Published in final edited form as:

AIDS. 2011 October 23; 25(16): 1951–1959. doi:10.1097/QAD.0b013e32834b34de.

Switching between raltegravir resistance pathways analyzed by deep sequencing

Rithun Mukherjee^a, Shane T. Jensen^b, Frances Male^a, Kyle Bittinger^a, Richard L. Hodinka^{c,d}, Michael D. Miller^e, and Frederic D. Bushman^a

^aDepartment of Microbiology, Perelman School of Medicine, University of Pennsylvania

^bDepartment of Statistics, The Wharton School, University of Pennsylvania

^cDepartment of Pediatrics, Perelman School of Medicine, University of Pennsylvania

^dClinical Virology Laboratory, Children's Hospital of Philadelphia, Philadelphia

^eDepartment of Infectious Diseases, Merck Research Laboratories, West Point, Pennsylvania, USA

Abstract

Objective—Our objective was to analyze the pathways leading to resistance of HIV to the integrase (IN) inhibitor raltegravir (RAL).

Design—Three HIV-infected individuals exhibiting RAL resistance pathway switching were characterized using longitudinal analysis of viral samples from plasma.

Methods—454/Roche pyrosequencing was used to generate approximately 74 000 sequence reads from the integrase coding region. Effects of error were controlled by denoising with Pyronoise, and by comparison to approximately 142 000 control reads from HIV_{NL4-3}. Viral lineages were modeled quantitatively using viral serial pathway analysis (vSPA).

Results—All three patients showed transitions from the N155H pathway to the Q148H/G140S pathway. Analysis with vSPA revealed complex pathways to the final genotype, probably involving both de-novo mutation and recombination. No reads contained both the N155H and Q148H drug resistance mutations (DRMs), indicating that the double mutant is not a prominent intermediate, consistent with low fitness. To characterize possible drug-resistant variants circulating prior to therapy, we sequenced approximately 70 000 reads from samples collected prior to initiating treatment. Although some preexisting drug-resistant variants were detected, N155H, the first major DRM present after initiating RAL therapy, was not detected.

Conclusion—The main DRMs are present at very low levels if at all prior to initiating therapy. We also outline general methods for deep sequence analysis of DRMs in longitudinal HIV samples.

Keywords

AIDS; deep sequencing; HIV; integrase; Pyronoise; raltegravir; viral serial pathway analysis

Introduction

The high mutation rate of HIV can result in rapid development of drug resistance [1,2]. For raltegravir (RAL), which inhibits DNA strand transfer by integrase (IN) [3-5], resistance mutations typically arise in the part of the integrase gene encoding the catalytic domain [6-8]. Three codons can mutate to generate primary resistance mutations, which encode Y143R/C/H, Q148H/R/K, and N155H. Each primary drug-resistance mutation (DRM) is associated with a preferred set of accessory mutations [9,10].

Patients who exhibited virologic failure following RAL treatment often switch from one resistance pathway to another [10,11]. Specifically, the N155H pathway is commonly replaced by the Q148H/R/K pathway, resulting in reduced susceptibility to RAL and improved viral replication capacity [9,10]. However, questions remain regarding the nature of the switch, such as whether resistance mutations were present before treatment, and the nature of intermediates during pathway switching.

Here we investigated three patients for whom conventional viral population genotyping identified an N155H to Q148H pathway switch. Longitudinal analysis of viral population evolution under RAL pressure was performed using 454/Roche pyrosequencing [12], which is well suited to tracking viral variants in complex populations [13-20]. We used the Pyronoise preclustering method for controlling error [21] and performed longitudinal analysis of serially sampled viral populations using viral serial pathway analysis (vSPA) [22] for rigorous lineage analysis. Particularly deep sequencing was also carried out on pretreatment samples and controls to investigate the abundance of possible preexisting resistance mutations.

