GENE•IT's GenomeQuest

Short Sequence Searches: Challenges and Solutions

Kamalakar Gulukota
Sr. Director, Content Development
Gene-IT, Inc.

Agenda

- Challenges with short sequences
- Example Workflow: HPV Primers
- Conclusion: The 3 “Legs” of Sequence Search: Content, Search, Analysis

Challenges with short sequences

- **Blast Heuristics**
  - **Word size: Blast will miss approximate hits**

```
Q:          1  ACAGGCACAGGCCACCGCA   19
               |||||||||.||||||||||
S:        156  ACAGGCACACGCCACCGCA  174
```

  - **Expectation value: Blast may even miss exact matches (e.g.: `Val-Gly-Val-Ala-Pro-Gly`)**

- **Setting appropriate search parameters**

- **Too many irrelevant hits: Need to filter**

Exact matches missed by Blast

AAP60213 | Sequence 52 from patent US 6533819 | NCBI GenPept | NCBI record
AAB78834 | I62674 Sequence 24 from patent US 5659041 | NCBI GenPept | NCBI record
AAB79011 | I63725 Sequence 25 from patent US 5662885 | NCBI GenPept | NCBI record
AAS33218 | Sequence 7 from patent US 6689359 | NCBI GenPept | NCBI record
AAT14681 | Sequence 52 from patent US 6699294 | NCBI GenPept | NCBI record
AAW05908 | Sequence 2 from patent US 6797697 | NCBI GenPept | NCBI record
AAW05910 | Sequence 4 from patent US 6797697 | NCBI GenPept | NCBI record
AAW05912 | Sequence 6 from patent US 6797697 | NCBI GenPept | NCBI record
AAW05907 | Sequence 1 from patent US 6797697 | NCBI GenPept | NCBI record
AAW05909 | Sequence 3 from patent US 6797697 | NCBI GenPept | NCBI record
AAW05911 | Sequence 5 from patent US 6797697 | NCBI GenPept | NCBI record
AAE96192 | Sequence 15 from patent US 6355776 | NCBI GenPept | NCBI record
AAE33561 | Sequence 25 from patent US 5976495 | NCBI GenPept | NCBI record
AAE27107 | Sequence 16 from patent US 5955055 | NCBI GenPept | NCBI record
AAE33373 | Sequence 16 from patent US 5972890 | NCBI GenPept | NCBI record
AAC80799 | I92230 Sequence 16 from patent US 5726153 | NCBI GenPept | NCBI record
AAC91462 | AR018237 Sequence 25 from patent US 5780006 | NCBI GenPept | NCBI record
AAC09976 | I71596 Sequence 24 from patent US 5681541 | NCBI GenPept | NCBI record
AAE10046 | Sequence 24 from patent US 5811394 | NCBI GenPept | NCBI record
ABL21266 | Sequence 1 from patent US 7125837 | NCBI GenPept | NCBI record
AAA76336 | Sequence 45 from patent US 5449761 | NCBI GenPept | NCBI record

Your query is identical to this sequence.

```
Align len= 6 aa, Errors= 0, Identity= 100%, Similarity= 100%
Query len= 6 aa, pos= 1-6 aa, Identity query= 100%, Nb gaps query= 0
Subject len= 6 aa, pos= 1-6 aa, Identity subject= 100%, Nb gaps subject= 0

Q:          1  VGVAPG  6
               ||||||
S:          1  VGVAPG  6
```

