BIOINFORMATION
Discovery at the interface of physical and biological sciences

open access www.bioinformation.net Database
Volume 8(8)

# HPV-QUEST: A highly customized system for automated HPV sequence analysis capable of processing Next Generation sequencing data set

**Li Yin[1]*, Jiqiang Yao[2], Brent P Gardner[1], Kaifen Chang[1], Fahong Yu[2] & Maureen M Goodenow[1]**

[1]Department of Pathology, Immunology and Laboratory Medicine, College of Medicine, University of Florida, Gainesville, FL; [2]Interdisciplinary Center for Biotechnology Research, University of Florida, Gainesville, FL; Li Yin – E-mail: yin@pathology.ufl.edu; Phone: 352-273-8288, Fax: 352-273-8284; *Corresponding author

Received April 18, 2012; Accepted April 19, 2012; Published April 30, 2012

**Abstract:**
Next Generation sequencing (NGS) applied to human papilloma viruses (HPV) can provide sensitive methods to investigate the molecular epidemiology of multiple type HPV infection. Currently a genotyping system with a comprehensive collection of updated HPV reference sequences and a capacity to handle NGS data sets is lacking. HPV-QUEST was developed as an automated and rapid HPV genotyping system. The web-based HPV-QUEST subtyping algorithm was developed using HTML, PHP, Perl scripting language, and MYSQL as the database backend. HPV-QUEST includes a database of annotated HPV reference sequences with updated nomenclature covering 5 genuses, 14 species and 150 mucosal and cutaneous types to genotype blasted query sequences. HPV-QUEST processes up to 10 megabases of sequences within 1 to 2 minutes. Results are reported in html, text and excel formats and display e-value, blast score, and local and coverage identities; provide genus, species, type, infection site and risk for the best matched reference HPV sequence; and produce results ready for additional analyses.

**Availability:** The tool is available for free access at http://www.ijbcb.org/HPV/

**Key Words:** Human papilloma virus, Genotyping; web-based, Blast search, Next Generation sequencing

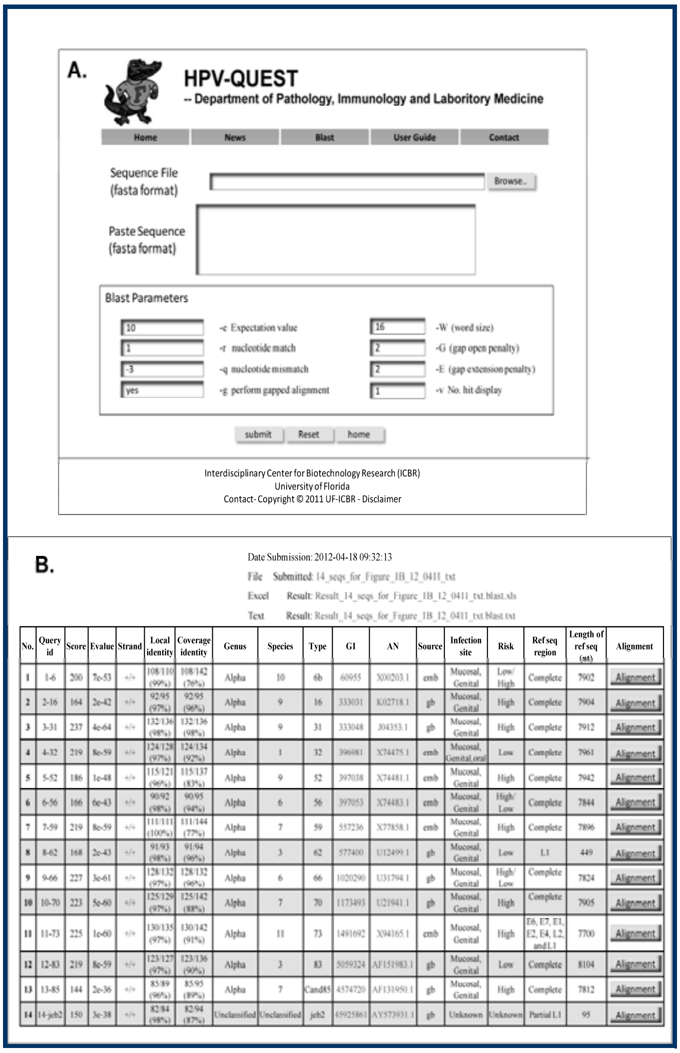
## Background:

Human papilloma virus (HPV), the most common sexually transmitted infection, causes cervical cancer in women, contributes to anogenital cancers in men, and is associated with oropharyngeal cancers and genital warts in men and women **[1]**. Currently, PCR-based assays are applied to identify HPV prevalence, ranges of oncogenic and nononcogenic HPV types, and incidence of multiple type infection **[2]**. Next Generation sequencing (NGS) technology provides increased sensitivity for in depth analysis of HPV types, although large datasets of HPV sequences present considerable barriers for analyses. Available automated HPV genotyping tools, including Virus Sequence Database **[3]**, REGA HPV Automated Subtyping Tool **[4]**, and NCBI blastn are limited by either a restricted number of reference sequences with incomplete annotation or outdated nomenclature, an inability to classify short sequences, or an inadequate capacity to analyze efficiently large sequence data sets. To accelerate HPV genotyping of high-throughput NGS data, an automated system including a comprehensive collection of HPV mucosal and cutaneous reference sequences with updated nomenclature was developed.

## Methodology:

The web-based HPV-QUEST subtyping system uses PHP/HTML language, MYSQL, as the database management system for blast searches is available freely on http://www.ijbcb.org/HPV/. HPV-QUEST is able to processes up to 10 megabases (Mb) of sequences (around 6,500 sequences of 100 bp) per run, returns results within one to two minutes, and displays up to 150 hits with the top hit as default.

HPV genotyping is based on sequences from the L1 region comprised of 1500 nucleotides. HPV-QUEST includes a new HPV database with updated nomenclature for 150 annotated cutaneous and mucosal HPV L1 sequences, representing 5 genuses, 14 species, and 150 types, compiled from complete genomes, subgenomic regions containing the L1 region, or L1 region **[5, 6]** from NCBI Genebank **[7]**, Virus Sequence Database **[8],** and Los Alamos HPV Sequence Database **[9].**



Sequence File (fasta format)

Paste Sequence (fasta format)

Blast Parameters

| | | | |
|---|---|---|---|
| 10 | -e Expectation value | 16 | -W (word size) |
| 1 | -r nucleotide match | 2 | -G (gap open penalty) |
| -3 | -q nucleotide mismatch | 2 | -E (gap extension penalty) |
| yes | -g perform gapped alignment | 1 | -v No. hit display |

Interdisciplinary Center for Biotechnology Research (ICBR)
University of Florida
Contact- Copyright © 2011 UF-ICBR - Disclaimer

Date Submission: 2012-04-18 09:32:13
File Submitted: 14_seqs_for_Figure_1B_12_0411_txt
Excel Result: Result_14_seqs_for_Figure_1B_12_0411_txt.blast.xls
Text Result: Result_14_seqs_for_Figure_1B_12_0411_txt.blast.txt