Methods

Deep sequencing of viral populations

Plasma samples were obtained from patients at multiple time-points and sequenced (Supplementary Tables 1-3, <http://links.lww.com/QAD/A165>). Details of sample preparation are in the Supplementary Methods section, <http://links.lww.com/QAD/A165> [23]. For the analysis below, we assume that losses during the viral isolation and cDNA synthesis steps are negligible (personal communication from manufacturer). All pyrosequence reads are available from our laboratory (at this writing, there is no public database accepting deep sequencing data). Sample metadata according to the MIMARKS standard [24] is available in Supplementary Table 4, <http://links.lww.com/QAD/A165>.

Bioinformatic analysis

All pyrosequence reads were filtered for exact matches to barcodes and primers, and to remove shorter reads (<200 bases long) and reads with more than two ambiguous base calls, which are often error-prone [25]. Output sequences were assigned to each patient/time-point combination using DNA bar codes embedded in the amplification primers. Details of data processing including implementation of Pyronoise and vSPA, as well as simulation-based statistical procedure for identifying DRMs can be found in the Supplementary Methods section, <http://links.lww.com/QAD/A165>.

Computation was carried out on a Dell Power Edge Cluster with 32 cores. Simulations and statistical tests were carried out using R.

Re-sampling statistics for numbers of starting genomes assayed in sequencing experiments after PCR amplification

Given an initial sample of n viral genomes, amplified by f PCR cycles for an amplification factor 2^f , and sequenced to yield s sequences, how many of the n original genomes will be detected among the s sequences? Let this quantity be defined by a random variable Y .

It can be shown that Y has a mean

$$E(Y) = n \left(1 - \left(1 - \frac{1}{n} \right)^s \right),$$

and variance

$$Var(Y) = E(Y) - E(Y)^2 + n(n-1) \times \left(1 - 2 \left(1 - \frac{1}{n} \right)^s + \left(1 - \frac{2}{n} \right)^s \right)$$

These formulae assume sampling with replacement, which is reasonable given the large number of amplified genomes from which we are sampling. This approximation is validated by comparisons in each case with 100 sampling simulations without replacement for 35 PCR cycles (Supplementary Table 2, <http://links.lww.com/QAD/A165>; values labeled 'Simulations' and 'Formula'). Further, the equations are independent of f in the limit of $2^f \rightarrow \infty$, which too was validated with simulations in the ranges of n and f studied (data not shown). These equations, thus, provide a concrete measure of the completeness of sampling and the extent of possible over-sampling due to oversequencing. To calculate the proportion of sequence reads corresponding to independent viral templates in the starting plasma sample, we used (Y/s) , which has mean $E(Y)/s$ and variance $(Var(Y)/s^2)$. Oversampling is then simply $1 - (Y/s)$.

Results

Participants studied

The three patients studied showed a drop in the viral load in the first few weeks of RAL treatment, followed by a rebound with development of resistance. Population genotyping with Sanger sequencing identified N155H as one of the first integrase DRMs to appear, but later the Q148H mutant predominated. For each patient, we studied a baseline sample, obtained just prior to RAL initiation, and several time-points after initiating therapy (Supplementary Tables 1 and 2, <http://links.lww.com/QAD/A165>).

Longitudinal samples assayed using pyrosequencing

HIV RNA was extracted from each plasma sample, reverse-transcribed, and PCR amplified to examine bases 3906 through 4288 (numbering of the HIV-1_{NL4-3} genome). This corresponds to integrase codons 45–171, which span the reported major DRM sites. PCR products were purified and then sequenced using 454/Roche pyrosequencing. RNA and DNA controls from HIV-1_{NL4-3} were also included in this experiment. All samples analyzed are summarized in Supplementary Table 1, <http://links.lww.com/QAD/A165>.

