DATA FORMATS 1 sequence data files

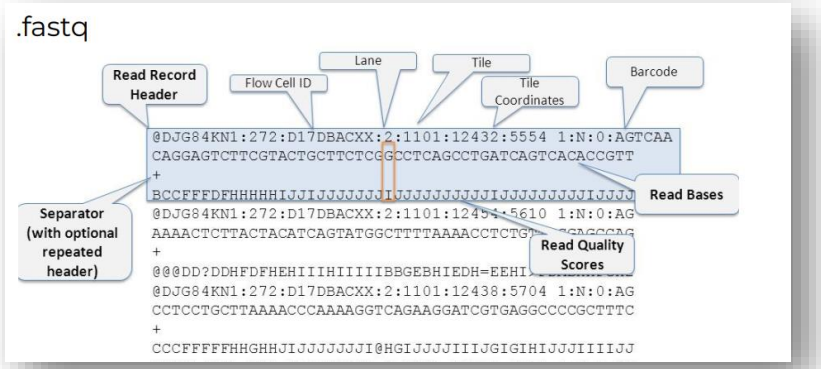
FASTQ

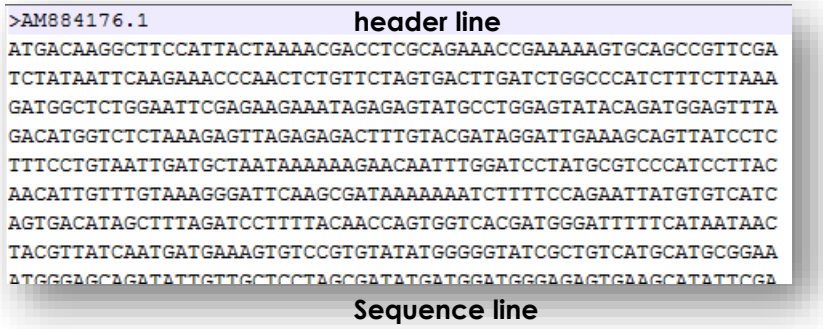
- Unaligned read sequences with base qualities

