Emerging tools for RNA structure analysis in polymorphic data

Jan Gorodkin
Center for non-coding RNA in Technology and Health (http://rth.dk)
University of Copenhagen

Content:

- Movitation
- Mutations in RNA structure
- Disease applications
- Perspectives

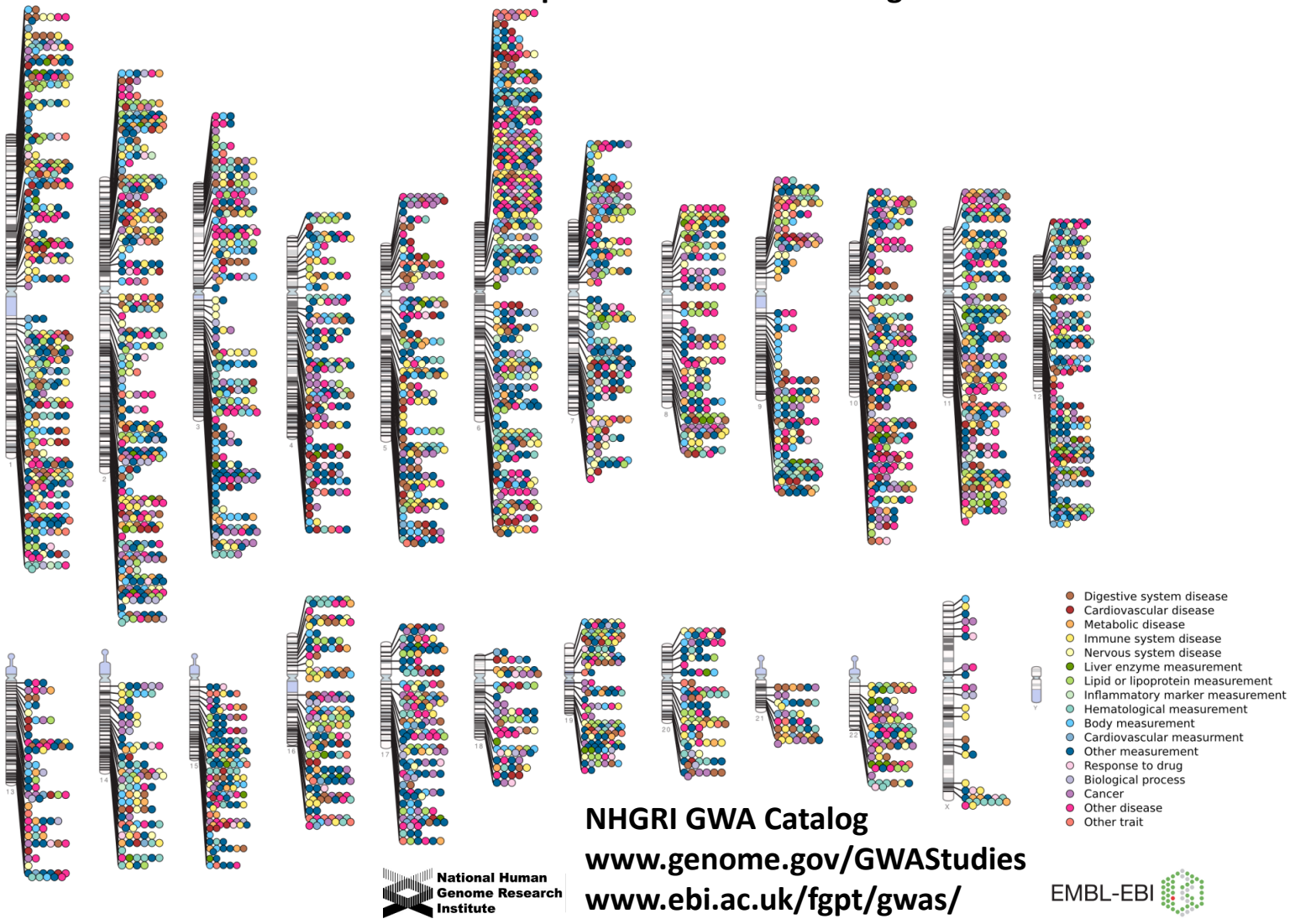
Single Nucleotide Polymorphisms (SNPs): Where are they?

- SNPs can direct phenotypes and diseases
- non-synonymous (ns) SNPs can alter a protein structure
- SNPs can induce/destroy microRNA target
- Probably far most disease studies aim at identifying nsSNPs

SNPs: Where are they?

Disease and trait associated SNPs†: 88% intronic or intergenic.

Published Genome-Wide Associations through 12/2013
Published GWA at p≤5X10^{-8} for 17 trait categories



† (Hindorff *et al.*, PNAS, 2009; MacArthur *et al.*, Nucl Acids Res, 2013)

SNPs: Where are they?

SNPs are outside coding regions
2,619 disease and trait associated SNPs of cancer GWAS loci[†]:

Classification	Approx percentages	Approx numbers
Intronic	40	1,047
Intergenic	32	838
Within non-coding seq of a gene	10	262
Upstream	8	210
Downstream	4	105
Non-synonymous coding	3	79
3' untranslated region	$\sim$1	26
Synonymous coding	$\sim$1	26
Unknown	$\sim$1	26

[†] (Freedman *et al.*, Nat. Genet., 2011)

The genome is potentially full of RNA structure

Recent independent studies indicates $>$ 10% of the genome is structured.

