at least one end located at a published or verified splice donor or acceptor were defined as introns. For each base of the HIV_{89.6} genome, the number of sequence reads in which that base occurred at the 5'-end (B) or 3'-end (C) of an intron is plotted for each cell type. Putative splice donors and acceptors were defined as loci that were found in at least 10 reads at the 5'- and 3'-ends of introns in sequence alignments from T-cell infections. Regions containing splice sites are enlarged for clarity. Asterisks indicate putative splice sites that are adjacent to dinucleotides other than the consensus GT and AG.

648 genes^{100,122,123}. Importantly for HIV, these include changes during T-cell activation^{124,125},
649 raising the question of how cell-specific splicing affects HIV replication. While most studies
650 of HIV-1 splicing have been conducted in cell lines using lab-adapted viral strains, limited
651 works in PBMCs from infected patients, monocytes and macrophages have suggested that
652 differences may indeed exist in relevant cell types^{107,120,126,127}. Moreover, human splicing
653 patterns differ between individuals, but such polymorphisms have not been investigated in
654 the context of HIV infection^{128,129}.

655 Here, we use deep sequencing to comprehensively characterize the transcriptome of an early
656 passage clinical isolate, HIV_{89.6}¹³⁰, in primary CD4⁺ T cells from seven human donors
657 and in the human osteosarcoma (HOS) cell line. Many deep sequencing techniques provide
658 short reads, which rarely query more than a single exon-exon junction. To distinguish
659 the full structure of HIV-1 mRNAs, which can contain several splice junctions, we used
660 Pacific Biosciences (PacBio) sequencing technology, which yields read lengths up to 10 kb¹³¹.
661 We used RainDance Technologies single-molecule PCR enrichment to preserve ratios of
662 RNAs during preparation of sequencing templates. We identified previously published and
663 novel HIV-1 transcripts and determined that HIV_{89.6} encodes a minimum of 109 different
664 splice forms. These included a new size class of transcripts, some of which contain novel
665 open reading frames (ORFs) that encode new proteins. We also found significant variation
666 between cell types, over time during infection of HOS cells and among individuals. These
667 data reveal unanticipated complexity and dynamics in HIV-1 message populations, begin
668 to clarify a little studied dimension of HIV-1 replication and suggest possible targets for
669 therapeutic interventions.

670 **3.3 Materials and methods**

671 **3.3.1 Cell culture and viral infections**

672 HIV_{89.6} was generated by transfection and subsequent expansion in SupT1 cells. Primary
673 T cells were isolated by the University of Pennsylvania Center for AIDS Research Im-

674 munology core and confirmed to be homozygous for the wild-type CCR5 allele as shown
675 in Supplementary Table S1 and described in Supplementary Methods. HOS-CD4-CCR5
676 cells^{132,133} were obtained through the AIDS Research and Reference Reagent Program,
677 Division of AIDS, NIAID, NIH from Dr Nathaniel Landau. Single round infections in T
678 cells and HOS-CD4-CCR5 cells were performed using standard methods (see Supplementary
679 Methods).

680 **3.3.2 RNA and reverse transcription**

681 Total cellular RNA was purified using the Illustra RNA kit (GE Life Sciences, Fairfield, CT,
682 USA) from 5×10^6 cells per infection. Viral cDNA was made using a reverse transcription
683 primer complementary to a sequence in U3 (RTprime, Supplementary Table S2). We used
684 Superscript III reverse transcriptase (Invitrogen) in the presence of RNaseOUT (Invitrogen)
685 to conduct first-strand cDNA synthesis from equal amounts of total cellular RNA from each
686 HOS-CD4-CCR5 time point (15.2 μ g) and from each T-cell infection (3 μ g) according to the
687 manufacturer's instructions for gene-specific priming of long cDNAs, and then treated with
688 RNaseH (Invitrogen). We checked for full reverse transcription of the longest (unspliced)
689 viral cDNAs by PCR using primers that bind in the first major intron of HIV_{89.6} (keo003,
690 keo004, Supplementary Table S2, data not shown).

691 **3.3.3 Bulk RT-PCR and cloning**

692 Transcripts were amplified from cellular RNA using the Onestep RT-PCR kit (Qiagen)
693 with primer pairs keo056/keo057 and keo058/keo059 (Supplementary Table S2) with the
694 following amplification: 5 cycles of 30 s at 94°C, 12 s at 56°C, 40 s at 72°C; then 30 cycles
695 of 30 s at 94°C, 14 s at 56°C, 40 s at 72°C; and finally 10 min at 72°C. For verification of
696 dynamic changes, primers F1.2 and R1.2 were used with 35 cycles of 30 s at 94°C, 30 s at
697 56°C and 45 s at 72°C followed by 10 min at 72°C. Products were resolved on agarose gels
698 (Nusieve 3:1, Lonza for verification of dynamic changes, Invitrogen for cloning) stained with
699 ethidium-bromide (Sigma) for visualization, or SYBR Safe DNA gel stain (Invitrogen) for

700 cloning (keo056/keo057 amplified material). DNA was purified using Qiaquick gel extraction
701 kit (Qiagen) and cloned using the TOPO TA cloning kit (Invitrogen). Plasmid DNA was
702 prepared using Qiaprep Spin Miniprep kit (Qiagen). Inserts were identified and verified
703 using Sanger sequencing. The cDNAs for *tat*^{8c}, *tat* (1 and 2 exon), *ref*, *rev* and *nef*, and the
704 transcript with exon structure 1-5-8c were cloned into the expression vector pIRES2-AcGFP1
705 (Clonetech) as described in Supplementary Methods.

706 **3.3.4 Assays of protein activity and HIV replication**

707 Activity and HIV replication assays were performed as described in Supplementary Methods.
708 Tat activity expressed from each cDNA was measured in TZM-bl cells¹³⁴ (gift of Dr Robert
709 W. Doms). Rev activity was assayed in HEK-293T cells co-transfected with pCMVGagPol-
710 RRE-R, a reporter plasmid from which Gag and Pol are expressed in a Rev-dependent
711 manner (gift of David Rekosh)¹³⁵. Intracellular and released supernatant p24 was measured
712 from cells transfected with expression constructs and infected with HIV_{89.6}.

713 **3.3.5 Western blotting**

714 HEK-293T cells were transfected with expression constructs and treated with MG132 (EMD
715 Chemicals) to inhibit the proteasome or DMSO (Supplementary Methods). Proteins were
716 detected by immunoblotting using a mouse antibody that recognizes the carboxy terminus
717 of HIV-1 Nef diluted 1:1000 in 5% milk (gift of Dr James Hoxie)¹³⁶. Horseradish peroxidase
718 (HRP)-conjugated secondary rabbit-anti-mouse antibody (p0260, DAKO) was used for
719 detection with SuperSignal West Pico Chemiluminescent Substrate (Thermo Scientific).
720 Beta-tubulin was used as a loading control, detected by the HRP-conjugated antibody
721 (ab21058, Abcam).

722 **3.3.6 Single-molecule amplification**

723 Amplification was performed by RainDance Technologies using a protocol similar to that
724 previously reported (detailed description in Supplementary Methods)¹³⁷. Amplification

725 was carried out in droplets to suppress competition between amplicons. PCR droplets
726 were generated on the RDT 1000 (RainDance Technologies) using the manufacturer's
727 recommended protocol. The custom primer libraries for this study contained 18 (HOS-CD4-
728 CCR5 cells) or 20 (primary T cells) PCR primer pairs designed to amplify different HIV
729 RNA isoforms (Supplementary Table S2).

730 **3.3.7 Single-molecule sequencing**

731 DNA amplification products from the RainDance PCR droplets were converted to SMRTbell
732 templates using the PacBio RS DNA Template Preparation Kit. Sequencing was performed by
733 Pacific Biosciences using the PacBio SMRT sequencing technology as described¹³¹. Sequence
734 information was acquired during real time as the immobilized DNA polymerase translocated
735 along the template molecule. Prior to sequence acquisition, hairpin adapters were ligated to
736 each DNA template end so that DNA polymerase could traverse DNA molecules multiple
737 times during rolling circle replication (SMRTbell template sequencing¹³⁸), allowing error
738 control by calculating the consensus ('circular consensus sequence' or CCS). For raw reads,
739 the average length was 2860 nt, and 10% were > 5000 nt. After condensing into consensus
740 reads, the mean read length was 249.5 nt, due to the use of a shorter Pacific Biosciences
741 sequencing protocol to accommodate the small size of many amplicons. Consensus reads of
742 1% were > 1100 nt. Sequencing data were collected in 45-min movies.

743 **3.3.8 Data analysis**

744 Raw reads were processed to produce CCSs. Raw reads were also retained to help in primer
745 identification and to avoid biasing against long reads. Reads were aligned against the human
746 genome using Blat⁷². Misprimed reads matching the RT primer, reads with a CCS length
747 shorter than 40 nt or raw length shorter than 100 nt and reads matching the human genome
748 were discarded. Filtered reads were aligned against the HIV_{89.6} reference genome. Potential
749 novel donors and acceptors were found by filtering putative splice junctions in the Blat
750 hits for a perfect sequence match 20 bases up- and downstream of the junction, ignoring

751 homopolymer errors, and requiring that one end of the junction be a known splice site. Local
752 maximums within a 5-nt span with > 9 such junctions were called as novel splice sites.

753 Filter-passed reads were aligned against all expected fragments based on primers and known
754 and novel junctions. Primers were identified in CCS reads by an edit distance ≤ 1 from
755 the primer in the start or end of the read, in raw reads by an edit distance ≤ 5 from a
756 concatenation of the primer, hairpin adapter and the reverse complement of the primer, and
757 in both types of reads by a Blat hit spanning an entire expected fragment.

758 Gaps in Blat hits were ignored if ≤ 10 bases long or in regions of likely poor read quality
759 ≤ 20 bases long where an inferred insertion of unmatched bases in the read occurred at the
760 same location as skipped bases in the reference. Any Blat hits with a gap > 10 nt remaining
761 in the query read were discarded. If HIV sequence was repeated in a given read (likely due
762 to PacBio circular sequencing), the alignments were collapsed into the union of the coverage.
763 Gaps in the HIV sequence found in uninterrupted query sequence were called as tentative
764 introns. Splice junctions were assigned to conserved or previously identified (published
765 or in this work) splice sites and reads appearing to contain donors or acceptors further
766 than 5 nt away from these sites were discarded. Reads with Blat hits outside the expected
767 primer range were discarded from that primer grouping. The assigned primer pair, observed
768 junctions and exonic sequence were used to assign each read to a given spliceform (specific
769 transcript structure) or set of possible spliceforms. Partial sequences that did not extend
770 through both primers were assigned to specific transcripts if the read contained enough
771 information to rule out all other spliceforms or if all other possible spliceforms contained
772 rare (< 1% usage) donors or acceptors (Supplementary Table S3). Otherwise, the read was
773 called indeterminate.

774 To calculate the ratios of transcripts within the partially spliced class, we counted the
775 number of reads for each assigned spliceform amplified by primer pair 1.3 and divided by the
776 total number of assigned partially spliced reads amplified with these primers (Supplementary
777 Figure S1 and Supplementary Table S2). Assigned sequences amplified with primer pairs

778 1.4 and 4.1 (full-length cDNAs, T cells only) were used to calculate ratios of transcripts
779 within each of the two completely splice classes (~ 2 and ~ 1 kb). To compare ratios of ~ 2
780 kb transcripts calculated within reads from primer pairs 1.4 and 4.1, we normalized ratios
781 from pair 4.1 to the *nef* 2 transcript (containing exons 1, 5 and 7). Due to size biases
782 inherent in the approach, we did not compare across size classes, and unspliced transcripts
783 were not included in ratio analysis. For all ratio analysis, transcripts including cryptic or
784 novel junctions were counted only if they appeared in at least five reads, otherwise they
785 were excluded from the analysis and from the count of total assigned reads.

786 To estimate the minimum total number of transcripts present, partial sequence reads were
787 included. Each exon-exon junction occurring in at least five reads and not previously assigned
788 to a particular transcript (Figure 3.2) was counted as evidence of an additional transcript
789 (47 additional junctions were detected, see Supplementary Table S4). If two such junctions
790 could conceivably occur in a single mRNA, we counted only one unless we could verify from
791 sequence reads that they were amplified from separate cDNAs, resulting in 31 additional
792 transcripts. The minimum transcript number calculated by a greedy algorithm treating
793 introns as events in a scheduling problem agreed with the above calculation.

794 Several groups have demonstrated tissue- and differentiation-specific splicing of cellular
795 genes^{100,122,123}. Importantly for HIV, these include changes during T-cell activation^{124,125},
796 raising the question of how cell-specific splicing affects HIV replication. While most studies
797 of HIV-1 splicing have been conducted in cell lines using lab-adapted viral strains, limited
798 works in PBMCs from infected patients, monocytes and macrophages have suggested that
799 differences may indeed exist in relevant cell types^{107,120,126,127}. Moreover, human splicing
800 patterns differ between individuals, but such polymorphisms have not been investigated in
801 the context of HIV infection^{128,129}.

802 For studies of transcript dynamics, reads from primer pairs 1.2, 1.3 and 1.4 containing
803 junctions between D1 or any donor and each of five mutually exclusive acceptors, A3, A4c,
804 A4a, A4b, A5 and A5a, were collected and their ratios calculated.

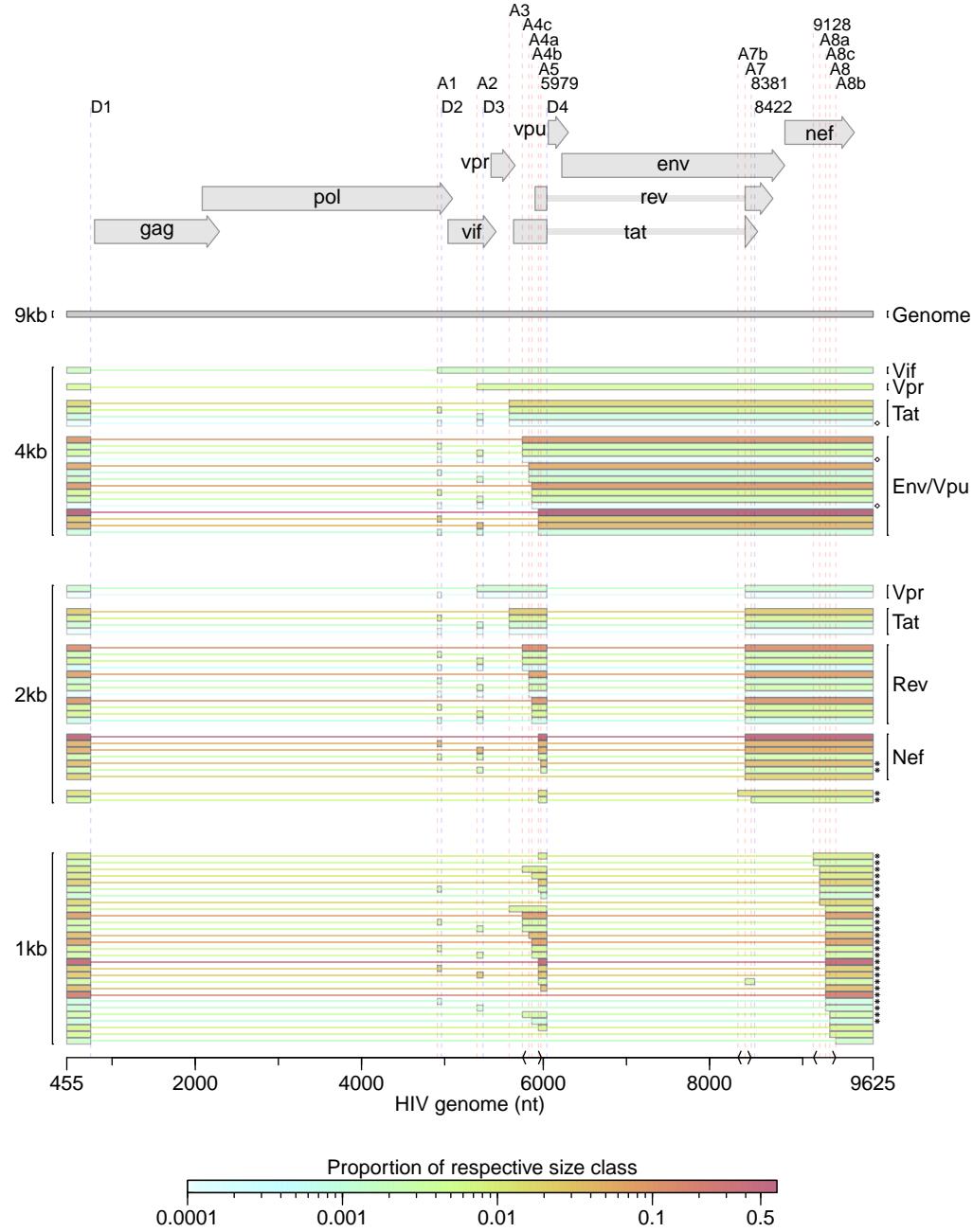


Figure 3.2: HIV_{89.6} transcripts in T cells for which the full message structure was determined are shown arranged by size class. Thick bars correspond to exons and thin lines to excised introns. For the well-conserved transcripts, encoded proteins are indicated. The relative abundance of each transcript within its size class is indicated by color. Asterisks denote transcripts that have not been reported previously to our knowledge. Of the 47 conserved HIV-1 transcripts, three were detected in fewer than five reads (indicated with ◊) and two messages were not detected and are not shown (one encoding Vpr and one encoding Env/Vpu). Depicted non-conserved transcripts (using novel or cryptic splice sites) were each detected in at least five independent sequence reads across samples from at least two different human T-cell donors.

805 **3.3.9 Statistical analysis**

806 Statistical modeling was performed using generalized linear modeling as described in Supple-
807 mentary Report S2. All analyses were performed in R 2.14.0 (R Development Core)⁶⁵.

808 **3.3.10 Data access**

809 Sequence data is available in the SRA database with the following accession numbers:
810 SRP014319.

811 **3.4 Results**

812 **3.4.1 Sequencing HIV-1 transcripts produced in primary T cells and HOS cells**

813 In order to characterize HIV-1 transcript populations, we prepared viral cDNA from primary
814 CD4⁺ T cells of seven different healthy human donors infected in vitro with HIV_{89.6}, an early
815 passage dual-tropic clade-B clinical isolate (Supplementary Figure S1, human donor data
816 in Supplementary Table S1)¹³⁰. We also studied HIV messages produced in infected HOS
817 cells engineered to express CD4 and CCR5 (HOS-CD4-CCR5) because these cells support
818 efficient HIV replication and engineered variants are widely used in HIV research. HOS
819 cells were harvested at 18, 24 and 48 hours post infection (hpi) to investigate longitudinal
820 changes during infection, and for comparison to 48 h infected T cells.

821 To preserve the relative proportions of template molecules while amplifying the cDNA, we
822 used RainDance Technologies' single-molecule micro-droplet based PCR¹³⁷. Droplet libraries
823 containing multiple overlapping primer pairs were designed to query all message forms and
824 allow later calculation of relative abundance (Supplementary Table S2 and Supplementary
825 Figure S1). Each primer was unique so that sequences could be assigned to a specific
826 primer pair, which helped reconstruct the origin of sequence reads and deduce message
827 structures. Amplified DNA products were sequenced using Single Molecule Real-Time
828 (SMRT) technology from Pacific Biosciences^{131,138}. We obtained 847 492 filtered reads of
829 amplified HIV-1 transcripts in primary CD4⁺ T cells and 89 350 in HOS cells. The longest

830 sequenced continuous stretch of HIV-1 cDNA was 2629 bp.

831 **3.4.2 Splice donors and acceptors**

832 We aligned PacBio reads containing HIV sequences to the HIV_{89.6} genome and identified
833 candidate introns as recurring gaps in our sequences. Using this approach, we observed
834 splicing at each of the widely conserved major splice donors and acceptors and several
835 published cryptic sites (Figure 3.1A, hereafter referred to by their identifications shown in
836 this figure, ‘D’ for donors, ‘A’ for acceptors).

837 In addition, we identified 13 putative novel splice sites: 2 donors and 11 acceptors (Figure
838 3.1 and Supplementary Table S3). In order to be selected as a bona fide splice site and
839 remove artifacts possibly created by recombination during sample preparation, we required
840 that the new acceptor or donor was observed spliced to previously reported splice donors or
841 acceptors in > 10 sequence reads in CD4⁺ T cells. The most frequently used novel splice site
842 was an acceptor that we have termed A8c because it lies near A8, A8a and A8b (discussed
843 in detail below). Additional novel sites are further discussed in Supplementary Report S1.

844 Most of the new splice sites adhered to consensus sequences for the standard spliceosome
845 (Supplementary Table S3). However, there appeared to be one splice donor upstream of
846 D1 with a cytidine in place of the usual uracil 2 nt downstream of the splice site. Similar
847 ‘GC donors’ appear in 1% of known splice junctions in humans¹³⁹. Of the novel splice
848 acceptors, three were preceded by dinucleotides other than the consensus AG. Alternative
849 dinucleotides are used infrequently as splice acceptors^{140–143}; however, it is possible that our
850 deep sequencing method allowed us to observe rare events.

851 **3.4.3 Structures of spliced HIV_{89.6} RNAs**

852 To quantify the populations of HIV-1 transcripts, we aligned all reads to the collection of
853 47 well-established spliced HIV-1 transcripts and detected 45 of them (Figure 3.2). We
854 additionally aligned reads to the HIV_{89.6} genome allowing all possible combinations of splice

855 junctions—canonical, cryptic or novel—determined from the sequencing data (Figure 3.1),
856 yielding an additional 32 complete transcripts, 19 of which were novel. The data also provide
857 evidence for more novel splice junctions but in incomplete sequences, implying the existence
858 of additional new transcripts (Supplementary Table S4 and Supplementary Report S1). The
859 full data set taken together provides evidence for least 109 different HIV_{89.6} transcripts in
860 primary T cells.

861 Amplification primers that isolated the two main classes of spliced messages allowed us to
862 determine the ratios of mRNAs in each (Figure 3.2 and Supplementary Table S5). Within
863 the partially spliced class of transcripts, *env/vpu*, *tat* (1-exon), *vpr* and *vif* messages existed
864 in an average ratio of 96:4:< 1:< 1 in CD4⁺ T cells. The ratio of *nef:rev:tat:vpr* within
865 the ~2 kb transcript class was 64:33:3:< 1. Consistent with previous reports, the most
866 abundant transcript in each class contained the splice junction from D1 to A5 (D1^A5)—an
867 *env/vpu* transcript contributing 64% of the partially spliced class, and a completely spliced
868 *nef* transcript contributing 47% of ~2 kb messages (Figure 3.2)^{105,144}. The relatively
869 low abundance of transcripts encoding Tat suggests that Tat sufficiently stimulates HIV
870 transcription elongation at low concentrations, or that the *tat* transcripts must be efficiently
871 translated. Due to biases inherent in the reverse transcription step, we could only compare
872 transcripts within each size class, and we note that our methods have not been validated
873 for empirical quantification. However, the ratios were roughly confirmed using overlapping
874 sequence reads obtained with alternate primer pairs and by end point RT-PCR analysis of
875 HIV-1 RNAs (data not shown).

876 Exons 2 and 3 are non-coding exons whose inclusion in transcripts other than *vif* and *vpr*
877 has no known function. We found that they were included in other messages infrequently,
878 each in ~7–8% of transcripts in the ~2 kb completely spliced class of transcripts and 5%
879 of partially spliced transcripts accumulating in T cells. This is consistent with previous
880 measurements in the partially spliced class but much lower than has been estimated for
881 completely spliced transcripts in HeLa cells, suggesting cell-type-specific splicing patterns

882 may influence inclusion of these exons¹⁰⁵.

883 **3.4.4 A novel ~1 kb class of completely spliced transcripts**

884 Primers placed near the 5'- and 3'-ends of the HIV_{89.6} genome amplified a second class of
885 completely spliced transcripts ~1 kb in length. In place of A7, these transcripts use a set of
886 little studied splice acceptors located ~800 bp downstream within the 3'-TR. Two groups
887 have previously observed splicing from D1 to acceptors A8, A8a and A8b in this region,
888 yielding messages of this size class in patient samples; however, none of these could be
889 translated to a protein of significant length^{107,111}. We determined the complete structure of
890 29 members of the 1-kb class (Figure 3.2 and Supplementary Table S5). The most abundant
891 messages observed in this class use the novel acceptor A8c to define their terminal exon. For
892 HIV89.6, acceptor A8c was used nearly as frequently as A7, which gives us the 2-kb class
893 of transcripts (Supplementary Table S3), and this was supported by end point RT-PCR
894 analysis (data not shown).

895 Acceptor A8c is not well conserved in HIV-1/SIVcpz (14%), although it is conserved in clade
896 G viruses (> 95%) and most HIV-2/SIVsmm genomes (86%)¹⁴⁵. This is due to the poor
897 conservation of an adenine at the wobble base position of the 123rd codon (proline) of the
898 Nef reading frame, which creates the AG dinucleotide generally required at splice acceptors.
899 Since any base at this position would code for proline, there does not seem to be strong
900 selection for a splice acceptor here. However, A8c is displaced from nearby well-conserved
901 (> 90%) cryptic acceptors A8a and A8b by multiples of 3 bp (12 and 21 bp, respectively),
902 so splicing to any of these three acceptors would create similar ORFs. All HIVs and SIVs
903 maintain at least one of these three acceptors, suggesting possible function¹⁴⁵. We confirmed
904 that the 1 kb transcripts using A8a, A8b and A8c were present in infected HOS and T cells
905 by end point RT-PCR using additional primer pairs and by Sanger sequencing of cloned
906 transcripts (Figure 3.3A and B; data not shown).

907 The 1-kb transcript containing exons 1, 4 and 8c (1-4-8c, where exon 8c begins at A8c

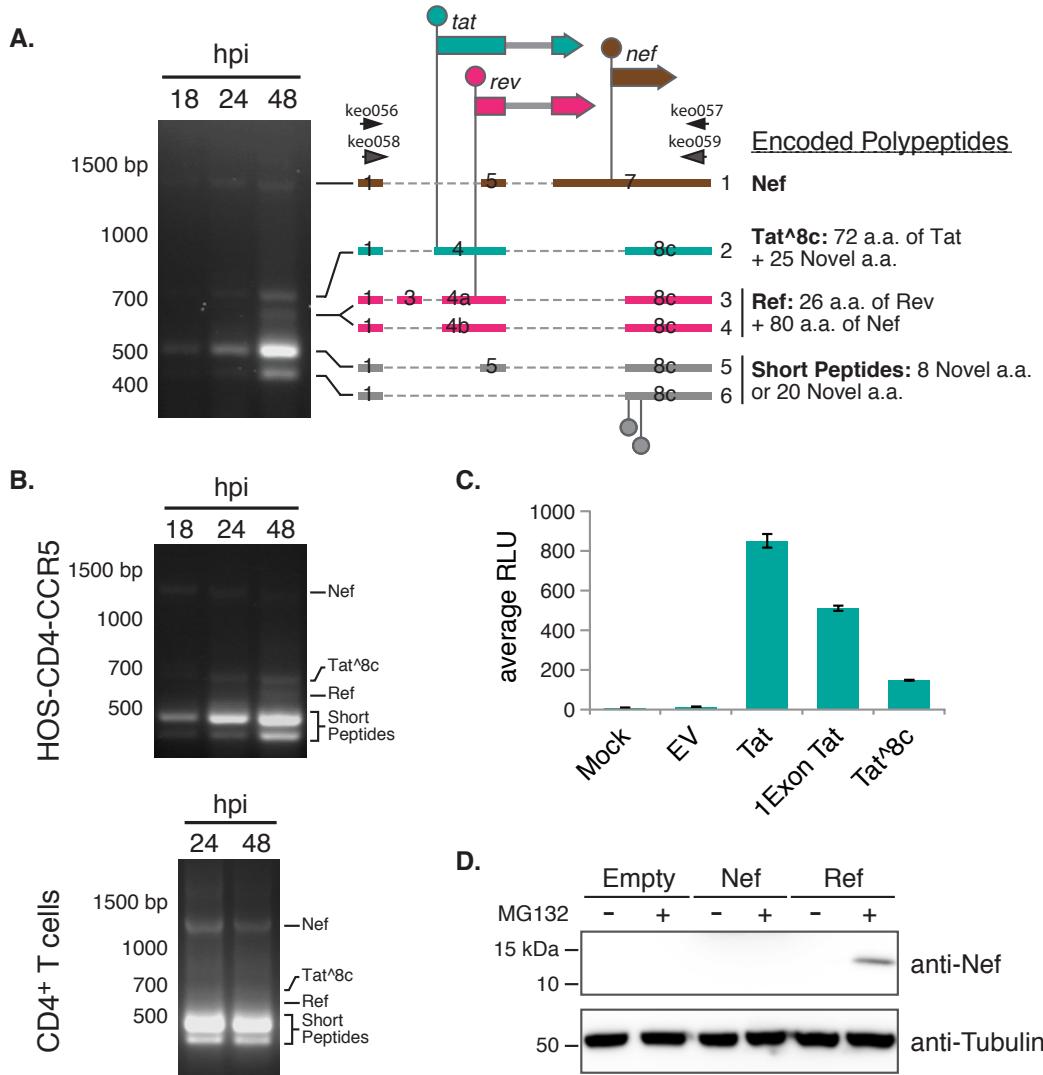


Figure 3.3: HIV_{89.6} transcripts were amplified by RT-PCR using RNA from infected HOS-CD4-CCR5 cells with primers keo056 and keo057. Major bands detected after gel electrophoresis were cloned from the 48 hpi sample and message structures determined by Sanger sequencing. Thick bars represent exons and dashed lines excised introns. Genes are shown above (not to scale) with start codons indicated by circles. Messages 1, 2, 4 and 5 were cloned into expression plasmids for activity assays. (B) Confirmation of presence of the ~1 kb message RNAs in HOS-CD4-CCR5 and primary CD4⁺ T cells (human donor 1, harvested 24 and 48 hpi). An independent primer pair (keo058 and keo059) was used to amplify transcripts by RT-PCR. (C) Tat activity was measured in Tzm-bl cells as Tat-dependent luciferase production after transient transfection with expression plasmids. (D) Western blot showing expression of protein of the predicted size for Ref (12.5 kb) in cells transfected with the Ref expression construct and treated with proteasome inhibitor MG132, detected by an antibody recognizing the carboxy-terminus of Nef. Expression plasmid encoding Nef was included to control for possible expression of partial Nef peptides or breakdown products from the Nef ORF.

908 and extends to the poly-adenylation site) encodes the first exon of Tat followed by 25
909 novel amino acids (termed Tat^{8c}). Tat^{8c} showed activity when overexpressed in cells
910 containing a Tat reporter construct (Figure 3.3C, nucleotide and amino acid sequences in
911 Supplementary Table S6). Transcripts with exon structures 1-4a/b/c-8c encode a novel
912 fusion of the amino-terminal 26 amino acids of Rev and the carboxy-terminal 80 amino acids
913 of Nef, hereafter referred to as Ref. We did not detect Rev activity on overexpression of
914 the *ref* transcript, and Ref did not appear to interfere with the normal function of Rev or
915 with HIV replication (Supplementary Figure S2). Ref was detectable by western blot using
916 antibodies targeting the C terminus of Nef after inhibition of the proteasome, suggesting
917 that the fusion is expressed but not stable (Figure 3.3D). Thus, Ref has the potential to
918 encode a new epitope potentially relevant in immune detection of HIV. The transcripts with
919 exon structures 1-5-8c and 1-8c encode at most a short peptide, and so are candidates for
920 acting as regulatory RNAs.

921 **3.4.5 Temporal dynamics of transcript populations**

922 To assess longitudinal variation, we investigated HIV_{89.6} transcript populations during the
923 course of a single round of infection in HOS-CD4-CCR5 cells. A sensitive method for
924 comparison among conditions involves quantifying utilization of six mutually exclusive splice
925 acceptors A3, A4c, A4a, A4b, A5 and a novel acceptor just downstream of A5 termed
926 A5a. Splicing at these acceptors determines the relative levels of messages encoding Tat
927 and Env/Vpu in the partially spliced class and messages encoding Tat, Rev and Nef in the
928 completely spliced class.

929 We observed longitudinal changes in the levels of these messages in HOS cells over 12–
930 48 h that were statistically significant ($p < 10^{-10}$; generalized linear model described in
931 Supplementary Report S2). This pattern was especially evident in junctions involving donor 1
932 spliced to each of these acceptors (Figure 3.4A). Most dramatically, transcripts with splicing
933 junctions between D1 and A3 (tat messages) increased with time ($p < 10^{-10}$), while D1^{8c}A4b
934 junctions (used in *env/vpu* or *rev* messages) were used reciprocally less ($p < 10^{-10}$). Such

935 kinetic changes affecting specific transcripts both with and without the Rev-response element
936 cannot be explained by the accumulation of Rev, and they may reflect differential transcript
937 stability or HIV-induced alterations to the host splicing machinery. Temporal changes in
938 HOS cells were confirmed using end point RT-PCR and analysis after electrophoresis on
939 ethidium-stained gels (Figure 3.4B).

940 **3.4.6 Cell-type-specific splicing patterns**

941 We also compared splicing between T cells and HOS cells and found significant cell type
942 differences ($p < 10^{-10}$). For example, while transcripts with D1^A5 junctions were dominant
943 in both cell types, messages using the D1^A4c splice junction (encoding Env/Vpu or Rev)
944 made up the bulk of the remaining transcripts in T cells but were a minor species in
945 HOS-CD4-CCR5 cells. Likewise, Tat messages (using A3), which were quite abundant in
946 HOS cells at all time points, contributed relatively little to populations of transcripts in
947 primary T cells harvested at 48 hpi (Figure 3.4A). We also used end point PCR and analysis
948 on ethidium-bromide-stained gels to confirm that the relative ratios of transcripts containing
949 junctions to A3, A4a, A4b and A4c were different in HOS and T cells (Figure 3.4B).

950 **3.4.7 Human variation in HIV-1 splicing**

951 Quantitative comparisons also revealed modest differences in splicing between primary CD4⁺
952 T cells isolated from different human donors that were statistically significant ($p < 10^{-10}$)
953 under a generalized linear model (Figure 3.4A). The magnitudes of predicted differences
954 were small, all < 33% and most < 10%.

955 **3.5 Discussion**

956 Use of single-molecule enrichment and long-read single-molecule sequencing has made possible
957 the most complete study to date of the composition of HIV-1 message populations, revealing
958 several new layers of regulation. Studies of the low-passage HIV89.6 isolate in a relevant cell
959 type showed numerous differences from studies of lab-adapted HIV strains in transformed

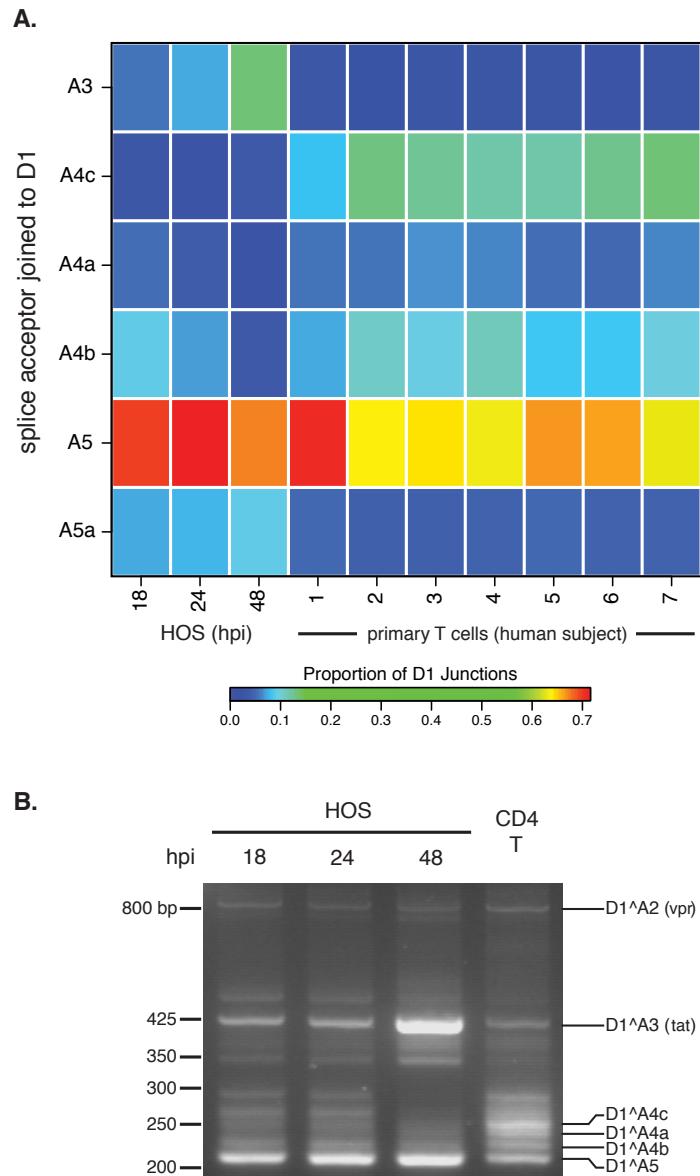


Figure 3.4: Temporal, cell type and donor variability in accumulation of HIV-1 messages. (A) In order to highlight changes in ratios of HIV-1 transcripts accumulating over time during infection and between HOS-CD4-CCR5 cells and primary T cells, we used PacBio read counts to calculate proportions of transcripts with splicing from the first major splice donor, D1, to each of the mutually exclusive acceptors: A3, A4c, A4a, A4c, A5 and the novel putative acceptor A5a. The heat map shows average data for T cell and HOS cell samples in columns with the color tiles indicating the proportion of D1 splicing to each of the mutually exclusive acceptors (rows), according to the color scale shown. (B) Reverse transcription and bulk PCR amplification of HIV_{89.6} transcripts from HOS cells and primary T cells from one human subject (subject 3) resolved by agarose gel electrophoresis and stained with ethidium bromide verified temporal and cell type changes shown in (A).

960 cell lines, highlighting the importance of studying the most relevant models. These data
961 also illustrate the limitations of gel-based assays for studying HIV-1 message population.
962 Multiple different combinations of HIV-1 exons yield mRNAs of similar sizes that are easily
963 confused in typical assays using gel electrophoresis. Thus, in many settings the more detailed
964 information provided by single-molecule amplification and single-molecule DNA sequencing
965 is more useful.

966 Using these methods, we have detected significant variations between HIV message pop-
967 ulations generated in T cells from different human donors. The differences were modest
968 compared to those observed between cell types or time points, perhaps not surprisingly
969 since any human polymorphisms strongly affecting mRNA processing might interfere with
970 normal gene expression. However, because tight calibration of message levels is important to
971 HIV-1, the observed differences in message ratios might affect HIV-1 acquisition or disease
972 progression. The variation in observed transcripts could also be affected by different kinetics
973 of infection in T cells from the different donors. In either case, these data suggest that human
974 polymorphisms may exist that affect HIV-1 message populations in infected individuals,
975 providing a new candidate mechanism connecting human genetic variation with measures of
976 HIV disease.

977 Sequences from the 89.6 viral strain revealed a class of small (~1 kb) completely spliced
978 transcripts, most contributed by splicing to a new poorly conserved acceptor A8c. These
979 encoded two new proteins, one of which had Tat activity, and we showed that another, a
980 Rev-Nef fusion termed Ref, could be detected in cells. HIV_{89.6} is a particularly cytotoxic virus
981 isolated from the CSF of a patient, and it forms unusually large syncitia in macrophages¹³⁰.
982 The abundance of 1-kb transcripts produced by this virus provides a possible explanation
983 for its unique properties. In addition to the novel acceptor A8c, we have also identified 3
984 putative novel splice donors and 11 putative novel acceptors, which require further studied
985 to clarify possible functions.

986 The wealth of new messages found here in HIV_{89.6} and in other HIV-1 isolates suggests there

may be ongoing evolution of novel splice sites and new ORFs. Because splice acceptors in HIV-1 are weak¹¹², mutations creating sequences that even slightly resemble the 3' splice site consensus may be occasionally recruited as novel acceptors, creating new mRNAs. In fact, new splice signals may evolve with relative ease—it has been estimated that reasonable matches to the consensus for splice donors, acceptors and branch-point sites occur within random sequence every 290, 490 and 24 bp, respectively¹⁴⁶, though sequence substitutions in HIV are usually also constrained by overlapping viral coding regions. We and others have observed appearance of novel exons within the major HIV-1 introns^{106,108,109}. Such long stretches of RNA relatively devoid of competing splice sites may be particularly poised to evolve new signals. On the other hand, most of the putative novel splice acceptors we observed clustered near previously identified acceptors in HIV-1, suggesting that conserved cis-acting splicing signals may recruit factors that act promiscuously on new nearby sequences. Clusters of splice sites might also provide redundancies that protect vital messages, as suggested previously^{147,148}. Frequent evolution of new splice sites may allow viruses to test out new combinations of exons, potentially yielding new RNAs and proteins, like those reported here. However, such novelty must compete with immune constraints—unstable novel polypeptides like Ref can be targeted to the proteasome and presented on MHC molecules as new epitopes for immune recognition.

HIV has likely evolved to produce calibrated message populations in T cells which seem to be altered with relative ease, as in infection in HOS cells, suggesting that therapeutic disruption of correct splicing may be feasible. A few studies have begun to explore small molecule therapy to disrupt HIV-1 splicing^{113,117}. Several factors could be responsible for the differences we observed between HOS and T cells, including hnRNP A/B and H, SC35, SF2/ASF and SRp40^{149,150}. Inhibition of SF2/ASF has already been shown to abrogate HIV-1 replication *in vitro*¹¹³. Thus the lability seen here for function of these factors suggests they may be attractive antiretroviral targets.

¹⁰¹³ **3.6 Acknowledgements**

¹⁰¹⁴ We would like to thank the University of Pennsylvania Center for AIDS Research (CFAR) for
¹⁰¹⁵ preparation of viral stocks and isolation of primary CD4⁺ T cells; James A. Hoxie, Ronald
¹⁰¹⁶ G. Collman, Jianxin You, Robert W. Doms, Paul Bates, David Rekosh and members of the
¹⁰¹⁷ Bushman laboratory for reagents, helpful discussion and technical expertise.

1018 **CHAPTER 4: Gene activity in primary T cells infected with HIV_{89.6}:**
1019 **intron retention and induction of distinctive genomic**
1020 **repeats**

This chapter is under review as:

S Sherrill-Mix, K Ocieja and F Bushman. Under Review.
Gene activity in primary T cells infected with HIV89.6: in-
tron retention and induction of distinctive genomic repeats.
Retrovirology

1021

KE Ocieja performed the infections and sequencing. I analyzed the data.
KE Ocieja, FD Bushman and I planned the overall study. I produced
the figures. FD Bushman and I wrote the paper.

