



# Sequence Coordinates

## 0- vs 1- base

Bob Milius, PhD  
Bioinformatics Research  
NMDP<sup>®</sup>/Be The Match<sup>®</sup>

# Two Things

## 1. How do we number the sequence?

- start with 0 or 1 ?

## 2. What do we include?

- Position includes the thing
  - Inclusive
  - Closed
  - [ ]
- Position excludes the thing
  - Exclusive
  - Open
  - ( )
- Example: these all mean the same thing
  - 0-base inclusive start, exclusive end
  - 0-base, half open
  - [0, )

# Two major systems in use

1-base 12345678

CAGGAGCA

0-base 01234567

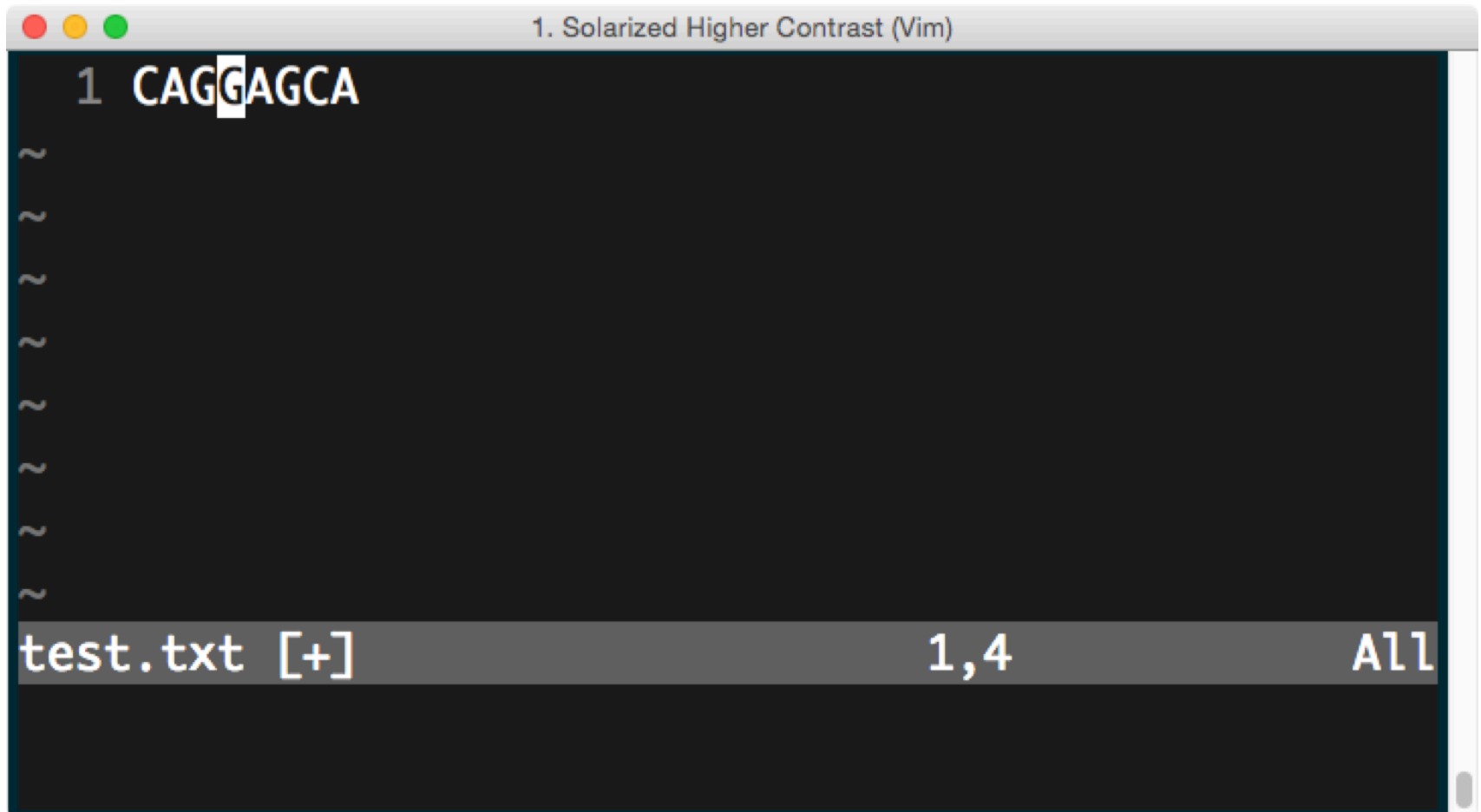
- **1-base, inclusive start, inclusive end**
  - Describing entire sequence : start = 1, end = 8
    - [1,8]
  - Describing GGAG subsequence: start = 3, end = 6
    - [3,6]
- **0-base, inclusive start, exclusive end**
  - Describing entire sequence: start = 0, end = 8
    - [0,8]
  - Describing GGAG subsequence: start = 2, end = 6
    - [2,6)

