Contamination of genomic databases by HIV-1 and its possible consequences. A study in Bioinformatics.

Miguel Romero Fernández-Bravo[†]
17 March 2014

Summary
The bioinformatics Basic Local Alignment Search Tools (BLASTN and TBLASTN) were used to search public databases for nucleic acid (NA) and amino acid (AA) sequences identical or similar to HIV-1 DNA and proteins. Several significant alignments were detected in a variety of non-HIV-1 taxa and other sources deposited in public genomic and protein repositories. Homologies between a number of HIV-1 proteins and those of *Candida*, *Cryptococcus* and *Schistosoma mansoni* are of uncertain significance and suggest the need for further analyses. The overwhelmingly likely cause for these data is contamination.

Introduction
During an exercise into the design of a diagnostic HIV-1 DNA microarray a preliminary literature search revealed reports of HIV-1 RNA in HIV-1 negative individuals not at risk of AIDS.[1,2] It was also noted that in adolescents and adults the US Centers for Disease Control and Prevention recommends "For HIV screening, HIV virologic [HIV-1 NA] tests should not be used in lieu of approved HIV antibody screening tests".[3,4] These data raise the prospect of diagnostic ambiguity – a possibility given further impetus by reports of HIV-1 group M subtype B DNA in several common malignancies – breast, gynaecological and prostate, excised from non-HIV-infected patients.[5,6] This prompted a database search for HIV-1 sequences using the Basic Local Alignment Search Tool (BLAST) algorithms.

Background
BLAST is an extensively researched and widely used bioinformatics software tool that calculates the degree of relatedness between the nucleotides of DNA sequences or the amino-acid sequences of different proteins. The sequences compared are referred to as the query and subject sequences. Of the several BLAST algorithms the

[†] MEng, BEng, DipHE, PgD in Bioinformatics. Student of MEng. in Telecommunication Engineering, Universitat Oberta de Catalunya, Barcelona, Spain.

This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License.

two used in this investigation were BLASTN and TBLASTN. BLASTN directly compares two NA sequences while TBLASTN first translates the subject DNA into an AA sequence (protein) in all six reading frames and then compares this with the query AA sequence. TBLASTN has improved sensitivity and biological significance compared to BLASTN.

Briefly BLAST generates and compares substrings of the query sequence with substrings of the subject sequence. Comparisons that exceed a certain threshold score ("hits") are ordered according to length and significance and returned in the form of a graph, table or set of alignments.[7] These document the:

1. Cover (percent of the query sequence that is aligned to the subject sequence).
2. Identities (number and percent exact matches between query and subject over the length of the coverage).
3. Similarities (number and percent exact matches plus conservative substitutions for AA) for the most similar regions. Conservative substitutions refers to AA with similar physiochemical properties, for example, leucine and isoleucine.
4. Expect value E. The E-value calculation is "a method to assess the statistical significance of the alignment" and is required to determine biologically relevant alignments[8] and separate these from "random background noise".[9] The E-value is the number of matches with the same score expected to occur by chance within a given database. For example, an E-value of 10^{-6} equates to one match expected by chance for every 1,000,000 entries in the database. The lower the E-value or the closer it is to zero the more significant the match.

Nucleotides alignments with E-value $<10^{-6}$ and identity >=70% are considered significant. Protein sequences more than 100 AA in length with alignment E-values $<10^{-3}$ and identity >=25% are also significant[10] and suggest homology. Homologous proteins "probably have the same ancestor, share the same structure, and have a similar biological function".[8] Homology is not to be confused with similarity or identity. Homology, although grounded on these variables, is either present or absent. For example, it is incorrect to state two proteins are 90% homologous.

Methods

Initially the 9719 base pair (bp) "Human immunodeficiency virus type 1 (HXB2), complete genome; HIV1/HTLV-III/LAV reference genome" (Accession Number K03455.1) was BLASTN tested against the *National Center for Biotechnology Information* (*NCBI*) and *European Nucleotide Archive* (*ENA*) DNA libraries and linked libraries. Because of the 10% and usually far greater variation in HIV-1 DNAs, all the alignments returned with HXB2 were retested (and confirmed) against a second full-length HIV-1 genome, the 9181 bp "Human immunodeficiency virus 1, complete genome" (Accession Number NC_001802.1). However searches using full-length HIV-1 genomes soon proved impractical because they form multiple alignments with

the thousands of HIV-1 DNA sequences deposited in the *NCBI*, *ENA* and affiliated databases. Despite the extensive filtering offered by the BLAST software it was not possible to remove these alignments. Fortunately the scientists at the Los Alamos National Laboratory maintain an extensive HIV database from which they have constructed a protein map of HIV-1 based on the individual, HXB2 NA coding sequences (cds, genes) for the HIV-1 proteins.[11] These sequences were substituted as query sequences and included those for the p17, p24, p32, p51, p66, gp41, gp120 and gp160 proteins. Each sequence was tested against the same libraries and the alignments returned then tested against the full-length HIV-1 sequences. Further subject sequences were obtained following BLASTN searches using three sets of 20-30 bp HIV-1 PCR primers[12-14] and the sequences returned similarly tested.