| No. | Query id | Score | Evalue | Strand | Local identity | Coverage identity | Genus | Species | Type | GI | AN | Source | Infection site | Risk | Ref seq region | Length of ref seq (nt) | Alignment |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 1-6 | 200 | 7e-53 | +/+ | 108/110 (99%) | 108/142 (76%) | Alpha | 10 | 6b | 60955 | X00203.1 | emb | Mucosal, Genital | Low/ High | Complete | 7902 | Alignment |
| 2 | 2-16 | 164 | 2e-42 | +/+ | 92/95 (97%) | 92/95 (96%) | Alpha | 9 | 16 | 333031 | K02718.1 | gb | Mucosal, Genital | High | Complete | 7904 | Alignment |
| 3 | 3-31 | 237 | 4e-64 | +/+ | 132/136 (98%) | 132/136 (98%) | Alpha | 9 | 31 | 333048 | J04353.1 | gb | Mucosal, Genital | High | Complete | 7912 | Alignment |
| 4 | 4-32 | 219 | 8e-59 | +/+ | 124/128 (97%) | 124/134 (92%) | Alpha | 1 | 32 | 396981 | X74475.1 | emb | Mucosal, Genital,oral | Low | Complete | 7961 | Alignment |
| 5 | 5-52 | 186 | 1e-48 | +/+ | 115/121 (96%) | 115/137 (83%) | Alpha | 9 | 52 | 397038 | X74481.1 | emb | Mucosal, Genital | High | Complete | 7942 | Alignment |
| 6 | 6-56 | 166 | 6e-43 | +/+ | 90/92 (98%) | 90/95 (94%) | Alpha | 6 | 56 | 397053 | X74483.1 | emb | Mucosal, Genital | High/ Low | Complete | 7844 | Alignment |
| 7 | 7-59 | 219 | 8e-59 | +/+ | 111/111 (100%) | 111/144 (77%) | Alpha | 7 | 59 | 557236 | X77858.1 | emb | Mucosal, Genital | High | Complete | 7896 | Alignment |
| 8 | 8-62 | 168 | 2e-43 | +/+ | 91/93 (98%) | 91/94 (96%) | Alpha | 3 | 62 | 577400 | U12499.1 | gb | Mucosal, Genital | Low | L1 | 449 | Alignment |
| 9 | 9-66 | 227 | 3e-61 | +/+ | 128/132 (97%) | 128/132 (96%) | Alpha | 6 | 66 | 1020290 | U31794.1 | gb | Mucosal, Genital | High/ Low | Complete | 7824 | Alignment |
| 10 | 10-70 | 223 | 5e-60 | +/+ | 125/129 (97%) | 125/142 (88%) | Alpha | 7 | 70 | 1173493 | U21941.1 | gb | Mucosal, Genital | High | Complete | 7905 | Alignment |
| 11 | 11-73 | 225 | 1e-60 | +/+ | 130/135 (97%) | 130/142 (91%) | Alpha | 11 | 73 | 1491692 | X94165.1 | emb | Mucosal, Genital | High | E6, E7, E1, E2, E4, L2, and L1 | 7700 | Alignment |
| 12 | 12-83 | 219 | 8e-59 | +/+ | 123/127 (97%) | 123/136 (90%) | Alpha | 3 | 83 | 5059324 | AF151983.1 | gb | Mucosal, Genital | Low | Complete | 8104 | Alignment |
| 13 | 13-85 | 144 | 2e-36 | +/+ | 85/89 (96%) | 85/95 (89%) | Alpha | 7 | Cand85 | 4574720 | AF131950.1 | gb | Mucosal, Genital | High | Complete | 7812 | Alignment |
| 14 | 14-jeb2 | 150 | 3e-38 | +/+ | 82/84 (98%) | 82/94 (87%) | Unclassified | Unclassified | jeb2 | 45925861 | AY573931.1 | gb | Unknown | Unknown | Partial L1 | 95 | Alignment |

**Figure 1**: Input and output files. **(A)** HPV-QUEST blast page. Users either paste or upload up to 10 Mb of sequences, chose desired parameters, click submit, and obtain the results in 1 to 2 minutes as html, excel or text files; **(B)** HPV-QUEST output. HPV-QUEST output includes a result page illustrating the No. (the query sequence serial number), Query id (fasta file header of the query sequence), Score (blast score), Evalue (expect value), Strand (+/+ or +/-), Local identity (percentage of matched nucleotides within alignment region), Coverage identity (percentage of nucleotides matched with reference sequence), Genus, Species, Type, GI (NCBI gene identification number), AN (NCBI accession number), Source (source of reference sequence), Infection site (mucosal or cutaneous or both), Risk (high or low or unknown), Ref seq region (reference sequence region in the genome), Length of ref seq (nt), and Alignment (alignment of query sequence with reference sequence). Date and time of submission is also displayed. Two result files in excel or text format are generated for download.

