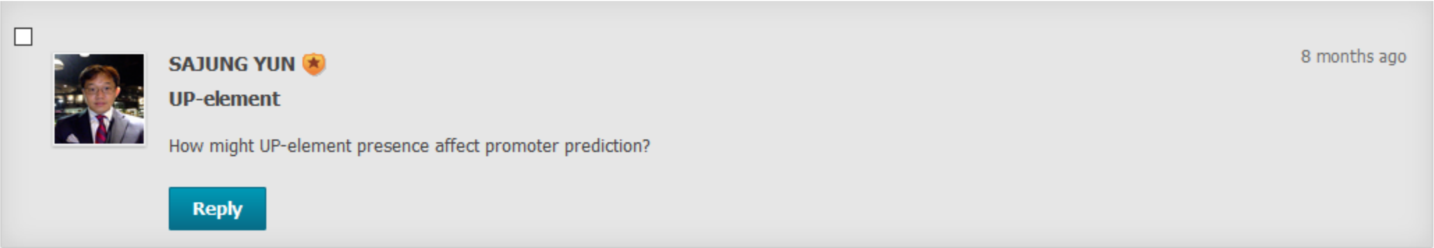
Unit 2



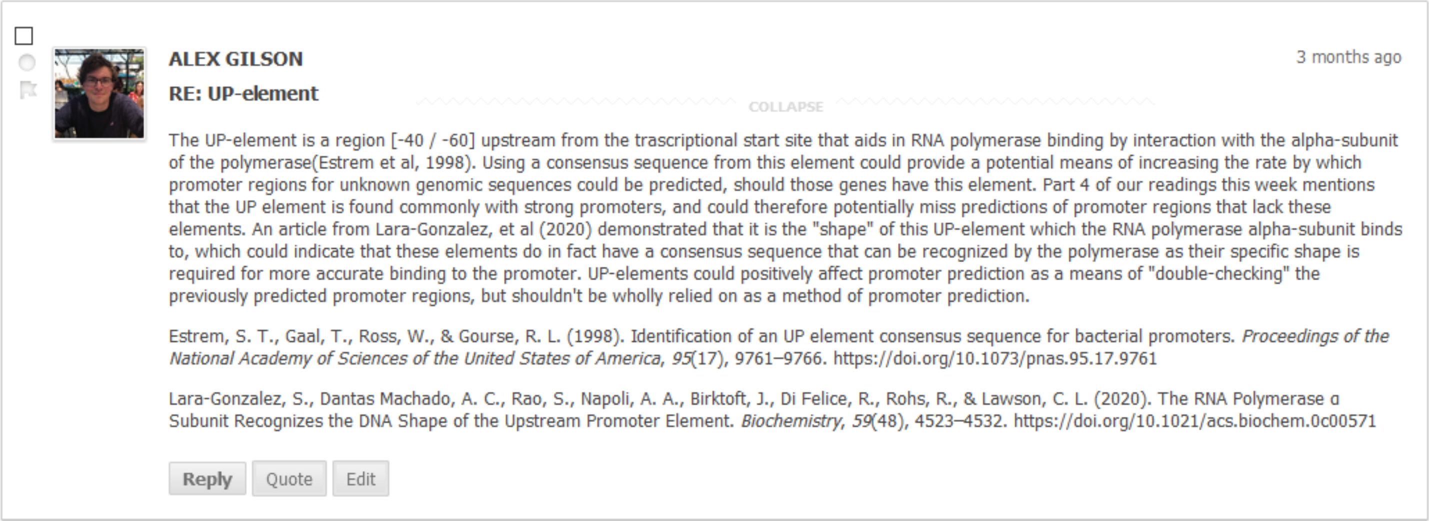
□

SAJUNG YUN

8 months ago

UP-element

How might UP-element presence affect promoter prediction?



