Monday, April 21, 2014

TE Gold-Standard Discussion

- A gold-standard TE annotation… Why?
  - What do we want to benchmark?
  - Why would it be useful?
    - should manual curation be the gold standard?
      - how to resolve differences?
      - sometimes tools can do a better job (?)
      - can we trust people
        - their annotations will vary in quality for many reasons)
  - Who would use it? Would you?
    - Not if it is hard to use.
      - Simple format
      - Easy to test tools
      - Wel formatted results
    - Not if their tools do badly
      - maybe, if they're forced to by community pressure
        - need critical mass
          - becomes de facto standard
      - better to use a carrot than a stick
        - = competition
          - would be good for new users to evaluate tools (e.g. specialty tools)
          - sub-categories of TEs
    - Would only work if a large enough portion of the community to buy in
    - Advantage over internal references in each group
      - better communication
        - e.g. easier to talk about at conferences
      - much easier for "small players"
        - similar approaches in phylogeny, etc
        - have no idea how to produce their own

- what would be the properties of an “ideal” TE annotation gold standard?

- how to choose regions?
    - species
        - diversity
        - some tools only work well in specific species
        - representatives of all major + a few crazy different ones
        - quality of gene annotation
        - assembled vs. non-assembled???
            - what would it look like?
            - no, it is a different problem
                - regions of different complexity (like tandem repeats), GC content, isochores, composition, TE density, TE families, centromere/telomere/arm, hetero-/euchromatin, quality/completeness of assembly, ploidy, genome size (total amount of TEs in genome), TE fragmentation, availability of closely related genomes / populations

            - how to annotate?
                - confidence score
                    - or, just only include things we are certain of?
                        - only way to be consistent with “gold” standard?
                        - no, but need to be clear about what is “gold”, silver, bronze, etc
                    - why not probability?
                        - can't evaluate probability on manual annotation (?)
                - start/end genome, TE
                - architecture of the TE
                    - TSDs
                    - termini
                    - internal … whatever
                        - binding sites for different things?
                            - maybe later?
                - age
                - alignment to consensus

- classification
  - families
    - how???
      - we as the annotation group decide?
      - is this a good time to standardize classification, or is it a different huge issue?
      -
      - is all classification arbitrary???
      - level?
        - class, superfamily, family… sub-family?
- weird stuff like transduplications/transductions
- other types of repeats???
- non-TE TE-like things
  - (need better name)
  - gene families
  - exapted TEs
  - similar to antiFAM (protein domain community)
  - in order to differentiate between things just missed (false negatives) from things excluded for good reason
- evidence

- Is it possible?????
  - =====>>>> BREAK….

- make a distinction btwn with vs. without a library?
  - no, really trying to match the final result
  - but if not, how to deal with generating a library
    - make a library but withhold it?
      - no, because you're going to be sad, quickly
-
- how do we update
  - it WILL have errors
    - including missing TEs

        - must be made clear
      - plus: will motivate people to find the errors
      - but if people do find errors, how can we agree?
    - wiki + periodic version updates

Etc.

- artificial vs. real
  - artificial too simplistic
  - too hard to simulate TEs realistically

Discussion #2

*** Is it needed? ***

- Q: what would it be used for?
  - A: to create benchmarks on our own tools
- Manual curation would be too much effort
- Most tools could just use:
  - simulated sequence, reverse genomes, fragmented / evolved
    - etc.
  -
  - Although not tools that use functional etc evidence.
  - If we do use synthetic sequences, we need to evaluate what their deficiencies may be
  - problem with generated sequences for evaluating false positives: no way of knowing which is best
  - false negatives: can use techniques like fragmenting / evolving known TEs
    - but doesn't find ghost sequences (?), unknown TEs
  - suggestion: use the best currently annotated genomes, plus synthetic for false positive detection
    - simulate only what is useful and relatively easy (?), not all evolution
    - eg drosophila, human, thaliana
      - not the whole genome: too arduous, may not be used
      - maybe 1 chromosome, or regions
    - possibly curate whole chromosomes(?) on these?
    - if new, TE-divergent organisms are sequenced, add them
    - use only for de novo, not improving annotation of same genomes

- q: experimental data?
  - yes, everything
  - give a higher score for experimentally validated TEs
- proposal should include required metrics output for benchmark
- process should be to have a website to upload annotation
  - automate evaluation
  - big problem: who would do it? (no answer)
    - maybe ppl who did similar with snips (ask Guillome)
- include mix of elements
  - eg difficult like Alu
- competition?
  - problem: might draw in new grad students w/ no experience

1. everyone submit proposals for benchmarks
2. platinum human genome
    1. highly analyzed, snp, etc
3. not to compete against each other

- upload proposals to website
  - Google Doc?
  - also upload actual benchmarks

  - what are the constraints, e.g. de novo vs homology
- metrics
- internal benchmarks
- (make it possible for everyone to comment)
- template for suggestions
-
- what is the “white paper”?
  - publish in Mobile DNA journal?
    - (opinion paper)

Discussions Last Day

Crowd-Sourcing

- 2 approaches: game-ification, workbench
-
- problems that can be solved
- library creation:
  - fragmentation is a common issue
    - need to manually join fragments together

- workbench
  - anyone with some knowledge of TEs should be able to use it

Databases
- need a centralized database for all TE sequences
- like NCBI
  - but NCBI doesn't want to deal with reconstructed sequences
- clearing house
  - links to other databases