Control of error is central to studies of the low abundance resistance mutations. Error in this study could originate from the reverse transcription step, the PCR step, or the sequence determination step. We used Pyronoise for control of sequencing error [21]. Pyronoise operates in a Bayesian framework to cluster sequence reads as light intensity values

(flowgrams) prior to interpretation as base calls. The resulting clusters are termed OTUs. The number of Pyronoise OTUs was then interpreted as the number of unique variants at each time-point (Supplementary Table 1, <http://links.lww.com/QAD/A165>). No single OTU occurred in more than one patient and no HIV-1_{NL4-3} control OTU was detected in any patient. Thus, there was no detectable contamination between samples.

To determine the optimal parameters for use in the Pyronoise analysis, we generated a first set of parallel HIV-1_{NL4-3} controls (443 reads; Supplementary Table 1, <http://links.lww.com/QAD/A165>), wherein the correct number of OTUs is known to be one (Fig. 1a and b). Under the parameters used, the two HIV-1_{NL4-3} RNA read sets (one each for forward and reverse direction sequencing) each yielded one OTU, whereas the HIV-1_{NL4-3} DNA sets yielded one and two OTUs. Examination of the output showed that an artifact at the extreme edge of the sequence reads led to the formation of two OTUs in one DNA control set. Thus, we conclude that Pyronoise processing resulted in effective though not perfect control of inflation of OTU numbers due to sequencing error. A much larger set of 141 582 HIV-1_{NL4-3} reads was generated for comparison to the deeply sequenced sets for pretreatment time-points and is discussed below.

We estimated the error for the reverse transcription step to be 2×10^{-4} base substitutions per nucleotide by comparing the HIV-1_{NL4-3} DNA and RNA sets. For PCR, the error rate (estimated by the manufacturer) is 1×10^{-5} base substitutions per nucleotide per replication cycle, which results in 1.75×10^{-4} base substitutions per nucleotide over 35 cycles of PCR. Thus, presequencing misincorporations add up to 3.75×10^{-4} per nucleotide. Comparison to the measured overall rate leaves approximately 8×10^{-4} base substitutions per nucleotide resulting from the pyrosequencing procedure. This is applicable to the highest quality portion of the reads over which such errors are uniform. In the 3' part of the reads, the error rate rises sharply, illustrated in Figure 1c and d with data generated from re-sequencing of HIV_{NL4-3}. The presence of homopolymers had no influence on base substitution frequency with our alignment protocol. Our pyrosequencing error is slightly lower than that in previous studies, which used full reads and not just high-quality regions [14,15,26].

An MSA was constructed with de-noised OTUs. Positions of codon polymorphisms within each patient over time are shown in Fig. 2a–c. The sequence reads were then translated *in silico*, and the known integrase DRM sites were abstracted to create an integrase resistance amino acid profile for each patient (Supplementary Figure 1, A–C, <http://links.lww.com/QAD/A165>).

Longitudinal analysis of drug resistance pathways

In the pyrosequence data, the majority of early primary mutations encoded N155H in all three patients, whereas the majority of later primary mutations encoded Q148H, confirming the pathway switch inferred from population genotyping. There was no evidence of any double mutant variants harboring both 155 and 148 pathway primary mutations. All DRMs detected by population genotyping (Supplementary Table 1, <http://links.lww.com/QAD/A165>) were also detected by deep sequencing. A methodological concern is that recombination during RT-PCR *in vitro* may link DRMs artifactually, but no recombinants encoding both substitutions were observed. We did not find evidence of preexisting integrase inhibitor-related primary mutations (positions 143, 148, and 155) prior to initiating therapy in this first pass analysis.

To track the evolution of drug-resistance lineages in a rigorous fashion, we used the vSPA algorithm [22]. For each sequence, a normalized distance vector over all other sequences is used to construct a correlation matrix. Sequences with more than a threshold correlation coefficient are clustered together based on a distribution of such matrices obtained from

permuted datasets, and clusters across serial samples are linked based on average genetic distance to yield a longitudinal phylogenetic network (Fig. 3).