□

ALEX GILSON

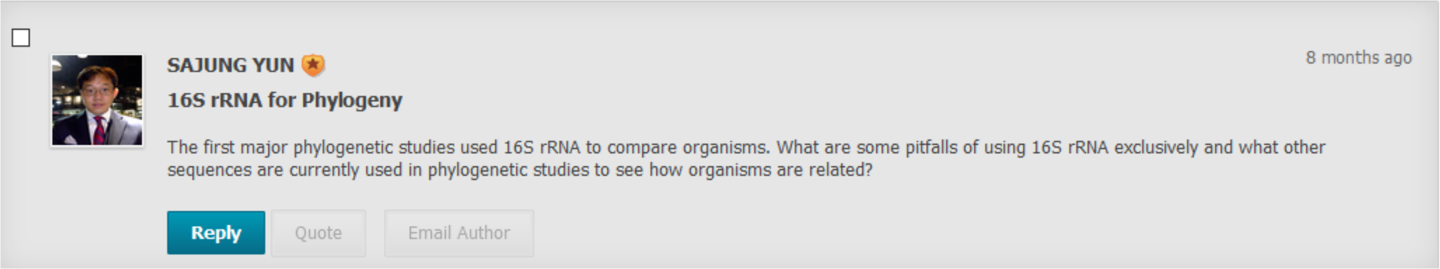
3 months ago

RE: UP-element

The UP-element is a region [-40 / -60] upstream from the trascriptional start site that aids in RNA polymerase binding by interaction with the alpha-subunit of the polymerase(Estrem et al, 1998). Using a consensus sequence from this element could provide a potential means of increasing the rate by which promoter regions for unknown genomic sequences could be predicted, should those genes have this element. Part 4 of our readings this week mentions that the UP element is found commonly with strong promoters, and could therefore potentially miss predictions of promoter regions that lack these elements. An article from Lara-Gonzalez, et al (2020) demonstrated that it is the "shape" of this UP-element which the RNA polymerase alpha-subunit binds to, which could indicate that these elements do in fact have a consensus sequence that can be recognized by the polymerase as their specific shape is required for more accurate binding to the promoter. UP-elements could positively affect promoter prediction as a means of "double-checking" the previously predicted promoter regions, but shouldn't be wholly relied on as a method of promoter prediction.

Estrem, S. T., Gaal, T., Ross, W., & Gourse, R. L. (1998). Identification of an UP element consensus sequence for bacterial promoters. *Proceedings of the National Academy of Sciences of the United States of America*, *95*(17), 9761–9766. https://doi.org/10.1073/pnas.95.17.9761

Lara-Gonzalez, S., Dantas Machado, A. C., Rao, S., Napoli, A. A., Birktoft, J., Di Felice, R., Rohs, R., & Lawson, C. L. (2020). The RNA Polymerase α Subunit Recognizes the DNA Shape of the Upstream Promoter Element. *Biochemistry*, *59*(48), 4523–4532. https://doi.org/10.1021/acs.biochem.0c00571



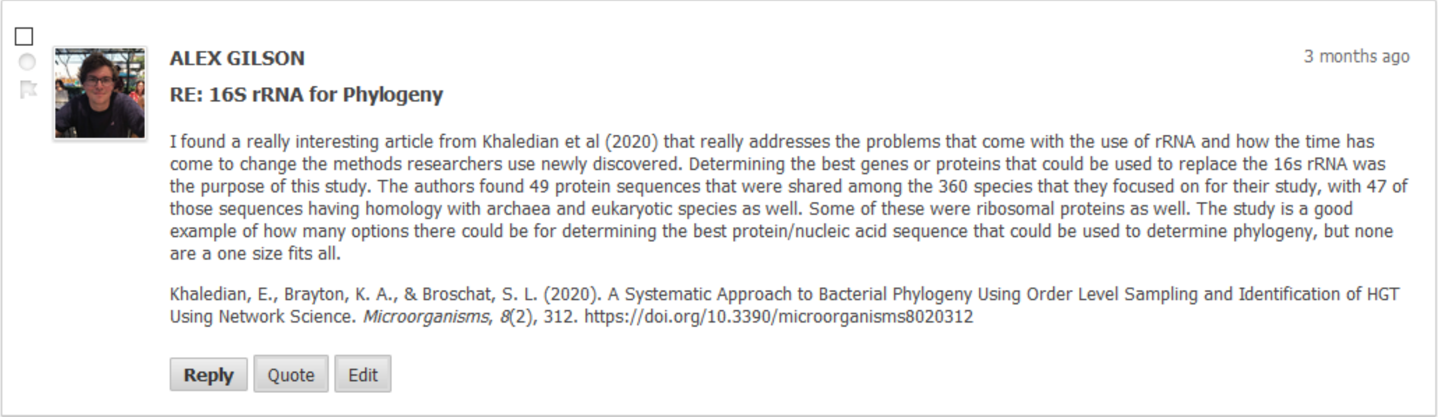
□

SAJUNG YUN

8 months ago

16S rRNA for Phylogeny

The first major phylogenetic studies used 16S rRNA to compare organisms. What are some pitfalls of using 16S rRNA exclusively and what other sequences are currently used in phylogenetic studies to see how organisms are related?