# Another way to think of it: character vs interval counting

character	1	2	3	4	5	6	7	8	
	C	A	G	G	A	G	C	A	
interval	0	1	2	3	4	5	6	7	8

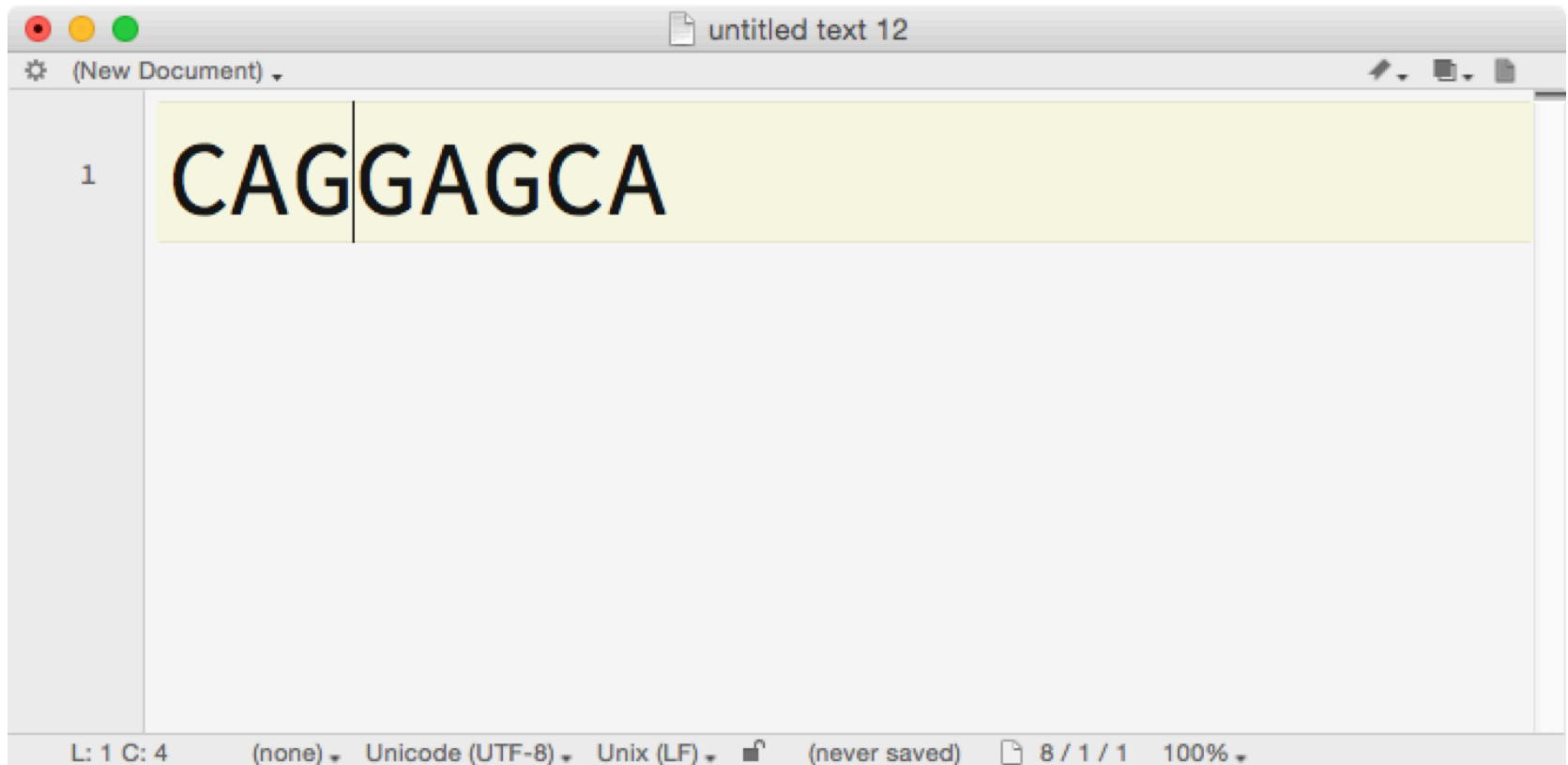
- Character = counts each base
- Interval = counts the spaces between each base
  - Also known as interbase counting

# Character based



```
1 CAGGAGCA
~
~
~
~
~
~
~
~
test.txt [+] 1,4 All
```

# Interval based



# 1-base, inclusive start, inclusive end

- Easier for humans to understand – it's how we think
- Many systems use this already
  - HGVS
  - VCF
  - ClinVar (uses HGVS)
  - Genbank files
  - UCSC genome browser
  - IPD-IMGT/HLA
- Akin to cursor position in early text editors
- Programming a little tricky
  - need to add or subtract 1 to calculations
  - $\text{length} = \text{end} - \text{start} + 1$

# 0-base, inclusive start, exclusive end (interval counting)

- Easier for computers to consume
- Many systems use this already  
(mostly newer and backend systems)
  - Global Alliance for Genomics and Health (GA4GH) API
  - ClinGen Data Model
  - Genbank database & ASN files
  - BED, BAM files
  - UCSC database
  - HML 1.0
- Akin to cursor positioning in modern text editors
- Programming easier
  - $\text{length} = \text{end} - \text{start}$



# What Do Programming Languages Use for Array Indexing?

---

- 1-based
  - FORTRAN, SASL, MATLAB, Smalltalk
- 0-based
  - C, Perl, Python, Java, Ruby, JavaScript

