GMAP: A Genomic Mapping and Alignment Program for mRNA and EST Sequences

Thomas D. Wu, M.D., Ph.D.
Department of Bioinformatics
Genentech, Inc.
South San Francisco, CA 94080

28 June 2005
General information about GMAP

Description of algorithm and test results:

**BIOINFORMATICS**  **ORIGINAL PAPER**

Sequence analysis

**GMAP: a genomic mapping and alignment program for mRNA and EST sequences**

Thomas D. Wu\(^1\,*\) and Colin K. Watanabe\(^2\)

\(^1\)Department of Bioinformatics and \(^2\)Department of Corporate Information Technology, Genentech, Inc., South San Francisco, CA 94080, USA

Received on February 24, 2004; revised on January 27, 2005; accepted on February 4, 2005

Advance Access publication February 22, 2005

Requirements: Unix, C compiler, Perl

Source code and sample databases: www.gene.com/share/gmap

Contact: Thomas Wu, Genentech, Inc., twu@gene.com
Outline

- Introduction
- Usage
- Algorithm
- Utility programs
- Basic setup
- Advanced setup
Genomic mapping and alignment

• Genomic mapping: Given a cDNA, find where it best aligns to an entire genome

• Genomic alignment: Given a cDNA and a genomic segment, provide a nucleotide-level correspondence for the exons of the cDNA to the genomic segment
Basic use

Map and align against entire genome:

```
gmap -d genome fastafile
```

or

```
cat fastafile | gmap -d genome
```

Align against a given genomic segment:

```
gmap -g genomic.fa cdna.fa
```

or

```
cat cdna.fa | gmap -g genomic.fa
```
gmap.aligntosegment

# Call: gmap -g genomic.fa -A cdna.fa

>AA015654 (198 bp)  ze29c11.rl Soares retina N2b4HR Homo sapiens cDNA clone IMAGE:36
0404 5' similar to gb:X01677 GLYCLERALDEHYDE 3-PHOSPHATE DEHYDROGENASE, LIVER (HUMAN);
mRNA sequence. 198 bp, mRNA, linear, EST 29-NOV-1996
Paths (1):
Path 1: query 10--198 (189 bp) => genomic 231--2,139 (1909 bp)
cDNA direction: sense
Genomic pos: 231--2,139 (+ strand)
Number of exons: 3
Coverage: 95.5
Percent identity: 97.4 (184 matches, 1 mismatches, 4 indels, 1 unknowns)
Non-intron gaps: 3 openings, 3 bases in cdna; 1 openings, 1 bases in genome
Translation: 10..194 (62 aa)
Mutations: S8* (TCA>TGA) [33]

Alignments:
Alignment for path 1:

<table>
<thead>
<tr>
<th>0</th>
<th>.</th>
<th>.</th>
<th>.</th>
<th>.</th>
<th>.</th>
<th>.</th>
<th>.</th>
</tr>
</thead>
<tbody>
<tr>
<td>aa.g</td>
<td>1 R A A E P K R S D T M G K V K V</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>231</td>
<td>CGCCCAACCC GAGCCACATC GTTCAGACACCATGGGGGAGAGTGAAGTC</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>50</td>
<td>GAGTCAACCGGTG...TAGATTTTGTATTTGGGGGGCTGTCACAG</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>100</td>
<td>AFGNGLKTVS</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>aa.c</td>
<td>1 R A A E P K R</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>17 G V N G</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>150</td>
<td>G GTTCACCAACCATGTGGGTG...GAGTTTACATGGTTCCAAATAGA</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>200</td>
<td>G TCA</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>aa.g</td>
<td>31 A A F N S G K V D I V A I N D P</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1592</td>
<td>G GTTCACCAACCATGTGGGTG...GAGTTTACATGGTTCCAAATAGA</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1952</td>
<td>GGTTCACCAACCATGTGGGTG...GAGTTTACATGGTTCCAAATAGA</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>100</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>aa.c</td>
<td>31 A A F N S G K V D I V A I N D P</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>150</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>200</td>
<td>61 S T</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2132</td>
<td>TTOGACCC</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>191</td>
<td>XTCCACCC</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>aa.c</td>
<td>61 S T</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
## Genomic mapping and alignment

<table>
<thead>
<tr>
<th></th>
<th>Mapping</th>
<th>Alignment</th>
</tr>
</thead>
<tbody>
<tr>
<td>dds/gap2</td>
<td></td>
<td>X</td>
</tr>
<tr>
<td>sim4</td>
<td></td>
<td>X</td>
</tr>
<tr>
<td>Spidey</td>
<td></td>
<td>X</td>
</tr>
<tr>
<td>GeneSeqer</td>
<td></td>
<td>X</td>
</tr>
<tr>
<td>MGAlign</td>
<td></td>
<td>X</td>
</tr>
<tr>
<td>ssaha</td>
<td>X</td>
<td></td>
</tr>
<tr>
<td>blat</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>squall</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>gmap</td>
<td>X</td>
<td>X</td>
</tr>
</tbody>
</table>
## Improvements in gene structure prediction

<table>
<thead>
<tr>
<th>Evidence</th>
<th>Microexon BM467080:555..571, chr +3</th>
<th>Low quality 3’ end BG118317:738..757, chr –12</th>
<th>Final exon AA036958:424..439, chr –1</th>
</tr>
</thead>
<tbody>
<tr>
<td>GMAP</td>
<td>GAGTCAGGTA...CAGCATCAGGTA...TAGACAA</td>
<td>GCTATATGAAAGGTA...CAGGAGATCCGG</td>
<td>AAGGTA...CAGGCTGATTCACCCC</td>
</tr>
<tr>
<td></td>
<td>GAGTCAG 393 CATCAG 198 ACAAA</td>
<td>GTTGTATG AGAG 522 GA ATCCGG</td>
<td>AAG 6772 GCTGATTCACCCC</td>
</tr>
<tr>
<td>BLAT</td>
<td>GAGTCAGGTA...CTTCATCTCA...TGtagacaa</td>
<td>Alignment ends at nt 651</td>
<td>AGGAAGG</td>
</tr>
<tr>
<td></td>
<td>GAGTCAG 48 CATC 543 AGACAA</td>
<td></td>
<td></td>
</tr>
<tr>
<td>dds/gap2</td>
<td>GAGTCAGGTA...CCTCATGTAGACAA</td>
<td>GCTATATGAAAGGATGTGTT...GATCCGG</td>
<td>AAGGT ATTTGTCCC</td>
</tr>
<tr>
<td></td>
<td>GAGTCAG 590 CATC AGACAA</td>
<td>GTTGTATGA GAGGAAT 523 CCGG</td>
<td>AAGGCTGATTCACCCC</td>
</tr>
<tr>
<td>GeneSequer</td>
<td>Same as GMAP</td>
<td>Same as GMAP</td>
<td>Same as GMAP</td>
</tr>
<tr>
<td>MGAlign</td>
<td>TCAG 595 AGACAA</td>
<td>Alignment skips nt 691 to 748</td>
<td>Same as GMAP</td>
</tr>
<tr>
<td></td>
<td>TCAG 4 AGACAA</td>
<td></td>
<td></td>
</tr>
<tr>
<td>SIM4</td>
<td>GAGTCAGGTA...CTGCCCTCATGTAGACAA</td>
<td>Alignment ends at nt 714</td>
<td>AAGG T ATTTGTCCC</td>
</tr>
<tr>
<td></td>
<td>GAGTCAG 586 CAT CAGACAA</td>
<td></td>
<td>AAGGCTGATTCACCCC</td>
</tr>
<tr>
<td>Spidey</td>
<td>GAGTCAGGTA...ACTTTGAA...TAGACAA</td>
<td>Alignment ends at nt 705</td>
<td>Same as GMAP</td>
</tr>
<tr>
<td></td>
<td>GAGTCAGCAGTAC 591 ACAAA</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
## Improvements in splice site prediction