In patient 1 at month 3 after initiating treatment, N155H predominated, though there were rare variants with Q148R and Q148H present (Figs 2a and 3a). Some but not all of the Q148H codons were associated with G140S (middle of month 3 panel in Figs 2a and 3a). Even though the most common substitution at position 148 at month 3 encodes Q148R, it does not occur together with any accessory mutations at codon 138 or 140. Two separate lineages were detected at all three time-points, distinguished by polymorphisms at codons encoding amino acids 124, 125, 129, 130, and 139 (Figs 2a and 3a). Each of the collections of DRMs (N155H, Q148R, and Q148H + G140S) was found on both backgrounds. For most of the mutations, it is simplest to assume that the mutations arose once and recombined onto the different backgrounds. However, for the G140S mutation, the codon is directly adjacent to the polymorphic codon 139, so in this case, recombination would need to break exactly between the two codons to generate the observed genotypes. Thus, independent mutation to generate the G140S substitution on the two backgrounds seems more likely.

Patient 2 also showed N155H switching to Q148H + G140S, but N155H was still detected at low abundance even after 8 months of therapy and a complex collection of intermediates were detected over the period sampled. After 3 months of therapy, N155H, Q148K, and Q148R all coexisted (Figs 2a and 3b). The Q148K + E138K combination was evident at month 3, and though this combination is reported to be a potent RAL escape variant [8], it was not detected subsequently. By month 4, only the N155H variants were detected, whereas by month 8 Y143R, Q148H, and N155H all were detected. At month 12, Q148H was the majority but N155H was still detectable, whereas Y143R was not. Patient 2 was the only participant in whom Y143R and N155H were detectable at later time-points. Tracking the origin of drug resistance lineages using vSPA indicated that all primary DRMs derived from a single ancestral cluster present before initiation of therapy. We also detected T97A, an accessory mutation for Y143R, and L74M and E92Q, accessory mutations for N155H, from the forward read sets (data not shown).

In patient 3, three clusters were detected at time 0, one of which went on to give rise to the N155H drug-resistant lineages by month 3. A different lineage derived from the same ancestral group emerged at month 5 and acquired the Q148H + G140S mutations, which persisted there-after. Q148H replaced N155H after 5 months.

In-depth analysis of pretreatment time-points

An important question in understanding the origin of RAL resistance is whether resistance mutations were present in the viral population prior to initiation of therapy. We, thus, carried out deeper sequencing analysis of the pretreatment time-points. RT-PCR products were sequenced in both directions, yielding 69 862 sequence reads, ranging from approximately 18 000 to approximately 26 000 for each of the three participants.

In designing such a study, it is desirable to avoid sequencing multiple PCR products derived from a single viral genome, as this would waste sequencing effort and potentially confuse the analysis. To minimize this potential problem, the number of viral RNA templates in each RT-PCR reaction was arranged to be greater than the number of pyrosequence reads ultimately determined (Supplementary Table 2, <http://links.lww.com/QAD/A165>). For each sample, 800 µl of plasma was used for RNA purification, and all was used for RT-PCR, so the predicted numbers of templates (assuming full recovery of RNA) were 27 440, 32 560, and 29 440 for patients 1-3, respectively. The numbers of sequences for each sample were lower, 24 937, 16 461, and 24 943 respectively, and distributed between the forward and reverse directions (Supplementary Table 2, <http://links.lww.com/QAD/A165>).

An important question centers on how many of our sequence reads corresponded to different viral genomes present in the initial preamplification sample, and how many were duplicates generated during PCR. We devised a general formula relating the number of viral genomes in the starting sample to the number observed in the sequence output (Methods section). For the three pretreatment samples, we calculated 72–93% of the sequence reads corresponded to independent viral genomes from the starting RNA sample (Supplementary Table 2, <http://links.lww.com/QAD/A165>).