# Easy to convert if all we are talking about is sequences and subsequences

character	1	2	3	4	5	6	7	8	
	C	A	G	G	A	G	C	A	
interval	0	1	2	3	4	5	6	7	8

Character  $\rightarrow$  Interval = subtract 1 from start

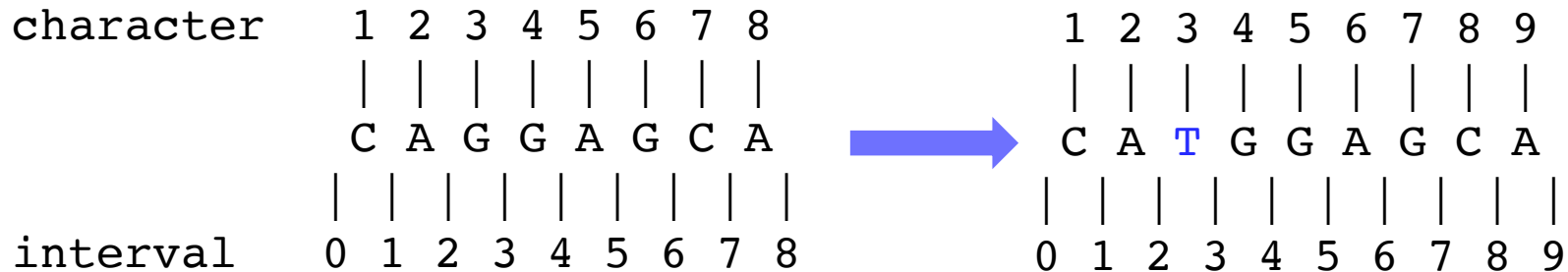
– Character (1-base)

- Start = 3
- End = 6
- [3,6]

– Interval (0-base)

- Start = 2
- End = 6
- [2,6)

# But trickier when describing variations



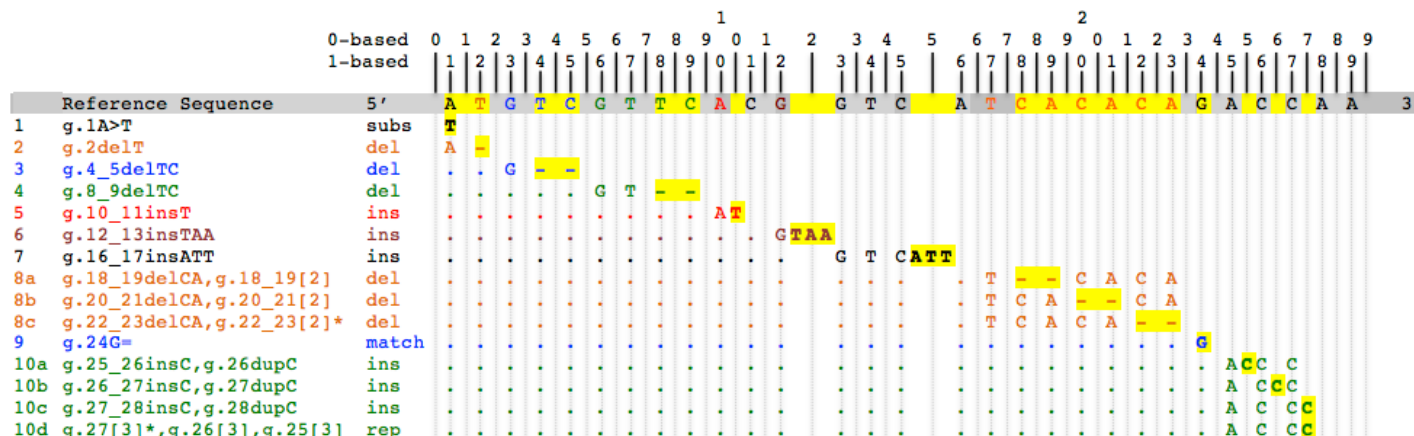
- How to describe an insertion?
- **Character:** If we say the insertion happen at position 3, we need to know if the insertion is before or after that position
  - If we have a rule that insertions are before a position, then we can put insertions at the beginning of a sequence but not the end.
  - Use substitution: eg. describe a substitution at position 3 from G to TG
  - HGVS and VCF have different rules in describing variants
- **Interval:** insertions are easily described, it happens in the space between the positions

# Great discussion describing variants with different coordinate systems

[http://datamodel.clinicalgenome.org/development/allele/discussion/coordinate\\_numbering.html](http://datamodel.clinicalgenome.org/development/allele/discussion/coordinate_numbering.html)

- The Alignment Method
    - Based on the numbering used in [VCF](#)
  - The Variant Method
    - Based on the numbering used in [HGVS expressions](#)
- 1-base
- 
- The Interval Method
    - Based on numbering intervals as in [BED files](#)
- 0-base