<table>
<thead>
<tr>
<th>Evidence</th>
<th>1 sequence difference BF669985:519..530, chr +X</th>
<th>2 sequence differences BF846255:69..80, chr +9</th>
<th>0 sequence differences BF591480:252..264, chr +11</th>
</tr>
</thead>
<tbody>
<tr>
<td>GMAP</td>
<td>CAGAAGGTATGT...CCTTAGATCCACA</td>
<td>ACTTGCCAAGGTAAAT...ATTAGGGA</td>
<td>ACATTGGAAGT...TTGTTTGGTGAC</td>
</tr>
<tr>
<td></td>
<td>CAGAAG 75 ATC ACA</td>
<td>AC TGCC AG 7812 GGTA</td>
<td>ACATTG 98 GGTGAC</td>
</tr>
<tr>
<td>BLAT</td>
<td>CAGAAGGTATGTAG...TAGATCCACA</td>
<td>ACTTGCCAAGGTA...TCATTAGGGTA</td>
<td>Same as GMAP</td>
</tr>
<tr>
<td></td>
<td>CAGAAGAT 76 CACA</td>
<td>ACT GCCA 7813 ACGGTA</td>
<td>Same as GMAP</td>
</tr>
<tr>
<td>dds/gap2</td>
<td>Same as GMAP</td>
<td>ACTTGCCAAGGTA...CATTTAGGGTA</td>
<td>Same as GMAP</td>
</tr>
<tr>
<td></td>
<td></td>
<td>ACT GCCA 7813 GGTA</td>
<td></td>
</tr>
<tr>
<td>GeneSequer</td>
<td>Same as GMAP</td>
<td>Misses 5’ exon</td>
<td>Same as GMAP</td>
</tr>
<tr>
<td>MGAlign</td>
<td>Same as GMAP</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SIM4</td>
<td>Same as GMAP</td>
<td>Same as GMAP</td>
<td></td>
</tr>
<tr>
<td>Spidey</td>
<td>CAGAAGGTATGTAGTG...GATCCACA</td>
<td>ACTTGCCAAGG...GTCATTAGGGTA</td>
<td></td>
</tr>
<tr>
<td></td>
<td>CAGAAGATCA 76 CA</td>
<td>ACTGG 7814 CAGGTA</td>
<td></td>
</tr>
</tbody>
</table>
## Comparison of BLAT and GMAP

<table>
<thead>
<tr>
<th>Feature</th>
<th>BLAT</th>
<th>GMAP</th>
</tr>
</thead>
<tbody>
<tr>
<td>Primary mode</td>
<td>Client/server</td>
<td>Standalone</td>
</tr>
<tr>
<td>Index lookup</td>
<td>RAM-based</td>
<td>File- or RAM-based</td>
</tr>
<tr>
<td>Startup time</td>
<td>Minutes</td>
<td>Seconds</td>
</tr>
<tr>
<td>Switching genomes</td>
<td>Multiple servers</td>
<td>Easy</td>
</tr>
<tr>
<td>Accuracy</td>
<td></td>
<td>Better</td>
</tr>
<tr>
<td>Speed</td>
<td></td>
<td>Faster</td>
</tr>
<tr>
<td>Memory usage</td>
<td>1.2 GB</td>
<td>128 MB</td>
</tr>
<tr>
<td>Licensing</td>
<td>Free to academics</td>
<td>Free</td>
</tr>
</tbody>
</table>
Outline

- Introduction
- Usage
- Algorithm
- Utility programs
- Basic setup
- Advanced setup
Features of GMAP

- Compressed alignment format (-Z)
- Batch mode (-B flag) with multithreading (-t)
- Microexon detection: Uses probability-based calculations
- Complex genomes: Handles alternate strains, alternate assemblies, and unmapped contigs
- Chimera detection (-x)
- Lookup of genomic map information (-m)
- Relative alignment of ESTs (-w)
- Cross-species alignment in many cases (-X)
# Call: gmap --help

Usage: gmap [OPTIONS...] <FASTA file>, or cat <FASTA file> | gmap [OPTIONS...]

Input options (must include -d or -g)
- `-D, --dir=directory` Genome directory
- `-d, --db=STRING` Genome database
- `-G, --genomefull` Use full genome (all ASCII chars allowed; built explicitly during setup), not compressed version
- `-g, --gseg=filename` User-suppled genomic segment

Computation options
- `-B, --batch=INT` Batch mode (1 = pre-load only indices; 2 = pre-load both indices and genome)
- `-L, --length` Max total intron length (default 1200000)
- `-x, --chimera_margin=INT` Amount of unaligned sequence that triggers search for a chimera (default off)
- `-l, --maponly` Report stage 1 only (genomic mapping only)
- `-X, --cross` Cross-species mode (changes parameters)
- `-w, --mutationref=filename` Mutation reference
- `-t, --nthreads=INT` Number of worker threads
- `-s, --altstrain` Search alternate strains in addition
- `-C, --chrsubsetfile=filename` User-supplied chromosome subset file
- `-c, --chrsubset=string` Chromosome subset to search
- `-U, --endmicro` Allow end microexons with arbitrarily long introns

Output options
- `-S, --summary` Show summary of alignments only
- `-A, --align` Show alignments
- `-3, --continuous` Show alignment in three continuous lines
- `-9, --diagnostic` Show diagnostics (alignment with *'s for stage 3)
- `-n, --npaths=INT` Maximum number of paths to show. If set to 0, prints two paths if chimera detected, else one.
- `-Z, --compress` Print output in compressed format
- `-O, --ordered` Print output in same order as input (relevant only if there is more than worker thread)
- `-5, --md5` Print MD5 checksum for each query sequence
- `-M, --mapdir=directory` Map directory
- `-m, --map=iitfile` Map file
- `-e, --mapexons` Map each exon separately
- `-b, --mapboth` Map both strands of genome
- `-o, --chimera_overlap` Overlap to show, if any, at chimera breakpoint
- `-f, --format=INT` Format for output (1 = PSL (BLAT) format, 2 = coords in table format)

Alignment output options
- `-E, --exons` Print cDNA exons
- `-P, --protein_dna` Print protein sequence (cDNA)
- `-Q, --protein_gen` Print protein sequence (genomic)
- `-F, --fulllength` Assume full-length protein, starting with Met
- `-T, --truncate` Truncate full-length protein, from Met to Stop. Implies -F flag.
- `-N, --nolengths` No intron lengths in alignment
- `-I, --invertmode=INT` Mode for alignments to (-) strand: 1=Invert and print sense (-) strand; 2=Invert and print antisense (+) strand.
- `-i, --introngap=INT` Nucleotides to show on each end of intron (default=3)
- `-l, --wraplength=INT` Wrap length for alignment (default=50)

Help options
- `-V, --version` Show version
# Call: gmap -d NHGD_R35 -Z cdna.fa | gmap_uncompress

>AA015654 (198 bp)

<table>
<thead>
<tr>
<th>Paths (1):</th>
</tr>
</thead>
<tbody>
<tr>
<td>Path 1: query 20..198 (198 bp) =&gt; chr 12:6514230..6516138 (1909 bp)</td>
</tr>
<tr>
<td>cDNA direction: sense</td>
</tr>
<tr>
<td>Genomic pos: NHGD_R35:2118024113..2118026021 (+ strand)</td>
</tr>
<tr>
<td>Number of exons: 3</td>
</tr>
<tr>
<td>Coverage: 95.5</td>
</tr>
<tr>
<td>Percent identity: 97.4</td>
</tr>
</tbody>
</table>

## Alignments:

### Alignment for path 1:

<table>
<thead>
<tr>
<th>Start</th>
<th>End</th>
<th>Identity</th>
<th>Accession</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>12:6514230</td>
<td>6516288</td>
<td>98%</td>
<td>AA015654 (198 bp)</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Methods used:</th>
</tr>
</thead>
<tbody>
<tr>
<td>-X:46056324 AAGTGGAATTGTT GCCATCAATGCCCATCATTGACACTCAAATCAT</td>
</tr>
<tr>
<td>100</td>
</tr>
<tr>
<td>120 AAGTGGAATTGTTGCCATCAATGGCCCATCATTGACACTCAAATCAT</td>
</tr>
</tbody>
</table>

# Another way to compress (same as using -Z flag)

## Call: gmap -d NHGD_R35 -A cdna.fa | gmap_compres

>AA015654 NHGD_R35 1/2 198 3 95.5 97.4 10--198 2118024113--2118026021 12:6514230--6516138 +

<table>
<thead>
<tr>
<th>Methods used:</th>
</tr>
</thead>
<tbody>
<tr>
<td>6514230 6516288 10 69 98</td>
</tr>
<tr>
<td>6515921 6516020 70 170 100</td>
</tr>
<tr>
<td>6516111 6516138 171 198 100</td>
</tr>
<tr>
<td>3 5X 10^X 3xG 36&gt; 60 1632</td>
</tr>
<tr>
<td>64^G 36&gt; 101 90</td>
</tr>
<tr>
<td>20X 7* 28</td>
</tr>
<tr>
<td>3080740920--3080740744 X:46056423--46056247 6247 -</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Methods used:</th>
</tr>
</thead>
<tbody>
<tr>
<td>4X 4xA 5xG 0^C 12xX 5xC 3xT 10xG 16^G 38</td>
</tr>
<tr>
<td>4xA 4xX 0xG 0^C 51xG 12xX 3xT 10xG 16^G 38</td>
</tr>
</tbody>
</table>