An example workflow

The problem: Find patented diagnostic primers to human papillomavirus

Step 1: Get the sequence of HPV genome

- **Keyword search in GenBank viral division**
- **“human papillomavirus”**

Step 2: Sequence Search for primers

Step 3: Filter the results

Step 4: Share results

Nucleotides | Proteins

Patents | Transcripts | **More ...**

human papillomavirus

Clear content

Search

- **All Nucleotide Sequence Databases**
  - **Patent Sequences**
  - Transcripts (21)
  - **GenBank**
    - RefSeq mRNA (21)
    - **GenBank Main Divisions (nt)**
      - GB BCT - Bacterial (157)
      - GB HTC - High-throughput cDNA (157)
      - GB INV - Invertebrate (157)
      - GB MAM - Other Mammalian (157)
      - GB PHG - Bacteriophage (157)
      - GB PLN - Plant / fungal / algal (157)
      - GB PRI - Primate (157)
      - GB ROD - Rodent (157)
      - GB SYN - Synthetic (157)
      - GB UNA - Unannotated (157)
      - GB VRL - Viral (157)
      - GB VRT - Other Vertebrate (157)
    - GB EST - Expressed Sequence Tags (157)
    - GB HTG - High-throughput Genomic sequences (157)
    - GB GSS - Genome Survey Sequences (157)
    - GB STS - Sequence Tagged Sites (157)
  - **Affymetrix Sequences**
  - ENSEMBL mRNA (42)
  - DrugBank nucleotide sequences (20051231)

Find sequences | Export Sequences | Launch Application | Display Settings

Match ◉ all of the following ○ any of the following

Description | matches | ... | More | Less | Reset | Apply

Display all sequences, 20 per page sequences 1-1 of 1

	Identifier	Accession	Gene Name	Description	Organism	Database name
☐ [3730]	AB027020	AB027020	E6	Human papillomavirus type 69 DNA, complete genome.	Human papillomavirus type 69	GB VRL - Viral

[3730] top

⊞ AB027020 | Human papillomavirus type 69 DNA, complete genome. | GB VRL - Viral | NCBI record

⊟ sequence

```
   1  CTTTTAACAA TCATAGTTTT ATAAAAGGGT GTAACCGAAG CGGTTTTAAC CGAAAACGGT
  61  GCATATAAAA GTAAAAGACA CAGCCATACA CAAAACCAGC TATGTTTCAA GATCCCAGAG
 121  AAAGACCACG AACGATACAT GAACTATGTG AAGCTTTGAA TACACCTTTG CAATCTTTGC
 181  AGGTACAGTG TGTATATTGC AAGAAAACAT TAGAATGGGC AGATGTATAT AACTTTGCAA
 241  TATGTGATTT AAGAATAGTG TATAGAAATG ATAGTGCATA TGGTGCATGT AAAAAATGTA
 301  TAATATTCTA TTCAAAAATA ATAGAATATA GACGCTACAC ATCGTCTGTG TATGGTGCAA
 361  CACTGGAAGC GCGTCCTAAA CGAAGTTTGT GTAATTTGTT AATAAGGTGT CATAGATGCC
 421  AAATACCATT GGGACCGGAA GAAAAACAGA GAATTGTGGA TGAAAAGCGA CGGTTCCATG
 481  AAATAGCAGG GTACTGGAAA GGGTTGTGCA CAAACTGCTG GAGACCAAGG CGCGAAGCAA
 541  CAGAAACACA AGTATAAATA ACAATGCATG GAGACACAAT TAATATACAG GATGTTATAT
 601  TAGATTTGGT GCCGCAACCC GAAATTGACC TACAGTGTTA CGAACAATTG GACTATGAAC
 661  AATTTGACAG CTCAGAGGAG GATGAAACAG ATAATGTGCG TAACCAGCAA GCCAGACAAG
 721  CTGAACAAGA AGCGTGTTAT AGAATAGAAG CGGAATGTTG TGTATGTAAT AGTATAGTGC
 781  AGCTAGCTGT ACTAAGCAGT CGACAGAACG TCCGAGCGGT GGAGCAGCTG CTGATGGGCG
 841  ACGTGAGTTT GGTGTGCCAC CAGTGTGCTA CATACTAAAC CTGCAATGGA CTGCGAAGGT
 901  ACAGATGGGG AGGGGTTGGG GTGTACAGGG TGGTTTTCAG TAGAAGCAAT AGTAGAAAAA
 961  CATACAGGAG AAACAATATC AGAAGATGAA ATAGAGTATA GTAGCGATAC AGGATCAGAC
1021  CTAATTGGAT TTATAGATGA TAGTAATATT AGTGATGGGG CAGAGCAACA GGTAGCGCAG
1081  GCATTGTTTC AGGCACAAGA AACACAAGCA AATAAGAAGG CAGTGCGTGC ATTAAAACGA
1141  AAGTTACTAG GTAGTCAGAA CAGCCCGTTG CAAGACATAA CAAATCAAAG CAACAGTCAG
```