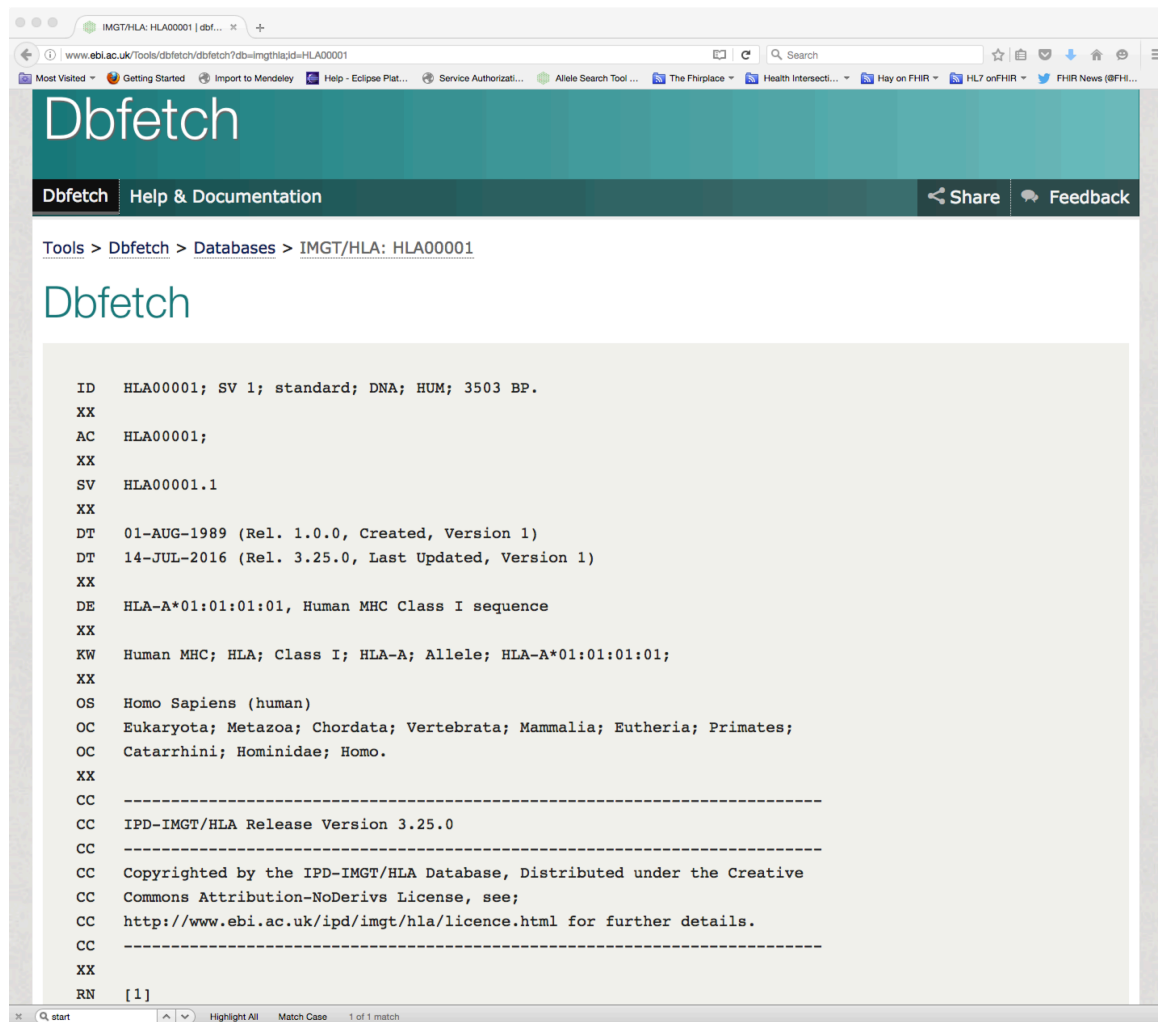
# How the different methods describe variation



	HGVS	Alignment Format (1-based)				Variant Format (1-based)				Interval Format (0-based)			
		Start	End	Ref Allele	Alt Allele	Start	End	Ref Allele	Alt Allele	Start	End	Ref Allele	Alt Allele
1	g.1A>T	1	1	A	T	1	1	A	T	0	1	A	T
2	g.2delT	1	2	AT	A	2	2	T	-	1	2	T	-
3	g.4_5delTC	3	5	GTC	G	4	5	TC	-	3	5	TC	-
4	g.8_9delTC	6	9	GTTC	GT	8	9	TC	-	7	9	TC	-
5	g.10_11insT	10	10	A	AT	10	11	-	T	10	10	-	T
6	g.12_13insTAA	12	12	G	GTAA	12	13	-	TAA	12	12	-	TAA
7	g.16_17insATT	13	15	GTC	GTCATT	15	16	-	ATT	15	15	-	ATT
8a	g.18_19delCA,g.18_19[2]	17	19	TCA	T	18	19	CA	-	17	19	CA	-
8b	g.20_21delCA,g.20_21[2]	17	21	TCACA	TCA	20	21	CA	-	19	21	CA	-
8c	g.22_23delCA,g.22_23[2]*	17	23	TCACACA	TCACA	22	23	CA	-	21	23	CA	-
9	g.24G=	24	24	G	G	24	24	G	G	23	24	G	G
10a	g.25_26insC,g.26dupC	25	26	A	AC	25	26	-	C	25	25	-	C
10b	g.26_27insC,g.27dupC	26	27	C	CC	26	27	-	C	26	26	-	C
10c	g.27_28insC,g.28dupC	27	28	C	CC	27	28	-	C	27	27	-	C
10d	g.27[3]*,g.26[3],g.25[3]	27	27	C	CC	27	28	-	C	27	27	-	C

\* These are the HGVS recommended representation for the canonically equivalent representations of item 8 and 10, respectively. These representations may appear in practice, but should be canonicalized so that they are seen as the same. HGVS recommends a right-justified representation and VCF recommends a left-justified representation, but neither is guaranteed.