| xT 11xA 5xX 7* 179 |

## Alignment for path 1:

<table>
<thead>
<tr>
<th>Start</th>
<th>End</th>
<th>Identity</th>
<th>Accession</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>12:6514230</td>
<td>6516247</td>
<td>93%</td>
<td>AA015654 (198 bp)</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Methods used:</th>
</tr>
</thead>
<tbody>
<tr>
<td>-X:46056247 (20--198) 93%</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Methods used:</th>
</tr>
</thead>
<tbody>
<tr>
<td>6514230 6516288 10 69 98</td>
</tr>
<tr>
<td>6515921 6516020 70 170 100</td>
</tr>
<tr>
<td>6516111 6516138 171 198 100</td>
</tr>
<tr>
<td>3 5X 10^X 3xG 36&gt; 60 1632</td>
</tr>
<tr>
<td>64^G 36&gt; 101 90</td>
</tr>
<tr>
<td>20X 7* 28</td>
</tr>
<tr>
<td>3080740920--3080740744 X:46056423--46056247</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Methods used:</th>
</tr>
</thead>
<tbody>
<tr>
<td>4X 4xA 5xG 0^C 12xX 5xC 3xT 10xG 16^G 38</td>
</tr>
<tr>
<td>4xA 4xX 0xG 0^C 51xG 12xX 3xT 10xG 16^G 38</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Methods used:</th>
</tr>
</thead>
<tbody>
<tr>
<td>-X:46056234 AAGTGGAATTGTT GCCATCAATGCCCATCATTGACACTCAAATCAT</td>
</tr>
<tr>
<td>100</td>
</tr>
<tr>
<td>120 AAGTGGAATTGTTGCCATCAATGGCCCATCATTGACACTCAAATCAT</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Methods used:</th>
</tr>
</thead>
<tbody>
<tr>
<td>6514230 6516288 10 69 98</td>
</tr>
<tr>
<td>6515921 6516020 70 170 100</td>
</tr>
<tr>
<td>6516111 6516138 171 198 100</td>
</tr>
<tr>
<td>3 5X 10^X 3xG 36&gt; 60 1632</td>
</tr>
<tr>
<td>64^G 36&gt; 101 90</td>
</tr>
<tr>
<td>20X 7* 28</td>
</tr>
<tr>
<td>3080740920--3080740744 X:46056423--46056247</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Methods used:</th>
</tr>
</thead>
<tbody>
<tr>
<td>4X 4xA 5xG 0^C 12xX 5xC 3xT 10xG 16^G 38</td>
</tr>
<tr>
<td>4xA 4xX 0xG 0^C 51xG 12xX 3xT 10xG 16^G 38</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Methods used:</th>
</tr>
</thead>
<tbody>
<tr>
<td>-X:46056247 (20--198) 93%</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Methods used:</th>
</tr>
</thead>
<tbody>
<tr>
<td>-X:46056234 AAGTGGAATTGTT GCCATCAATGCCCATCATTGACACTCAAATCAT</td>
</tr>
<tr>
<td>100</td>
</tr>
<tr>
<td>120 AAGTGGAATTGTTGCCATCAATGGCCCATCATTGACACTCAAATCAT</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Methods used:</th>
</tr>
</thead>
<tbody>
<tr>
<td>6514230 6516288 10 69 98</td>
</tr>
<tr>
<td>6515921 6516020 70 170 100</td>
</tr>
<tr>
<td>6516111 6516138 171 198 100</td>
</tr>
<tr>
<td>3 5X 10^X 3xG 36&gt; 60 1632</td>
</tr>
<tr>
<td>64^G 36&gt; 101 90</td>
</tr>
<tr>
<td>20X 7* 28</td>
</tr>
<tr>
<td>3080740920--3080740744 X:46056423--46056247</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Methods used:</th>
</tr>
</thead>
<tbody>
<tr>
<td>4X 4xA 5xG 0^C 12xX 5xC 3xT 10xG 16^G 38</td>
</tr>
<tr>
<td>4xA 4xX 0xG 0^C 51xG 12xX 3xT 10xG 16^G 38</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Methods used:</th>
</tr>
</thead>
<tbody>
<tr>
<td>-X:46056247 (20--198) 93%</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Methods used:</th>
</tr>
</thead>
<tbody>
<tr>
<td>-X:46056234 AAGTGGAATTGTT GCCATCAATGCCCATCATTGACACTCAAATCAT</td>
</tr>
<tr>
<td>100</td>
</tr>
<tr>
<td>120 AAGTGGAATTGTTGCCATCAATGGCCCATCATTGACACTCAAATCAT</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Methods used:</th>
</tr>
</thead>
<tbody>
<tr>
<td>6514230 6516288 10 69 98</td>
</tr>
<tr>
<td>6515921 6516020 70 170 100</td>
</tr>
<tr>
<td>6516111 6516138 171 198 100</td>
</tr>
<tr>
<td>3 5X 10^X 3xG 36&gt; 60 1632</td>
</tr>
<tr>
<td>64^G 36&gt; 101 90</td>
</tr>
<tr>
<td>20X 7* 28</td>
</tr>
<tr>
<td>3080740920--3080740744 X:46056423--46056247</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Methods used:</th>
</tr>
</thead>
<tbody>
<tr>
<td>4X 4xA 5xG 0^C 12xX 5xC 3xT 10xG 16^G 38</td>
</tr>
<tr>
<td>4xA 4xX 0xG 0^C 51xG 12xX 3xT 10xG 16^G 38</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Methods used:</th>
</tr>
</thead>
<tbody>
<tr>
<td>-X:46056247 (20--198) 93%</td>
</tr>
</tbody>
</table>
# Batch mode 1: pre-loads just the indices
# Call: gmap -d NHGD_R35 -B 1 -Z cdnas.fa

Pre-loading index db every 2048 indices/page.

# Batch mode 2: pre-loads indices and the genome
# Call: gmap -d NHGD_R35 -B 1 -Z cdnas.fa

Pre-loading index db every 2048 indices/page.

# Batch mode with multithreading (3 threads)
# Alignments will be printed as they are computed
# Call: gmap -d NHGD_R35 -B 1 -t 3 -Z cdnas.fa

# Batch mode with multithreading and ordered output
# Alignments will be printed in same order as input
# Call: gmap -d NHGD_R35 -B 1 -t 3 -Z -O cdnas.fa
# Call: gmap -d NHGD_R35 -A cdna.fa

>AA641131 (182 bp)  nr28nh1.r1 NCI_CGAP_Pr3 Homo sapiens cDNA clone IMAGE:1169347, mRNA sequence. 182 bp, mRNA, linear, EST 27-OCT-1997

Paths (1):
Path 1: query 3--182 (180 bp) => chr 3:178,816,080--178,829,972 (11893 bp)
cDNA direction: sense
Genomic pos: NHGD_R35:670,694,445--670,706,337 (+ strand)
Accessions: NT_005612.14:83,830,536--83,842,428 (out of 100530261 bp)
Number of exons: 3
Coverage: 98.9
Percent identity: 98.9 (178 matches, 2 mismatches, 0 indels, 0 unknowns)
Translation: 5..170 (59 aa)
Mutations: S38R (AGT>AGA) [118]

Alignments:
Alignment for path 1:

```
+3:178818080-178818197 (3-120)  98% ->
+3:178822676-178822678 (121-123) 100% ->
+3:178829912-178829972 (124-182) 100%

aa.g 1  V F  P  A  C  E  M  P  A  G  S  H  P  E  S
+3:178818080
GAGTTCCACCTCCACGCTGAGATGCTGCAAGGGACCAATCTTGAGTCC

3  GACTTCCACCTCCACGCTGAGATGCTGCAAGGGACCAATCTTGAGTC

aa.c 1  F P  A  C  E  M  P  A  G  S  H  P  E  S

50  .  .  .  .  .  .  .  .  .  .  .  .  .  .  .  .  .
aa.g 17  S  E  L  T  Q  C  L  W  G  R  N  F  P  V  N  Y
+3:178818130
TCGGAATGACACAGTGCTTGGGTAGAAATTCACAGTGAAATGCTTA

53  TCGGAATGACACAGTGCTTGGGTAGAAATTCACAGTGAAATGCTTA

aa.c 17  S  E  L  T  Q  C  L  W  G  R  N  F  P  V  N  Y

100  .  .  .  .  .  .  .  .  .  .  .  .  .  .  .  .  .
aa.g 34  L  N  E  F  K  R  K  R  K  H  F  H
+3:178818180
CTTAAATGGAATTCAGTTAAAGTG...AGGAGGTG...GAGAAAAATTCACAGTGAAATGCTTA

103  CTTAAATGGAATTCAGAAA

aa.c 34  L  N  E  F  K  R  K  R  K  H  F  H

150  .  .  .  .  .  .  .  .  .  .  .  .  .  .  .  .  .
aa.g 45  G  R  G  N  R  K  F  Q  Q  I  S  N  S  W
+3:178829925
ATGAAAGAGAAACAGGAAATTCACAGTTAGCAGGTGATG

135  ATGAAAGAGAAACAGGAAATTCACAGTTAGCAGGTGATG
```

```
aa.c 45  G  R  G  N  R  K  F  Q  Q  I  S  N  S  W
```
GMAP can handle alternate strains (e.g., NCBI mouse has contigs from 9 strains)