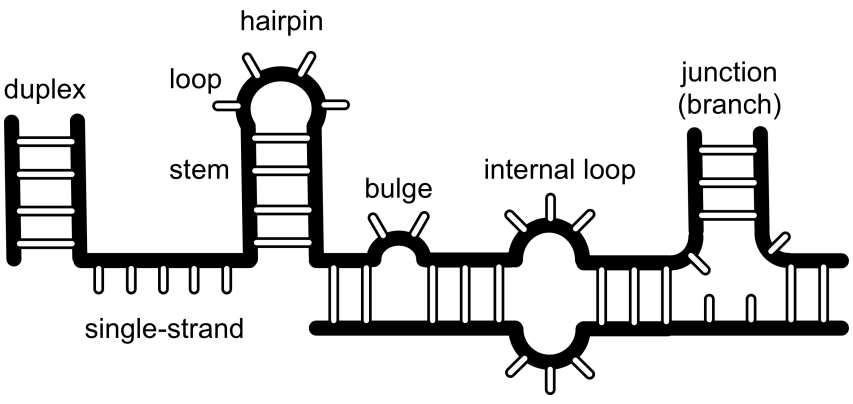
- *In silico* study of mammalian genome : $\sim$ 13% [†]
- $\sim$15% of all transcribed Single Nucleotides Variants (SNVs) locally alter the RNA structure in human[*]
- $> 10,000$ transcripts structured in *A. Thaliana*[‡]

The current analyses point in the direction that a non-neglible amount of the transcriptome make up structured RNA.

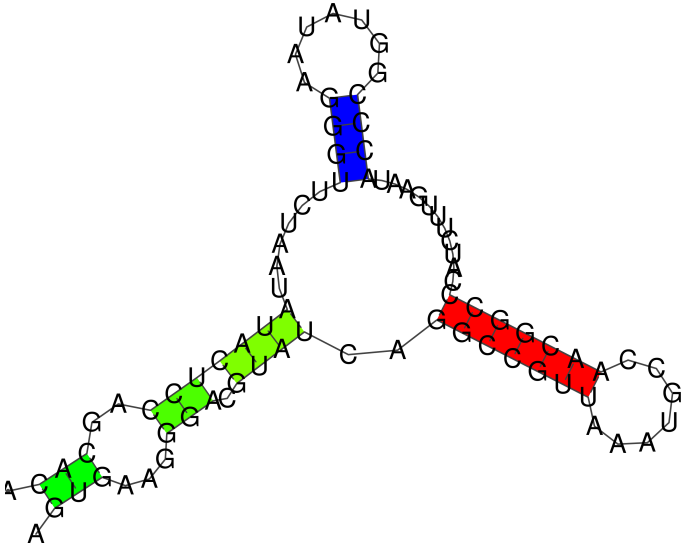
[†] (Smith *et al.*, Nucl Acids Res, 2013); [*] (Wan *et al.*, Nature, 2014); [‡] (Ding *et al.*, Nature, 2014)

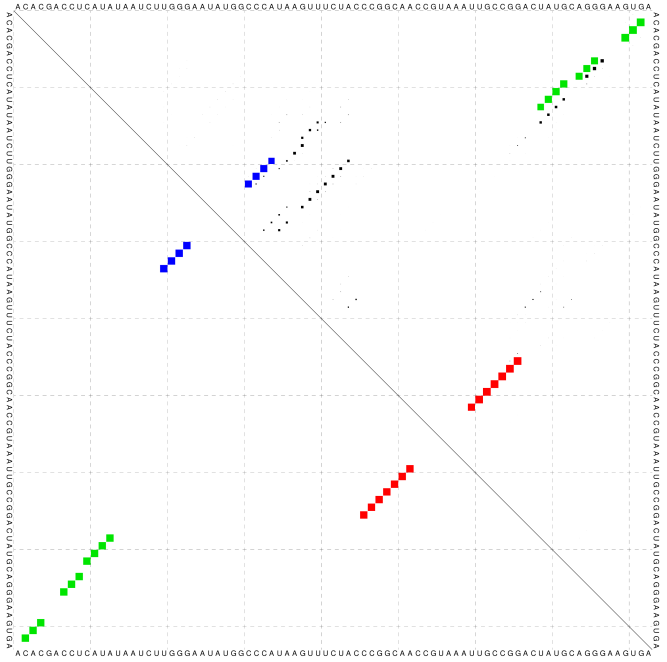
Computational folding of RNA sequences

Contributions from structural components
Folding is an optimization problem



(From Zwieb, Meth Mol Biol, 2014)