□

ALEX GILSON

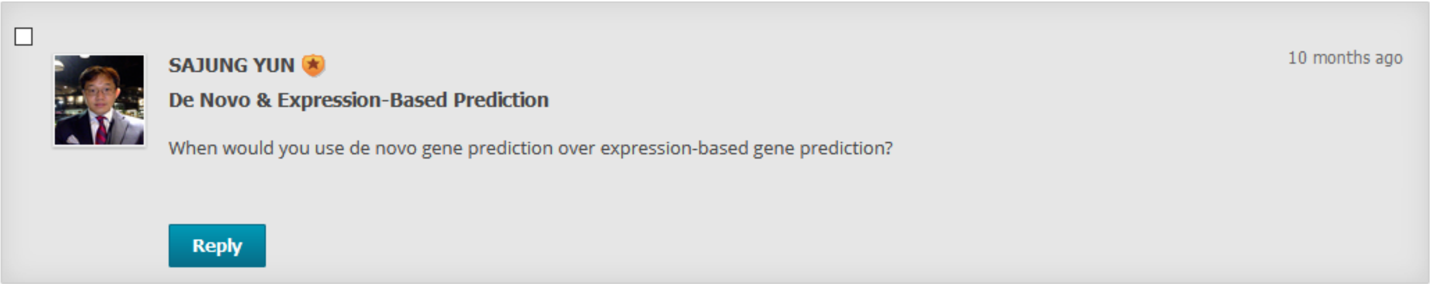
3 months ago

RE: 16S rRNA for Phylogeny

I found a really interesting article from Khaledian et al (2020) that really addresses the problems that come with the use of rRNA and how the time has come to change the methods researchers use newly discovered. Determining the best genes or proteins that could be used to replace the 16s rRNA was the purpose of this study. The authors found 49 protein sequences that were shared among the 360 species that they focused on for their study, with 47 of those sequences having homology with archaea and eukaryotic species as well. Some of these were ribosomal proteins as well. The study is a good example of how many options there could be for determining the best protein/nucleic acid sequence that could be used to determine phylogeny, but none are a one size fits all.

Khaledian, E., Brayton, K. A., & Broschat, S. L. (2020). A Systematic Approach to Bacterial Phylogeny Using Order Level Sampling and Identification of HGT Using Network Science. *Microorganisms*, *8*(2), 312. https://doi.org/10.3390/microorganisms8020312