## HPV-QUEST Input and Output:

HPV-QUEST is password protected. User can obtain a log-in password for free access by visiting the website. Sequence files as large as 10 Mb containing forward or complementary reverse sequences are entered in fasta format by copy/pasting or uploading files. Pre-blast sequence cleaning to remove low quality reads is recommended. Blast parameters with suggested default values include: –e (expect value, default = 10), -r (nucleotide match, default = 1), -q (nucleotide mismatch, default = -3), -g (perform gapped alignment, default = yes), -W (word size, default = 16), -G (gap open penalty, default = 2), -E (gap extension penalty, default = 2), and -v (number of hits display, default = 1) **(Figure 1A).** A confirmation with submitted file name, date and time of submission and a link to view results and retrieve reports is generated and e-mailed to the user.

A set of Perl scripts is applied to parse the program output files and produce a result page in HTML format, and a report in both text- and excel-format containing: No. (the query sequence serial number), Query id (fasta file header of the query sequence), Score (blast score), Evalue (expect value), Strand (+/+ or +/-), Local identity (percentage of matched nucleotides within alignment region), Coverage identity (percentage of nucleotides matched with reference sequence), Genus, Species, Type, GI (NCBI gene identification number), AN (NCBI accession number), Source (source of reference sequence), Infection site (mucosal or cutaneous or both), Risk (high or low or unknown), Ref seq region (reference sequence region in the genome), Length of ref seq (nt), and Alignment (alignment of query sequence with reference sequence) (Figure 1B). The original query sequences are included in the report to eliminate the need to match query sequences with correspondent results. Query sequences failing to align with any known reference sequences in the HPV-QUEST are designated as "nd". Any sequences that fail to blast, have low local identity, or with an e-value >1e-15 are considered as low quality, new recombination, or new genotype.

## Testing and validation:

HPV-QUEST version 1.0 was tested and validated in two ways. Firstly, reference sequences used to construct the database were blasted against themselves. The typing was 100% correct, and all e-values were 0 with local or coverage identities of 100%. Secondly, a test dataset of 18,000 quality HPV pyrosequences, generated by PGMY9/11 and GP5+/6+ primers using Titanium Amplicon Pyrosequencing technology from DNA extracted from genital swabs of 15 asymptomatic men recruited in an international study cohort, was processed by using HPV-QUEST and the results compared with typing by traditional NCBI blastn with an cutoff evalue of 1e-15 **[10].** HPV genotypes and frequency distribution by using HPV-QUEST coincided with results from NCBI blastn with significantly shorter processing time (less than 30 minutes versus more than 40 hours) to produce results ready for analysis.

### Caveats and Future development:

Although new HPV types are discovered continuously, HPV classification and nomenclature are updated periodically by the Reference Center for Human Papillomaviruses at the German Cancer Research Center in Heidelberg, which will be used to update HPV-QUEST. Version 2.0 will include HPV subgenomic regions other than L1, reference sequences for non-human papilloma viruses, and extensive data sets generated by next generation sequencing technology.

### Acknowledgment:

University of Florida Cancer Center /Moffitt Cancer Center Collaborative Initiative; NIH R01CA098803; Stephaney W. Holloway University Chair for AIDS Research.

### References:

**[1]** Lu B *et al*. *Cancer Epidemiol Biomarkers Prev*. 2001 **20**: 990 [PMID: 21378268].
**[2]** Plummer M *et al*. *J Infect Dis*. 2011 **203**: 891 [PMID: 21402540].
**[3]** http://kcdc.labkm.net/vsd/main/index.jsp.
**[4]** http://www.bioafrica.net/subtypetool/html/indexhpv.html.
**[5]** de Villiers EM *et al*. *Virology*. **324**: 17 [PMID: 15183049].
**[6]** Bernard HU *et al*. *Virology*. 2010 **401**: 70 [PMID: 20206957].
**[7]** http://www.ncbi.nlm.nih.gov/Taxonomy/Browser/wwwtax.cgi?mode=Tree&id=151340&lvl=3&lin=f&keep=1&srchmode=1&unlock.
**[8]** http://kcdc.labkm.net/vsd/database/gene_search_7.jsp?orgId=7&reset=1.
**[9]** http://hpv-web.lanl.gov/.
**[10]** Giuliano AR *et al*. *Cancer Epidemiol Biomarkers Prev*. 2008 **17**: 2036 [PMID: 18708396].

**Edited by P Kangueane**
**Citation**: **Kalita *et al.*** Bioinformation 8(8): 388-390 (2012)