1022 **4.1 Abstract**

1023 Background: HIV infection has been reported to alter cellular gene activity, but published
1024 studies have commonly assayed transformed cell lines and lab-adapted HIV strains, yielding
1025 inconsistent results. Here we carried out a deep RNA-Seq analysis of primary human T cells
1026 infected with the low passage HIV isolate HIV_{89.6}.

1027 Results: Seventeen percent of cellular genes showed altered activity 48 hours after infection.
1028 In a meta-analysis including four other studies, our data differed from studies of transcription
1029 after HIV infection of cell lines but showed more parallels with infections of primary cells.
1030 We found a global trend toward retention of introns after infection, suggestive of a novel
1031 cellular response to infection. HIV_{89.6} infection was also associated with activation of human
1032 endogenous retroviruses (HERVs) and several retrotransposons, of interest as possible novel
1033 antigens that could serve as vaccine targets. The most highly activated group of HERVs
1034 was a subset of the ERV-9, a group not reported previously to be induced by HIV. Analysis
1035 showed that activation was associated with a particular variant of an ERV-9 long terminal
1036 repeat that contains an indel near the U3-R border. These data also allowed quantification of
1037 >70 splice forms of the HIV_{89.6} RNA and specified the main types of chimeric HIV_{89.6}-host

1038 RNAs. Comparison to 147,281 integration site sequences from the same infected cells allowed
1039 quantification of authentic versus artifactual chimeric reads (0.1% of the total), showing
1040 that 5' read-in, splicing out of HIV_{89.6} from the D4 donor and 3' read-through were the most
1041 common HIV_{89.6}-host cell chimeric RNA forms.

1042 Conclusions: Analysis of RNA abundance after infection of primary T cells with the low
1043 passage HIV_{89.6} isolate disclosed multiple novel features of HIV-host interactions, notably
1044 intron retention and induction of transcription of distinctive retrotransposons and endogenous
1045 retroviruses.

1046 4.2 Background

1047 HIV replication requires integration of a cDNA copy of the viral RNA genome into cellular
1048 chromosomes, followed by transcription and splicing to yield viral mRNA. Alternative
1049 splicing allows the small 9.1 kb HIV genome to generate at least 108 mRNA transcripts
1050 encoding at least 9 proteins and polyproteins^{98,105,110,112,152,153}. During replication, HIV
1051 also reprograms cellular transcription and splicing. For example, the virus-encoded Vpr
1052 protein arrests the cell cycle^{154–157} and the viral Tat protein binds to P-TEFb and alters
1053 transcript at the HIV promoter and some cellular promoters^{158–163}.

1054 Multiple studies suggest that cells detect HIV infection and respond by inducing inter-
1055 feron-regulated, apoptotic and stress response pathways^{88,164–171}. Several studies have also
1056 suggested that HIV infection disrupts normal cellular splicing pathways^{127,171}. However,
1057 results have varied with many experimental parameters, including target cell type, HIV
1058 isolate and the duration of infection. Many of the published studies focused on infections
1059 with lab-adapted HIV strains in transformed cell lines^{88,164,171–174}, and so results may not
1060 be fully reflective of infections in patients.

1061 In this study, we sought to generate data more resembling HIV replication in patients
1062 by analyzing transcriptional responses after infection of primary T cells with HIV_{89.6}, a
1063 low passage patient isolate¹³⁰. This represents a continuation of a long term effort to

1064 understand HIV-host cell interactions at the transcriptional level that began with analysis
1065 of transcription by HIV_{89.6} in primary T cells using Pacific Biosciences long read single
1066 molecule sequencing⁹⁸. Our strategy here was to analyze a single time after infection in
1067 depth, analyzing over 1 billion sequence reads from HIV_{89.6} infected and uninfected host
1068 cells. These data were then combined with 147,281 unique integration site sequences from
1069 the same infections and the Pacific Biosciences data on HIV_{89.6} transcription to 1) elucidate
1070 effects of HIV infection on host cell mRNA abundances and splicing, 2) characterize viral
1071 message structure in detail and 3) probe the nature of the chimeras formed between host
1072 cell and viral RNAs.

1073 **4.3 Methods**

1074 **4.3.1 Cell culture and viral infections**

1075 HIV_{89.6} stocks were generated by the University of Pennsylvania Center for Aids Research.
1076 293T cells were transfected with a plasmid encoding an HIV_{89.6} provirus, and harvested virus
1077 was passaged in SupT1 cells once. Viral stocks were quantified by measuring p24 antigen
1078 content. Primary CD4⁺ T cells were isolated by the University of Pennsylvania Center
1079 for AIDS research Immunology Core from apheresis product from a single healthy male
1080 donor (ND365) using the RosetteSep Human CD4⁺ T Cell Enrichment Cocktail (StemCell
1081 Technologies).

1082 T cells were stimulated for 3 days at 0.5×10^6 cells per milliliter in R10 media (RPMI 1640
1083 with GlutaMAX (Invitrogen) supplemented with 10% FBS (Sigma-Aldrich) with 100 units
1084 U/mL recombinant IL2 (Novartis) + 5 μ g/mL PHA-L (Sigma-Aldrich)). Cells were infected
1085 in triplicate and mock infections were performed in duplicate. For each infection, 6.6×10^6
1086 cells were mixed with 1.32 μ g HIV_{89.6} in a total volume of 2.25 mL. Infection mixtures was
1087 split into three wells of a 6 well plate for spinoculation at 1200 g for 2 hr at 37°C. Cells were
1088 incubated an additional 2 hr at 37°C. Cells were then pooled into flasks and volume was
1089 increased to a total of 12 mL. Spreading infection was allowed to proceed 48 hr at 37°C,

1090 after which cells were harvested. 1×10^6 cells were harvested for flow cytometry, and 6×10^6
1091 cells were pelleted following two washes in PBS for nucleic acid extraction. Genomic DNA
1092 and total RNA were isolated from 6×10^6 T cells per infection using the AllPrep DNA/RNA
1093 Mini Kit (Qiagen) with Qiashredder columns (Qiagen) for homogenization according to the
1094 manufacturer's instructions. DNA was eluted in 140 μL elution buffer. RNA samples were
1095 treated with DNase prior to elution in 40 μL water.

1096 **4.3.2 Analysis of HIV_{89.6} integration sites in primary T cells**

1097 Integration site sequences were determined for DNA fractions from the above infections
1098 after ligation mediated PCR⁷³. A total of 147,281 unique integration site sequences were
1099 determined. An analysis of integration site distributions for these samples was reported in
1100 Berry et al.⁷³.

1101 **4.3.3 mRNA sequencing**

1102 Messenger RNA was isolated and amplified from purified total cellular RNA (3 μL or
1103 approximately 9 μg from each uninfected sample, 25 μL or approximately 3 μg from each
1104 infected sample) using the Illumina TruSeq RNA sample preparation kit according to
1105 manufacturer's protocol. SuperScript III (Invitrogen) was used for reverse transcription.
1106 Each sample was tagged with a separate barcode and sequenced on an Illumina HiSeq 2000
1107 using 100-bp paired-end chemistry.

1108 **4.3.4 Flow cytometry**

1109 To assess percent infected cells, 1×10^6 cells per infection were stained for flow cytometry.
1110 All staining incubations were at room temperature. Cells were first washed in PBS and
1111 then twice in FACS wash buffer (PBS, 2.5% FBS, 2 mM EDTA). Cells were fixed and
1112 permeabilized with CytoFix/CytoPerm (BD) for 20 minutes and washed with Perm-Wash
1113 Buffer (BD) before staining with anti-HIV-Gag-PE (Beckman Coulter) for 60 min. Finally
1114 cells were washed in FACS wash buffer and resuspended in 3% PFA. Samples were run

1115 on a LSRII (BD) and analyzed with FlowJo 8.8.6 (Treestar). Cells were gated as follows:
1116 lymphocytes (SSC-A by FSC-A), then singlets (FSC-A by FSC-H), then by Gag expression
1117 (FSC-A by Gag).

1118 **4.3.5 Analysis**

1119 Reads were aligned to the human genome using a combination of BLAT⁷² and Bowtie¹⁷⁵
1120 through the Rum pipeline¹⁷⁶. Estimates of fragments per kilobase of transcript per million
1121 mapped reads and changes in expression for cellular genes were calculated by Cufflinks⁷⁴.
1122 Reads found to contain sequence similar to the HIV genome using a suffix tree algorithm were
1123 aligned against the HIV_{89.6} genome using BLAT⁷². All statistical analyses were performed
1124 in R 3.1.2⁶⁵. RNA-Seq reads from Chang et al.⁸⁸ were downloaded from the Sequence Read
1125 Archive (SRP013224) and aligned using the Rum pipeline.

1126 Gene lists were obtained from the supplementary materials of four other studies of differential
1127 gene expression during HIV infection^{88,169,174,177}. We called genes differentially expressed
1128 in Li et al.¹⁷⁷ if they had a reported $p < 0.01$ or in Lefebvre et al.¹⁷⁴, Chang et al.⁸⁸
1129 and Imbeault et al.¹⁶⁹ if they had an adjusted $p < 0.05$. We called genes as differentially
1130 expressed in our own study if the adjusted $p < 0.01$. For the comparison of differentially
1131 expressed genes regardless of direction in figure 4.1 (below the diagonal), it was unclear
1132 exactly how many genes were studied in each study so we assumed a background of the
1133 14,192 genes (the number of genes which could be tested for significance in our data).

1134 We obtained transcriptional profiles comparing immune cell subsets from the Molecular
1135 Signatures Database¹⁷⁸. MSigDB set names from the MSigDB used in Figure 4.2A were
1136 GSE10325 LUPUS CD4 TCELL VS LUPUS BCELL, GSE10325 CD4 TCELL VS MYELOID,
1137 GSE10325 CD4 TCELL VS BCELL, GSE10325 LUPUS CD4 TCELL VS LUPUS MYELOID,
1138 GSE3982 MEMORY CD4 TCELL VS TH1, GSE22886 CD4 TCELL VS BCELL NAIVE,
1139 GSE11057 CD4 CENT MEM VS PBMC, GSE11057 CD4 EFF MEM VS PBMC, GSE3982
1140 MEMORY CD4 TCELL VS TH2 and GSE11057 PBMC VS MEM CD4 TCELL and in

1141 Figure 4.2B were GSE36476 CTRL VS TSST ACT 72H MEMORY CD4 TCELL OLD,
1142 GSE10325 CD4 TCELL VS LUPUS CD4 TCELL, GSE22886 NAIVE CD4 TCELL VS 12H
1143 ACT TH1, GSE3982 CENT MEMORY CD4 TCELL VS TH1, GSE17974 CTRL VS ACT
1144 IL4 AND ANTI IL12 48H CD4 TCELL, GSE24634 IL4 VS CTRL TREATED NAIVE CD4
1145 TCELL DAY5, GSE24634 NAIVE CD4 TCELL VS DAY10 IL4 CONV TREG, GSE1460
1146 CD4 THYMOCYTE VS THYMIC STROMAL CELL and GSE1460 INTRATHYMIC T
1147 PROGENITOR VS NAIVE CD4 TCELL ADULT BLOOD.

1148 We downloaded the RepeatMasker track from the UCSC genome browser¹⁷⁹ and used the
1149 SAMtools library¹⁸⁰ to assign reads to the repeat regions. HERV-K age estimates were
1150 obtained from the supplementary materials of Subramanian et al.¹⁸¹.

1151 We used a Bayesian estimate of the ratio of expression in uninfected and HIV infected
1152 samples to account for sampling effort and differing expression in genomic regions. We
1153 modeled the observed counts as a binomial distribution with a flat beta prior ($\alpha = 1, \beta = 1$)
1154 separately for uninfected and infected samples. We then Monte Carlo sampled the two
1155 posterior distribution to estimate the posterior distribution of the ratio. For introns, the
1156 number of binomial successes was set to the number of reads mapped to the intron and the
1157 number of trials was the total number of reads observed in the genes overlapping that intron.
1158 For repeat regions, the number of binomial successes was set to the number of reads mapped
1159 to that region and the number of trials was the total number of reads mapped to the human
1160 genome.

1161 To estimate determinants of LTR12C expression, we fit a logistic regression for which
1162 LTR12C increased in expression with HIV_{89.6} infection (95% Bayesian credible interval
1163 >1) on to characteristics of the LTR12C regions. We extracted all the LTR12C regions
1164 from the human genome and determined the U3-R boundary using a ends free alignment of
1165 the previously reported U3-R border^{182–186} against the sequences. Regions less than 1,000
1166 bases long were discarded. Previous studies disagreed about the location of the LTR12C
1167 transcription start site and it appears that transcription may start in several places^{183,184}.

1168 We took the 5' most site that had agreement between studies (transcription starting with
1169 TGGCAACCC). We split the sequences into short, medium and long length classes based
1170 on an indel about 70 bases upstream from the transcription start site. For each length class,
1171 we generated a consensus sequence and counted the Levenshtein edit distance between the
1172 consensuses and each corresponding sequence. We also counted the number of NFY motifs
1173 (CCAAT or ATTGG), MZF1 motifs (GTGGGG) and GATA2 motifs (GATA or TATC)
1174 in the entire U3 region or checked in any of the three motifs was present in the 150 bases
1175 upstream of the TSS. A final regression model was selected using stepwise regression with
1176 an AIC cutoff of 5. For display, the LTR12C sequences were aligned with MUSCLE¹⁸⁷.

1177 The abundance of the HIV RNA size classes was estimated as described in Additional File
1178 5. These estimates were then multiplied by the within size class proportions estimated by
1179 Ocwieja et al.⁹⁸ using PacBio sequencing of HIV_{89.6} to yield proportions over 78 measured
1180 HIV_{89.6} RNAs.

1181 4.4 Results

1182 4.4.1 Infections studied

1183 HIV_{89.6}, a clade B primary clinical isolate¹³⁰, was used to infect primary CD4⁺ T cells from
1184 a single human donor in three replicate infections. For comparison, two additional replicates
1185 from the same donor were mock infected. Samples were harvested after 48 hours of infection,
1186 which allowed for widespread infection in the primary T cell cultures, though some cells may
1187 be infected secondarily by viruses produced in the first round. Thus cultures probably were
1188 not tightly synchronized but did have extensive representation of infected primary T cells.
1189 From these samples, we obtained 1,161,705,678 101-bp reads from primary CD4⁺ T cells
1190 from a single donor; 1,021,207,853 were mapped to the human genome and 24,783,844 to
1191 the HIV_{89.6} provirus (Table 4.1). Below we first discuss the influence of infection on cellular
1192 gene activity and RNA splicing, then analyze HIV RNAs and lastly analyze chimeras formed
1193 between HIV and cellular RNAs.

Sample	Infection rate (%)	Reads	Human reads	HIV reads	% HIV	% HIV in infected
Uninfected-1	—	232,450,106	212,391,460	—	—	—
Uninfected-2	—	235,048,212	203,760,783	—	—	—
Infected-1	37.5	234,378,088	199,871,662	10,219,315	4.86	13.0
Infected-2	26	226,078,422	198,436,507	7,322,556	3.56	13.7
Infected-3	21	233,750,850	205,747,441	7,241,973	3.40	16.2

Table 4.1: Samples used in this study, their infection rates and sequencing depth.

1194 **4.4.2 Changes in gene activity in primary T cells upon infection with HIV_{89.6}**

1195 Changes in host cell gene expression have been reported during HIV infection^{88,164–172,174,188}
 1196 and differences in expression have been observed associated with the stage¹⁷⁷ and progres-
 1197 sion¹⁸⁹ of disease. Here we observed significant changes in gene expression (false discovery
 1198 rate corrected $q < 0.01$) in 3,142 genes, 17.1% of expressed cellular genes (Additional file 1).
 1199 The genes with most extreme increases, all $>6\times$ fold higher, during HIV infection included
 1200 IFI44L, RSAD2, HMOX1, MX1, USP18, IGJ, OAS1, CMPK2, DDX60, IFI44, IFI6, IFNG
 1201 and CCL3. All of these have been reported to be involved in innate immunity¹⁹⁰ or are
 1202 interferon inducible¹⁹¹, highlighting a strong innate immune response in the cells studied.
 1203 Genes with the largest decreases, all $>3\times$ fold lower, were GNG4, GPA33, IL6R, CCR8,
 1204 RORC, AFF2 and CCR2.

1205 Many gene ontology categories were significantly enriched for differentially expressed genes
 1206 (Additional file 2). Notably upregulated with infection were genes involved in apoptosis,
 1207 immune responses and cytokine production (all $q < 10^{-4}$) and down-regulated were genes
 1208 involved in viral gene expression, nonsense-mediated decay and translation elongation and
 1209 termination (all $q < 10^{-19}$). These changes suggest that the cells responded to HIV infection
 1210 with the induction of inflammatory, interferon regulated and apoptotic responses, patterns
 1211 posited from several previous studies^{88,164–170,173,174,192}. Several genes were activated that
 1212 were characteristic of other hematopoietic lineages, e.g. hemoglobin β , CD8, CD20 and
 1213 CD117, while several CD4 $^+$ T cell specific genes, e.g. CD4 and CD3, were downregulated,
 1214 potentially consistent with de-differentiation of infected and bystander cells. We return to

Cell type	HIV type	Differentially expressed genes (Up/Down)	Study
Primary CD4 ⁺ T	HIV _{89.6}	3393 (1756/1637)	This study
Primary CD4 ⁺ T	NL4-3 BAL-IRES-HSA	228 (182/46)	Imbeault et al. ¹⁶⁹
Lymph node biopsies	Acute infection	448 (383/65)	Li et al. ¹⁷⁷
SupT1	HIV _{LAI}	4997 (2666/2331)	Chang et al. ⁸⁸
SupT1	NL4-3Δenv-eGFP/VSV-G	579 (212/367)	Lefebvre et al. ¹⁷⁴

Table 4.2: Data from this study and four others used for meta-analysis of human gene expression changes during HIV infection

1215 this point in the discussion.

1216 **4.4.3 Comparison of transcriptional profiles from HIV_{89.6} infection of primary**
 1217 **T cells to data on HIV infection in other cell types**

1218 We sought to identify the transcriptional responses that were most conserved upon HIV
 1219 infection and so collected and analyzed data from four other studies of transcription in
 1220 HIV-infected cells (Table 4.2). These included two studies of infection of the SupT1 cell
 1221 line^{88,174}, a study of primary CD4⁺ T cells¹⁶⁹ and a study of lymphatic tissue in acutely
 1222 viremic patients¹⁷⁷. Genes were scored as increased or decreased in activity after infection,
 1223 and the amount of agreement was compared among the different studies.

1224 No gene was called as differentially expressed in all five studies. Eight genes were differentially
 1225 expressed in the same direction in 4 out of 5 studies; AQP3 and EPHX2 were down-regulated
 1226 with HIV infection and CD70, EGR1, FOS, ISG20, RGS16 and SAMD9L were up-regulated.
 1227 A full listing is provided in Additional file 4. Several of the up-regulated genes are known to
 1228 be interferon inducible, again emphasizing the role of innate immune pathways.

1229 For each pair of studies, we compared whether they agreed on the identities of differentially
 1230 expressed genes and whether they agreed on the direction of change (Figure 4.1). The
 1231 estimated alterations in gene activity showed notable differences in the responses to infection
 1232 in primary cells versus the SupT1 cell line. The two SupT1 studies were significantly similar
 1233 ($p < 10^{-15}$) to each other but were not significantly associated (Lefebvre et al.¹⁷⁴, $p = 0.2$)
 1234 or were negatively associated (Chang et al.⁸⁸, $p = 10^{-7}$) with data from lymphatic tissue in

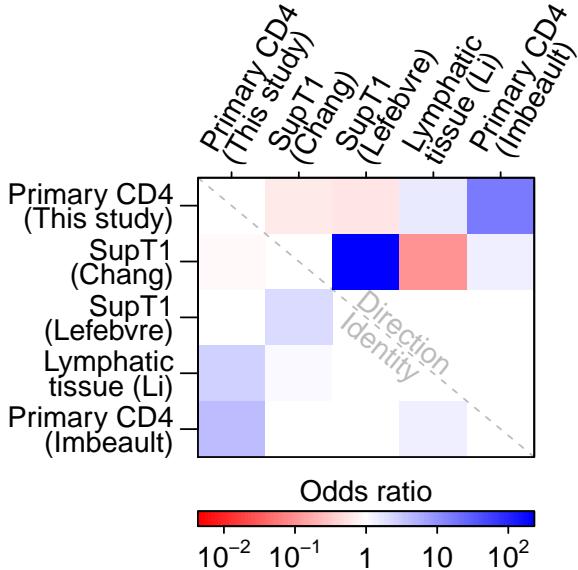


Figure 4.1: Comparisons among studies quantifying cellular gene expression after HIV infection. For each pair of studies, the association between up- and down-regulation calls was measured for genes identified by both studies as differentially expressed (above the diagonal). As another comparison, we also measured the agreement between studies for which genes were called differentially expressed regardless of direction (below the diagonal). The color scale shows the conservative (i.e. closest to 1) boundary of the confidence interval of the odds ratio with blue indicating a positive association and red a negative association between studies. For confidence intervals overlapping 1, the value was set to 1. Therefore all colored squares indicate significant associations.

acute HIV patients. The primary T cell study reported here was significantly associated with the second study in primary cells ($p < 10^{-15}$) and with a study of lymphatic tissue from patients acutely infected with HIV ($p = 0.003$). Our primary T cell data was negatively associated with the SupT1 studies (both $p < 10^{-3}$). This documents significant differences in responses to HIV infection between infected primary cells and SupT1 cells and suggests that results of infections in primary cells more closely align with actual acute HIV infections in patients. SupT1 cells might be expected to respond to infection differently than primary cells since they have several nonsynonymous mutations in innate immunity genes¹⁹³, have blocks in immune signaling pathways¹⁹⁴ and fail to activate many interferon stimulated genes during HIV infection¹⁷⁰.

4.4.4 Comparison of the HIV infected cell transcriptional profiles to additional experimental T cell profiles

To investigate the transcriptional changes in more depth, we compared the results of the five studies of HIV infection to transcriptional profiles comparing immune cell subsets available at the Molecular Signatures Database (MSigDB)¹⁷⁸. The MSigDB reports genes that are

1250 increased or decreased in relative expression for each of 185 pairs of transcriptional profiles
1251 involving CD4⁺ T cells. We compared the lists of affected genes in each pair to genes altered
1252 in activity by HIV infection. Those pairs of studies with the most significant associations
1253 with HIV_{89.6} data are show in Figure 4.2A. For comparison, the associations with the four
1254 other HIV transcriptional profiling studies mentioned above are shown as well.

1255 The most significant associations for our data showed gene expression in HIV_{89.6}-infected
1256 cells moving away from typical T cell expression patterns and towards patterns more similar
1257 to B cells, myeloid cells and bulk peripheral blood mononuclear cells (all Fisher's $p < 10^{-15}$)
1258 (Figure 4.2A). These changes were also seen, although to a lesser extent, in the Imbeault
1259 et al.¹⁹⁵ study which also used primary CD4⁺ T cells.

1260 For comparison, we also extracted those profiles most strongly associated with the transcrip-
1261 tional data on lymphatic tissue of HIV patients¹⁷⁷. The profiles showed patterns similar to
1262 strongly stimulated T cells, autoimmune disease and to the Th1 T cell subset (all $p < 0.01$)
1263 (Figure 4.2B). Our data in primary CD4⁺ T cells paralleled the changes seen in lymphatic
1264 tissue. These transcriptional changes again highlights the strong immune response generated
1265 by HIV infection in primary cells.

1266 4.4.5 Intron retention

1267 Cells respond to infection by shutting down macromolecular synthesis at multiple levels^{196–200},
1268 so we investigated whether cells also showed perturbations in splicing efficiency after infection.
1269 As a probe, we created a database of cellular genomic regions annotated exclusively as exons
1270 or introns in all spliceforms in the UCSC gene database⁸⁴ and quantified expression in these
1271 regions in infected and uninfected cells. We found a significant increase in intronic sequences
1272 relative to exonic sequence (Wilcoxon $p < 10^{-15}$) (Figure 4.3A). This increase in intronic
1273 sequence was reproducible between replicates in our study (Kendall's $\tau=0.42$, $p < 10^{-15}$)
1274 (Figure 4.3B). We reanalyzed RNA-Seq data from Chang et al.⁸⁸ and also documented intron
1275 retention which correlated with the changes seen in our data (Kendall's $\tau=0.12$, $p < 10^{-15}$)

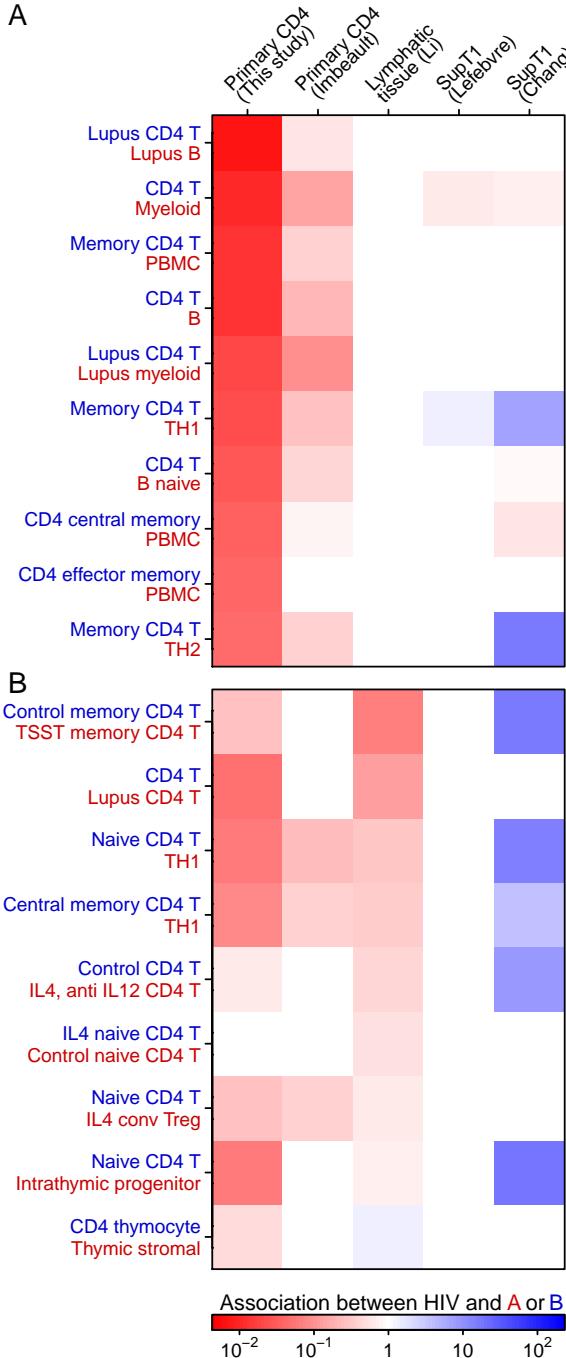


Figure 4.2: Comparisons of the effect of HIV infection on gene expression to studies comparing subsets of immune cells. The MSigDB database was used to extract 185 sets of differentially expressed genes from pairs of transcriptional profiling studies of immune cell subsets involving CD4⁺ T cells. For each pair of studies, we used Fisher's exact test to measure the association between up- and down-regulation calls for genes identified as differentially expressed in both our HIV study and the comparator immune subsets. A) The transcriptional profiles with strongest associations with changes observed in our study of HIV_{89.6} infection of primary T cells. Blue indicates a positive association between changes seen in HIV infected cells and the first immune subset (text colored blue) while red indicates a positive association with the second immune subset (text colored red). The color scale shows the conservative (i.e. closest to 1) boundary of the confidence interval of the odds ratio. For confidence intervals overlapping 1, the value was set to 1. Therefore all colored squares indicate significant associations. B) As in A, but showing the transcriptional profiles most strongly associated with changes observed in lymph node biopsies from acutely infected patients¹⁷⁷.

1276 (Figure 4.3C).

1277 A possible artifactual explanation for enrichment of intronic sequences could involve greater
1278 DNA contamination in the infected cells samples. That is, if the relative amount of DNA
1279 differed between treatments, the amount of apparent intronic sequences could also differ
1280 due to sequencing of contaminating DNA. To examine whether DNA contamination was
1281 abundant in our samples, we compiled a collection of 27 large gene desert regions, defined
1282 here as 1) regions outside the centrosome and first and last cytoband, 2) containing less than
1283 1% unknown sequence, 3) containing no genes annotated in UCSC genes⁸⁴, 4) containing
1284 no repeats annotated in the repeatMasker database⁹⁰ and 5) spanning more than 100 kb.
1285 No reads were mapped to these 41 Mb of gene deserts in any sample, arguing against
1286 explanations based on DNA contamination. Thus these data indicate that intron retention
1287 was increased in these cell populations upon HIV infection, revealing a previously undisclosed
1288 aspect of the host cell transcriptional response to infection.

1289 Previous studies have reported changes in the expression and localization of splicing factors
1290 with HIV infection^{127,201,202}. In our data, HIV_{89.6} infection significantly altered the expression
1291 of genes involved in RNA splicing ($p = 2 \times 10^{-7}$) and nonsense-mediated decay ($p < 10^{-15}$).
1292 Genes related to nonsense-mediated decay genes showed a strong pattern of lowered RNA
1293 abundance, with 71 out of 118 annotated genes significantly lower in expression after infection.
1294 These patterns suggest potential mechanisms for the intron retention observed here.

1295 **4.4.6 Induction of transcription from HERVs and LINEs by HIV_{89.6} infection**

1296 HIV infection has been reported to induce expression of certain HERVs, particularly HERV-
1297 K^{203–205}, and LINE and Alu transposable elements²⁰⁶, providing candidate markers of
1298 infection and possible vaccine targets. Thus we analyzed our data in primary T cells infected
1299 with HIV_{89.6} to investigate the expression of HERVs, LINEs and other repeated sequences.
1300 Figure 4.4A shows a comparison of the association between changes in expression with
1301 HIV_{89.6} infection and the various genomic repeat types over varying levels of differential

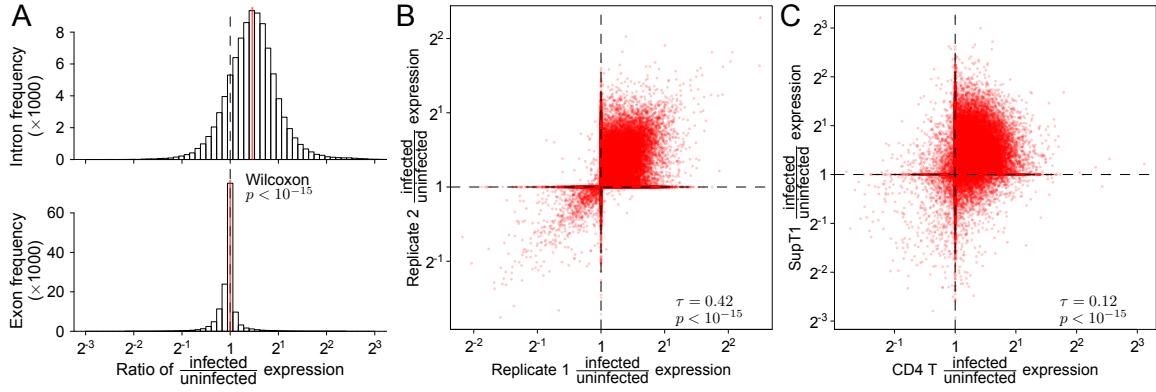


Figure 4.3: Changes in the abundance of intronic regions with HIV infection. Expression of intronic and exonic regions was quantified as the proportion of reads mapping within the intron/exon out of the total reads mapping to the transcription units overlapping that intron/exon. A) Comparison of the ratios of expression between infected and uninfected replicates in exclusively intronic or exonic regions of transcription units. B) Reproducibility of intron retention between replicates. Each point quantifies the change in expression with HIV infection for a specific intronic region. The x-axis shows changes in gene activity accompanying infection for one set of replicates (Infected-1 and Infected-2 vs. Uninfected-1) and the y-axis shows the same data for different replicates (Infected-3 vs. Uninfected-2). C) Reproducibility of intron retention between studies. The plot is arranged as in B but with all data from our study combined on the x-axis and corresponding data from Chang et al.⁸⁸ on the y-axis.

1302 expression. At high levels of expression, ERV-9 (odds ratio at $4\times$ expression: 152, 95%
 1303 CI: 82.5–259) and its long terminal repeat LTR12C (odds ratio at $4\times$ expression: 144, 95%
 1304 CI: 98.2–207) are the only repeats highly associated with upregulation during HIV infection.
 1305 Looking at genomic repeats with any significant increase, the expression of many recently
 1306 acquired genomic repeats, including L1HS, LTR5_Hs (a human specific LTR of HERV-K),
 1307 AluYa5, AluYg6 and SVA_D and SVA_F, were associated with HIV_{89.6} infection (Figure
 1308 4.4B).

1309 We saw a relationship between the age of genomic repeats and its likelihood of being induced
 1310 by HIV_{89.6} infection. The most highly enriched repeats were associated with relatively
 1311 recent hominid-specific repeat classes as annotated by the RepeatMasker database (repeat
 1312 classes with $p < 10^{-50}$ odds ratio: 31.6, 95% CI: 8.88–112). In HERV-K (HML-2), the
 1313 most recently active endogenous retrovirus in the human genome^{181,207,208}, we saw that
 1314 integrations unique to the human genome¹⁸¹ were more likely to be differentially expressed

1315 than older HERV-Ks (odds ratio: 5.38, 95% CI: 1.93–16.0).

1316 Previous RNA-Seq studies of cellular expression during HIV infection in transformed cell
1317 lines did not report increases in HERV mRNA^{88,174}. To investigate this difference, we
1318 downloaded and analyzed the RNA-Seq data from Chang et al.⁸⁸, which quantified gene
1319 activity in transformed SupT1 cells infected with a lab-adapted strain of HIV. We found a
1320 much higher level of HERV expression in their data in both HIV infected cells and uninfected
1321 controls than in primary cells (Figure 4.4C). We suspect that in SupT1 cells, as with many
1322 cancerous cells^{209–213}, the baseline expression of transposons and endogenous retroviruses is
1323 higher than in primary cells, masking further induction by HIV infection.

1324 We observed heterogeneous expression among ERV-9/LTR12C sequences and so investigated
1325 the primary sequence determinants. We observed that ERV-9/LTR12C has three variants of
1326 differing length in the U3 region just upstream of the transcription start site (Figure 4.5A),
1327 an important region for transcription initiation¹⁸³. The U3 region of LTR12C also contains
1328 multiple motifs for transcription factors NFY, GATA2 and MZF1¹⁸⁶. To clarify factors
1329 affecting expression levels, we counted the number of motifs matching these transcription
1330 factors, assigned each LTR12C to one of the length classes, counted the number of mutations
1331 away from the consensus for that length class and checked for integration in a transcription
1332 unit. We then carried out a regression analysis to test the effects of these variables on
1333 LTR12C differential expression. We found that HIV_{89.6} induced transcription was more
1334 likely with the fewer mutations away from consensus, the number of locations matching the
1335 NFY transcription factor binding motif (CCAAT) and LTRs containing the short length
1336 variant of the 3' U3 region. The presence of a MZF1 motif near the transcription start site
1337 decreased transcription (Figure 4.5B).

1338 4.4.7 HIV mRNA synthesis and splicing

1339 Over 24 million Illumina reads mapped to HIV_{89.6}, yielding an average coverage of over
1340 240,000-fold. Reads mapping to HIV_{89.6} comprised between 3.4–4.8% of mapped reads in

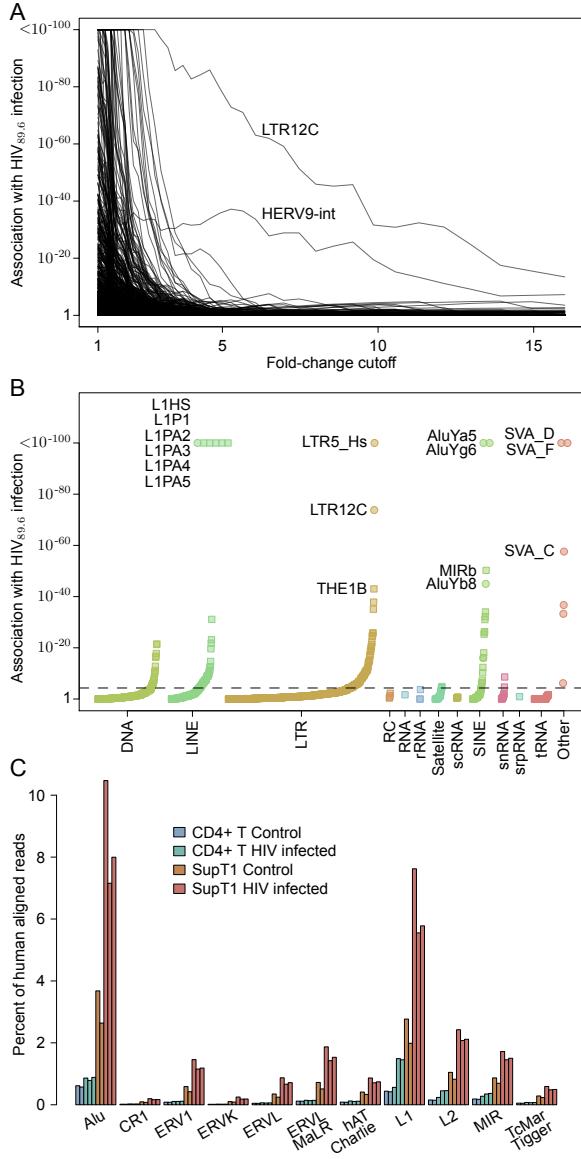


Figure 4.4: Repeat categories enriched upon infection with HIV. A) The association of repeat regions differentially expressed after HIV_{89.6} infection of primary T cells observed for varying thresholds of differential expression. The threshold used to call a gene differentially expressed based on the Bayesian posterior median was varied and Fisher's exact test was used to assess whether any genomic repeats had a significant association with this differential expression. Note that only ERV-9 (annotated as HERV9-int in the RepeatMasker database) and its corresponding long terminal repeat LTR12C were significantly associated with large changes in expression. B) Enrichment of repeat categories in regions differentially expressed (Bayesian 95% credible interval >1) between HIV-infected and control CD4⁺ T cells. The repeated sequences are ordered on the x-axis by the extent of induction within each class, the y-axis shows the p-value for upregulation after infection. The dashed line indicates a Bonferroni corrected p value of 0.05. (C) The proportion of human mapped reads that align within classes of genomic repeats for data from primary CD4⁺ T cells from this study and SupT1 cells from Chang et al.⁸⁸. A single read mapping multiple times to a given category was only counted once.

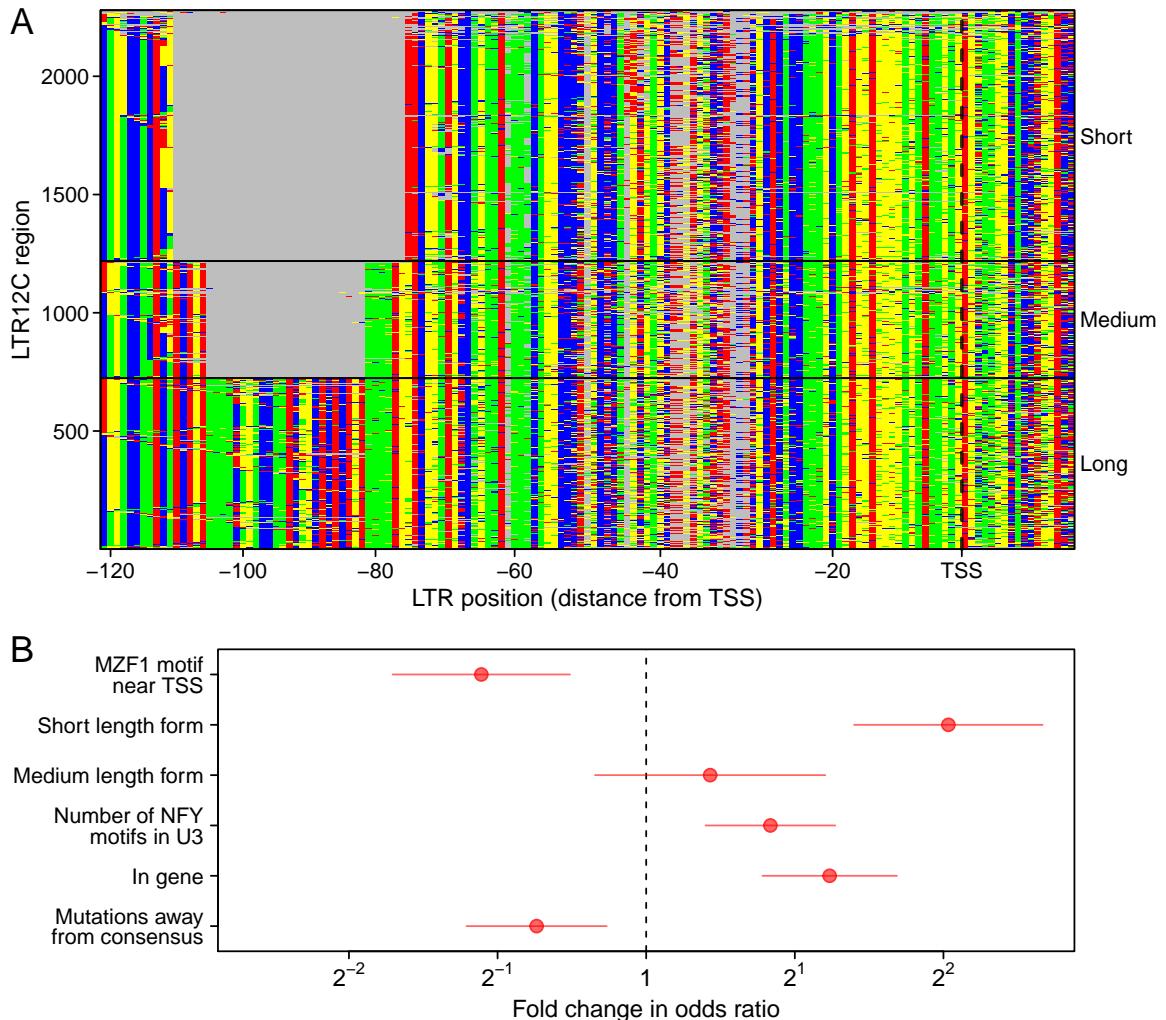


Figure 4.5: Characteristics of LTR12C sequences associated with induction upon infection of primary T cells with HIV_{89.6}. A) An alignment of the 3' end of the U3 region of repeats annotated as ERV-9 LTR12C. Each row is a LTR sequence and each column a base in that sequence colored by nucleotide identity. Three distinct classes are visible with a short, medium and long form. Mutations away from the consensus can also be seen. B) The coefficients (points) and ± 1.96 standard errors (horizontal lines) of a logistic regression comparing differential expression of LTR12C to the presence of MZF1 and NFY motifs, short/medium/long length alternate forms of the U3-R region, mutations away from the consensus for each length form and integration inside a transcription unit. The coefficient shown for mutations away from consensus is for a 10 mutation difference and the coefficient shown for NFY motifs is for a change of 5 additional motifs. All other coefficients are for binary values.