Sequencing Instrument

FASTA

- plain sequence unaligned





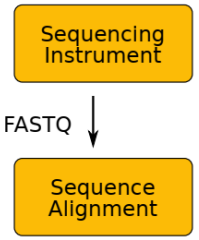
DATA FORMATS 2 mapping files

FASTQ

- Unaligned read sequences with base qualities

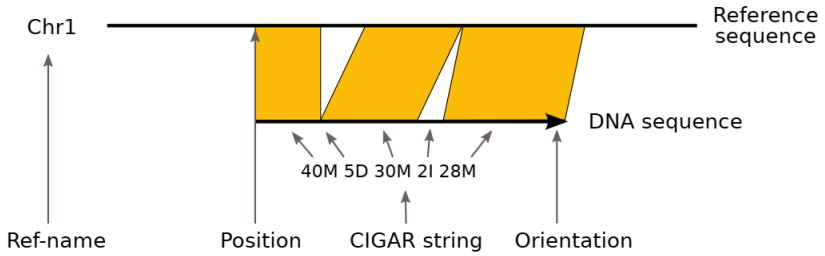
SAM/BAM

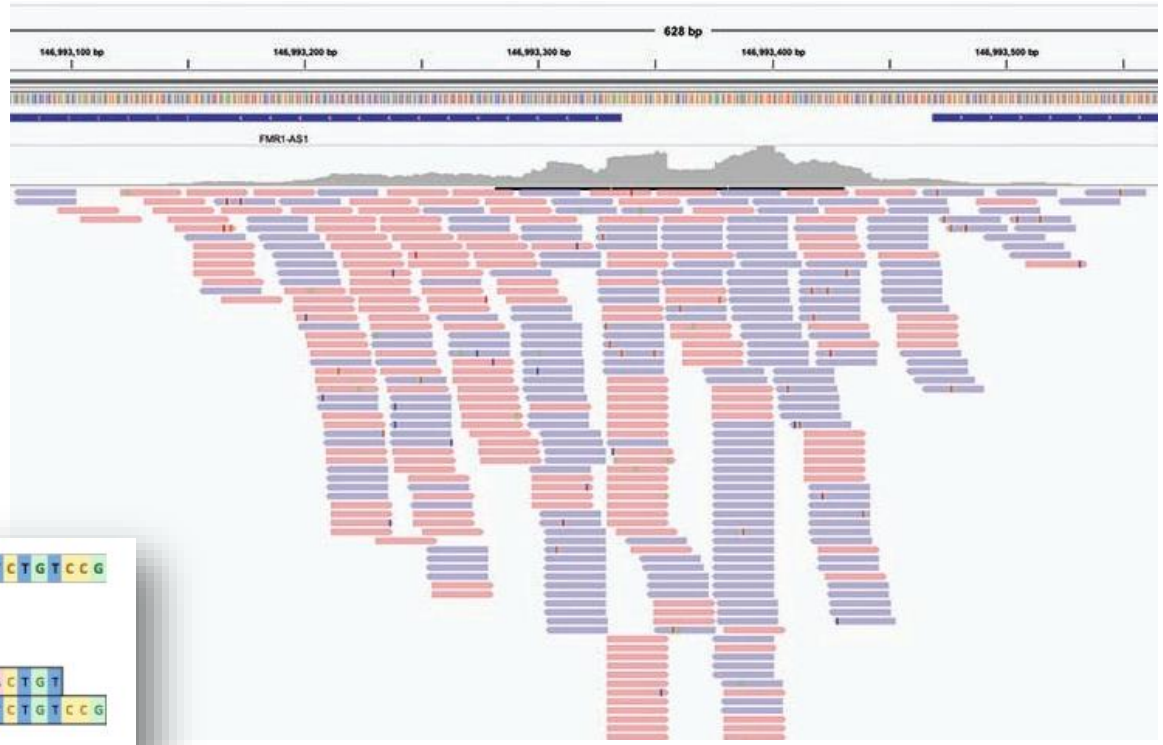
- Unaligned or aligned reads
- Text and binary formats



SAM (Sequence Alignment/Map) format

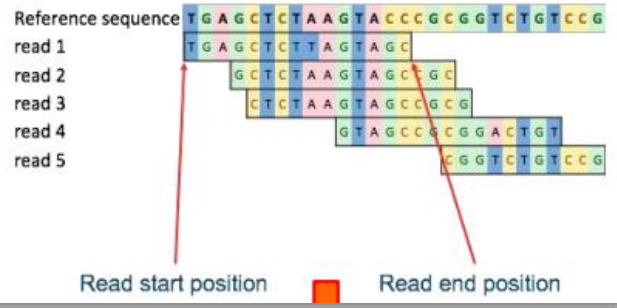
- Unified format for storing read alignments to a reference genome
- Developed by the 1000 Genomes Project group (2009)
- One record (a single DNA fragment alignment) per line describing alignment between fragment and reference
- 11 fixed columns + optional key:type:value tuples





BAM stores all of the data

- Every read base
- Every base quality
- Using a single conventional compression technique for all types of data



DATA FORMATS 3 variant files

FASTQ

- Unaligned read sequences with base qualities

SAM/BAM

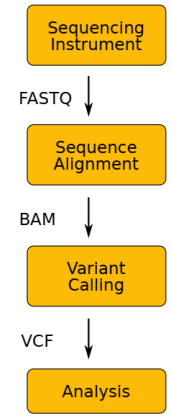
- Unaligned or aligned reads
- Text and binary formats