Because the genetic code is redundant, an AA sequence "hides" alternative NA coding sequences for any particular protein. (For example, there are many alignments between the proteins of both AIDS causing retroviruses HIV-1 and HIV-2 although their DNAs do not align). Since a principal *raison d'être* for NA sequences is the assembly of proteins, in addition to NA searches a limited number of translated BLASTN searches were performed over a range of taxa and other database AA sequences using the HXB2 NA protein coding regions as subject sequences.

In lieu of controls in the conventional sense of experimental science alignment searches against HIV-1 DNA were conducted using the genomes of several RNA viruses of similar length to the HIV-1 genome. Those investigated were polio, rubella, mumps, measles, hepatitis B, hepatitis C, Coxsakie and human and porcine enterovirus species.

Results

Tables 1, 2, and 3 document significant similarities between HIV-1 DNA sequences and a diverse set non-HIV-1 DNAs. These include (in alphabetical order) *Bienertia sinuspersici* (a plant belonging to the Amaranth family), *Bombus insularis* Contig1484.Boin (bees), *Camellia sinensis* (the camellia plant), Candidate division TM7 single-cell TM7a_contig_2274, (an uncultured bacterial phylum from the human subgingival crevice), *Ctenogobiops feroculus* microsatellite DNA (the fierce shrimpgoby coral reef fish), *Culex pipiens quinquefasciatus* (the southern house mosquito and vector of lymphatic filariasis and avian malaria), *Dirofilaria immitis* (a parasitic heartworm spread by mosquitos), *Homo sapiens neanderthalensis* ("a 38,000-year-old Neanderthal fossil that is exceptionally free of contamination from modern human DNA"[15]), *Homo sapiens* chromosome 8, a human genomic sequence flanking an endonuclease *Not***I** site, *Human T-lymphotropic virus* I, *Leishmania major* (a protozoan pathogen and cause of cutaneous leishmaniasis), *Locusta migratoria* (the migratory locust), *Methyloversatilis universalis* (a betaproteobacterium), *Mus musculus* (the house mouse), *Neurospora intermedia* (an *Ascomycete* fungus used in the production of the food staple oncom), *Nicotiana tabacum* (the tobacco plant), *Oryza sativa Indica* (long-grained rice), *Phaseolus coccineus* (runner bean),

Plasmodium vivax (a protozoal parasite and cause of benign tertian malaria), the RAK alpha gene (present in breast and other human cancers), *Reticulitermes flavipes* (eastern subterranean termite) gut metagenome, *Sesamum indicum* (the sesame plant), *Setaria italica* (the foxtail millet plant), *Sorghum bicolor* (sorghum), *Streptomyces* species (actinobacteria), an uncultured fungus ribosomal RNA gene and *Zea mays* (maize).

Table 4 shows alignments returned from TBLASTN searches using HIV-1 proteins against translated NA sequences of *Cryptococcus neoformans* (yeast), *Candida albicans* (yeast), *Schistosoma mansoni* (a human trematode parasite), the human genome *Not*I flanking site, an uncultured fungus sequence and *Homo sapiens* chromosomes 3 and 7.

Table 5 shows alignments between HXB2 NA sequences and *Brucella melitensis* and *Ceratocystis fimbriata* (a fungal plant pathogen). The accession numbers of these organisms have been recently removed by NCBI.

No alignments were returned with the several RNA virus "control" genomes tested.

Discussion

It is well known that HIV genomes are extraordinarily diverse.[16] In fact such diversity necessitates the construction of consensus sequences. "A consensus sequence is a sequence of the most common nucleotide or amino acid at each position in an alignment. We [the Los Alamos National Laboratory] generally use a 50% cut-off, such that at least 50% of the sequences have the same character at this position, or else we replace the character with a question mark…Another way to create a consensus is to take the most frequently occurring character, even if it is not the majority".[17] Given the background for this study it was decided to search for alignments using the HXB2 consensus sequence. This is the standard reference strain HIV-1/HTLV-III/LAV[18] dating from the first HIV-1 isolates.[19]