Unit 3

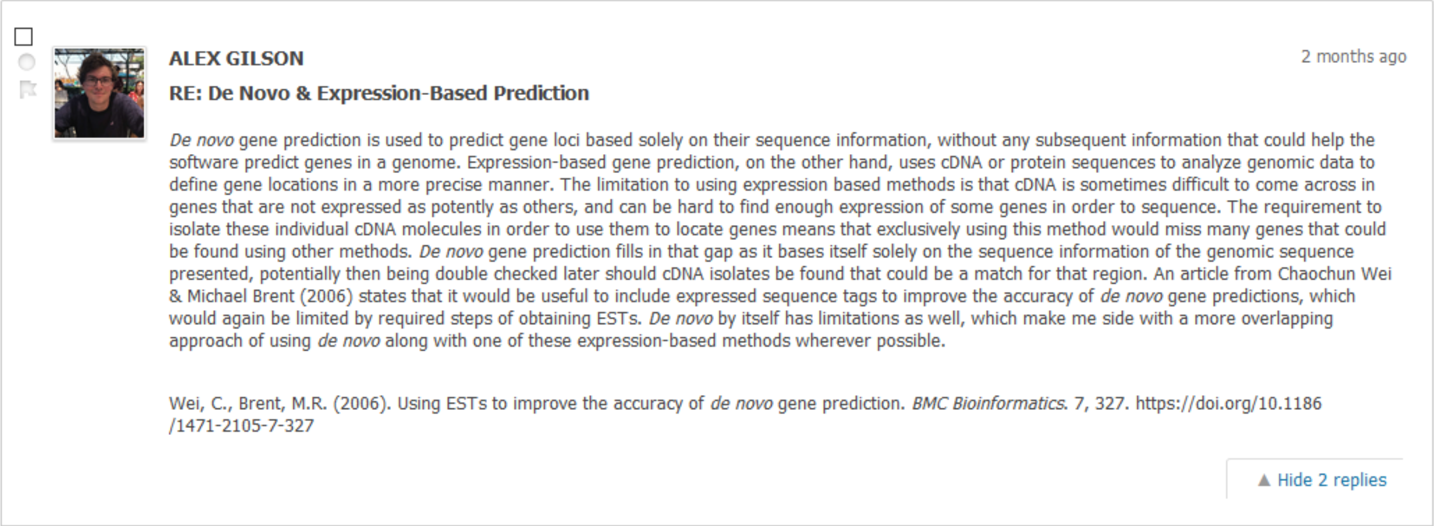


SAJUNG YUN

10 months ago

De Novo & Expression-Based Prediction

When would you use de novo gene prediction over expression-based gene prediction?



ALEX GILSON

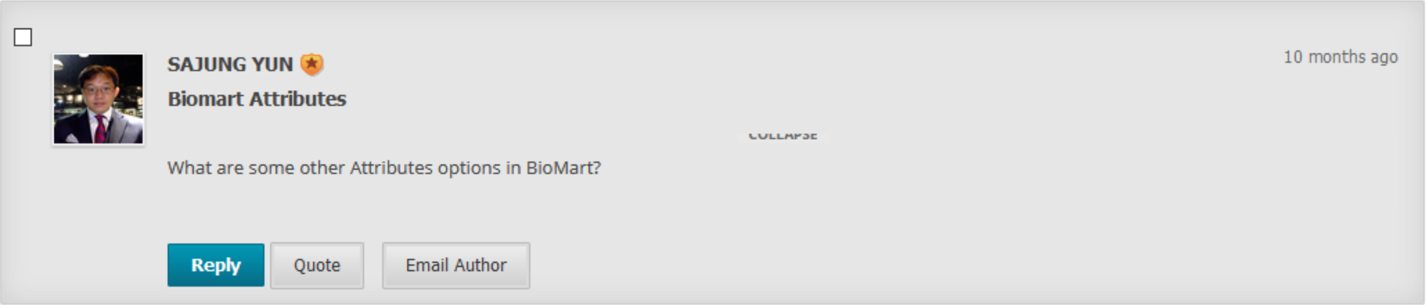
2 months ago

RE: De Novo & Expression-Based Prediction

De novo gene prediction is used to predict gene loci based solely on their sequence information, without any subsequent information that could help the software predict genes in a genome. Expression-based gene prediction, on the other hand, uses cDNA or protein sequences to analyze genomic data to define gene locations in a more precise manner. The limitation to using expression based methods is that cDNA is sometimes difficult to come across in genes that are not expressed as potently as others, and can be hard to find enough expression of some genes in order to sequence. The requirement to isolate these individual cDNA molecules in order to use them to locate genes means that exclusively using this method would miss many genes that could be found using other methods. *De novo* gene prediction fills in that gap as it bases itself solely on the sequence information of the genomic sequence presented, potentially then being double checked later should cDNA isolates be found that could be a match for that region. An article from Chaochun Wei & Michael Brent (2006) states that it would be useful to include expressed sequence tags to improve the accuracy of *de novo* gene predictions, which would again be limited by required steps of obtaining ESTs. *De novo* by itself has limitations as well, which make me side with a more overlapping approach of using *de novo* along with one of these expression-based methods wherever possible.

Wei, C., Brent, M.R. (2006). Using ESTs to improve the accuracy of *de novo* gene prediction. *BMC Bioinformatics*. 7, 327. https://doi.org/10.1186/1471-2105-7-327