<table>
<thead>
<tr>
<th>Reference strain (C57BL/6J)</th>
<th>Alternate strain (129_substrain)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 __________________________</td>
<td>1 __________________________</td>
</tr>
<tr>
<td>2 __________________________</td>
<td>2 __________________________</td>
</tr>
<tr>
<td>3 __________________________</td>
<td>3 __________________________</td>
</tr>
<tr>
<td>4 __________________________</td>
<td>4 __________________________</td>
</tr>
<tr>
<td>5 __________________________</td>
<td>5 __________________________</td>
</tr>
<tr>
<td>6 __________________________</td>
<td>6 __________________________</td>
</tr>
<tr>
<td>7 __________________________</td>
<td>7 __________________________</td>
</tr>
<tr>
<td>8 __________________________</td>
<td>8 __________________________</td>
</tr>
<tr>
<td>9 __________________________</td>
<td>9 __________________________</td>
</tr>
<tr>
<td>10 __________________________</td>
<td>10 __________________________</td>
</tr>
<tr>
<td>11 __________________________</td>
<td>11 __________________________</td>
</tr>
<tr>
<td>12 __________________________</td>
<td>12 __________________________</td>
</tr>
<tr>
<td>13 __________________________</td>
<td>13 __________________________</td>
</tr>
<tr>
<td>14 __________________________</td>
<td>14 __________________________</td>
</tr>
<tr>
<td>15 __________________________</td>
<td>15 __________________________</td>
</tr>
<tr>
<td>16 __________________________</td>
<td>16 __________________________</td>
</tr>
<tr>
<td>17 __________________________</td>
<td>17 __________________________</td>
</tr>
<tr>
<td>18 __________________________</td>
<td>18 __________________________</td>
</tr>
<tr>
<td>19 __________________________</td>
<td>19 __________________________</td>
</tr>
<tr>
<td>MT X _________________________</td>
<td>MT X _________________________</td>
</tr>
<tr>
<td>Y __________________________</td>
<td>Y __________________________</td>
</tr>
</tbody>
</table>
GMAP can handle extra chromosomes and alternate versions of chromosomes

<table>
<thead>
<tr>
<th>1</th>
<th>1U -</th>
<th>Unmapped or partially mapped contigs (concatenate them into “U” chromosomes)</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>2U -</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>3U -</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>4U -</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>5U -</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>6U -</td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>7U -</td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>8U -</td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>9U -</td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>10U -</td>
<td></td>
</tr>
<tr>
<td>11</td>
<td>11U -</td>
<td></td>
</tr>
<tr>
<td>12</td>
<td>12U -</td>
<td></td>
</tr>
<tr>
<td>13</td>
<td>13U -</td>
<td></td>
</tr>
<tr>
<td>14</td>
<td>14U -</td>
<td></td>
</tr>
<tr>
<td>15</td>
<td>15U -</td>
<td></td>
</tr>
<tr>
<td>16</td>
<td>16U -</td>
<td></td>
</tr>
<tr>
<td>17</td>
<td>17U -</td>
<td></td>
</tr>
<tr>
<td>18</td>
<td>18U -</td>
<td></td>
</tr>
<tr>
<td>19</td>
<td>19U -</td>
<td></td>
</tr>
<tr>
<td>MT</td>
<td>UnU</td>
<td></td>
</tr>
<tr>
<td>X</td>
<td>XU</td>
<td></td>
</tr>
<tr>
<td>Y</td>
<td>YU</td>
<td></td>
</tr>
</tbody>
</table>

Alternate version of chromosome 16 (from Celera)

Needs to be renamed (e.g., 16C)
### The concept of chromosomal subsets

<table>
<thead>
<tr>
<th>chrsubset “all”</th>
<th>chrsubset “vanilla”</th>
<th>chrsubset “chr10”</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>10</td>
</tr>
<tr>
<td>1U</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>3</td>
<td></td>
</tr>
<tr>
<td>2U</td>
<td>4</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>5</td>
<td></td>
</tr>
<tr>
<td>3U</td>
<td>6</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>7</td>
<td></td>
</tr>
<tr>
<td>4U</td>
<td>8</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>9</td>
<td></td>
</tr>
<tr>
<td>5U</td>
<td>10</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td></td>
<td></td>
</tr>
<tr>
<td>6U</td>
<td></td>
<td></td>
</tr>
<tr>
<td>7</td>
<td></td>
<td></td>
</tr>
<tr>
<td>7U</td>
<td></td>
<td></td>
</tr>
<tr>
<td>8</td>
<td></td>
<td></td>
</tr>
<tr>
<td>8U</td>
<td></td>
<td></td>
</tr>
<tr>
<td>9</td>
<td></td>
<td></td>
</tr>
<tr>
<td>9U</td>
<td></td>
<td></td>
</tr>
<tr>
<td>10</td>
<td></td>
<td></td>
</tr>
<tr>
<td>10U</td>
<td></td>
<td></td>
</tr>
<tr>
<td>11</td>
<td></td>
<td></td>
</tr>
<tr>
<td>11U</td>
<td></td>
<td></td>
</tr>
<tr>
<td>12</td>
<td></td>
<td></td>
</tr>
<tr>
<td>12U</td>
<td></td>
<td></td>
</tr>
<tr>
<td>13</td>
<td></td>
<td></td>
</tr>
<tr>
<td>13U</td>
<td></td>
<td></td>
</tr>
<tr>
<td>14</td>
<td></td>
<td></td>
</tr>
<tr>
<td>14U</td>
<td></td>
<td></td>
</tr>
<tr>
<td>15</td>
<td></td>
<td></td>
</tr>
<tr>
<td>15U</td>
<td></td>
<td></td>
</tr>
<tr>
<td>16</td>
<td></td>
<td></td>
</tr>
<tr>
<td>16U</td>
<td></td>
<td></td>
</tr>
<tr>
<td>17</td>
<td></td>
<td></td>
</tr>
<tr>
<td>17U</td>
<td></td>
<td></td>
</tr>
<tr>
<td>18</td>
<td></td>
<td></td>
</tr>
<tr>
<td>18U</td>
<td></td>
<td></td>
</tr>
<tr>
<td>19</td>
<td></td>
<td></td>
</tr>
<tr>
<td>19U</td>
<td></td>
<td></td>
</tr>
<tr>
<td>MT</td>
<td></td>
<td></td>
</tr>
<tr>
<td>UnU</td>
<td></td>
<td></td>
</tr>
<tr>
<td>X</td>
<td></td>
<td></td>
</tr>
<tr>
<td>XU</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Y</td>
<td></td>
<td></td>
</tr>
<tr>
<td>YU</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Alignments:
Alignment for path 1:
+7:27338303-27338700 (1-402) 99%
0

<table>
<thead>
<tr>
<th>aa.g</th>
<th>1</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>M</td>
</tr>
</tbody>
</table>
+7:27338353 TATAATAAATTATCTAAAGCAAATATCATCATTGGCTCTGAAATGCGT
      | 1  |
50

<table>
<thead>
<tr>
<th>aa.c</th>
<th>1</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>R</td>
</tr>
</tbody>
</table>
+7:27338403 TAATATAATAATTATCTAAAGCAAATATCATCATTGGCTCTGAAATGCGT
      | 3  |
100

Alignments:
Alignment for path 2:
+7:27338650 AGTAATT XCACTGCATTT XCTCTTCTCCTTCTCAATCAGT
      | 1  |
400

|
---

Mutations: H2R (CAT>CGT) [99], E17V (GAA>GTA) [144]
# Call: gmap -d NHGD_R35 -A
# Note: assumes that chromosome subset "vanilla", which excludes unmapped contigs, has been
# defined in file NHGD_R35.chrsuibset and set as the first entry (default)

>AA663092 (167 bp) ab73d01.sl Stratagene fetal retina 937202 Homo sapiens cDNA clone
e IMAGE:852577 3', mRNA sequence. 167 bp, mRNA, linear, EST 12-NOV-1997

Paths (0):

# Call: gmap -d NHGD_R35 -c all -A
# Note: assumes that chromosome subset "all", which includes unmapped contigs, has been
# defined in file NHGD_R35.chrsuibset

>AA663092 (167 bp) ab73d01.sl Stratagene fetal retina 937202 Homo sapiens cDNA clone
’e IMAGE:852577 3', mRNA sequence. 167 bp, mRNA, linear, EST 12-NOV-1997