CRAM

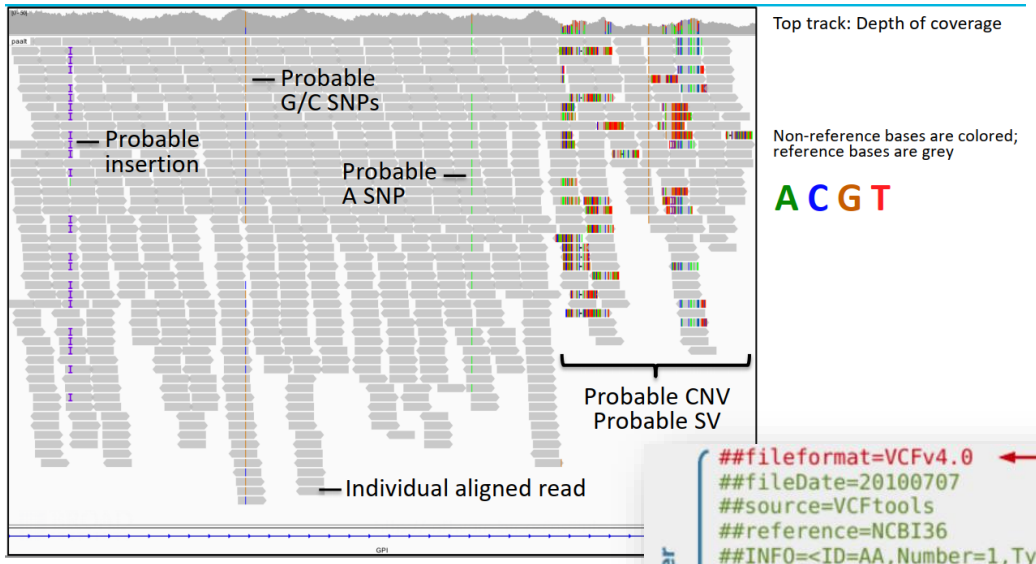
- Better compression than BAM

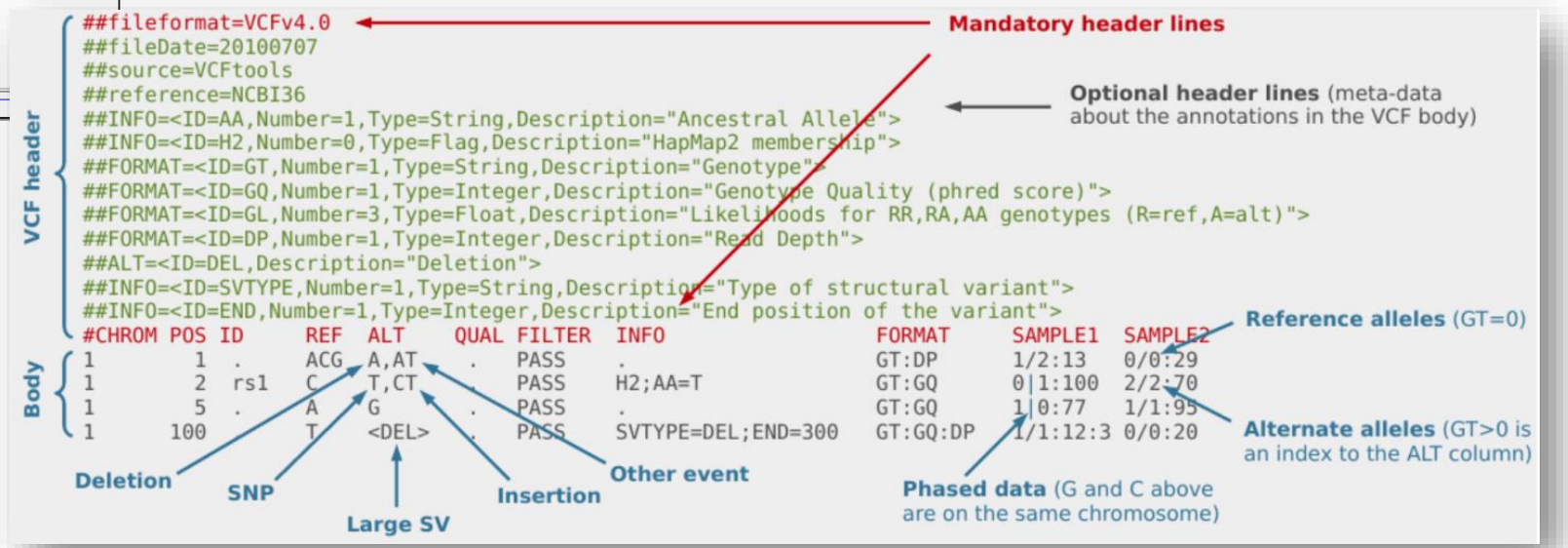
VCF/BCF

- Flexible variant call format
- Arbitrary types of sequence variation
- SNPs, indels, structural variations



