PNAS
www.pnas.org

Supplementary Information for

Combined Near Full-Length HIV-1 Sequencing and Integration Site Analysis Informs Intra-Patient Viral Dynamics and Reconstruction of Replication-Competent Viral Ancestors

Sean C. Patro[1], Leah D. Brandt[2], Michael J. Bale[1], Elias K. Halvas[2], Kevin W. Joseph[2], Wei Shao[3], Xiaolin Wu[3], Shuang Guo[3], Ben Murrell[4], Ann Wiegand[1], Jonathan Spindler[1], Castle Raley[3], Christopher Hautman[3], Michele Sobolewski[2], Christine M. Fennessey[5], Wei-Shau Hu[1], Brian Luke[3], Jenna M. Hasson[1], Aurelie Niyongabo[1], Adam A. Capoferri[1], Brandon F. Keele[5], Jeff Milush[6], Rebecca Hoh[6], Steven G. Deeks[6], Frank Maldarelli[1], Stephen H. Hughes[1], John M. Coffin[7], Jason W. Rausch[8], John W. Mellors[2], Mary F. Kearney[1]

Co-corresponding Authors:

John M. Coffin
Department of Molecular Biology and Microbiology
Tufts University School of Medicine
Boston, MA 02111
617 636 6528
john.coffin@tufts.edu

Sean C. Patro
HIV Replication and Dynamics Program
National Cancer Institute at Frederick
1050 Boyles Street, Building 535, Room 134
Frederick, Maryland 21702
301 846 5215
sean.patro@nih.gov

www.pnas.org/cgi/doi/10.1073/pnas.1910334116

This PDF file includes:

Supplementary text
Figures S1 to S2
Legends for Figures S1 to S2
Tables S1 to S2

Supplementary Information Text

Supplemental Methods

Multiple Displacement Amplification

MDA conditions described here are modified from previously established protocols (1-4). Supplies and specialized reagents used for MDA reactions include phi29 DNA polymerase and 10x reaction buffer (New England BioLabs), trehalose (Sigma; T5251-10G), and Zymo-10 columns (Zymo Research; D4011). The random hexamer primer (p6N), purchased from Integrated DNA Technologies, contains phosphorothioate linkages at the ultimate and penultimate 3'-terminal positions, and is 5' phosphorylated (IDT synthesis code: /5phos/NNNN*N*N). 8x D-solution (3.2 M KOH, 80 mM EDTA), 1x N-solution (0.4 M Trizma-HCl). 0.8 M trehalose, 0.5 M KCl, and 10 mM dNTPs were prepared prior to dilution and dispensation of extracted cellular DNA into 96-well plates and stored at -20°C. 8x D-solution should be remade at least once every six months, stored in small aliquots and diluted to 1x immediately prior to use. For each MDA plate, extracted cellular DNA was diluted in 5ml Tris-HCL pH 8.0 to a concentration of 0.3 proviruses/2 µL solution as determined by the standard SGS endpoint, and 2 µL of this dilution transferred to each well of a 96 well plate. Genomic DNA in each sample was chemically denatured by addition of 2µL 1x D-solution followed by incubation at room temperature for 3-6 minutes. Reactions were neutralized by addition of 4µL 1x N-solution and immediately placed on ice to minimize partial renaturation of genomic DNA. The nucleic acid mixtures (NA-mixes) were supplemented with 2µL 1 nmol/µL p6N primer and kept on ice during preparation of the Enzyme Mixture (E mix). Sufficient E-mix was prepared to permit the addition of 30 µL to all NA-mixes, with each aliquot containing the following components: 4µL 10x phi29 DNA pol reaction buffer, 4µL 0.5 M KCl, 4µL 10 mM dNTPs, 10µL 0.8 M trehalose, 7µL water and 1 µL phi29 DNA polymerase. E-mixes were assembled and dispensed to NA-mixes on ice, and MDA reactions incubated at 30°C for 18hr in a thermal cycler utilizing a heated lid to minimize condensation. Completed reactions were incubated at 65°C for 10min to denature the polymerase and stored at -20°C, as necessary.