Paths (1):

Path 1: query 1--167 (167 bp) => chr 17u:638,655--638,487 (-169 bp)
cDNA direction: indeterminate
Genomic pos: NHGD_R35:2,734,186,332--2,734,186,164 (- strand)
Accessions: NT_079556.1:54,369--54,537 (out of 142595 bp)
Number of exons: 1
Coverage: 100.0
Percent identity: 98.8 (167 matches, 0 mismatches, 2 indels, 0 unknowns)
Non-intron gaps: 0 openings, 0 bases in cdna; 1 openings, 2 bases in genome
Translation: 1..163 (56 aa)
Mutations: delX40 [117]

Alignments:

Alignment for path 1:

```
-17u:638655-638487 (1-167) 100%

0 . . . . . . . . . . . . . . .
-17u:638655
CTGUGGACCTGCAGTCTTCGTTAGGAAACAGAGTCTTCGACCCCTGCACTGCCAGCA
1 CTGUGGACCTGCAGTCTTCGTTAGGAAACAGAGTCTTCGACCCCTGCACTGCCAGCA

-17u:638605
AAAGCGGTTAAGCGATGCGACAGAGCTCGGCTCGACGACAGCACAGAGCTC
1 AAAGCGGTTAAGCGATGCGACAGAGCTCGGCTCGACGACAGCACAGAGCTC

-17u:638555
CATGCATCGGAGGGTCTTCNACTCGGTTCGAGAAGAACACGGCCCTCCGC
101 CATGCATCGGAGGGTCTTCNACTCGGTTCGAGAAGAACACGGCCCTCCGC

-17u:638505
ATGCACTCGGAGGGTCTCCG
149 ATGCACTCGGAGGGTCTCCG
```

aa.g 1 L W D C S L R T V S A A V T W Q Q
aa.c 1 L W D C S L R T V S A A V T W Q Q

```
18 N G V K Q C T R V C V D D S Q S P
-17u:638605
AAAGCGGTTAAGCGATGCGACAGAGCTCGGCTCGACGACAGCACAGAGCTC
1 AAAGCGGTTAAGCGATGCGACAGAGCTCGGCTCGACGACAGCACAGAGCTC

18 N G V K Q C T R V C V D D S Q S P
```
GMAP has procedures for detecting and mapping chimeras

BLAT alignment: BF743818

GMAP alignment: BF743818
# Call: gmap -d NHGD_R35 -x 50 -A cdna.fa
# Note: chimera mode is currently being improved, and will be available in a forthcoming release of GMAP

>BG183185 (392 bp) RST2056 Athersys RAGE Library Homo sapiens cDNA, mRNA sequence. 392 bp, mRNA, linear, EST 21-Apr-2001
Paths (2): *** Possible chimera with exon-exon boundary (sense) at 255 (donor_prob = 0.999, acceptor_prob = 0.997) -- alternative has 150 matches, 1 mismatches, 1 indels **

Path 1: query 105--255 (151 bp) => chr 6:89,568,052--89,536,212 (3184 bp) cdna direction: sense
Genomic pos: NHGD_R35:1,154,828,926--1,154,797,086 (- strand)
Accessions: NT_007299.12:27,299,664--27,331,504 (out of 33500716 bp)
Number of exons: 2
Coverage: 38.5
Percent identity: 98.7 (150 matches, 1 mismatches, 1 indels, 0 unknowns)
Non-intron gaps: 0 openings, 0 bases in cdna; 1 openings, 1 bases in genome
Translation: 105..253 (51 aa)
Mutations:

Path 2: query 256--392 (137 bp) => chr 12:50,787,430--50,787,566 (137 bp) cdna direction: indeterminate
Genomic pos: NHGD_R35:2,162,297,313--2,162,297,449 (+ strand)
Accessions: NT_029419.10:14,644,469--14,644,605 (out of 38627316 bp)
Number of exons: 1
Coverage: 34.9
Percent identity: 96.4 (132 matches, 5 mismatches, 0 indels, 0 unknowns)
Translation: 138..387 (45 aa)
Mutations: S34N (AGC>AAC) [356]

Alignments:
Alignment for path 1:
-6:89568052--89568002 (105-154) 98% ->
-6:89536312--89536212 (155-255) 100%

Alignment for path 2:
-6:50787430-50787566 (256-392) 96%

aa.g 1 A K E V S H E M D G L I F Q P T G
-6:89568052 GCCCAAGAGGTGAGCGATGAGAAATGGAGAATATATATATATATGCTCAGTGGG
aa.c 1 A T E V S H E M D G L I F Q P T G

aa.g 18 K Y K P G R C D D I L K W K
-6:89568002 AGTA...TAGAAATACAAACCCTTGCTGATGTGATGATATGATTTGGGAATTTATG
aa.c 18 K Y K P G R C D D I L K W K

aa.g 32 F P S L N S V D F R L K I T R M
-6:89536272 AGCTCCGCGATCTGAAATTTCTGTGATTTCTGCTCATAAAATAACAAGAATG
aa.c 32 F P S L N S V D F R L K I T R M

aa.g 48 G G E G
-6:89536222 GGAGGAGGAGG
aa.c 48 G G E G

Alignment for path 2:
Retrieval of relevant genomic map information

Overlapping intervals are retrieved in logarithmic time using interval index trees (IITs)
gmap.mapinfo

Non-intron gaps: 2 openings, 2 bases in cdna; 0 openings, 0 bases in genome
Mutations: R15G (AGG>GGG) [83], S19T (AGG>ACC) [96], V21A (GTT>GCT) [102], V26G (GTT>GCT) [117]

Maps:
Map hits for path 1 (23):
  affy.affy.map  NGU133A.AFFX-HUGMAPDH/M33197.5_at gb|M33197; M33197 Human glyceraldehyde-3-phosphate dehydrogenase (GAPDH) mRNA, complete cds (_5, _M, _3 represent transcript regions 5 prime, Middle, and 3 prime respectively)
  affy.affy.map  NGU133B.AFFX-HUGMAPDH/M33197.5_at gb|M33197; M33197 Human glyceraldehyde-3-phosphate dehydrogenase (GAPDH) mRNA, complete cds (_5, _M, _3 represent transcript regions 5 prime, Middle, and 3 prime respectively)

Map hits for path 2 (0):

---

# Call: gmap -d NHGD_R35 -m refseq.mrna.map cdna.fa
# Note: assumes that refseq.mrna.map.iit has been built with iit_store and installed in NHGD_R35.map

>AA015654 (198 bp)  ze29cl2lr Soares retina N2b4HR Homo sapiens cDNA clone IMAGE:36
0404 5' similar to gb:X01677 GLYCERALDEHYDE 3-PHOSPHATE DEHYDROGENASE, LIVER (HUMAN);
mRNA sequence. 198 bp, mRNA, linear, EST 29-NOV-1996
Paths (2):
Path 1: query 10--198 (189 bp) => chr 12:6,514,230--6,516,138 (1909 bp)
cDNA direction: sense
Genomic pos: NHGD_R35:2,118,024,113--2,118,026,021 (+ strand)
Accessions: NT_009759.115:6,498,230--6,500,138 (out of 7043293 bp)
Number of exons: 3
Coverage: 95.5
Percent identity: 97.4 (184 matches, 1 mismatches, 4 indels, 1 unknowns)
Non-intron gaps: 3 openings, 3 bases in cdna; 1 openings, 1 bases in genome
Translation: 10..194 (62 aa)
Mutations: S8* (TCA>TGA) [33]

Path 2: query 20--198 (179 bp) => chr X:46,056,423--46,056,247 (-177 bp)
cDNA direction: indeterminate
Genomic pos: NHGD_R35:3,080,740,920--3,080,740,744 (- strand)
Accessions: NT_079573.2:9,149,802--9,149,978 (out of 12039763 bp)
Number of exons: 1
Coverage: 90.4
Percent identity: 93.8 (166 matches, 9 mismatches, 2 indels, 2 unknowns)
Non-intron gaps: 2 openings, 2 bases in cdna; 0 openings, 0 bases in genome
Translation: 41..194 (52 aa)
Mutations: R15G (AGG>GGG) [83], S19T (AGG>ACC) [96], V21A (GTT>GCT) [102], V26G (GTT>GCT) [117]

Maps:
Map hits for path 1 (1):
  refseq.mrna.map  NM_002046.2 Homo sapiens glyceraldehyde-3-phosphate dehydrogenase (GAPDH), mRNA
Map hits for path 2 (0):

---

# Call: gmap -d NHGD_R35 -m affy.affy.map cdna.fa
# Note: assumes that affy.affy.map.iit has been built with iit_store and installed in NHGD_R35.map