Unit 4

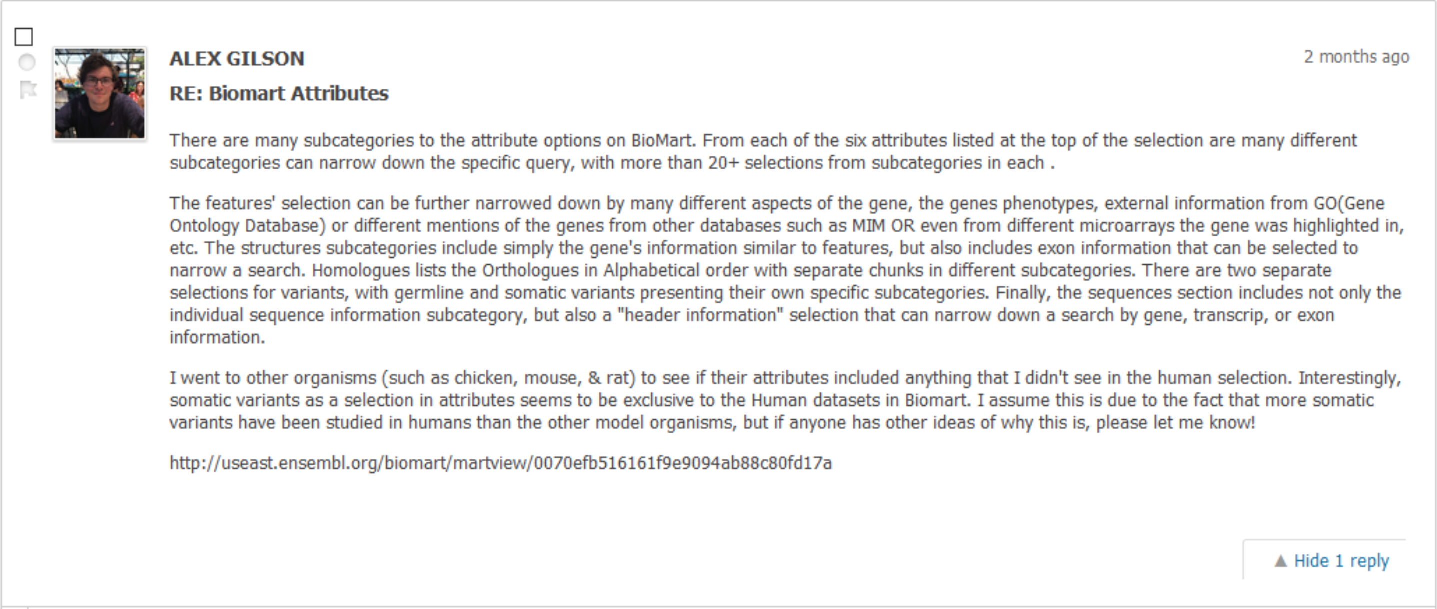


SAJUNG YUN

10 months ago

Biomart Attributes

What are some other Attributes options in BioMart?



ALEX GILSON

2 months ago

RE: Biomart Attributes

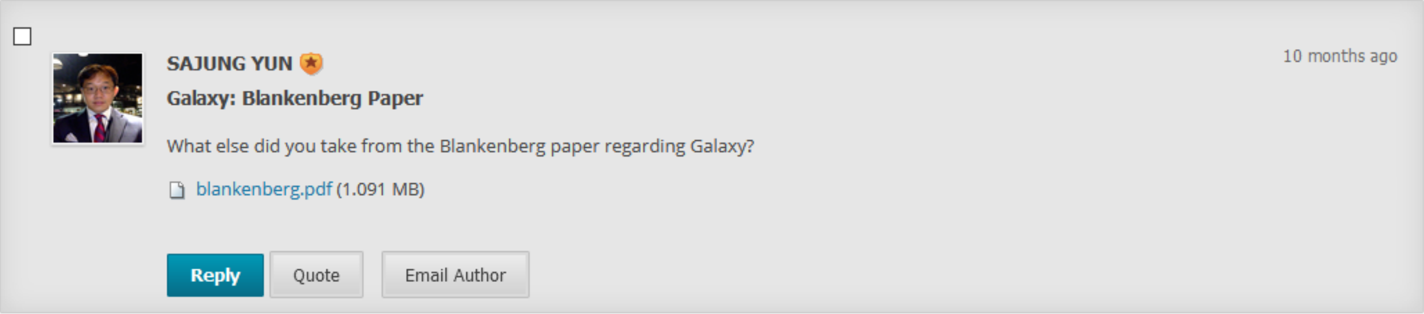
There are many subcategories to the attribute options on BioMart. From each of the six attributes listed at the top of the selection are many different subcategories can narrow down the specific query, with more than 20+ selections from subcategories in each .

The features' selection can be further narrowed down by many different aspects of the gene, the genes phenotypes, external information from GO(Gene Ontology Database) or different mentions of the genes from other databases such as MIM OR even from different microarrays the gene was highlighted in, etc. The structures subcategories include simply the gene's information similar to features, but also includes exon information that can be selected to narrow a search. Homologues lists the Orthologues in Alphabetical order with separate chunks in different subcategories. There are two separate selections for variants, with germline and somatic variants presenting their own specific subcategories. Finally, the sequences section includes not only the individual sequence information subcategory, but also a "header information" selection that can narrow down a search by gene, transcrip, or exon information.