Although the NA alignments reported here relate to only a tiny fraction of all published DNA sequences, at first sight they and their diversity are startling. Examples are:

1. A 201 bp segment of *Streptomyces* DNA bearing a 99% (198/199) alignment with HIV-1 DNA between base pairs 9521-9719. This segment is a partial cds for the HIV-1 nef protein.
2. The same 201 bp bearing a 99% (200/201) alignment between base pairs 436-636. This segment represents approximately 30% of the HXB2 5' end LTR.
3. A 675 bp segment of *Culex pipens* (mosquito) DNA bearing a 97% (655/675) alignment with HIV-1 DNA between base pairs 1-669 (12 non-identical NA and 8 gaps) This segment includes the complete coding region for HIV-1 p17 and a partial cds for the Pr55 *gag* precursor protein. (Interestingly in 1986 "HIV

proviral DNA was reported in various insects [including mosquitos] from Central Africa"[20]).

4. A 510 bp segment of human DNA with a (89%) (453/510) alignment with HIV-1 DNA between base pairs 3135-3641. This segment is a cds for the HIV-1 Pol polyprotein p66/p51 (reverse transcriptase) region.
5. A 67 bp segment of human chromosome 8 DNA that has a 99% (66/67) alignment with HIV-1 DNA between base pairs 6348-6414. This segment is a partial cds for HIV-1 proteins p120 and p160.

However if one is cognisant of the *National Cancer Institute* report that "almost one-fifth of nonprimate genomes in public databases contain stretches of human DNA sequence",[21] these findings may not be totally unexpected. Given the prevalence of HIV-1 nucleic acid testing, as well as the sensitivity of PCR, a similar situation may operate. Contamination, an ever present potential hazard may result from amplification of a mere few million sequences introduced for example from a nearby laboratory on a seemingly innocent fomite. There is evidence that varicella-zoster virus DNA[22] and *Pneumocystis carinii* DNA[23] can be found in air samples in a high proportion of hospital wards including rooms that do not house patients. It is conceivable that HIV RNA could likewise become airborne or transferred to a laboratory by a person who has come into contact directly or indirectly with a patient, health care provider or another person in a hospital setting. Contamination may be introduced at several places including at source (during isolation of taxon or other DNA) or the PCR or sequencing steps (during which DNAs from different sources may combine as chimeras). The results in Table 5 with *Brucella melitensis* and *Ceratocystis fimbriata* are considered definite evidence of contamination particularly as the coverage between *Brucella* and HIV-1 exceeded 5800 base pairs and no such alignments have been deposited in other *Brucella melitensis* accessions. The recent removal of the Table 5 Accession numbers by the NCBI appear to confirm this view.

To guard against contamination the NCBI screens each submitted genome to identify contaminating sequences that may be present for artificial or biological reasons. According to the NCBI, screening consists of BLAST searches of submitted sequences against the chromosomes of unrelated organisms, against primate- or rodent-specific repeats and against an enhanced common contaminants database that contains vector sequences, bacterial insertion sequences and *E. coli*.[24] Unfortunately NCBI does not include HIV sequences in its quality control screening.

Notwithstanding, there are several reasons why contamination may not be the universal explanation for these data. They include (a) there is no actual proof of contamination; (b) sequences are deposited by laboratories conducting research unrelated to HIV; (c) some sequences are reported by laboratories at foremost institutions; (d) laboratories undertake forensic precautions to exclude and neutralise contamination; (e) no HIV-1 DNA alignments were found in the nine sets of "control",

non-HIV-1 RNA virus genomes similar in length to the HIV-1 genome (data not shown); (f) using the HIV-1 gp41-derived primers SK68 and SK69 HIV-1 sequences were reported in malignant tissues of patients in the absence of HIV-1 infection.[25] In the latter contamination is virtually impossible since the same HIV-1 DNA was not detected in any of the non-malignant control tissues adjacent to the excised neoplasms.

Several of the AA alignments in Table 4 are also worthy of mention. For example, the 93% and 100% identities between HXB2 envelope proteins and an 856 bp "uncultured fungus" translated DNA from "composting dairy manure" reported in 2013 by the *Plant Sciences Division*, University of Idaho. And the 68% and 73% identities between HXB2 p66 and p51 (reverse transcriptase) and a 597 bp human genomic sequence flanking a *Not***I** site reported in 2001 by the *Microbiology and Tumorbiology Centre*, Karolinska Institute.[26] There are also alignments signifying homology between HXB2 and several human pathogens including the AIDS-related *Candida* and *Cryptococcus*. The significance of these alignments is uncertain but since both organisms are globally prevalent and exposure often occurs in early childhood they may underlie some instances of unconfirmed reactive HIV-1 ELISAs and/or cross-reacting Western blot bands. Similar remarks may apply to *Schistosoma mansoni* in Africa and other parts of the world where Schistosomiasis is endemic.[27]