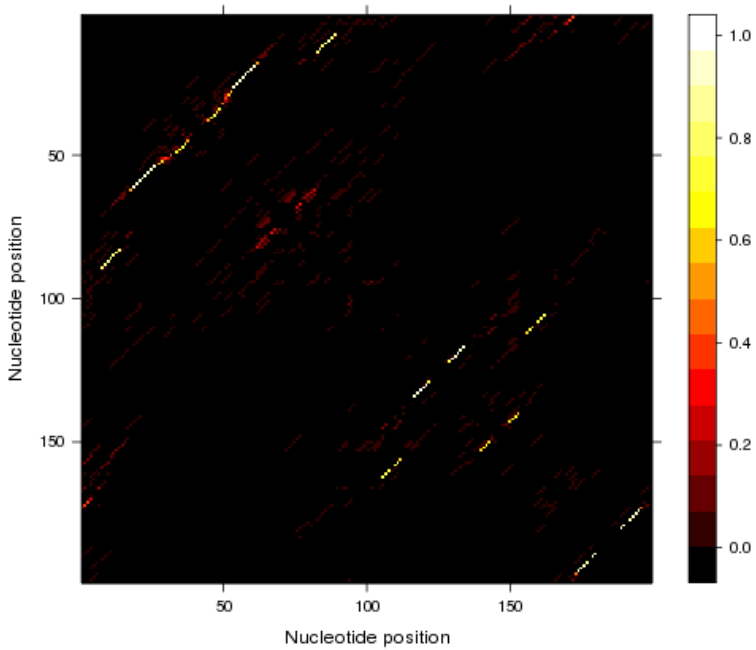
Exploit the *ensemble* of all structures
Computing a dot-plot

(From Hofacker, Meth Mol Biol, 2014)

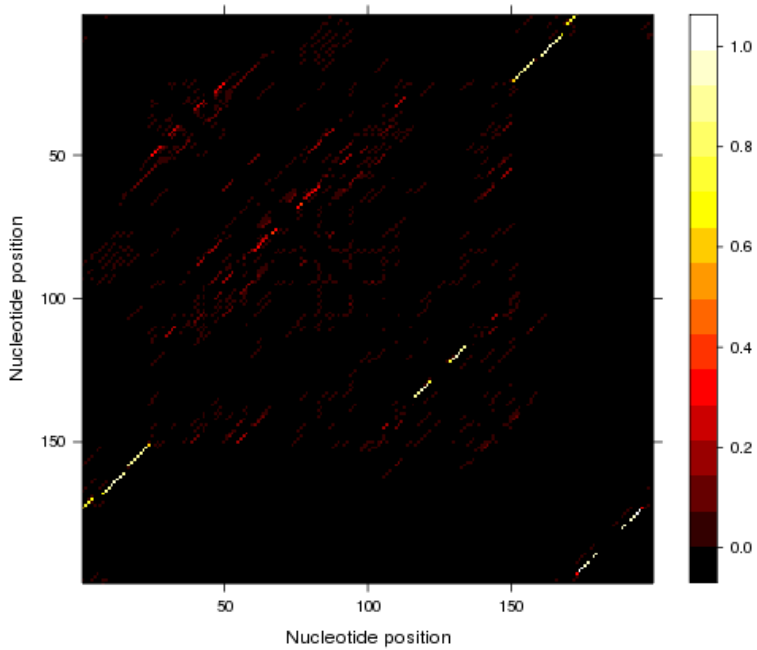
Effect of mutations in RNA sequences

Global structural change: SNP could change the base-pair probabilities of the global RNA structure.

Example: SNP C14G in 5'UTR of the FTL gene (in an IRE hairpin)