An example workflow

The problem: Find patented diagnostic primers to human papillomavirus

Step 1: Get the sequence of HPV genome

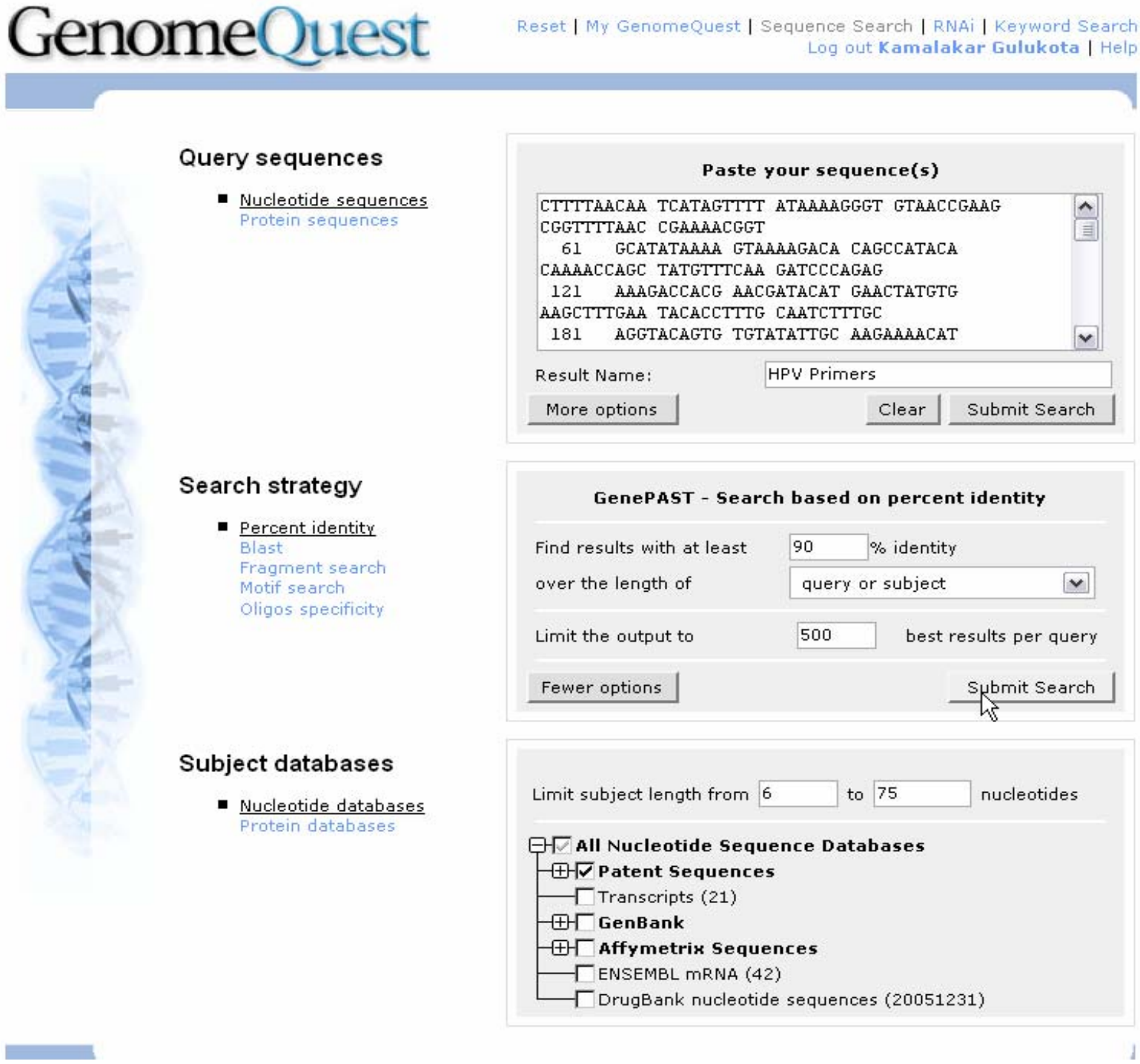
- **Keyword search in GenBank viral division**
- **"human papillomavirus"**