Integration sites assay (ISA)

Prior to ISA, MDA reactions were purified using Zymo10 DNA clean and concentration (Zymo; D4011). ISA was performed as previous described (5-8) using patient virus-specific primers to the 5' and 3' LTRs. All integration sites are reported using the 3' LTR mapped to hg19. The primer sequences are as follows:

Outer PCR Primers:
Generic
5'LTR: None
3'LTR: TGTGACTCTGGTAACTAGAGATCCCTC

Patient 1
5’LTR: TCAGGGAAGTAGCCTTGTGTGTGGT
3’LTR: TGTGACTCTGGTAACTAGAGATCCCTC

1683
5’LTR: CTGATCCCTGGCCCTGGTGTGT
3’LTR: TGTGACTCTGGTACTAGAGATCCCTC

2669
5’LTR: TCAGGGAAGTAGCCTTGTGTGTGGT
3’LTR: TGTGACTCTGGTAACTAGAGATCCCTC

3162
5’LTR: CTAGTCCCTGGCCCTGGTGTGT
3’LTR: GTGACTCTGGTAACTAGAGATCCCTC

R-09
5’LTR: AAGTAGCCTTGTGTGTGGTAGACC
3’LTR: AGAGATCCCTCAGACCATTTTAGT

Nested PCR Primers:
Patient 1
5’LTR:
AATGATACGGCGACCACCGAGATCTACACTATAGCCTACACTCTTTCCCTACACGACGCTCTTCC
GATCTNNNNNNACGTCTCGTCCTGTCTTTTCTGGGAGTGAACTA
3’LTR:
AATGATACGGCGACCACCGAGATCTACACATAGAGGCACACTCTTTCCCTACACGACGCTCTTCC
GATCTNNNNNNATAGAGGCCCCTTTTAGTCAGTGTGGAAAATC

1683
5’LTR:
AATGATACGGCGACCACCGAGATCTACACTATAGCCTACACTCTTTCCCTACACGACGCTCTTCC
GATCTNNNNNNATATACACTCTTGCCTTTGCTGGGAGTAAATTA
3’LTR:
AATGATACGGCGACCACCGAGATCTACACTATAGCCTACACTCTTTCCCTACACGACGCTCTTCC
GATCTNNNNNNGACACTGACCCTTTTAGTCAGTGTGAAAATC

2669
5’LTR:
AATGATACGGCGACCACCGAGATCTACACTAATCTTAACACTCTTTCCCTACACGACGCTCTTCC
GATCTNNNNNNCGTTACTACTTGTCTTTTCTGGGAGTGAATTA

3’LTR:
AATGATACGGCGACCACCGAGATCTACACGGCTCTGAACACTCTTTCCCTACACGACGCTCTTCC
GATCTNNNNNNCTGCGTGTCCCTTTTAGTCAGTGTGGAAAATC

3162
5’LTR:
AATGATACGGCGACCACCGAGATCTACACGGCTCTGAACACTCTTTCCCTACACGACGCTCTTCC
GATCTNNNNNNCTGCGTGTTCTTGTCTTTTCTGGGACCAAACTA
3’LTR:
AATGATACGGCGACCACCGAGATCTACACCCTATCCTACACTCTTTCCCTACACGACGCTCTTCC
GATCTNNNNNNTAGCGAGTCAATTTAGTCAGTGTGGAAAATC

R-09
5’LTR:
AATGATACGGCGACCACCGAGATCTACACCAGGACGTACACTCTTTCCCTACACGACGCTCTTCC
GATCTNNNNNNAGAGTCACCTTGTCTCTTTTGGGAGTAAATTA
3’LTR:
AATGATACGGCGACCACCGAGATCTACACCCTATCCTACACTCTTTCCCTACACGACGCTCTTCC
GATCTNNNNNNTAGCGAGTCATTTTAGTCAGTGTAGAAAATC

Near-Full-Length Sanger Sequencing from MDA wells

Near-full-length (NFL) amplification was performed with 2 µL of a 1:8 dilution of input DNA from an MDA well. PCR reactions (10 µL) were performed using the Ranger enzyme (BIOLINE-BIO-25052). PCR1, denoted NFL1, was used to generate a 9kb amplicon from U5-5’LTR to U5-3’LTR (primers in Table S2). NFL1 product, diluted 1:5, was used as template for nested PCR to generate four 2.9-3.2kb genome-spanning amplicons, denoted F1, F2, F3, and F4 (Figure S2). Alternatively, a 9kb NFL2 amplicon was generated. Target amplicons sizes were confirmed by ethidium bromide gel visualization and sequenced by Sanger sequencing or PacBio sequencing. The resulting data were used to generate the sequence of an NFL genome.