wild-type

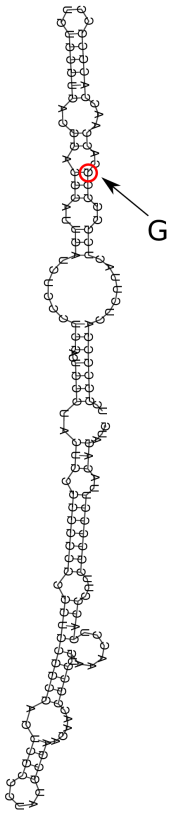


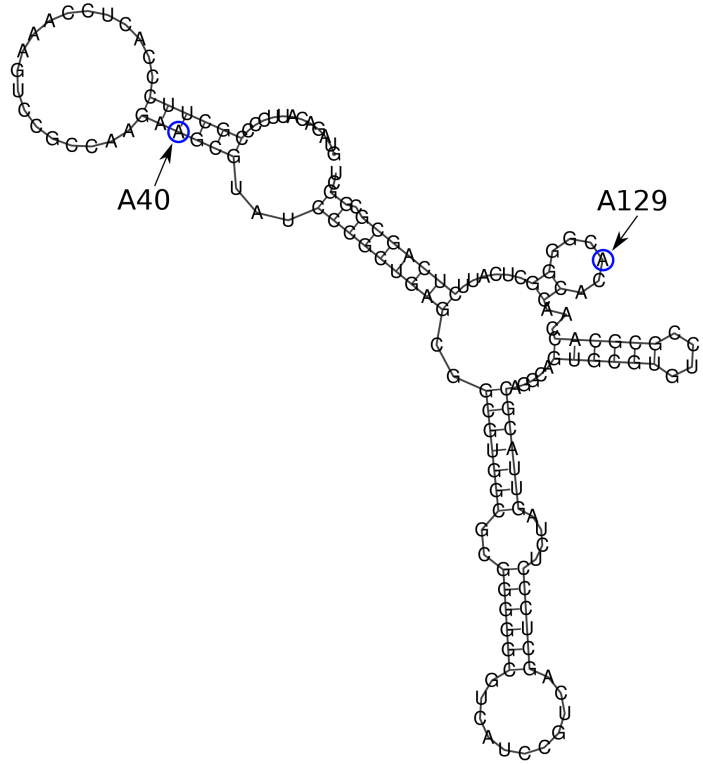
mutant

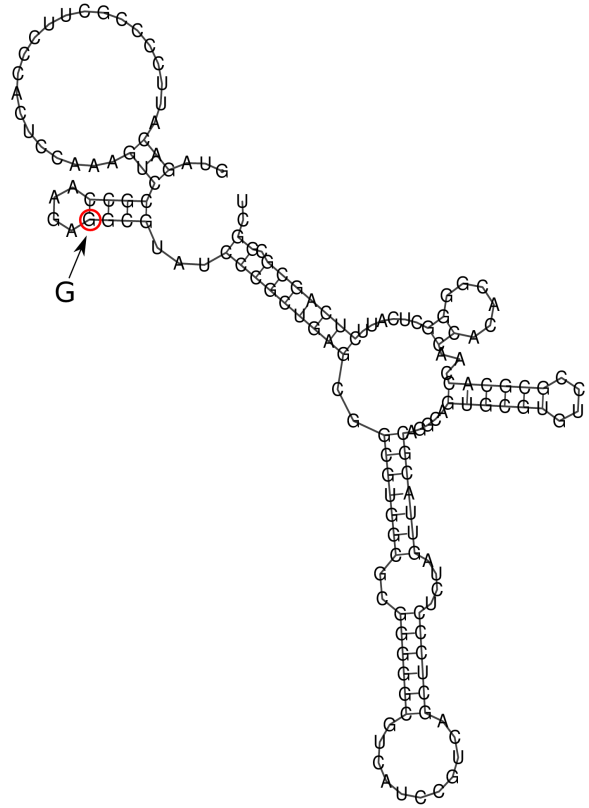
Structural changes in IRE - aberrant FTL (ferritin, light polypeptide) gene regulation - hereditary hyperferritinemia-cataract syndrome

Effect of mutations in RNA sequences

Global versus local







mutant: A129G

wild-type

mutant: A40G

Small local structural change in functional motifs can have striking effect on the RNA functions[†].

[†] (Westerhout et al., 2005; Abbink et al., 2008; Hemert et al., 2008; Grover et al., 2011)

Motivation

- Impact of SNPs in non-coding RNA structure and function.
- Existing methods detect *global* changes
  - `RNAmute`[a,b]
  - `RDMAS`[c]
  - `RNAmutants`[d,e]
  - `SNPfold`[f]
- Overcome limitations by searching for *local structural changes*.
- `remuRNA`[g]: Entropy based measure. Local version by average windows sorrounding the SNP.

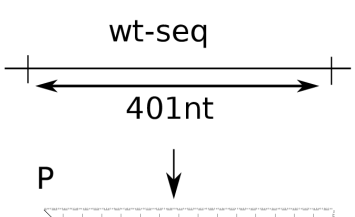
---

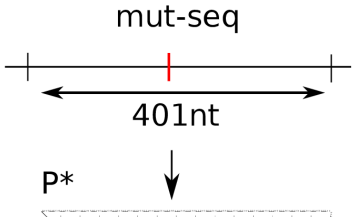
[a](Barash, Nucl Acids Res, 2003); [b](Churkin and Barash, BMC Bioinform, 2006); [c](Shu et al., BMC Bioinform, 2006); [d](Waldispuhl et al., PLoS Comput Biol 2008); [e](Waldispuhl et al., Nucleic Acids Res 2009); [f](Halvorsen et al., PLoS Genet, 2010); [g](Salari et al., Nucl Acids Res, 2012)

Pipeline concept

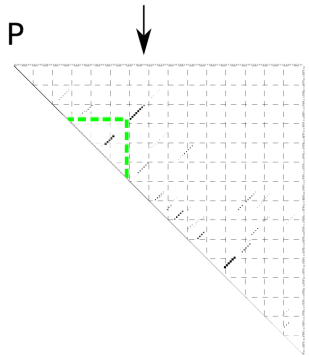
RNAsnp† detection of locally changed structure.

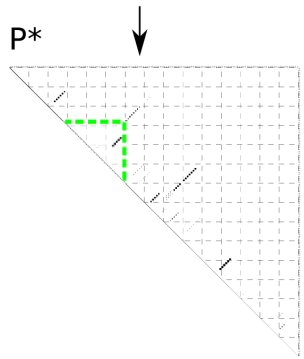
Generate input sequences



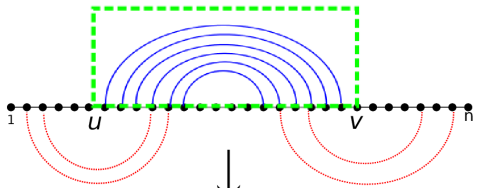


compute base pairing probability matrix (using `RNAfold -p`)





select local regions with self-contained base pairs



compute difference in base pairing probabilities between wt and mut

Euclidean distance:

$$d^2_{[u,v]}(P,P^*) = \sum_{i=u}^{v}\sum_{j=i}^{v}(P_{ij}\text{-}P^*_{ij})^2$$

Pearson cor. coeff.:

$$r_{[u,v]}(P,P^*) = \frac{cov(P,P^*)}{\sqrt{var(P)var(P^*)}}$$

where, P_{ij} - probaility of base *i* being paired with *j*

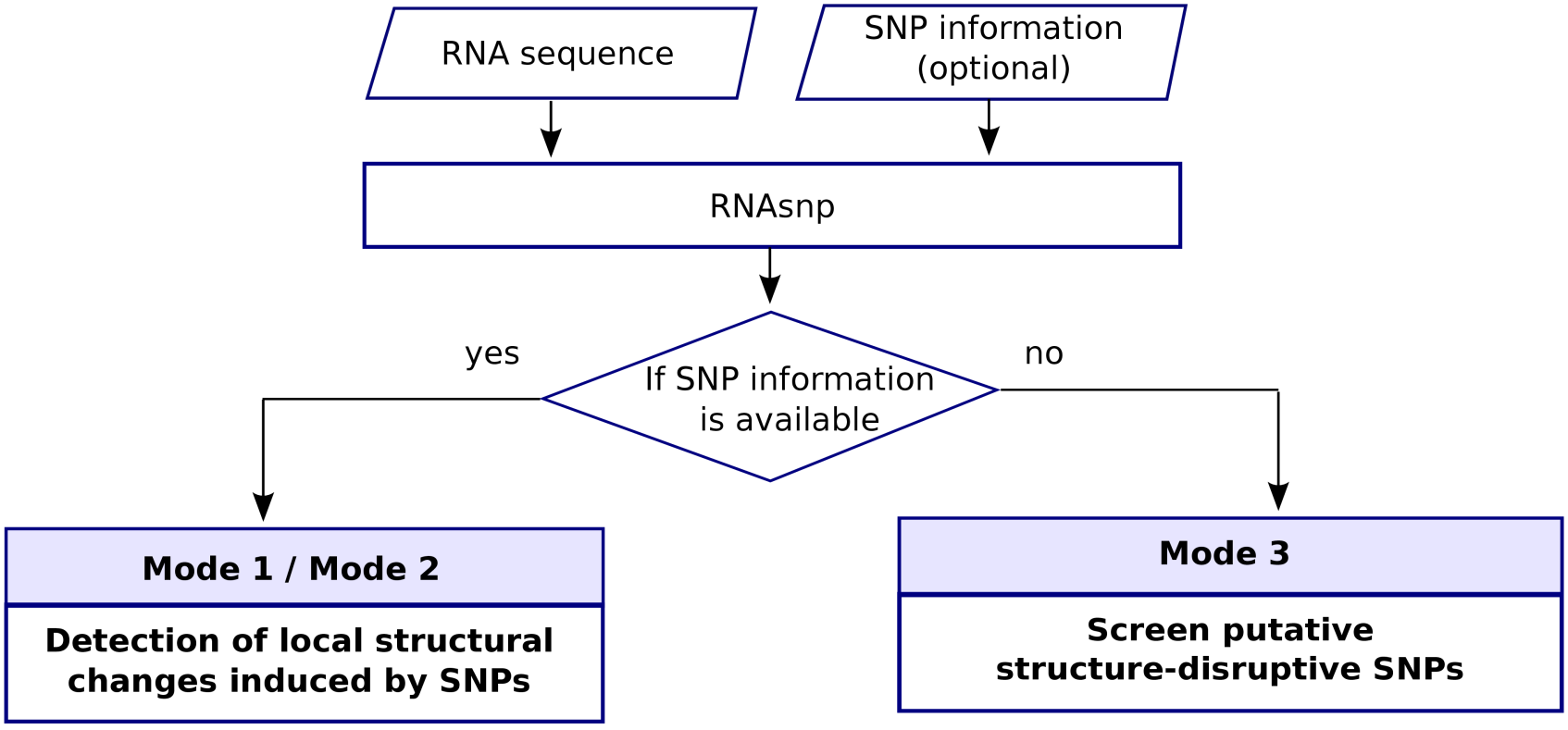
report the local region with maximum difference and its significance

$$[k,l] = \arg\max_{u,v} d^2_{[u,v]}(P,P^*)$$

$$[k,l] = \arg\min_{u,v} r_{[u,v]}(P,P^*)$$

† Sabarinathan, Tafer, Seemann, Hofacker, Stadler, Gorodkin. Hum Mutat, 2013

RNAsnp pipeline



Mode 1 - based on global folding method (`RNAfold`)
Mode 2 - based on local folding method (`RNAplfold`)
Mode 3 - combination of mode 1 and 2.

The SNP effects are quantified in terms of empirical *P*-value
P-values: Empirically (∼156 CPU years)

