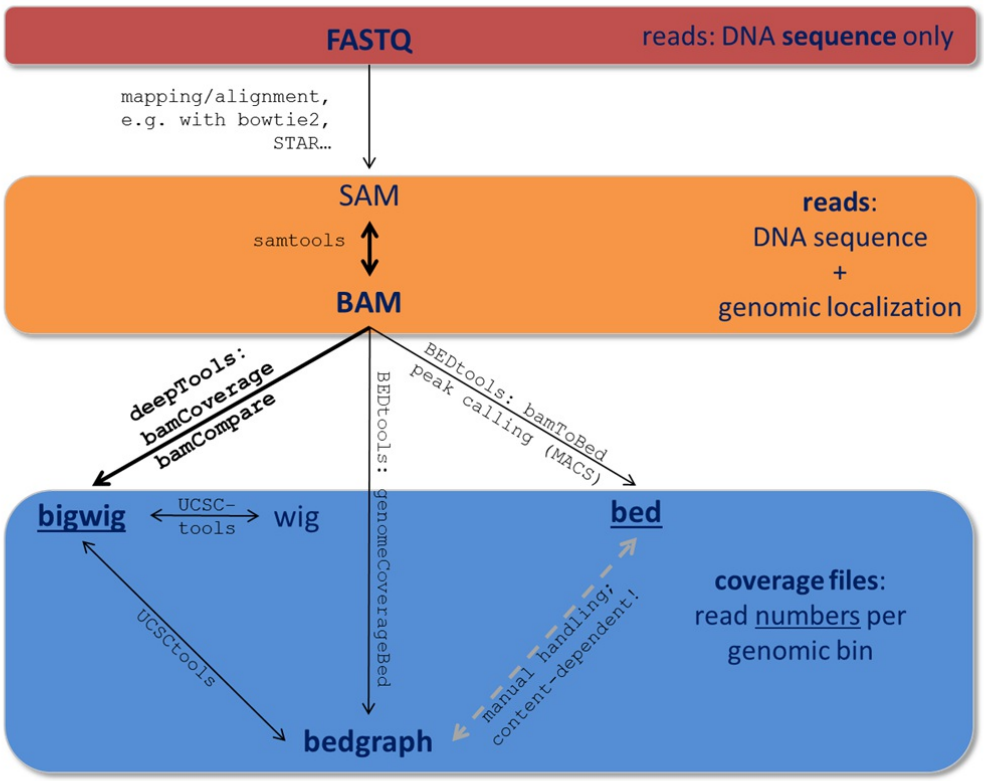
# File Formats

Data obtained from next-generation sequencing data must be processed several times. Most of the processing steps are aimed at extracting only those information that are truly needed for a specific down-stream analysis and to discard all the redundant entries. Therefore, **specific data formats are often associated with different steps of a data processing pipeline**. These associations, however, are by no means binding, but you should understand what kind of data is represented in which data format - this will help you to select the correct tools further down the road.

Here, we just want to give very brief key descriptions of the file, for elaborate information we will link to external websites. Be aware, that the sorting here is purely alphabetically, not according to their usage within an analysis pipeline that is depicted here.
For more information on the different tool collections mentioned in the figure, please check the following links:

- deepTools wiki: http://github.com/fidelram/deepTools/wiki
- samtools: http://samtools.sourceforge.net/http://samtools.sourceforge.net/
- UCSCtools download: http://hgdownload.cse.ucsc.edu/admin/exe/
- BEDtools: http://bedtools.readthedocs.org/en/latest/




---

BAM

- typical file extension: .bam
- *binary* file format (complement to SAM)
- contains information about sequenced reads *after alignment* to a reference genome
- each line = 1 mapped read, with information about:
  - its mapping quality (how certain is the read alignment to this particular genome locus?)
  - its sequencing quality (how well was each base pair detected during sequencing?)
  - its DNA sequence

    - its location in the genome
    - etc.
- highly recommended format for storing data
- to make a BAM file human-readable, one can, for example, use the program samtools view (from UCSC tools)
- for more information, see below for the definition of SAM files

#### bed

- typical file extension: .bed
- text file
- used for genomic intervals, e.g. genes, peak regions etc.
- actually, there is a rather strict definition of the format that can be found at UCSC
- for deepTools, the first 3 columns are important: chromosome, start position of the region, end position of the genome
- do not confuse it with the bedGraph format (eventhough they are quite similar)

#### bedGraph

- typical file extension: .bg, .bedgraph
- text file
- similar to BED file (not the same!), it can ONLY contain 4 columns and 4th column MUST be a score
- again, read the UCSC description for more details

#### bigWig

- typical file extension: .bw, .bigwig
- *binary* version of a bedGraph file
- usually contains 4 columns: chromosome, start of genomic bin, end of genomic bin, score
- the score can be anything, e.g. an average read coverage
- UCSC description for more details

#### FASTQ

- typical file extension: .fastq, fq
- text file, often gzipped (--> .fastq.gz)
- contains raw read information (e.g. base calls, sequencing quality measures etc.), but not information about where in the genome the read originated from

#### SAM

- typical file extension: .sam
- should be the result of an alignment of sequenced reads to a reference genome
- each line = 1 mapped read, with information about its mapping quality, its sequence, its location in the genome etc.
- it is recommended to generate the binary (compressed) version of this file format: BAM
- for more information, see the SAM specification

---

**Fidel Ramírez, Friederike Dündar, Sarah Diehl, Björn A. Grüning, Thomas Manke**

*Bioinformatics Group, Max-Planck-Institute of Immunobiology and Epigenetics, Freiburg & Department of Computer Science, University of Freiburg*

Web server (incl. sample data): http://deepTools.ie-freiburg.mpg.de
Code: https://github.com/fidelram/deepTools
Wiki & Tutorials: https://github.com/fidelram/deepTools/wiki