1341 the infected samples (Table 4.1). Assuming HIV-infected cells contain the same amount of
1342 mRNA as uninfected cells and adjusting for rates of infection ranging between 21–37.5%
1343 (Table 4.1), we estimate that HIV transcripts comprise between 13.0–16.2% of the total
1344 polyadenylated mRNA nucleotides in infected cells 48 hours after initial infection. This
1345 parallels previous estimates of around 10%²¹⁴ at 48 hours postinfection, 38% at 24 hours⁸⁸
1346 or 30% after 72 hours¹⁶⁴.

1347 Over 47,257 single reads spanned previously reported HIV splice junctions, allowing a
1348 quantitative assessment of donor and acceptor utilization (Figure 4.7A). As expected from
1349 previous studies^{98,105}, the most abundant junctions were D1-A5 and D4-A7. We confirmed
1350 the use of unusual splice acceptors A8c and A5a, previously reported in HIV_{89.6}⁹⁸. In our
1351 data, we also see a higher abundance of D1-A1 and D1-A2 splice junctions than might
1352 be expected^{98,105}, although previous studies reported proportional abundance within size
1353 classes, making comparisons between size classes uncertain.

1354 A 3' bias is apparent in our sequencing data (Additional file 5). This could be due to the
1355 poly-A capture step of the protocol where any break in the RNA would result in distal
1356 5' sequences being lost²¹⁵. We used sequence reads from the large unspliced HIV intron
1357 1 to measure this bias using a regression of the log of the number of fragments with a
1358 5'-most end starting at a given position against the distance of that position from the
1359 viral polyadenylation site, yielding an estimated probability of breakage of 0.021% per base
1360 (Additional file 5). Given this rate of termination, there is only a 14% chance of reaching
1361 the 5' end of the 9171 nt unspliced HIV genome ($(1 - 0.00021)^{9171}$).

1362 Ocwieja et al.⁹⁸ determined the relative abundance of HIV_{89.6} of similarly sized transcripts
1363 using PacBio single molecule sequencing, but were not able to estimate the relative abundance
1364 of all transcripts due to a sequencing bias favoring shorter transcripts. For this reason,
1365 relative abundances could only be specified within message size classes (i.e. the 4 kb, 2 kb
1366 and unexpectedly a 1 kb size class as well) and the overall quantitative abundances were
1367 unknown. The RNA-Seq data reported here are unable to determine complete transcript

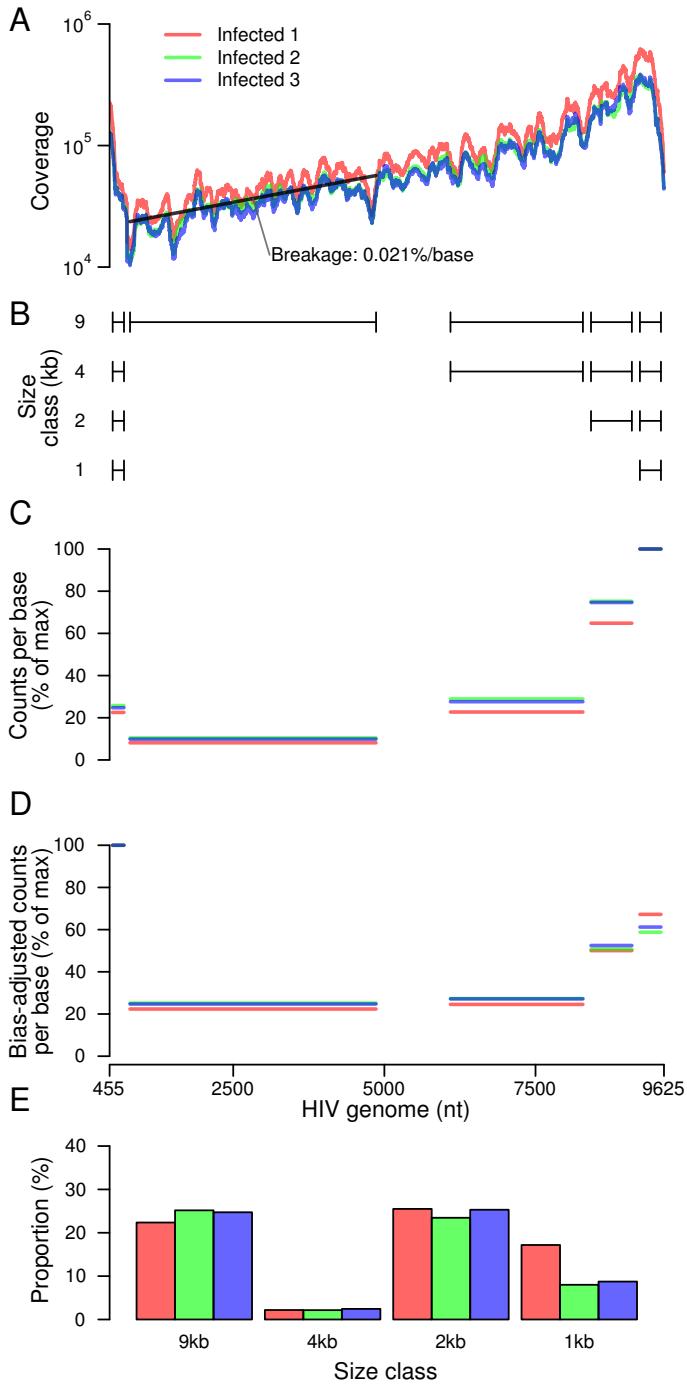


Figure 4.6: Estimating relative abundance of HIV_{89.6} message size classes using RNA-Seq data.

A) RNA-Seq coverage of the HIV_{89.6} genome for the replicates in this study. Each replicate is indicated by a different color. The HIV genome is shown on the x-axis and the number of reads that aligned to each position is shown on the y-axis. Black line indicates the 0.021% coverage decrease per base distance from the 3' end of the mRNA estimated from a least squares fit on the read counts in the first intron.

B) Diagram of the segments of the HIV_{89.6} RNA present in each of 9 kb, 4 kb, 2 kb and 1 kb size class.

C) The proportion of reads mapped to each of the segments of the HIV_{89.6} genome shown in B adjusted by the length of the segment. Each replicate is shown by a different color.

D) Corrected representation of RNA segments from the different size classes. Because cDNA synthesis was primed from the polyA tail, more 3' sequences are recovered preferentially. Using the bias estimate from A, we adjusted each genome segment by the inverse of the bias predicted based on its distance from the 3' end of the mRNA. Corrected proportions for the indicated RNA segments are shown colored by replicate.

E) The proportion of each size class was inferred using the estimates in D by calculating the difference between segments. Replicates are indicated by color.

abundance because the short read length does not allow reconstruction of multiply spliced messages but do permit estimation of size class abundances after correcting for 3' bias (Additional file 5). Thus the PacBio data reported by Ocwieja et al.⁹⁸ and the Illumina data reported here can be combined together to determine complete relative abundance of all HIV_{89.6} transcripts (Figure 4.7B).

The most abundant HIV mRNAs were the unspliced HIV genome (37.6%), a transcript encoding Nef (D1-A5-D4-A7: 15.5%), two 1 kb size class transcripts (D1-A5-D4-A8c: 10.6%, D1-A8c: 4.9%) and two Rev-encoding transcripts (D1-A4c-D4-A7: 4.2%, D1-A4b-D4-A7: 3.1%). The function of this large amount of 1 kb transcript is unknown. These two 1 kb transcripts do not appear to encode significant open reading frames although other 1 kb transcripts can encode a Rev-Nef fusion⁹⁸.

Using these abundances, we can estimate the number of HIV_{89.6} genomes in these primary T cells 48 hours after infection. To determine the proportion of the mRNA nucleotides from viral transcripts, we multiplied the estimated abundances by their transcript lengths. Unspliced genome transcripts appear to form 79% of the mRNA nucleotides from HIV_{89.6} transcripts. Assuming T cells contain at least 0.1 pg of mRNA then an infected cell should contain at least 0.011 pg of unspliced HIV transcript ($0.1\text{pg} \times 0.14 \frac{\text{HIV mRNA nt}}{\text{cell mRNA nt}} \times 0.79 \frac{\text{unspliced mRNA nt}}{\text{HIV mRNA nt}}$) or, assuming 9171 bases of RNA weigh about 5×10^{-6} pg, at least 2200 HIV genomes at 48 hour post infection. This estimate roughly agrees with previous estimates of HIV production per cell^{214,216,217}.

4.4.8 Human-HIV chimeric reads

The suggestion that HIV integration may disrupt cellular cancer-associated genes and thereby promote cell proliferation^{218–221} has focused attention on the range of novel message types formed when HIV integrates within transcription units^{33,102,222–224}. Chimeric reads containing HIV and cellular sequence are also of clinical interest due to the potential of lentiviral vectors to trigger oncogenesis in gene therapy patients through insertional

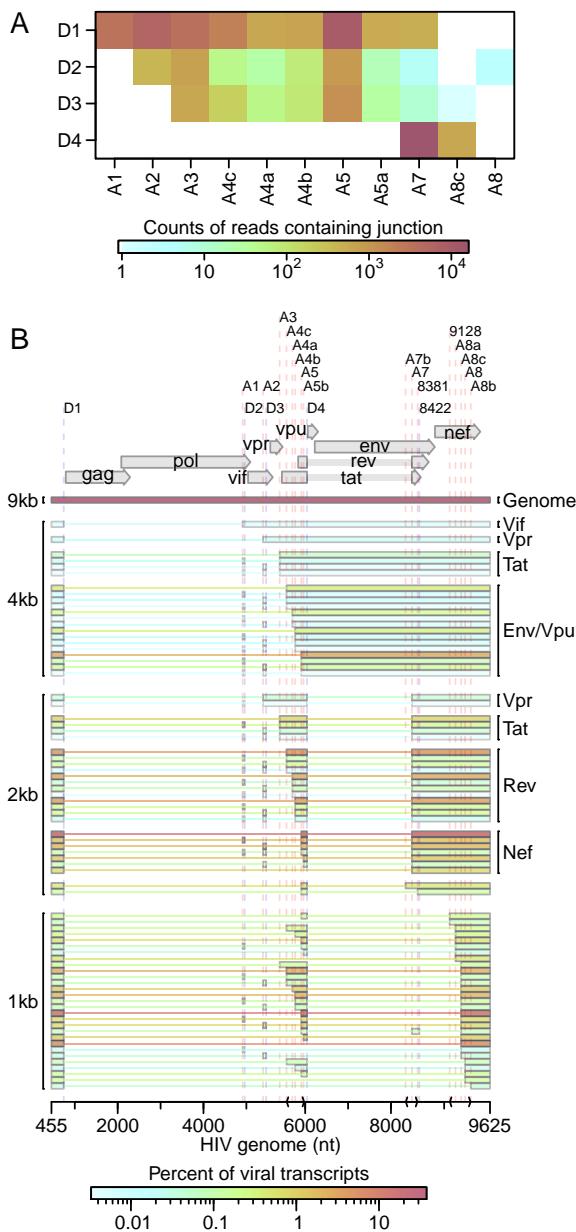


Figure 4.7: Transcription and splicing of the HIV_{89.6} RNA. A) Junctions between HIV splice donors and acceptors observed in the RNA-Seq data. Acceptors are shown as the columns and donors as the rows with the coloring indicating the frequency of each pairing. B) The relative abundance of all HIV_{89.6} transcripts as determined by a combination of PacBio sequencing⁹⁸ and Illumina sequencing. Message structures were generated by targeted long read single molecule sequencing, which allowed association of multiple splice junctions in single sequence reads. The Illumina short read sequencing allowed normalization of message abundances between size classes. The inferred HIV message population is shown colored by relative abundance.

1394 mutagenesis^{225–228}.

1395 In our data, 80,045 reads contained sequences matching to both HIV and human genomic
1396 DNA, but a considerable complication arises because chimeras can be formed artifactually
1397 during the preparation of libraries for sequence analysis^{229–236}. Many of the chimeric
1398 sequences in our data contained junctions between the HIV and human sequence where the
1399 ends of the human and HIV sequence were similar and potentially complementary (Figure
1400 4.8A). This raises the concern that some of these chimeras could be products of in vitro
1401 recombinations during the reverse transcription, amplification and sequencing processes.
1402 Template switching between sequences with shared similarity is a well established property
1403 of retroviral reverse transcriptase enzymes used in RNA-Seq library preparation^{237–239}.
1404 Priming off incomplete transcripts during DNA synthesis is another potential source of
1405 chimeric transcripts^{229,230,240,241}. Failing to account for chimeras can hinder interpretation
1406 of deep sequencing data^{231–236}.

1407 Also consistent with artificial chimera formation, 7,354 reads (9.2% of chimeric messages)
1408 contained HIV sequences joined to human mitochondrial sequences, yet HIV proviruses have
1409 not previously been found integrated in mitochondrial DNA¹⁰². To probe this further, we
1410 used ligation-mediated PCR to recover integration site junctions from the same infected cell
1411 populations analyzed by RNA seq, yielding 147,281 unique integration sites (Figure 4.8B)⁷³.
1412 No integrations in mitochondrial DNA were detected. We conclude that chimeric HIV-
1413 mitochondrial sequence reads in the RNA-seq data represent artifacts of library construction
1414 and so used these chimeras as an assay to evaluate subsequent data filtering steps. We
1415 reasoned that reads without sequence similarity at junctions between human and HIV
1416 mapping were less likely to be artifacts caused by template switching. Filtering to only reads
1417 where no overlap and no unknown intervening sequence was present between human and HIV
1418 portions left 2181 junctions and reduced the proportion of reads containing mitochondrial
1419 DNA to 2.4%. Of the remaining HIV-human chimeric reads, the HIV portion of 605 sequences
1420 bordered the 3' or 5' end of HIV or an HIV splice donor or acceptor. Filtering to these

1421 more likely authentic junctions left only 2 (0.3%) chimeric reads containing mitochondrial
1422 sequence. This decrease in likely mitochondrial artifacts suggests that the filtering was
1423 effective. The high rate of mitochondrial chimeras in the unfiltered sequences raises the
1424 concern that artifacts may easily distort results in studies using similar amplification and
1425 sequencing techniques.

1426 Chimeric messages composed of HIV and cellular RNA sequences can be formed by cellular
1427 gene transcription reading into the integrated provirus, by HIV transcription reading out
1428 through the viral polyadenylation site or by splicing between human and viral splice sites.
1429 In our filtered data, the predominant forms appear to be derived from reading through the
1430 HIV polyadenylation signal into the surrounding DNA (78%), splicing out of the viral D4
1431 splice donor to join to human slice acceptors (17%) and reading into the HIV 5' LTR from
1432 human sequence (4.0%) (Figure 4.8C). No splice site other than D4 had more than two
1433 chimeric reads observed.

1434 The filtered chimeric reads had many traits consistent with biological chimera formation.
1435 The reads containing HIV D4 joined to human sequences had the characteristics expected of
1436 splicing—72.1% of the chimeric junctions mapped to known human acceptors and 96.1%
1437 mapped to a location immediately preceded by the AG consensus of human mRNA acceptors.
1438 The reads containing the 5' or 3' LTR border were almost exclusively (93%) found in
1439 transcription units, with odds of being in a gene 2.3-fold (95% CI: 1.6–3.2×) higher than
1440 integration sites from the same sample. The 5' or 3' chimeras were also more likely to be
1441 located in an exon than integration sites even after excluding any integration or chimera not
1442 located in a transcription unit (odds ratio: 2.1×, 95% CI: 1.6–2.6×).

1443 We next compared whether the human and viral segments of chimeric reads agreed or
1444 disagreed in orientation (i.e. strand transcribed) for reads with the human portion mapped
1445 within annotated transcription units. The sequencing technique used here does not preserve
1446 strand information, but we can check whether the strand of a sequence read agrees or
1447 disagrees with the annotated gene strand and compare this to the observed strand of the

1448 HIV portion of the read. We found a strong association between the orientation of the
1449 human and HIV portions of chimeric reads within 3' and 5' chimeras (odds ratio: 6.2 \times ,
1450 95% CI: 3.9–10.2 \times). This highly significant enrichment of HIV and human genes in the
1451 same orientation (Fisher's exact test $p < 10^{-15}$) might indicate that antisense HIV RNA
1452 is rapidly degraded by a response to double-stranded RNA or that polymerases oriented
1453 in opposing directions interfere with one another during elongation. Chimeras involving
1454 HIV splice donor D4 were even more highly enriched for matching orientations (odds ratio:
1455 52.5 \times , 95% CI: 12.1–307 \times) suggesting that pairing with human splice acceptors may add
1456 an additional constraint on the orientation of D4 chimeric reads.

1457 Based on these data, we can propose a lower bound on the relative abundance of chimeras. If
1458 we assume that our filtering removed nearly all artifacts so that we have few false positives,
1459 then our estimate should be lower than the true proportion of chimeras. In our data, only
1460 $\frac{604}{12,689,879} = 0.0048\%$ of reads containing sequence mapping to HIV also contained identifiable
1461 chimeric junctions. However, this is an underestimate because in an HIV-derived mRNA, any
1462 fragment of the sequence will be mappable to HIV, while for a chimeric sequence only a read
1463 spanning the HIV-human junction will allow identification of a chimera. If we assume that
1464 25 bases of sequence are necessary to map to human or HIV sequence, then, with the 100-bp
1465 reads used here, only read fragments starting between 75- and 25-bp downstream of the
1466 chimeric junction will be identifiable. If we assume the average chimeric mRNA sequences is
1467 at least 2 kb long, then a read from a chimeric sequence has at most a $\frac{50}{2000} = 2.5\%$ chance
1468 of containing a mappable junction. Thus, a lower bound for the proportion of HIV mRNA
1469 that also contain human-derived sequences is 0.2% ($\frac{0.0048\%}{2.5\%}$). Looking only at splicing from
1470 HIV donor D4, we saw 16,843 reads containing a junction from D4 to an HIV acceptor and
1471 104 reads from D4 to human sequence. Thus, in our data, 0.6% of D4 splice products form
1472 junctions with human acceptors instead of HIV acceptors.

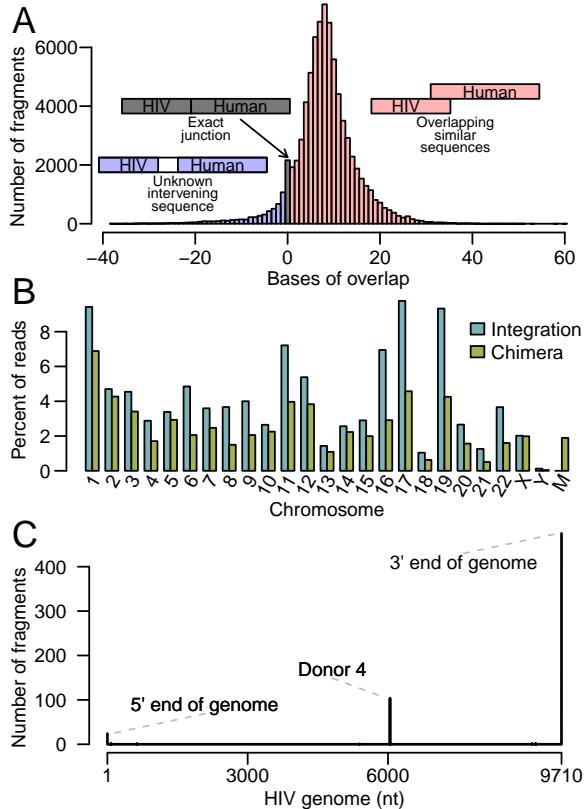


Figure 4.8: Analysis of chimeric RNA sequences containing both human and HIV sequences. A) The length of overlapping sequence (regions of complementarity potentially favoring chimera formation) matching both human and HIV at inferred chimeric junctions. The x-axis shows the length of the overlap and the y-axis shows the frequency of chimeric junctions with the indicated extent of overlap. B) Chromosomal distribution of uniquely mapping HIV integration sites from the same infections of primary T cells and comparison to uniquely mapping human sequences in chimeric reads observed in RNA-Seq. Note that the mitochondrial genome, denoted as M, has no authentic integration sites but does have extensive matches to chimeric junctions found in the RNA-Seq data. C) Counts of the location in the HIV genome of the HIV-human junctions in filtered chimeric reads.

1473 4.5 Discussion

1474 Here we used RNA-Seq to analyze mRNA accumulation and splicing in primary T cells
 1475 infected with the low passage isolate HIV_{89.6}. We did not carry out dense time series
 1476 analysis, compare different human cell donors or compare different perturbations of the
 1477 infections—instead, we focused on generating a dense data set at a single time point. We
 1478 analyzed replicate infected cell and control samples to allow discrimination of within-condition
 1479 versus between-condition variation and assessed differences using a series of bioinformatic
 1480 approaches. Many previous studies have used microarray technology or RNA-Seq to study
 1481 gene activity in HIV-infected cells^{88,164–174}, usually analyzing infections of transformed cell
 1482 lines or laboratory adapted strains of HIV-1. Here we present what is to our knowledge
 1483 the deepest RNA-Seq data set reported for infection in primary T cells using a low passage
 1484 HIV isolate (HIV_{89.6}). This data set was paired with a set of 147,281 unique integration
 1485 site sequences extracted from the same infections, which were critical to our ability to

1486 quality control chimeric reads. An advantage of studies using cell lines and laboratory
1487 adapted strains is that often a high percent of cell infection can be achieved, whereas in
1488 this study we achieved only ~30% infection. However, we report distinctive features of the
1489 transcriptional response not seen in studies of HIV infections in cell lines. Novel in this
1490 study are 1) identification of intron retention as a consequence of HIV infection, 2) the
1491 finding of activation of ERV-9/LTR12C after HIV infection, 3) generation of a quantitative
1492 account of the structures and abundances of over 70 HIV_{89.6} messages and 4) clarification of
1493 the predominant types of HIV-host transcriptional chimeras. These findings are discussed
1494 below.

1495 Broad changes in host cell mRNA abundances were evident after infection, with over 17% of
1496 expressed genes changing significantly in activity. Changes included expected response to
1497 viral infection, apoptosis and T cell activation. Although it is not possible here to separate
1498 the response of infected and bystander cells, this study highlights the drastic changes in
1499 cellular expression caused by HIV-1 infection. In a meta-analysis including four previously
1500 published studies, no gene was detected as differentially expressed in all five studies and
1501 only a handful of genes appeared in four out of five studies. Further analysis showed that
1502 expression changes appear to be cell type specific, raising concerns that studies using cell
1503 lines may not fully reflect host cell responses in *in vivo* infections.

1504 Unexpectedly, intronic sequences were more common in the RNA-Seq data from cells after
1505 HIV_{89.6} infection than in mock infected cells. The mechanism is unclear. It is possible
1506 that the splicing machinery is reduced in activity after 48 hours of infection, perhaps as a
1507 part of the antiviral response of infected and bystander cells. HIV infection does appear to
1508 alter expression and localization of some splicing factors^{127,202}. In addition, we saw a large
1509 reduction in the abundance of mRNA from nonsense-mediated decay related genes, perhaps
1510 indicating that RNA surveillance is loosened thus allowing more unspliced or aberrantly
1511 spliced transcripts. Alternatively, fully spliced mRNAs might be more rapidly degraded after
1512 infection, possibly by interferon-mediated induction of RNaseL²⁴². A speculative possibility

1513 is that HIV_{89.6} encodes a factor that alters cellular splicing or promotes mRNA degradation
1514 to optimize splicing and translation of viral messages.

1515 Infection resulted in increased expression of specific cellular repeated sequences. HERVs, in
1516 particular HERV-K, have previously been observed to show increased RNA accumulation with
1517 HIV infection^{203–205,243} and possibly represent vaccine targets because of their production of
1518 distinctive proteins^{209,243–247}. Here, though we saw modest increases in HERV-K expression,
1519 ERV-9 had the greatest change in expression (33 LTR12C and 14 ERV-9 annotated regions
1520 with greater than 4× change in expression). Previous RNA-Seq studies of HIV infection in
1521 cell lines did not report increases in HERV expression^{88,174} but this difference is likely due
1522 to a much higher baseline expression of HERVs in transformed cell lines. We also observed
1523 increases in LINE and Alu element transcription, as has been reported previously²⁰⁶, and
1524 expression changes in ERV-9/LTR12C expression associated with transcription factor motifs
1525 and U3 variants.

1526 Many of the repeated sequence elements that were induced by HIV_{89.6} infection are relatively
1527 recently integrated in the human genome. The reason for this pattern is unclear. It may
1528 be that older elements have accumulated more mutations, resulting in an inactivation of
1529 transcriptional signals. Alternatively, perhaps the elements that are induced have been
1530 recruited for transcriptional control of cellular functions, so that their transcriptional activity
1531 is preserved evolutionarily^{185,248,249}.

1532 Comparison of results of sequencing HIV_{89.6} messages using long-read single molecule
1533 sequencing (Pacific Biosciences) and dense short read sequencing (Illumina data reported
1534 here) allowed a full quantitative accounting of more than 70 HIV_{89.6} splice forms. The full
1535 length unspliced HIV RNA comprised 37.6% of all messages, corresponding to about 2000
1536 genomes per cell. Notably abundant messages included those encoding Nef (D1-A5-D4-A7:
1537 15.5%) and two Rev-encoding transcripts (D1-A4c-D4-A7: 4.2%, D1-A4b-D4-A7: 3.1%).
1538 The full set of messages is summarized in Figure 4.7B. Our previous analysis revealed an
1539 unusually prominent 1 kb size class. HIV_{89.6} encodes a rare splice acceptor (A8c) within Nef

1540 responsible for formation of the short messages. Our data indicated that two members of the
1541 1-kb size class, D1-A5-D4-A8c and D1-A8c, accounted for 10.6% and 4.9% of all messages.
1542 The 1 kb size class as a whole accounted for fully 20% of messages. Most HIV/SIV variants
1543 appear to encode an acceptor near this position, suggesting a potential unknown function
1544 for these short spliced forms^{98,107,111}.

1545 After filtering, we detected a sizeable number of apparently authentic chimeras containing
1546 both HIV and cellular sequences, allowing comparison to examples of host-cell modification
1547 by integration. Mechanisms of insertional activation have been studied intensively in animal
1548 models of transformation and in adverse events in human gene therapy. One of the most
1549 common mechanisms involves insertion of a retroviral enhancer near a cellular promoter,
1550 so that the rate of initiation is increased and normal cellular messages are increased in
1551 abundance. However, another common mechanism involves formation of chimeric messages
1552 involving both cellular and viral/vector sequences. In HIV infection, examples of insertion
1553 in the Bach2 and MKL2 genes have been associated with long term persistence of particular
1554 cell clones^{218–221}. In these cells, proviruses were integrated within the cellular transcription
1555 unit, and the transcriptional direction of the integrated provirus was the same as that of
1556 Bach2 or MKL2. This would allow formation of a fusion of the 5' HIV sequences with 3'
1557 Bach2 sequences, potentially involving the most common events seen here (either 3' read out
1558 or splicing from HIV D4 to a cellular exon). However, a closely studied example of clonal
1559 expansion in a successful lentiviral vector gene therapy for beta-thalassemia was associated
1560 with expansion of a cell clone harboring an integrated vector within the transcription unit
1561 of HMGA2. In this case the message spliced into the vector and terminated, removing
1562 a negative regulatory sequence normally present in the 3' end HMGA2 message²²⁵. A
1563 targeted study in vitro of chimeric message formation by lentiviral vectors showed examples
1564 of multiple types of read-in and -out and splice-in and -out²²⁷, which may have been more
1565 frequent and more varied than for HIV_{89.6} proviruses studied here. The lack of splicing or
1566 reading into HIV in this study may be a reflection of the high rate HIV transcription in
1567 these infected cells—because HIV was so highly expressed, there would be more opportunities

1568 for polymerase to splice out of or read through the HIV genome than to read or splice in.
1569 The vast majority of HIV proviruses in expanded clones in well-suppressed patients now
1570 appear to be defective²²¹—going forward, it will be of interest to investigate whether these
1571 HIV proviruses are damaged in ways that promote formation of chimeric transcripts.

1572 Lastly, we note that several features of the transcriptional response to HIV_{89.6} infection were
1573 suggestive of de-differentiation away from T cell specific expression patterns. The increase
1574 in expression of cellular HERVs and LINEs is characteristic of cells in early development.
1575 Specific HERVs and transposons, including ERV-9/LTR12C and HERV-K, have been
1576 implicated in regulating gene activity early in development^{185,248,250–253}. Several genes
1577 related to other hematopoietic cell types showed elevated RNA abundance after HIV_{89.6}
1578 infection. These data are of interest given the finding that patients undergoing long term
1579 ART can contain long lived T cell clones that may contribute to the latent reservoir^{221,254–257}.
1580 Possibly the transcriptional responses seen in infected primary T cells here are reflective
1581 of processes leading to formation of the long-lived latently-infected cells with stem-like
1582 properties.

1583 4.6 Conclusions

1584 Infections of primary T cells with a low passage HIV isolate show several distinctive features
1585 compared with previously published data using T cell lines and/or lab-adapted HIV strains.
1586 We found strong changes in expression in genes related to immune response and apoptosis
1587 similar to studies of HIV infection in patient samples and primary cells but different from
1588 studies performed in SupT1 cell lines. Notable changes after infection included intron
1589 retention and activation of recently integrated retrotransposons and endogenous retroviruses,
1590 in particular LTR12C/ERV-9. We also present complete absolute estimation of over 70
1591 messages from HIV_{89.6} and specify the major virus-host chimeras as read out from the 3'
1592 end of the provirus and splicing from viral splice donor 4 to cellular acceptors.

1593 **4.7 Availability of supporting data**

1594 RNA-Seq reads from this study are available at the Sequence Read Archive under accession
1595 number SRP055981. The integration site data is available at the Sequence Read Archive
1596 under accession number SRP057555.

1597 **4.8 Acknowledgements**

1598 We would like to thank the University of Pennsylvania Center for AIDS Research (P30
1599 AI045008) for preparation of viral stocks and isolation of primary CD4⁺ T cells; Ronald
1600 G. Collman and members of the Bushman laboratory for reagents, helpful discussion and
1601 technical expertise. This work was funded by NIH grant R01 AI052845, the HIV Immune
1602 Networks Team (HINT) consortium P01 AI090935 and NRSA computational genomics
1603 training grant T32 HG000046.

1604 **CHAPTER 5: A reverse transcription loop-mediated isothermal**
1605 **amplification assay optimized to detect multiple HIV**
1606 **subtypes**

This chapter was originally published as:

KE Ocwieja, S Sherrill-Mix, C Liu, J Song, H Bau and FD Bushman. 2015. A reverse transcription loop-mediated isothermal amplification assay optimized to detect multiple HIV subtypes. *PLoS One*, 10:e0117852. doi: 10.1371/journal.pone.0117852

1607 KE Ocwieja, C Liu, H Bau, FD Bushman and I conceived and designed the experiments. KE Ocwieja, C Liu and J Song performed the experiments. K Ocwieja, J Song and I analyzed the data. I produced the figures. KE Ocwieja, C Liu, H Bau, FD Bushman and I wrote the paper.
Supporting information are available at <http://journals.plos.org/plosone/article?id=10.1371/journal.pone.0117852#sec011>

1608 **5.1 Abstract**

1609 Diagnostic methods for detecting and quantifying HIV RNA have been improving, but
1610 efficient methods for point-of-care analysis are still needed, particularly for applications in
1611 resource-limited settings. Detection based on reverse-transcription loop-mediated isothermal
1612 amplification (RT-LAMP) is particularly useful for this, because when combined with
1613 fluorescence-based DNA detection, RT-LAMP can be implemented with minimal equipment
1614 and expense. Assays have been developed to detect HIV RNA with RT-LAMP, but existing
1615 methods detect only a limited subset of HIV subtypes. Here we report a bioinformatic study
1616 to develop optimized primers, followed by empirical testing of 44 new primer designs. One
1617 primer set (ACeIN-26), targeting the HIV integrase coding region, consistently detected
1618 subtypes A, B, C, D, and G. The assay was sensitive to at least 5000 copies per reaction for
1619 subtypes A, B, C, D, and G, with Z-factors of above 0.69 (detection of the minor subtype F

1620 was found to be unreliable). There are already rapid and efficient assays available for detecting
1621 HIV infection in a binary yes/no format, but the rapid RT-LAMP assay described here has
1622 additional uses, including 1) tracking response to medication by comparing longitudinal
1623 values for a subject, 2) detecting of infection in neonates unimpeded by the presence of
1624 maternal antibody, and 3) detecting infection prior to seroconversion.

1625 5.2 Introduction

1626 Despite the introduction of efficient antiretroviral therapy, HIV infection and AIDS continue
1627 to cause a worldwide health crisis²⁵⁹. Methods for detecting HIV infection have improved
1628 greatly with time²⁶⁰—today rapid assays are available that can detect HIV infection in a
1629 yes-no format using a home test kit that detects antibodies in saliva. Viral load assays that
1630 quantify viral RNA with quick turn-around time are widely available in the developed world.
1631 However, quantitative viral load assays are not commonly available with actionable time
1632 scales in much of the developing world. This motivates the development of new rapid and
1633 quantitative assays that can be used at the point of care with minimal infrastructure^{261,262}.

1634 One simple and quantitative detection method involves reverse transcription-based loop
1635 mediated isothermal amplification (RT-LAMP)²⁶³. In this method, a DNA copy of the viral
1636 RNA is generated by reverse transcriptase, and then isothermal amplification is carried out to
1637 increase the amount of total DNA. Primer binding sites are chosen so that a series of strand
1638 displacement steps allow continuous synthesis of DNA without requiring thermocycling.
1639 Reaction products can be detected by adding an intercalating dye to reaction mixtures
1640 that fluoresces only when bound to DNA, allowing quantification of product formation by
1641 measurement of fluorescence intensity. Such assays can potentially be packaged in simple
1642 self-contained devices and read out with no technology beyond a cell phone.

1643 RT-LAMP assays for HIV-1 have been developed previously and reported to show high
1644 sensitivity and specificity for subtype B, the most common HIV strain in the developed
1645 world^{262,264,265}. Another recent study reported RT-LAMP primer set optimized for the

1646 detection of HIV variants circulating in China²⁶⁶, and another on confirmatory RT-LAMP
1647 for group M viruses²⁶⁷. Assays have also been developed for HIV-2²⁶⁸. A complication
1648 arises in using available RT-LAMP assays due to the variation of HIV genomic sequences
1649 among the HIV subtypes^{269,270}, so that an RT-LAMP assay optimized for one viral subtype
1650 may not detect viral RNA of another subtype²⁷¹. Tests presented below show that many
1651 RT-LAMP assays are efficient for detecting subtype B, for which they were designed, but
1652 often performed poorly on other subtypes. Subtype C infects the greatest number of people
1653 worldwide, including in Sub-Saharan Africa, where such RT-LAMP assays would be most
1654 valuable, motivating optimization for subtype C. Several additional non-B subtypes are also
1655 responsible for significant burdens of disease world-wide²⁷².

1656 Here we present the development of an RT-LAMP assay capable of detecting HIV-1 subtypes
1657 A, B, C, D, and G. We first carried out a bioinformatic analysis to identify regions conserved
1658 in all the HIV subtypes. We then tested 44 different combinations of RT-LAMP primers
1659 targeting this region in over 700 individual assays, allowing identification of a primer set
1660 (ACeIN-26) that was suitable for detecting these subtypes. We propose that the optimized
1661 RT-LAMP assay may be useful for quantifying HIV RNA copy numbers in point-of-care
1662 applications in the developing world, where multiple different subtypes may be encountered.

1663 5.3 Methods

1664 5.3.1 Viral strains used in this study

1665 Viral strains tested included HIV-1 92/UG/029 (Uganda) (subtype A, NIH AIDS Reagent
1666 program reagent number 1650), HIV-1 THRO (subtype B, plasmid derived, University of
1667 Pennsylvania CFAR)²⁷³, CH269 (subtype C, plasmid derived, University of Pennsylvania
1668 CFAR)²⁷³, UG0242 (subtype D, University of Pennsylvania CFAR), 93BRO20 (subtype F,
1669 University of Pennsylvania CFAR), HIV-1 G3 (subtype G, NIH AIDS Reagent program
1670 reagent number 3187)²⁷⁴.

1671 Viral stocks were prepared by transfection and infection. Culture supernatants were cleared

1672 of cellular debris by centrifugation at 1500g for 10 min. The supernatant containing virus
1673 was then treated with 100 U DNase (Roche) per 450 uL virus for 15 min at 30°C. RNA was
1674 isolated using the QiaAmp Viral RNA mini kit (Qiagen GmbH, Hilden, Germany). RNA
1675 was eluted in 80 uL of the provided elution buffer and stored at -80°C.

1676 Concentration of viral RNA copies was calculated from p24 capsid antigen capture assay
1677 results provided by the University of Pennsylvania CFAR or the NIH AIDS-reagent program.
1678 In calculating viral RNA copy numbers, we assumed that all p24 was incorporated in virions,
1679 all RNA was recovered completely from stocks, 2 genomes were present per virion, 2000 p24
1680 molecules per viral particle, and the molecular weight of HIV-1 p24 was 25.6 kDa.

1681 **5.3.2 Assays**

1682 RT-LAMP reaction mixtures (15 μ L) contained 0.2 μ M each of primers F3 and B3 (if a
1683 primer set used multiple B3 primers, mixture contained 0.2 μ M of each); 1.6 μ M each of FIP
1684 and BIP primers (if a primer set had multiple FIP primes, reaction mixture contained 0.8
1685 μ M of each FIP primer); and 0.8 μ M each of LoopF and LoopB primers; 7.5 μ L OptiGene
1686 Isothermal Mastermix ISO-100nd (Optigene, UK), ROX reference dye (0.15 μ L from a 50X
1687 stock), EvaGreen dye (0.4 μ L from a 20X stock; Biotium, Hayward, CA); HIV RNA in 4.7
1688 μ L; AMV reverse transcriptase (10U/ μ L) 0.1 μ L and water to 15 μ L In most cases where
1689 two primer sets were combined, the total primer concentration within the reaction was
1690 doubled such that the above individual primer molarities were maintained. For the mixture
1691 ACeIN-26+F-IN (S2 Table, line 46), the total primer concentration was not doubled—the
1692 F-IN primer set comprised 25% of the total primer concentration, and the ACeIN-26 primer
1693 set comprised 75% of the total primer concentration with the ratios of primers listed above
1694 preserved. This mixture was combined 1:1 with the ACe-PR primer set (S2 Table, line 47)
1695 such that total primer concentration in the final mixture was doubled.

1696 Amplification was measured using the 7500-Fast Real Time PCR system from Applied
1697 Biosystems with the following settings: 1 minute at 62°C; 60 cycles of 30 seconds at 62°C

1698 and 30 seconds at 63°C. Data was collected every minute. Product structure was assessed
1699 using dissociation curves which showed denaturation at 83°C. Products from selected
1700 amplification reactions were analyzed by agarose gel electrophoresis and showed a ladder of
1701 low molecular weight products (data not shown).

1702 Product synthesis was quantified as the cycle of threshold for 10% amplification. Z-factors²⁷⁵
1703 were calculated from tests of 24 replicates using the ACeIN26 primer set in assays with viral
1704 RNA of each subtype. No detection after 60 min was given a value of 61 min in the Z-factor
1705 calculation.

1706 **5.4 Results**

1707 **5.4.1 Testing published RT-LAMP primer sets against multiple HIV subtypes**

1708 We first assessed the performance of existing RT-LAMP assays on RNA samples from
1709 multiple HIV subtypes. We obtained viral stocks from HIV subtypes A, B, C, D, F, and
1710 G, estimated the numbers of virions per ml, and extracted RNA. RNAs were mixed with
1711 RT-LAMP reagents which included the six RT-LAMP primers, designated F3, B3, FIP, BIP,
1712 LF and LB²⁶³. Reactions also contained reverse transcriptase, DNA polymerase, nucleotides
1713 and the intercalating fluorescent EvaGreen dye, which yields a fluorescent signal upon DNA
1714 binding. DNA synthesis was quantified as the increase in fluorescence intensity over time,
1715 which yielded a typical curve describing exponential growth with saturation (examples are
1716 shown below). Results are expressed as threshold times (T_t) for achieving 10% of maximum
1717 fluorescence intensity at the HIV RNA template copy number tested.

1718 In initial tests, published primer sets targeting the HIV-1 subtype B coding regions for
1719 capsid (CA), protease (PR), and reverse transcriptase (RT) (named B-CA, B-PR and B-RT)
1720 were assayed in reactions with RNAs from four of the subtypes. Results with each primer set
1721 tested are shown in Figure 5.1 in heat map format, where each tile summarizes the results of
1722 tests of 5000 RNA copies. Primers and their groupings into sets are summarized in S1 and
1723 S2 Tables, average assay results are in S3 Table, and raw assay data is in S4 Table. Assays

1724 (Figure 5.1, top) with the B-CA, B-PR and B-RT primer sets detected subtypes B and D
1725 at 5000 RNA copies with threshold times less than 20 min. However, assays with B-CA
1726 and B-RT detected subtypes C and F with threshold times > 50 min, indicating inefficient
1727 amplification and the potential for poor separation between signal and noise. B-PR did
1728 not detect subtype C at all. In an effort to improve the breadth of detection, we first tried
1729 mixing the B-PR primers, which detected clade F (albeit with limited efficiency) with the
1730 B-CA and B-RT primers (Figure 5.1 and S3 and S4 Tables). In neither case did this provide
1731 coverage of all four clades tested. We thus did not test these primer sets on RNAs from the
1732 remaining subtypes and instead sought to develop primer sets targeting different regions of
1733 the HIV genome.

1734 **5.4.2 Primer design strategy**

1735 To design primers that detected multiple HIV subtypes efficiently, we analyzed alignments
1736 of HIV genomes (downloaded from the Los Alamos National Laboratory site²⁶⁹) for regions
1737 with similarity across most viruses, revealing that a segment of the pol gene encoding
1738 IN was particularly conserved (Figure 5.2A). A total of six primers are required for each
1739 RT-LAMP assay²⁶³. We used the EIKEN primer design tool to identify an initial primer set
1740 targeting this region. In further analysis, positions in the alignments were identified within
1741 primer landing sites that commonly contained multiple different bases. Primer positions
1742 were manually adjusted to avoid these bases when possible, and when necessary mixtures
1743 were formulated containing each of these commonly occurring bases (S1 and S2 Tables).
1744 An extensive series of variants targeting the IN coding region was tested empirically in
1745 assays containing RNAs from multiple subtypes (5000 RNA copies per reaction, over 700
1746 total assays; S3 and S4 Tables). Based on initial results, primers were further modified by
1747 adjusting the primer position or addition of locked nucleic acids as described below.