Step 2: Sequence Search for primers

- **Query sequence: HPV Genome**
- **Algorithm: GenePast (or Kerr) at 90% ID**
- **Subject: Short sequences from a comprehensive patent archive**

Step 3: Filter the results

Step 4: Share results


GenomeQuest
Reset | My GenomeQuest | Sequence Search | RNAi | Keyword Search
Log out Kamalakar Gulukota | Help
Query sequences
Nucleotide sequences
Protein sequences
Paste your sequence(s)
CTTTTAACAA TCATAGTTTT ATAAAAGGGT GTAACCGAAG
CGGTTTTAAC CGAAAACGGT
61 GCATATAAAA GTAAAAGACA CAGCCATACA
CAAAACCAGC TATGTTTCAA GATCCCAGAG
121 AAAGACCACG AACGATACAT GAACTATGTG
AAGCTTTGAA TACACCTTTG CAATCTTTGC
181 AGGTACAGTG TGTATATTGC AAGAAAACAT
Result Name:
HPV Primers
More options
Clear
Submit Search
Search strategy
Percent identity
Blast
Fragment search
Motif search
Oligos specificity
GenePAST - Search based on percent identity
Find results with at least
90
% identity
over the length of
query or subject
Limit the output to
500
best results per query
Fewer options
Submit Search
Subject databases
Nucleotide databases
Protein databases
Limit subject length from
6
to
75
nucleotides
All Nucleotide Sequence Databases
Patent Sequences
Transcripts (21)
GenBank
Affymetrix Sequences
ENSEMBL mRNA (42)
DrugBank nucleotide sequences (20051231)

Result Filtering | Export Sequences | Launch Application | Create a Report | Display Settings

Match ◉ all of the following ○ any of the following

Alignment % identity | greater than | [ ]

More | Less | Reset | Apply

Display | all | alignments, | 20 | per page | alignments 1-20 of 500 | next> | Go to Page | [ ]

My GenomeQuest > HPV Primers > query

Redo this search

	Query View	Nb diff.	% id query	% id subject	% Id alignment	Alignment length	Length subject
[1]		0	0.4	100	100	33	33
	US20040265794-0016		JAMES P ZELLER MARSHALL GERSTEIN & BORUN 6300 SEARS TOWERS 233 SOUTH WACKER DRIVE CHICAGO, IL 60606-6357 (US)	US20040265794			
[2]		0	0.4	100	100	30	30
	WO2006004365-0113		BIOMEDLAB, CO.;	WO2006004365			
[3]		0	0.4	100	100	30	30
	EP1302550-0459		KING CAR FOOD INDUSTRIAL CO., LTD. (TW)	EP1302550			
	JP2002360271-0488		KING CAR FOOD INDUSTRIAL CO LTD.	JP2002360271			
[4]		0	0.4	100	100	30	30
	EP1609874-0034		BIOANALISI CENTRO SUD S.N.C. DI PERSEU SINIBALDO EC. (IT)	EP1609874			
[5]		0	0.4	100	100	30	30
	WO2006004365-0112		BIOMEDLAB, CO.;	WO2006004365			