Near-Full-Length PacBio Sequencing from MDA wells

Amplicon DNA generated from PCR was purified using PB AMPure beads (Pacific Biosciences). Quality, concentration, and size were determined using the Nanodrop 8000 Spectrophotometer (Thermo Scientific) and the BioAnalyzer 2100 (Agilent). Each set of samples was combined into an equimolar pool and subjected to library preparation following the PacBio standard 2 kb protocol, using a-tailing and overhang adapter ligation instead of blunt ligation. Each completed pool was then loaded onto a SMRT Cell for sequencing. The libraries were sequenced on either the PacBio RSII (Pacific Biosciences) using P6/C4 chemistry, a 0.15 nM MagBead loading concentration,

and 360-minute movie lengths, or on the PacBio Sequel (Pacific Biosciences) using the diffusion method, v2.0 or v2.1 chemistry, a 4 pM loading concentration, and 600 minute movie lengths, both located at the Frederick National Laboratory for Cancer Research (Frederick, MD, USA).

Quantitative Viral Outgrowth Assay on PBMC from donor R-09

Quantitative viral outgrowth assays (QVOA) were performed, as described previously (9), with the following modifications: CD4+ cells were isolated from Cryopreserved PBMC by negative selection (Stemcell Technologies), serially diluted 3-fold from 1,000,000 cells per well to 30,000 cells per well, and seeded in individual wells in 6 replicates. Cells were stimulated with PHA overnight and a 10-fold excess of allogeneic irradiated feeder cells obtained from uninfected donors was added. CD4+ lymphoblasts from HIV-1-negative donors were added, and virus co-cultures were carried out for 3 weeks. Supernatants were tested weekly to measure HIV-1 p24 antigen by ELISA (Perkin Elmer). Aliquots of the supernatants were also frozen for sequencing the recovered viruses. The frequency of HIV- 1 infected cells carrying a recoverable infectious provirus (IUPM) was determined by a maximum likelihood method, as previously described (10).

Pipeline for analysis of HIV single-genome PacBio Sequencing

Fastq files containing PacBio CCS reads were processed using the Robust Amplicon Denoising algorithm (11), which attempts to extract error-free variants from complex amplicon mixtures. These denoised variants were mapped, using USEARCH (12) to an HIV-1 reference sequence (consensus B) to retain only HIV-1 reads, which were then manually stitched together (merging overlapping ends) to construct near-full-length contigs.

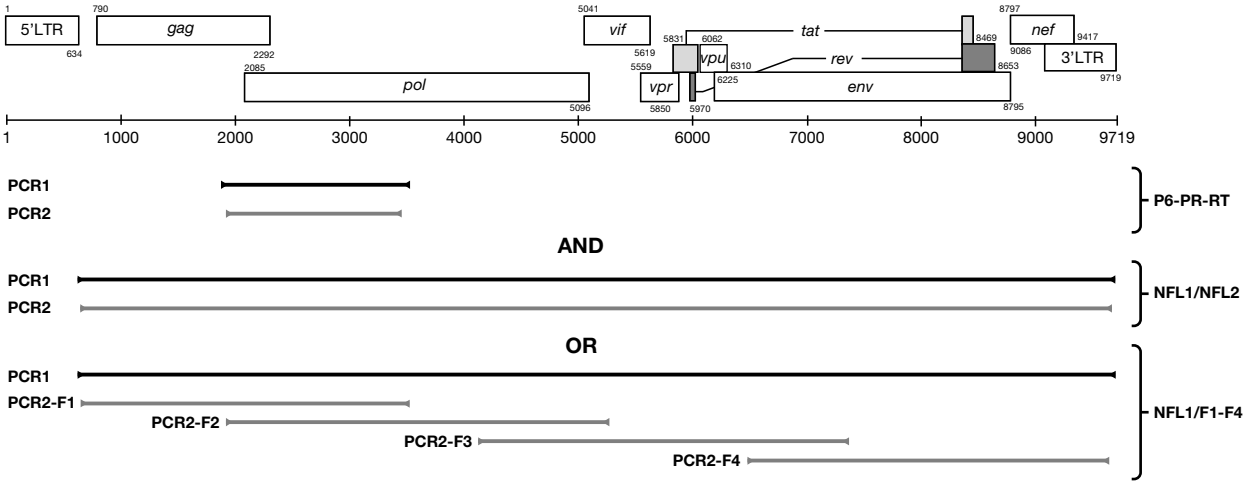

Figure S1. Near-Full-Length (NFL) Amplification and Sequencing. PCR amplification strategies used to screen MDA wells and to amplify and sequence near-full-length proviruses by either amplifying as one (9kb) or four (3kb) overlapping amplicons.