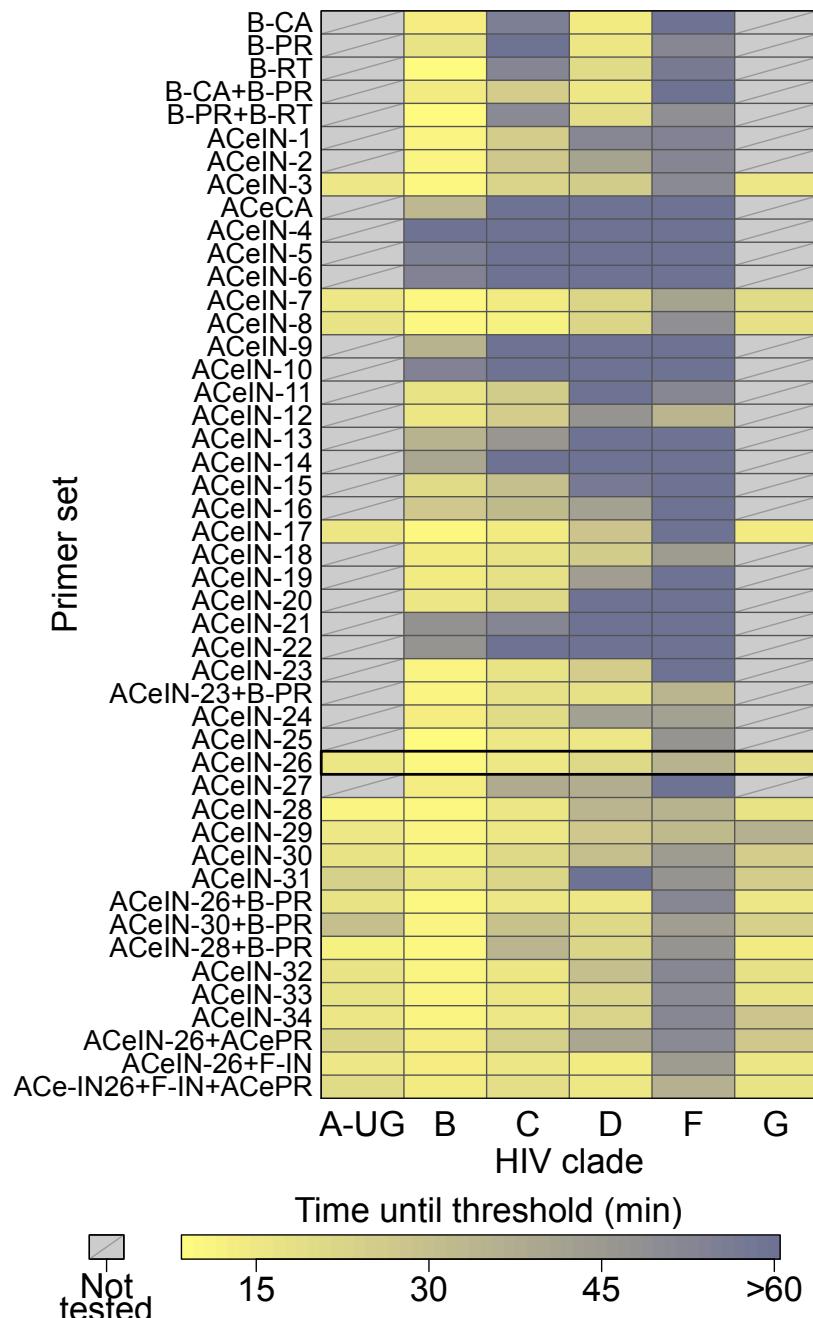


Figure 5.1: Summary of amplification results for all the RT-LAMP primer sets tested in this study. The data is shown as a heat map, with more intense yellow coloring indicating shorter amplification times (key at bottom). Primer sets tested are named along the left of the figure. Primer sequences, and their organization into LAMP primer sets, are catalogued in S1 and S2 Tables. The raw data and averaged data are collected in S3 and S4 Tables. ACeIN-26 primer set (highlighted) had one of the best performances across the subtypes and a relatively simple primer design.

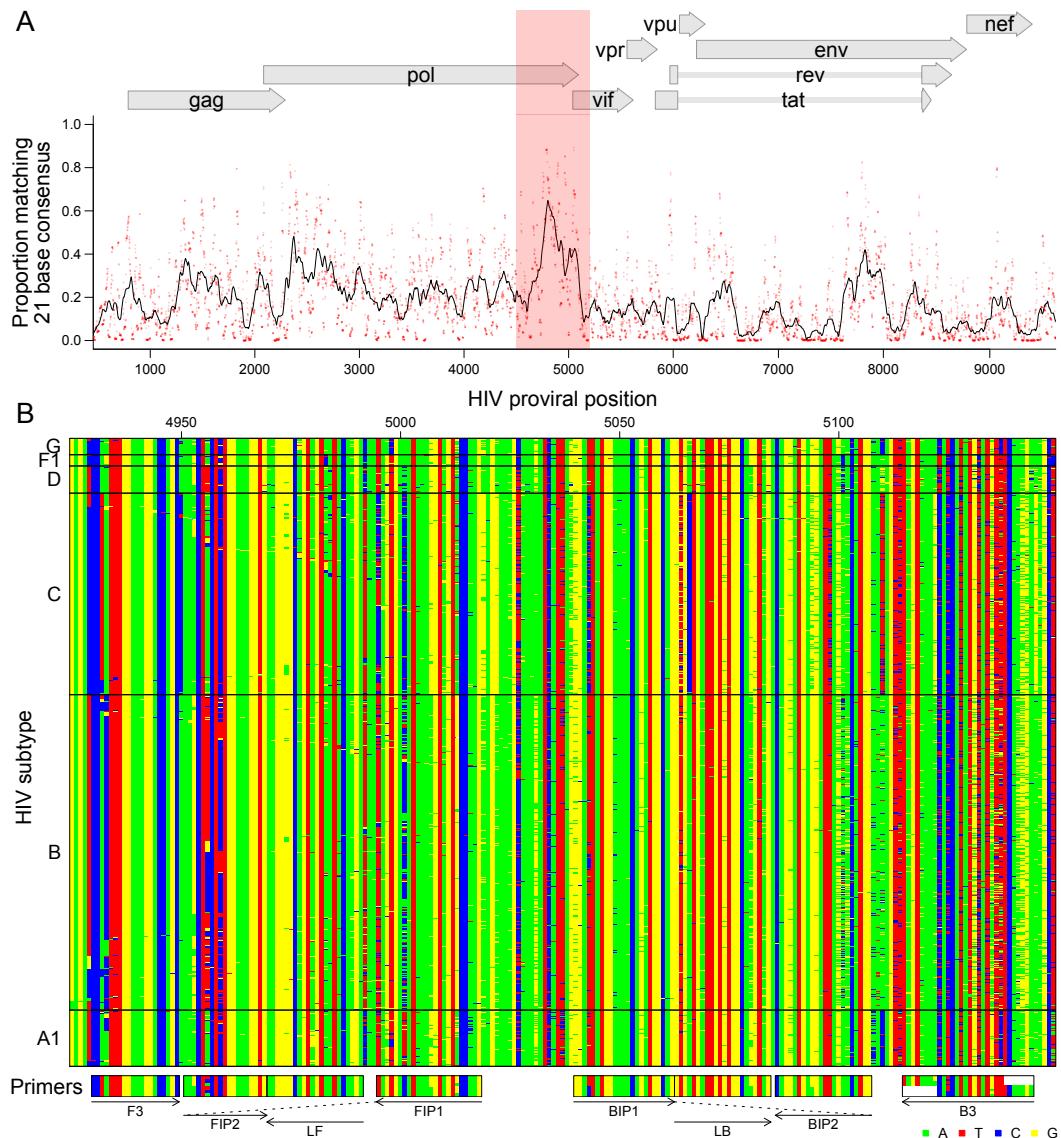


Figure 5.2: Bioinformatic analysis to design subtype-agnostic RT-LAMP primers. A) Conservation of sequence in HIV. HIV genomes ($n = 1340$) from the Los Alamos National Laboratory collection were aligned and conservation calculated. The x-axis shows the coordinate on the HIV genome, the y-axis shows the proportion of sequences matching the consensus for each 21 base segment of the genome (red points). The black line shows a 101 base sliding average over these proportions. The vertical red shading shows the region targeted for LAMP primer design that was used as input into the EIKEN primer design tool. Numbering is relative to the HIV_{89.6} sequence. B) Aligned genomes, showing the locations of the ACeIN-26 primers. Sequences are shown with DNA bases color-coded as shown at the lower right. Each row indicates an HIV sequence and each column a base in that sequence. Horizontal lines separate the HIV subtypes (labeled at left). Arrows indicate the strand targeted by each primer. Primers targeting the negative strand of the virus are shown as reverse compliments for ease of viewing.

1748 **5.5 Testing different primer designs**

1749 Our first design, ACeIN-1 (“ACe” for “All Clade” and “IN” for “integrase”), targeted the
1750 HIV IN coding region and contained multiple bases at selected sites to broaden detection
1751 (Figure 5.1). ACeIN-2 and-3 have primers (B3) with slightly different landing sites. Tests
1752 showed that the mixture of primers allowed amplification with a shorter threshold time than
1753 did either alone (Figure 5.1).

1754 We also tried to design a new primer set to the CA coding region (Figure 5.1, ACeCA)
1755 but found that the set only amplified clade B, and not efficiently. Thus this design was
1756 abandoned.

1757 ACeIN-3 through-6 were altered by inserting a polyT sequence between the two different
1758 sections of FIP and BIP in various combinations, a modification introduced with the goal of
1759 improving primer folding, but these designs performed quite poorly (Figure 5.1).

1760 Because the FIP primer appeared to bind the region with most variability among clades, we
1761 tried variations that bound to several nearby regions. These were tried with and without
1762 the polyT containing BIP and FIP primers in various combinations (Figure 5.1, ACeIN-7
1763 through-22). We also tried mixing all of the variations of FIP together (ACeIN-23; S2 Table).
1764 The ACeIN-23 primer set was tried as a mixture with the B-PR set to try to capture clade
1765 F, yielding a relatively effective primer set (Figure 5.1, ACeIN-23+B-PR).

1766 In an effort to increase affinity, an additional G/C pair was added to F3 and tested with
1767 various other IN primers (Figure 5.1, ACeIN-24 through-31). Testing showed improvement,
1768 with ACeIN-26 showing particularly robust amplification.

1769 In a second effort to increase primer affinities, we substituted locked nucleic acids (LNAs) for
1770 selected bases that were particularly highly conserved among subtypes (Figure 5.1, ACeIN-30,
1771 -31, -32, -33, and-34). Some improvement was shown over the non-LNA containing bases.
1772 However, the ACeIN-26 primer set was as effective as or better than any LNA containing

1773 primer sets.

1774 In further tests, the ACeIN-26, -28 and-30 primers were tested combined with the ACePR
1775 primer set (a slightly modified version of the B-PR primer set, S2 Table, row 2, designed
1776 to accommodate a wider selection of HIV-1 subtypes) but no improvement was seen and
1777 efficiency may even have fallen for some subtypes. We also designed a primer set that
1778 matched exactly to the targeted sequences found in the problematic subtype F, and mixed
1779 this set with the ACeIN-26 primers. However, no improvement was seen (Figure 5.1, mixtures
1780 with F-IN set). Mixing the ACeIN-26 primers with both the ACePR and F-specific primers
1781 did yield effective primer sets (Figure 5.1, ACeIN26+F-IN and ACeIN26+F-IN+ACePR).
1782 However, amplification efficiency was not greatly improved over the ACeIN-26 primer set, so
1783 we proceeded with the simpler ACeIN-26 primer set (Figure 5.2B) in further studies.

1784 **5.5.1 Performance of the optimized RT-LAMP assay**

1785 The ACeIN-26 RT-LAMP primer set was next tested to determine the minimum concentration
1786 of RNA detectable under the reaction conditions studied (Figure 5.3). RNA template amounts
1787 were titrated and time to detection quantified. Tests showed detection after less than 20
1788 min of incubation for 50 copies of subtypes A or B, detection after less than 30 min for 5000
1789 copies for C, D, and G, and detection after less than 20 min for 50,000 copies for F.

1790 For clinical implementation the reliability of an assay is critical. This is commonly sum-
1791 marized as a Z-factor²⁷⁵, which takes into account both the separation in means between
1792 positive and negative samples and the variance in measurement of each. An assay with
1793 a Z-factor above 0.5 is judged to be an excellent assay. Z-factors for detection of each of
1794 the subtypes at 5000 RNA copies per reaction were > 0.50 for subtypes A, B, C, D, and
1795 G, respectively (Figure 5.4, n = 24 replicates per test). Detection of subtype F at 5000
1796 copies per reaction was sporadic, showing a much lower Z-factor. Therefore our ACeIN-26
1797 RT-LAMP primer set appears well suited to detect 5000 copies of subtypes A, B, C, D and
1798 G.

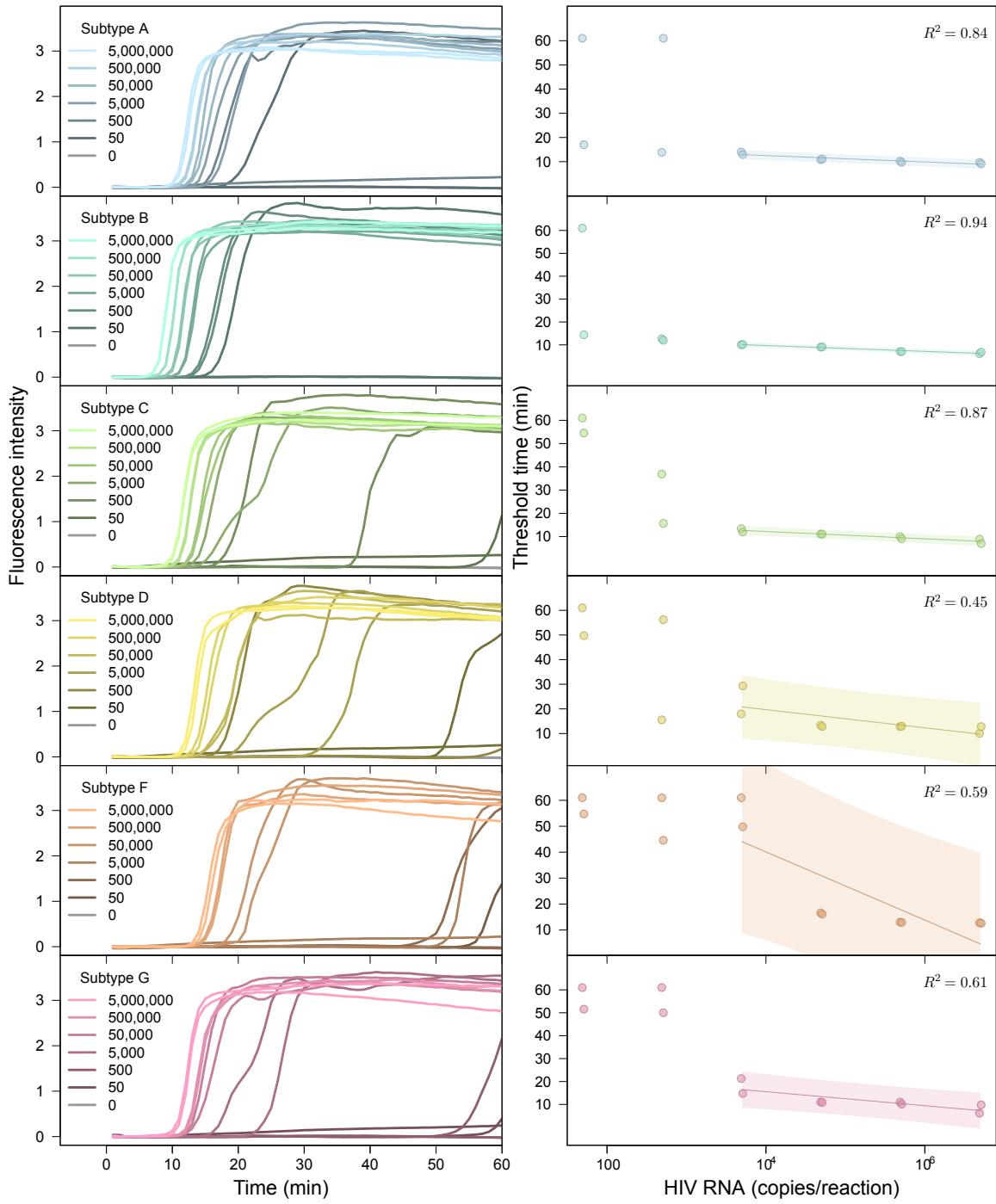


Figure 5.3: Performance of the AceIN-26 primer set with different starting RNA concentrations. Tests of each subtype are shown as rows. In each lettered panel, the left shows the raw accumulation of fluorescence signal (y-axis) as a function of time (x-axis); the right panel shows the threshold time (y-axis) as a function of log RNA copy number (x-axis) added to the reaction. In the right hand panels, values were dithered where two points overlapped to allow visualization of both.

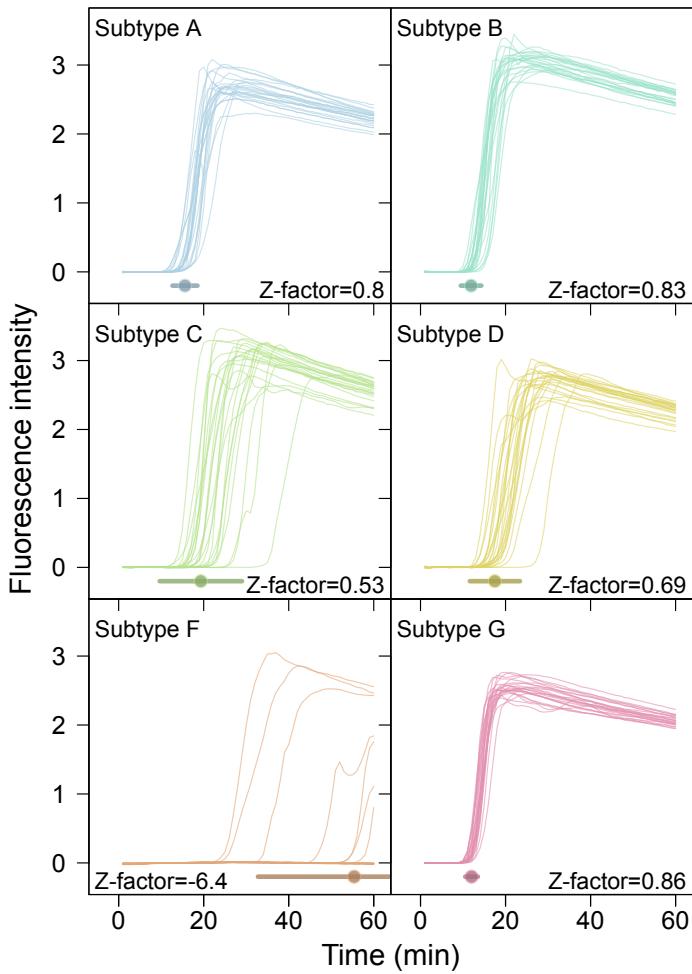


Figure 5.4: Examples of time course assays, displaying replicate tests of RT-LAMP primer set ACeIN-26 tested over six HIV subtypes, used in Z-factor calculations. A total of 5000 RNA copies were tested in each 15 μ L reaction. Time is shown on the x-axis, Fluorescence intensity on the y-axis. Replicates are distinguished using an arbitrary color code. Z-factor values and standard deviations are shown on each panel.

1799 **5.6 Discussion**

1800 Here we present an RT-LAMP assay optimized to identify multiple HIV subtypes. Infections
1801 with subtype B predominate in most parts of the developed world, but elsewhere other
1802 subtypes are more common²⁷². Thus nucleic acid-based assays for use in the developing
1803 world need to query HIV subtypes more broadly. Previously reported RT-LAMP assays,
1804 while effective at detecting subtype B, commonly showed poor ability to detect at least some
1805 of the HIV subtypes, including C, which is common in the developing world (Figure 5.1).
1806 Here we first carried out an initial bioinformatic survey to identify regions conserved across
1807 all HIV subtypes that could serve as binding sites for RT-LAMP primers. We then tested
1808 primer sets targeting these regions empirically for efficiency. Testing 44 different primer
1809 sets revealed that assays containing ACeIN-26 were effective in detecting 5000 copies of
1810 RNA from subtypes A, B, C, D, and G within 30 minutes of incubation. For these five
1811 subtypes, the times of incubation to reach the threshold times were not too different, which
1812 simplifies interpretation when the subtype in the sample is unknown. Regardless of the
1813 efficiency, these assays can be applied to longitudinal studies of changes in viral load within
1814 an individual. We propose that RT-LAMP assays based on the ACeIN-26 primer set can be
1815 useful world-wide for assaying HIV-1 viral loads in infected patients.

1816 There are several limitations to our study. Subtypes A, B, C, D, and G were detected
1817 efficiently and showed Z-factors above 0.5, but subtype F was detected reliably only with
1818 higher template amounts, probably due to more extensive mismatches with the ACeIN-26
1819 primer set. Subtype F is estimated, however, to comprise only 0.59% of all infections
1820 globally²⁷², though it is common in some regions. For many of the common circulating
1821 recombinant forms, such as AE and BC, the target site for ACeIN-26 is from a subtype
1822 known to be efficiently detected, though in some cases the efficiency of detection is not easy
1823 to predict and will need to be tested. We did not test subtypes beyond A, B, C, D, F and
1824 G, and we did not attempt to assess multiple different variants within each subtype. Thus,
1825 while we do know that our RT-LAMP assays are more widely applicable than many of those

1826 reported previously, we do not know whether they are able to detect all strains efficiently. In
1827 addition, although we carried out more than 700 assays in this study, there remain multiple
1828 parameters that could be optimized further, such as primer concentrations, salt type and
1829 concentration, temperature, and divalent metal concentrations, so there are likely further
1830 opportunities for improvement. Also, possible effects of RNA quality on assay performance
1831 were not tested rigorously.

1832 A particularly important parameter for further optimization is primer sequence. Several
1833 groups have recently published primer sets optimized for broad detection of different HIV
1834 lineages^{266,267}, offering opportunities for creating sophisticated primer blends with increased
1835 breadth of detection. However, in developing such mixtures, it will be important to monitor
1836 for possible complicating interactions of primers with each other. As an example of ongoing
1837 development of mixtures, we found that addition of another primer to the ACeIN-26 set
1838 that was matched to a common subtype C lineage allowed improved detection of subtype C
1839 variants (S1 Report). In order to improve detection of subtype F, which was suboptimal with
1840 ACeIN-26, additional primer sets could be mixed to specifically target subtype F, though
1841 the ones we tried so far did not work well. It will be useful to explore the performance of
1842 broader primer mixtures in future work.

1843 Today rapid assays are available that can report infection efficiently, for example by detecting
1844 anti-HIV antibodies in oral samples—however, the nucleic acid-based method presented here
1845 has additional potential uses. We envision combining the RT-LAMP assay with simple
1846 point-of-care devices for purifying blood plasma²⁶¹ and quantitative analysis of accumulation
1847 of fluorescent signals²⁷⁶. In one implementation of the technology, cell phones could be used
1848 to capture and analyze results, thereby minimizing equipment costs. Point-of-care devices are
1849 available facilitating the concentration of viral RNA from blood plasma or saliva²⁷⁶ to allow
1850 the detection of the 1000 RNA copy threshold that the WHO defines as virological treatment
1851 failure (World Health Organization, Consolidated ARV guidelines, June 2013). Together,
1852 these methods will allow assessment of parameters beyond just the presence/absence of

1853 infection. Quantitative RT-LAMP assays should allow tracking of responses to medication,
1854 detection in neonates (where immunological tests are confounded by presence of maternal
1855 antibody), and early detection before seroconversion.

1856 **5.7 Acknowledgments**

1857 We are grateful to members of the Bushman and Bau laboratory for help and suggestions.

CHAPTER 6: Conclusions and future directions

1859 In this dissertation, we described studies characterizing HIV-1 latency, expression and
1860 alternative splicing and host cell response to infection. We then developed point-of-care
1861 methods for the detection of infection and quantification of viral load. These projects suggest
1862 many avenues for continuing research.

1863 6.1 Latency and integration location

1864 In Chapter 2, we showed that the chromosomal location of integration affects proviral latency
1865 but the mechanisms appear to differ between cell culture models. Similarly a recent study
1866 of nine cell culture models found that no single model reliably predicted the performance of
1867 activating compounds in *ex vivo* tests of latently infected cells from HIV patients²⁷⁷. This
1868 suggests that either some cell culture models do not accurately reflect latency in patients or
1869 that there are diverse subsets of cells with differing mechanisms of latency within patients.

1870 Cell culture models are currently used to screen potentially therapeutic compounds^{277,278}. If
1871 some cell culture models are not representative of *in vivo* conditions then potential treatments
1872 may be discarded or marked for development erroneously. Further comparisons between
1873 additional cell culture models and additional replicates of existing models might allow
1874 discrimination between batch/lab effects and reveal patterns between models. Comparison
1875 with cells extracted from patients or infected lab animals might offer a gold standard
1876 comparison although it is difficult to obtain large amounts of cells and difficult to distinguish
1877 defective provirus from latent provirus in such populations.

1878 Various treatments are now being considered for the reactivation of latent provirus²⁷⁷. To
1879 further understand the mechanisms of these treatments, it would be informative to compare
1880 the features of latent provirus induced by a given treatment to latent viable provirus
1881 remaining uninduced. Repeated cell sorting and integration site sequencing might provide
1882 insight on mechanism. For example, one could first sort out cells with active provirus, then

1883 treat with the potential latency modulator and sort out cells with newly active provirus and
1884 then treat with a strong inducer or alternative stimuli and sort out cell with newly activated
1885 provirus. This would give subsets of cells where latent proviruses had been activated by
1886 treatment and cells with provirus which were not activated by treatment but still inducible.
1887 Synergies between treatments could be assessed and the location of integration sites could
1888 be determined and used to locate patterns of genomic features correlated with induction for
1889 each treatment.

1890 Current efforts at “shock and kill” therapy, inducing latent virus to activate and then
1891 eliminating infected cells, focus on histone deacetylase inhibitors. If there are diverse
1892 mechanisms of latency within patients then much of the latent reservoir may remain
1893 unactivated by single-target therapies. Clinical trials with histone deacetylase inhibitors
1894 have shown some small increases in viral RNA but little decrease in the latent reservoir of
1895 HIV^{60,279–281}. It appears that the majority of viable latent provirus from patient cells are
1896 not reactivated by current therapies²⁸². These results are particularly worrisome because a
1897 functional cure for HIV will likely require a greater than 10,000-fold reduction of the latent
1898 reservoir²⁸³.

1899 In Chapter 2, we used publicly available genomic data. Perhaps there is some chromosomal
1900 feature with a strong association with latency but the data is not currently available or
1901 varies greatly between cell populations. More varieties of annotations are rapidly becoming
1902 available^{284–288}. Decreasing sequencing costs^{289–291} may also make it feasible to measure
1903 more epigenetic features in the exact cell population of interest. Repeating analyses similar
1904 to Chapter 2, perhaps by simply rerunning the reproducible report in Appendix A.2 with
1905 new data, would allow any new features to be monitored for correlations with latency.

1906 **6.2 HIV-1 alternative splicing**

1907 In Chapters 3 and 4, we showed that HIV RNA spliceforms are more diverse than previously
1908 appreciated and estimated the abundances of viral spliceforms. We also showed that splicing

1909 at some splice sites vary between host subjects, between cell types and over the course of
1910 infection. Further characterization of viral splicing would be beneficial to the study and
1911 treatment of HIV-1 especially as there were some technical limitations to our research that
1912 might be improved upon using current techniques.

1913 We studied HIV splicing using droplet PCR¹³⁷ and a set of customized primer in Chapter
1914 3 and bulk sequencing of cellular mRNA in Chapter 4. Sequencing biases and difficulties
1915 determining full length transcripts from short reads hindered characterization of HIV
1916 sequencing. One alternative to these techniques is the targeted capture and enrichment^{292,293}
1917 of HIV-specific sequences. Using probes targeted to conserved regions of HIV, similar to
1918 finding conserved regions for primers as in Chapter 5, would allow for enrichment of viral
1919 reads without the biases induced by primer-based PCR while still allowing for efficient use
1920 of sequencing effort.

1921 The research in Chapter 3 was also limited by a short read bias in the PacBio sequencing.
1922 PacBio sequencing has improved²⁹⁴ and additional long read sequencers have been devel-
1923 oped^{295–297}. In addition, Illumina MiSeq sequencers can now produce 25 million paired 300
1924 bp reads in a single run^{298,299} and better spliceform estimation methods are being devel-
1925 oped^{300,301}. These improved sequencing techniques might allow for more straightforward
1926 analysis of new samples and verification of our previous results.

1927 RNA transcribed antisense to the canonically expressed strand of HIV have been ob-
1928 served^{174,302–307}. These transcripts may be translated to proteins^{308,309} that trigger immune
1929 response in infected individuals^{307,308,310}. Our sequencing techniques were designed only for
1930 the HIV positive strand (Chapter 3) or did not preserve strand information (Chapter 4).
1931 Strand-specific sequencing^{311,312} of multiple HIV strains under varying cellular conditions
1932 would clarify the identity of these transcripts.

1933 Cryptic polypeptides encoding epitopes recognized through major histocompatibility complex
1934 type I also appear to be generated from alternative reading frames in the sense strand of

1935 the virus^{313,314}. Ribosome profiling^{315–317} of infected cells might reveal whether transcripts
1936 generated through alternative splicing or antisense expression are likely to be translated.
1937 These cryptic transcripts could offer new opportunities in vaccine design^{307,310,318,319} but
1938 first their abundance, identity and conservation across strains of HIV must be ascertained.

1939 We observed that splicing varies over the course of infection, between human subjects and
1940 between cell types. Further sampling could reveal additional patterns in these splicing
1941 changes.

1942 Long-lived reservoir of HIV infected cells exist in both macrophages^{320,321} and resting central
1943 memory CD4 T cells^{34,35,40,322,323}. It may be difficult to obtain enough viral RNA from
1944 resting CD4 cells³²² but macrophages provide an interesting target. Splicing changes due to
1945 differing abundances of splice factors have been reported in macrophages¹²⁷. Characterization
1946 of splicing in these important reservoirs might aid in the understanding of latency.

1947 We quantified the splicing of a single clinical isolate and showed unexpected diversity. Most
1948 previous studies of HIV splicing have been performed with lab-adapted strains¹⁰⁵. Additional
1949 studies could determine if the high number of transcripts seen here is an anomaly and whether
1950 additional cryptic splice sites and novel proteins or epitopes exist. In addition, an important
1951 subset of HIV are the founder viruses transmitted between hosts^{324,325}. These viruses are
1952 not well studied and perhaps their splicing and gene expression differ from the rest of the
1953 viral swarm of infected patients. Comparisons to splicing in other retroviral taxa might
1954 highlight evolution and adaptation in this viral lineage.

1955 Disruption of RNA processing can drastically reduce viral replication^{149,326–329}. Small
1956 molecules that inhibit cellular SR splicing proteins and disrupt viral splicing show promise as
1957 antiretroviral therapies^{113,330–332}. Characterization of splicing in cells treated with splicing
1958 inhibitors could reveal potential escape pathways and optimal combinations of drug therapies.

1959 **6.3 Host expression during HIV infection**

1960 In Chapter 4, we saw many changes in host expression and splicing in HIV infected cells
1961 including intron retention and strong changes in apoptotic and innate immunity genes.
1962 We focused on generating a dense data set at a single time point and subject to allow
1963 discrimination of within-condition versus between-condition variation. Further sampling
1964 using more human subjects and time points, improved sequencing techniques, alternative
1965 culturing and extraction and more viral strains would clarify and extend these patterns.

1966 In our primary cell infections, only about 25% of cells were infected with HIV. This makes
1967 it difficult to distinguish between the responses of bystander and infected cells. In addition,
1968 changes in expression due to cellular response to infection are confounded with changes
1969 due to hijacking of cellular controls by the virus. For example, bystander cell death has
1970 been suggested as a major driver of HIV pathogenesis^{333,334} but our data do not make it
1971 clear whether bystander or infected cells are undergoing apoptosis. Cell pull-down with a
1972 labelled HIV strain¹⁹⁵ or an anti-Env antibody³³⁵ or flow cytometry with a labelled antibody
1973 targeting HIV antigen^{54,336} might allow the separation of bystander and infected cells.

1974 Additionally, abortive infections can drive cell death^{334,337} so our populations might be a
1975 mix of three responses; cells responding to a progressive infection, cells responding to an
1976 aborted infection and cells responding to neighbor cell infections. A useful control might be
1977 to infect cells with integrase-deficient virions to guarantee that all infections are aborted.
1978 This would provide a good measure of innate immune response and the effect of abortive
1979 infections undiluted by productive HIV infection and help to deconvolute the patterns seen
1980 in mixed populations.

1981 HIV infection appeared to increase the abundance of intronic sequences. We observed a
1982 significant decrease of nonsense-mediated decay-related genes so perhaps these transcripts
1983 escape degradation due to decreased cellular RNA surveillance. Alternatively, HIV Vpr
1984 protein has been reported to disrupt nuclear integrity and allow mixing of nuclear and

1985 cytoplasmic components³³⁸. These sequences might represent incompletely spliced mRNA
1986 that escaped into the cytoplasm before processing. Infection with a Vpr-deficient HIV virus
1987 and separate isolation of RNA from nuclear and cytoplasmic compartments^{339–341} would
1988 test these hypotheses.

1989 We saw that chimeric sequences were almost entirely derived from read-in or -out from
1990 viral long terminal repeats or splicing from the viral splice donor D4 to human acceptors.
1991 With this knowledge, we could use targeted amplification of these three sites, analogous
1992 to integration site sequencing^{73,102,222}, on cellular cDNA to get a much deeper and cleaner
1993 sampling of chimera formation. Comparison of these data to deeply sequenced integration
1994 site data from the same samples might reveal associations between integration location and
1995 chimera formation.

1996 MicroRNA are small RNAs that block translation through base pairing with complementary
1997 mRNA^{342–344}. Viral derived microRNA, perhaps in part from Dicer processing of the struc-
1998 tured trans-activation response element of HIV^{304,345–347}, may suppress HIV expression^{348–350}
1999 and inhibit apoptosis³⁴⁷ but the presence of such microRNA is controversial^{351,352}. HIV
2000 may suppress silencing by microRNA^{349,353,354} but this is also controversial³⁵². Cellular mi-
2001 croRNA may have antiviral effects^{355,356} or be exploited by HIV to enhance replication^{357–361}
2002 or promote latency^{362,363} but there seems to be disagreement on which microRNA are in-
2003 volved among different studies³⁶⁴. High-throughput genome-wide assays of small RNA^{174,192}
2004 from primary cells infected with patient isolates would help clarify these debates.

2005 **6.4 LAMP PCR and lab-on-a-chip**

2006 In Chapter 5, we report a loop-mediated isothermal amplification system using primers
2007 optimized to detect most subtypes of HIV-1. An alternative to a single broadly targeted
2008 primer set would be to design separate primer sets targeted specifically to each subtype so
2009 that a positive amplification would then be able to discriminate viral subtype. Different viral
2010 subtypes can have different rates of disease progression^{365–368}, transmission dynamics^{369–371}

2011 and response to treatment^{372–374}. Simple low-cost devices with multiple reactions chambers
2012 could be used to both identify viral subtype, estimate viral load^{375,376} and allow more
2013 informed treatment decisions.

2014 A LAMP chip with subtype-specific primers would also allow the detection of intersubtype
2015 superinfections. Superinfection of a single individual with multiple distinct strains of HIV is
2016 common in high risk individuals^{256,377–380} and the general population³⁸¹. Superinfection with
2017 a phenotypically different strain of HIV can lead to disease progression^{382–387} or drug resis-
2018 tance³⁸⁸. Superinfection also allows recombination between divergent strains^{377,383,384,386,389}
2019 and this rapid exchange of genetic information can lead to more fit recombinant strains and
2020 worsen the global epidemic^{384,390–393}. LAMP detection of superinfection could allow early
2021 intervention and suppression in superinfected individuals.

2022 The techniques described in Chapter 5 also allow for rapid development of detection assays
2023 for novel pathogens. For example, in a recent outbreak in West Africa, Zaire ebolavirus
2024 has infected over 26,000 confirmed, probable and suspected cases and caused over 11,000
2025 reported deaths^{394–396}. Early detection and quarantine are essential to the control of this
2026 epidemic³⁹⁷. Amplification of Ebolavirus nucleic acid through polymerase chain reaction is
2027 the best diagnostic test currently available but the necessary resources are often not available
2028 in these resource-poor regions^{398,399}. Antigen-based tests are quicker and available at the
2029 point-of-care but are not as accurate or sensitive as polymerase chain reaction tests and are
2030 still in limited supply³⁹⁹. Loop-mediated isothermal amplification offers the potential for
2031 rapid, sensitive and efficient detection of Ebolavirus RNA but available LAMP primers⁴⁰⁰ do
2032 not match the current outbreak strain. Using sequences from the recent outbreak^{394,401} and
2033 the methods described in Chapter 5, we designed primers to match all known Zaire ebolavirus
2034 (Figure 6.1). These primer combined with simple lab-on-a-chip devices for purifying blood
2035 plasma²⁶¹ and imaging fluorescent signals^{276,375} could allow rapid point-of-care detection of
2036 Ebolavirus.

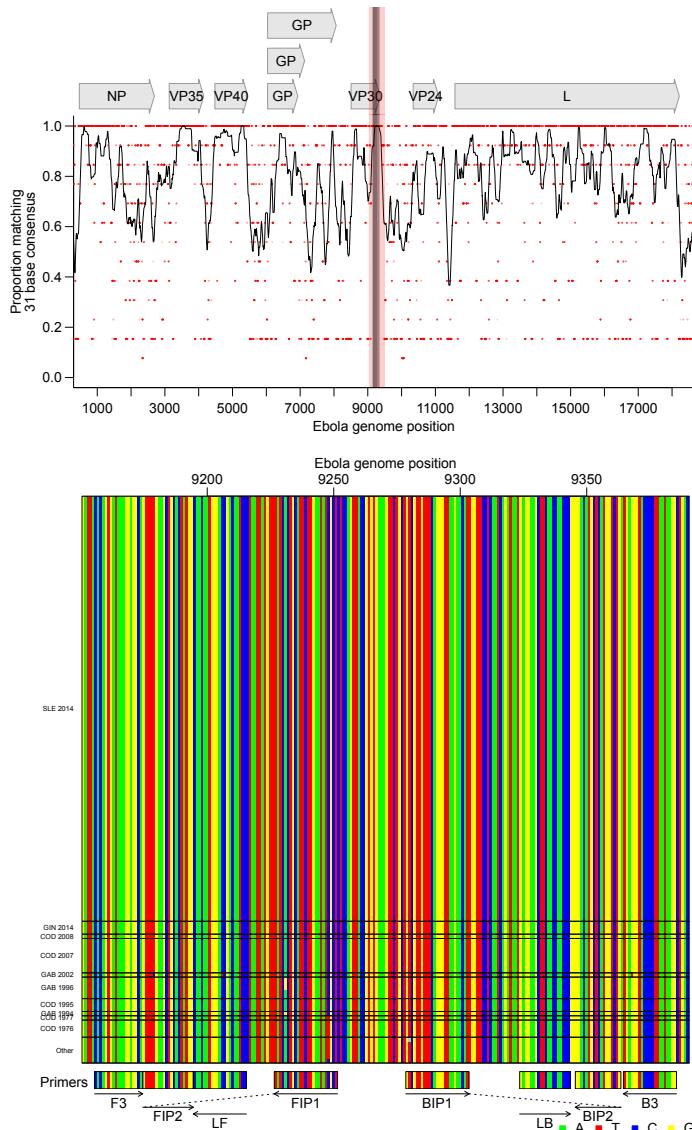


Figure 6.1: Bioinformatic analysis to design Ebola RT-LAMP primers. A) Conservation of sequence in Ebola. Ebola genomes ($n = 131$) from Genbank and sequences from the recent Zaire Ebolavirus outbreak³⁹⁴ were aligned and conservation calculated. The x-axis shows the coordinate on the Ebola genome, the y-axis shows the proportion of sequences matching the consensus for each 21 base segment of the genome (red points). The black line shows a 101 base sliding average over these proportions. The vertical red shading shows the region targeted for LAMP primer design that was used as input into the EIKEN primer design tool and grey shading indicates the area covered by the optimized primer set. Numbering is relative to the Ebola Mayinga sequence. B) Aligned genomes, showing the locations of the LAMP primers. Sequences in the grey-shaded region in A are shown, with DNA bases color-coded as shown at the lower right. Each row indicates an HIV sequence and each column a base in that sequence. Horizontal lines separate Ebolavirus outbreaks (labeled at left). Arrows indicate the strand targeted by each primer. Primers targeting the negative strand of the virus are shown as reverse compliments for ease of viewing.

2037 **6.4.1 Conclusions**

2038 These studies contribute to the study and treatment of HIV-1 by revealing aspects of latency,
2039 expression and host response. They highlight the importance of primary cell models and
2040 the effects that host cell can have on viral processes. With rapidly increasing sequencing
2041 throughput, studies like those presented here offer the opportunity for a deeper and broader
2042 understanding of HIV-1 biology and host response and further development of diagnostics
2043 and therapeutics.

2044 APPENDIX A.1 : Generalized linear models of changes in
2045 use of mutually exclusive HIV-1 splice
2046 acceptors

2047 Reads splicing from D1 to one of five mutually exclusive acceptors, D3, D4c, D4a, D4b,
2048 D5, and D5a, in three primers, 1.2, 1.3 and 1.4, were collected. Since these data are based
2049 on counts, we modeled them as Poisson distributed with an extra variance term allowing
2050 for additional variance using a quasi-Poisson generalized linear model with log link. We
2051 accounted for differences in sequencing effort by including the total number of D1 to mutually
2052 exclusive acceptors reads in each primer-sample as an offset. Differences in the read counts
2053 a) over time,b) between human donor and c) cell type were analyzed separately. A term
2054 was included for each acceptor and its interaction with the variable of interest. The models
2055 included primer and replicate terms and their individual interactions with acceptor to
2056 account for any confounding factors.

2057 A.1.1 HOS vs T Cells

2058 R command:

```
2059 glm(count~offset(log(total)) + acceptor:primer + acceptor:  
2060      isHos  
2061 + acceptor, data = mutEx[mutEx$time == 48,],  
2062 family = 'quasipoisson')
```

2063 Difference between HOS and T cells may be confounded by run differences between early
2064 sequencing and later sequencing. Verification by agarose gel (Figure 3.4b) suggest that these
2065 differences are likely biological.