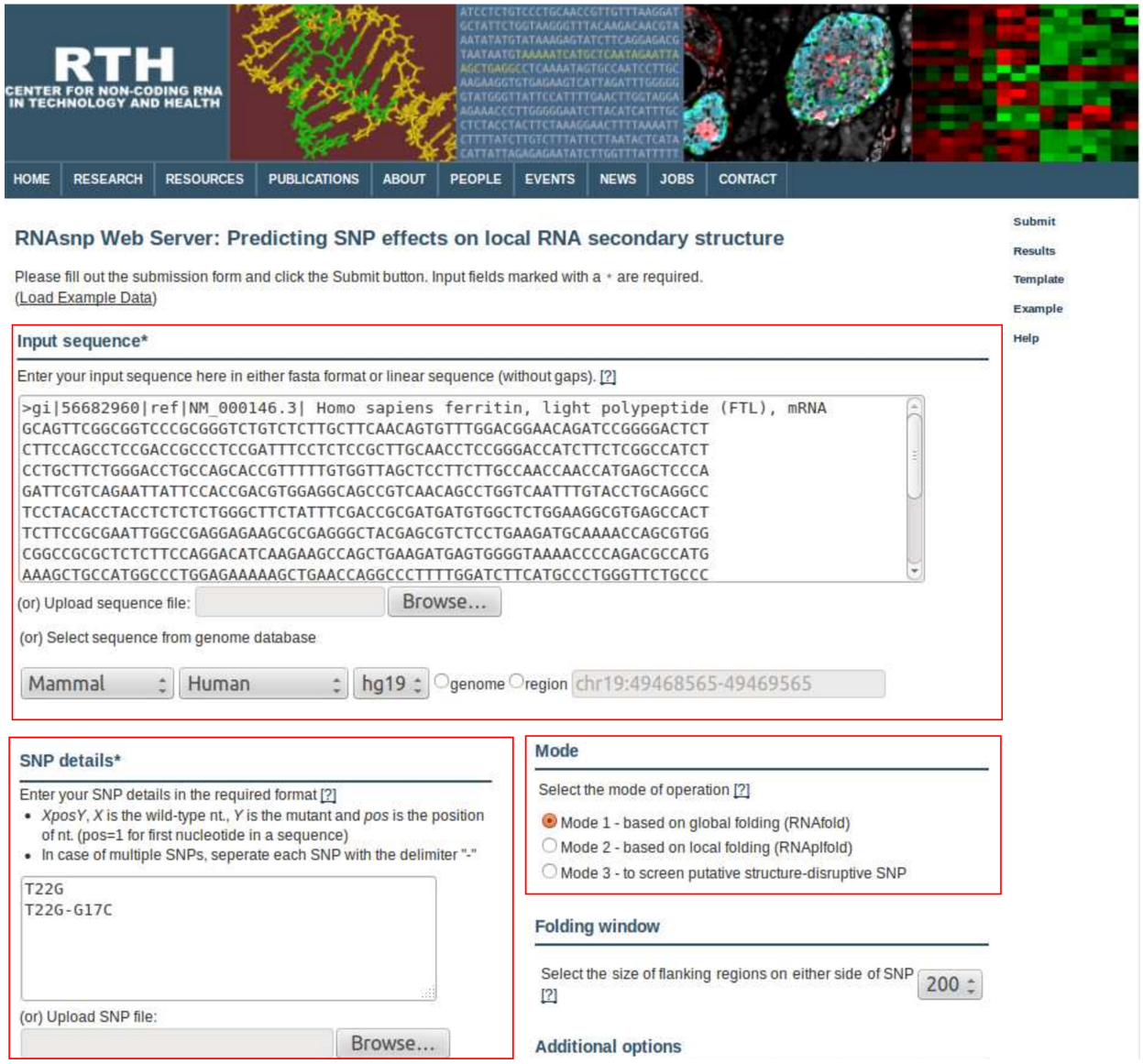
Predicting structural effects of disease associated SNPs

Overlap of 20 candidates by d_{max} and r_{min} of which `SNPfold` overlap 3 (grey).

Disease/phenotype	Gene	HGMD Accession	UTR	Genbank Accession	NTs	SNP	*p*-value $p(d_{max})$	*p*-value $p(r_{min})$
Pseudohypoaldosteronism	NR3C2	CR030126	5	NM_000901	5898	C362G	0.017	0.022
Hypertension	EDN2	CR994679	3	NM_001956	1243	G999A	0.036	0.021
Obesity	CNR1	CR073542	3	NM_033181	5373	A3777G	0.032	0.036
Myocardial infarction	GP1BA	CR022116	5	NM_000173	2463	U71C	0.040	0.037
Colorectal cancer	INSR	CR082021	3	NM_001079817	9023	A4326G	0.042	0.030
Graves'disease	FCRL3	CR067134	5	NM_052939	3019	G282C	0.011	0.042
Increased triglyceride levels	ABCA1	CR025352	5	NM_005502	10502	C126G	0.044	0.022
Insulinresis.hypertension	RETN	CR032443	3	NM_020415	478	G435A	0.045	0.043
Cartilage-Hairhypoplasia	RMRP	CR063417	ncRNA	NR_003051	268	A215G	0.048	0.027
Hypercholesterolaemia	LDLR	CR971948	5	NM_000527	5283	C174A	0.025	0.048
Glaucoma	CYP1B1	CR032431	5	NM_000104	5153	C118U	0.063	0.036
Reduced transcriptional activity	NR3C1	CR016150	5	NM_001024094	6787	C274A	0.044	0.063
HDL cholesterol levels	LIPG	CR032437	3	NM_006033	4141	A2237G	0.051	0.065
FactorVIIdeficiency	F7	CR090334	5	NM_019616	3059	U8C	0.066	0.042
HaemophiliaA	F8	CR070421	5	NM_000132	9035	G60A	0.074	0.010
Cartilage-Hairhypoplasia	RMRP	CR064472	ncRNA	NR_003051	268	U10C	0.076	0.024
VonHippel-Lindau syndrome	VHL	CR011856	3	NM_000551	4560	C862G	0.076	0.065
Obesity	SLC6A14	CR035766	3	NM_007231	4564	C2238G	0.078	0.062
Spasticparaplegia31	REEP1	CR082030	3	NM_022912	3853	C764U	0.033	0.081
Hyperferritinaemia-cataract syndrome	FTL	CR061334	5	NM_000146	871	U22G	0.052	0.097