An example workflow

The problem: Find patented diagnostic primers to human papillomavirus

Step 1: Get the sequence of HPV genome

- Keyword search in GenBank viral division
- "human papillomavirus"

Step 2: Sequence Search for primers

- Query sequence: HPV Genome
- Algorithm: GenePast (or Kerr) at 90% ID
- Subject: Short sequences from a comprehensive patent archive

Step 3: Filter the results

- **Remove weak alignments**
- **Show applications with priority**
- **Remove "mega" patents**
- **Remove if not claimed**
- **Remove unrelated patents**

Step 4: Share results

Result Filtering | Export Sequences | Launch Application | Create a Report | Display Settings

Match (•) all of the following ( ) any of the following

Number of differences	less than	3		
Priority date	is earlier than	2004	September	6
Number of SEQ (nuc/prot)	less than	1000		
Patent sequence location	contains	claim		
Alignment % identity	greater than			

...

Reset | Apply

Alignment properties
- Alignment % identity
- Alignment % similarity
- Alignment length
- Number of differences
- Query % identity
- Query start position
- Query stop position
- Subject % identity
- Subject start position
- Subject stop position

Subject text annotations
- Abstract
- All text
- Claims
- Comments
- Organism
- Patent assignee
- Patent inventors
- Patent title

alignments, 20 per page — alignments 1-20 of 500 next> — Go to Page

My … > query — Redo this search

	Nb diff.	% id query	% id subject	% Id alignment	Alignment length	Length subject
	0	0.4	100	100	33	33
		JAMES P ZELLER MARSHALL GERSTEIN & BORUN 6300 SEARS TOWERS 233 SOUTH WACKER DRIVE CHICAGO, IL 60606-6357 (US)	US20040265794			
[2]	0	0.4	100	100	30	30
WO2006004365-0113		BIOMEDLAB, CO.;	WO2006004365			
[3]	0	0.4	100	100	30	30
EP1302550-0459		KING CAR FOOD INDUSTRIAL CO., LTD. (TW)	EP1302550			
JP2002360271-0488		KING CAR FOOD INDUSTRIAL CO LTD.	JP2002360271			
[4]	0	0.4	100	100	30	30
EP1609874-0034		BIOANALISI CENTRO SUD	EP1609874			

Result Filtering | Export Sequences | Launch Application | Create a Report | Display Settings

Match ◉ all of the following ○ any of the following

Number of differences	less than	3		
Priority date	is earlier than	2004	September	6
Number of SEQ (nuc/prot)	less than	1000		
Patent sequence location	contains	claim		...
Claims	contains	papilloma diagnostic primer PCR		...

More | Less | Reset | Apply

Display all alignments, 20 per page alignments 1-14 of 14

My GenomeQuest > HPV Primers > query Redo this search

	Query View	Nb diff.	% id query	% id subject	% Id alignment	Alignment length	Length subject
☐ [1]		0	0.4	100	100	33	33
☐	US20040265794-0016		JAMES P ZELLER MARSHALL GERSTEIN & BORUN 6300 SEARS TOWERS 233 SOUTH WACKER DRIVE CHICAGO, IL 60606-6357 (US)	US20040265794			
☐ [9]		0	0.3	100	100	26	26
☐	US20050175989-0314		MATHEWS, COLLINS, SHEPHERD & MCKAY, SUITE 306 100 THANET CIRCLE PRINCETON, NJ 08540-3674	US20050175989			
☐ [39]		0	0.3	100	100	23	23
☐	US6482588-0159		INNOGENETICS S.A. BELGIUM	US6482588			
☐	WO9914377-0159		INNOGENETICS NV (BE);	WO9914377			

An example workflow

The problem: Find patented diagnostic primers to human papillomavirus

Step 1: Get the sequence of HPV genome

- Keyword search in GenBank viral division
- "human papillomavirus"