Variable	Df	Deviance	Resid. Df	Resid. Dev	F	Pr(>F)
NULL	395	138 330				
acceptor	5	133 985	390	4345	9004	$<2.2 \times 10^{-16}$
acceptor:primer	12	751	378	3594	21.03	$<2.2 \times 10^{-16}$
acceptor:isHos	6	2466	372	1127	138.1	$<2.2 \times 10^{-16}$

2066 So after accounting for primer-acceptor bias, the difference between HOS and T cells is
 2068 significant.

2069 The interesting terms in the model are:

Variable	Estimate	Std. Error	t value	Pr(> t)
acceptorA3:isHosTRUE	1.4717	0.065 86	22.35	$<2.2 \times 10^{-16}$
acceptorA4a:isHosTRUE	-0.9449	0.1246	-7.583	2.73×10^{-13}
2070 acceptorA4b:isHosTRUE	-0.9285	0.1059	-8.767	$<2.2 \times 10^{-16}$
acceptorA4c:isHosTRUE	-1.228	0.1066	-11.51	$<2.2 \times 10^{-16}$
acceptorA5:isHosTRUE	0.090 82	0.026 08	3.483	0.000 555
acceptorA5a:isHosTRUE	0.6308	0.079 40	7.945	2.33×10^{-14}

2071 So it appears A3 is up; A4c, A4a and A4b are down; A5 is up a little and A5a up in HOS.

2072 A.1.2 HOS Over Time

2073 R command:

```
2074 glm(value~offset(log(total)) + acceptor + acceptor:primer
2075 + acceptor:time, data=mutEx[mutEx$isHos , ],
2076 family = 'quasipoisson')
```

2077 Looking only within HOS, we see a significant linear effect of time:

Variable	Df	Deviance	Resid. Df	Resid. Dev	F	Pr(>F)
NULL	53	17962				
acceptor	5	17710	48	252.2	6698	$<2.2 \times 10^{-16}$
acceptor:primer	12	18.0	36	234.2	2.834	0.01018
acceptor:time	6	217.8	30	16.4	68.65	3.57×10^{-16}

2079 We are assuming that a particular acceptor will have the same change in all three primers
 2080 here.

2081 The interesting terms are:

Variable	Estimate	Std. Error	t value	Pr(> t)
acceptorA3:time	0.02477	0.001778	13.93	1.22×10^{-14}
acceptorA4a:time	-0.01621	0.002812	-5.765	2.69×10^{-6}
acceptorA4b:time	-0.02526	0.002271	-11.12	3.62×10^{-12}
acceptorA4c:time	0.015867	0.003050	5.202	1.32×10^{-5}
acceptorA5:time	-0.001918	0.0006313	-3.038	0.0049
acceptorA5a:time	0.004919	0.001969	2.499	0.0182

2083 So A3, A4c and A5a increase over time and A4a, A4b and A5 decrease over time. All of
 2084 these coefficients are with a log link and linear and so multiplicative. That means that for
 2085 example A3 will increase 2.5%/hour ($\exp(.0247)$) or equivalently 81% (1.025^{24}) over 24hours.

2086 A.1.3 Between Human Comparison

2087 R command:

```
2088 glm(value~offset(log(total)) + acceptor + acceptor:run
2089 + acceptor:primer + acceptor:subject ,
2090 data=mutEx[!mutEx$ishos,], family = 'quasipoisson')
```

2091 In humans, we added a term to account for any potential run bias between the three
 2092 replicates. Subject refers to the seven human blood donors from which T cells were collected:

Variable	Df	Deviance	Resid. Df	Resid. Dev	F	Pr(>F)
NULL	377	128 430				
acceptor	5	126 446	372	1985	19 598	$<2.2 \times 10^{-16}$
acceptor:run	12	136	360	1849	8.792	1.77×10^{-14}
acceptor:primer	12	850	348	998	54.91	$<2.2 \times 10^{-16}$
acceptor:subject	36	597	312	401	12.86	$<2.2 \times 10^{-16}$

2093 So after accounting for any run and primer bias, subject ID has a statistically significant
 2094 effect on our observed counts. If we compare everything to subject 7, the interesting terms
 2095 are:

Variable	Estimate	Std. Error	t value	Pr(> t)
acceptorA3:subject6	-0.001 399	0.072 86	-0.019	0.9847
acceptorA4a:subject6	-0.112 90	0.049 44	-2.284	0.023 07
acceptorA4b:subject6	-0.054 33	0.040 38	-1.345	0.1795
acceptorA4c:subject6	0.028 29	0.033 60	0.842	0.4005
acceptorA5:subject6	0.016 83	0.016 00	1.051	0.2939
acceptorA5a:subject6	-0.030 85	0.060 92	-0.506	0.6129
acceptorA3:subject5	-0.077 67	0.074 23	-1.046	0.2962
acceptorA4a:subject5	-0.1144	0.049 82	-2.296	0.0223
acceptorA4b:subject5	-0.0684	0.040 90	-1.672	0.0956
acceptorA4c:subject5	-0.085 85	0.034 75	-2.471	0.0140
acceptorA5:subject5	0.038 88	0.016 16	2.406	0.0167
acceptorA5a:subject5	0.078 77	0.060 38	1.304	0.1930
acceptorA3:subject4	-0.1849	0.095 78	-1.931	0.0544
acceptorA4a:subject4	0.071 86	0.057 91	1.241	0.2156
acceptorA4b:subject4	0.126 20	0.047 14	2.677	0.0078
acceptorA4c:subject4	-0.100 21	0.043 03	-2.329	0.0205
acceptorA5:subject4	-0.001 16	0.019 69	-0.059	0.9531
2097 acceptorA5a:subject4	0.023 46	0.073 53	0.319	0.7499
acceptorA3:subject3	-0.003 51	0.086 65	-0.041	0.9677
acceptorA4a:subject3	0.071 07	0.055 64	1.277	0.2024
acceptorA4b:subject3	0.006 46	0.046 99	0.138	0.8907
acceptorA4c:subject3	-0.063 34	0.040 76	-1.554	0.1212
acceptorA5:subject3	0.010 52	0.018 87	0.557	0.5776
acceptorA5a:subject3	-0.070 95	0.072 85	-0.974	0.3309
acceptorA3:subject2	-0.2329	0.091 76	-2.539	0.0116
acceptorA4a:subject2	0.024 05	0.056 43	0.426	0.6702
acceptorA4b:subject2	0.1107	0.045 35	2.441	0.0152
acceptorA4c:subject2	0.021 76	0.039 52	0.551	0.5823
acceptorA5:subject2	-0.003 760	0.018 69	-0.201	0.8407
acceptorA5a:subject2	-0.1608	0.073 51	-2.187	0.0295
acceptorA3:subject1	0.095 36	0.065 56	1.454	0.1468
acceptorA4a:subject1	0.029 32	0.044 31	0.662	0.5087
acceptorA4b:subject1	-0.2144	0.038 43	-5.578	5.28×10^{-8}
acceptorA4c:subject1	-0.3974	0.033 85	-11.74	$<2.2 \times 10^{-16}$
acceptorA5:subject1	0.091 44	0.014 70	6.221	1.58×10^{-9}
acceptorA5a:subject1	0.027 47	0.055 94	0.491	0.6238

2098 So there were small but significant effects between subjects especially between subject 1 and
 2099 subjects 2–7. A potential confounder is that T cells were collected from apheresis product in

²¹⁰⁰ subject 1 and from whole blood in subjects 2–7 although why this would affect later assays
²¹⁰¹ is unknown.

2102 APPENDIX A.2 : Reproducible report of HIV integration
2103 sites and latency analysis

2104 A.2.1 Supplementary data

2105 Additional File 2 is a gzipped csv file that includes a row for each uniquely mapped provirus
2106 and its surrounding genomic annotations. The csv file should have 12436 rows (excluding
2107 header) with 6252 expressed and 6184 latent proviruses.

```
integrationData <- read.csv("AdditionalFile2.csv.gz",  
  stringsAsFactors = FALSE)  
  
nrow(integrationData)  
  
## [1] 12436  
  
table(integrationData$isLatent)  
  
##  
## FALSE TRUE  
## 6252 6184
```

2108 A.2.2 Lasso regression

2109 The lasso regressions take a while to run so I've turned down the number of cross validations
2110 here (set **eval=FALSE** below to completely skip this step). Leave one out and 480-fold cross
2111 validation were used in the paper but processing may take a few days without parallel
2112 processing. Lasso regression requires the R **glmnet** package.

```
notFitColumns <- c("id", "chr", "pos", "strand", "sample", "
```

```
isLatent")
```

```
samples <- unique(as.character(integrationData$sample))
```

```
sampleMatrix <- do.call(cbind, lapply(samples, function(x)
```

```

integrationData$sample ==
x))

colnames(sampleMatrix) <- gsub(" ", "_", samples)

interact <- function(predMatrix, columns, addNames = NULL) {
  out <- do.call(cbind, lapply(1:ncol(columns), function(x)
    predMatrix *
    columns[, x]))
  if (!is.null(addNames)) {
    if (length(addNames) != ncol(columns)) {
      stop(simpleError("Names not same length as columns
"))
    }
    colnames(out) <- sprintf("%s_%s", rep(addNames, each =
      ncol(predMatrix)),
      rep(colnames(predMatrix), length(addNames)))
  }
  return(out)
}

fitData <- as.matrix(integrationData[, !colnames(
  integrationData) %in%
  notFitColumns])

fitData2 <- as.matrix(cbind(interact(fitData, sampleMatrix,
  colnames(sampleMatrix)),
  fitData, sampleMatrix))

```

```
library(glmnet)

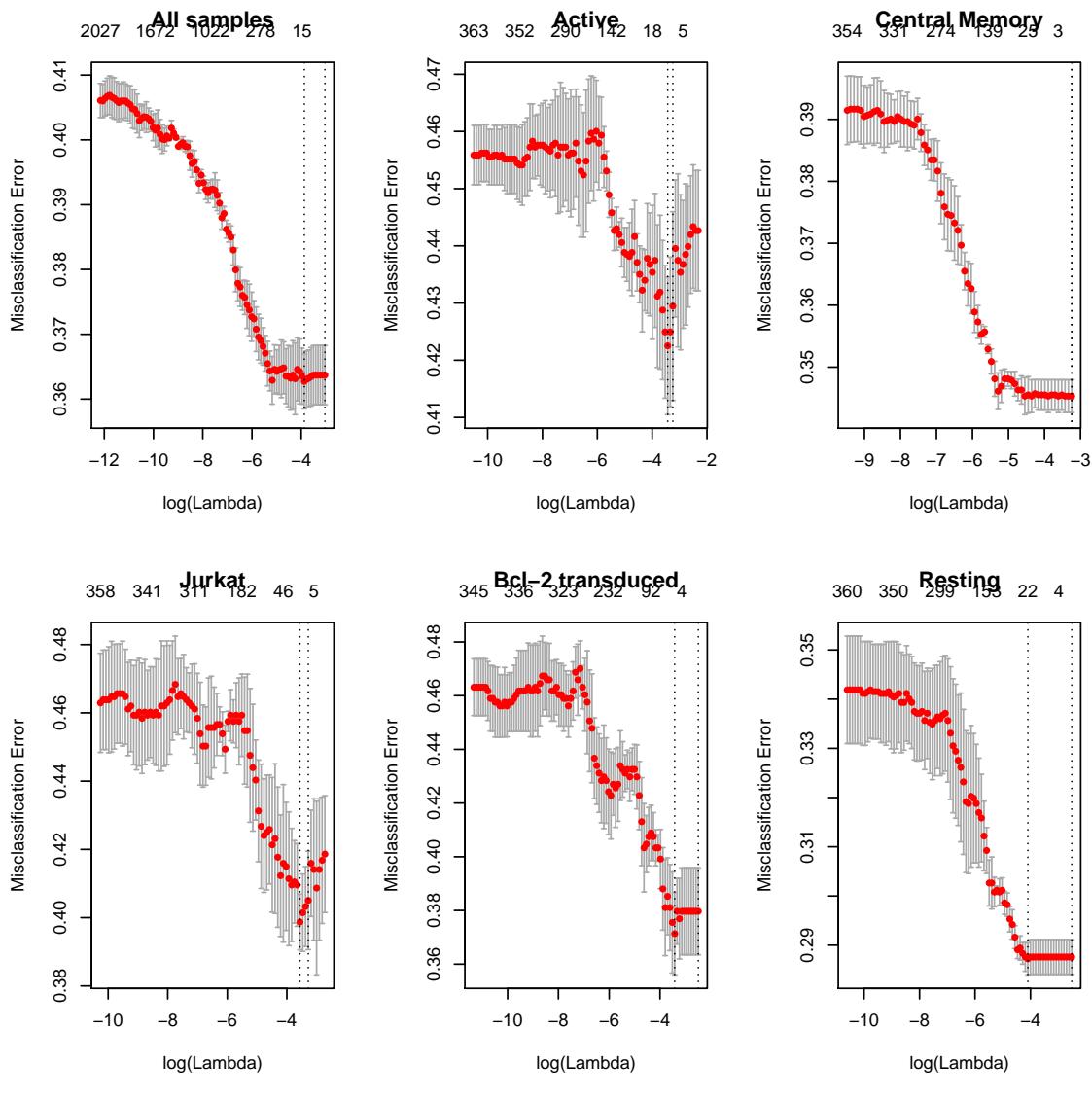
penalties <- rep(1, ncol(fitData2))

penalties[ncol(fitData2) - (ncol(sampleMatrix):1) + 1] <- 0

lassoFit <- cv.glmnet(fitData2, integrationData$isLatent ,
family = "binomial",
type.measure = "class", nfolds = 3, penalty.factor =
penalties)

seperateFits <- lapply(samples, function(x) cv.glmnet(fitData[
integrationData$sample ==
x, ], integrationData$isLatent[integrationData$sample ==
x], family = "binomial", type.measure = "class", nfolds =
3))

names(seperateFits) <- samples
```



2113

2114 **A.2.3 Correlation**

2115 We looked for correlation between the genomic variables and expression status of the
 2116 proviruses.

```
corMat <- apply(fitData, 2, function(x) sapply(samples,
```

```

function(y) {
  selector <- integrationData$sample == y
  if (sd(x[selector]) == 0)
    return(0)
  isLatent <- integrationData[selector, "isLatent"]
  cor(as.numeric(isLatent), x[selector], method = "spearman"
  ""))
})

quantile(corrMat, seq(0, 1, 0.1))

##          0%           10%          20%          30%
## -0.185223020 -0.081555830 -0.048938130 -0.030895834
##          40%           50%          60%          70%
## -0.018053321 -0.005613895  0.003580982  0.017822483
##          80%           90%         100%
##  0.036694554   0.062003356  0.170642314

```

2117 If we looked for genomic variables consistently correlated or anti-correlated with proviral
 2118 expression status with an FDR q-value less than 0.01, no variable was significantly correlated
 2119 in more than 3 samples.

```
pMat <- apply(fitData, 2, function(x) sapply(samples, function
```

```

(y) {

  selector <- integrationData$sample == y
  if (sd(x[selector]) == 0)
    return(NA)

  isLatent <- integrationData[selector, "isLatent"]
  cor.test(as.numeric(isLatent), x[selector], method =
    spearman",
    exact = FALSE)$p.value
}))

adjustPMat <- pMat

adjustPMat[, ] <- p.adjust(pMat, "fdr")

downPMat <- upPMat <- adjustPMat

downPMat [corMat > 0] <- 1

upPMat [corMat < 0] <- 1

table(apply(upPMat < 0.01 & !is.na(upPMat), 2, sum))

##
##      0     1     2     3
## 298   27   38   10

table(apply(downPMat < 0.01 & !is.na(downPMat), 2, sum))

##
##      0     1     2     3
## 216   36   63   58

```

2120 **A.2.4 RNA expression**

2121 We fit a logistic regression to a polynomial of log RNA-Seq reads within 5000 bases from
2122 Jurkat cells for the Jurkat sample and T cells for the rest.

```
rna <- ifelse(integrationData$sample == "Jurkat",
```

```
integrationData$log_jurkatRNA ,
```

```
integrationData$rna_5000)
```

```
rna2 <- rna^2
```

```
rna3 <- rna^3 #
```

```
rna4 <- rna^4
```

```
glmData <- data.frame(isLatent = integrationData$isLatent ,
```

```
sample = integrationData$sample ,
```

```
 rna, rna2, rna3, rna4)
```

```
glmMod <- glm(isLatent ~ sample * rna + sample * rna2 + sample
```

*

```
 rna3 + sample * rna4, data = glmData, family = "binomial")
```

```
summary(glmMod)
```

##

```
## Call:
```

```
## glm(formula = isLatent ~ sample * rna + sample * rna2 +
```

sample *

```
##      rna3 + sample * rna4, family = "binomial", data =
```

```
glmData)
```

##

```
## Deviance Residuals:
```

Min 1Q Median 3Q Max

```
## -2.2899 -0.9864 -0.8676 1.0960 1.6007
```

##

```
## Coefficients:
```

```
##           Estimate Std. Error z value
```

```
## (Intercept) 1.7623655 0.2138859 8.240
```

```
## sampleBcl-2 transduced      -2.1625912   0.7061524   -3.062
```

```
## sampleCentral Memory -2.5010063 0.2437685 -10.260
```

```
## sampleJurkat      -2.0800202   0.2836871   -7.332
```

```
## sampleResting          0.7840481   0.3312247   2.367
```

```
## rna -0.6567268 0.2344422 -2.801
```

```
## rna2          0.1387703   0.0770589   1.801
```

```
## rna3 -0.0167219 0.0094076 -1.777
```

```
## rna4          0.0007572   0.0003845   1.969
```

```
## sampleBcl-2 transduced:rna    0.5750186    0.6366537    0.903
```

```
## sampleCentral Memory:rna      0.9067758  0.2750955  3.296
```

```
## sampleJurkat:rna          0.5294036   0.3867163   1.369
```

```
## sampleResting:rna          0.0366276   0.3436248   0.107
```

```
## sampleBcl-2 transduced:rna2 -0.0369353 0.1878816 -0.197
```

```
## sampleCentral Memory:rna2    -0.2106715   0.0915492   -2.301
```

```
## sampleJurkat: rna2      -0.0766215   0.1641153   -0.467
```

```
## sampleResting:rna2 -0.0760450 0.1086998 -0.700
```

```
## sampleBcl-2 transduced:rna3 0.0032503 0.0213743 0.152
```

```
## sampleCentral Memory:rna3      0.0237064   0.0112661   2.104
```

```
## sampleJurkat: rna3          0.0042183   0.0263910   0.160
```

```
## sampleResting:rna3          0.0153132   0.0128711   1.190
```

```
## sampleBcl-2 transduced:rna4 -0.0002532 0.0008267 -0.306
```

```
## sampleCentral Memory: rna4      -0.0009877   0.0004627   -2.135
```

```
## sampleJurkat: rna4          0.0001725   0.0014215   0.121
```

```
## sampleResting:rna4 -0.0008049 0.0005119 -1.572
```

##

$\Pr(>|z|)$

```
## (Intercept) < 2e-16 ***
```

```
## sampleBcl-2 transduced      0.00219 **
```

```
## sampleCentral Memory < 2e-16 ***
```

```
## sampleJurkat      2.27e-13 ***
```

```
## sampleResting          0.01793 *
```

```
## rna          0.00509 **
```

```
## rna2          0.07173 .
```

```
## rna3          0.07549 .
```

```
## rna4          0.04891 *
```

```

## sampleBcl-2 transduced:rna      0.36643
## sampleCentral Memory:rna      0.00098 ***
## sampleJurkat:rna              0.17101
## sampleResting:rna             0.91511
## sampleBcl-2 transduced:rna2   0.84415
## sampleCentral Memory:rna2     0.02138 *
## sampleJurkat:rna2            0.64059
## sampleResting:rna2           0.48419
## sampleBcl-2 transduced:rna3   0.87913
## sampleCentral Memory:rna3     0.03536 *
## sampleJurkat:rna3            0.87301
## sampleResting:rna3           0.23415
## sampleBcl-2 transduced:rna4   0.75939
## sampleCentral Memory:rna4     0.03280 *
## sampleJurkat:rna4            0.90339
## sampleResting:rna4           0.11585
## ---
## Signif. codes:
## 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
## Null deviance: 17240  on 12435  degrees of freedom
## Residual deviance: 15874  on 12411  degrees of freedom
## AIC: 15924
##
## Number of Fisher Scoring iterations: 4

```

2123 **A.2.5 Strand orientation**

2124 We used a Fisher's exact test to check if silent/inducible proviruses were enriched when
2125 integrated in the same strand orientation as cellular genes.

```
selector <- integrationData$inGene == 1
```

```
strandTable <- with(integrationData[selector, ], table(ifelse(
```

isLatent ,

```
"Silent/Inducible", "Active"), ifelse(inGeneSameStrand ==
```

```
1 , "Same" , "Diff") , sample))
```

```
apply(strandTable, 3, fisher.test)
```

```
## $Active
```

##

```
## Fisher's Exact Test for Count Data
```

##

```
## data: array(newX[, i], d.call, dn.call)
```

```
## p-value = 0.06061
```

```
## alternative hypothesis: true odds ratio is not equal to 1
```

```
## 95 percent confidence interval:
```

```
## 0.7219466 1.0081995
```

```
## sample estimates:
```

```
## odds ratio
```

```
## 0.8532127
```

##

##

```
## $`Bcl-2 transduced`
```

##

```
## Fisher's Exact Test for Count Data
```

##

```
## data: array(newX[, i], d.call, dn.call)
```

```
## p-value = 2.177e-05
```

```
## alternative hypothesis: true odds ratio is not equal to 1
```

```
## 95 percent confidence interval:
```

```
## 1.446896 2.872562
```

```
## sample estimates:
```

```
## odds ratio
```

2.036148

##

##

```
## $`Central Memory`
```

##

```
## Fisher's Exact Test for Count Data
```

##

```
## data: array(newX[, i], d.call, dn.call)
```

```
## p-value = 0.2907
```

```
## alternative hypothesis: true odds ratio is not equal to 1
```

```
## 95 percent confidence interval:
```

```
## 0.9386167 1.2320238
```

```
## sample estimates:
```

```
## odds ratio
```

1.07529

##

```
##  
## $Jurkat  
##  
## Fisher's Exact Test for Count Data  
##  
## data: array(newX[, i], d.call, dn.call)  
## p-value = 0.1674  
## alternative hypothesis: true odds ratio is not equal to 1  
## 95 percent confidence interval:  
## 0.9207548 1.5699893  
## sample estimates:  
## odds ratio  
## 1.202007  
##  
##  
## $Resting  
##  
## Fisher's Exact Test for Count Data  
##  
## data: array(newX[, i], d.call, dn.call)  
## p-value = 0.5732  
## alternative hypothesis: true odds ratio is not equal to 1  
## 95 percent confidence interval:  
## 0.7825231 1.1405158  
## sample estimates:  
## odds ratio  
## 0.9447415
```

2126 **A.2.6 Acetylation**

2127 To reduce correlation between acetylation marks, we generated the first ten principal
2128 components of the acetylation data and ran a logistic regression against them. We compared
2129 the cross validated performance of this regression with a base model only including which
2130 dataset the integration site came from. The cross-validation here has been reduced for
2131 efficiency but 480-fold cross-validation was used in the paper.

```
acetyl <- integrationData[, !grepl("logDist", colnames(
```

```
integrationData)) &
```

```
grepl("ac", colnames(integrationData))]
```

```
acetylPCA <- princomp(acetyl)
```

```
cumsum(acetylPCA$sdev[1:10]^2/sum(acetylPCA$sdev^2))
```

Comp . 1 Comp . 2 Comp . 3 Comp . 4 Comp . 5 Comp . 6

```
## 0.5947268 0.6786611 0.7267433 0.7610502 0.7833616 0.7964470
```

Comp .7 Comp .8 Comp .9 Comp .10

```
## 0.8093295 0.8215027 0.8299358 0.8372584
```

```
cv.glm <- function(model, K = nrow(thisData), subsets = NULL)
```

{

```
modelCall <- model$call
```

```
thisData <- eval(modelCall$data)
```

```
n <- nrow(thisData)
```

```
if (is.null(subsets))
```

```
subsets <- split(1:n, sample(rep(1:K, length.out = n)))
```

)

```
preds <- lapply(subsets, function(outGroup) {
```

```
subsetData <- thisData[-outGroup, , drop = FALSE]
```

```
predData <- thisData[outGroup, , drop = FALSE]
```

```
thisModel <- modelCall
```

```
thisModel$data <- subsetData
```

```
return(predict(eval(thisModel), predData))
```

})

```
pred <- unlist(preds)[order(unlist(subsets))]
```

```
subsetId <- rep(1:K, sapply(subsets, length))[order(unlist
```

```
(subsets))]
```

```
return(data.frame(pred, subsetId))
```

}

```
inData <- data.frame(isLatent = integrationData$isLatent ,
```

```
sample = as.factor(integrationData$sample),
```

```
acetylPCA$score[, 1:10])
```

```
modelPreds <- cv.glm(glm(isLatent ~ sample + Comp.1 + Comp.2 +
```

Comp.3 + Comp.4 + Comp.5 + Comp.6 + Comp.7 + Comp.8 + Comp

. 9 +

```
Comp.10, family = "binomial", data = inData), K = 5)
```

```
basePreds <- cv.glm(glm(isLatent ~ sample, family = "binomial
```

```

" ,
  data = inData), subsets = split(1:nrow(inData),
  modelPreds$subsetId),
  K = 5)

modelCorrect <- sum((modelPreds$pred > 0) ==
  integrationData$isLatent)
baseCorrect <- sum((basePreds$pred > 0) ==
  integrationData$isLatent)

prop.test(c(baseCorrect, modelCorrect), rep(nrow(
  integrationData),
  2))

##
##      2-sample test for equality of proportions with
##      continuity correction
##
## data: c(baseCorrect, modelCorrect) out of rep(nrow(
##   integrationData), 2)
## X-squared = 0.00017372, df = 1, p-value = 0.9895
## alternative hypothesis: two.sided
## 95 percent confidence interval:
##   -0.01187726  0.01219890
## sample estimates:
##   prop 1   prop 2
## 0.6362978 0.6361370

```

2132 **A.2.7 Gene deserts**

2133 We used Fisher's exact test to look for an association between integration outside a gene
2134 and proviral expression status.

```
geneTable <- table(integrationData$isLatent ,
```

```
integrationData$inGene ,
```

```
integrationData$sample)
```

```
apply(geneTable, 3, fisher.test)
```

```
## $Active
```

##

```
## Fisher's Exact Test for Count Data
```

##

```
## data: array(newX[, i], d.call, dn.call)
```

```
## p-value < 2.2e-16
```

```
## alternative hypothesis: true odds ratio is not equal to 1
```

```
## 95 percent confidence interval:
```

```
## 0.3629548 0.5446204
```

```
## sample estimates:
```

```
## odds ratio
```

0.4452621

##

##

```
## $`Bcl-2 transduced`
```

##

```
## Fisher's Exact Test for Count Data
```

##

```
## data: array(newX[, i], d.call, dn.call)
```

```
## p-value = 0.1052
```

```
## alternative hypothesis: true odds ratio is not equal to 1
```

```
## 95 percent confidence interval:
```

```
## 0.9203418 2.3478599
```

```
## sample estimates:
```

```
## odds ratio
```

1.472224

##

##

```
## $`Central Memory`
```

##

```
## Fisher's Exact Test for Count Data
```

##

```
## data: array(newX[, i], d.call, dn.call)
```

```
## p-value = 0.7803
```

```
## alternative hypothesis: true odds ratio is not equal to 1
```

```
## 95 percent confidence interval:
```

```
## 0.8525329 1.1253952
```

```
## sample estimates:
```

```
## odds ratio
```

0.9791165

##

```
##  
## $Jurkat  
##  
##      Fisher's Exact Test for Count Data  
##  
## data: array(newX[, i], d.call, dn.call)  
## p-value = 0.5443  
## alternative hypothesis: true odds ratio is not equal to 1  
## 95 percent confidence interval:  
## 0.7909269 1.6167285  
## sample estimates:  
## odds ratio  
## 1.127836  
##  
##  
## $Resting  
##  
##      Fisher's Exact Test for Count Data  
##  
## data: array(newX[, i], d.call, dn.call)  
## p-value = 3.071e-08  
## alternative hypothesis: true odds ratio is not equal to 1  
## 95 percent confidence interval:  
## 0.4384828 0.6864112  
## sample estimates:  
## odds ratio  
## 0.5500205
```

2135 We used a two-sample t-test to investigate whether there was a significant difference in
2136 distance to the nearest gene between expressed and silent/inducible proviruses integrated
2137 outside genes.

```
geneDistData <- integrationData[!integrationData$inGene, c("
```

isLatent",

```
"logDist_nearest", "sample")]
```

```
by(geneDistData, geneDistData$sample, function(x) t.test(
```

logDist_nearest ~

```
isLatent , data = x))
```

```
## geneDistData$sample: Active
```

##

```
## Welch Two Sample t-test
```

##

```
## data: logDist_nearest by isLatent
```

```
## t = -2.4539, df = 287.73, p-value = 0.01472
```

```
## alternative hypothesis: true difference in means is not
```

equal to 0

```
## 95 percent confidence interval:
```

```
## -0.80738340 -0.08867607
```

```
## sample estimates:
```

```
## mean in group FALSE  mean in group TRUE
```

##

9.608737

10.056767

##

```
## geneDistData$sample: Bcl-2 transduced
```

##

```
## Welch Two Sample t-test
```

##

```
## data: logDist_nearest by isLatent
```

```
## t = 0.40978, df = 86.2, p-value = 0.683
```

```
## alternative hypothesis: true difference in means is not
```

equal to 0

```
## 95 percent confidence interval:
```

```
## -0.6309351 0.9586004
```

```
## sample estimates:
```

```
## mean in group FALSE  mean in group TRUE
```

##

9.036872

8.873039

##

```
## geneDistData$sample: Central Memory
```

##

```
## Welch Two Sample t-test
```

##

```
## data: logDist_nearest by isLatent
```

```
## t = -0.07188, df = 861.61, p-value = 0.9427
```

```
## alternative hypothesis: true difference in means is not
```

equal to 0

```
## 95 percent confidence interval:
```

```
## -0.2371374 0.2203819
```

```
## sample estimates:
```

```
## mean in group FALSE  mean in group TRUE
```

##

10.19225

10.20063

##

```
## geneDistData$sample: Jurkat
```

```
##  
##      Welch Two Sample t-test  
##  
## data: logDist_nearest by isLatent  
## t = -1.8217, df = 139.56, p-value = 0.07064  
## alternative hypothesis: true difference in means is not  
## equal to 0  
## 95 percent confidence interval:  
## -1.26342086 0.05167979  
## sample estimates:  
## mean in group FALSE mean in group TRUE  
## 9.925782 10.531652  
##  
## -----  
## geneDistData$sample: Resting  
##  
##      Welch Two Sample t-test  
##  
## data: logDist_nearest by isLatent  
## t = -5.1275, df = 193.49, p-value = 7.096e-07  
## alternative hypothesis: true difference in means is not  
## equal to 0  
## 95 percent confidence interval:  
## -1.2687917 -0.5638568  
## sample estimates:  
## mean in group FALSE mean in group TRUE  
## 9.489931 10.406255
```

2138 To check for a relationship between silent/inducible status and distance to CpG islands, we
2139 used a two sample t-test on the logged distance and saw a significant difference between
2140 silent/inducible and expressed proviruses (before accounting for a correlation between being
2141 near CpG islands and in genes)

```
t.test(integrationData$logDist_cpg ~ integrationData$isLatent)
```

```

## Welch Two Sample t-test
##
## data: integrationData$logDist_cpg by
## integrationData$isLatent
## t = -2.0233, df = 12381, p-value = 0.04306
## alternative hypothesis: true difference in means is not
## equal to 0
## 95 percent confidence interval:
## -0.105657514 -0.001675563
## sample estimates:
## mean in group FALSE mean in group TRUE
## 10.16362 10.21728

sapply(unique(integrationData$sample), function(x) with(
  integrationData[integrationData$sample ==
    x, ], p.adjust(t.test(logDist_cpg ~ isLatent)$p.value,
    method = "bonferroni",
    n = 5)))

## Active Central Memory Jurkat
## 0.512040457 1.000000000 1.000000000
## Bcl-2 transduced Resting
## 1.000000000 0.005866539

```

2142 Many CpG islands are found near genes. To account for this relationship, we used an ANOVA
 2143 test including whether the integration site was inside a gene prior to including CpG islands.

2144 After including integration inside genes, CpG islands were not significantly associated with
2145 silent/inducible status of the proviruses with all samples grouped or individually after
2146 Bonferroni correction for multiple comparisons.

```
anova(with(integrationData, glm(isLatent ~ I(logDist_nearest
```

==

```
0) + logDist_cpg, family = "binomial")) , test = "Chisq")
```

```
## Analysis of Deviance Table
```

##

```
## Model: binomial, link: logit
```

##

```
## Response: isLatent
```

##

```

## Terms added sequentially (first to last)

## 

##                               Df Deviance Resid. Df Resid. Dev

## NULL                           12435      17240

## I(logDist_nearest == 0)    1   26.2682     12434      17213

## logDist_cpg                  1     1.1328     12433      17212

##                                     Pr(>Chi)

## NULL

## I(logDist_nearest == 0)  2.971e-07 ***

## logDist_cpg                  0.2872

## ---

## Signif. codes:

## 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

sapply(unique(integrationData$sample), function(x) {

  p.adjust(anova(with(integrationData[integrationData$sample

  ==

  x, ] , glm(isLatent ~ I(logDist_nearest == 0) + logDist_cpg , family = "binomial")) , test = "Chisq")["logDist_cpg" , "Pr(>Chi)"] , method = "bonferroni" , n = 5)

})

##          Active   Central Memory           Jurkat

## 1.0000000 1.0000000 1.0000000

## Bcl-2 transduced             Resting

## 1.0000000 0.2007788

```

2147 A.2.8 Alphoid repeats

2148 When analyzing repetitive elements, we treated each read as an independent observation and
2149 included reads with multiple alignments to the genome. Additional File 3 is a gzipped csv file
2150 containing a row for each read with multiple alignments and one row for each dereplicated
2151 integration site with a single alignment with the count variable indicating the number of
2152 reads dereplicated to that integration site. There should be 26,190 rows (excluding header)
2153 with 14,494 rows of expressed provirus and 11,696 rows of silent/inducible provirus.

```
repeats <- read.csv("AdditionalFile3.csv.gz", check.names =  
  FALSE,  
  stringsAsFactors = FALSE)  
  
nrow(repeats)  
  
## [1] 26190  
  
summary(repeats$isLatent)  
  
##      Mode   FALSE    TRUE     NA 's  
##  logical  14494    11696       0  
  
notRepeatColumns <- c("id", "isLatent", "sample", "count")
```

2154 To analyze whether there was an association between proviral expression status and integra-
2155 tion within alphoid repeats, we used Fisher's exact test with a Bonferroni correction for five
2156 samples. For comparison, we looked at the association between proviral expression and the
2157 other repeats in the RepeatMasker database. We did not Bonferroni correct for the multiple
2158 repeat types so that the repeats could be compared with the analysis of alphoid repeats (for
2159 which we had an a priori hypothesis for an association with latency).

```
dummyX <- rep(c(TRUE, FALSE), 2)
```

```
dummyY <- rep(c(TRUE, FALSE), each = 2)
```

```
repeatData <- repeats[, !colnames(repeats) %in%
```

```
notRepeatColumns]
```

```
repeatData <- repeatData[, apply(repeatData, 2, sum) > 0]
```

```
testRepeats <- function(x, repeats) {
```

```
sapply(samples, function(thisSample, repeats) {
```

```

    selector <- repeats$sample == thisSample
    repLatent <- rep(repeats$isLatent[selector],
                      repeats$count[selector])
    repRepeat <- rep(x[selector], repeats$count[selector])
    fisher.test(table(c(dummyX, repLatent), c(dummyY,
                                                repRepeat)) -
                1)$p.value
  }, repeats)
}

repeatPs <- apply(repeatData, 2, testRepeats, repeats[, notRepeatColumns])

table(apply(repeatPs * 5 < 0.05, 2, sum))

## 
##      0      1      2      3
##  611   76   15     1

which(apply(repeatPs * 5 < 0.05, 2, sum) >= 3)

## ALR/Alpha
##          178

p.adjust(repeatPs[, "ALR/Alpha"], "bonferroni")

##           Active   Central Memory        Jurkat
##      5.026890e-02   3.940207e-03   1.027189e-08
## Bcl-2 transduced           Resting
##      1.000000e+00   2.424896e-02

```

2160 **A.2.9 Neighbors**

2161 We looked at all pairs of viruses on the same chromosome separated by no more than a
2162 given distance, e.g. 100 bases, either with all samples pooled or split between within sample
2163 pairs or between sample pairs.

```
allNeighbors <- data.frame(id1 = 0, id2 = 0)[0, ]
```

```
ids <- 1:nrow(integrationData)
```

```

for (chr in unique(integrationData$chr)) {
  chrSelector <- integrationData$chr == chr
  neighborPairs <- data.frame(id1 = rep(ids[chrSelector],
    sum(chrSelector)),
    id2 = rep(ids[chrSelector], each = sum(chrSelector)))
  neighborPairs <- neighborPairs[neighborPairs$id1 <
    neighborPairs$id2,
  ]
  allNeighbors <- rbind(allNeighbors, neighborPairs)
}

allNeighbors$dist <- abs(integrationData$pos[allNeighbors$id1] -
  -
  integrationData$pos[allNeighbors$id2])

allNeighbors$latent1 <- integrationData$isLatent [
  allNeighbors$id1]

allNeighbors$latent2 <- integrationData$isLatent [
  allNeighbors$id2]

allNeighbors$sample1 <- integrationData$sample [
  allNeighbors$id1]

allNeighbors$sample2 <- integrationData$sample [
  allNeighbors$id2]

allNeighbors <- allNeighbors[allNeighbors$dist <= 1e+06, ]

```

2164 The expected number of matching pairs was calculated as $\sum_{j \in \text{samples}} n_{j,d}(\theta_{j,d}\theta_{\neg j,d} + (1 -$
2165 $\theta_{j,d})(1 - \theta_{\neg j,d}))$ for between sample, $\sum_{j \in \text{samples}} n_{j,d}(\theta_{j,d}^2 + (1 - \theta_{j,d})^2)$ for within sample and
2166 $n_d(\theta_d^2 + (1 - \theta_d)^2)$ for all pairs, where $n_{j,d}$ is the number of pairs of proviruses separated by
2167 no more than d base pairs where the first provirus is from sample j , $\theta_{j,d}$ is the proportion of
2168 silent/inducible proviruses in sample j appearing in at least one pair of proviruses separated
2169 by less than d base pairs and $\neg j$ means all samples except sample j .

```
dists <- unique(round(10^seq(1, 6, 1)))
```

```
pairings <- do.call(rbind, lapply(dists, function(x,
```

```
allNeighbors) {
```

```
inSelector <- allNeighbors$dist <= x &
```

```
allNeighbors$sample1 ==
```

```
allNeighbors$sample2
```

```
outSelector <- allNeighbors$dist <= x &
```

```
allNeighbors$sample1 !=
```

```
allNeighbors$sample2
```

```
allSelector <- allNeighbors$dist <= x
```

```
out <- data.frame(dist = x, observedIn = sum(allNeighbors[
```

```
inSelector ,
```

```
"latent1"] == allNeighbors[inSelector, "latent2"]),
```

```
observedOut = sum(allNeighbors[outSelector,
```

```
"latent1"] == allNeighbors[outSelector, "latent2"]),
```

```
observedAll = sum(allNeighbors[allSelector, "latent1"]
```

==

```
allNeighbors[allSelector, "latent2"])), totalIn =
```

```
    sum(inSelector),
```

```
totalOut = sum(outSelector), totalAll = sum(
```

```
allSelector))
```

```
out$expectedIn <- sum(with(allNeighbors[inSelector, ],
```

```
sapply(samples,
```

```
function(x) {
```

```
inLatent <- c(latent1[sample1 == x], latent2[
```

```
sample2 ==
```

```
x]) [!duplicated(c(id1[sample1 == x], id2[
```

```
sample2 ==
```

x]))]

```
if (length(inLatent) == 0) return(0)
```

```
return(sum(sample1 == x) * (mean(inLatent)^2 +
```

```
mean(!inLatent)^2))
```

})))

```
out$expectedOut <- sum(with(allNeighbors[outSelector, ],
```

```
sapply(samples, function(x) {
```

```

inLatent <- c(latent1[sample1 == x], latent2[
  sample2 ==
  x])[!duplicated(c(id1[sample1 == x], id2[
    sample2 ==
    x]))]

outLatent <- c(latent1[sample1 != x], latent2[
  sample2 !=
  x])[!duplicated(c(id1[sample1 != x], id2[
    sample2 !=
    x]))]

if (length(inLatent) == 0) return(0)

return(sum(sample1 == x) * (mean(inLatent) * mean(
  outLatent) +
  mean(!inLatent) * mean(!outLatent)))
  }))

out$expectedAll <- sum(with(allNeighbors[, allSelector, ],
{
  allLatent <- c(latent1, latent2)[!duplicated(c(id1
  ,
  id2))]

  return(length(latent1) * (mean(allLatent)^2 + mean
  (!allLatent)^2))
  }))

  return(out)
}, allNeighbors))

rownames(pairings) <- pairings$dist

```

2170 To look for more matches than expected by random pairing between neighboring proviruses,
2171 we used a one sample Z-test of proportion to compare the observed number of matching
2172 pairs with the expected proportion of pairs.

```
combinations <- c(All = "All", `Between sample` = "Out", `
```

```
Within sample` = "In")
```

```
lapply(combinations, function(x, pairing) {
```

```
vars <- sprintf(c("observed%s", "expected%s", "total%s"),
```

x)

```
expectedProb <- pairing[, vars[2]]/pairing[, vars[3]]
```

```
prop.test(pairing[, vars[1]], pairing[, vars[3]], p =
```

```
expectedProb)
```

```
}, pairings ["100", ])
```

\$All

##

```
## 1-sample proportions test with continuity correction
```

##

```
## data: pairing[, vars[1]] out of pairing[, vars[3]], null
```

```
probability expectedProb
```

```
## X-squared = 13.002, df = 1, p-value = 0.0003111
```

```
## alternative hypothesis: true p is not equal to 0.5000141
```

```
## 95 percent confidence interval:
```

```
## 0.5586837 0.6962353
```

```
## sample estimates:
```

p

0.63

##

##

```
## $`Between sample`
```

```

##  

##      1-sample proportions test with continuity correction  

##  

## data:  pairing[, vars[1]] out of pairing[, vars[3]], null  

##       probability expectedProb  

## X-squared = 0.21919, df = 1, p-value = 0.6397  

## alternative hypothesis: true p is not equal to 0.4836763  

## 95 percent confidence interval:  

##  0.3570532 0.5572662  

## sample estimates:  

##  

##          p  

## 0.4554455  

##  

##  

##  

## $`Within sample`  

##  

##      1-sample proportions test with continuity correction  

##  

## data:  pairing[, vars[1]] out of pairing[, vars[3]], null  

##       probability expectedProb  

## X-squared = 24.446, df = 1, p-value = 7.644e-07  

## alternative hypothesis: true p is not equal to 0.5561437  

## 95 percent confidence interval:  

##  0.7140170 0.8776751  

## sample estimates:  

##  

##          p  

## 0.8080808

```

2173 A.2.10 Compiling this document

2174 This document was generated using R's Sweave function ([http://en.wikipedia.org/
wiki/Sweave](http://en.wikipedia.org/wiki/Sweave)). If you would like to regenerate this document, download Additional Files
2175 2, 3 and 4 from Sherrill-Mix et al.³³ and make sure the files are all in the same directory
2176 and named AdditionalFile2.csv.gz, AdditionalFile3.csv.gz and AdditionalFile4.Rnw. Then
2177 compile by going to that directory and using the commands:

2179 R CMD Sweave AdditionalFile4.Rnw
2180 pdflatex AdditionalFile4.tex

2181

2182 Note that you will need R and L^AT_EX (and the R package glmnet if you would like to rerun
2183 the lasso regressions) installed.