Conclusions

Contamination is the most likely explanation for these data. If HIV nucleic acid contamination is a prevalent risk similar to that of human DNA then there is a possibility for HIV nucleic acid to contaminate HIV testing facilities. This has the potential to complicate direct virologic diagnosis and the monitoring of HIV infection. It would prove especially problematic when least desired, for example, when testing is mandated to clarify indeterminate serology in low risk individuals or complement the screening of serologically negative blood donations. It is hoped this report will stimulate other students of bioinformatics to confirm these data and further elucidate the extent to which HIV-1 NA and AA sequences are related to the DNA and proteins of humans and other taxa. If these data are not wholly explained by contamination then students or others more familiar with the biological sciences may be in a position to propose an alternative explanation.

Table 1: HIV-1 nucleic acid BLASTN identity search against full-length HXB2 (9719 bp)

Subject (bp)	Cover	Identity	E-Value	Data
Plasmodium vivax (various sequences and lengths)	–	97%	2e-6	Link 0
Streptomyces sp. AA4 (8475969)	5%	99%	3e-100	Link 1
Methyloversatilis universalis FAM5 Contig00007 (779)	100%	96%	0.0	Link 2
Methyloversatilis universalis FAM5 Contig00009 (1122)	100%	89%	0.0	Link 3
Methyloversatilis universalis FAM5 Contig00006 (649)	99%	91%	0.0	Link 4
Uncultured fungus (856)	52%	96%	0.0	Link 5
Oryza sativa cv. LYP9 (432)	96%	98%	0.0	Link 6
Neurospora intermedia (various sequences and lengths)	–	99%	1e-91	Link 7
Mus musculus (148)	46%	100%	3e-33	Link 8
Setaria italica (720)	91%	77%	3e-99	Link 9
Bienertia sinuspersici (1216)	12%	84%	2e-37	Link 10
Homo sapiens chromosome 8 (125)	53%	99%	1e-30	Link 11
Human T-lymphotropic virus I (404)	97%	88%	5e-134	Link 12
Locusta migratoria (462)	92%	98%	0.0	Link 13
Phaseolus coccineus (107)	90%	93%	2e-97	Link 14
Nicotiana tabacum (383)	80%	90%	2e-113	Link 15
Camellia sinensis (883)	100%	96%	0.0	Link 16
Bombus insularis Contig1484.Boin (232)	99%	93%	3e-94	Link 17
Homo sapiens genomic sequence flanking *Not*I site *(597)*	84%	89%	0.0	Link 18
Homo sapiens neanderthalensis fossil long bone DSASCWG01DMKL3 (70)	91%	94%	9e-25	Link 19
Homo sapiens neanderthalensis fossil long bone DSASCWG01CJM12 (114)	56%	98%	6e-29	Link 20
Homo sapiens neanderthalensis fossil long bone DSASCWG01BK7EM (105)	57%	100%	2e-28	Link 21
Homo sapiens neanderthalensis fossil long bone *DSASCWG02IG82R* (94)	64%	97%	3e-26	Link 22

Sequence from: http://www.hiv.lanl.gov/content/sequence/HIV/REVIEWS/HXB2.html

Table 2. HIV-1 nucleic acid BLASTN identity search against primer derived subject sequences I

Primer	Subject (bp)	Cover	Identity	E.Value	Data
JA17	*Ctenogobiops feroculus* microsatellite (85)	100%	96%	2e-34	Link 23
JA20	*Zea mays* (330)	95%	96%	9e-146	Link 24
JA18	*Zea mays* (330)	95%	96%	9e-146	Link 24
JA13	*Dirofilaria immitis* (2631)	100%	91%	0.0	Link 25
SK68	RAK gene alpha (142)	100%	95%	3e-62	Link 26
	Dirofilaria immitis (2631)	100%	91%	0.0	Link 25
SK69	*Dirofilaria immitis* (2631)	100%	91%	0.0	Link 25
Msf12b	*Culex pipiens quinquefasciatus* (934)	72%	97%	0.0	Link 27
F2nst	*Sesamum indicum* (770)	98%	99%	0.0	Link 28
	Culex pipiens quinquefasciatus (934)	72%	97%	0.0	Link 27
POLintF1	*Zea mays* (330)	95%	96%	9e-146	Link 24
POLoutF1	Candidate division *TM7* single-cell TM7a_contig_2274 (868)	99%	94%	0.0	Link 29
VIF-VPUinR1	*Dirofilaria immitis* (2631)	100%	91%	0.0	Link 25
	Leishmania major (527)	95%	97%	0.0	Link 30
VIF-VPUoutR1	*Phaseolus coccineus* (99)	54%	96%	8e-22	Link 31
	Methyloversatilis universalis FAM5 (649)	99%	91%	0.0	Link 4
	Leishmania major (527)	95%	97%	0.0	Link 30
Nefyn05	*Methyloversatilis universalis* FAM5 contig00009 (1122)	100%	89%	0.0	Link 3
UNINEF 7’	*Streptomyces* sp. AA4 (8475969)	5%	99%	3e-100	Link 1
	Streptomyces sp AA4 cont1.30 (21080)	5%	99%	8e-103	Link 32
ProRT	*Homo sapiens* genomic sequence flanking *Not*I site (597)	84%	89%	0.0	Link 18