Step 2: Sequence Search for primers

- Query sequence: HPV Genome
- Algorithm: GenePast (or Kerr) at 90% ID
- Subject: Short sequences from a comprehensive patent archive

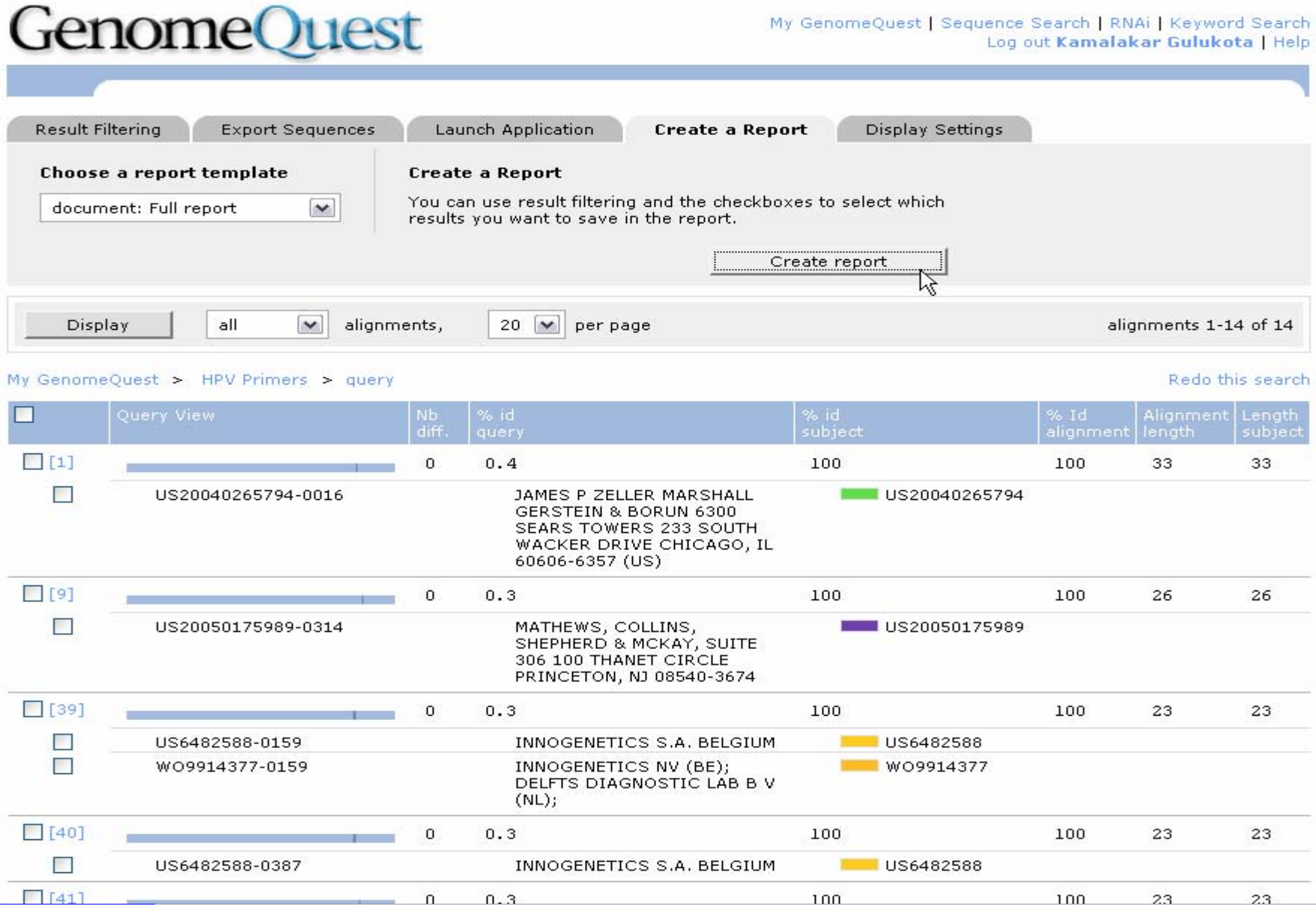
Step 3: Filter the results

- Remove weak alignments
- Show applications with priority
- Remove "mega" patents
- Remove if not claimed
- Remove unrelated patents