>AA015654 (198 bp)  ze29cl2lr Soares retina N2b4HR Homo sapiens cDNA clone IMAGE:36
0404 5' similar to gb:X01677 GLYCERALDEHYDE 3-PHOSPHATE DEHYDROGENASE, LIVER (HUMAN);
mRNA sequence. 198 bp, mRNA, linear, EST 29-NOV-1996
Paths (2):
Path 1: query 10--198 (189 bp) => chr 12:6,514,230--6,516,138 (1909 bp)
cDNA direction: sense
Genomic pos: NHGD_R35:2,118,024,113--2,118,026,021 (+ strand)
Accessions: NT_009759.115:6,498,230--6,500,138 (out of 7043293 bp)
Number of exons: 3
Coverage: 95.5
Percent identity: 97.4 (184 matches, 1 mismatches, 4 indels, 1 unknowns)
Non-intron gaps: 3 openings, 3 bases in cdna; 1 openings, 1 bases in genome
Translation: 10..194 (62 aa)
Mutations: S8* (TCA>TGA) [33]

Path 2: query 20--198 (179 bp) => chr X:46,056,423--46,056,247 (-177 bp)
cDNA direction: indeterminate
Genomic pos: NHGD_R35:3,080,740,920--3,080,740,744 (- strand)
Accessions: NT_079573.23:9,149,802--9,149,978 (out of 12039763 bp)
Number of exons: 1
Coverage: 90.4
Percent identity: 93.8 (166 matches, 9 mismatches, 2 indels, 2 unknowns)
Relative alignment of ESTs

Provide both a full-length mRNA and one or more ESTs

- Use this alignment to mark codon boundaries on the genome.
- Use the marked codon boundaries on the genome to translate the ESTs correctly.
# Call: gmap -d NHGD_R35 -A -w fulllength.fa cdna.fa
# or    cat cdna.fa | gmap -d NHGD_R35 -A -w fulllength.fa
# Note that fulllength.fa is assumed to translate from Met to Stop
# Here fulllength.fa is NM_002046.1, as found in the gmap.mapinfo example
# Compare this protein translation with that in gmap.example

>AA015654 (198 bp)  ze29c11.r1 Soares retina N2b4HR Homo sapiens cDNA clone IMAGE:36
0404 5' similar to gb:X01677 GLYCERALDEHYDE 3-PHOSPHATE DEHYDROGENASE, LIVER (HUMAN);, mRNA sequence. 198 bp, mRNA, linear, EST 29-NOV-1996
Paths (1):
  Path 1: query 10--198 (189 bp) => chr 12:6,514,230--6,516,138 (1909 bp)
cDNA direction: sense
Genomic pos: NHGD_R35:2,118,024,113--2,118,026,021 (+ strand)
Number of exons: 3
Coverage: 95.5
Percent identity: 97.4 (184 matches, 1 mismatches, 4 indels, 1 unknowns)
Non-intron gaps: 3 openings, 3 bases in cDNA; 1 openings, 1 bases in genome
Protein coords: 1..53
Mutations:

<table>
<thead>
<tr>
<th>aa.g</th>
<th>aa.c</th>
</tr>
</thead>
</table>
| 0    | 1
| 1    | M    |
| 50   | 7    |
| 100  | 21   |
| 150  | 37   |

Mutations:

<table>
<thead>
<tr>
<th>Position</th>
<th>Mutation</th>
<th>aa.g</th>
<th>aa.c</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>aa.g</td>
<td>aa.c</td>
</tr>
<tr>
<td>1</td>
<td>M</td>
<td>G</td>
<td>G</td>
</tr>
<tr>
<td>50</td>
<td>G</td>
<td>V</td>
<td>V</td>
</tr>
<tr>
<td>100</td>
<td>A</td>
<td>A</td>
<td>A</td>
</tr>
<tr>
<td>150</td>
<td>F</td>
<td>I</td>
<td>I</td>
</tr>
</tbody>
</table>

Mutations:

<table>
<thead>
<tr>
<th>Position</th>
<th>Mutation</th>
<th>aa.g</th>
<th>aa.c</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>6514230</td>
<td>6514230</td>
</tr>
<tr>
<td></td>
<td></td>
<td>12</td>
<td>12</td>
</tr>
<tr>
<td></td>
<td></td>
<td>6514278</td>
<td>6514278</td>
</tr>
<tr>
<td></td>
<td></td>
<td>1632</td>
<td>1632</td>
</tr>
<tr>
<td></td>
<td></td>
<td>90</td>
<td>90</td>
</tr>
<tr>
<td></td>
<td></td>
<td>191</td>
<td>191</td>
</tr>
</tbody>
</table>

Mutations:
Cross-species alignment with GMAP

GMAP being used to align Brassica ESTs to Arabidopsis genome
(by Nicolas Tsesmetzis, John Innes Centre, UK)
Outline

• Introduction
• Usage
• Algorithm
• Utility programs
• Basic setup
• Advanced setup
Indexing step: Store genome as long oligomers

Genome

Oligos (24-mers)

Store each oligo in an index table for fast lookup

<table>
<thead>
<tr>
<th>24-mer</th>
<th>Genomic loc</th>
</tr>
</thead>
<tbody>
<tr>
<td>24-mer</td>
<td>Genomic loc</td>
</tr>
<tr>
<td>24-mer</td>
<td>Genomic loc</td>
</tr>
<tr>
<td>24-mer</td>
<td>Genomic loc</td>
</tr>
</tbody>
</table>
Genomic mapping: Finding candidate genomic segments

Note: this is the basic idea. The actual algorithm uses a *minimal sampling strategy* to confirm candidates and to resolve multiple candidates.
SSAHA (Ning et al., 2001) indexing scheme

Index offset file
- 4 bytes
- 0
- index
- index+1
- $4^k$

Genomic position file
- 4 bytes per position
- position
- position
- position
- ...
# Oligomer choice

<table>
<thead>
<tr>
<th>Program</th>
<th>Oligomer size</th>
<th>Uniqueness</th>
<th>SSAHA index size</th>
</tr>
</thead>
<tbody>
<tr>
<td>BLAT</td>
<td>11-mers</td>
<td>0.1%</td>
<td>16.8 MB</td>
</tr>
<tr>
<td>SSAHA</td>
<td>14-mers</td>
<td>22.5%</td>
<td>1.1 GB</td>
</tr>
<tr>
<td>GMAP</td>
<td>24-mers</td>
<td>97.0%</td>
<td>1.1 PB</td>
</tr>
</tbody>
</table>
Oligomer counts in the unmasked human genome

Total number of oligomers observed

Total number of distinct oligomers

Possible space of distinct oligomers

Distinct oligomers that are unique

Distinct oligomers that are repeated

Oligomer size

Number of oligomers

Total number of oligomers observed

Total number of distinct oligomers

Possible space of distinct oligomers

Distinct oligomers that are unique

Distinct oligomers that are repeated

Oligomer size

Number of oligomers
One possible implementation: Hash table

- **Hash offset file**
  - 4 bytes
  - 0
  - hashvalue
  - hashvalue + 1
  - $2^{24}$

- **Genomic position file**
  - 6 bytes
  - 2 bytes
  - 4 bytes per position

<table>
<thead>
<tr>
<th>oligomer</th>
<th>count</th>
<th>position</th>
<th>...</th>
<th>position</th>
</tr>
</thead>
<tbody>
<tr>
<td>oligomer</td>
<td>count</td>
<td>position</td>
<td>...</td>
<td>position</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>oligomer</td>
<td>count</td>
<td>position</td>
<td>...</td>
<td>position</td>
</tr>
</tbody>
</table>

- Hash offset file contains hashvalues and oligo offsets.
- Genomic position file stores oligomer counts, positions, and possibly other metadata.
- The hash table is used to efficiently look up genomic positions based on hash values.
A faster implementation: Double lookup of 12-mers

Look for genomic positions that are 12 nucleotides apart.
Approximate alignment via oligomer chaining

Build a detailed index table of overlapping 8-mers in the genomic segment

Scan 8-mers from the cDNA sequence to find mappings to the genomic segment

Perform dynamic programming to find the best chain of mappings:
Nucleotide-level alignment: Filling in the gaps

cDNA

end sequence alignment

global alignment

sandwich alignment

end sequence alignment

Gap due to mismatch

Gap due to intron

constrain here

find intron boundaries

constrain here

genome
Sandwich dynamic programming for finding splice sites
Probabilistic microexon detection

- Microexons have been found in 0.5–1.6% of mRNA sequences in various species. Volfovsky and colleagues (2003) have developed a method of scanning for microexons.