Primers from:

http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2170516/

http://www.biotechniques.com/multimedia/archive/00007/98242bm11_7387a.pdf

Table 3. HIV-1 nucleic acid BLASTN identity search against primer derived subject sequences II

Primer	Subject (bp)	Cover	Identity	E.Value	Data
E10	*Methyloversatilis universalis* FAM5 (649)	99%	91%	0.0	Link 4
E60	*Reticulitermes flavipes* metagenome (632)	98%	96%	0.0	Link 33
	Reticulitermes flavipes metagenome TS23-D12 (738)	100%	95%	0.0	Link 34
	Reticulitermes flavipes metagenome TS23-F11 (619)	100%	92%	0.0	Link 35
	Reticulitermes flavipes metagenome TS23-G7 (528)	98%	91%	0.0	Link 36
E70	*Reticulitermes flavipes* metagenome (632)	98%	96%	0.0	Link 33
	Reticulitermes flavipes metagenome TS23-D12 (738)	100%	95%	0.0	Link 34
	Reticulitermes flavipes metagenome TS23-F11 (619)	100%	92%	0.0	Link 35
E260	*Sorghum bicolor* (772)	54%	98%	9e-169	Link 37
E145	*Reticulitermes flavipes* metagenome (632)	98%	96%	0.0	Link 33
	Reticulitermes flavipes metagenome TS23-D12 (738)	100%	95%	0.0	Link 34
	Candidate division TM7 single-cell TM7a_contig_2274 (868)	100%	92%	0.0	Link 35

Primers from: http://www.hiv.lanl.gov/content/sequence/HIV/COMPENDIUM/1995/PART-III/3.pdf

Table 4: TBLASTN HIV-1 amino acid sequences identity search against specified taxa

Sequence	Subject (bp)	Cover	Identity	E-Value	Data
HIV p31 (288aa)	*Schistosoma mansoni* (6914)	90%	28%	2e-23	Link 38
HIV p24 (231aa)	*Homo sapiens* chromosome 7, GRCh37.p13 (159138663)	60%	37%	6e-11	Link 39
	Homo sapiens chromosome 3, GRCh37.p13 (198022430)	51%	36%	4e-09	Link 40
HIV p55 (500aa)	*Schistosoma mansoni* (6914)	46%	29%	2e-17	Link 41
HIV p51 (440aa)	*Homo sapiens* genomic sequence surrounding *Not***I** site (597)	38%	71%	2e-81	Link 42
	Cryptococcus neoformans var. grubii H99 chromosome 12 (774062)	61%	25%	3e-11	Link 43
	Schistosoma mansoni (6914)	64%	36%	4e-46	Link 44
HIV polPolyprotein (1003aa)	*Homo sapiens* genomic sequence surrounding *Not***I** site (597)	16%	71%	4e-78	Link 45
	Candida albicans genomic DNA, chromosome 7 (949626)	46%	29%	1e-09	Link 46
	Homo sapiens chromosome 7, GRCh37.p13 (159138663)	94%	26%	1e-63	Link 47
HIV p41 (345aa)	Uncultured fungus (856)	23%	93%	4e-45	Link 48
HIV p120 (511aa)	Uncultured fungus *(856)*	7%	100%	4e-14	Link 49
HIV p160 (856aa)	Uncultured fungus (856)	14%	93%	1e-53	Link 50
HIV p66 (560aa)	*Homo sapiens* genomic sequence surrounding *Not***I** site (597)	31%	68%	7e-81	Link 51
	Schistosoma mansoni (6914)	50%	36%	8e-46	Link 52
	Homo sapiens chromosome 7, GRCh37.p13 (159138663)	99%	28%	2e-48	Link 53

Sequence from: http://www.hiv.lanl.gov/content/sequence/HIV/REVIEWS/HXB2.html

Table 5: HIV-1 nucleic acid BLASTN identity search against full-length HXB2. Erased databases.