Step 4: Share results

Analyze sequences on your desktop

Create Reports



GOresID	Query View	Nb diff.	% id query	% id subject	% Id alignment	Alignment length	Length subject
[1]		0	0.4	100	100	33	33
result 1	US20040265794-0016		JAMES P ZELLER MARSHALL GERSTEIN & BORUN 6300 SEARS TOWERS 233 SOUTH WACKER DRIVE CHICAGO, IL 60606-6357 (US)	US20040265794			
[9]		0	0.3	100	100	26	26
result 13	US20050175989-0314		MATHEWS, COLLINS, SHEPHERD & MCKAY, SUITE 306 100 THANET CIRCLE PRINCETON, NJ 08540-3674	US20050175989			
[39]		0	0.3	100	100	23	23
result 50	US6482588-0159		INNOGENETICS S.A. BELGIUM	US6482588			
result 51	WO9914377-0159		INNOGENETICS NV (BE); DELFTS DIAGNOSTIC LAB B V (NL);	WO9914377			
[40]		0	0.3	100	100	23	23
result 52	US6482588-0387		INNOGENETICS S.A. BELGIUM	US6482588			
[41]		0	0.3	100	100	23	23
result 53	US20050175989-0312		MATHEWS, COLLINS, SHEPHERD & MCKAY, SUITE 306 100 THANET CIRCLE PRINCETON, NJ 08540-3674	US20050175989			

US20040265794 Patent SEQ ID NO found: 16

Priority date	20010914 KR20010056827; 20010918 WO2001KR01562;
Patent title	Genotyping kit for diagnosis of human papilloma virus infection
Abstract	The present invention relates to a genotyping kit for diagnosis of detecting the human papillomavirus (HPV) infection, probes for genotyping the HPV, and DNA chips including the probes. Also, the present invention relates to a method for diagnosis of HPV infection. BACKGROUND OF THE INVENTION [0001] (a) Field of the Invention [0002] The present invention relates to a genotyping kit and method for diagnosis of human papillomavirus (HPV) infection, probes for genotyping the HPV, and DNA chips including the probes. More specifically, the present invention relates to a genotyping kit for detecting human papillomaviruses from clinical samples of infected patients using a DNA chip, a process for preparing the said DNA chip, and a method for diagnosis of HPV infection using the genotyping kit.
Patent assignee	JAMES P ZELLER MARSHALL GERSTEIN & BORUN 6300 SEARS TOWERS 233 SOUTH WACKER DRIVE CHICAGO, IL 60606-6357 (US)
Patent inventors	Yoon; Sung-Wook (Seoul Republic of Korea); Park; Tae-Shin (Seoul Republic of Korea); Kim; Jeong-Mi (Seoul Republic of Korea); Park; Mi-Sun (Busan Republic of Korea); CORRESPONDENCE_ADDRESS: James P Zeller Marshall Gerstein & Borun 6300 Sears Towers 233 South Wacker Drive Chicago, IL 60606-6357 (US)
Patent family	Equivalents: EP1434873; JP2005503177; Family: WO03027323; KR2003027178; EP1434873; CN1558954; JP2005503177;
Claimed SEQ ID	1-41; 45-46;
Number of SEQ (nuc/prot)	54
Claims	What is claimed is: **1.** A probe comprising a nucleotide sequence which can complementarily bind to DNA of Human Papillomavirus (HPV) and is

Conclusions

Content
Sequence databases

Search
Algorithms

Analysis
Whittle down results

- These are the 3 “legs” of sequence search
- Putting all 3 “legs” together makes sequence search efficient
- This is especially important for short sequence search