BIBLIOGRAPHY

- 2185 [1] RS Yalow and SA Berson. 1960. Immunoassay of endogenous plasma insulin in man.
2186 *J Clin Invest*, 39:1157–1175. doi: 10.1172/JCI104130
- 2187 [2] E Engvall and P Perlmann. 1971. Enzyme-linked immunosorbent assay (ELISA).
2188 Quantitative assay of immunoglobulin G. *Immunochemistry*, 8:871–874. doi: 10.1016/
2189 0019-2791(71)90454-X
- 2190 [3] BK Van Weemen and AH Schuurs. 1971. Immunoassay using antigen-enzyme conju-
2191 gates. *FEBS Lett*, 15:232–236. doi: 10.1016/0014-5793(71)80319-8
- 2192 [4] F Barré-Sinoussi, JC Chermann, F Rey, MT Nugeyre, S Chamaret, J Gruest, C Dau-
2193 guet, C Axler-Blin, F Vézinet-Brun et al. 1983. Isolation of a T-lymphotropic retrovirus
2194 from a patient at risk for acquired immune deficiency syndrome (AIDS). *Science*, 220:
2195 868–871
- 2196 [5] RC Gallo, PS Sarin, EP Gelmann, M Robert-Guroff, E Richardson, VS Kalyanaraman,
2197 D Mann, GD Sidhu, RE Stahl et al. 1983. Isolation of human T-cell leukemia
2198 virus in acquired immune deficiency syndrome (AIDS). *Science*, 220:865–867. doi:
2199 10.1126/science.6601823
- 2200 [6] M Popovic, MG Sarngadharan, E Read and RC Gallo. 1984. Detection, isolation, and
2201 continuous production of cytopathic retroviruses (HTLV-III) from patients with AIDS
2202 and pre-AIDS. *Science*, 224:497–500. doi: 10.1126/science.6200935
- 2203 [7] JA Levy, AD Hoffman, SM Kramer, JA Landis, JM Shimabukuro and LS Oshiro.
2204 1984. Isolation of lymphocytopathic retroviruses from San Francisco patients with
2205 AIDS. *Science*, 225:840–842. doi: 10.1126/science.6206563
- 2206 [8] B Safai, MG Sarngadharan, JE Groopman, K Arnett, M Popovic, A Sliski, J Schüpbach
2207 and RC Gallo. 1984. Seroepidemiological studies of human T-lymphotropic retrovirus
2208 type III in acquired immunodeficiency syndrome. *Lancet*, 1:1438–1440. doi: 10.1016/
2209 S0140-6736(84)91933-0
- 2210 [9] MG Sarngadharan, M Popovic, L Bruch, J Schüpbach and RC Gallo. 1984. Antibodies
2211 reactive with human T-lymphotropic retroviruses (HTLV-III) in the serum of patients
2212 with AIDS. *Science*, 224:506–508. doi: 10.1126/science.6324345
- 2213 [10] H Towbin, T Staehelin and J Gordon. 1979. Electrophoretic transfer of proteins from
2214 polyacrylamide gels to nitrocellulose sheets: procedure and some applications. *Proc
2215 Natl Acad Sci U S A*, 76:4350–4354
- 2216 [11] Centers for Disease Control. 1985. Provisional Public Health Service inter-agency
2217 recommendations for screening donated blood and plasma for antibody to the virus
2218 causing acquired immunodeficiency syndrome. *MMWR Morb Mortal Wkly Rep*, 34:1–5

- 2219 [12] DS Burke and RR Redfield. 1986. False-positive Western blot tests for antibodies to
2220 HTLV-III. *JAMA*, 256:347. doi: 10.1001/jama.1986.03380030049013
- 2221 [13] DS Burke, JF Brundage, RR Redfield, JJ Damato, CA Schable, P Putman, R Visintine
2222 and HI Kim. 1988. Measurement of the false positive rate in a screening program
2223 for human immunodeficiency virus infections. *N Engl J Med*, 319:961–964. doi:
2224 10.1056/NEJM198810133191501
- 2225 [14] RJ Chappel, KM Wilson and EM Dax. 2009. Immunoassays for the diagnosis of HIV:
2226 meeting future needs by enhancing the quality of testing. *Future Microbiol*, 4:963–982.
2227 doi: 10.2217/fmb.09.77
- 2228 [15] B Weber, EH Fall, A Berger and HW Doerr. 1998. Reduction of diagnostic window
2229 by new fourth-generation human immunodeficiency virus screening assays. *J Clin
2230 Microbiol*, 36:2235–2239
- 2231 [16] B Weber, L Görtler, R Thorstensson, U Michl, A Mühlbacher, P Bürgisser, R Vil-
2232 laescusa, A Eiras, C Gabriel et al. 2002. Multicenter evaluation of a new automated
2233 fourth-generation human immunodeficiency virus screening assay with a sensitive
2234 antigen detection module and high specificity. *J Clin Microbiol*, 40:1938–1946. doi:
2235 10.1128/JCM.40.6.1938-1946.2002
- 2236 [17] WJ Kassler, C Haley, WK Jones, AR Gerber, EJ Kennedy and JR George. 1995.
2237 Performance of a rapid, on-site human immunodeficiency virus antibody assay in a
2238 public health setting. *J Clin Microbiol*, 33:2899–2902
- 2239 [18] Centers for Disease Control and Prevention. 1998. Update: HIV counseling and testing
2240 using rapid tests—United States, 1995. *MMWR Morb Mortal Wkly Rep*, 47:211–215
- 2241 [19] Centers for Disease Control and Prevention. 2002. Approval of a new rapid test for
2242 HIV antibody. *MMWR Morb Mortal Wkly Rep*, 51:1051–1052
- 2243 [20] D Gallo, JR George, JH Fitchen, AS Goldstein and MS Hindahl. 1997. Evaluation of a
2244 system using oral mucosal transudate for HIV-1 antibody screening and confirmatory
2245 testing. OraSure HIV Clinical Trials Group. *JAMA*, 277:254–258. doi: 10.1001/jama.
2246 1997.03540270080030
- 2247 [21] KP Delaney, BM Branson, A Uniyal, PR Kerndt, PA Keenan, K Jafa, AD Gardner,
2248 DJ Jamieson and M Bulterys. 2006. Performance of an oral fluid rapid HIV-1/2
2249 test: experience from four CDC studies. *AIDS*, 20:1655–1660. doi: 10.1097/01.aids.
2250 0000238412.75324.82
- 2251 [22] C Semá Baltazar, C Raposo, IV Jani, D Shodell, D Correia, C Gonçalves da Silva,
2252 M Kalou, H Patel and B Parekh. 2014. Evaluation of performance and acceptability
2253 of two rapid oral fluid tests for HIV detection in Mozambique. *J Clin Microbiol*, 52:
2254 3544–3548. doi: 10.1128/JCM.01098-14

- 2255 [23] TC Granade, BS Parekh, SK Phillips and JS McDougal. 2004. Performance of the
2256 OraQuick and Hema-Strip rapid HIV antibody detection assays by non-laboratorians.
2257 *J Clin Virol*, 30:229–232. doi: 10.1016/j.jcv.2003.12.006
- 2258 [24] N Pant Pai, J Sharma, S Shivkumar, S Pillay, C Vadnais, L Joseph, K Dheda and
2259 RW Peeling. 2013. Supervised and unsupervised self-testing for HIV in high- and
2260 low-risk populations: a systematic review. *PLoS Med*, 10:e1001414. doi: 10.1371/
2261 journal.pmed.1001414
- 2262 [25] C Hart, G Schochetman, T Spira, A Lifson, J Moore, J Galphin, J Sninsky and CY Ou.
2263 1988. Direct detection of HIV RNA expression in seropositive subjects. *Lancet*, 2:
2264 596–599. doi: 10.1016/S0140-6736(88)90639-3
- 2265 [26] CY Ou, S Kwok, SW Mitchell, DH Mack, JJ Sninsky, JW Krebs, P Feorino, D Warfield
2266 and G Schochetman. 1988. DNA amplification for direct detection of HIV-1 in DNA of
2267 peripheral blood mononuclear cells. *Science*, 239:295–297. doi: 10.1126/science.3336784
- 2268 [27] EF Long. 2011. HIV screening via fourth-generation immunoassay or nucleic acid
2269 amplification test in the United States: a cost-effectiveness analysis. *PLoS One*, 6:
2270 e27625. doi: 10.1371/journal.pone.0027625
- 2271 [28] SA Fiscus, B Cheng, SM Crowe, L Demeter, C Jennings, V Miller, R Respass,
2272 W Stevens and FfCHIVRAVLAWG . 2006. HIV-1 viral load assays for resource-
2273 limited settings. *PLoS Med*, 3:e417. doi: 10.1371/journal.pmed.0030417
- 2274 [29] S Wang, F Xu and U Demirci. 2010. Advances in developing HIV-1 viral load assays
2275 for resource-limited settings. *Biotechnol Adv*, 28:770–781. doi: 10.1016/j.biotechadv.
2276 2010.06.004
- 2277 [30] P Mee, KL Fielding, S Charalambous, GJ Churchyard and AD Grant. 2008. Evaluation
2278 of the WHO criteria for antiretroviral treatment failure among adults in South Africa.
2279 *AIDS*, 22:1971–1977. doi: 10.1097/QAD.0b013e32830e4cd8
- 2280 [31] JJG van Oosterhout, L Brown, R Weigel, JJ Kumwenda, D Mzinganjira, N Saukila,
2281 B Mhango, T Hartung, S Phiri and MC Hosseinipour. 2009. Diagnosis of antiretroviral
2282 therapy failure in Malawi: poor performance of clinical and immunological WHO
2283 criteria. *Trop Med Int Health*, 14:856–861. doi: 10.1111/j.1365-3156.2009.02309.x
- 2284 [32] MC Hosseinipour, JJG van Oosterhout, R Weigel, S Phiri, D Kamwendo, N Parkin,
2285 SA Fiscus, JAE Nelson, JJ Eron and J Kumwenda. 2009. The public health approach
2286 to identify antiretroviral therapy failure: high-level nucleoside reverse transcriptase
2287 inhibitor resistance among Malawians failing first-line antiretroviral therapy. *AIDS*,
2288 23:1127–1134. doi: 10.1097/QAD.0b013e32832ac34e
- 2289 [33] S Sherrill-Mix, MK Lewinski, M Famiglietti, A Bosque, N Malani, KE Ocwieja,
2290 CC Berry, D Looney, L Shan et al. 2013. HIV latency and integration site placement
2291 in five cell-based models. *Retrovirology*, 10:90. doi: 10.1186/1742-4690-10-90

- 2292 [34] TW Chun, D Finzi, J Margolick, K Chadwick, D Schwartz and RF Siliciano. 1995. In
2293 vivo fate of HIV-1-infected T cells: quantitative analysis of the transition to stable
2294 latency. *Nat Med*, 1:1284–1290
- 2295 [35] TW Chun, L Carruth, D Finzi, X Shen, JA DiGiuseppe, H Taylor, M Hermankova,
2296 K Chadwick, J Margolick et al. 1997. Quantification of latent tissue reservoirs and
2297 total body viral load in HIV-1 infection. *Nature*, 387:183–188. doi: 10.1038/387183a0
- 2298 [36] RT Davey, N Bhat, C Yoder, TW Chun, JA Metcalf, R Dewar, V Natarajan, RA Lem-
2299 picki, JW Adelsberger et al. 1999. HIV-1 and T cell dynamics after interruption of
2300 highly active antiretroviral therapy (HAART) in patients with a history of sustained
2301 viral suppression. *Proc Natl Acad Sci U S A*, 96:15109–15114
- 2302 [37] DD Richman, DM Margolis, M Delaney, WC Greene, D Hazuda and RJ Pomerantz.
2303 2009. The challenge of finding a cure for HIV infection. *Science*, 323:1304–1307. doi:
2304 10.1126/science.1165706
- 2305 [38] D Finzi, J Blankson, JD Siliciano, JB Margolick, K Chadwick, T Pierson, K Smith,
2306 J Lisziewicz, F Lori et al. 1999. Latent infection of CD4+ T cells provides a mechanism
2307 for lifelong persistence of HIV-1, even in patients on effective combination therapy.
2308 *Nat Med*, 5:512–517. doi: 10.1038/8394
- 2309 [39] JD Siliciano, J Kajdas, D Finzi, TC Quinn, K Chadwick, JB Margolick, C Kovacs,
2310 SJ Gange and RF Siliciano. 2003. Long-term follow-up studies confirm the stability
2311 of the latent reservoir for HIV-1 in resting CD4+ T cells. *Nat Med*, 9:727–728. doi:
2312 10.1038/nm880
- 2313 [40] D Finzi, M Hermankova, T Pierson, LM Carruth, C Buck, RE Chaisson, TC Quinn,
2314 K Chadwick, J Margolick et al. 1997. Identification of a reservoir for HIV-1 in patients
2315 on highly active antiretroviral therapy. *Science*, 278:1295–1300. doi: 10.1126/science.
2316 278.5341.1295
- 2317 [41] LS Weinberger, RD Dar and ML Simpson. 2008. Transient-mediated fate determination
2318 in a transcriptional circuit of HIV. *Nat Genet*, 40:466–470. doi: 10.1038/ng.116
- 2319 [42] A Singh, B Razooky, CD Cox, ML Simpson and LS Weinberger. 2010. Transcriptional
2320 bursting from the HIV-1 promoter is a significant source of stochastic noise in HIV-1
2321 gene expression. *Biophys J*, 98:L32–L34. doi: 10.1016/j.bpj.2010.03.001
- 2322 [43] BS Razooky and LS Weinberger. 2011. Mapping the architecture of the HIV-1 Tat
2323 circuit: A decision-making circuit that lacks bistability and exploits stochastic noise.
2324 *Methods*, 53:68–77. doi: 10.1016/j.ymeth.2010.12.006
- 2325 [44] HJ Muller. 1930. Types of visible variations induced by X-rays in *Drosophila*. *J Genet*,
2326 22:299–334

- 2327 [45] M Gaszner and G Felsenfeld. 2006. Insulators: exploiting transcriptional and epigenetic
2328 mechanisms. *Nat Rev Genet*, 7:703–713. doi: 10.1038/nrg1925
- 2329 [46] A Jordan, P Defechereux and E Verdin. 2001. The site of HIV-1 integration in
2330 the human genome determines basal transcriptional activity and response to Tat
2331 transactivation. *EMBO J*, 20:1726–1738. doi: 10.1093/emboj/20.7.1726
- 2332 [47] A Jordan, D Bisgrove and E Verdin. 2003. HIV reproducibly establishes a latent
2333 infection after acute infection of T cells in vitro. *EMBO J*, 22:1868–1877. doi:
2334 10.1093/emboj/cdg188
- 2335 [48] R Pearson, YK Kim, J Hokello, K Lassen, J Friedman, M Tyagi and J Karn. 2008. Epigenetic
2336 silencing of human immunodeficiency virus (HIV) transcription by formation of
2337 restrictive chromatin structures at the viral long terminal repeat drives the progressive
2338 entry of HIV into latency. *J Virol*, 82:12291–12303. doi: 10.1128/JVI.01383-08
- 2339 [49] F Romerio, MN Gabriel and DM Margolis. 1997. Repression of human immunodeficiency
2340 virus type 1 through the novel cooperation of human factors YY1 and LSF. *J
2341 Virol*, 71:9375–9382
- 2342 [50] JJ Coull, F Romerio, JM Sun, JL Volker, KM Galvin, JR Davie, Y Shi, U Hansen
2343 and DM Margolis. 2000. The human factors YY1 and LSF repress the human immunodeficiency
2344 virus type 1 long terminal repeat via recruitment of histone deacetylase 1.
2345 *J Virol*, 74:6790–6799. doi: 10.1128/JVI.74.15.6790-6799.2000
- 2346 [51] G He and DM Margolis. 2002. Counterregulation of chromatin deacetylation and
2347 histone deacetylase occupancy at the integrated promoter of human immunodeficiency
2348 virus type 1 (HIV-1) by the HIV-1 repressor YY1 and HIV-1 activator Tat. *Mol Cell
2349 Biol*, 22:2965–2973. doi: 10.1128/MCB.22.9.2965-2973.2002
- 2350 [52] MK Lewinski, D Bisgrove, P Shinn, H Chen, C Hoffmann, S Hannenhalli, E Verdin,
2351 CC Berry, JR Ecker and FD Bushman. 2005. Genome-wide analysis of chromosomal
2352 features repressing human immunodeficiency virus transcription. *J Virol*, 79:6610–6619.
2353 doi: 10.1128/JVI.79.11.6610-6619.2005
- 2354 [53] L Shan, HC Yang, SA Rabi, HC Bravo, NS Shroff, RA Irizarry, H Zhang, JB Margolick,
2355 JD Siliciano and RF Siliciano. 2011. Influence of host gene transcription level and
2356 orientation on HIV-1 latency in a primary-cell model. *J Virol*, 85:5384–5393. doi:
2357 10.1128/JVI.02536-10
- 2358 [54] MJ Pace, EH Graf, LM Agosto, AM Mexas, F Male, T Brady, FD Bushman and
2359 U O'Doherty. 2012. Directly infected resting CD4+ T cells can produce HIV Gag
2360 without spreading infection in a model of HIV latency. *PLoS Pathog*, 8:e1002818. doi:
2361 10.1371/journal.ppat.1002818
- 2362 [55] T Lenasi, X Contreras and BM Peterlin. 2008. Transcriptional interference antagonizes

- 2363 proviral gene expression to promote HIV latency. *Cell Host Microbe*, 4:123–133. doi:
2364 10.1016/j.chom.2008.05.016
- 2365 [56] Y Han, YB Lin, W An, J Xu, HC Yang, K O'Connell, D Dordai, JD Boeke, JD Silicano
2366 and RF Siliciano. 2008. Orientation-dependent regulation of integrated HIV-1
2367 expression by host gene transcriptional readthrough. *Cell Host Microbe*, 4:134–146.
2368 doi: 10.1016/j.chom.2008.06.008
- 2369 [57] L Shan, K Deng, NS Shroff, CM Durand, SA Rabi, HC Yang, H Zhang, JB Margolick,
2370 JN Blankson and RF Siliciano. 2012. Stimulation of HIV-1-specific cytolytic T
2371 lymphocytes facilitates elimination of latent viral reservoir after virus reactivation.
2372 *Immunity*, 36:491–501. doi: 10.1016/j.jimmuni.2012.01.014
- 2373 [58] D Boehm, V Calvanese, RD Dar, S Xing, S Schroeder, L Martins, K Aull, PC Li,
2374 V Planelles et al. 2013. BET bromodomain-targeting compounds reactivate HIV from
2375 latency via a Tat-independent mechanism. *Cell Cycle*, 12:452–462. doi: 10.4161/cc.
2376 23309
- 2377 [59] A Savarino, A Mai, S Norelli, SE Daker, S Valente, D Rotili, L Altucci, AT Palamara
2378 and E Garaci. 2009. “Shock and kill” effects of class I-selective histone
2379 deacetylase inhibitors in combination with the glutathione synthesis inhibitor buthionine
2380 sulfoximine in cell line models for HIV-1 quiescence. *Retrovirology*, 6:52. doi:
2381 10.1186/1742-4690-6-52
- 2382 [60] NM Archin, AL Liberty, AD Kashuba, SK Choudhary, JD Kuruc, AM Crooks,
2383 DC Parker, EM Anderson, MF Kearney et al. 2012. Administration of vorinostat
2384 disrupts HIV-1 latency in patients on antiretroviral therapy. *Nature*, 487:482–485. doi:
2385 10.1038/nature11286
- 2386 [61] A Bosque and V Planelles. 2009. Induction of HIV-1 latency and reactivation in
2387 primary memory CD4+ T cells. *Blood*, 113:58–65. doi: 10.1182/blood-2008-07-168393
- 2388 [62] A Bosque and V Planelles. 2011. Studies of HIV-1 latency in an ex vivo model that uses
2389 primary central memory T cells. *Methods*, 53:54–61. doi: 10.1016/j.ymeth.2010.10.002
- 2390 [63] X Wu, Y Li, B Crise and SM Burgess. 2003. Transcription start regions in the
2391 human genome are favored targets for MLV integration. *Science*, 300:1749–1751. doi:
2392 10.1126/science.1083413
- 2393 [64] RS Mitchell, BF Beitzel, ARW Schroder, P Shinn, H Chen, CC Berry, JR Ecker and
2394 FD Bushman. 2004. Retroviral DNA integration: ASLV, HIV, and MLV show distinct
2395 target site preferences. *PLoS Biol*, 2:e234. doi: 10.1371/journal.pbio.0020234
- 2396 [65] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation
2397 for Statistical Computing, Vienna, Austria, 2012
- 2398 [66] C Berry, S Hannenhalli, J Leipzig and FD Bushman. 2006. Selection of target sites

- 2399 for mobile DNA integration in the human genome. *PLoS Comput Biol*, 2:e157. doi:
2400 10.1371/journal.pcbi.0020157
- 2401 [67] GP Wang, A Ciuffi, J Leipzig, CC Berry and FD Bushman. 2007. HIV integration
2402 site selection: analysis by massively parallel pyrosequencing reveals association with
2403 epigenetic modifications. *Genome Res*, 17:1186–1194. doi: 10.1101/gr.6286907
- 2404 [68] H Mochizuki, JP Schwartz, K Tanaka, RO Brady and J Reiser. 1998. High-titer
2405 human immunodeficiency virus type 1-based vector systems for gene delivery into
2406 nondividing cells. *J Virol*, 72:8873–8883
- 2407 [69] Y Han, K Lassen, D Monie, AR Sedaghat, S Shimoji, X Liu, TC Pierson, JB Margolick,
2408 RF Siliciano and JD Siliciano. 2004. Resting CD4+ T cells from human
2409 immunodeficiency virus type 1 (HIV-1)-infected individuals carry integrated HIV-
2410 1 genomes within actively transcribed host genes. *J Virol*, 78:6122–6133. doi:
2411 10.1128/JVI.78.12.6122-6133.2004
- 2412 [70] G Plesa, J Dai, C Baytop, JL Riley, CH June and U O'Doherty. 2007. Addition of
2413 deoxynucleosides enhances human immunodeficiency virus type 1 integration and 2LTR
2414 formation in resting CD4+ T cells. *J Virol*, 81:13938–13942. doi: 10.1128/JVI.01745-07
- 2415 [71] N Malani. hiReadsProcessor R package. URL <http://github.com/malnirav/hিReadsProcessor>
- 2416
- 2417 [72] WJ Kent. 2002. BLAT—the BLAST-like alignment tool. *Genome Res*, 12:656–664. doi:
2418 10.1101/gr.229202
- 2419 [73] CC Berry, K Ocwieja, N Malani and FD Bushman. 2014. Comparing DNA in-
2420 tegration site clusters with Scan Statistics. *Bioinformatics*, 30:1493–1500. doi:
2421 10.1093/bioinformatics/btu035
- 2422 [74] C Trapnell, BA Williams, G Pertea, A Mortazavi, G Kwan, MJ van Baren, SL Salzberg,
2423 BJ Wold and L Pachter. 2010. Transcript assembly and quantification by RNA-Seq
2424 reveals unannotated transcripts and isoform switching during cell differentiation. *Nat
Biotechnol*, 28:511–515. doi: 10.1038/nbt.1621
- 2426 [75] J Ernst and M Kellis. 2010. Discovery and characterization of chromatin states
2427 for systematic annotation of the human genome. *Nat Biotechnol*, 28:817–825. doi:
2428 10.1038/nbt.1662
- 2429 [76] AS Hinrichs, D Karolchik, R Baertsch, GP Barber, G Bejerano, H Clawson, M Diekhans,
2430 TS Furey, RA Harte et al. 2006. The UCSC Genome Browser Database: update 2006.
2431 *Nucleic Acids Res*, 34:D590–D598. doi: 10.1093/nar/gkj144
- 2432 [77] KR Rosenbloom, CA Sloan, VS Malladi, TR Dreszer, K Learned, VM Kirkup,
2433 MC Wong, M Maddren, R Fang et al. 2013. ENCODE data in the UCSC Genome
2434 Browser: year 5 update. *Nucleic Acids Res*, 41:D56–D63. doi: 10.1093/nar/gks1172

- 2435 [78] J Han, SG Park, JB Bae, J Choi, JM Lyu, SH Park, HS Kim, YJ Kim, S Kim and
2436 TY Kim. 2012. The characteristics of genome-wide DNA methylation in naïve CD4+
2437 T cells of patients with psoriasis or atopic dermatitis. *Biochem Biophys Res Commun*,
2438 422:157–163. doi: 10.1016/j.bbrc.2012.04.128
- 2439 [79] LR Meyer, AS Zweig, AS Hinrichs, D Karolchik, RM Kuhn, M Wong, CA Sloan,
2440 KR Rosenbloom, G Roe et al. 2013. The UCSC Genome Browser database: extensions
2441 and updates 2013. *Nucleic Acids Res*, 41:D64–D69. doi: 10.1093/nar/gks1048
- 2442 [80] Z Wang, C Zang, JA Rosenfeld, DE Schones, A Barski, S Cuddapah, K Cui, TY Roh,
2443 W Peng et al. 2008. Combinatorial patterns of histone acetylations and methylations
2444 in the human genome. *Nat Genet*, 40:897–903. doi: 10.1038/ng.154
- 2445 [81] A Barski, S Cuddapah, K Cui, TY Roh, DE Schones, Z Wang, G Wei, I Chepelev and
2446 K Zhao. 2007. High-resolution profiling of histone methylations in the human genome.
2447 *Cell*, 129:823–837. doi: 10.1016/j.cell.2007.05.009
- 2448 [82] Z Wang, C Zang, K Cui, DE Schones, A Barski, W Peng and K Zhao. 2009. Genome-
2449 wide mapping of HATs and HDACs reveals distinct functions in active and inactive
2450 genes. *Cell*, 138:1019–1031. doi: 10.1016/j.cell.2009.06.049
- 2451 [83] DE Schones, K Cui, S Cuddapah, TY Roh, A Barski, Z Wang, G Wei and K Zhao.
2452 2008. Dynamic regulation of nucleosome positioning in the human genome. *Cell*, 132:
2453 887–898. doi: 10.1016/j.cell.2008.02.022
- 2454 [84] F Hsu, WJ Kent, H Clawson, RM Kuhn, M Diekhans and D Haussler. 2006. The UCSC
2455 Known Genes. *Bioinformatics*, 22:1036–1046. doi: 10.1093/bioinformatics/btl048
- 2456 [85] J Friedman, T Hastie and R Tibshirani. 2010. Regularization paths for generalized
2457 linear models via coordinate descent. *J Stat Softw*, 33:1–22
- 2458 [86] IH Greger, F Demarchi, M Giacca and NJ Proudfoot. 1998. Transcriptional interference
2459 perturbs the binding of Sp1 to the HIV-1 promoter. *Nucleic Acids Res*, 26:1294–1301
- 2460 [87] A De Marco, C Biancotto, A Knezevich, P Maiuri, C Vardabasso and A Marcello.
2461 2008. Intragenic transcriptional cis-activation of the human immunodeficiency virus 1
2462 does not result in allele-specific inhibition of the endogenous gene. *Retrovirology*, 5:98.
2463 doi: 10.1186/1742-4690-5-98
- 2464 [88] ST Chang, P Sova, X Peng, J Weiss, GL Law, RE Palermo and MG Katze. 2011.
2465 Next-generation sequencing reveals HIV-1-mediated suppression of T cell activation
2466 and RNA processing and regulation of noncoding RNA expression in a CD4+ T cell
2467 line. *MBio*, 2. doi: 10.1128/mBio.00134-11
- 2468 [89] JS Waye and HF Willard. 1987. Nucleotide sequence heterogeneity of alpha satellite
2469 repetitive DNA: a survey of alphoid sequences from different human chromosomes.
2470 *Nucleic Acids Res*, 15:7549–7569. doi: 10.1093/nar/15.18.7549

- 2471 [90] J Jurka, VV Kapitonov, A Pavlicek, P Klonowski, O Kohany and J Walichiewicz. 2005.
2472 Repbase Update, a database of eukaryotic repetitive elements. *Cytogenet Genome Res*,
2473 110:462–467. doi: 10.1159/000084979
- 2474 [91] E Verdin, P Paras and C Van Lint. 1993. Chromatin disruption in the promoter of
2475 human immunodeficiency virus type 1 during transcriptional activation. *EMBO J*, 12:
2476 3249–3259
- 2477 [92] C Van Lint, S Emiliani, M Ott and E Verdin. 1996. Transcriptional activation and
2478 chromatin remodeling of the HIV-1 promoter in response to histone acetylation. *EMBO J*, 15:1112–1120
- 2480 [93] KG Lassen, KX Ramyar, JR Bailey, Y Zhou and RF Siliciano. 2006. Nuclear retention
2481 of multiply spliced HIV-1 RNA in resting CD4+ T cells. *PLoS Pathog*, 2:e68. doi:
2482 10.1371/journal.ppat.0020068
- 2483 [94] M Dieudonné, P Maiuri, C Biancotto, A Knezevich, A Kula, M Lusic and A Marcello.
2484 2009. Transcriptional competence of the integrated HIV-1 provirus at the nuclear
2485 periphery. *EMBO J*, 28:2231–2243. doi: 10.1038/emboj.2009.141
- 2486 [95] RF Siliciano and WC Greene. 2011. HIV Latency. *Cold Spring Harb Perspect Med*, 1:
2487 a007096. doi: 10.1101/cshperspect.a007096
- 2488 [96] M Lusic, B Marini, H Ali, B Lucic, R Luzzati and M Giacca. 2013. Proximity to
2489 PML nuclear bodies regulates HIV-1 latency in CD4+ T cells. *Cell Host Microbe*, 13:
2490 665–677. doi: 10.1016/j.chom.2013.05.006
- 2491 [97] LM Mansky and HM Temin. 1995. Lower in vivo mutation rate of human immunodefici-
2492 ency virus type 1 than that predicted from the fidelity of purified reverse transcriptase.
2493 *J Virol*, 69:5087–5094
- 2494 [98] KE Ocwieja, S Sherrill-Mix, R Mukherjee, R Custers-Allen, P David, M Brown, S Wang,
2495 DR Link, J Olson et al. 2012. Dynamic regulation of HIV-1 mRNA populations
2496 analyzed by single-molecule enrichment and long-read sequencing. *Nucleic Acids Res*,
2497 40:10345–10355. doi: 10.1093/nar/gks753
- 2498 [99] Q Pan, O Shai, LJ Lee, BJ Frey and BJ Blencowe. 2008. Deep surveying of alternative
2499 splicing complexity in the human transcriptome by high-throughput sequencing. *Nature
2500 Genetics*, 40:1413–1415. doi: 10.1038/ng.259
- 2501 [100] ET Wang, R Sandberg, S Luo, I Khrebtukova, L Zhang, C Mayr, SF Kingsmore,
2502 GP Schroth and CB Burge. 2008. Alternative isoform regulation in human tissue
2503 transcriptomes. *Nature*, 456:470–476. doi: 10.1038/nature07509
- 2504 [101] F Pagani, M Raponi and FE Baralle. 2005. Synonymous mutations in CFTR exon
2505 12 affect splicing and are not neutral in evolution. *Proc Natl Acad Sci U S A*, 102:
2506 6368–6372. doi: 10.1073/pnas.0502288102

- 2507 [102] C Wang, Y Mitsuya, B Gharizadeh, M Ronaghi and SR W. 2007. Characterization of
2508 mutation spectra with ultra-deep pyrosequencing: Application to HIV-1 drug resistance.
2509 *Genome Research*, 17:1195–1201. doi: 10.1101/gr.6468307
- 2510 [103] K Wang, R Wernersson and S Brunak. 2011. The strength of intron donor splice sites
2511 in human genes displays a bell-shaped pattern. *Bioinformatics*, 27:3079–3084. doi:
2512 10.1093/bioinformatics/btr532
- 2513 [104] Y Barash, JA Calarco, W Gao, Q Pan, X Wang, O Shai, BJ Blencowe and BJ Frey.
2514 2010. Deciphering the splicing code. *Nature*, 465:53–59. doi: 10.1038/nature09000
- 2515 [105] DF Purcell and MA Martin. 1993. Alternative splicing of human immunodeficiency
2516 virus type 1 mRNA modulates viral protein expression, replication, and infectivity. *J
2517 Virol*, 67:6365–6378
- 2518 [106] DM Benko, S Schwartz, GN Pavlakis and BK Felber. 1990. A novel human immunod-
2519 eficiency virus type 1 protein, tev, shares sequences with tat, env, and rev proteins. *J
2520 Virol*, 64:2505–2518
- 2521 [107] C Carrera, M Pinilla, L Pérez-Alvarez and MM Thomson. 2010. Identification of
2522 unusual and novel HIV type 1 spliced transcripts generated in vivo. *AIDS Res Hum
2523 Retroviruses*, 26:815–820. doi: 10.1089/aid.2010.0011
- 2524 [108] M Lützelberger, LS Reinert, AT Das, B Berkhout and J Kjems. 2006. A novel splice
2525 donor site in the gag-pol gene is required for HIV-1 RNA stability. *J Biol Chem*, 281:
2526 18644–18651. doi: 10.1074/jbc.M513698200
- 2527 [109] J Salfeld, HG Gttlinger, RA Sia, RE Park, JG Sodroski and WA Haseltine. 1990. A
2528 tripartite HIV-1 tat-env-rev fusion protein. *EMBO J*, 9:965–970
- 2529 [110] S Schwartz, BK Felber, DM Benko, EM Fenyö and GN Pavlakis. 1990. Cloning and
2530 functional analysis of multiply spliced mRNA species of human immunodeficiency
2531 virus type 1. *J Virol*, 64:2519–2529
- 2532 [111] J Smith, A Azad and N Deacon. 1992. Identification of two novel human immunodefi-
2533 ciency virus type 1 splice acceptor sites in infected T cell lines. *J Gen Virol*, 73 (Pt
2534 7):1825–1828
- 2535 [112] CM Stoltzfus. 2009. Chapter 1. Regulation of HIV-1 alternative RNA splicing and its
2536 role in virus replication. *Adv Virus Res*, 74:1–40. doi: 10.1016/S0065-3527(09)74001-1
- 2537 [113] N Bakkour, YL Lin, S Maire, L Ayadi, F Mahuteau-Betzer, CH Nguyen, C Mettling,
2538 P Portales, D Grierson et al. 2007. Small-molecule inhibition of HIV pre-mRNA
2539 splicing as a novel antiretroviral therapy to overcome drug resistance. *PLoS Pathog*, 3:
2540 1530–1539. doi: 10.1371/journal.ppat.0030159
- 2541 [114] AL Brass, DM Dykxhoorn, Y Benita, N Yan, A Engelman, RJ Xavier, J Lieberman

- 2542 and SJ Elledge. 2008. Identification of host proteins required for HIV infection through
2543 a functional genomic screen. *Science*, 319:921–926. doi: 10.1126/science.1152725
- 2544 [115] JA Jablonski and M Caputi. 2009. Role of cellular RNA processing factors in human
2545 immunodeficiency virus type 1 mRNA metabolism, replication, and infectivity. *J Virol*,
2546 83:981–992. doi: 10.1128/JVI.01801-08
- 2547 [116] R König, Y Zhou, D Elleder, TL Diamond, GMC Bonamy, JT Irelan, CY Chiang,
2548 BP Tu, PDD Jesus et al. 2008. Global analysis of host-pathogen interactions that
2549 regulate early-stage HIV-1 replication. *Cell*, 135:49–60. doi: 10.1016/j.cell.2008.07.032
- 2550 [117] A Tranell, S Tingsborg, EM Feny and S Schwartz. 2011. Inhibition of splicing by
2551 serine-arginine rich protein 55 (SRp55) causes the appearance of partially spliced HIV-1
2552 mRNAs in the cytoplasm. *Virus Res*, 157:82–91. doi: 10.1016/j.virusres.2011.02.010
- 2553 [118] H Zhou, M Xu, Q Huang, AT Gates, XD Zhang, JC Castle, E Stec, M Ferrer,
2554 B Strulovici et al. 2008. Genome-scale RNAi screen for host factors required for HIV
2555 replication. *Cell Host Microbe*, 4:495–504. doi: 10.1016/j.chom.2008.10.004
- 2556 [119] Y Zhu, G Chen, F Lv, X Wang, X Ji, Y Xu, J Sun, L Wu, YT Zheng and G Gao. 2011.
2557 Zinc-finger antiviral protein inhibits HIV-1 infection by selectively targeting multiply
2558 spliced viral mRNAs for degradation. *Proc Natl Acad Sci U S A*, 108:15834–15839.
2559 doi: 10.1073/pnas.1101676108
- 2560 [120] MJ Saltarelli, E Hadziyannis, CE Hart, JV Harrison, BK Felber, TJ Spira and
2561 GN Pavlakis. 1996. Analysis of human immunodeficiency virus type 1 mRNA splicing
2562 patterns during disease progression in peripheral blood mononuclear cells from infected
2563 individuals. *AIDS Res Hum Retroviruses*, 12:1443–1456. doi: 10.1089/aid.1996.12.1443
- 2564 [121] E Delgado, C Carrera, P Nebreda, A Fernández-García, M Pinilla, V García, L Prez-
2565 Ivarez and MM Thomson. 2012. Identification of new splice sites used for generation
2566 of rev transcripts in human immunodeficiency virus type 1 subtype C primary isolates.
2567 *PLoS One*, 7:e30574. doi: 10.1371/journal.pone.0030574
- 2568 [122] P Grabowski. 2011. Alternative splicing takes shape during neuronal development.
2569 *Curr Opin Genet Dev*, 21:388–394. doi: 10.1016/j.gde.2011.03.005
- 2570 [123] M Llorian and CWJ Smith. 2011. Decoding muscle alternative splicing. *Curr Opin*
2571 *Genet Dev*, 21:380–387. doi: 10.1016/j.gde.2011.03.006
- 2572 [124] JY Ip, A Tong, Q Pan, JD Topp, BJ Blencowe and KW Lynch. 2007. Global analysis of
2573 alternative splicing during T-cell activation. *RNA*, 13:563–572. doi: 10.1261/rna.457207
- 2574 [125] JD Topp, J Jackson, AA Melton and KW Lynch. 2008. A cell-based screen for splicing
2575 regulators identifies hnRNP LL as a distinct signal-induced repressor of CD45 variable
2576 exon 4. *RNA*, 14:2038–2049. doi: 10.1261/rna.1212008

- 2577 [126] S Sonza, HP Mutimer, K O'Brien, P Ellery, JL Howard, JH Axelrod, NJ Deacon,
2578 SM Crowe and DFJ Purcell. 2002. Selectively reduced tat mRNA heralds the decline
2579 in productive human immunodeficiency virus type 1 infection in monocyte-derived
2580 macrophages. *J Virol*, 76:12611–12621
- 2581 [127] D Dowling, S Nasr-Esfahani, CH Tan, K O'Brien, JL Howard, DA Jans, DF j Purcell,
2582 CM Stoltzfus and S Sonza. 2008. HIV-1 infection induces changes in expression of
2583 cellular splicing factors that regulate alternative viral splicing and virus production in
2584 macrophages. *Retrovirology*, 5:18. doi: 10.1186/1742-4690-5-18
- 2585 [128] J Hull, S Campino, K Rowlands, MS Chan, RR Copley, MS Taylor, K Rockett,
2586 G Elvidge, B Keating et al. 2007. Identification of common genetic variation that
2587 modulates alternative splicing. *PLoS Genet*, 3:e99. doi: 10.1371/journal.pgen.0030099
- 2588 [129] T Kwan, D Benovoy, C Dias, S Gurd, D Serre, H Zuzan, TA Clark, A Schweitzer,
2589 MK Staples et al. 2007. Heritability of alternative splicing in the human genome.
2590 *Genome Res*, 17:1210–1218. doi: 10.1101/gr.6281007
- 2591 [130] R Collman, JW Balliet, SA Gregory, H Friedman, DL Kolson, N Nathanson and
2592 A Srinivasan. 1992. An infectious molecular clone of an unusual macrophage-tropic
2593 and highly cytopathic strain of human immunodeficiency virus type 1. *J Virol*, 66:
2594 7517–7521
- 2595 [131] J Eid, A Fehr, J Gray, K Luong, J Lyle, G Otto, P Peluso, D Rank, P Baybayan
2596 et al. 2009. Real-time DNA sequencing from single polymerase molecules. *Science*,
2597 323:133–138. doi: 10.1126/science.1162986
- 2598 [132] H Deng, R Liu, W Ellmeier, S Choe, D Unutmaz, M Burkhardt, P Di Marzio, S Marmon,
2599 RE Sutton et al. 1996. Identification of a major co-receptor for primary isolates of
2600 HIV-1. *Nature*, 381:661–666. doi: 10.1038/381661a0
- 2601 [133] NR Landau and DR Littman. 1992. Packaging system for rapid production of murine
2602 leukemia virus vectors with variable tropism. *J Virol*, 66:5110–5113
- 2603 [134] X Wei, JM Decker, S Wang, H Hui, JC Kappes, X Wu, JF Salazar-Gonzalez,
2604 MG Salazar, JM Kilby et al. 2003. Antibody neutralization and escape by HIV-
2605 1. *Nature*, 422:307–312. doi: 10.1038/nature01470
- 2606 [135] N Srinivasakumar, N Chazal, C Helga-Maria, S Prasad, ML Hammarskjöld and
2607 D Rekosh. 1997. The effect of viral regulatory protein expression on gene delivery by
2608 human immunodeficiency virus type 1 vectors produced in stable packaging cell lines.
2609 *J Virol*, 71:5841–5848
- 2610 [136] DC Shugars, MS Smith, DH Glueck, PV Nantermet, F Seillier-Moiseiwitsch and
2611 R Swanstrom. 1993. Analysis of human immunodeficiency virus type 1 nef gene
2612 sequences present in vivo. *J Virol*, 67:4639–4650