I went to other organisms (such as chicken, mouse, & rat) to see if their attributes included anything that I didn't see in the human selection. Interestingly, somatic variants as a selection in attributes seems to be exclusive to the Human datasets in Biomart. I assume this is due to the fact that more somatic variants have been studied in humans than the other model organisms, but if anyone has other ideas of why this is, please let me know!

http://useast.ensembl.org/biomart/martview/0070efb516161f9e9094ab88c80fd17a

Unit 5



SAJUNG YUN

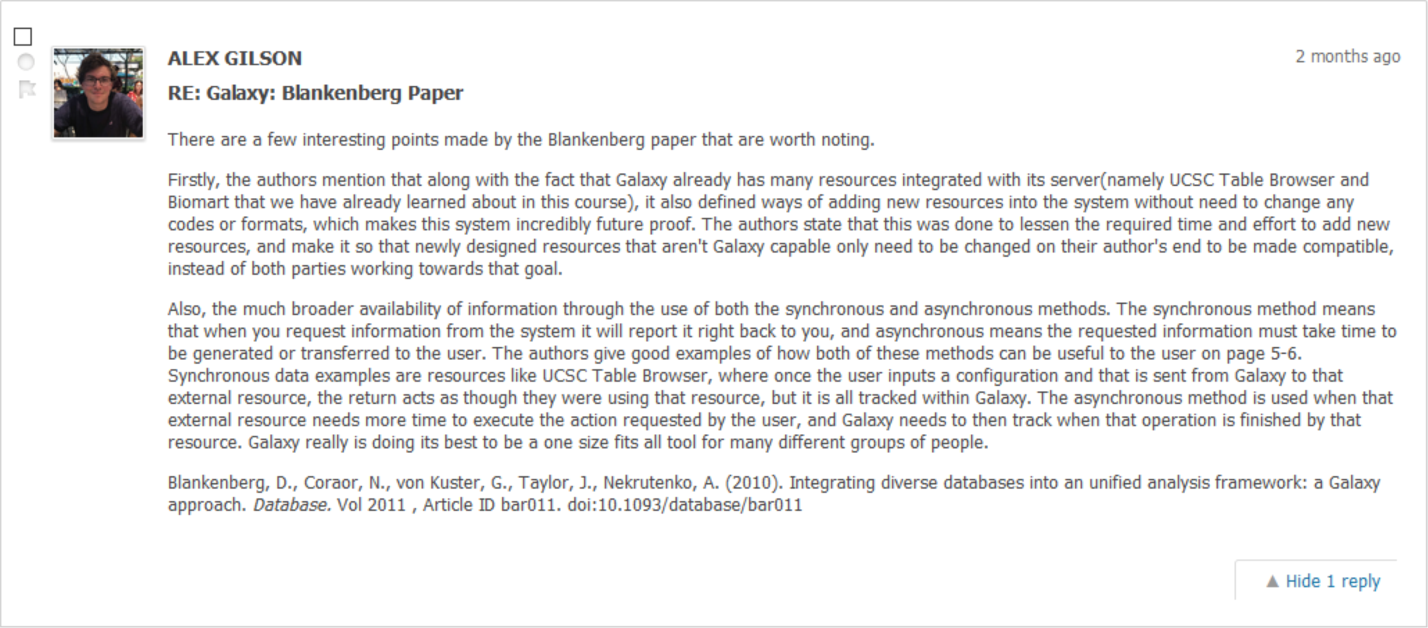
10 months ago

Galaxy: Blankenberg Paper

What else did you take from the Blankenberg paper regarding Galaxy?

blankenberg.pdf (1.091 MB)

https://pubmed.ncbi.nlm.nih.gov/21531983/



ALEX GILSON

2 months ago

RE: Galaxy: Blankenberg Paper

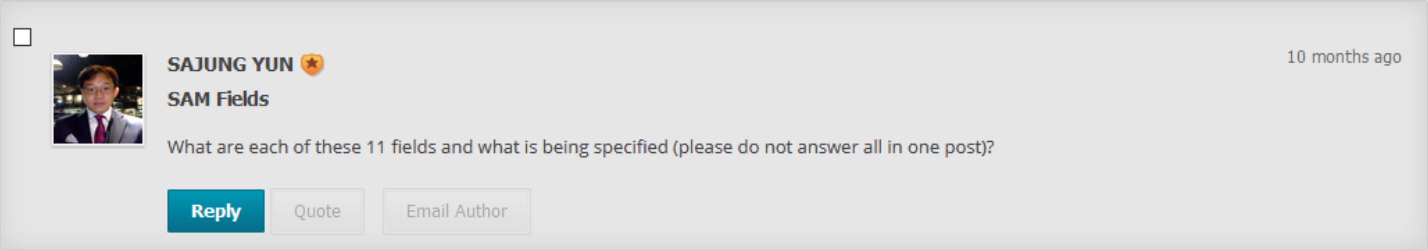
There are a few interesting points made by the Blankenberg paper that are worth noting.

