Query = SNORD116

* 30 SNORD116 variants identified
  + ? long poly-A /-T stretches ?!
* SNORD116 has no conserved D’ site
* Binding typically downstream of C or upstream of D sites, i.e. no in center region
  + → RRI constraints ?! eg. seed location in ranges CD’ and C’D
  + → maybe consensus SNORD116 structure as accessibility and location constraint

Target = ???

* Plus and minus strand of
* RRI screen in ???
  + Human genome (all chromosomes)
  + Transcriptome
  + Specific genes
  + Spliceosome
* Sequences already available/read @Stamm lab?

snoRNA target scan software/approaches checked? (Hakim Tafer, Jana Hertel, Peter Stadler)

* RNAplex-based PLEXY - <http://www.bioinf.uni-leipzig.de/Software/PLEXY/>
* PhD thesis Stephanie Kehr - <https://ul.qucosa.de/api/qucosa%3A15187/attachment/ATT-0/>
* Snotarget - <http://bpg.utoledo.edu/~shuhao/lab/DOWNLOADS_WEB/23.pdf> (dead)
* Via blast with specific constraints on interaction site <https://www.frontiersin.org/articles/10.3389/fpls.2021.731484/full>

Output

* Mfe
* Subopts (multiple regions)
* Combination of both snoRNA interacting regions, ie C2D + C’2D’?
* Spot probability vector
  + for all targets?
  + For each snoRNA ?
  + Later for specific targets?

######################

1. 30 snornas
2. 3020 gene → get sequences (6040 am ende)
   1. Plus strand
   2. Minus strand
   3. Jeweils +- 1000nt context
   4. → ggf. Einschränken via längen cutoff
3. Run gUUgle
   1. Ggf. query sequences vorn 15nt und hinten um 11 kürzen, damit match zwischen den boxen und letzte CD box intakt
   2. Min match length 9 (ggf. Count stats für versch. thresholds)
      1. Stats on
      2. Grep “^MatchLength:” und gleich bzippen (d.h länge und positionsinfo)
   3. Grep match statistics
   4. Decide based on stats ob wir nochmal min=8 rechnen oder gar min=10 aus aktuellem output rausziehen
4. Identify target regions of interest (merge für alle queries)
   1. Merge aller guugle hits +- 20nt
   2. Target region = region +- 200nt (plfold window für accessibility)
5. RNAplfold für target-regions rechnen
   1. Je target region speichern
6. Für jeden guugle hit:
   1. Rechne intarna mit
      1. Target-constraint = +200 bis -200 vor ende
      2. Target pu values from RNAplfold file
      3. Target-start-index auf region start setzten (für output indexing
      4. + Rigide constraints
      5. + ggf. Helix-based computation
7. Intarna output mergen und p-values rechnen
8. Beste treffer inspizieren/visualisieren
   1. P-value cut off
   2. E cut off
   3. …

####################

Conda env

* Intarna
* Gcc + guugle-build dependencies