# The Zeroth law of bioinformatics[1]

The software programmer and his wife relocate to a new house in the city. Counting the cardboard boxes the movers dropped off, they discover one box is missing.

"How can this be?" the wife wonders. "I'm sure I've packed forty boxes. Are you sure you've counted them right?".

"Yes, thirty nine it is", says the husband. "Don't believe me? let's count them together: zero, one, two, three,..."

## *The Problem*

Here's the start of chromosome 1 (hg18):

```
>chr1
taaccctaaccctaaccctaaccctaaccctaaccctaaccctaacccta
accctaaccctaaccctaaccctaaccctaaccctaaccctaaccctaac
```

What should be the genomic position for the first nucleotide (the yellow 't') ?
Answering 1 (one) would make you a AWK programmer.
Answering 0 (zero) would make you a C programmer.

Let's consider genomic location of the first ten nucleotides in chromosome 1: `taaccctaac`

With zero-based notation, we would use the following genomic location:

```
chr1 0    9
```

With one-based notation, we would use the following genomic location:

```
chr1 1    10
```

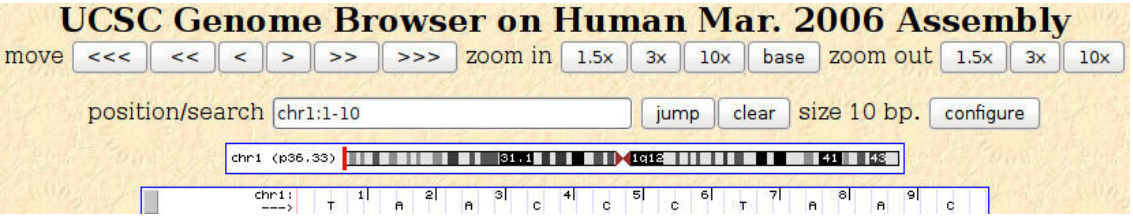
The question is:

Which program / database / browser / other public resouce uses what notation ?

1 Compare with Fowler's Zeroth law of Thermodynamics (http://en.wikipedia.org/wiki/Zeroth_law_of_thermodynamics) and Asimov's Zeroth law of robotics (http://en.wikipedia.org/wiki/Laws_of_Robotics#Zeroth_Law_added)

## *UCSC Browser*

When ***displaying*** genomic positions, the UCSC browser uses one-based notation:




Note that for the genomic location `chr:1-10` we get the the first ten nucleotides `taaccctaac`.

In the UCSC browser, there is no position 'zero'. Requesting the genomic location `chr:0-9`, the UCSC browser will display genomic location `chr:1-9` (ignoring our zero value).

## *UCSC Browser's tables*

Unlike the UCSC's web interface, all the tables and data files in the UCSC browser use zero-based notation for the *start* coordinate, but a one-based notation for the *end* coordinate.

When downloading raw data (using UCSC's table-browser), the following warning is displayed (when you click on the describe table schema button):

**Schema for sno/miRNA - C/D and H/ACA Box snoRNAs, scaRNAs, and microRNAs**

**Database:** hg18 **Primary Table:** wgRna **Row Count:** 1,059
**Format description:** Browser extensible data

| field | example | SQL type | info | description |
|---|---|---|---|---|
| bin | 585 | smallint(6) | range | Indexing field to speed chromosome range queries. |
| chrom | chr1 | varchar(255) | values | Reference sequence chromosome or scaffold |
| chromStart | 20228 | int(10) unsigned | range | Start position in chromosome |
| chromEnd | 20366 | int(10) unsigned | range | End position in chromosome |
| name | hsa-mir-1302-2 | varchar(255) | values | Name of item |
| score | 960 | int(10) unsigned | range | Score from 0-1000 |
| strand | + | char(1) | values | + or - |
| thickStart | 0 | int(10) unsigned | range | Start of where display should be thick (start codon) |
| thickEnd | 0 | int(10) unsigned | range | End of where display should be thick (stop codon) |
| type | miRna | varchar(255) | values | |

**Sample Rows**

| bin | chrom | chromStart | chromEnd | name | score | strand | thickStart | thickEnd | type |
|---|---|---|---|---|---|---|---|---|---|
| 585 | chr1 | 20228 | 20366 | hsa-mir-1302-2 | 960 | + | 0 | 0 | miRna |
| 593 | chr1 | 1092346 | 1092441 | hsa-mir-200b | 960 | + | 0 | 0 | miRna |
| 593 | chr1 | 1093105 | 1093195 | hsa-mir-200a | 960 | + | 0 | 0 | miRna |
| 593 | chr1 | 1094247 | 1094330 | hsa-mir-429 | 960 | + | 0 | 0 | miRna |
| 611 | chr1 | 3467118 | 3467214 | hsa-mir-551a | 480 | - | 0 | 0 | miRna |
| 654 | chr1 | 9134313 | 9134423 | hsa-mir-34a | 480 | - | 0 | 0 | miRna |
| 680 | chr1 | 12489886 | 12490038 | ACA59 | 960 | + | 0 | 0 | HAcaBox |
| 730 | chr1 | 19096151 | 19096229 | hsa-mir-1290 | 480 | - | 0 | 0 | miRna |
| 746 | chr1 | 21187393 | 21187512 | hsa-mir-1256 | 480 | - | 0 | 0 | miRna |
| 798 | chr1 | 28033498 | 28033664 | ACA35 | 960 | + | 0 | 0 | scaRna |

*Note: all start coordinates in our database are 0-based, not 1-based. See explanation here.*

The explanation “here” (http://genome.ucsc.edu/FAQ/FAQtracks#tracks1) says:

**Database/browser start coordinates differ by 1 base**

**Question:**
"I am confused about the start coordinates for items in the refGene table. It looks like you need to add "1" to the starting point in order to get the same start coordinate as is shown by the Genome Browser. Why is this the case?"
**Response:**
Our internal database representations of coordinates always have a zero-based start and a one-based end. We add 1 to the start before displaying coordinates in the Genome Browser. Therefore, they appear as one-based start, one-based end in the graphical display. The refGene.txt file is a database file, and consequently is based on the internal representation.