We next investigated the numbers of variants present in the pretreatment samples. To assess reproducibility, the results from the first pass low coverage analysis of the pretreatment samples were compared to the second much deeper analysis. OTU sets were reconstructed and analyzed for the high accuracy 5' part of the reads for both experiments. We plotted the proportion of reads contributed by each OTU within each participant in the two experiments (Fig. 4a). OTUs mostly fell near a line from lower left to upper right, indicating rough reproducibility between experiments. To assess concordance, we divided OTUs into those in the upper 95% of abundance or lower 5% in each experiment (Fig. 4a, indicated by red lines) and asked how many OTUs were concordant between experiments (that is present in the lower left and upper right quadrants). We found that 4050 out of 4062 or 99.7% of OTUs were concordant between experiments by this measure.

454/Roche pyrosequence determination is error-prone, making identification of low-level DRMs challenging. We used two strategies to distinguish low-level authentic DRMs in patient samples from spurious DRMs arising from error. In the first, we determined an additional 141 582 reads for the same amplicon on a homogeneous HIV_{NL4-3} template, calculated the frequency of false calls of DRMs on HIV_{NL4-3}, and then compared these frequencies to those observed in patient baseline samples. In the second strategy, we measured the error frequencies in the HIV_{NL4-3} data and then used the frequencies to model the expected frequency of spurious DRMs in the patient samples. The first approach has the advantage of modeling all aspects of the sequence acquisition process, whereas the second method allows larger numbers of *in silico* generated control sequences to be compared to the patient sequences. For both, only the reverse reads were used because these had the highest quality sequence over the relevant 143–155 coding positions.

Figure 4b shows the *P* values for detection of the major DRMs at amino acids 143, 148, and 155 in pretreatment patient samples when compared to the HIV_{NL4-3} empirical control. The results are compared for both the raw sequence data and de-noised data. A few substitutions show potential significance, including Y143H in all participants and Q148R in participant 3 (Fig. 4b, pink and red shading). However, after correction for multiple comparisons, only Y143H in participant 3 maintained significance (Fig. 4b, red shading).

A similar analysis was carried out comparing the pretreatment patient sequencing data to simulations of DRM accumulation due to error (Supplementary Table 5, <http://links.lww.com/QAD/A165>). For this, the error rate of 1.2×10^{-3} was used to generate a set of codons equal to the number sampled for each patient at each codon. A total of 10 000 such simulations were carried out, and the relationship of the observed data to simulated data assessed. Following correction for multiple comparisons, Y143H attained significance in both patient 1 and 3, but no other positions survived the test for multiple comparisons.

In all three patients, N155H, which was the first DRM to arise after Raltegravir therapy, as well as Q148H, which was the majority primary DRM after pathway switch, were not significantly enriched in the pretreatment samples. Thus, preexisting mutations did not contribute detectably to treatment failure.

Discussion

We analyzed the evolutionary dynamics of RAL DRMs for three patients who showed a switch from the N155H to the Q148H pathway. We describe an analytical pipeline based on Pyronoise and vSPA that may be useful for longitudinal tracking of DRMs in the future, though we note that computational feasibility becomes an issue with larger datasets. The pathways leading to the final states were complex, probably involving multiple rounds of point mutation to generate new variants and recombination to assemble variants on a single genome.

Following our initial survey, we sequenced deeply into the pretreatment time-points to assess whether genomes containing DRMs that became abundant after treatment were detectable before treatment. Comparisons to control sequences from HIV_{NL4-3} or to results of simulations showed that most DRMs were not convincingly detectable above the error background. We did find significant enrichment for Y143H in one participant by all measures and possible enrichment in a second (marginally significant relative to HIV_{NL4-3} and significant relative to simulated data). The Y143C and Q148R substitutions were marginally significant by some measures but did not survive correction for multiple comparisons. Codoner *et al.* [26] also investigated RAL resistance using 454 sequencing and reported detection of Y143H, Y143C, and Q148R prior to initiation of therapy. Our findings taken together support the idea that Y143H, and possibly Y143C and Q148R, are authentic replication-competent polymorphisms that are present in viral populations in the absence of RAL.