# If you are using IPD-IMGT/HLA as a reference...

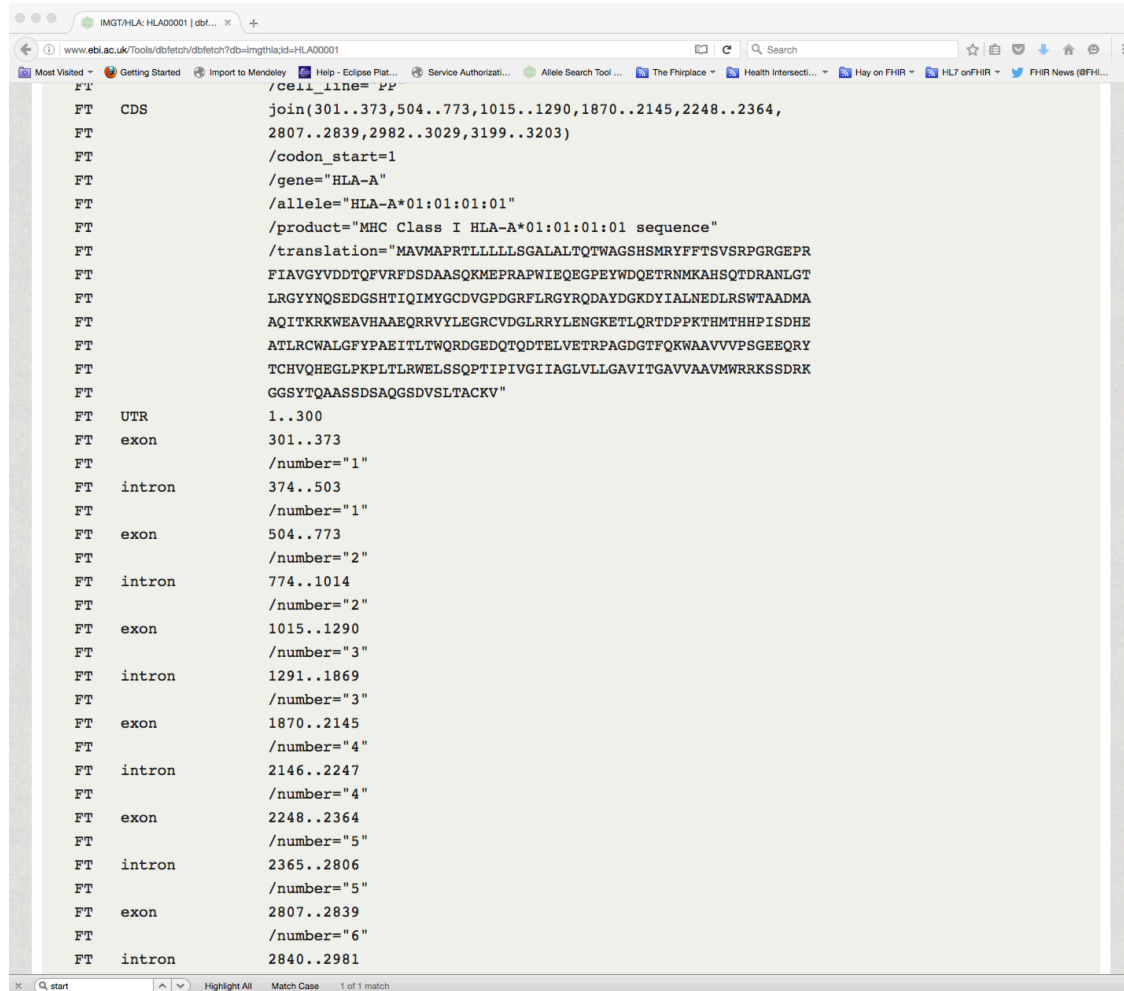


The screenshot shows a web browser window with the URL [www.ebi.ac.uk/Tools/dbfetch/dbfetch?db=imgthla;id=HLA00001](http://www.ebi.ac.uk/Tools/dbfetch/dbfetch?db=imgthla;id=HLA00001). The page title is "Dbfetch" and the breadcrumb trail is "Tools > Dbfetch > Databases > IMGT/HLA: HLA00001". The main content area displays the following text:

```
ID   HLA00001; SV 1; standard; DNA; HUM; 3503 BP.
XX
AC   HLA00001;
XX
SV   HLA00001.1
XX
DT   01-AUG-1989 (Rel. 1.0.0, Created, Version 1)
DT   14-JUL-2016 (Rel. 3.25.0, Last Updated, Version 1)
XX
DE   HLA-A*01:01:01:01, Human MHC Class I sequence
XX
KW   Human MHC; HLA; Class I; HLA-A; Allele; HLA-A*01:01:01:01;
XX
OS   Homo Sapiens (human)
OC   Eukaryota; Metazoa; Chordata; Vertebrata; Mammalia; Eutheria; Primates;
OC   Catarrhini; Hominidae; Homo.
XX
CC   -----
CC   IPD-IMGT/HLA Release Version 3.25.0
CC   -----
CC   Copyrighted by the IPD-IMGT/HLA Database, Distributed under the Creative
CC   Commons Attribution-NoDerivs License, see;
CC   http://www.ebi.ac.uk/ipd/imgt/hla/licence.html for further details.
CC   -----
XX
RN   [1]
```

The browser's address bar and tabs are visible at the top, and the page has a teal header with "Dbfetch" and navigation links like "Help & Documentation", "Share", and "Feedback".