Specifications maintained by the Global Alliance for Genomics and Health





DATA FORMATS 4 annotation files

EMBL

Two-character line code indicates the type of information contained in the line

Header

Key

Qualifier

Annotation

Sequence

```
ID   ECRSMA     standard; DNA; PRO; 500 BP.
XX
AC   L40173;
XX
SV   L40173.1
XX
DT   10-AUG-1995 (Rel. 44, Created)
DT   04-MAR-2000 (Rel. 63, Last updated, Version 4)
XX
DE   Erwinia carotovora repressor (rsmA) gene, complete cds.
XX
KW   repressor; rsmA gene.
XX
OS   Pectobacterium carotovorum
OC   Bacteria; Proteobacteria; Gammaproteobacteria; Enterobacteriaceae;
OC   Pectobacterium.
XX
RN   [1]
RP   1-500
RA   Cui Y., Chatterjee A., Liu Y., Dumenyo C.K., Chatterjee A.K.;
RT   "Identification of a global repressor gene, rsmA, of Erwinia carotovora
RT   subsp. carotovora that controls extracellular enzymes,
RT   N-(3-oxohexanoyl)-L-homoserine lactone, and pathogenicity in soft-rotting
RT   Erwinia spp";
RL   J. Bacteriol. 177(17):0-0(1995).
XX
DR   GOA; Q47620; Q47620.
DR   SWISS-PROT; Q47620; CSRA_ERWCA.
XX
FH   Key             Location/Qualifiers
FH
FT   source          1..500
FT                   /db_xref="taxon:554"
FT                   /organism="Pectobacterium carotovorum"
FT                   /strain="71"
FT                   /sub_species="carotovora"
FT   -10_signal      107..112
FT                   /gene="rsmA"
FT   RBS             235..239
FT                   /gene="rsmA"
FT   CDS             246..431
FT                   /codon_start=1
FT                   /db_xref="GOA:Q47620"
FT                   /db_xref="SWISS-PROT:Q47620"
FT                   /note="putative"
FT                   /transl_table=11
FT                   /gene="rsmA"
FT                   /function="global repressor"
FT                   /protein_id="AAA74502.1"
FT                   /translation="MLILTRRVGETLIIGDEVTVTVLGVKGNQVRIGVNAPKEVSVHRE
FT                   EIYQRIQAEKSQPTSY"
XX
SQ   Sequence 500 BP; 140 A; 101 C; 120 G; 139 T; 0 other;
     ggatccggca agcaggatag aaagtgtgtt accttcagat attctgaagc tttacatgct        60
     cagttctgtt gttgtgataa caaaagcaca agctactgat atcgactaaa ctaacaagta       120
     gtgacaaacc ggagtgtgat ggtgtggtta taccatcgtc taggtttacg ttttcacagc       180
     acatgatgga taatggcggg gagacagaga gaccgcactc tttataatct ttcaaggagc       240
     aaagaatgct tattttgact cgtcgagttg gcgaaaccct catcatcggc gatgaggtaa       300
     cggttaccgt attaggagtg aaaggcaacc aggtgcgtat tggtgttaat gcacctaaag       360
     aggtttctgt ccaccgtgaa gagatctatc agcgtattca ggccgaaaaa tctcaaccaa       420
     cgtcatattg attgacaatg cgtctcgtgt tcgcgggacg caattgttat ttccggtttt       480
     tccccacac atttatcgat                                                   500
//
```