One of the main questions in initiating this analysis was whether the first resistance mutation to arise, N155H, was present prior to initiating therapy. We found a single read for this DRM over all patients, which is readily attributable to error. Thus, the primary DRM that appeared immediately after treatment initiation was not detectable before treatment initiation, despite sequencing to a depth of approximately 19 000 reads in the reverse direction, wherein the DRM positions are of highest quality. Similarly for Q148H, the major form at the last time-point, only three reads were detected, a number also attributable to error. This is consistent with the idea that viruses with N155H or Q148H substitutions are considerably less fit than wild-type and so are very low in abundance in the absence of pressure from RAL.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

The authors are grateful to Warren J. Ewens, Department of Biology, University of Pennsylvania and members of the Bushman laboratory for help and suggestions, especially Rebecca Custers-Allen for assistance with the 454/Roche Junior sequencing workup.

R.M., S.T.J., M.D.M., and F.D.B. designed the study and experiments; R.M., F.M., K.B., and R.L.H. conducted experiments; R.M., K.B., S.T.J., and F.D.B. analyzed results and wrote the article.

The present work was supported by a grant from Merck Research Laboratories and NIH grant R01 AI052845 to F.D.B.

References

1. Coffin JM. HIV population dynamics in vivo: implications for genetic variation, pathogenesis, and therapy. *Science*. 1995; 267:483–486. [PubMed: 7824947]