Subject (bp)	Cover	Identity	E-Value	Data
Brucella melitensis bv. 1 M111(5859)	94%	91%	0.0	Link 54
Brucella melitensis bv.1 M28-12 (4911)	98%	91%	0.0	Link 55
Brucella melitensis bv.1 str. M5 (3176)	92%	90%	0.0	Link 56
Brucella melitensis bv.1 str. S2 (2084)	79%	92%	0.0	Link 57
Ceratocystis fimbriata CBS 114723 contig00925 (389)	98%	91%	4e-150	Link 58

Sequence from: http://www.hiv.lanl.gov/content/sequence/HIV/REVIEWS/HXB2.html

REFERENCES

1. de Mendoza C, Holguin A, Soriano V. False positive for HIV using commercial viral load quantification assays. *AIDS.* 1998;12:2076-2077.
2. Rich JD, Merriman NA, Mylonakis E, *et al*. Misdiagnosis of HIV infection by HIV-1 plasma viral load testing: A case series. *Ann Intern Med.* 1999;130:37-39.
3. Schneider E, Whitmore S, Glynn K, Dominguez K, Mitsch A, McKenna MT. Revised surveillance case definitions for HIV infection among adults, adolescents, and children aged< 18 months and for HIV infection and AIDS among children aged 18 months to< 13 years—United States, 2008. *MMWR Recomm Rep.* 2008;57(10):1-12. http://www.cdc.gov/mmWr/preview/mmwrhtml/rr5710a1.htm
4. Roche COBAS® AmpliPrep/COBAS® TaqMan® HIV-1 Test, version 2.0 (v2.0) 2012. Roche Diagnostics Corporation. Indianapolis USA.
5. Rakowicz-Szulczynska EM. Viral markers RAK in early diagnosis and therapy of breast, ovarian, uterine, and prostate cancers. In: Diamandis E, Fritsche H, Lilja H, Chan D, Schwartz M, eds. *Tumor markers: Physiology, Pathobiology, Technology, and Clinical Applications*: American Association for Clinical Chemistry; 2002.
6. Rakowicz-Szulczynska EM, Jackson B, Szulczynska AM, Smith M. Human immunodeficiency virus type 1-like DNA sequences and immunoreactive viral particles with unique association with breast cancer. *Clin Diagn Lab Immunol.* Sep 1998;5(5):645-653.
7. Agostino M. Introduction to the BLAST Suite and BLASTN. *Practical Bioinformatics*: Garland Science; 2012.
8. Bordoli L. Similarity Searches on Sequence Databases: BLAST, FASTA. Swiss Institute of Bioinformatics 2003: http://www.ch.embnet.org/CoursEMBnet/Basel03/slides/BLAST_FASTA.pdf.
9. The NCBI Handbook [Internet]. 2nd ed: National Center for Biotechnology Information (US); 2013: http://www.ncbi.nlm.nih.gov/books/NBK143764/.
10. Baxevanis A. Current topics in genome analysis 2012, Biological sequence analysis I. *National Human Genome Research Institute* 2012. http://www.youtube.com/watch?v=Ud_6VpX5AgI&list=PLF09DBAA3E24C5068
11. Korber B, Foley B, Kuiken C, Pillai S, Sodroski J. HIV sequence database: Numbering Positions in HIV Relative to HXB2CG. 2008; http://www.hiv.lanl.gov/content/sequence/HIV/REVIEWS/HXB2.html.
12. Sanders-Buell E, Salminen MO, McCutchan FE. Sequencing primers for HIV-1. *The human retroviruses and AIDS Compendium on Line.* Vol January. USA: US Government; 1996:15-21.
13. Zimmermann K, Mannhalter J. Comparable sensitivity and specificity of nested PCR and single-stage PCR using a thermally activated DNA polymerase. *Biotechniques.* 1998;24(2):222-224.
14. Nadai Y, Eyzaguirre LM, Constantine NT, *et al*. Protocol for nearly full-length sequencing of HIV-1 RNA from plasma. *PLoS ONE.* 2008;3(1):e1420.
15. Green RE, Krause J, Ptak SE, *et al*. Analysis of one million base pairs of Neanderthal DNA. *Nature.* 2006;444(7117):330-336.
16. Robertson DL, Hahn BH, Sharp PM. Recombination in AIDS viruses. *J Mol Evol.* 1995;40(3):249-259.
17. http://www.hiv.lanl.gov/content/sequence/HIV/FAQ.html#CONSENSUS