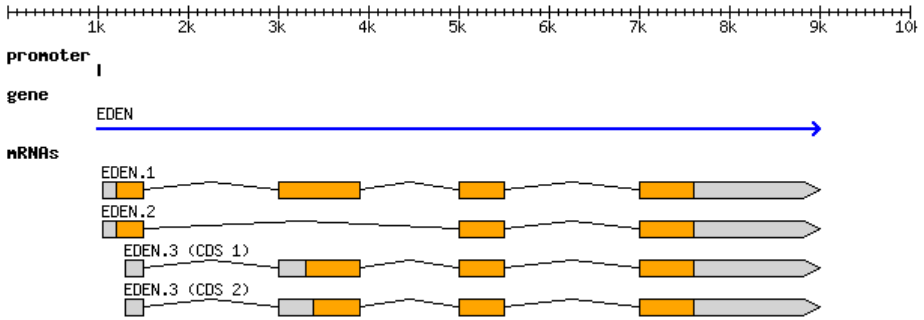
Genbank

```
LOCUS       ERWRSMA                  500 bp    DNA     linear   BCT 19-AUG-1995
DEFINITION  Erwinia carotovora repressor (rsmA) gene, complete cds.
ACCESSION   L40173
VERSION     L40173.1  GI:927031
KEYWORDS    repressor; rsmA gene.
SOURCE      Pectobacterium carotovorum
  ORGANISM  Pectobacterium carotovorum
            Bacteria; Proteobacteria; Gammaproteobacteria; Enterobacteriaceae;
            Pectobacterium.
REFERENCE   1  (bases 1 to 500)
  AUTHORS   Cui,Y., Chatterjee,A., Liu,Y., Dumenyo,C.K. and Chatterjee,A.K.
  TITLE     Identification of a global repressor gene, rsmA, of Erwinia
            carotovora subsp. carotovora that controls extracellular enzymes,
            N-(3-oxohexanoyl)-L-homoserine lactone, and pathogenicity in
            soft-rotting Erwinia spp
  JOURNAL   J. Bacteriol. 177(17) (1995) In press
COMMENT     Original source text: Erwinia carotovora (strain 71, sub_species
            carotovora) DNA.
FEATURES             Location/Qualifiers
     source          1..500
                     /organism="Pectobacterium carotovorum"
                     /strain="71"
                     /sub_species="carotovora"
                     /db_xref="taxon:554"
     gene            107..431
                     /gene="rsmA"
     -10_signal      107..112
                     /gene="rsmA"
     RBS             235..239
                     /gene="rsmA"
     CDS             246..431
                     /gene="rsmA"
                     /function="global repressor"
                     /note="putative"
                     /codon_start=1
                     /transl_table=11
                     /protein_id="AAA74502.1"
                     /db_xref="GI:927032"
                     /translation="MLILTRRVGETLIIGDEVTVTVLGVKGNQVRIGVNAPKEVSVHR
                     EEIYQRIQAEKSQPTSY"
BASE COUNT     140 a    101 c    120 g    139 t
ORIGIN
        1 ggatccggca agcaggatag aaagtgtgtt accttcagat attctgaagc tttacatgct
       61 cagttctgtt gttgtgataa caaaagcaca agctactgat atcgactaaa ctaacaagta
      121 gtgacaaacc ggagtgtgat ggtgtggtta taccatcgtc taggtttacg ttttcacagc
      181 acatgatgga taatggcggg gagacagaga gaccgcactc tttataatct ttcaaggagc
      241 aaagaatgct tattttgact cgtcgagttg gcgaaaccct catcatcggc gatgaggtaa
      301 cggttaccgt attaggagtg aaaggcaacc aggtgcgtat tggtgttaat gcacctaaag
      361 aggtttctgt ccaccgtgaa gagatctatc agcgtattca ggccgaaaaa tctcaaccaa
      421 cgtcatattg attgacaatg cgtctcgtgt tcgcgggacg caattgttat ttccggtttt
      481 tccccacac atttatcgat
//
```

GFF, GTF

- tabular feature mapping and description data

GFF




```
0  ##gff-version 3.2.1
1  ##sequence-region ctg123 1 1497228
2  ctg123 . gene            1000  9000  .  +  .  ID=gene00001;Name=EDEN
3  ctg123 . TF_binding_site 1000  1012  .  +  .  ID=tfbs00001;Parent=gene00001
4  ctg123 . mRNA            1050  9000  .  +  .  ID=mRNA00001;Parent=gene00001;Name=EDEN.1
5  ctg123 . mRNA            1050  9000  .  +  .  ID=mRNA00002;Parent=gene00001;Name=EDEN.2
6  ctg123 . mRNA            1300  9000  .  +  .  ID=mRNA00003;Parent=gene00001;Name=EDEN.3
7  ctg123 . exon            1300  1500  .  +  .  ID=exon00001;Parent=mRNA00003
8  ctg123 . exon            1050  1500  .  +  .  ID=exon00002;Parent=mRNA00001,mRNA00002
9  ctg123 . exon            3000  3902  .  +  .  ID=exon00003;Parent=mRNA00001,mRNA00003
10 ctg123 . exon            5000  5500  .  +  .  ID=exon00004;Parent=mRNA00001,mRNA00002,mRNA00003
11 ctg123 . exon            7000  9000  .  +  .  ID=exon00005;Parent=mRNA00001,mRNA00002,mRNA00003
12 ctg123 . CDS             1201  1500  .  +  0  ID=cds00001;Parent=mRNA00001;Name=edenprotein.1
13 ctg123 . CDS             3000  3902  .  +  0  ID=cds00001;Parent=mRNA00001;Name=edenprotein.1
14 ctg123 . CDS             5000  5500  .  +  0  ID=cds00001;Parent=mRNA00001;Name=edenprotein.1
15 ctg123 . CDS             7000  7600  .  +  0  ID=cds00001;Parent=mRNA00001;Name=edenprotein.1
16 ctg123 . CDS             1201  1500  .  +  0  ID=cds00002;Parent=mRNA00002;Name=edenprotein.2
17 ctg123 . CDS             5000  5500  .  +  0  ID=cds00002;Parent=mRNA00002;Name=edenprotein.2
18 ctg123 . CDS             7000  7600  .  +  0  ID=cds00002;Parent=mRNA00002;Name=edenprotein.2
19 ctg123 . CDS             3301  3902  .  +  0  ID=cds00003;Parent=mRNA00003;Name=edenprotein.3
20 ctg123 . CDS             5000  5500  .  +  1  ID=cds00003;Parent=mRNA00003;Name=edenprotein.3
```