2. Coffin, JM.; Hughes, SH.; Varmus, HE. Retroviruses. Cold Spring Harbor Laboratory Press; Cold Spring Harbor: 1997.
3. Espeseth AS, Felock P, Wolfe A, Witmer M, Grobler J, Anthony N, et al. HIV-1 integrase inhibitors that compete with the target DNA substrate define a unique strand transfer conformation for integrase. *Proc Natl Acad Sci U S A*. 2000; 97:11244–11249. [PubMed: 11016953]
4. Grobler JA, Stillmock K, Hu B, Witmer M, Felock P, Espeseth AS, et al. Diketo acid inhibitor mechanism and HIV-1 integrase: implications for metal binding in the active site of phosphotransferase enzymes. *Proc Natl Acad Sci U S A*. 2002; 99:6661–6666. [PubMed: 11997448]
5. Hazuda DJ, Felock P, Witmer M, Wolfe A, Stillmock K, Grobler JA, et al. Inhibitors of strand transfer that prevent integration and inhibit HIV-1 replication in cells. *Science*. 2000; 287:646–650. [PubMed: 10649997]
6. Myers RE, Pillay D. Analysis of natural sequence variation and covariation in human immunodeficiency virus type 1 integrase. *J Virol*. 2008; 82:9228–9235. [PubMed: 18596095]
7. Cooper DA, Steigbigel RT, Gatell JM, Rockstroh JK, Katlama C, Yeni P, et al. Subgroup and resistance analyses of raltegravir for resistant HIV-1 infection. *N Engl J Med*. 2008; 359:355–365. [PubMed: 18650513]
8. Goethals O, Clayton R, Van Ginderen M, Vereycken I, Wagemans E, Geluykens P, et al. Resistance mutations in human immunodeficiency virus type 1 integrase selected with elvitegravir confer reduced susceptibility to a wide range of integrase inhibitors. *J Virol*. 2008; 82:10366–10374. [PubMed: 18715920]
9. Fransen S, Gupta S, Danovich R, Hazuda D, Miller M, Witmer M, et al. Loss of raltegravir susceptibility by human immunodeficiency virus type 1 is conferred via multiple nonoverlapping genetic pathways. *J Virol*. 2009; 83:11440–11446. [PubMed: 19759152]
10. Fransen S, Karmochkine M, Huang W, Weiss L, Petropoulos CJ, Charpentier C. Longitudinal analysis of raltegravir susceptibility and integrase replication capacity of human immunodeficiency virus type 1 during virologic failure. *Antimicrob Agents Chemother*. 2009; 53:4522–4524. [PubMed: 19667293]
11. Charpentier C, Karmochkine M, Laureillard D, Tisserand P, Belec L, Weiss L, et al. Drug resistance profiles for the HIV integrase gene in patients failing raltegravir salvage therapy. *HIV Med*. 2008; 9:765–770. [PubMed: 18651855]
12. Margulies M, Egholm M, Altman WE, Attiya S, Bader JS, Bemben LA, et al. Genome sequencing in microfabricated high-density picolitre reactors. *Nature*. 2005; 437:376–380. [PubMed: 16056220]
13. Zagordi O, Geyrhofer L, Roth V, Beerenwinkel N. Deep sequencing of a genetically heterogeneous sample: local haplotype reconstruction and read error correction. *J Comput Biol*. 2010; 17:417–428. [PubMed: 20377454]
14. Zagordi O, Klein R, Daumer M, Beerenwinkel N. Error correction of next-generation sequencing data and reliable estimation of HIV quasispecies. *Nucleic Acids Res*. 2010; 38:7400–7409. [PubMed: 20671025]
15. Wang C, Mitsuya Y, Gharizadeh B, Ronaghi M, Shafer RW. Characterization of mutation spectra with ultra-deep pyrosequencing: application to HIV-1 drug resistance. *Genome Res*. 2007; 17:1195–1201. [PubMed: 17600086]
16. Eriksson N, Pachter L, Mitsuya Y, Rhee SY, Wang C, Gharizadeh B, et al. Viral population estimation using pyrosequencing. *PLoS Comput Biol*. 2008; 4:e1000074. [PubMed: 18437230]
17. Simen BB, Simons JF, Hullsiek KH, Novak RM, Macarthur RD, Baxter JD, et al. Low-abundance drug-resistant viral variants in chronically HIV-infected, antiretroviral treatment-naïve patients significantly impact treatment outcomes. *J Infect Dis*. 2009; 199:693–701. [PubMed: 19210162]
18. Le T, Chiarella J, Simen BB, Hanczaruk B, Egholm M, Landry ML, et al. Low-abundance HIV drug-resistant viral variants in treatment-experienced persons correlate with historical antiretroviral use. *PLoS One*. 2009; 4:e6079. [PubMed: 19562031]
19. Johnson JA, Li JF, Wei X, Lipscomb J, Irlbeck D, Craig C, et al. Minority HIV-1 drug resistance mutations are present in antiretroviral treatment-naïve populations and associate with reduced treatment efficacy. *PLoS Med*. 2008; 5:e158. [PubMed: 18666824]

20. Hoffmann C, Minkah N, Leipzig J, Wang G, Arens MQ, Tebas P, et al. DNA bar coding and pyrosequencing to identify rare HIV drug resistance mutations. *Nucleic Acids Res.* 2007; 35:e91. [PubMed: 17576693]
21. Quince C, Lanzen A, Curtis TP, Davenport RJ, Hall N, Head IM, et al. Accurate determination of microbial diversity from 454 pyrosequencing data. *Nat Methods.* 2009; 6:639–641. [PubMed: 19668203]
22. Hasegawa N, Sugiura W, Shibata J, Matsuda M, Ren F, Tanaka H. Inferring within-patient HIV-1 evolutionary dynamics under anti-HIV therapy using serial virus samples with vSPA. *BMC Bioinformatics.* 2009; 10:360. [PubMed: 19863822]
23. Edgar RC. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.* 2004; 32:1792–1797. [PubMed: 15034147]
24. Yilmaz P, Kottmann R, Field D, Knight R, Cole JR, Amaral-Zettler L, et al. Minimum information about a marker gene sequence (MIMARKS) and minimum information about any (x) sequence (MIXS) specifications. *Nat Biotechnol.* 2011; 29:415–420. [PubMed: 21552244]
25. Huse SM, Huber JA, Morrison HG, Sogin ML, Welch DM. Accuracy and quality of massively parallel DNA pyrosequencing. *Genome Biol.* 2007; 8:R143. [PubMed: 17659080]
26. Codoner FM, Pou C, Thielen A, Garcia F, Delgado R, Dalmau D, et al. Dynamic escape of preexisting raltegravir-resistant HIV-1 from raltegravir selection pressure. *Antiviral Res.* 2010; 88:281–286. [PubMed: 20883724]