Firstly, the authors mention that along with the fact that Galaxy already has many resources integrated with its server(namely UCSC Table Browser and Biomart that we have already learned about in this course), it also defined ways of adding new resources into the system without need to change any codes or formats, which makes this system incredibly future proof. The authors state that this was done to lessen the required time and effort to add new resources, and make it so that newly designed resources that aren't Galaxy capable only need to be changed on their author's end to be made compatible, instead of both parties working towards that goal.

Also, the much broader availability of information through the use of both the synchronous and asynchronous methods. The synchronous method means that when you request information from the system it will report it right back to you, and asynchronous means the requested information must take time to be generated or transferred to the user. The authors give good examples of how both of these methods can be useful to the user on page 5-6. Synchronous data examples are resources like UCSC Table Browser, where once the user inputs a configuration and that is sent from Galaxy to that external resource, the return acts as though they were using that resource, but it is all tracked within Galaxy. The asynchronous method is used when that external resource needs more time to execute the action requested by the user, and Galaxy needs to then track when that operation is finished by that resource. Galaxy really is doing its best to be a one size fits all tool for many different groups of people.

Blankenberg, D., Coraor, N., von Kuster, G., Taylor, J., Nekrutenko, A. (2010). Integrating diverse databases into an unified analysis framework: a Galaxy approach. *Database.* Vol 2011 , Article ID bar011. doi:10.1093/database/bar011

Unit 7



□

SAJUNG YUN

10 months ago

SAM Fields

What are each of these 11 fields and what is being specified (please do not answer all in one post)?

ALEX GILSON

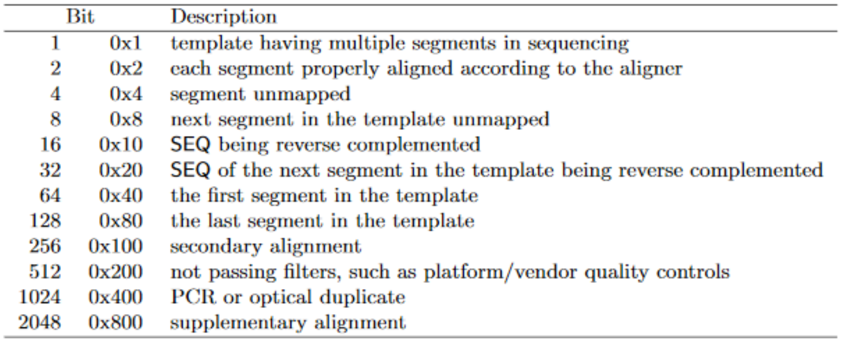
2 months ago

RE: SAM Fields

According to the Heng Li et al(2009) article, the 11 mandatory fields are present for each alignment line in a SAM file, though if info is missing they can be left with a */0 to represent unavailable data. Not much more is stated by the article in terms of what these fields fully represent.

I found a specification sheet from samtools' github titled "Sequence Alignment/Map Format Specification" by "The SAM/BAM Format Specification Working group. This was originally published in 2009 but is continuously updated, with the last version added on Jan 7, 2021. They give a similar table to the Heng Li article, but also give a few definitions for the 11 different mandatory fields.

For example, FLAG represents bitwise FLAGs, and the spec sheet gives the following information about the bits that are included in this field:

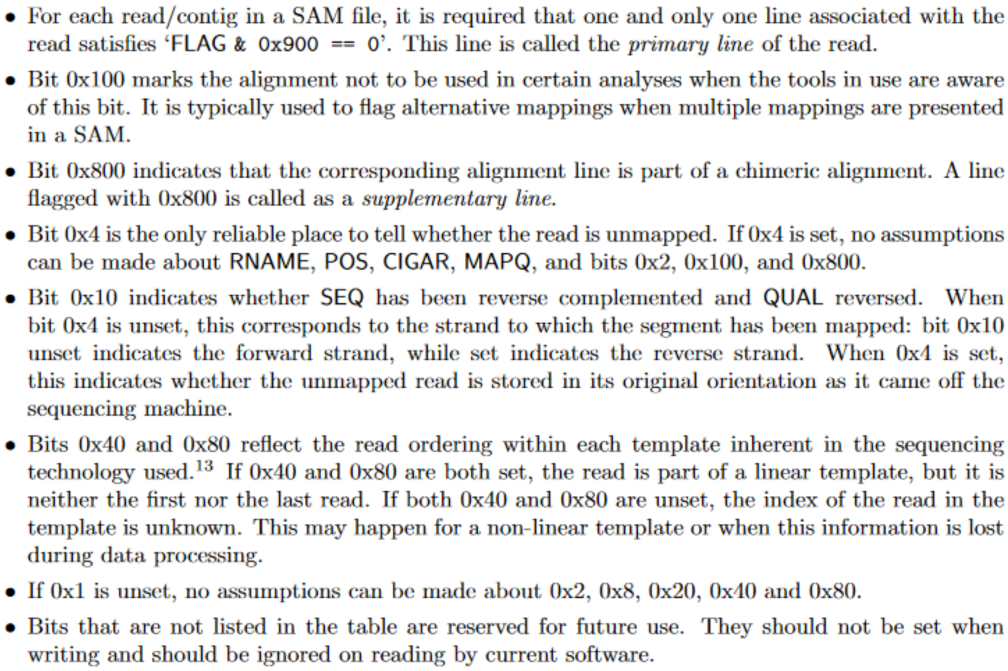


Bit		Description
1	0x1	template having multiple segments in sequencing
2	0x2	each segment properly aligned according to the aligner
4	0x4	segment unmapped
8	0x8	next segment in the template unmapped
16	0x10	SEQ being reverse complemented
32	0x20	SEQ of the next segment in the template being reverse complemented
64	0x40	the first segment in the template
128	0x80	the last segment in the template
256	0x100	secondary alignment
512	0x200	not passing filters, such as platform/vendor quality controls
1024	0x400	PCR or optical duplicate
2048	0x800	supplementary alignment

table retrieved from:

https://samtools.github.io/hts-specs/SAMv1.pdf

The authors and editors then go into great detail about each following this table:



- For each read/contig in a SAM file, it is required that one and only one line associated with the read satisfies 'FLAG & 0x900 == 0'. This line is called the *primary line* of the read.
- Bit 0x100 marks the alignment not to be used in certain analyses when the tools in use are aware of this bit. It is typically used to flag alternative mappings when multiple mappings are presented in a SAM.
- Bit 0x800 indicates that the corresponding alignment line is part of a chimeric alignment. A line flagged with 0x800 is called as a *supplementary line*.
- Bit 0x4 is the only reliable place to tell whether the read is unmapped. If 0x4 is set, no assumptions can be made about RNAME, POS, CIGAR, MAPQ, and bits 0x2, 0x100, and 0x800.
- Bit 0x10 indicates whether SEQ has been reverse complemented and QUAL reversed. When bit 0x4 is unset, this corresponds to the strand to which the segment has been mapped: bit 0x10 unset indicates the forward strand, while set indicates the reverse strand. When 0x4 is set, this indicates whether the unmapped read is stored in its original orientation as it came off the sequencing machine.
- Bits 0x40 and 0x80 reflect the read ordering within each template inherent in the sequencing technology used.[13] If 0x40 and 0x80 are both set, the read is part of a linear template, but it is neither the first nor the last read. If both 0x40 and 0x80 are unset, the index of the read in the template is unknown. This may happen for a non-linear template or when this information is lost during data processing.
- If 0x1 is unset, no assumptions can be made about 0x2, 0x8, 0x20, 0x40 and 0x80.
- Bits that are not listed in the table are reserved for future use. They should not be set when writing and should be ignored on reading by current software.

retrieved from:

https://samtools.github.io/hts-specs/SAMv1.pdf

I haven't worked extensively with SAM files, and I am not sure if this much information is necessary to know about a field of one line for a SAM file, but it is helpful to understand what "FLAG" could mean when there are no other clues.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R., & 1000 Genome Project Data Processing Subgroup (2009). The Sequence Alignment/Map format and SAMtools. *Bioinformatics (Oxford, England)*, *25*(16), 2078–2079. https://doi.org/10.1093/bioinformatics/btp352

https://samtools.github.io/hts-specs/SAMv1.pdf

Reply Quote Edit