- **GMAP** incorporates a probabilistic version of this method. For a microexon with \( e \) nucleotides, the probability \( p \) that the microexon matches somewhere in an intron of length \( L \) and is surrounded by two canonical introns is

\[
p = 1 - \left[ 1.0 - \left( \frac{1}{4} \right)^m \right]^L
\]

\[
m = \frac{\log(1 - (1 - p)^{1/L})}{\log(4)}
\]

Here \( m = e + 8 \) to include the two GT-AG pairs.
Outline

• Introduction
• Usage
• Algorithm
• Utility programs
• Basic setup
• Advanced setup
Utility programs included with GMAP package

- get-genome
  - retrieves genomic intervals
  - translates chromosomal coordinates
  - lists chromosomal and contig intervals
- an implementation of interval index trees
  - iit_store, iit_get, iit_dump
- compression and uncompression routines
  - gmap_compress, gmap_uncompress
- gmap_setup, gmapindex: process genomes
get-genome --help
Usage: get-genome [OPTIONS...] -d genome [genome:range, or
get-genome [OPTIONS...] -d genome chromosome:range, or
get-genome [OPTIONS...] -d genome contig:range]
where
  range is startposition..endposition (endpos < startpos means - strand)
  or startposition+length (+ strand)
  or startposition+-length (- strand)

Input options
-D, --dir=STRING  Data directory
-d, --db=STRING  Database (e.g., NHGD)
-R, --rel=STRING  Release of database

Output options
-S, --altstrain  Show sequence for all strains (in addition to reference)
-s, --strain=STRING  Show sequence ...
-G, --fullgenome  Use full (uncompressed) version of genome
-h, --header=STRING  Desired header line

Dump options
-L, --chromosomes  List all chromosomes with universal coordinates
-l, --contigs  List all contigs with universal coordinates

Help options
-V, --version  Show version
-?, --help  Show this help message

ala> get-genome -d NHGD_R35 12:6,514,000..6,514,100
>NHGD_R35:2118023883--2118023983AAGACGGGCGGAGAGAAACCCGGGAGGCTAGGGACGGCCTGAAGGCGGCAGGGGCGGGCGCAGGCCGGATGTGTTCGCGCCGCTGCGGGGTGGGCCCGGGC
ala> get-genome -d NHGD_R35 12:6,514,000..6,514,100
>NHGD_R35:2118023883--2118023983 rcTCACCTGGCGACGCAAAAGAAGATGCGGCTGACTGTCGAACAGGAGGAGCAGAGAGCGAAGCGGGAGGCTGCGGGCTCAATTTATAGAAACCGGGGGCGC
ala> get-genome -d NHGD_R35 -C 12:6,514,000..6,514,100
2118023883--2118023983  12:6514000      12:6514100
Interval index trees are used extensively in GMAP.

In a database of $n$ intervals, can retrieve all $k$ intervals that overlap a query interval in $O(k + \lg n)$ time.
# The input to the IIT utilities is a FASTA file, where each header
# contains a segment name, start position, end position, and optional tag.

Segments may have zero or more lines of annotation. Annotations may
be arbitrarily long, so they may contain sequences, as shown in segment B.

C 6 10
This segment has two lines of annotation.
The next one has none.

# This command creates the file iit-example.iit
ala> iit-store -o iit-example iit-example.fa

# Can retrieve all intervals that overlap 10..20
ala> iit_get iit-example iit-example.fa

# Can retrieve interval by name
ala> iit_get iit-example A

# Can retrieve all intervals that overlap the point 17
ala> iit_get iit-example 17

# Can restrict results to a particular type
ala> iit_get iit-example 10 20 129_substrain
ala> iit_get iit-example 10 20 129_substrain

# Can dump contents of an .iit file
ala> iit_dump iit-example.iit

# Can restrict results to a particular type
ala> iit_get iit-example 10 20 129_substrain
ala> iit_get iit-example 10 20 129_substrain

# Can retrieve interval by name
ala> iit_get iit-example A

Segments may have zero or more lines of annotation. Annotations may
be arbitrarily long, so they may contain sequences, as shown in segment B.
Outline

- Introduction
- Usage
- Algorithm
- Utility programs
- Basic setup
- Advanced setup
How to compile the GMAP package

1. Download package from Web site

2. Edit config.site file
   - prefix: binaries will install into prefix/bin
   - with_gmapdb: location of genome directory

3. Build the program
   ./configure
   make
   make check
   make install
How to process a genome for GMAP

1. Download fasta files for the genome

2. Create coords.txt file by running either
   – md_coords seq_contig.md [for NCBI]
   – fa_coords fastafile... [e.g., UCSC, Ensembl]

3. Run gmap_setup -d genome fastafile...

4. Follow instructions given by gmap_setup to move the files to the genome directory
Now look at coords.txt (you may edit it manually if you wish) and then run

gmap_setup -d <db_name> <fastafiles>

where you decide on the db_name, and the fastafiles contain the sequences for the contigs listed in /usr/seqdb/hum/ncbi/seq_contig.md

Note that gmap_setup can take a while to index a large genome

Contig mapping information has been written to file coords.txt.

For future reference, some of this setup process can be specified on the command-line, as follows:

```bash
md_coords -o coords.txt -c 6,2,3,4,5,7 -U 1 -A 1 /usr/seqdb/hum/ncbi/seq_contig.md
```
Chromosome 21 has universal coordinates 2937981576..2984925898
Chromosome 22 has universal coordinates 2984925899..3034460608
Chromosome MT has universal coordinates 3034460609..3034667926
Chromosome X has universal coordinates 3034684498..3189508761
Chromosome XU has universal coordinates 3189508762..3191027929
Chromosome Y has universal coordinates 3191027930..3248729620

Making genome file...
Writing alternate strain file...
Done writing alternate strain file
Genome length is 3248729620 nt
Trying to allocate 304568403*4 bytes of memory...succeeded. Building genome in memory.
Writing contig NT_034403.3 to universal coordinates 145784053..146093292
Writing contig NT_034398.4 to universal coordinates 143352668..143634199
Writing contig NT_037485.3 to universal coordinates 25572994..28427660

... 
Writing contig NT_079597 to universal coordinates 1550693209..1554641063
Writing contig NG_002392 to universal coordinates 1097772297..1097911478 (alternate at strain DR52)
Writing contig NG_002433 to universal coordinates 1097764272..1097914718 (alternate at strain DR53)

Making index offsets file...
Indexing offsets of oligomers in genome (every 6), position 1000000
Indexing offsets of oligomers in genome (every 6), position 2000000
Indexing offsets of oligomers in genome (every 6), position 3000000

... 
Indexing offsets of oligomers in genome (every 6), position 3246000000
Indexing offsets of oligomers in genome (every 6), position 3247000000
Indexing offsets of oligomers in genome (every 6), position 3248000000

Making index positions file...
Trying to allocate 503479310*4 bytes of memory...succeeded. Building positions in memory.
Indexing positions of oligomers in genome (every 6), position 1000000
Indexing positions of oligomers in genome (every 6), position 2000000
Indexing positions of oligomers in genome (every 6), position 3000000

... 
Indexing positions of oligomers in genome (every 6), position 3246000000
Indexing positions of oligomers in genome (every 6), position 3247000000
Indexing positions of oligomers in genome (every 6), position 3248000000

Making chrsubset file...
Making version file...

Computation started at Sat Jun 25 07:44:20 2005
Computation ended at Sat Jun 25 08:35:11 2005
======================================================================================
Genome db files have been written in this directory
To install them for use by gmap, do the following
  mkdir /usr/gmap/data/genomes/NHGID_R35
  mv -f NHGD_R35.* /usr/gmap/data/genomes/NHGID_R35
  chmod 644 /usr/gmap/data/genomes/NHGID_R35/NHGID_R35.*
======================================================================================
# contig coordinates
altstrain