18. Korber B, Gaschen B, Yusim K, Thakallapally R, Kesmir C, Detours V. Evolutionary and immunological implications of contemporary HIV-1 variation. *Br Med Bull.* 2001;58(1):19-42.
19. Shaw GM, Hahn BH, Arya S, Groopman JE, Gallo RC, Wong-Staal F. Molecular characterization of human T-cell leukemia (lymphotropic) virus type III in the acquired immune deficiency syndrome. *Science.* 1984;226:1165-1171.
20. Becker JL, Hazan U, Nugeyre MT, *et al*. Infection of insect cell lines by the HIV virus, an agent of AIDS, and a demonstration of insects of African origin infected by this virus]. *C R Acad Sci III.* 1986;303(8):303-306. http://www.ncbi.nlm.nih.gov/pubmed/3094848
21. Tuma R. Genome Sequencing in Patient Care: Preventing Contamination is Crucial. *J Natl Cancer Inst.* 2011;103(11):847-848.
22. Sawyer MH, Chamberlin CJ, Wu YN, Aintablian N, Wallace MR. Detection of varicella-zoster virus DNA in air samples from hospital rooms. *J Infect Dis.* Jan 1994;169(1):91-94.
23. Bartlett MS, Vermund SH, Jacobs R, *et al*. Detection of Pneumocystis carinii DNA in air samples: likely environmental risk to susceptible persons. *J Clin Microbiol.* Oct 1997;35(10):2511-2513.
24. Contamination in Sequence Databases. http://www.ncbi.nlm.nih.gov/tools/vecscreen/contam/
25. Rakowicz-Szulczynska EM, Jackson B, Snyder W. Prostate, breast and gynecological cancer markers RAK with homology to HIV-1. *Cancer Lett.* Feb 27 1998;124(2):213-223.
26. Kutsenko AS, Gizatullin RZ, Al-Amin AN, *et al*. *Not***I** flanking sequences: a tool for gene discovery and verification of the human genome. *Nucleic Acids Res.* 2002;30(14):3163-3170.
27. Everett DB, Baisely KJ, McNerney R, *et al*. Association of schistosomiasis with false-positive HIV test results in an African adolescent population. *J Clin Microbiol.* 2010;48(5):1570-1577.