# If you are using IPD-IMGT/HLA as a reference...



The screenshot shows a web browser window with the URL [www.ebi.ac.uk/Tools/dbfetch/dbfetch?db=imgthla;id=HLA00001](http://www.ebi.ac.uk/Tools/dbfetch/dbfetch?db=imgthla;id=HLA00001). The page displays the genomic structure and sequence of the HLA-A\*01:01:01:01 allele. The structure is shown as a list of features (FT) with their coordinates and descriptions. The sequence is shown as a single line of text.

```
FT CDS join(301..373,504..773,1015..1290,1870..2145,2248..2364,
FT 2807..2839,2982..3029,3199..3203)
FT /codon_start=1
FT /gene="HLA-A"
FT /allele="HLA-A*01:01:01:01"
FT /product="MHC Class I HLA-A*01:01:01:01 sequence"
FT /translation="MAVMAPRTL LLLSGALALTQTWAGSHSMRYFFTSVSRPGRGEPR
FT FIAGVYVDDTQFVRFDSDAASQKMEPRAPWIEQGPPEYWDQETRNMKASHQTDRLNLT
FT LRGYYNQSEDSHTIQIMYGCDVGPDRFLRGYRQDAYDGKDYIALNEDLSWTAADMA
FT AQITKRKWEAVHAAEQRRVYLEGRCDVGLRRYLENGKETLQRTDPPKTHMTHHPISDHE
FT ATLRCWALGFYPAEITLTWQRDGEDQTQDTELVETRPAGDGTQKWAAVVVPSEGEQRY
FT TCHVQHEGLPKPLTLRWELSSQPTIPVGIAGLVLLGAVITGAVVAVMWRKSSDRK
FT GGSYTAASSDSDAQGSVDVSLTACKV"
FT UTR 1..300
FT exon 301..373
FT /number="1"
FT intron 374..503
FT /number="1"
FT exon 504..773
FT /number="2"
FT intron 774..1014
FT /number="2"
FT exon 1015..1290
FT /number="3"
FT intron 1291..1869
FT /number="3"
FT exon 1870..2145
FT /number="4"
FT intron 2146..2247
FT /number="4"
FT exon 2248..2364
FT /number="5"
FT intron 2365..2806
FT /number="5"
FT exon 2807..2839
FT /number="6"
FT intron 2840..2981
```

# From IPD-IMGT/HLA to 0-base

```

FT   UTR      1..300
FT   exon     301..373
FT           /number="1"
FT   intron   374..503
FT           /number="1"
FT   exon     504..773
FT           /number="2"
FT   intron   774..1014
FT           /number="2"
FT   exon     1015..1290
FT           /number="3"
FT   intron   1291..1869
FT           /number="3"
FT   exon     1870..2145
FT           /number="4"
FT   intron   2146..2247
FT           /number="4"
FT   exon     2248..2364
FT           /number="5"
FT   intron   2365..2806
FT           /number="5"
FT   exon     2807..2839
FT           /number="6"
FT   intron   2840..2981
FT           /number="6"
FT   exon     2982..3029
FT           /number="7"
FT   intron   3030..3198
FT           /number="7"
FT   exon     3199..3203
FT           /number="8"
FT   UTR      3204..3503
    
```



```

FT   UTR      0..300
FT   exon     300..373
FT           /number="1"
FT   intron   373..503
FT           /number="1"
FT   exon     503..773
FT           /number="2"
FT   intron   773..1014
FT           /number="2"
FT   exon     1014..1290
FT           /number="3"
FT   intron   1290..1869
FT           /number="3"
FT   exon     1869..2145
FT           /number="4"
FT   intron   2145..2247
FT           /number="4"
FT   exon     2247..2364
FT           /number="5"
FT   intron   2364..2806
FT           /number="5"
FT   exon     2806..2839
FT           /number="6"
FT   intron   2840..2981
FT           /number="6"
FT   exon     2981..3029
FT           /number="7"
FT   intron   3029..3198
FT           /number="7"
FT   exon     3198..3203
FT           /number="8"
FT   UTR      3203..3503
    
```



# From IPD-IMGT/HLA to 0-base

```

FT   UTR      1..300
FT   exon     301..373
FT           /number="1"
FT   intron   374..503
FT           /number="1"
FT   exon     504..773
FT           /number="2"
FT   intron   774..1014
FT           /number="2"
FT   exon     1015..1290
FT           /number="3"
FT   intron   1291..1869
FT           /number="3"
FT   exon     1870..2145
FT           /number="4"
FT   intron   2146..2247
FT           /number="4"
FT   exon     2248..2364
FT           /number="5"
FT   intron   2365..2806
FT           /number="5"
FT   exon     2807..2839
FT           /number="6"
FT   intron   2840..2981
FT           /number="6"
FT   exon     2982..3029
FT           /number="7"
FT   intron   3030..3198
FT           /number="7"
FT   exon     3199..3203
FT           /number="8"
FT   UTR      3204..3503
    
```



```

FT   UTR      0..300
FT   exon     300..373
FT           /number="1"
FT   intron   373..503
FT           /number="1"
FT   exon     503..773
FT           /number="2"
FT   intron   773..1014
FT           /number="2"
FT   exon     1014..1290
FT           /number="3"
FT   intron   1290..1869
FT           /number="3"
FT   exon     1869..2145
FT           /number="4"
FT   intron   2145..2247
FT           /number="4"
FT   exon     2247..2364
FT           /number="5"
FT   intron   2364..2806
FT           /number="5"
FT   exon     2806..2839
FT           /number="6"
FT   intron   2840..2981
FT           /number="6"
FT   exon     2981..3029
FT           /number="7"
FT   intron   3029..3198
FT           /number="7"
FT   exon     3198..3203
FT           /number="8"
FT   UTR      3203..3503
    
```