NT_077402.1 1:176290
NT_077911.1 1:217281..257582
NT_077912.1 1:357583..511231
NT_077913.3 1:561232..1539795
NT_077914.2 1:56109796..1601817
NT_004350.17 1:16501818..2666382
NT_004351.16 1:2716383..3846740
NT_004547.16 1:33067641..5336899
NT_012197.17 1:53969000..13052128
NT_077382.3 1:13065219..13302468
NT_004873.16 1:13352469..16870964
NT_004610.17 1:16920965..25522933
NT_037485.3 1:25572934..28427660
NT_004538.16 1:28627661..29692175
NT_004511.17 1:29697176..45842628
NT_032997.7 1:45896269..103575297
NT_019723.17 1:103625928..120409198
NT_086586.1 1:120459199..120648377
NT_077389.3 1:120698738..121097476
NT_079485.2 1:141387477..141551846
NT_077392.2 1:141851837..141979909
NT_077931.1 1:142029100..142199768
NT_004434.17 1:142299769..143302667
NT_034398.4 1:143352669..143634199
NT_034400.3 1:143684200..144750022
NT_077936.2 1:144800023..154169503
NT_079483.1 1:145290931..154365379
NT_034401.5 1:145513740..145734052
NT_034403.3 1:145784053..146093292
NT_086599.1 1:146134293..146272716
NT_004487.17 1:146322719..202654346
NT_086602.17 1:202704365..202875288
NT_021877.17 1:202952289..219608088
NT_077939.1 1:219658089..220008054
NT_004559.12 1:220130452..231518252
NT_004836.16 1:231568253..245217961
NT_032968.7 1:245227962..245442847
NT_079497.2 1:245448339..245519972
NT_079496.1 1:245762559..257922865
NT_079486.1 1:282676..480423
NT_077970.1 1:480424..689365
NT_079501.1 1:689366..731866
NT_079796.3 1:731869..2369608
NT_086608.1 1:2369609..2514794
NT_079490.1 1:2514795..2540788
NT_077964.3 1:2540789..2733140
NT_079489.1 1:2733141..2760672
NT_077962.1 1:2760673..2874732
NT_086609.1 1:2874729..301543
NT_079487.1 1:301544..3063368
NT_077949.1 1:3063369..3247131
NT_077567.3 1:3247132..3510193
NT_077569.2 1:3510194..3677111
NT_008705.15 1:3677111..4804234

---

NT_001807.4 MT:1..16571
NT_086925.1 X:1..34821
NT_078115.2 X:8482..171384
NT_028413.7 X:251385..101755
NT_086929.1 X:1067558..1104113
Opening file /usr/seqdb/hum/ensembl/Homo_sapiens.NCB135.apr.dna.chromosome.1.fa
Header line: >1 dna:chromosome chromosome:NCBI35:1:1:245522847:1
Processing contig 1 at chromosomal coordinates 1:1..245522847 (length = 245522847 nt)

Opening file /usr/seqdb/hum/ensembl/Homo_sapiens.NCB135.apr.dna.chromosome.10.fa
Header line: >10 dna:chromosome chromosome:NCBI35:10:1:135413628:1
Processing contig 10 at chromosomal coordinates 10:1..135413628 (length = 135413628 nt)

Opening file /usr/seqdb/hum/ensembl/Homo_sapiens.NCB135.apr.dna.chromosome.11.fa
Header line: >11 dna:chromosome chromosome:NCBI35:11:1:134452384:1
Processing contig 11 at chromosomal coordinates 11:1..134452384 (length = 134452384 nt)

Opening file /usr/seqdb/hum/ensembl/Homo_sapiens.NCB135.apr.dna.chromosome.9.fa
Header line: >9 dna:chromosome chromosome:NCBI35:9:1:138429268:1
Processing contig 9 at chromosomal coordinates 9:1..138429268 (length = 138429268 nt)

Opening file /usr/seqdb/hum/ensembl/Homo_sapiens.NCB135.apr.dna.chromosome.DR52.fa
Header line: >DR52 dna:chromosome chromosome:NCBI35:DR52:32511423:32650604:1
Processing contig DR52 at chromosomal coordinates DR52:32511423..32650604 (length = 170972699 nt)

Opening file /usr/seqdb/hum/ensembl/Homo_sapiens.NCB135.apr.dna.chromosome.DR53.fa
Header line: >DR53 dna:chromosome chromosome:NCBI35:DR53:32503398:32653844:1
Processing contig DR53 at chromosomal coordinates DR53:32503398..32653844 (length = 170972699 nt)

*** Possible errors: ***
First contig in chromosome DR52 starts at position 32511423
First contig in chromosome DR53 starts at position 32503398
First contig in chromosome Y starts at position 2692882

Some of the errors above may be addressed by specifying the contigs to be on alternate strains of existing chromosomes, rather than on independent alternate chromosomes.
You may make the appropriate changes in coords.txt, by adding an alternate strain in column 3, and specifying an existing chromosome in column 2.

Contig mapping information has been written to file coords.txt.
3 possible errors were found (listed above)

Now look at coords.txt (you may edit it manually if you wish)
and then run

gmap_setup -d <db_name> <fastafiles>
where you decide on the db_name, and the fastafiles are the same you used in running this fa_coords program.
# contig coordinates altstrain

1 1:1..245522847
2 1:1..133413628
3 1:1..134452384
4 1:1..132449811
5 1:1..1414234780
6 1:1..106368585
7 1:1..100338915
8 1:1..88827254
9 1:1..78774742
10 1:1..76117153
11 1:1..63811651
12 2:1..243018229
13 2:1..62435964
14 2:1..46944323
15 2:1..49554710
16 3:1..199505740
17 3:1..191411218
18 4:1..180857866
19 5:1..170975699
20 6:1..158628139
21 7:1..146274826
22 8:1..138429268
23 9:1..138429268

#DR52 DR52:32511423..203484121
#DR53 DR53:32503398..203476096
MT MT:1..16571
X X:1..154824264
Y Y:2692882..60394572
Outline

• Introduction
• Usage
• Algorithm
• Utility programs
• Basic setup
• Advanced setup
Processing a genome for GMAP

FASTA files of contigs or chromosomes

seq_contig.md file from NCBI

fa_coords or md_coords (or create/edit manually)

cords.txt

fa_coords

NT_039169.4 1:3000001..22662952
NT_039170.4 1:22712953..75359766
NT_039200.1 1:60944100..61121618  NOD

md_coords

. .

coords.txt

Gmap_setup (which calls gmapindex)

Genome files (May be used on any computer architecture)
Organization of the GMAP genome database

During build of GMAP package, we specified a GMAPDB directory

Example: /usr/local/share/gmap

This directory contains one subdirectory per genome. We suggest including the release as part of the genome name. To point to the most recent release as a default, can create symbolic links:

- NHGD  ->  NHGD_R35/
- NHGD_R32/
- NHGD_R33/
- NHGD_R34/
- NHGD_R35/
- NMGD  ->  NMGD_R34/
- NMGD_R32/
- NMGD_R33/
- NMGD_R34
- arath  ->  arath_R1/
- arath_R1

Users may then point to the default version using -d NHGD, for example, to gmap or get-genome.
## Genome files created by GMAP_SETUP

Directory `$(GMAPDB)/NMGD_R34/`

<table>
<thead>
<tr>
<th>Size</th>
<th>File Name</th>
</tr>
</thead>
<tbody>
<tr>
<td>47798933</td>
<td>NMGD_R34.altstrain.iit</td>
</tr>
<tr>
<td>76</td>
<td>NMGD_R34.altstrain.type</td>
</tr>
<tr>
<td>1534</td>
<td>NMGD_R34.chromosome</td>
</tr>
<tr>
<td>2558</td>
<td>NMGD_R34.chromosome.iit</td>
</tr>
<tr>
<td>503</td>
<td>NMGD_R34.chrsbset</td>
</tr>
<tr>
<td>2081080</td>
<td>NMGD_R34.contig</td>
</tr>
<tr>
<td>2318283</td>
<td>NMGD_R34.contig.iit</td>
</tr>
<tr>
<td>1054620024</td>
<td>NMGD_R34.genomecomp</td>
</tr>
<tr>
<td>67108868</td>
<td>NMGD_R34.idxoffsets</td>
</tr>
<tr>
<td>1820516148</td>
<td>NMGD_R34.idxpositions</td>
</tr>
<tr>
<td>9</td>
<td>NMGD_R34.version</td>
</tr>
</tbody>
</table>

Entries highlighted in **red** may be edited, to change run-time behavior.
The following genome files can be customized to change the run-time behavior of gmap:

The .version file specifies how the genome is labeled in the output:

```
ala> more NMGD_R34.version
NMGD_R34
```

The .chrsubset file defines chromosomal subsets. The first one is the default. A blank line signifies all chromosomes. A line beginning with a "+" sign means to include the following chromosomes. A line beginning with a "-" sign means to exclude the following chromosomes:

```
ala> more NMGD_R34.chrsubset
>vanilla
+1,2,3,4,5,6,7,8,9,10,11,12,13,14,15,16,17,18,19,MT,X,Y
>all
>chr1
+1
>chr1U
+1U
>chr2
+2
>chr2U
+2U
```

...
General information about GMAP

Description of algorithm and test results:

Sequence analysis

**GMAP: a genomic mapping and alignment program for mRNA and EST sequences**

Thomas D. Wu¹,* and Colin K. Watanabe²

¹Department of Bioinformatics and ²Department of Corporate Information Technology, Genentech, Inc., South San Francisco, CA 94080, USA

Received on February 24, 2004; revised on January 27, 2005; accepted on February 4, 2005
Advance Access publication February 22, 2005

Requirements: Unix, C compiler, Perl

Source code and sample databases: www.gene.com/share/gmap

Contact: Thomas Wu, Genentech, Inc., twu@gene.com