We use this particular internal representation because it simplifies coordinate arithmetic, i.e. it eliminates the need to add or subtract 1 at every step. Unfortunately, it does create some confusion when the internal representation is exposed or when we forget to add 1 before displaying a start coordinate. However, it saves us from much trickier bugs.

In summary, if you use a database dump file but would prefer to see the one-based start coordinates, you will always need to add 1 to each start coordinate.

## *Uploading Custom Tracks (BED files) to UCSC Browser*

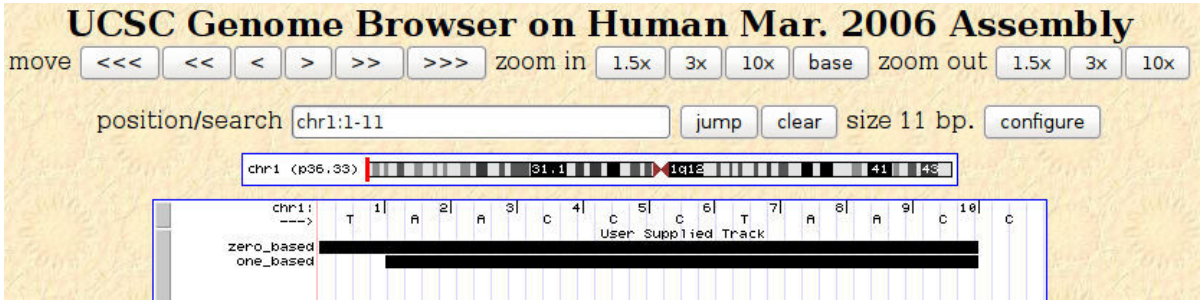
BED files (http://genome.ucsc.edu/FAQ/FAQformat#format1) are textual files containing genomic locations. As such, UCSC Genome Browser treats BED files as database files, and assumes they have a zero-based starting coordinate. When displaying the custom track, the browser will add 1 to all start coordinates (but not to the end coordinates).

Example:

Consider the following BED file:

```
chr1    0    10    zero_based
chr1    1    10    one_based
```

Uploading and displaying it in the UCSC genome browser:




Notes:

- The start coordinate for both sequences was incremented by 1:
  `zero_based` starts on position 1,
  `one_based` starts on position 2.
- The end coordinate for both sequences was not altered.
- The first sequence (`zero_based`) is 10 nucleotides long.
- The second sequence (`one_based`) is 9 nucleotides long.

## *Sanger's miRBase*

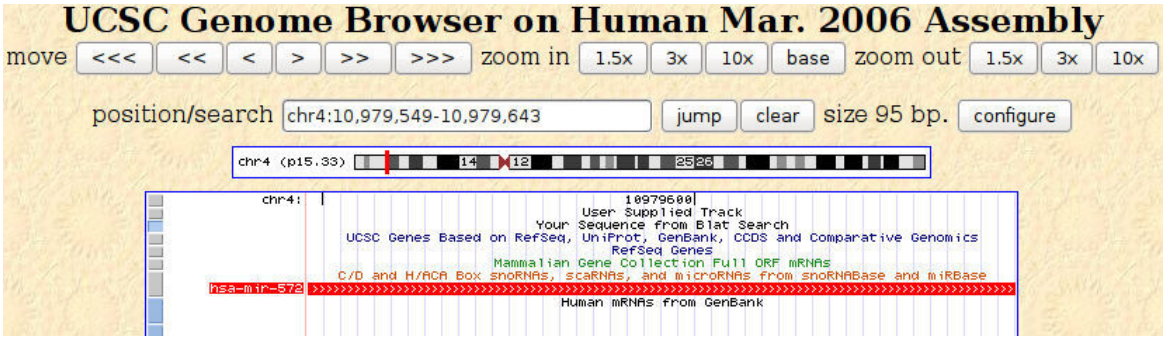
On sanger's miRBase website (http://microrna.sanger.ac.uk/sequences/ftp.shtml), the GFF files with the genomic locations use one-based notation for both start and end coordinates.

An Example with hsa-mir-572

Relevant line from `hsa.gff` (downloaded from miRBase):

```
4  .  miRNA  10979549  10979643  .  +  .  ACC="MI0003579"; ID="hsa-mir-572";
```

hsa-mir-572 in UCSC Genome browser:




Relevant line from UCSC sno/miRNA track (using the table browser):

```
668   chr4   10979548   10979643   hsa-mir-572   960   +   0   0   miRna
```

Notes:

- miRBase uses one-based notation for both start, end coordinates
- UCSC displays data using one-based notation for both start, end coordinates
- UCSC stores data using zero-based notation for the start coordinate ( 10979548 in UCSC vs. 10979549 in miRBase).