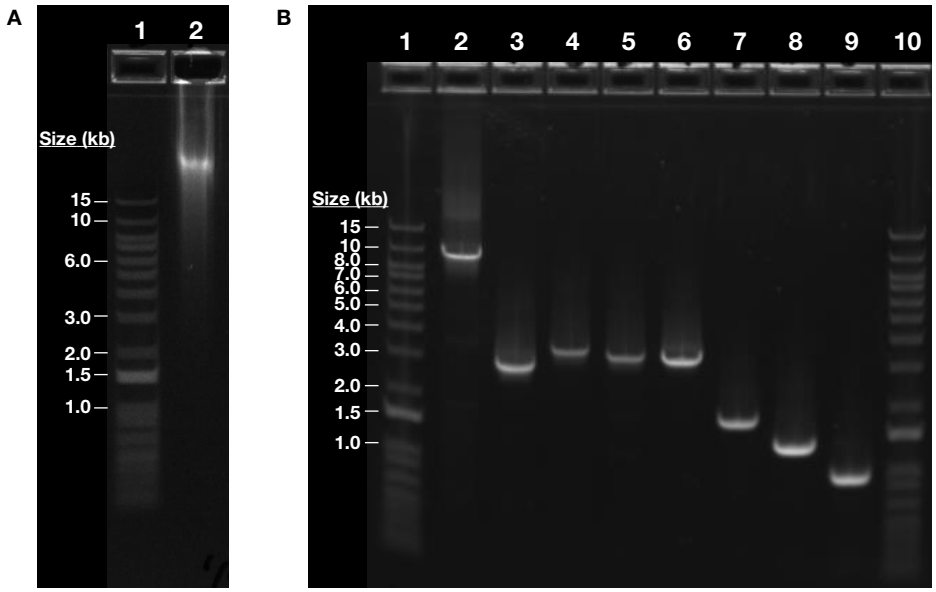

Figure S2. MDA and NFL Amplification Products. (A) MDA product (lane 2) fractionated on 0.8% agarose eGel (Invitrogen) with E-Gel 1 Kb Plus Ladder, Invitrogen (lane 1). (B) AMBI-1 near-full-length PCR product (lane 2). PCR products from fragment 1 (bp 634-3528), fragment 2 (bp 1871-5267), fragment 3 (bp 4133-7349), and fragment 4 (bp 6445-9643) (lanes 3-6). PCR products from P6-PR-RT (bp 1871-3438), integrase (bp 4133-5267), and partial envelope (bp 6445-7349) (lanes 7-9).

Table S1. Integration Sites Not Within Groups of Identical P6-PR-RT Sequences

PID	Integration site (hg19)	Chromosome	Gene	Provirus orientation relative to gene
1	32540529	6	HLA-DRB1	intergenic
	47990522	21	DIP2A	intergenic
1683	8525730	1	RERE	intergenic
	114652168	2	ACTR3	against
	128513393	9	PBX3	against
	11883248	11	USP47	with
	44856659	15	SPG11	against
	67070298	16	CBFB	with
	123070995[a], 104667510[a]	6, 7	PKIB, MLL5	intergenic, with
2669	151614674	1	SNX27	with
	57646144	3	FAM116A	against
	128593815	9	PBX3	with
	6692768	12	CHD4	against
	59361732	15	RNF111	against
	153393371	X	OPN1LW	intergenic
	139905034[a], 100562724[a]	5, 14	ANKHD1, EVL	against, against
3162	40419651	17	STAT5B	with
	30600340[a], 53712797[a]	6, 12	ATAT1, AAAS	against, with