# Exon 2 of HLA-A\*01:01:01:01 (HLA00001.1)

CACATCCGTGTCCCGGCCCGGCCGCGGGGAGCCCCGCTTCATCG  
CCGTGGGCTACGTGGACGACACGCAGTTCGTGCGGTTCGACAGC  
GACGCCGCGAGCCAGAAGATGGAGCCGCGGGGCGCCGTGGATAGA  
GCAGGAGGGGCGCGAGTATTGGGACCAGGAGACACGGAATATGA  
AGGCCCACTCACAGACTGACCGAGCGAACCTGGGGACCCTGCGC  
GGCTACTACAACCAGAGCGAGGACG

- IPD-IMGT/HLA (uses 1-base)
  - start = 504, end=773
  - [504,773]
- In HML (uses 0-base, interval)
  - start= 503, end= 773
  - [503,773)

# Describing a reference from IPD-IMGT/HLA in HML 1—base to 0-base conversion

```
<consensus-sequence date="2016-09-01">
  <reference-database
    name="IPD-IMGT/HLA"
    description="IPD-IMGT/HLA Database"
    version="3.25"
    availability="public"
    curated="true"
    uri="http://www.ebi.ac.uk/ipd/imgt/hla/" >
    <reference-sequence
      id="ref1"
      name="HLA-A*01:01:01:01"
      start="0" end="3503"
      accession="HLA00001.1"
      uri="http://www.ebi.ac.uk/Tools/dbfetch/dbfetch?db=imgthla;id=HLA00001.1"/>
    </reference-database>
  <consensus-sequence-block
    reference-sequence-id="ref1"
    start="503" end="773"
    description="Exon 2 of HLA-A*01:01:01:01"
    phase-set="1">
    <sequence>
CACATCCGTGTCCCGGCCCGCGGGGAGCCCCGCTTCATCGCCGTGGGCTACGTGGACGACACGCAGTTCGTGCGGTTTCGACAGCGACGCCGCGAGCCAGAAGATGGAGCCGCG
GGCGCCGTGGATAGAGCAGGAGGGGCCGGAGTATTGGGACCAGGAGACACGGAATATGAAGGCCCACTCACAGACTGACCGAGCGAACCTGGGGACCCTGCGCGGCTACTACAACCA
GAGCGAGGACG
    </sequence>
  </consensus-sequence-block>
</consensus-sequence>
```

# Describing a reference from IPD-IMGT/HLA in HML 1–base to 0-base conversion

```
<reference-sequence
  id="ref1"
  name=" HLA-A*01:01:01:01"
  start="0" end="3503"
  accession="HLA00001.1"
  uri="http://www.ebi.ac.uk/Tools/dbfetch/dbfetch?db=imgthla;id=HLA00001.1"/>
</reference-database>
<consensus-sequence-block
  reference-sequence-id="ref1"
  start="503" end="773"
  description="Exon 2 of HLA-A*01:01:01:01"
  phase-set="1">
  <sequence>CACATCCGTG...GGACG</sequence>
</consensus-sequence-block>
```

# Exon 2 of HLA-A\*01:01:01:01 (HLA00001.1) with C→T substitution

CATATCCGTGTCCCGGCCCGGCCGCGGGGAGCCCCGCTTCATCG  
CCGTGGGCTACGTGGACGACACGCAGTTCGTGCGGTTCGACAGC  
GACGCCGCGAGCCAGAAGATGGAGCCGCGGGGCGCCGTGGATAGA  
GCAGGAGGGGCGCGAGTATTGGGACCAGGAGACACGGAATATGA  
AGGCCCACTCACAGACTGACCGAGCGAACCTGGGGACCCTGCGC  
GGCTACTACAACCAGAGCGAGGACG

- **SequenceBlock** is always relative to the reference
- In HML (uses 0-base, interval)
  - start= 503, end= 773

# Exon 2 of HLA-A\*01:01:01:01 (HLA00001.1) with C→T substitution

CA**T**ATCCGTGTCCCGGCCCGGCCGCGGGGAGCCCCGCTTCATCG  
CCGTGGGCTACGTGGACGACACGCAGTTCGTGCGGTTCGACAGC  
GACGCCGCGAGCCAGAAGATGGAGCCGCGGGGCGCCGTGGATAGA  
GCAGGAGGGGCGCGAGTATTGGGACCAGGAGACACGGAATATGA  
AGGCCCACTCACAGACTGACCGAGCGAACCTGGGGACCCTGCGC  
GGCTACTACAACCAGAGCGAGGACG

- **SequenceBlock** is always relative to the reference
- In HML (uses 0-base, interval)
  - start= **503**, end= **773**

```
<variant  
  reference-bases="C"  
  alternate-bases="T"  
  start="505" end="506" />
```

# Describing a reference from IPD-IMGT/HLA in HML 1–base to 0-base conversion

```
<reference-sequence
  id="ref1"
  name=" HLA-A*01:01:01:01"
  start="0" end="3503"
  accession="HLA00001.1"
  uri="http://www.ebi.ac.uk/Tools/dbfetch/dbfetch?db=imgthla;id=HLA00001.1"/>
</reference-database>
<consensus-sequence-block
  reference-sequence-id="ref1"
  start="503" end="773"
  description="Exon 2 of HLA-A*01:01:01:01"
  phase-set="1">
  <sequence>CACATCCGTG...GGACG</sequence>
  <variant reference-bases="C" alternate-bases="T" start="505" end="506">
</consensus-sequence-block>
```