seqId source feat.type start end score strand frame attributes (; separated)

DATA FORMATS 4
multiple alignment & phylogenetic tree files

Fasta multiple alignment
- aligned fasta entries

Clustal (.aln)
- multiple sequence alignment

Newick tree format
- (nodes and distances)

```
CLUSTAL FORMAT for T-COFFEE r479 [http://www.tcoffee.org] [MODE:  ], CPU=0.00 sec, SCORE=90, Nseq=8, Len=83

1PHT            YQYRALYDYKKEREEDIDLHLGDILTVNKGSLVALGFSDGQEARPEEI----------GWLNGYNETTGERGDFPGTYVEYIG
1BB9            FKVQAQHDYTATDTDELQLKAGDVVLVIP-------FQNP----EEQDEGWLMGVKESDWNQHK-ELEKCRGVFPENFTERVQ
1UHC            QVYFAVYTFKARNPNELSVSANQKLKILE-------FKDV----TGNT----------EWWLAE--VNGKKGYVPSNYIRKTE
1YCS            GVIYALWDYEPQNDDELPMKEGDCMTIIH-------REDE----D-EI----------EWWWA--RLNDKEGYVPRNLLGLYP
1OOT            PKAVALYSFAGEESGDLPFRKGDVITILKK-------S------DSQN----------DWWTG--RVNGREGIFPANYVE-LV
1ABO            NLFVALYDFVASGDNTLSITKGEKLRVLG-------YNH-------NG----------EWCEAQ--TKNGQGWVPSNYITPVN
1FYN            TLFVALYDYEARTEDDLSFHKGEKFQILN-------SS--------EG----------DWWEARSLTTGETGYIPSNYVAPVD
1QCF            IIVVALYDYEAIHHEDLSFQKGDQMVVLE-------E---------SG----------EWWKARSLATRKEGYIPSNYVARVD
                    *   :       : .  .: . :                   .           *          * .* .
```

```
>1PHT
YQYRALYDYKKEREEDIDLHLGDILTVNKGSLVALGFSDGQEARPEEI----------GWLNGYNETTGERGDFPGTYVEYIG
>1BB9
FKVQAQHDYTATDTDELQLKAGDVVLVIP-------FQNP----EEQDEGWLMGVKESDWNQHK-ELEKCRGVFPENFTERVQ
>1UHC
QVYFAVYTFKARNPNELSVSANQKLKILE-------FKDV----TGNT----------EWWLAE--VNGKKGYVPSNYIRKTE
>1YCS
GVIYALWDYEPQNDDELPMKEGDCMTIIH-------REDE----D-EI----------EWWWA--RLNDKEGYVPRNLLGLYP
>1OOT
PKAVALYSFAGEESGDLPFRKGDVITILKK-------S------DSQN----------DWWTG--RVNGREGIFPANYVE-LV
>1ABO
NLFVALYDFVASGDNTLSITKGEKLRVLG-------YNH-------NG----------EWCEAQ--TKNGQGWVPSNYITPVN
>1FYN
TLFVALYDYEARTEDDLSFHKGEKFQILN-------SS--------EG----------DWWEARSLTTGETGYIPSNYVAPVD
>1QCF
IIVVALYDYEAIHHEDLSFQKGDQMVVLE-------E---------SG----------EWWKARSLATRKEGYIPSNYVARVD
```

vi) (A:0.01,B:0.02,(C:0.01,D:0.03)Int1:0.01)[1.5];[Tree name]