(a few) Common genomic data formats

...

Fasta and FastQ

>MCHU - Calmodulin - Human, rabbit, bovine, rat, and chicken
ADQLTEEQIAEFKEAFSLFDKDGDGTITTKELGTVMRSLGQNPTEAELQDMINEVDADGNGTID
FPEFLTMMARKMKDTDSEEEIREAFRVFDKDGNGYISAAELRHVMTNLGEKLTDEEVDEMIREA
DIDGDGQVNYEEFVQMMTAK*

@SEQ_ID
GATTTGGGGTTCAAAGCAGTATCGATCAAATAGTAAATCCATTTGTTCAACTCACAGTTT
+
!"*((((***+))%%%++)(%%%%).1***-+*"))**55CCF>>>>>>CCCCCCC65

https://en.wikipedia.org/wiki/FASTA_format

https://en.wikipedia.org/wiki/FASTA_format

SAM/BAM/CRAM

@HD VN:1.5 SO:coordinate

@SQ SN:ref LN:45

r001 99 ref 7 30 8M2I4M1D3M = 37 39 TTAGATAAAGGATACTG *

r002 0 ref 9 30 3S6M1P1I4M * 0 0 AAAAGATAAGGATA *

r003 0 ref 9 30 5S6M * 0 0 GCCTAAGCTAA * SA:Z:ref,29,-,6H5M,17,0;

r004 0 ref 16 30 6M14N5M * 0 0 ATAGCTTCAGC *

r003 2064 ref 29 17 6H5M * 0 0 TAGGC * SA:Z:ref,9,+,5S6M,30,1;

r001 147 ref 37 30 9M = 7 -39 CAGCGGCAT * NM:i:1

https://github.com/samtools/hts-specs/blob/master/SAMv1.pdf

VCF

```
##fileformat=VCFv4.0
##fileDate=20110705
##reference=1000GenomesPilot-NCBI37
##phasing=partial
##INFO=<ID=NS,Number=1,Type=Integer,Description="Number of Samples With Data">
##INFO=<ID=DP,Number=1,Type=Integer,Description="Total Depth">
##INFO=<ID=AF,Number=.,Type=Float,Description="Allele Frequency">
##INFO=<ID=AA,Number=1,Type=String,Description="Ancestral Allele">
##INFO=<ID=DB,Number=0,Type=Flag,Description="dbSNP membership, build 129">
##INFO=<ID=H2,Number=0,Type=Flag,Description="HapMap2 membership">
##FILTER=<ID=q10,Description="Quality below 10">
##FILTER=<ID=s50,Description="Less than 50% of samples have data">
##FORMAT=<ID=GQ,Number=1,Type=Integer,Description="Genotype Quality">
##FORMAT=<ID=GT,Number=1,Type=String,Description="Genotype">
##FORMAT=<ID=DP,Number=1,Type=Integer,Description="Read Depth">
##FORMAT=<ID=HQ,Number=2,Type=Integer,Description="Haplotype Quality">
#CHROM POS    ID        REF  ALT     QUAL FILTER INFO                              FORMAT      Sample1         Sample2         Sample3
2     4370   rs6057    G    A       29   .      NS=2;DP=13;AF=0.5;DB;H2           GT:GQ:DP:HQ 0|0:48:1:52,51  1|0:48:8:51,51  1/1:43:5:.,.
2     7330   .         T    A       3    q10    NS=5;DP=12;AF=0.017               GT:GQ:DP:HQ 0|0:46:3:58,50  0|1:3:5:65,3    0/0:41:3
2     110696 rs6055    A    G,T     67   PASS   NS=2;DP=10;AF=0.333,0.667;AA=T;DB GT:GQ:DP:HQ 1|2:21:6:23,27 2|1:2:0:18,2   2/2:35:4
2     130237 .         T    .       47   .      NS=2;DP=16;AA=T                   GT:GQ:DP:HQ 0|0:54:7:56,60  0|0:48:4:56,51  0/0:61:2
2     134567 microsat1 GTCT G,GTACT 50   PASS   NS=2;DP=9;AA=G                    GT:GQ:DP    0/1:35:4        0/2:17:2        1/1:40:3
```

https://en.wikipedia.org/wiki/Variant_Call_Format

BED

Required fields

The first three fields in each feature line are required:

1. **chrom** - name of the chromosome or scaffold.
2. **chromStart** - Start position of the feature in standard chromosomal coordinates (i.e. first base is 0).
3. **chromEnd** - End position of the feature in standard chromosomal coordinates

```
chr1  213941196  213942363
chr1  213942363  213943530
chr1  213943530  213944697
chr2  158364697  158365864
```

Optional fields

name, score, strand, thickStart, thickEnd, itemRgb, blockCount, blockSizes, blockStarts

https://genome.ucsc.edu/FAQ/FAQformat#format1

GTF

Fields

Fields **must** be tab-separated. Also, all but the final field in each feature line must contain a value; "empty" columns should be denoted with a '.'

1. **seqname** - name of the chromosome or scaffold; chromosome names can be given with or without the 'chr' prefix. **Important note**: the seqname must be one used within Ensembl, i.e. a standard chromosome name or an Ensembl identifier such as a scaffold ID, without any additional content such as species or assembly. See the example GFF output below.
2. **source** - name of the program that generated this feature, or the data source (database or project name)
3. **feature** - feature type name, e.g. Gene, Variation, Similarity
4. **start** - Start position of the feature, with sequence numbering starting at 1.
5. **end** - End position of the feature, with sequence numbering starting at 1.
6. **score** - A floating point value.
7. **strand** - defined as + (forward) or - (reverse).
8. **frame** - One of '0', '1' or '2'. '0' indicates that the first base of the feature is the first base of a codon, '1' that the second base is the first base of a codon, and so on..
9. **attribute** - A semicolon-separated list of tag-value pairs, providing additional information about each feature.

http://asia.ensembl.org/info/website/upload/gff.html

Plink - PED/MAP

Test.ped:

#FID IID PID MID SEX Phen

F1 I1 0 0 1 0 G G 2 2 C C

F1 I2 0 0 2 0 A A 0 0 A C

F1 I3 F1 F2 1 2 0 0 1 2 A C

F2 I1 0 0 1 0 A A 2 2 0 0

Test.map:

chr snp gpos bp

1 snp1 0 1

1 snp2 0 2

1 snp3 0 3

http://zzz.bwh.harvard.edu//plink/data.shtml#ped

‘Oxford’ - GEN/SAMPLE

```
SNP1 rs1 1000 A C 1 0 0 1 0 0

SNP2 rs2 2000 G T 1 0 0 0 1 0

SNP3 rs3 3000 C T 1 0 0 0 1 0

SNP4 rs4 4000 C T 0 1 0 0 1 0

SNP5 rs5 5000 A G 0 1 0 0 0 1
```

```
ID_1 ID_2 missing cov_1 cov_2 cov_3 cov_4 pheno1 bin1
0 0 0 D D C C P B
1 1 0.007 1 2 0.0019 -0.008 1.233 1
2 2 0.009 1 2 0.0022 -0.001 6.234 0
3 3 0.005 1 2 0.0025 0.0028 6.121 1
4 4 0.007 2 1 0.0017 -0.011 3.234 1
5 5 0.004 3 2 -0.012 0.0236 2.786 0
```

http://www.stats.ox.ac.uk/~marchini/software/gwas/file_format.html

Plain Text - .csv .tsv .txt