# Exon 2 of HLA-A\*01:01:01:01 (HLA00001.1) with C deletion

CA**CA**TCCGTGTCCCGGCCCGGCCGCGGGGAGCCCCGCTTCATCG  
CCGTGGGCTACGTGGACGACACGCAGTTCGTGCGGTTCGACAGC  
GACGCCGCGAGCCAGAAGATGGAGCCGCGGGCGCCGTGGATAGA  
GCAGGAGGGGCCGGAGTATTGGGACCAGGAGACACGGAATATGA  
AGGCCCACTCACAGACTGACCGAGCGAACCTGGGGACCCTGCGC  
GGCTACTACAACCAGAGCGAGGACG

- **consensus-sequence-block coordinate is always relative to the reference**
- In HML (uses 0-base, interval)
  - start= **503**, end= **773**



# Exon 2 of HLA-A\*01:01:01:01 (HLA00001.1) with C deletion

CA**A**TCCGTGTCCCGGCCCGGCCGCGGGGAGCCCCGCTTCATCG  
CCGTGGGCTACGTGGACGACACGCAGTTCGTGCGGTTCGACAGC  
GACGCCGCGAGCCAGAAGATGGAGCCGCGGGGCGCCGTGGATAGA  
GCAGGAGGGGCCGGAGTATTGGGACCAGGAGACACGGAATATGA  
AGGCCCACTCACAGACTGACCGAGCGAACCTGGGGACCCTGCGC  
GGCTACTACAACCAGAGCGAGGACG

- **consensus-sequence-block coordinate is always relative to the reference**
- In HML (uses 0-base, interval)
  - start= **503**, end= **773**

```
<variant  
  reference-bases="CA"  
  alternate-bases="A"  
  start="505" end="507" />
```

# Describing a reference from IPD-IMGT/HLA in HML 1–base to 0-base conversion

```
<reference-sequence
  id="ref1"
  name=" HLA-A*01:01:01:01"
  start="0" end="3503"
  accession="HLA00001.1"
  uri="http://www.ebi.ac.uk/Tools/dbfetch/dbfetch?db=imgthla;id=HLA00001.1"/>
</reference-database>
<consensus-sequence-block
  reference-sequence-id="ref1"
  start="503" end="773"
  description="Exon 2 of HLA-A*01:01:01:01"
  phase-set="1">
  <sequence>CACATCCGTG...GGACG</sequence>
  <variant reference-bases="CA" alternate-bases="A" start="505" end="507">
</consensus-sequence-block>
```

# A word about what's allowed in <sequence> in HML

```
<xs:simpleType name="iupac-bases">
  <xs:annotation>
    <xs:documentation>
      Nucleotide bases representing sequence ambiguity. Primary nucleotides: A, C, G,
      T (DNA). "Wildcard" nucleotides: M, R, W, S, Y, K, V, H, D, B, X, N. Wildcard
      nucleotides may be used if they are acceptable in the context in which they
      appear. The default is to use all upper case letters. The full specification of
      the IUPAC codes may be found here:
      (http://nar.oxfordjournals.org/content/13/9/3021.short) Cornish-Bowden A.
      Nomenclature for incompletely specified bases in nucleic acid sequences:
      recommendations 1984. Nucleic Acids Res. 1985; 13:3021-3030. The bases of the
      sequence string are restricted to the upper and lower case versions of the
      nucleotides specified above. Data: ---- - Nucleotide sequence in DNA alphabet
      (string, required)
    </xs:documentation>
  </xs:annotation>
  <xs:restriction base="xs:string">
    <xs:pattern value="([\sACGTUMRWSYKVHDBXNacgtumrwsykvhdbxn])+"/>
    <xs:minLength value="1"/>
  </xs:restriction>
</xs:simpleType>
```

No gaps!!

# Things to remember

---

- HML 1.0 uses the 0-base, interval counting system
- To convert a 1-base, closed system to 0-based interval, just subtract 1 from the start
- Reference sequence must be gapless, unambiguous, and can be easily dereferenced
- Consensus-sequence-block coordinates are relative to the reference coordinates
- Variant coordinates are relative to the reference coordinates
- Variations are treated as substitutions
- Given a reference sequence, and a window into that reference, and a variant description in that window, we can easily reconstruct the actual sequence being reported

# More reading...

- ClinGen discussion on coordinate numbering
  - [http://datamodel.clinicalgenome.org/development/allele/discussion/coordinate\\_numbering.html](http://datamodel.clinicalgenome.org/development/allele/discussion/coordinate_numbering.html)
- Question: What Are The Advantages/Disadvantages Of One-Based Vs. Zero-Based Genome Coordinate Systems
  - <https://www.biostars.org/p/6373/>
- Tutorial: Cheat Sheet For One-Based Vs Zero-Based Coordinate Systems
  - <https://www.biostars.org/p/84686/>
- Coordinate Transforms
  - [http://genomewiki.ucsc.edu/index.php/Coordinate\\_Transforms](http://genomewiki.ucsc.edu/index.php/Coordinate_Transforms)
- Genome Coordinate Conventions
  - <http://alternateallele.blogspot.com/2012/03/genome-coordinate-conventions.html>
- Genome Coordinate Cheat Sheet
  - <http://alternateallele.blogspot.com/2012/03/genome-coordinate-cheat-sheet.html>

# And more...

- Global Alliance for Genomics and Health (GA4GH)
  - Discussion
    - <https://github.com/ga4gh/schemas/issues/121>
  - API
    - <https://ga4gh-schemas.readthedocs.io/en/latest/schemas/common.proto.html>
  -
- Ensembl is using the GA4GH API as part of its RESTful services
  - <https://rest.ensembl.org/documentation/info/gavariants>



# Thank you!

Questions?