- 2613 [137] R Tewhey, JB Warner, M Nakano, B Libby, M Medkova, PH David, SK Kotsopoulos,
2614 ML Samuels, JB Hutchison et al. 2009. Microdroplet-based PCR enrichment for
2615 large-scale targeted sequencing. *Nat Biotechnol*, 27:1025–1031. doi: 10.1038/nbt.1583
- 2616 [138] KJ Travers, CS Chin, DR Rank, JS Eid and SW Turner. 2010. A flexible and efficient
2617 template format for circular consensus sequencing and SNP detection. *Nucleic Acids
2618 Res*, 38:e159. doi: 10.1093/nar/gkq543
- 2619 [139] TA Thanaraj and F Clark. 2001. Human GC-AG alternative intron isoforms with
2620 weak donor sites show enhanced consensus at acceptor exon positions. *Nucleic Acids
2621 Res*, 29:2581–2593. doi: 10.1093/nar/29.12.2581
- 2622 [140] M Aebi, H Hornig, RA Padgett, J Reiser and C Weissmann. 1986. Sequence require-
2623 ments for splicing of higher eukaryotic nuclear pre-mRNA. *Cell*, 47:555–565
- 2624 [141] M Burset, IA Seledtsov and VV Solovyev. 2000. Analysis of canonical and non-
2625 canonical splice sites in mammalian genomes. *Nucleic Acids Res*, 28:4364–4375. doi:
2626 10.1093/nar/28.21.4364
- 2627 [142] M Burset, IA Seledtsov and VV Solovyev. 2001. SpliceDB: database of canonical
2628 and non-canonical mammalian splice sites. *Nucleic Acids Res*, 29:255–259. doi:
2629 10.1093/nar/29.1.255
- 2630 [143] N Sheth, X Roca, ML Hastings, T Roeder, AR Krainer and R Sachidanandam. 2006.
2631 Comprehensive splice-site analysis using comparative genomics. *Nucleic Acids Res*, 34:
2632 3955–3967. doi: 10.1093/nar/gkl556
- 2633 [144] JC Guatelli, TR Gingeras and DD Richman. 1990. Alternative splice acceptor
2634 utilization during human immunodeficiency virus type 1 infection of cultured cells. *J
2635 Virol*, 64:4093–4098
- 2636 [145] C Kuiken, B Foley, T Leitner, C Apetrei, B Hahn, I Mizrahi, J Mullins, A Rambaut,
2637 S Wolinsky and B Korber. 2010. HIV Sequence Compendium 2010. Theoretical Biology
2638 and Biophysics Group, Los Alamos National Laboratory, New Mexico. URL <http://www.hiv.lanl.gov/content/sequence/HIV/COMPENDIUM/2010compendium.html>
- 2640 [146] C Burge, T Tuschl and P Sharp. 1999. Splicing of precursors to mRNAs by the
2641 spliceosomes. *Cold Spring Harbor Monograph Archive*, 37. doi: 10.1101/087969589.37.
2642 525
- 2643 [147] TEM Abbink and B Berkhout. 2008. RNA structure modulates splicing efficiency at
2644 the human immunodeficiency virus type 1 major splice donor. *J Virol*, 82:3090–3098.
2645 doi: 10.1128/JVI.01479-07
- 2646 [148] K Verhoef, PS Bilodeau, JL van Wamel, J Kjems, CM Stoltzfus and B Berkhout. 2001.
2647 Repair of a Rev-minus human immunodeficiency virus type 1 mutant by activation of
2648 a cryptic splice site. *J Virol*, 75:3495–3500. doi: 10.1128/JVI.75.7.3495-3500.2001

- 2649 [149] M Caputi, M Freund, S Kammler, C Asang and H Schaal. 2004. A bidirectional
2650 SF2/ASF- and SRp40-dependent splicing enhancer regulates human immunodeficiency
2651 virus type 1 rev, env, vpu, and nef gene expression. *J Virol*, 78:6517–6526. doi:
2652 10.1128/JVI.78.12.6517-6526.2004
- 2653 [150] AM Zahler, CK Damgaard, J Kjems and M Caputi. 2004. SC35 and heterogeneous
2654 nuclear ribonucleoprotein A/B proteins bind to a juxtaposed exonic splicing
2655 enhancer/exonic splicing silencer element to regulate HIV-1 tat exon 2 splicing. *J Biol
2656 Chem*, 279:10077–10084. doi: 10.1074/jbc.M312743200
- 2657 [151] S Sherrill-Mix, K Ocwieja and F Bushman. Under Review. Gene activity in primary
2658 T cells infected with HIV89.6: intron retention and induction of distinctive genomic
2659 repeats. *Retrovirology*
- 2660 [152] S Wain-Hobson, P Sonigo, O Danos, S Cole and M Alizon. 1985. Nucleotide sequence
2661 of the AIDS virus, LAV. *Cell*, 40:9–17. doi: 10.1016/0092-8674(85)90303-4
- 2662 [153] SK Arya, C Guo, SF Josephs and F Wong-Staal. 1985. Trans-activator gene of human
2663 T-lymphotropic virus type III (HTLV-III). *Science*, 229:69–73
- 2664 [154] J He, S Choe, R Walker, P Di Marzio, DO Morgan and NR Landau. 1995. Human
2665 immunodeficiency virus type 1 viral protein R (Vpr) arrests cells in the G2 phase of
2666 the cell cycle by inhibiting p34cdc2 activity. *J Virol*, 69:6705–6711
- 2667 [155] JB Jowett, V Planelles, B Poon, NP Shah, ML Chen and IS Chen. 1995. The human
2668 immunodeficiency virus type 1 vpr gene arrests infected T cells in the G2 + M phase
2669 of the cell cycle. *J Virol*, 69:6304–6313
- 2670 [156] ME Rogel, LI Wu and M Emerman. 1995. The human immunodeficiency virus type 1
2671 vpr gene prevents cell proliferation during chronic infection. *J Virol*, 69:882–888
- 2672 [157] WC Goh, ME Rogel, CM Kinsey, SF Michael, PN Fultz, MA Nowak, BH Hahn
2673 and M Emerman. 1998. HIV-1 Vpr increases viral expression by manipulation of
2674 the cell cycle: a mechanism for selection of Vpr in vivo. *Nat Med*, 4:65–71. doi:
2675 10.1038/nm0198-065
- 2676 [158] RA Marciniak and PA Sharp. 1991. HIV-1 Tat protein promotes formation of
2677 more-processive elongation complexes. *EMBO J*, 10:4189–4196
- 2678 [159] P Wei, ME Garber, SM Fang, WH Fischer and KA Jones. 1998. A novel CDK9-
2679 associated C-type cyclin interacts directly with HIV-1 Tat and mediates its high-affinity,
2680 loop-specific binding to TAR RNA. *Cell*, 92:451–462. doi: 10.1016/S0092-8674(00)
2681 80939-3
- 2682 [160] S Kanazawa, T Okamoto and BM Peterlin. 2000. Tat competes with CIITA for the
2683 binding to P-TEFb and blocks the expression of MHC class II genes in HIV infection.
2684 *Immunity*, 12:61–70. doi: 10.1016/S1074-7613(00)80159-4

- 2685 [161] M Barboric, JHN Yik, N Czudnochowski, Z Yang, R Chen, X Contreras, M Geyer,
2686 B Matija Peterlin and Q Zhou. 2007. Tat competes with HEXIM1 to increase the
2687 active pool of P-TEFb for HIV-1 transcription. *Nucleic Acids Res*, 35:2003–2012. doi:
2688 10.1093/nar/gkm063
- 2689 [162] SK O'Brien, H Cao, R Nathans, A Ali and TM Rana. 2010. P-TEFb kinase complex
2690 phosphorylates histone H1 to regulate expression of cellular and HIV-1 genes. *J Biol
2691 Chem*, 285:29713–29720. doi: 10.1074/jbc.M110.125997
- 2692 [163] L Muniz, S Egloff, B Ughy, BE Jády and T Kiss. 2010. Controlling cellular P-TEFb
2693 activity by the HIV-1 transcriptional transactivator Tat. *PLoS Pathog*, 6:e1001152.
2694 doi: 10.1371/journal.ppat.1001152
- 2695 [164] J Corbeil, D Sheeter, D Genini, S Rought, L Leoni, P Du, M Ferguson, DR Masys,
2696 JB Welsh et al. 2001. Temporal gene regulation during HIV-1 infection of human
2697 CD4+ T cells. *Genome Res*, 11:1198–1204. doi: 10.1101/gr.180201
- 2698 [165] CH Woelk, F Ottone, CR Plotkin, P Du, CD Royer, SE Rought, J Lozach, R Sasik,
2699 RS Kornbluth et al. 2004. Interferon gene expression following HIV type 1 infection
2700 of monocyte-derived macrophages. *AIDS Res Hum Retroviruses*, 20:1210–1222. doi:
2701 10.1089/0889222042545009
- 2702 [166] MD Hyrcza, C Kovacs, M Loutfy, R Halpenny, L Heisler, S Yang, O Wilkins, M Os-
2703 trowski and SD Der. 2007. Distinct transcriptional profiles in ex vivo CD4+ and CD8+
2704 T cells are established early in human immunodeficiency virus type 1 infection and
2705 are characterized by a chronic interferon response as well as extensive transcriptional
2706 changes in CD8+ T cells. *J Virol*, 81:3477–3486. doi: 10.1128/JVI.01552-06
- 2707 [167] JQ Wu, DE Dwyer, WB Dyer, YH Yang, B Wang and NK Saksena. 2008. Transcrip-
2708 tional profiles in CD8+ T cells from HIV+ progressors on HAART are characterized
2709 by coordinated up-regulation of oxidative phosphorylation enzymes and interferon
2710 responses. *Virology*, 380:124–135. doi: 10.1016/j.virol.2008.06.039
- 2711 [168] AJ Smith, Q Li, SW Wietgrefe, TW Schacker, CS Reilly and AT Haase. 2010. Host
2712 genes associated with HIV-1 replication in lymphatic tissue. *J Immunol*, 185:5417–5424.
2713 doi: 10.4049/jimmunol.1002197
- 2714 [169] M Imbeault, K Giguère, M Ouellet and MJ Tremblay. 2012. Exon level transcriptomic
2715 profiling of HIV-1-infected CD4(+) T cells reveals virus-induced genes and host
2716 environment favorable for viral replication. *PLoS Pathog*, 8:e1002861. doi: 10.1371/
2717 journal.ppat.1002861
- 2718 [170] P Mohammadi, S Desfarges, I Bartha, B Joos, N Zangger, M Muoz, HF Gnethard,
2719 N Beerewinkel, A Telenti and A Ciuffi. 2013. 24 hours in the life of HIV-1 in a T cell
2720 line. *PLoS Pathog*, 9:e1003161. doi: 10.1371/journal.ppat.1003161
- 2721 [171] X Peng, P Sova, RR Green, MJ Thomas, MJ Korth, S Proll, J Xu, Y Cheng, K Yi

- 2722 et al. 2014. Deep sequencing of HIV-infected cells: insights into nascent transcription
2723 and host-directed therapy. *J Virol*, 88:8768–8782. doi: 10.1128/JVI.00768-14
- 2724 [172] R Mitchell, CY Chiang, C Berry and F Bushman. 2003. Global analysis of cellular
2725 transcription following infection with an HIV-based vector. *Mol Ther*, 8:674–687. doi:
2726 10.1016/S1525-0016(03)00215-6
- 2727 [173] C de la Fuente, F Santiago, L Deng, C Eadie, I Zilberman, K Kehn, A Maddukuri,
2728 S Baylor, K Wu et al. 2002. Gene expression profile of HIV-1 Tat expressing cells: a
2729 close interplay between proliferative and differentiation signals. *BMC Biochem*, 3:14.
2730 doi: 10.1186/1471-2091-3-14
- 2731 [174] G Lefebvre, S Desfarges, F Uyttendaele, M Muoz, N Beerenswinkel, J Rougemont,
2732 A Telenti and A Ciuffi. 2011. Analysis of HIV-1 expression level and sense of
2733 transcription by high-throughput sequencing of the infected cell. *J Virol*, 85:6205–6211.
2734 doi: 10.1128/JVI.00252-11
- 2735 [175] B Langmead, C Trapnell, M Pop and SL Salzberg. 2009. Ultrafast and memory-efficient
2736 alignment of short DNA sequences to the human genome. *Genome Biology*, 10:R25.
2737 doi: 10.1186/gb-2009-10-3-r25
- 2738 [176] GR Grant, MH Farkas, AD Pizarro, NF Lahens, J Schug, BP Brunk, CJ Stoeckert,
2739 JB Hogenesch and EA Pierce. 2011. Comparative analysis of RNA-Seq alignment
2740 algorithms and the RNA-Seq unified mapper (RUM). *Bioinformatics*, 27:2518–2528.
2741 doi: 10.1093/bioinformatics/btr427
- 2742 [177] Q Li, AJ Smith, TW Schacker, JV Carlis, L Duan, CS Reilly and AT Haase. 2009.
2743 Microarray analysis of lymphatic tissue reveals stage-specific, gene expression signatures
2744 in HIV-1 infection. *J Immunol*, 183:1975–1982. doi: 10.4049/jimmunol.0803222
- 2745 [178] A Subramanian, P Tamayo, VK Mootha, S Mukherjee, BL Ebert, MA Gillette,
2746 A Paulovich, SL Pomeroy, TR Golub et al. 2005. Gene set enrichment analysis: a
2747 knowledge-based approach for interpreting genome-wide expression profiles. *Proc Natl
2748 Acad Sci U S A*, 102:15545–15550. doi: 10.1073/pnas.0506580102
- 2749 [179] WJ Kent, CW Sugnet, TS Furey, KM Roskin, TH Pringle, AM Zahler and D Haussler.
2750 2002. The human genome browser at UCSC. *Genome Res*, 12:996–1006. doi: 10.1101/
2751 gr.229102
- 2752 [180] H Li, B Handsaker, A Wysoker, T Fennell, J Ruan, N Homer, G Marth, G Abecasis,
2753 R Durbin and GPDPS . 2009. The Sequence Alignment/Map format and SAMtools.
2754 *Bioinformatics*, 25:2078–2079. doi: 10.1093/bioinformatics/btp352
- 2755 [181] RP Subramanian, JH Wildschutte, C Russo and JM Coffin. 2011. Identification,
2756 characterization, and comparative genomic distribution of the HERV-K (HML-2) group
2757 of human endogenous retroviruses. *Retrovirology*, 8:90. doi: 10.1186/1742-4690-8-90

- 2758 [182] G La Mantia, D Maglione, G Pengue, A Di Cristofano, A Simeone, L Lanfrancone
2759 and L Lania. 1991. Identification and characterization of novel human endogenous
2760 retroviral sequences preferentially expressed in undifferentiated embryonal carcinoma
2761 cells. *Nucleic Acids Res*, 19:1513–1520
- 2762 [183] G La Mantia, B Majello, A Di Cristofano, M Strazzullo, G Minchietti and L Lania.
2763 1992. Identification of regulatory elements within the minimal promoter region of the
2764 human endogenous ERV9 proviruses: accurate transcription initiation is controlled by
2765 an Inr-like element. *Nucleic Acids Res*, 20:4129–4136. doi: 10.1093/nar/20.16.4129
- 2766 [184] KE Plant, SJ Routledge and NJ Proudfoot. 2001. Intergenic transcription in the
2767 human beta-globin gene cluster. *Mol Cell Biol*, 21:6507–6514. doi: 10.1128/MCB.21.
2768 19.6507-6514.2001
- 2769 [185] J Ling, W Pi, R Bollag, S Zeng, M Keskintepe, H Saliman, S Krantz, B Whitney and
2770 D Tuan. 2002. The solitary long terminal repeats of ERV-9 endogenous retrovirus are
2771 conserved during primate evolution and possess enhancer activities in embryonic and
2772 hematopoietic cells. *J Virol*, 76:2410–2423. doi: 10.1128/jvi.76.5.2410-2423.2002
- 2773 [186] X Yu, X Zhu, W Pi, J Ling, L Ko, Y Takeda and D Tuan. 2005. The long terminal
2774 repeat (LTR) of ERV-9 human endogenous retrovirus binds to NF-Y in the assembly
2775 of an active LTR enhancer complex NF-Y/MZF1/GATA-2. *J Biol Chem*, 280:35184–
2776 35194. doi: 10.1074/jbc.M508138200
- 2777 [187] RC Edgar. 2004. MUSCLE: a multiple sequence alignment method with reduced time
2778 and space complexity. *BMC Bioinformatics*, 5:113. doi: 10.1186/1471-2105-5-113
- 2779 [188] M Rotger, KK Dang, J Fellay, EL Heinzen, S Feng, P Descombes, KV Shianna, D Ge,
2780 HF Gnathard et al. 2010. Genome-wide mRNA expression correlates of viral control
2781 in CD4+ T-cells from HIV-1-infected individuals. *PLoS Pathog*, 6:e1000781. doi:
2782 10.1371/journal.ppat.1000781
- 2783 [189] M Rotger, J Dalmau, A Rauch, P McLaren, SE Bosinger, R Martinez, NG Sandler,
2784 A Roque, J Liebner et al. 2011. Comparative transcriptomics of extreme phenotypes
2785 of human HIV-1 infection and SIV infection in sooty mangabey and rhesus macaque.
2786 *J Clin Invest*, 121:2391–2400. doi: 10.1172/JCI45235
- 2787 [190] K Breuer, AK Foroushani, MR Laird, C Chen, A Sribnaia, R Lo, GL Winsor, REW
2788 Hancock, FSL Brinkman and DJ Lynn. 2013. InnateDB: systems biology of innate
2789 immunity and beyond—recent updates and continuing curation. *Nucleic Acids Res*, 41:
2790 D1228–D1233. doi: 10.1093/nar/gks1147
- 2791 [191] I Rusinova, S Forster, S Yu, A Kannan, M Masse, H Cumming, R Chapman and
2792 PJ Hertzog. 2013. Interferome v2.0: an updated database of annotated interferon-
2793 regulated genes. *Nucleic Acids Res*, 41:D1040–D1046. doi: 10.1093/nar/gks1215
- 2794 [192] ST Chang, MJ Thomas, P Sova, RR Green, RE Palermo and MG Katze. 2013.

- 2795 Next-generation sequencing of small RNAs from HIV-infected cells identifies phased
2796 microrna expression patterns and candidate novel microRNAs differentially expressed
2797 upon infection. *MBio*, 4:e00549–e00512. doi: 10.1128/mBio.00549-12
- 2798 [193] Z Kalender Atak, K De Keersmaecker, V Gianflicci, E Geerdens, R Vandepoel,
2799 D Pauwels, M Porcu, I Lahortiga, V Brys et al. 2012. High accuracy mutation
2800 detection in leukemia on a selected panel of cancer genes. *PLoS One*, 7:e38463. doi:
2801 10.1371/journal.pone.0038463
- 2802 [194] ES Patel and LJ Chang. 2012. Synergistic effects of interleukin-7 and pre-T cell
2803 receptor signaling in human T cell development. *J Biol Chem*, 287:33826–33835. doi:
2804 10.1074/jbc.M112.380113
- 2805 [195] M Imbeault, M Ouellet and MJ Tremblay. 2009. Microarray study reveals that HIV-1
2806 induces rapid type-I interferon-dependent p53 mRNA up-regulation in human primary
2807 CD4+ T cells. *Retrovirology*, 6:5. doi: 10.1186/1742-4690-6-5
- 2808 [196] S Iwase, Y Furukawa, J Kikuchi, M Nagai, Y Terui, M Nakamura and H Yamada.
2809 1997. Modulation of E2F activity is linked to interferon-induced growth suppression
2810 of hematopoietic cells. *J Biol Chem*, 272:12406–12414. doi: 10.1074/jbc.272.19.12406
- 2811 [197] RW Johnstone, JA Kerry and JA Trapani. 1998. The human interferon-inducible
2812 protein, IFI 16, is a repressor of transcription. *J Biol Chem*, 273:17172–17177. doi:
2813 10.1074/jbc.273.27.17172
- 2814 [198] BR Williams. 1999. PKR; a sentinel kinase for cellular stress. *Oncogene*, 18:6112–6120.
2815 doi: 10.1038/sj.onc.1203127
- 2816 [199] CV Ramana, N Grammatikakis, M Chernov, H Nguyen, KC Goh, BR Williams and
2817 GR Stark. 2000. Regulation of c-myc expression by IFN-gamma through Stat1-
2818 dependent and -independent pathways. *EMBO J*, 19:263–272. doi: 10.1093/emboj/19.
2819 2.263
- 2820 [200] SL Liang, D Quirk and A Zhou. 2006. RNase L: its biological roles and regulation.
2821 *IUBMB Life*, 58:508–514. doi: 10.1080/15216540600838232
- 2822 [201] F Maldarelli, C Xiang, G Chamoun and SL Zeichner. 1998. The expression of the
2823 essential nuclear splicing factor SC35 is altered by human immunodeficiency virus
2824 infection. *Virus Res*, 53:39–51
- 2825 [202] A Monette, L Ajamian, M López-Lastra and AJ Mouland. 2009. Human immun-
2826 odeficiency virus type 1 (HIV-1) induces the cytoplasmic retention of heterogeneous
2827 nuclear ribonucleoprotein A1 by disrupting nuclear import: implications for HIV-1
2828 gene expression. *J Biol Chem*, 284:31350–31362. doi: 10.1074/jbc.M109.048736
- 2829 [203] R Contreras-Galindo, P López, R Vélez and Y Yamamura. 2007. HIV-1 infection

- 2830 increases the expression of human endogenous retroviruses type K (HERV-K) in vitro.
2831 *AIDS Res Hum Retroviruses*, 23:116–122. doi: 10.1089/aid.2006.0117
- 2832 [204] R Contreras-Galindo, MH Kaplan, S He, AC Contreras-Galindo, MJ Gonzalez-
2833 Hernandez, F Kappes, D Dube, SM Chan, D Robinson et al. 2013. HIV infection
2834 reveals widespread expansion of novel centromeric human endogenous retroviruses.
2835 *Genome Res*, 23:1505–1513. doi: 10.1101/gr.144303.112
- 2836 [205] N Bhardwaj, F Maldarelli, J Mellors and JM Coffin. 2014. HIV-1 infection leads to
2837 increased transcription of human endogenous retrovirus HERV-K (HML-2) proviruses
2838 in vivo but not to increased virion production. *J Virol*, 88:11108–11120. doi: 10.1128/
2839 JVI.01623-14
- 2840 [206] RB Jones, H Song, Y Xu, KE Garrison, AA Buzdin, N Anwar, DV Hunter, S Mujib,
2841 V Mihajlovic et al. 2013. LINE-1 retrotransposable element DNA accumulates in
2842 HIV-1-infected cells. *J Virol*, 87:13307–13320. doi: 10.1128/JVI.02257-13
- 2843 [207] P Medstrand and DL Mager. 1998. Human-specific integrations of the HERV-K
2844 endogenous retrovirus family. *J Virol*, 72:9782–9787
- 2845 [208] C Macfarlane and P Simmonds. 2004. Allelic variation of HERV-K(HML-2) endogenous
2846 retroviral elements in human populations. *J Mol Evol*, 59:642–656. doi: 10.1007/
2847 s00239-004-2656-1
- 2848 [209] K Büscher, U Trefzer, M Hofmann, W Sterry, R Kurth and J Denner. 2005. Expression
2849 of human endogenous retrovirus K in melanomas and melanoma cell lines. *Cancer
2850 Res*, 65:4172–4180. doi: 10.1158/0008-5472.CAN-04-2983
- 2851 [210] G Howard, R Eiges, F Gaudet, R Jaenisch and A Eden. 2008. Activation and
2852 transposition of endogenous retroviral elements in hypomethylation induced tumors in
2853 mice. *Oncogene*, 27:404–408. doi: 10.1038/sj.onc.1210631
- 2854 [211] RC Iskow, MT McCabe, RE Mills, S Torene, WS Pittard, AF Neuwald, EG Van
2855 Meir, PM Vertino and SE Devine. 2010. Natural mutagenesis of human genomes by
2856 endogenous retrotransposons. *Cell*, 141:1253–1261. doi: 10.1016/j.cell.2010.05.020
- 2857 [212] E Lee, R Iskow, L Yang, O Gokcumen, P Haseley, LJ Luquette, 3rd, JG Lohr,
2858 CC Harris, L Ding et al. 2012. Landscape of somatic retrotransposition in human
2859 cancers. *Science*, 337:967–971. doi: 10.1126/science.1222077
- 2860 [213] SW Criscione, Y Zhang, W Thompson, JM Sedivy and N Neretti. 2014. Transcriptional
2861 landscape of repetitive elements in normal and cancer human cells. *BMC Genomics*,
2862 15:583. doi: 10.1186/1471-2164-15-583
- 2863 [214] AW Whisnant, HP Bogerd, O Flores, P Ho, JG Powers, N Sharova, M Stevenson,
2864 CH Chen and BR Cullen. 2013. In-depth analysis of the interaction of HIV-1

- 2865 with cellular microRNA biogenesis and effector mechanisms. *MBio*, 4:e000193. doi:
2866 10.1128/mBio.00193-13
- 2867 [215] NF Lahens, IH Kavakli, R Zhang, K Hayer, MB Black, H Dueck, A Pizarro, J Kim,
2868 R Irizarry et al. 2014. IVT-seq reveals extreme bias in RNA sequencing. *Genome Biol*,
2869 15:R86. doi: 10.1186/gb-2014-15-6-r86
- 2870 [216] RD Hockett, JM Kilby, CA Derdeyn, MS Saag, M Sillers, K Squires, S Chiz, MA Nowak,
2871 GM Shaw and RP Bucy. 1999. Constant mean viral copy number per infected cell in
2872 tissues regardless of high, low, or undetectable plasma HIV RNA. *J Exp Med*, 189:
2873 1545–1554. doi: 10.1084/jem.189.10.1545
- 2874 [217] RJ De Boer, RM Ribeiro and AS Perelson. 2010. Current estimates for HIV-1
2875 production imply rapid viral clearance in lymphoid tissues. *PLoS Comput Biol*, 6:
2876 e1000906. doi: 10.1371/journal.pcbi.1000906
- 2877 [218] T Ikeda, J Shibata, K Yoshimura, A Koito and S Matsushita. 2007. Recurrent HIV-1
2878 integration at the BACH2 locus in resting CD4+ T cell populations during effective
2879 highly active antiretroviral therapy. *J Infect Dis*, 195:716–725. doi: 10.1086/510915
- 2880 [219] TA Wagner, S McLaughlin, K Garg, CYK Cheung, BB Larsen, S Styrvak, HC Huang,
2881 PT Edlefsen, JI Mullins and LM Frenkel. 2014. Proliferation of cells with HIV
2882 integrated into cancer genes contributes to persistent infection. *Science*, 345:570–573.
2883 doi: 10.1126/science.1256304
- 2884 [220] F Maldarelli, X Wu, L Su, FR Simonetti, W Shao, S Hill, J Spindler, AL Ferris,
2885 JW Mellors et al. 2014. Specific HIV integration sites are linked to clonal expansion
2886 and persistence of infected cells. *Science*, 345:179–183. doi: 10.1126/science.1254194
- 2887 [221] LB Cohn, IT Silva, TY Oliveira, RA Rosales, EH Parrish, GH Learn, BH Hahn,
2888 JL Czartoski, MJ McElrath et al. 2015. HIV-1 integration landscape during latent
2889 and active infection. *Cell*, 160:420–432. doi: 10.1016/j.cell.2015.01.020
- 2890 [222] ARW Schröder, P Shinn, H Chen, C Berry, JR Ecker and F Bushman. 2002. HIV-1
2891 integration in the human genome favors active genes and local hotspots. *Cell*, 110:
2892 521–529. doi: 10.1016/S0092-8674(02)00864-4
- 2893 [223] T Brady, YN Lee, K Ronen, N Malani, CC Berry, PD Bieniasz and FD Bushman.
2894 2009. Integration target site selection by a resurrected human endogenous retrovirus.
2895 *Genes Dev*, 23:633–642. doi: 10.1101/gad.1762309
- 2896 [224] B Marini, A Kertesz-Farkas, H Ali, B Lucic, K Lisek, L Manganaro, S Pongor,
2897 R Luzzati, A Recchia et al. 2015. Nuclear architecture dictates HIV-1 integration site
2898 selection. *Nature*. doi: 10.1038/nature14226
- 2899 [225] M Cavazzana-Calvo, E Payen, O Negre, G Wang, K Hehir, F Fusil, J Down, M Denaro,

- 2900 T Brady et al. 2010. Transfusion independence and HMGA2 activation after gene
2901 therapy of human β -thalassaemia. *Nature*, 467:318–322. doi: 10.1038/nature09328
- 2902 [226] S Hacein-Bey-Abina, A Garrigue, GP Wang, J Soulier, A Lim, E Morillon, E Clappier,
2903 L Caccavelli, E Delabesse et al. 2008. Insertional oncogenesis in 4 patients after
2904 retrovirus-mediated gene therapy of SCID-X1. *J Clin Invest*, 118:3132–3142. doi:
2905 10.1172/JCI35700
- 2906 [227] A Moiani, Y Paleari, D Sartori, R Mezzadra, A Miccio, C Cattoglio, F Cocchiarella,
2907 MR Lidonnici, G Ferrari and F Mavilio. 2012. Lentiviral vector integration in the
2908 human genome induces alternative splicing and generates aberrant transcripts. *J Clin
2909 Invest*, 122:1653–1666. doi: 10.1172/JCI61852
- 2910 [228] D Cesana, J Sgualdino, L Rudilosso, S Merella, L Naldini and E Montini. 2012. Whole
2911 transcriptome characterization of aberrant splicing events induced by lentiviral vector
2912 integrations. *J Clin Invest*, 122:1667–1676. doi: 10.1172/JCI62189
- 2913 [229] S Pääbo, DM Irwin and AC Wilson. 1990. DNA damage promotes jumping between
2914 templates during enzymatic amplification. *J Biol Chem*, 265:4718–4721
- 2915 [230] SJ Odelberg, RB Weiss, A Hata and R White. 1995. Template-switching during DNA
2916 synthesis by *Thermus aquaticus* DNA polymerase I. *Nucleic Acids Res*, 23:2049–2057.
2917 doi: 10.1093/nar/23.11.2049
- 2918 [231] XC Zeng and SX Wang. 2002. Evidence that BmTXK beta-BmKCT cDNA from
2919 Chinese scorpion *Buthus martensii* Karsch is an artifact generated in the reverse
2920 transcription process. *FEBS Lett*, 520:183–4; author reply 185
- 2921 [232] B Tasic, CE Nabholz, KK Baldwin, Y Kim, EH Rueckert, SA Ribich, P Cramer,
2922 Q Wu, R Axel and T Maniatis. 2002. Promoter choice determines splice site selection
2923 in protocadherin alpha and gamma pre-mRNA splicing. *Mol Cell*, 10:21–33
- 2924 [233] M Geiszt, K Lekstrom and TL Leto. 2004. Analysis of mRNA transcripts from the
2925 NAD(P)H oxidase 1 (Nox1) gene. Evidence against production of the NADPH oxidase
2926 homolog-1 short (NOH-1S) transcript variant. *J Biol Chem*, 279:51661–51668. doi:
2927 10.1074/jbc.M409325200
- 2928 [234] J Cocquet, A Chong, G Zhang and RA Veitia. 2006. Reverse transcriptase template
2929 switching and false alternative transcripts. *Genomics*, 88:127–131. doi: 10.1016/j.
2930 geno.2005.12.013
- 2931 [235] CJ McManus, JD Coolon, MO Duff, J Eipper-Mains, BR Graveley and PJ Wittkopp.
2932 2010. Regulatory divergence in *Drosophila* revealed by mRNA-seq. *Genome Res*, 20:
2933 816–825. doi: 10.1101/gr.102491.109
- 2934 [236] B Cogné, R Snyder, P Lindenbaum, JB Dupont, R Redon, P Moullier and A Leger.

- 2935 2014. NGS library preparation may generate artifactual integration sites of AAV
2936 vectors. *Nat Med*, 20:577–578. doi: 10.1038/nm.3578
- 2937 [237] E Gilboa, SW Mitra, S Goff and D Baltimore. 1979. A detailed model of reverse
2938 transcription and tests of crucial aspects. *Cell*, 18:93–100. doi: 10.1016/0092-8674(79)
2939 90357-X
- 2940 [238] GX Luo and J Taylor. 1990. Template switching by reverse transcriptase during DNA
2941 synthesis. *J Virol*, 64:4321–4328
- 2942 [239] J Houseley and D Tollervey. 2010. Apparent non-canonical trans-splicing is generated by
2943 reverse transcriptase in vitro. *PLoS One*, 5:e12271. doi: 10.1371/journal.pone.0012271
- 2944 [240] A Meyerhans, JP Vartanian and S Wain-Hobson. 1990. DNA recombination during
2945 PCR. *Nucleic Acids Res*, 18:1687–1691
- 2946 [241] DJG Lahr and LA Katz. 2009. Reducing the impact of PCR-mediated recombination
2947 in molecular evolution and environmental studies using a new-generation high-fidelity
2948 DNA polymerase. *Biotechniques*, 47:857–866. doi: 10.2144/000113219
- 2949 [242] W Al-Ahmadi, L Al-Haj, FA Al-Mohanna, RH Silverman and KSA Khabar. 2009.
2950 RNase L downmodulation of the RNA-binding protein, HuR, and cellular growth.
2951 *Oncogene*, 28:1782–1791. doi: 10.1038/onc.2009.16
- 2952 [243] RB Jones, KE Garrison, S Mujib, V Mihaiovic, N Aidarus, DV Hunter, E Martin,
2953 VM John, W Zhan et al. 2012. HERV-K-specific T cells eliminate diverse HIV-1/2
2954 and SIV primary isolates. *J Clin Invest*, 122:4473–4489. doi: 10.1172/JCI64560
- 2955 [244] K Boller, O Janssen, H Schuldes, RR Tönjes and R Kurth. 1997. Characterization of
2956 the antibody response specific for the human endogenous retrovirus HTDV/HERV-K.
2957 *J Virol*, 71:4581–4588
- 2958 [245] KE Garrison, RB Jones, DA Meiklejohn, N Anwar, LC Ndhlovu, JM Chapman,
2959 AL Erickson, A Agrawal, G Spotts et al. 2007. T cell responses to human endogenous
2960 retroviruses in HIV-1 infection. *PLoS Pathog*, 3:e165. doi: 10.1371/journal.ppat.
2961 0030165
- 2962 [246] R Tandon, D SenGupta, LC Ndhlovu, RGS Vieira, RB Jones, VA York, VA Vieira,
2963 ER Sharp, AA Wiznia et al. 2011. Identification of human endogenous retrovirus-
2964 specific T cell responses in vertically HIV-1-infected subjects. *J Virol*, 85:11526–11531.
2965 doi: 10.1128/JVI.05418-11
- 2966 [247] D SenGupta, R Tandon, RGS Vieira, LC Ndhlovu, R Lown-Hecht, CE Ormsby, L Loh,
2967 RB Jones, KE Garrison et al. 2011. Strong human endogenous retrovirus-specific T
2968 cell responses are associated with control of HIV-1 in chronic infection. *J Virol*, 85:
2969 6977–6985. doi: 10.1128/JVI.00179-11

- 2970 [248] W Pi, Z Yang, J Wang, L Ruan, X Yu, J Ling, S Krantz, C Isales, SJ Conway et al.
2971 2004. The LTR enhancer of ERV-9 human endogenous retrovirus is active in oocytes
2972 and progenitor cells in transgenic zebrafish and humans. *Proc Natl Acad Sci U S A*,
2973 101:805–810. doi: 10.1073/pnas.0307698100
- 2974 [249] XHF Zhang and LA Chasin. 2006. Comparison of multiple vertebrate genomes reveals
2975 the birth and evolution of human exons. *Proc Natl Acad Sci USA*, 103:13427–13432.
2976 doi: 10.1073/pnas.0603042103
- 2977 [250] FA Santoni, J Guerra and J Luban. 2012. HERV-H RNA is abundant in human
2978 embryonic stem cells and a precise marker for pluripotency. *Retrovirology*, 9:111. doi:
2979 10.1186/1742-4690-9-111
- 2980 [251] NV Fuchs, S Loewer, GQ Daley, Z Izsvák, J Löwer and R Löwer. 2013. Human
2981 endogenous retrovirus K (HML-2) RNA and protein expression is a marker for human
2982 embryonic and induced pluripotent stem cells. *Retrovirology*, 10:115. doi: 10.1186/
2983 1742-4690-10-115
- 2984 [252] A Fort, K Hashimoto, D Yamada, M Salimullah, CA Keya, A Saxena, A Bonetti,
2985 I Voineagu, N Bertin et al. 2014. Deep transcriptome profiling of mammalian stem
2986 cells supports a regulatory role for retrotransposons in pluripotency maintenance. *Nat
2987 Genet*, 46:558–566. doi: 10.1038/ng.2965
- 2988 [253] J Wang, G Xie, M Singh, AT Ghanbarian, T Raskó, A Szvetnik, H Cai, D Besser,
2989 A Prigione et al. 2014. Primate-specific endogenous retrovirus-driven transcription
2990 defines naive-like stem cells. *Nature*, 516:405–409. doi: 10.1038/nature13804
- 2991 [254] B Joos, M Fischer, H Kuster, SK Pillai, JK Wong, J Böni, B Hirscher, R Weber,
2992 A Trkola et al. 2008. HIV rebounds from latently infected cells, rather than from
2993 continuing low-level replication. *Proc Natl Acad Sci U S A*, 105:16725–16730. doi:
2994 10.1073/pnas.0804192105
- 2995 [255] TP Brennan, JO Woods, AR Sedaghat, JD Siliciano, RF Siliciano and CO Wilke.
2996 2009. Analysis of human immunodeficiency virus type 1 viremia and provirus in resting
2997 CD4+ T cells reveals a novel source of residual viremia in patients on antiretroviral
2998 therapy. *J Virol*, 83:8470–8481. doi: 10.1128/JVI.02568-08
- 2999 [256] TA Wagner, JL McKernan, NH Tobin, KA Tapia, JI Mullins and LM Frenkel. 2013.
3000 An increasing proportion of monotypic HIV-1 DNA sequences during antiretroviral
3001 treatment suggests proliferation of HIV-infected cells. *J Virol*, 87:1770–1778. doi:
3002 10.1128/JVI.01985-12
- 3003 [257] MF Kearney, J Spindler, W Shao, S Yu, EM Anderson, A O’Shea, C Rehm, C Poethke,
3004 N Kovacs et al. 2014. Lack of detectable HIV-1 molecular evolution during suppressive
3005 antiretroviral therapy. *PLoS Pathog*, 10:e1004010. doi: 10.1371/journal.ppat.1004010
- 3006 [258] KE Ocwieja, S Sherrill-Mix, C Liu, J Song, H Bau and FD Bushman. 2015. A

- 3007 reverse transcription loop-mediated isothermal amplification assay optimized to detect
3008 multiple HIV subtypes. *PLoS One*, 10:e0117852. doi: 10.1371/journal.pone.0117852
- 3009 [259] CJL Murray, KF Ortblad, C Guinovart, SS Lim, TM Wolock, DA Roberts,
3010 EA Dansereau, N Graetz, RM Barber et al. 2014. Global, regional, and national inci-
3011 dence and mortality for HIV, tuberculosis, and malaria during 1990-2013: a systematic
3012 analysis for the Global Burden of Disease Study 2013. *Lancet*, 384:1005–1070. doi:
3013 10.1016/S0140-6736(14)60844-8
- 3014 [260] KA Sollis, PW Smit, S Fiscus, N Ford, M Vitoria, S Essajee, D Barnett, B Cheng,
3015 SM Crowe et al. 2014. Systematic review of the performance of HIV viral load
3016 technologies on plasma samples. *PLoS One*, 9:e85869. doi: 10.1371/journal.pone.
3017 0085869
- 3018 [261] C Liu, M Mauk, R Gross, FD Bushman, PH Edelstein, RG Collman and HH Bau.
3019 2013. Membrane-based, sedimentation-assisted plasma separator for point-of-care
3020 applications. *Anal Chem*, 85:10463–10470. doi: 10.1021/ac402459h
- 3021 [262] KA Curtis, DL Rudolph, I Nejad, J Singleton, A Beddoe, B Weigl, P LaBarre and
3022 SM Owen. 2012. Isothermal amplification using a chemical heating device for point-of-
3023 care detection of HIV-1. *PLoS One*, 7:e31432. doi: 10.1371/journal.pone.0031432
- 3024 [263] T Notomi, H Okayama, H Masubuchi, T Yonekawa, K Watanabe, N Amino and
3025 T Hase. 2000. Loop-mediated isothermal amplification of DNA. *Nucleic Acids Res*,
3026 28:E63
- 3027 [264] KA Curtis, DL Rudolph and SM Owen. 2008. Rapid detection of HIV-1 by reverse-
3028 transcription, loop-mediated isothermal amplification (RT-LAMP). *J Virol Methods*,
3029 151:264–270. doi: 10.1016/j.jviromet.2008.04.011
- 3030 [265] KA Curtis, DL Rudolph and SM Owen. 2009. Sequence-specific detection method for
3031 reverse transcription, loop-mediated isothermal amplification of HIV-1. *J Med Virol*,
3032 81:966–972. doi: 10.1002/jmv.21490
- 3033 [266] Y Zeng, X Zhang, K Nie, X Ding, BZ Ring, L Xu, L Dai, X Li, W Ren et al. 2014.
3034 Rapid quantitative detection of Human immunodeficiency virus type 1 by a reverse
3035 transcription-loop-mediated isothermal amplification assay. *Gene*, 541:123–128. doi:
3036 10.1016/j.gene.2014.03.015
- 3037 [267] N Hosaka, N Ndembí, A Ishizaki, S Kageyama, K Numazaki and H Ichimura. 2009.
3038 Rapid detection of human immunodeficiency virus type 1 group M by a reverse
3039 transcription-loop-mediated isothermal amplification assay. *J Virol Methods*, 157:
3040 195–199. doi: 10.1016/j.jviromet.2009.01.004
- 3041 [268] KA Curtis, PL Niedzwiedz, AS Youngpairoj, DL Rudolph and SM Owen. 2014. Real-
3042 time detection of HIV-2 by reverse transcription-loop-mediated isothermal amplification.
3043 *J Clin Microbiol*, 52:2674–2676. doi: 10.1128/JCM.00935-14

