***Slide 01***

The Whole Topic of this presentation covers the **FASTA** and **FASTQ** file formats, including a discussion of **PHRED** Scores and their representation.

<>

***Slide 02***

This will, mercifully, be split into two parts, the first covering just the **FASTA** and **FASTQ** file formats.

<>

***Slide 03***

First **FASTA** format.

<>

Each sequence in a **FASTA** format file commences with a line that starts with a “*Greater Than*” sign “**>**”.

<>

Followed immediately by a **Sequence Identifier** or **Name**.

<>

The **Sequence Identifier** is deemed to be ended by the first **White Space** (**Space**, **Tab** or **End Of Line**) encountered.

<>

Anything after the **Sequence Identifier** on the first line is interpreted as **Sequence Annotation**.

<>

Until the start of a **New Sequence** is encountered (that is another line beginning with a “*Greater Than*” sign “**>**”), all characters are interpreted as the **Sequence** itself.

<>

A **FASTA Sequence File** may contain arbitrarily many **FASTA Formatted Sequences**.

<>

***Slide 04***

There are some minor historic variants of **FASTA Format**.

Some, reflecting the limitations of line length imposed by various species of **Punched Cards** (yes, this format really is that old!) .

<>

Such anomalies are rarely, if ever, a problem in these enlightened times.

<>

***Slide 05***

**FASTA Format** is designed to store **Sequence Data** for **Analysis**. **Minimal Annotation** only is required.

The format can be used for **DNA**, **RNA** or **Protein** sequences but does not include any means of determining which type of sequence is being stored.

<>

Especially if sequences include **Ambiguity Codes**, guessing the type of sequence from its composition can be problematic.

<>

It is largely left to the analysis software to determine the type of sequence, usually by context.

<>

**FASTQ Format** is adapted from **FASTA Forma**t.

The format is extended to meet the requirements of **Sequencing Reads**, with **Call Quality Estimates** and just as much **Annotation** as is necessary.

<>

At least for **FASTQ Format Sequences**, things are a little simpler in that the sequence components cannot be **Amino Acids**.

<>

***Slide 06***

First some cosmetic “embellishments”?

Introduced to establish **FASTQ** as “different” possibly?

<>

The “*Greater Than*” sign (“**>**”) is replaced by an “**@**” character. Jolly Good?

<>

An additional **Annotation Line** is allowed after the **Sequence.**

This **Extra Annotation Line** is distinguished by starting with a **Plus Character** (“**+**”).

<>

The disappointing truth is, that nobody ever seems to want to use this extra opportunity to record forever the life stories of their sequences?

So, almost without exception, these worthy lines exist only as stunted enigmas, forever holding unto themselves the fascinating tales they were born to spread over an anxiously awaiting world!

<>

Currently, **Read Sequences** are, generally, relatively short.

This is one reason why they are stored all on just one line.

<>

***Slide 07***

For **Sequencing Reads** to be analysed sensibly, it is essential that the **Base Call Quality** of every element of each **Read** is determined and recorded.

That is, as well as knowing the **Base** in every position of a **Read**, it is necessary to record some measure of the certainty with which that **Base** is known.

<>

The **Base Call Quality** measures are recorded in a fourth line for each **Read**.

Each **Bases Call Quality** is recorded as a single character.

There is one **Base Call Quality Character** for every **Called Base**.

<>

***Slide 08***

There are a considerable number of possible B**ase Call Quality Values**.

Each must be assigned a unique printable character.

Thus, most possible printable characters are required to represent the full range of possible **Base Call Quality Values** … including the “**@**” character and the **Plus Sign** (“**+**”).

<>

If **Read Sequences**, and so **Base Call Quality Values**, were allowed to occupy multiple lines, ambiguous interpretations could arise.

Specifically, **Base Call Quality Lines** beginning with either an “**@**” character or a **Plus Sign** (“**+**”) would cause problems.

<>

A **Base Call Quality Line** starting with an “**@**” character might be misinterpreted as the start of a new **Read Sequence**.

The **Base Call Quality line** following might be misinterpreted as a rather strange **Read Sequence**!

<>

The problem would be exacerbated if a subsequent **Base Call Quality Line** happened to start with a **Plus Sign** (“**+**”) and so look like a **Supplementary Annotation Line**.

<>

***Slide 09***

The solution is to insist that **Read Sequences** never occupy more than one line, however long they are.

**Base Call Qualities** must also be confined to a single line.

Thus, a **Sequencing Read Record** in a **FASTQ Format File** always occupies ***Exactly* 4 Lines**.

<>

***Slide 10***

Enough for one bite! But remember there is **Part 2** still to come in which the wonders of **PHRED Scores** and there **Representation** will be revealed.

But now, it is time for tea and quiet contemplation.