Genome representation concepts

- At the simplest level of abstraction the genome is represented by a one dimensional "space" (lines)
- Genome is two stranded → a line corresponds to each strand
- Each strand has a polarity → each line has a direction
- Strands (lines) are paired
- The smallest unit is one base → one integer on the number line
- Annotations (features) are segments (coordinates) on each line

Genetic coordinates – brief overview

DNA two stranded and directional
But there is only one coordinate system

Coordinate systems

- 0 based → 0, 1, 2, ... 9
- 1 based → 1, 2, 3, ... 10

Typically

- 0 based are non-inclusive 10:20 → [10, 20)
- 1 based include both ends 10:20 → [10, 20]
Comparing coordinate systems

<table>
<thead>
<tr>
<th></th>
<th>1 based indexing</th>
<th>0 based indexing</th>
</tr>
</thead>
<tbody>
<tr>
<td>Third element</td>
<td></td>
<td></td>
</tr>
<tr>
<td>First ten</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Second ten</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Third ten</td>
<td></td>
<td></td>
</tr>
<tr>
<td>One base long interval starting at 10</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Length of an interval</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Five elements starting at 1000</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Empty interval</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Vote for what you think is better

Fundamental interval formats

- SAM/BAM – Sequence Alignment Map
- VCF/BCF → for variant calls
- BED/GFF → Gene Annotation representation
- BEDgraph, Wiggle → values over intervals

What is a genomic feature?

- Feature: a genomic region (interval) associated with a certain annotation (description).

  Typical attributes to describe a feature

  1. chromosome
  2. start
  3. end
  4. strand
  5. name

Why do we have so many variants? There is no good rational reason ... history I guess

Values on intervals

- A single value characterizes an entire interval → score (value) for the interval

- Continuous values → different value for each base of the interval → analogous to a series of 1bp long intervals

Different data representation formats
Two commonly used formats

• **BED** – UCSC genome browser → 0 based non-inclusive → also used to display tracks in the genome browser (US “standard”), (variants: **bigBed, bedgraph**)

• **GFF** – Sanger institute in Great Britain → 1 based inclusive indexing system (“European standard”), (variants: **GTF, GFF 2.0**)

---

**BED format**

Search for BED format

Tab separated 3 required and 9 optional columns. Lower numbered filed must be filled.

1. **chrom** (name of the chromosome, sequence id)
2. **chromStart** (starting position on the chromosome)
3. **chromEnd** (end position of the chromosome, note this base is not included)
4. **name** (feature name)
5. **score** (between 0 and 1000)
6. **strand** (+ or -)
7. **thickStart** (the starting position at which the feature is drawn thickly)
8. **thickEnd** (the ending position at which the feature is drawn thickly)
9. **itemRGB** (RGB color → 255, 0, 0 display color of the data contained)
10. **blockCount** (the number of blocks (exons) in the BED line.)
11. **blockSizes** (a comma-separated list of the block sizes)
12. **blockStarts** (a comma-separated list of the block starts)

These files may also take a track definition line that is visualization specific

---

**BedGraph Format**

Tab separated 4 required columns.

1. **chrom** (name of the chromosome, sequence id)
2. **chromStart** (starting position on the chromosome)
3. **chromEnd** (end position of the chromosome, note this base is not included)
4. **dataValue** (value of the data for that region)
GFF format

Search for GFF3 → http://www.sequenceontology.org/gff3.shtml

Tab separated with 9 columns. Missing attributes may be replaced with a dot → .

1. Seqid (usually chromosome)
2. Source (where is the data coming from)
3. Type (usually a term from the sequence ontology)
4. Start (interval start relative to the seqid)
5. End (interval end relative to the seqid)
6. Score (the score of the feature, a floating point number)
7. Strand (+ or −)
8. Phase (used to indicate reading frame for coding sequences)
9. Attributes (semicolon separated attributes → Name=ABC(ID=1))

Wiggle format

- two versions → fixed step and variable step each trying to optimize the amount of data storage

<table>
<thead>
<tr>
<th>fixedStep</th>
<th>chrom=chr1</th>
<th>start=100</th>
<th>step=1</th>
</tr>
</thead>
<tbody>
<tr>
<td>10</td>
<td>15</td>
<td>11</td>
<td>22</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>variableStep</th>
<th>chrom=chr1</th>
</tr>
</thead>
<tbody>
<tr>
<td>100 10</td>
<td>101 15</td>
</tr>
<tr>
<td>102 11</td>
<td>103 22</td>
</tr>
<tr>
<td>variableStep</td>
<td>chrom=chr2</td>
</tr>
<tr>
<td>2000 23</td>
<td>2005 40</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>

Wiggle is an nasty format – it looks simpler than it is – please avoid

We may have data in different coordinate systems!

Being “one off” is one of the most common errors in bioinformatics.

Conversion from GFF to BED

(start, end) → (start − 1, end)

Conversion from BED to GFF

(start, end) → (start + 1, end)

Not that there will be differences when selecting positions that depend on the END coordinate!

Handling coordinates relative to intervals

What are the coordinate of the base preceding and following the interval

GFF [start, end] → preceding base is at start - 1
BED [start, end] → preceding base is at start - 1

GFF [start, end] → next base is at end + 1
BED [start, end] → next base is at end
Representing interval relationships

- We have a gene with three splicing variants

<table>
<thead>
<tr>
<th>Start at 1000</th>
<th>Ends at 8000</th>
</tr>
</thead>
<tbody>
<tr>
<td>each exon is 1kb and is separated by 1kb</td>
<td></td>
</tr>
</tbody>
</table>

How to represent this in a data format?

Data representation

- Both BED and GFF files can represent them

- Two common versions of GFF → GTF2 and GFF3
  (note: tool documentation can often wrong and shows a weird combination of these two formats)

- In GFF the content of the ATTRIBUTE (9th) column specifies the relationship between features

GTF/GFF formats

GTF attributes:

- gene_id value;
  a globally unique identifier for the genomic source of the transcript
- transcript_id value;
  a globally unique identifier for the predicted transcript.

  gene_id “G1” transcript_id “T1”

GFF attributes:

  ID=exon1; Parent=T1

See the GFF3 site for exact specification of these means.

Important: More than one parent may be listed!

Example interval as GTF

A distinct line is entered for each exon, repeated for each transcript

11/13/14
The same exon may be part of different transcripts (parents)

---

**Example interval as GFF 3**

```plaintext
1 #gff-version 3
2 chr1 demo exon 1000 2000 . + . Parent=T1,T2,T3;
3 chr1 demo exon 3000 4000 . + . Parent=T1,T2;
4 chr1 demo exon 5000 6000 . + . Parent=T1,T3;
5 chr1 demo exon 7000 8000 . + . Parent=T1,T2,T3;
```

---

**Example interval in BED**

```plaintext
chr1 999 0000 21 0 . 999 0000 20,000 4 3000 18000 18000 18000 18000 18000
```

---

**Visualizing in IGV**

---

**Homework 24**

- Create and visualize in IGV an interval file that contains three splice variants of a 1 kb long gene with 5 exons.
- Show the file and a screenshot