- 3044 [269] C Kuiken, H Yoon, W Abfalterer, B Gaschen, C Lo and B Korber. 2013. Viral
3045 genome analysis and knowledge management. *Methods Mol Biol*, 939:253–261. doi:
3046 10.1007/978-1-62703-107-3_16
- 3047 [270] M Manak, S Sina, B Anekella, I Hewlett, E Sanders-Buell, V Ragupathy, J Kim,
3048 M Vermeulen, SL Stramer et al. 2012. Pilot studies for development of an HIV subtype
3049 panel for surveillance of global diversity. *AIDS Res Hum Retroviruses*, 28:594–606.
3050 doi: 10.1089/AID.2011.0271
- 3051 [271] J Louwagie, FE McCutchan, M Peeters, TP Brennan, E Sanders-Buell, GA Eddy,
3052 G van der Groen, K Fransen, GM Gershay-Damet and R Deleys. 1993. Phylogenetic
3053 analysis of gag genes from 70 international HIV-1 isolates provides evidence for multiple
3054 genotypes. *AIDS*, 7:769–780
- 3055 [272] L Buonaguro, ML Tornesello and FM Buonaguro. 2007. Human immunodeficiency virus
3056 type 1 subtype distribution in the worldwide epidemic: pathogenetic and therapeutic
3057 implications. *J Virol*, 81:10209–10219. doi: 10.1128/JVI.00872-07
- 3058 [273] NF Parrish, F Gao, H Li, EE Giorgi, HJ Barbian, EH Parrish, L Zajic, SS Iyer,
3059 JM Decker et al. 2013. Phenotypic properties of transmitted founder HIV-1. *Proc
3060 Natl Acad Sci U S A*, 110:6626–6633. doi: 10.1073/pnas.1304288110
- 3061 [274] AG Abimiku, TL Stern, A Zwandor, PD Markham, C Calef, S Kyari, WC Saxinger,
3062 RC Gallo, M Robert-Guroff and MS Reitz. 1994. Subgroup G HIV type 1 isolates
3063 from Nigeria. *AIDS Res Hum Retroviruses*, 10:1581–1583
- 3064 [275] Zhang, Chung and Oldenburg. 1999. A simple statistical parameter for use in evaluation
3065 and validation of high throughput screening assays. *J Biomol Screen*, 4:67–73. doi:
3066 10.1177/108705719900400206
- 3067 [276] C Liu, E Geva, M Mauk, X Qiu, WR Abrams, D Malamud, K Curtis, SM Owen
3068 and HH Bau. 2011. An isothermal amplification reactor with an integrated isolation
3069 membrane for point-of-care detection of infectious diseases. *Analyst*, 136:2069–2076.
3070 doi: 10.1039/c1an00007a
- 3071 [277] CA Spina, J Anderson, NM Archin, A Bosque, J Chan, M Famiglietti, WC Greene,
3072 A Kashuba, SR Lewin et al. 2013. An in-depth comparison of latent HIV-1 reactivation
3073 in multiple cell model systems and resting CD4+ T cells from aviremic patients. *PLoS
3074 Pathog*, 9:e1003834. doi: 10.1371/journal.ppat.1003834
- 3075 [278] S Xing, CK Bullen, NS Shroff, L Shan, HC Yang, JL Manucci, S Bhat, H Zhang,
3076 JB Margolick et al. 2011. Disulfiram reactivates latent HIV-1 in a Bcl-2-transduced
3077 primary CD4+ T cell model without inducing global T cell activation. *J Virol*, 85:
3078 6060–6064. doi: 10.1128/JVI.02033-10
- 3079 [279] G Lehrman, IB Hogue, S Palmer, C Jennings, CA Spina, A Wiegand, AL Landay,

- 3080 RW Coombs, DD Richman et al. 2005. Depletion of latent HIV-1 infection in vivo: a
3081 proof-of-concept study. *Lancet*, 366:549–555. doi: 10.1016/S0140-6736(05)67098-5
- 3082 [280] NM Archin, M Cheema, D Parker, A Wiegand, RJ Bosch, JM Coffin, J Eron, M Cohen
3083 and DM Margolis. 2010. Antiretroviral intensification and valproic acid lack sustained
3084 effect on residual HIV-1 viremia or resting CD4+ cell infection. *PLoS One*, 5:e9390.
3085 doi: 10.1371/journal.pone.0009390
- 3086 [281] AM Spivak, A Andrade, E Eisele, R Hoh, P Bacchetti, NN Bumpus, F Emad,
3087 R Buckheit, 3rd, EF McCance-Katz et al. 2014. A pilot study assessing the safety
3088 and latency-reversing activity of disulfiram in HIV-1-infected adults on antiretroviral
3089 therapy. *Clin Infect Dis*, 58:883–890. doi: 10.1093/cid/cit813
- 3090 [282] AR Cillo, MD Sobolewski, RJ Bosch, E Fyne, M Piatak, Jr, JM Coffin and JW Mellors.
3091 2014. Quantification of HIV-1 latency reversal in resting CD4+ T cells from patients
3092 on suppressive antiretroviral therapy. *Proc Natl Acad Sci U S A*, 111:7078–7083. doi:
3093 10.1073/pnas.1402873111
- 3094 [283] AL Hill, DIS Rosenbloom, F Fu, MA Nowak and RF Siliciano. 2014. Predicting the
3095 outcomes of treatment to eradicate the latent reservoir for HIV-1. *Proc Natl Acad Sci
3096 U S A*, 111:13475–13480. doi: 10.1073/pnas.1406663111
- 3097 [284] ENCODE Project Consortium. 2012. An integrated encyclopedia of DNA elements in
3098 the human genome. *Nature*, 489:57–74. doi: 10.1038/nature11247
- 3099 [285] T Barrett, SE Wilhite, P Ledoux, C Evangelista, IF Kim, M Tomashevsky, KA Marshall,
3100 KH Phillippy, PM Sherman et al. 2013. NCBI GEO: archive for functional
3101 genomics data sets—update. *Nucleic Acids Res*, 41:D991–D995. doi: 10.1093/nar/gks1193
- 3102 [286] D Karolchik, GP Barber, J Casper, H Clawson, MS Cline, M Diekhans, TR Dreszer,
3103 PA Fujita, L Guruvadoo et al. 2014. The UCSC Genome Browser database: 2014
3104 update. *Nucleic Acids Res*, 42:D764–D770. doi: 10.1093/nar/gkt1168
- 3105 [287] M Goldman, B Craft, T Swatloski, M Cline, O Morozova, M Diekhans, D Haussler
3106 and J Zhu. 2015. The UCSC Cancer Genomics Browser: update 2015. *Nucleic Acids
3107 Res*, 43:D812–D817. doi: 10.1093/nar/gku1073
- 3108 [288] F Cunningham, MR Amode, D Barrell, K Beal, K Billis, S Brent, D Carvalho-Silva,
3109 P Clapham, G Coates et al. 2015. Ensembl 2015. *Nucleic Acids Res*, 43:D662–D669.
3110 doi: 10.1093/nar/gku1010
- 3111 [289] ML Metzker. 2010. Sequencing technologies - the next generation. *Nat Rev Genet*, 11:
3112 31–46. doi: 10.1038/nrg2626
- 3113 [290] ER Mardis. 2011. A decade's perspective on DNA sequencing technology. *Nature*, 470:
3114 198–203. doi: 10.1038/nature09796

- 3115 [291] K Wetterstrand. 2015. DNA Sequencing Costs: Data from the NHGRI Genome
3116 Sequencing Program (GSP). URL www.genome.gov/sequencingcosts
- 3117 [292] DP Depledge, AL Palser, SJ Watson, IYC Lai, ER Gray, P Grant, RK Kanda,
3118 E Leproust, P Kellam and J Breuer. 2011. Specific capture and whole-genome
3119 sequencing of viruses from clinical samples. *PLoS One*, 6:e27805. doi: 10.1371/journal.
3120 pone.0027805
- 3121 [293] TR Mercer, MB Clark, J Crawford, ME Brunck, DJ Gerhardt, RJ Taft, LK Nielsen,
3122 ME Dinger and JS Mattick. 2014. Targeted sequencing for gene discovery and
3123 quantification using RNA CaptureSeq. *Nat Protoc*, 9:989–1009. doi: 10.1038/nprot.
3124 2014.058
- 3125 [294] JJ Mosher, B Bowman, EL Bernberg, O Shevchenko, J Kan, J Korlach and LA Kaplan.
3126 2014. Improved performance of the PacBio SMRT technology for 16S rDNA sequencing.
3127 *J Microbiol Methods*, 104:59–60. doi: 10.1016/j.mimet.2014.06.012
- 3128 [295] AS Mikhayev and MMY Tin. 2014. A first look at the Oxford Nanopore MinION
3129 sequencer. *Mol Ecol Resour*, 14:1097–1102. doi: 10.1111/1755-0998.12324
- 3130 [296] M Jain, IT Fiddes, KH Miga, HE Olsen, B Paten and M Akeson. 2015. Improved
3131 data analysis for the MinION nanopore sequencer. *Nat Methods*, 12:351–356. doi:
3132 10.1038/nmeth.3290
- 3133 [297] A Kilianski, JL Haas, EJ Corriveau, AT Liem, KL Willis, DR Kadavy, CN Rosenzweig
3134 and SS Minot. 2015. Bacterial and viral identification and differentiation by amplicon
3135 sequencing on the MinION nanopore sequencer. *Gigascience*, 4:12. doi: 10.1186/
3136 s13742-015-0051-z
- 3137 [298] S Jünemann, FJ Sedlazeck, K Prior, A Albersmeier, U John, J Kalinowski, A Mellmann,
3138 A Goesmann, A von Haeseler et al. 2013. Updating benchtop sequencing performance
3139 comparison. *Nat Biotechnol*, 31:294–296. doi: 10.1038/nbt.2522
- 3140 [299] Illumina, Inc. 2015. System specification sheet: MiSeq system. URL <http://www.illumina.com/products/miseq-reagent-kit-v3.html>
- 3141
- 3142 [300] D Rossell, C Stephan-Otto Attolini, M Kroiss and A Stöcker. 2014. Quantifying
3143 alternative splicing from paired-end RNA-sequencing data. *Ann Appl Stat*, 8:309–330.
3144 doi: 10.1214/13-AOAS687
- 3145 [301] N Bray, H Pimentel, P Melsted and L Pachter. 2015. Near-optimal RNA-Seq
3146 quantification. *arXiv preprint*, page 1505.02710
- 3147 [302] NL Michael, MT Vahey, L d'Arcy, PK Ehrenberg, JD Mosca, J Rappaport and RR Red-
3148 field. 1994. Negative-strand RNA transcripts are produced in human immunodeficiency
3149 virus type 1-infected cells and patients by a novel promoter downregulated by Tat. *J Virol*, 68:979–987
3150

- 3151 [303] S Landry, M Halin, S Lefort, B Audet, C Vaquero, JM Mesnard and B Barbeau.
3152 2007. Detection, characterization and regulation of antisense transcripts in HIV-1.
3153 *Retrovirology*, 4:71. doi: 10.1186/1742-4690-4-71
- 3154 [304] NCT Schopman, M Willemsen, YP Liu, T Bradley, A van Kampen, F Baas, B Berkhout
3155 and J Haasnoot. 2012. Deep sequencing of virus-infected cells reveals HIV-encoded
3156 small RNAs. *Nucleic Acids Res*, 40:414–427. doi: 10.1093/nar/gkr719
- 3157 [305] M Kobayashi-Ishihara, M Yamagishi, T Hara, Y Matsuda, R Takahashi, A Miyake,
3158 K Nakano, T Yamochi, T Ishida and T Watanabe. 2012. HIV-1-encoded antisense
3159 RNA suppresses viral replication for a prolonged period. *Retrovirology*, 9:38. doi:
3160 10.1186/1742-4690-9-38
- 3161 [306] S Saayman, A Ackley, AMW Turner, M Famiglietti, A Bosque, M Clemson, V Planelles
3162 and KV Morris. 2014. An HIV-encoded antisense long noncoding RNA epigenetically
3163 regulates viral transcription. *Mol Ther*, 22:1164–1175. doi: 10.1038/mt.2014.29
- 3164 [307] CT Berger, A Llano, JM Carlson, ZL Brumme, MA Brockman, S Cedeño, PR Harrigan,
3165 DE Kaufmann, D Heckerman et al. 2015. Immune screening identifies novel T cell
3166 targets encoded by antisense reading frames of HIV-1. *J Virol*, 89:4015–4019. doi:
3167 10.1128/JVI.03435-14
- 3168 [308] LB Ludwig, JL Ambrus, KA Krawczyk, S Sharma, S Brooks, CB Hsiao and
3169 SA Schwartz. 2006. Human Immunodeficiency Virus-Type 1 LTR DNA contains an
3170 intrinsic gene producing antisense RNA and protein products. *Retrovirology*, 3:80. doi:
3171 10.1186/1742-4690-3-80
- 3172 [309] C Torresilla, É Larocque, S Landry, M Halin, Y Coulombe, JY Masson, JM Mesnard
3173 and B Barbeau. 2013. Detection of the HIV-1 minus-strand-encoded antisense protein
3174 and its association with autophagy. *J Virol*, 87:5089–5105. doi: 10.1128/JVI.00225-13
- 3175 [310] A Bansal, J Carlson, J Yan, OT Akinsiku, M Schaefer, S Sabbaj, A Bet, DN Levy,
3176 S Heath et al. 2010. CD8 T cell response and evolutionary pressure to HIV-1
3177 cryptic epitopes derived from antisense transcription. *J Exp Med*, 207:51–59. doi:
3178 10.1084/jem.20092060
- 3179 [311] JZ Levin, M Yassour, X Adiconis, C Nusbaum, DA Thompson, N Friedman, A Gnirke
3180 and A Regev. 2010. Comprehensive comparative analysis of strand-specific RNA
3181 sequencing methods. *Nat Methods*, 7:709–715. doi: 10.1038/nmeth.1491
- 3182 [312] J Podnar, H Deiderick, G Huerta and S Hunicke-Smith. 2014. Next-generation
3183 sequencing RNA-Seq library construction. *Curr Protoc Mol Biol*, 106:4.21.1–4.21.19.
3184 doi: 10.1002/0471142727.mb0421s106
- 3185 [313] S Cardinaud, A Moris, M Février, PS Rohrlich, L Weiss, P Langlade-Demoyen,
3186 FA Lemonnier, O Schwartz and A Habel. 2004. Identification of cryptic MHC I-

- 3187 restricted epitopes encoded by HIV-1 alternative reading frames. *J Exp Med*, 199:
3188 1053–1063. doi: 10.1084/jem.20031869
- 3189 [314] CT Berger, JM Carlson, CJ Brumme, KL Hartman, ZL Brumme, LM Henry,
3190 PC Rosato, A Piechocka-Trocha, MA Brockman et al. 2010. Viral adaptation to
3191 immune selection pressure by HLA class I-restricted CTL responses targeting epitopes
3192 in HIV frameshift sequences. *J Exp Med*, 207:61–75. doi: 10.1084/jem.20091808
- 3193 [315] NT Ingolia, S Ghaemmaghami, JRS Newman and JS Weissman. 2009. Genome-wide
3194 analysis in vivo of translation with nucleotide resolution using ribosome profiling.
3195 *Science*, 324:218–223. doi: 10.1126/science.1168978
- 3196 [316] NT Ingolia, LF Lareau and JS Weissman. 2011. Ribosome profiling of mouse embryonic
3197 stem cells reveals the complexity and dynamics of mammalian proteomes. *Cell*, 147:
3198 789–802. doi: 10.1016/j.cell.2011.10.002
- 3199 [317] NT Ingolia. 2014. Ribosome profiling: new views of translation, from single codons to
3200 genome scale. *Nat Rev Genet*, 15:205–213. doi: 10.1038/nrg3645
- 3201 [318] NJ Maness, AD Walsh, SM Piaskowski, J Furlott, HL Kolar, AT Bean, NA Wilson
3202 and DI Watkins. 2010. CD8+ T cell recognition of cryptic epitopes is a ubiquitous
3203 feature of AIDS virus infection. *J Virol*, 84:11569–11574. doi: 10.1128/JVI.01419-10
- 3204 [319] A Bet, EA Maze, A Bansal, S Sterrett, A Gross, S Graff-Dubois, A Samri, A Guihot,
3205 C Katlama et al. 2015. The HIV-1 antisense protein (ASP) induces CD8 T cell responses
3206 during chronic infection. *Retrovirology*, 12:15. doi: 10.1186/s12977-015-0135-y
- 3207 [320] S Koenig, HE Gendelman, JM Orenstein, MC Dal Canto, GH Pezeshkpour, M Yung-
3208 bluth, F Janotta, A Aksamit, MA Martin and AS Fauci. 1986. Detection of AIDS
3209 virus in macrophages in brain tissue from AIDS patients with encephalopathy. *Science*,
3210 233:1089–1093. doi: 10.1126/science.3016903
- 3211 [321] S Sonza, HP Mutimer, R Oelrichs, D Jardine, K Harvey, A Dunne, DF Purcell, C Birch
3212 and SM Crowe. 2001. Monocytes harbour replication-competent, non-latent HIV-1 in
3213 patients on highly active antiretroviral therapy. *AIDS*, 15:17–22
- 3214 [322] M Hermankova, JD Siliciano, Y Zhou, D Monie, K Chadwick, JB Margolick, TC Quinn
3215 and RF Siliciano. 2003. Analysis of human immunodeficiency virus type 1 gene
3216 expression in latently infected resting CD4+ T lymphocytes in vivo. *J Virol*, 77:
3217 7383–7392. doi: 10.1128/JVI.77.13.7383-7392.2003
- 3218 [323] N Soriano-Sarabia, RE Bateson, NP Dahl, AM Crooks, JD Kuruc, DM Margolis and
3219 NM Archin. 2014. Quantitation of replication-competent HIV-1 in populations of
3220 resting CD4+ T cells. *J Virol*, 88:14070–14077. doi: 10.1128/JVI.01900-14
- 3221 [324] BF Keele, EE Giorgi, JF Salazar-Gonzalez, JM Decker, KT Pham, MG Salazar, C Sun,
3222 T Grayson, S Wang et al. 2008. Identification and characterization of transmitted and

- 3223 early founder virus envelopes in primary HIV-1 infection. *Proc Natl Acad Sci USA*,
3224 105:7552–7557. doi: 10.1073/pnas.0802203105
- 3225 [325] JF Salazar-Gonzalez, MG Salazar, BF Keele, GH Learn, EE Giorgi, H Li, JM Decker,
3226 S Wang, J Baalwa et al. 2009. Genetic identity, biological phenotype, and evolutionary
3227 pathways of transmitted/founder viruses in acute and early HIV-1 infection. *J Exp
3228 Med*, 206:1273–1289. doi: 10.1084/jem.20090378
- 3229 [326] MP Wentz, BE Moore, MW Cloyd, SM Berget and LA Donehower. 1997. A naturally
3230 arising mutation of a potential silencer of exon splicing in human immunodeficiency
3231 virus type 1 induces dominant aberrant splicing and arrests virus production. *J Virol*,
3232 71:8542–8551
- 3233 [327] JM Madsen and CM Stoltzfus. 2005. An exonic splicing silencer downstream of the 3'
3234 splice site A2 is required for efficient human immunodeficiency virus type 1 replication.
3235 *J Virol*, 79:10478–10486. doi: 10.1128/JVI.79.16.10478-10486.2005
- 3236 [328] S Paca-Uccaralertkun, CK Damgaard, P Auewarakul, A Thitithanyanont,
3237 P Suphaphiphat, M Essex, J Kjems and TH Lee. 2006. The effect of a single nucleotide
3238 substitution in the splicing silencer in the tat/rev intron on HIV type 1 envelope expres-
3239 sion. *AIDS Research & Human Retroviruses*, 22:76–82. doi: 10.1089/aid.2006.22.76
- 3240 [329] D Mandal, Z Feng and CM Stoltzfus. 2008. Gag-processing defect of human immuno-
3241 deficiency virus type 1 integrase E246 and G247 mutants is caused by activation of
3242 an overlapping 5' splice site. *J Virol*, 82:1600–1604. doi: 10.1128/JVI.02295-07
- 3243 [330] T Fukuhara, T Hosoya, S Shimizu, K Sumi, T Oshiro, Y Yoshinaka, M Suzuki,
3244 N Yamamoto, LA Herzenberg et al. 2006. Utilization of host SR protein kinases
3245 and RNA-splicing machinery during viral replication. *Proc Natl Acad Sci USA*, 103:
3246 11329–11333. doi: 10.1073/pnas.0604616103
- 3247 [331] R Wong, A Balachandran, AY Mao, W Dobson, S Gray-Owen and A Cochrane. 2011.
3248 Differential effect of CLK SR Kinases on HIV-1 gene expression: potential novel targets
3249 for therapy. *Retrovirology*, 8:47. doi: 10.1186/1742-4690-8-47
- 3250 [332] RW Wong, A Balachandran, MA Ostrowski and A Cochrane. 2013. Digoxin suppresses
3251 HIV-1 replication by altering viral RNA processing. *PLoS Pathog*, 9:e1003241. doi:
3252 10.1371/journal.ppat.1003241
- 3253 [333] TH Finkel, G Tudor-Williams, NK Banda, MF Cotton, T Curiel, C Monks, TW Baba,
3254 RM Ruprecht and A Kupfer. 1995. Apoptosis occurs predominantly in bystander cells
3255 and not in productively infected cells of HIV- and SIV-infected lymph nodes. *Nat Med*,
3256 1:129–134. doi: 10.1038/nm0295-129
- 3257 [334] G Doitsh, NLK Galloway, X Geng, Z Yang, KM Monroe, O Zepeda, PW Hunt,
3258 H Hatano, S Sowinski et al. 2014. Cell death by pyroptosis drives CD4 T-cell depletion
3259 in HIV-1 infection. *Nature*, 505:509–514. doi: 10.1038/nature12940

- 3260 [335] B Bahbouhi and L al Harthi. 2004. Enriching for HIV-infected cells using anti-gp41
3261 antibodies indirectly conjugated to magnetic microbeads. *Biotechniques*, 36:139–147
- 3262 [336] S Hrvatin, F Deng, CW O'Donnell, DK Gifford and DA Melton. 2014. MARIS:
3263 method for analyzing RNA following intracellular sorting. *PLoS One*, 9:e89459. doi:
3264 10.1371/journal.pone.0089459
- 3265 [337] KM Monroe, Z Yang, JR Johnson, X Geng, G Doitsh, NJ Krogan and WC Greene.
3266 2014. IFI16 DNA sensor is required for death of lymphoid CD4 T cells abortively
3267 infected with HIV. *Science*, 343:428–432. doi: 10.1126/science.1243640
- 3268 [338] CM de Noronha, MP Sherman, HW Lin, MV Cavrois, RD Moir, RD Goldman and
3269 WC Greene. 2001. Dynamic disruptions in nuclear envelope architecture and integrity
3270 induced by HIV-1 Vpr. *Science*, 294:1105–1108. doi: 10.1126/science.1063957
- 3271 [339] M Wilkinson. 1988. A rapid and convenient method for isolation of nuclear, cytoplasmic
3272 and total cellular RNA. *Nucleic Acids Res*, 16:10934. doi: 10.1093/nar/16.22.1093
- 3273 [340] HW Trask, R Cowper-Sal-lari, MA Sartor, J Gui, CV Heath, J Renuka, AJ Higgins,
3274 P Andrews, M Korc et al. 2009. Microarray analysis of cytoplasmic versus whole cell
3275 RNA reveals a considerable number of missed and false positive mRNAs. *RNA*, 15:
3276 1917–1928. doi: 10.1261/rna.1677409
- 3277 [341] BW Solnestam, H Stranneheim, J Hällman, M Käller, E Lundberg, J Lundeberg and
3278 P Akan. 2012. Comparison of total and cytoplasmic mRNA reveals global regulation by
3279 nuclear retention and miRNAs. *BMC Genomics*, 13:574. doi: 10.1186/1471-2164-13-574
- 3280 [342] M Lagos-Quintana, R Rauhut, W Lendeckel and T Tuschl. 2001. Identification of
3281 novel genes coding for small expressed RNAs. *Science*, 294:853–858. doi: 10.1126/
3282 science.1064921
- 3283 [343] V Ambros. 2004. The functions of animal microRNAs. *Nature*, 431:350–355. doi:
3284 10.1038/nature02871
- 3285 [344] P Landgraf, M Rusu, R Sheridan, A Sewer, N Iovino, A Aravin, S Pfeffer, A Rice,
3286 AO Kamphorst et al. 2007. A mammalian microRNA expression atlas based on small
3287 RNA library sequencing. *Cell*, 129:1401–1414. doi: 10.1016/j.cell.2007.04.040
- 3288 [345] Z Klase, P Kale, R Winograd, MV Gupta, M Heydarian, R Berro, T McCaffrey
3289 and F Kashanchi. 2007. HIV-1 TAR element is processed by Dicer to yield a viral
3290 micro-RNA involved in chromatin remodeling of the viral LTR. *BMC Mol Biol*, 8:63.
3291 doi: 10.1186/1471-2199-8-63
- 3292 [346] DL Ouellet, I Plante, P Landry, C Barat, ME Janelle, L Flamand, MJ Tremblay
3293 and P Provost. 2008. Identification of functional microRNAs released through
3294 asymmetrical processing of HIV-1 TAR element. *Nucleic Acids Res*, 36:2353–2365.
3295 doi: 10.1093/nar/gkn076

- 3296 [347] Z Klase, R Winograd, J Davis, L Carpio, R Hildreth, M Heydarian, S Fu, T McCaffrey,
3297 E Meiri et al. 2009. HIV-1 TAR miRNA protects against apoptosis by altering cellular
3298 gene expression. *Retrovirology*, 6:18. doi: 10.1186/1742-4690-6-18
- 3299 [348] S Omoto, M Ito, Y Tsutsumi, Y Ichikawa, H Okuyama, EA Brisibe, NK Saksena and
3300 YR Fujii. 2004. HIV-1 nef suppression by virally encoded microRNA. *Retrovirology*, 1:
3301 44. doi: 10.1186/1742-4690-1-44
- 3302 [349] R Triboulet, B Mari, YL Lin, C Chable-Bessia, Y Bennasser, K Lebrigand, B Cardinaud,
3303 T Maurin, P Barbry et al. 2007. Suppression of microRNA-silencing pathway by
3304 HIV-1 during virus replication. *Science*, 315:1579–1582. doi: 10.1126/science.1136319
- 3305 [350] C Chable-Bessia, O Meziane, D Latreille, R Triboulet, A Zamborlini, A Wagschal,
3306 JM Jacquet, J Reynes, Y Levy et al. 2009. Suppression of HIV-1 replication by
3307 microRNA effectors. *Retrovirology*, 6:26. doi: 10.1186/1742-4690-6-26
- 3308 [351] S Pfeffer, A Sewer, M Lagos-Quintana, R Sheridan, C Sander, FA Grässer, LF van
3309 Dyk, CK Ho, S Shuman et al. 2005. Identification of microRNAs of the herpesvirus
3310 family. *Nat Methods*, 2:269–276. doi: 10.1038/nmeth746
- 3311 [352] J Lin and BR Cullen. 2007. Analysis of the interaction of primate retroviruses with
3312 the human RNA interference machinery. *J Virol*, 81:12218–12226. doi: 10.1128/JVI.
3313 01390-07
- 3314 [353] Y Bennasser, SY Le, M Benkirane and KT Jeang. 2005. Evidence that HIV-1
3315 encodes an siRNA and a suppressor of RNA silencing. *Immunity*, 22:607–619. doi:
3316 10.1016/j.immuni.2005.03.010
- 3317 [354] S Qian, X Zhong, L Yu, B Ding, P de Haan and K Boris-Lawrie. 2009. HIV-1
3318 Tat RNA silencing suppressor activity is conserved across kingdoms and counteracts
3319 translational repression of HIV-1. *Proc Natl Acad Sci U S A*, 106:605–610. doi:
3320 10.1073/pnas.0806822106
- 3321 [355] TL Sung and AP Rice. 2009. miR-198 inhibits HIV-1 gene expression and replication
3322 in monocytes and its mechanism of action appears to involve repression of cyclin T1.
3323 *PLoS Pathog*, 5:e1000263. doi: 10.1371/journal.ppat.1000263
- 3324 [356] G Swaminathan, F Rossi, LJ Sierra, A Gupta, S Navas-Martín and J Martín-García.
3325 2012. A role for microRNA-155 modulation in the anti-HIV-1 effects of Toll-like
3326 receptor 3 stimulation in macrophages. *PLoS Pathog*, 8:e1002937. doi: 10.1371/journal.
3327 ppat.1002937
- 3328 [357] HS Zhang, TC Wu, WW Sang and Z Ruan. 2012. MiR-217 is involved in Tat-
3329 induced HIV-1 long terminal repeat (LTR) transactivation by down-regulation of
3330 SIRT1. *Biochim Biophys Acta*, 1823:1017–1023. doi: 10.1016/j.bbamcr.2012.02.014
- 3331 [358] HS Zhang, XY Chen, TC Wu, WW Sang and Z Ruan. 2012. MiR-34a is involved in Tat-

- 3332 induced HIV-1 long terminal repeat (LTR) transactivation through the SIRT1/NF κ B
3333 pathway. *FEBS Lett*, 586:4203–4207. doi: 10.1016/j.febslet.2012.10.023
- 3334 [359] K Chiang, H Liu and AP Rice. 2013. miR-132 enhances HIV-1 replication. *Virology*,
3335 438:1–4. doi: 10.1016/j.virol.2012.12.016
- 3336 [360] E Orecchini, M Doria, A Michienzi, E Giuliani, L Vassena, SA Ciafrè, MG Farace and
3337 S Galardi. 2014. The HIV-1 Tat protein modulates CD4 expression in human T cells
3338 through the induction of miR-222. *RNA Biol*, 11:334–338. doi: 10.4161/rna.28372
- 3339 [361] L Farberov, E Herzig, S Modai, O Isakov, A Hizi and N Shomron. 2015. MicroRNA-
3340 mediated regulation of p21 and TASK1 cellular restriction factors enhances HIV-1
3341 infection. *J Cell Sci*, 128:1607–1616. doi: 10.1242/jcs.167817
- 3342 [362] J Huang, F Wang, E Argyris, K Chen, Z Liang, H Tian, W Huang, K Squires,
3343 G Verlinghieri and H Zhang. 2007. Cellular microRNAs contribute to HIV-1 latency in
3344 resting primary CD4+ T lymphocytes. *Nat Med*, 13:1241–1247. doi: 10.1038/nm1639
- 3345 [363] K Chiang, TL Sung and AP Rice. 2012. Regulation of cyclin T1 and HIV-1 Replication
3346 by microRNAs in resting CD4+ T lymphocytes. *J Virol*, 86:3244–3252. doi: 10.1128/
3347 JVI.05065-11
- 3348 [364] K Chiang and AP Rice. 2012. MicroRNA-mediated restriction of HIV-1 in resting
3349 CD4+ T cells and monocytes. *Viruses*, 4:1390–1409. doi: 10.3390/v4091390
- 3350 [365] PJ Kanki, DJ Hamel, JL Sankalé, C Hsieh, I Thior, F Barin, SA Woodcock, A Guèye-
3351 Ndiaye, E Zhang et al. 1999. Human immunodeficiency virus type 1 subtypes differ in
3352 disease progression. *J Infect Dis*, 179:68–73. doi: 10.1086/314557
- 3353 [366] P Kaleebu, N French, C Mahe, D Yirrell, C Watera, F Lyagoba, J Nakiyingi, A Rutebe-
3354 mberwa, D Morgan et al. 2002. Effect of human immunodeficiency virus (HIV) type 1
3355 envelope subtypes A and D on disease progression in a large cohort of HIV-1-positive
3356 persons in Uganda. *J Infect Dis*, 185:1244–1250. doi: 10.1086/340130
- 3357 [367] JM Baeten, B Chohan, L Lavreys, V Chohan, RS McClelland, L Certain, K Mandaliya,
3358 W Jaoko and J Overbaugh. 2007. HIV-1 subtype D infection is associated with faster
3359 disease progression than subtype A in spite of similar plasma HIV-1 loads. *J Infect
3360 Dis*, 195:1177–1180. doi: 10.1086/512682
- 3361 [368] N Kiwanuka, O Laeyendecker, M Robb, G Kigozi, M Arroyo, F McCutchan, LA Eller,
3362 M Eller, F Makumbi et al. 2008. Effect of human immunodeficiency virus Type 1
3363 (HIV-1) subtype on disease progression in persons from Rakai, Uganda, with incident
3364 HIV-1 infection. *J Infect Dis*, 197:707–713. doi: 10.1086/527416
- 3365 [369] B Renjifo, P Gilbert, B Chaplin, G Msamanga, D Mwakagile, W Fawzi, M Essex,
3366 TV and HIVS Group. 2004. Preferential in-utero transmission of HIV-1 subtype C as
3367 compared to HIV-1 subtype A or D. *AIDS*, 18:1629–1636

- 3368 [370] GC John-Stewart, RW Nduati, CM Rousseau, DA Mbori-Ngacha, BA Richardson,
3369 S Rainwater, DD Panteleeff and J Overbaugh. 2005. Subtype C Is associated with
3370 increased vaginal shedding of HIV-1. *J Infect Dis*, 192:492–496. doi: 10.1086/431514
- 3371 [371] W Huang, SH Eshleman, J Toma, S Fransen, E Stawiski, EE Paxinos, JM Whitcomb,
3372 AM Young, D Donnell et al. 2007. Coreceptor tropism in human immunodeficiency virus
3373 type 1 subtype D: high prevalence of CXCR4 tropism and heterogeneous composition
3374 of viral populations. *J Virol*, 81:7885–7893. doi: 10.1128/JVI.00218-07
- 3375 [372] J Snoeck, R Kantor, RW Shafer, K Van Laethem, K Deforche, AP Carvalho, B Wyn-
3376 hoven, MA Soares, P Cane et al. 2006. Discordances between interpretation algorithms
3377 for genotypic resistance to protease and reverse transcriptase inhibitors of human
3378 immunodeficiency virus are subtype dependent. *Antimicrob Agents Chemother*, 50:
3379 694–701. doi: 10.1128/AAC.50.2.694-701.2006
- 3380 [373] PJ Easterbrook, M Smith, J Mullen, S O’Shea, I Chrystie, A de Ruiter, ID Tatt,
3381 AM Geretti and M Zuckerman. 2010. Impact of HIV-1 viral subtype on disease
3382 progression and response to antiretroviral therapy. *J Int AIDS Soc*, 13:4. doi: 10.1186/
3383 1758-2652-13-4
- 3384 [374] AU Scherrer, B Ledergerber, V von Wyl, J Böni, S Yerly, T Klimkait, P Bürgisser,
3385 A Rauch, B Hirscher et al. 2011. Improved virological outcome in White patients
3386 infected with HIV-1 non-B subtypes compared to subtype B. *Clin Infect Dis*, 53:
3387 1143–1152. doi: 10.1093/cid/cir669
- 3388 [375] C Liu, MM Sadik, MG Mauk, PH Edelstein, FD Bushman, R Gross and HH Bau.
3389 2014. Nuclemeter: a reaction-diffusion based method for quantifying nucleic acids
3390 undergoing enzymatic amplification. *Sci Rep*, 4:7335. doi: 10.1038/srep07335
- 3391 [376] MG Mauk, C Liu, M Sadik and HH Bau. 2015. Microfluidic devices for nucleic acid
3392 (NA) isolation, isothermal NA amplification, and real-time detection. *Methods Mol
3393 Biol*, 1256:15–40. doi: 10.1007/978-1-4939-2172-0_2
- 3394 [377] A Piantadosi, B Chohan, V Chohan, RS McClelland and J Overbaugh. 2007. Chronic
3395 HIV-1 infection frequently fails to protect against superinfection. *PLoS Pathog*, 3:e177.
3396 doi: 10.1371/journal.ppat.0030177
- 3397 [378] RLR Powell, MM Urbanski, S Burda, T Kinge and PN Nyambi. 2009. High frequency
3398 of HIV-1 dual infections among HIV-positive individuals in Cameroon, West Central
3399 Africa. *J Acquir Immune Defic Syndr*, 50:84–92. doi: 10.1097/QAI.0b013e31818d5a40
- 3400 [379] K Ronen, CO McCoy, FA Matsen, DF Boyd, S Emery, K Odem-Davis, W Jaoko,
3401 K Mandaliya, RS McClelland et al. 2013. HIV-1 superinfection occurs less frequently
3402 than initial infection in a cohort of high-risk Kenyan women. *PLoS Pathog*, 9:e1003593.
3403 doi: 10.1371/journal.ppat.1003593
- 3404 [380] AD Redd, D Ssemwanga, J Vandepitte, SK Wendel, N Ndembí, J Bukenya,

- 3405 S Nakubulwa, H Grosskurth, CM Parry et al. 2014. Rates of HIV-1 superinfection
3406 and primary HIV-1 infection are similar in female sex workers in Uganda. *AIDS*,
3407 28:2147–2152. doi: 10.1097/QAD.0000000000000365
- 3408 [381] AD Redd, CE Mullis, D Serwadda, X Kong, C Martens, SM Ricklefs, AAR Tobian,
3409 C Xiao, MK Grabowski et al. 2012. The rates of HIV superinfection and primary HIV
3410 incidence in a general population in Rakai, Uganda. *J Infect Dis*, 206:267–274. doi:
3411 10.1093/infdis/jis325
- 3412 [382] S Jost, MC Bernard, L Kaiser, S Yerly, B Hirschel, A Samri, B Autran, LE Goh and
3413 L Perrin. 2002. A patient with HIV-1 superinfection. *N Engl J Med*, 347:731–736. doi:
3414 10.1056/NEJMoa020263
- 3415 [383] G Fang, B Weiser, C Kuiken, SM Philpott, S Rowland-Jones, F Plummer, J Kimani,
3416 B Shi, R Kaul et al. 2004. Recombination following superinfection by HIV-1. *AIDS*,
3417 18:153–159
- 3418 [384] G Blick, RM Kagan, E Coakley, C Petropoulos, L Maroldo, P Greiger-Zanolungo,
3419 S Gretz and T Garton. 2007. The probable source of both the primary multidrug-
3420 resistant (MDR) HIV-1 strain found in a patient with rapid progression to AIDS and
3421 a second recombinant MDR strain found in a chronically HIV-1-infected patient. *J
3422 Infect Dis*, 195:1250–1259. doi: 10.1086/512240
- 3423 [385] GS Gottlieb, DC Nickle, MA Jensen, KG Wong, RA Kaslow, JC Shepherd, JB Margolick
3424 and JI Mullins. 2007. HIV type 1 superinfection with a dual-tropic virus
3425 and rapid progression to AIDS: a case report. *Clin Infect Dis*, 45:501–509. doi:
3426 10.1086/520024
- 3427 [386] H Streeck, B Li, AFY Poon, A Schneidewind, AD Gladden, KA Power, D Daskalakis,
3428 S Bazner, R Zuniga et al. 2008. Immune-driven recombination and loss of control
3429 after HIV superinfection. *J Exp Med*, 205:1789–1796. doi: 10.1084/jem.20080281
- 3430 [387] O Clerc, S Colombo, S Yerly, A Telenti and M Cavassini. 2010. HIV-1 elite controllers:
3431 beware of super-infections. *J Clin Virol*, 47:376–378. doi: 10.1016/j.jcv.2010.01.013
- 3432 [388] DM Smith, JK Wong, GK Hightower, CC Ignacio, KK Koelsch, CJ Petropoulos,
3433 DD Richman and SJ Little. 2005. HIV drug resistance acquired through superinfection.
3434 *AIDS*, 19:1251–1256
- 3435 [389] M Pernas, C Casado, R Fuentes, MJ Pérez-Elías and C López-Galíndez. 2006. A dual
3436 superinfection and recombination within HIV-1 subtype B 12 years after primoinfection.
3437 *J Acquir Immune Defic Syndr*, 42:12–18. doi: 10.1097/01.qai.0000214810.65292.73
- 3438 [390] DL Robertson, PM Sharp, FE McCutchan and BH Hahn. 1995. Recombination in
3439 HIV-1. *Nature*, 374:124–126. doi: 10.1038/374124b0
- 3440 [391] F Gao, E Bailes, DL Robertson, Y Chen, CM Rodenburg, SF Michael, LB Cummins,

- 3441 LO Arthur, M Peeters et al. 1999. Origin of HIV-1 in the chimpanzee Pan troglodytes
3442 troglodytes. *Nature*, 397:436–441. doi: 10.1038/17130
- 3443 [392] BH Hahn, GM Shaw, KM De Cock and PM Sharp. 2000. AIDS as a zoonosis: scientific
3444 and public health implications. *Science*, 287:607–614. doi: 10.1126/science.287.5453.607
- 3445 [393] MH Malim and M Emerman. 2001. HIV-1 sequence variation: drift, shift, and
3446 attenuation. *Cell*, 104:469–472. doi: 10.1016/S0092-8674(01)00234-3
- 3447 [394] SK Gire, A Goba, KG Andersen, RSG Sealfon, DJ Park, L Kanneh, S Jalloh, M Momoh,
3448 M Fullah et al. 2014. Genomic surveillance elucidates Ebola virus origin and transmis-
3449 sion during the 2014 outbreak. *Science*, 345:1369–1372. doi: 10.1126/science.1259657
- 3450 [395] WHO Ebola Response Team. 2014. Ebola virus disease in West Africa—the first 9
3451 months of the epidemic and forward projections. *N Engl J Med*, 371:1481–1495. doi:
3452 10.1056/NEJMoa1411100
- 3453 [396] World Health Organization. 2015. Ebola situation report: 13
3454 May 2014. URL <http://apps.who.int/ebola/en/current-situation/ebola-situation-report-13-may-2015>
- 3455 [397] G Chowell and H Nishiura. 2014. Transmission dynamics and control of Ebola virus
3456 disease (EVD): a review. *BMC Med*, 12:196. doi: 10.1186/s12916-014-0196-0
- 3458 [398] AS Fauci. 2014. Ebola—underscoring the global disparities in health care resources. *N
3459 Engl J Med*, 371:1084–1086. doi: 10.1056/NEJMp1409494
- 3460 [399] World Health Organization. 2015. Interim guidance on the use of rapid Ebola antigen
3461 detection tests. URL <http://www.who.int/csr/resources/publications/ebola/ebola-antigen-detection/en/>
- 3463 [400] Y Kurosaki, A Takada, H Ebihara, A Grolla, N Kamo, H Feldmann, Y Kawaoka and
3464 J Yasuda. 2007. Rapid and simple detection of Ebola virus by reverse transcription-
3465 loop-mediated isothermal amplification. *J Virol Methods*, 141:78–83. doi: 10.1016/j.jviromet.2006.11.031
- 3467 [401] T Hoenen, D Safronetz, A Groseth, KR Wollenberg, OA Koita, B Diarra, IS Fall,
3468 FC Haidara, F Diallo et al. 2015. Mutation rate and genotype variation of Ebola virus
3469 from Mali case sequences. *Science*, 348:117–119. doi: 10.1126/science.aaa5646