[a]Mixture of two proviruses within single MDA well

Table S2. Primers for NFL PCR amplification and Sanger sequencing

Amplification Primers		
Name	**PCR Reaction**	**Primer Sequence**
611(+)	NFL1	AGTCAGTGTGGAAAATCTCT*A*G
9675(-)	NFL1	GAGGGATCTCTAGTTACCAG*A*G
618(+)	NFL2	GTGGAAAATCTCTAGCAGT*G*G
9662(-)	NFL2	TTACCAGAGTCACACAACAG*A*C
634(+)	F1	AGTGGCGCCCGAACAGGGAC
3500(-)	F1	CTATTAAGTATTTTGATGGGTCATAA
1870(+)	F2	GAGTTTTGGCTGAGGCAATGAG
5248(-)	F2	TCTCCTGTATGCAGACCCCA
4133(+)	F3	GGAAAAGGTCTACCTGGCATG
E125(-)	F3	CAATTTCTGGGTCCCCTCCTGAGG
E30(+)	F4	GTGTACCCACAGACCCCAGCCCACAAG
R-519(-)	F4	GCACTCAAGGCAAGCTTTATTGAGGCTTA
Amplification Primers		
Name	**PCR Reaction**	**Primer Sequence**
3996(+)	WIPF1 3' 1	CATCTAGCTTTGCAGGATTCG
WIPF1 +329(-)	WIPF1 3' 1	TGACAGAAAGTTAATCGGTGAACG
4133(+)	WIPF1 3' 2	GGAAAAGGTCTACCTGGCATG
WIPF1 +199(-)	WIPF1 3' 2	CTTCCTTGGTAAGCCTTTCCCAA
1849(+)	XPO6 3' 1	GATGACAGCATGTCAGGGAG
XPO +371(-)	XPO6 3' 1	GAGGTCCTCAAAACACTTTGTAAG
1870(+)	XPO6 3' 2	GAGTTTTGGCTGAGGCAATGAG
XPO +279(-)	XPO6 3' 2	CTGTGATGAACCCATTACATATTAACAG
1849(+)	TMCO5A 3' 1	GATGACAGCATGTCAGGGAG
TMC +252(-)	TMCO5A 3' 1	GCTAGAATGACTCACAGAACTCAC
1870(+)	TMCO5A 3' 2	GAGTTTTGGCTGAGGCAATGAG
TMC +177(-)	TMCO5A 3' 2	GATTAAATCAGCCAAGGGAAGAACAC
Sequencing Primers		
Name	**PCR Reaction**	**Primer Sequence**
1061(+) For1	Seq	GGATAGAGGTAAAAGACACCAA
1488(+) For2	Seq	AAGTGACATAGCAGGAACTACTAG
2012(+) For3	Seq	CTAGGAAAAAGGGCTGTTGGAAATG
2385(+) For4	Seq	AAAATGATAGGGGGAATTGGAGGTTT
2869(+) For5	Seq	CAGTACTAGATGTGGGGGATGCATA
3298(+) For6	Seq	ACAGCTGGACTGTCAATGATATACA
3676(+) For7	Seq	CCACAGAAAGCATAGTAATATGGGG
4162(+) For8	Seq	CACACAAAGGGATTGGAGGAAATGA
6552(+) For14	Seq	TATGGGACCAAAGCCTAAAGCCATGTG
6961(+) For15	Seq	CAGCACAGTACAATGTACACATGGAA
7807(+) For17	Seq	AGCAGCAGGAAGCACTATGGGCGC
8257(+) For18	Seq	CATATCAAATTGGCTGTGGTATAT
8517(+) For19	Seq	GGAACCTGTGCCTCTTCAGCTACC
9021(+) For20	Seq	CCACAGGTACCTTTAAGACCAATGAC
895(-) Rev1	Seq	AATTTTTCCAGCTCCCTGCTTGCCCA
1294(-) Rev2	Seq	CTGATAATGCTGAAAACATGGGTAT
1606(-) Rev3	Seq	GACAGGGCTATACATTCTTACTAT
2557(-) Rev5	Seq	TTACTGGTACAGTTTCAATAGGAC
5451(-) Rev12	Seq	AGAGATCCTACCTTGTTATGTCCT
5955(-) Rev13	Seq	CTTCCTGCCATAGGAGATGCCTAAG
6497(-) Rev14	Seq	ACCATGTTATTTTTCCACATGTTAAA
7001(-) Rev15	Seq	CTGCCATTTAACAGCAGTTGAGTTGA
7531(-) Rev16	Seq	ATGGGAGGGGCATACATTGCT
7943(-) Rev17	Seq	CCTGGAGCTGTTTAATGCCCCAGAC
8352(-) Rev18	Seq	GGTGAGTATCCCTGCCTAACTCTAT