RNAsnp web server

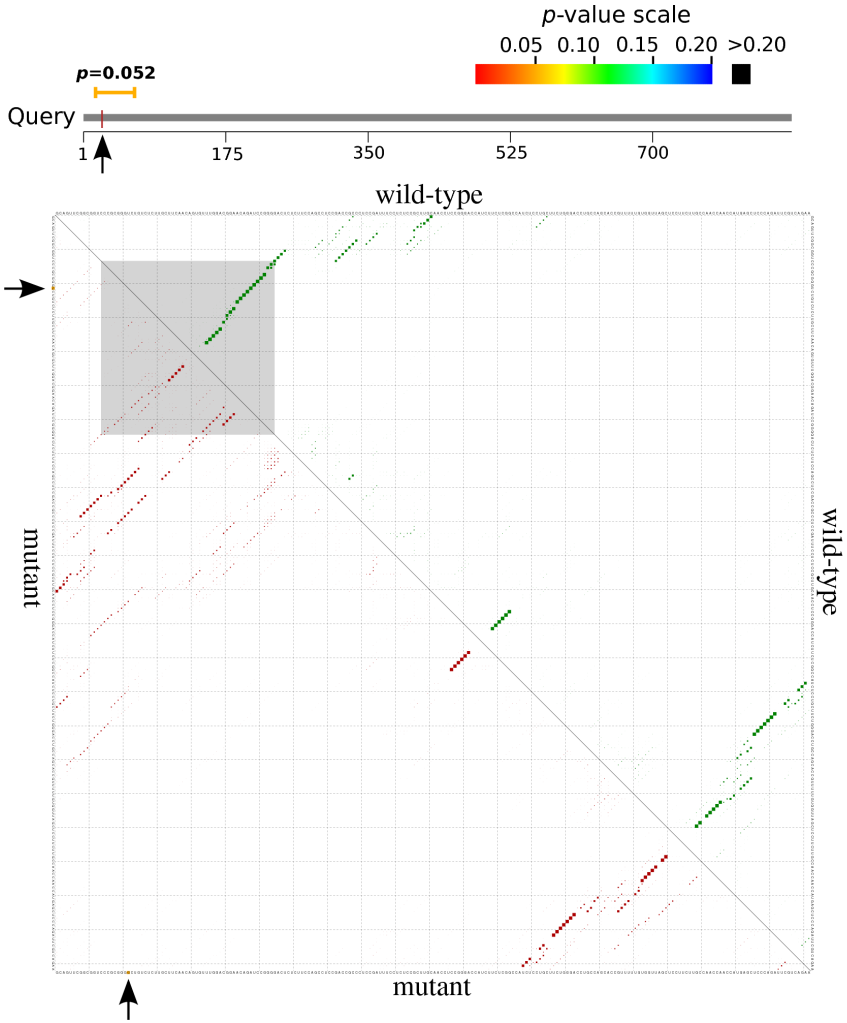
RNAsnp Web server input[1] (http://rth.dk/resources/rnasnp)

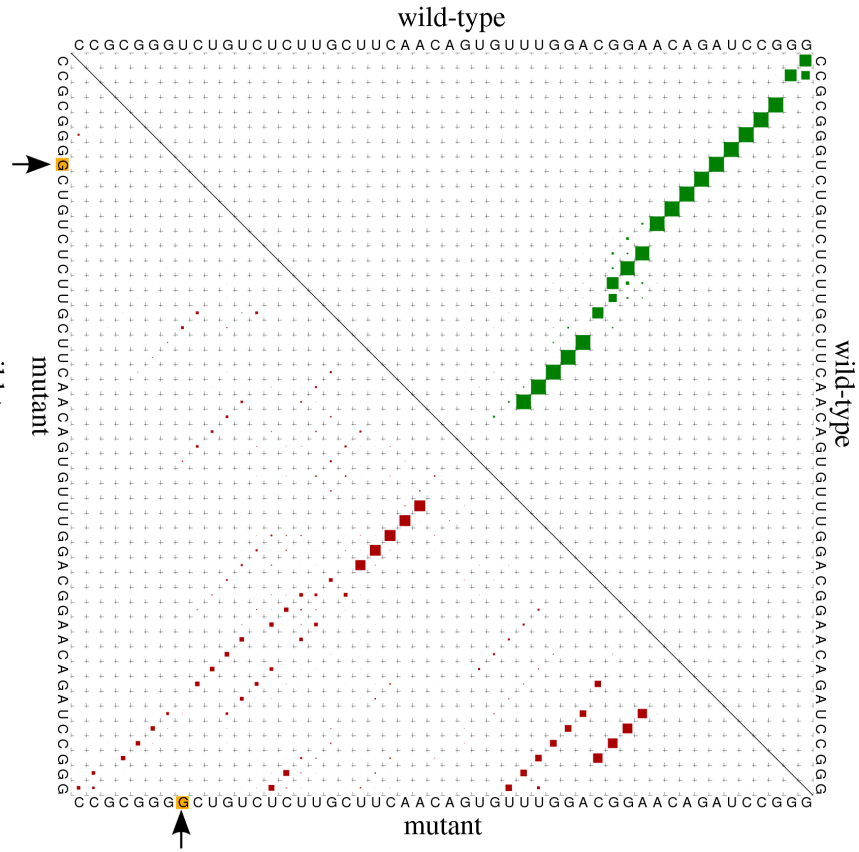


[1] (Sabarinathan *et al.*, Nucl Acids Res, (Web Server Issue), 2013)