LINKS

Link 0: https://www.dropbox.com/s/n268kzverbppnin/Plasmodium_HIV%20.pdf
Link 1: https://www.dropbox.com/s/d1fyvpq9nczqqsf/Streptomyces_HIV.pdf
Link 2: https://www.dropbox.com/s/6cx60f2fa7tgiin/Methyloversatilis_HIV.pdf
Link 3: https://www.dropbox.com/s/awgphvub7ba1vcs/Methy_009_HIV.pdf
Link 4: https://www.dropbox.com/s/vk8g0balfnqn941/MethylContig0006_HIV.pdf
Link 5: https://www.dropbox.com/s/2ypfcfurr9vs2e3/unculturedFungus_HIV.pdf
Link 6: https://www.dropbox.com/s/1bhovt6786uyww1/Rice.pdf
Link 7: https://www.dropbox.com/s/6f117njjq159df9/Neurospora_HIV.pdf
Link 8: https://www.dropbox.com/s/s3du02t9z3cdlq0/Mus_musculus_HIV.pdf
Link 9: https://www.dropbox.com/s/i83p0adtmjhq7mm/Setaria_Italica_HIV.pdf
Link 10: https://www.dropbox.com/s/a0lardii1bffvws/Bienertia%20sinuspersici_HIV.pdf
Link 11: https://www.dropbox.com/s/emc3yqx4hvh8f09/HumanChr8_HIV.pdf
Link 12: https://www.dropbox.com/s/m9aawgkfj9wkyrs/HTLV1_HIV.pdf
Link 13: https://www.dropbox.com/s/41t9bb06jt5w1f3/Locusta_HIV.pdf
Link 14: https://www.dropbox.com/s/23ut1j3p7wiv12s/runnerBean_HIV.pdf
Link 15: https://www.dropbox.com/s/lf5ve5f9r7ovdn8/Nicotiana_Tabacum_HIV.pdf
Link 16: https://www.dropbox.com/s/ai1kjdrn42czgnu/TEA_HIV.pdf
Link 17: https://www.dropbox.com/s/7c4rvi3f5orusfm/Bee_Bombus_insularis_HIV.pdf
Link 18: https://www.dropbox.com/s/0os7f658oq0uvko/NotI_HIV.pdf
Link 19: https://www.dropbox.com/s/2q21rr1zkgbv3u9/Neandertal_HIV.pdf
Link 20: https://www.dropbox.com/s/b2b3e0s2h7vi96p/Neandertal2_HIV.pdf
Link 21: https://www.dropbox.com/s/42027jhkx1hk4jq/Neanderthal3_HIV.pdf
Link 22: https://www.dropbox.com/s/4zardd44xiyndgc/Neanderthal4_HIV.pdf
Link 23: https://www.dropbox.com/s/lup5vwlhhv5f76b/Coral_Fish_HIV.pdf
Link 24: https://www.dropbox.com/s/ntzznmwzc0aox2a/ZeaMays_HIV.pdf
Link 25: https://www.dropbox.com/s/g0s3bs7wxbxqd07/Dirofilaria%20immitis.pdf
Link 26: https://www.dropbox.com/s/9cxs0gqes3crc4h/RAK_alpha_gene.pdf
Link 27: https://www.dropbox.com/s/jp2nskj6qy9jz5s/Culex_HIV.pdf
Link 28: https://www.dropbox.com/s/dd1e51lx14emxc5/Sesamum_HIV.pdf
Link 29: https://www.dropbox.com/s/qpbqospgho6h6y3/TM7_cell_1541.pdf
Link 30: https://www.dropbox.com/s/2wyyu2ood57tf9k/Leishmania.pdf
Link 31: https://www.dropbox.com/s/arjo9xmpw1cs3u0/VIF_Bean.pdf
Link 32: https://www.dropbox.com/s/nim4juhn25r3dx0/Streptomyces_2.pdf
Link 33: https://www.dropbox.com/s/6ruconkqcjzghfn/Termite_1_HIV.pdf
Link 34: https://www.dropbox.com/s/o09gp3vvy48ia2n/Termites_2_HIV.pdf
Link 35: https://www.dropbox.com/s/lck0uo7dexh7fmv/Termites_3_HIV.pdf
Link 36: https://www.dropbox.com/s/5m3p9ts793w08qn/Termites_4_HIV.pdf
Link 37: https://www.dropbox.com/s/x0e2xm8f6jwttum/Sorgum_bicolor_HIV.pdf
Link 38: https://www.dropbox.com/s/tc4mv60hguhe0wj/Schistosoma_p31Integrase.pdf
Link 39: https://www.dropbox.com/s/qnjj4ft3m1y9b2f/p24_Actin.pdf
Link 40: https://www.dropbox.com/s/jsy1or4akhfps95/HIVp24_present_Chr3_Human.pdf
Link 41: https://www.dropbox.com/s/wplwmpr7d5htq8c/Schistosoma_p55.pdf
Link 42: https://www.dropbox.com/s/t5t7vuop1d8rn6f/NotI_p51.pdf
Link 43: https://www.dropbox.com/s/p8mq9itg8idsmw7/p51_Cryptococcus.pdf
Link 44: https://www.dropbox.com/s/dk2r1vbfrol0oxi/Schistosoma_p51RT.pdf
Link 45: https://www.dropbox.com/s/x4fkw9b2izwpalx/NotI_polpolyprotein.pdf
Link 46: https://www.dropbox.com/s/ouq5f1gyw53s921/PolPolyprotein_Candida.pdf
Link 47: https://www.dropbox.com/s/5a1o2f896m0say3/pol_polyprotein_ACTIN.pdf
Link 48: https://www.dropbox.com/s/a4t8fnm9m69p7zk/HIV_p41_unculturedfungus.pdf
Link 49: https://www.dropbox.com/s/e6ujezml6jrfp66/HIV_p120_unculturedfungus.pdf
Link 50: https://www.dropbox.com/s/pra6uvizzg46vzs/HIV_P160_Unculturedfungus.pdf
Link 51: https://www.dropbox.com/s/x6kuvwp1tkcwrra/NotI_p66.pdf

Link 52: https://www.dropbox.com/s/0g0c9l8hby15vqp/Schistosoma_p66.pdf
Link 53: https://www.dropbox.com/s/lv3l5op84up6o4p/p66_actin.pdf
Link 54: https://www.dropbox.com/s/jmverzdxjl7lqkr/BrucellaM111_HIV.pdf
Link 55: https://www.dropbox.com/s/8byxrph7dykx9fa/BrucellaM28_HIV.pdf
Link 56: https://www.dropbox.com/s/qvx2ho1tod6q9v0/BrucellaM5_HIV.pdf
Link 57: https://www.dropbox.com/s/yfiy3nq5xdy15t1/BrucellaS2_HIV.pdf
Link 58: https://www.dropbox.com/s/mqn0ui1v7v3lm61/Ceratocystisfimbriata.pdf