Fig. 1. Results of denoising control data using Pyronoise

RNA from HIV-1_{NL4-3} particles was purified, amplified by RT PCR, and analyzed by 454/Roche pyrosequencing. After filtering for error-prone reads, sequences were de-noised using Pyronoise, which preclusters reads based on light intensity values for each flow prior to interpretation as base calls. Alignments of a subset of reads are shown before denoising (a) and after denoising (b). The effects of denoising were also compared at each position in the HIV-1_{NL4-3} sequences studied (c and d). Shown are sequences from the forward reads. The base position is shown on the x-axis and the error rate is shown on the y-axis. Positions are assigned a homopolymer index of 1-5 (refer key) based on their location on a homopolymer. For example, a position with index 3 carries a base that occupies the third place on a homopolymer with a minimum length of 3. Results are compared before denoising (c) and after denoising (d). The error rates in the high-quality parts of the reads were 1.2×10^{-3} before denoising and 5×10^{-4} after denoising.

Fig. 2. Sequence analysis of HIV populations in three participants treated with raltegravir and undergoing pathway switches from N155H to Q148 + G140S

Nucleic acid sequences are shown for the three participants: (a) patient 1; (b) patient 2; and (c) patient 3. For economy of display, only codons with polymorphisms over the time course studied are shown. The display shows the sequence of each operational taxonomic unit (OTU) detected without reference to their abundance. The numbers to the left of each panel show the cumulative number of OTUs and the numbers to the right indicate the number of months since initiating therapy. A key to nucleic acid designations is shown at the bottom.

Fig. 3. Evolutionary network of mutations following raltegravir treatment inferred using viral serial pathway analysis

(a) Patient 1. (b) Patient 2. (c) Patient 3. The color of the cluster corresponds to the frequency of its membership from among sequences in that time-point (see color code). Viral serial pathway analysis (vSPA) reports all mutations relative to a baseline (in this case, month 0) ancestral cluster that arise at a level $2\times$ or more in a descendent cluster. Primary mutations are in red, accessory ones in black, and others in green (shown as the encoded amino acids). Mutations marked with an asterisk are not known to be associated with any primary drug resistance mutation (DRM). Clusters can contain one or many operational taxonomic units (OTUs). The proportion of OTUs within a cluster carrying a particular mutation is indicated in parenthesis.

Fig. 4. Analysis of possible drug resistance mutation substitutions present in participants prior to initiation of raltegravir therapy

(a) Analysis of reproducibility of operational taxonomic unit (OTU) recovery in duplicate analyses of the pretreatment samples. In this study, the pretreatment time-points were analyzed twice, once in the low depth first pass analysis and then again in our deeper re-sequencing study, allowing a comparison of the OTUs recovered. The x and y-axes show the relative proportion of each OTU in the two experiments. The red lines indicate the 5% abundance cut-off. To compare consistency, we asked for how many OTUs did the placement in less than 5% or more than 5% abundance cut-off groups agree (that is, in how many cases were samples in the lower left or upper right quadrants, indicating concordance). A total of 4050 OTUs showed concordance, or 99.7% of the total. (b) Table showing *P* values for the comparison of drug resistance mutations (DRMs) at 143, 148, and 155 in the pretreatment samples to error rate measured with the HIV-1_{NL4-3} sequence controls. Analysis was carried out both before and after denoising with Pyronoise.