* Phosphorothioate bond

References

1. Dean FB, Nelson JR, Giesler TL, & Lasken RS (2001) Rapid amplification of plasmid and phage DNA using Phi 29 DNA polymerase and multiply-primed rolling circle amplification. *Genome Res* 11(6):1095-1099.
2. Dean FB, Hosono S, Fang L, Wu X, Faruqi AF, Bray-Ward P, Sun Z, Zong Q, Du Y, Du J, Driscoll M, Song W, Kingsmore SF, Egholm M, & Lasken RS (2002) Comprehensive human genome amplification using multiple displacement amplification. *Proceedings of the National Academy of Sciences of the United States of America* 99(8):5261-5266.
3. Pinard R, de Winter A, Sarkis GJ, Gerstein MB, Tartaro KR, Plant RN, Egholm M, Rothberg JM, & Leamon JH (2006) Assessment of whole genome amplification-induced bias through high-throughput, massively parallel whole genome sequencing. *BMC Genomics* 7:216.
4. Pan X, Urban AE, Palejev D, Schulz V, Grubert F, Hu Y, Snyder M, & Weissman SM (2008) A procedure for highly specific, sensitive, and unbiased whole-genome amplification. *Proceedings of the National Academy of Sciences of the United States of America* 105(40):15499-15504.
5. Maldarelli F, Wu X, Su L, Simonetti FR, Shao W, Hill S, Spindler J, Ferris AL, Mellors JW, Kearney MF, Coffin JM, & Hughes SH (2014) HIV latency. Specific HIV integration sites are linked to clonal expansion and persistence of infected cells. *Science* 345(6193):179-183.
6. Simonetti FR, Sobolewski MD, Fyne E, Shao W, Spindler J, Hattori J, Anderson EM, Watters SA, Hill S, Wu X, Wells D, Su L, Luke BT, Halvas EK, Besson G, Penrose KJ, Yang Z, Kwan RW, Van Waes C, Uldrick T, Citrin DE, Kovacs J, Polis MA, Rehm CA, Gorelick R, Piatak M, Keele BF, Kearney MF, Coffin JM, Hughes SH, Mellors JW, & Maldarelli F (2016) Clonally expanded CD4+ T cells can produce infectious HIV-1 in vivo. *Proceedings of the National Academy of Sciences of the United States of America* 113(7):1883-1888.
7. Shao W, Shan J, Kearney MF, Wu X, Maldarelli F, Mellors JW, Luke B, Coffin JM, & Hughes SH (2016) Retrovirus Integration Database (RID): a public database for retroviral insertion sites into host genomes. *Retrovirology* 13(1):47.
8. McManus WR, Bale MJ, Spindler J, Wiegand A, Musick A, Patro SC, Sobolewski MD, Musick VK, Anderson EM, Cyktor JC, Halvas EK, Shao W, Wells D, Wu X, Keele BF, Milush JM, Hoh R, Mellors JW, Hughes SH, Deeks SG, Coffin JM, & Kearney MF (2019) HIV-1 in lymph nodes is maintained by cellular proliferation during antiretroviral therapy. *J Clin Invest* 130.
9. Siliciano JD & Siliciano RF (2005) Enhanced culture assay for detection and quantitation of latently infected, resting CD4+ T-cells carrying replication-competent virus in HIV-1-infected individuals. *Methods Mol Biol* 304:3-15.
10. Rosenbloom DI, Elliott O, Hill AL, Henrich TJ, Siliciano JM, & Siliciano RF (2015) Designing and Interpreting Limiting Dilution Assays: General Principles and Applications to the Latent Reservoir for Human Immunodeficiency Virus-1. *Open Forum Infect Dis* 2(4):ofv123.

11. Kumar V, Vollbrecht T, Chernyshev M, Mohan S, Hanst B, Bavafa N, Lorenzo A, Ketteringham R, Eren K, Golden M, Oliveira MF, & Murrell B (2018) Long-read amplicon denoising. *bioRxiv*:383794.
12. Edgar RC (2010) Search and clustering orders of magnitude faster than BLAST. *Bioinformatics* 26(19):2460-2461.