RNAsnp web server

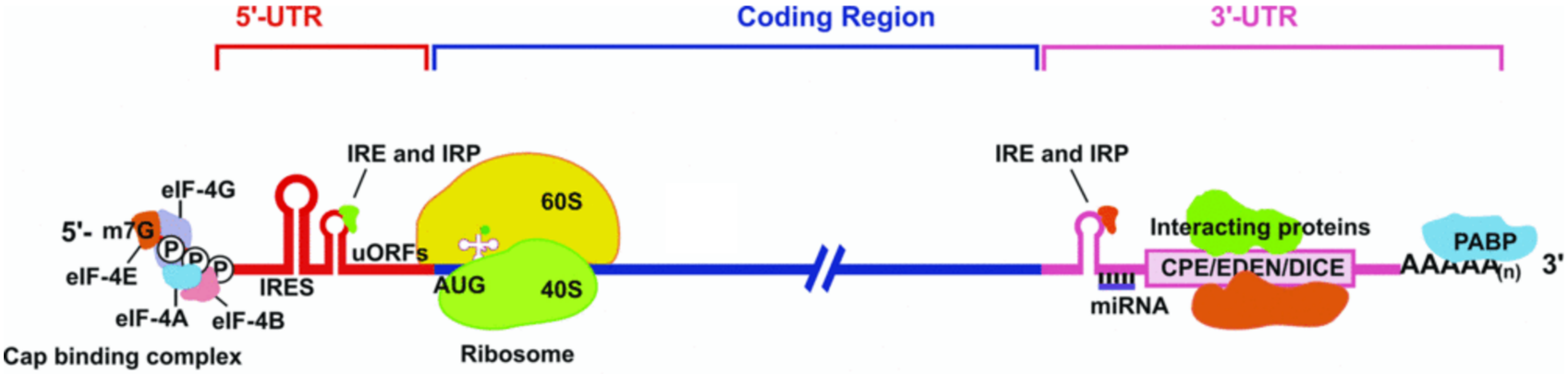
RNAsnp Web server output[1]





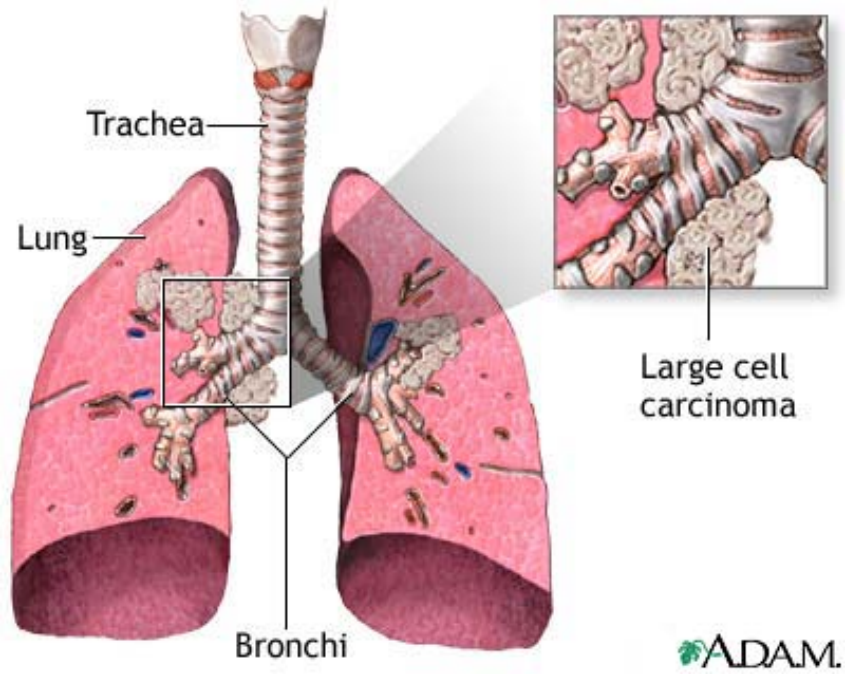
[1] (Sabarinathan *et al.*, Nucl Acids Res, (Web Server Issue), 2013)

RNA structure in protein coding genes



Analysis SNPs in UTRs expressed in lung cancer

Non-small cell lung cancer (NSCLC) is the most common form of lung cancer (activating mutations in *KRAS* oncogene).



(http://lungcancer.ucla.edu/adm_lung_cancer_nonsm.html)

→ Transcriptome-wide sequencing of lung (adenocarcinoma) tumors. *KRAS*)

→ SNVs effects predicted for coding regions (and splice sites).

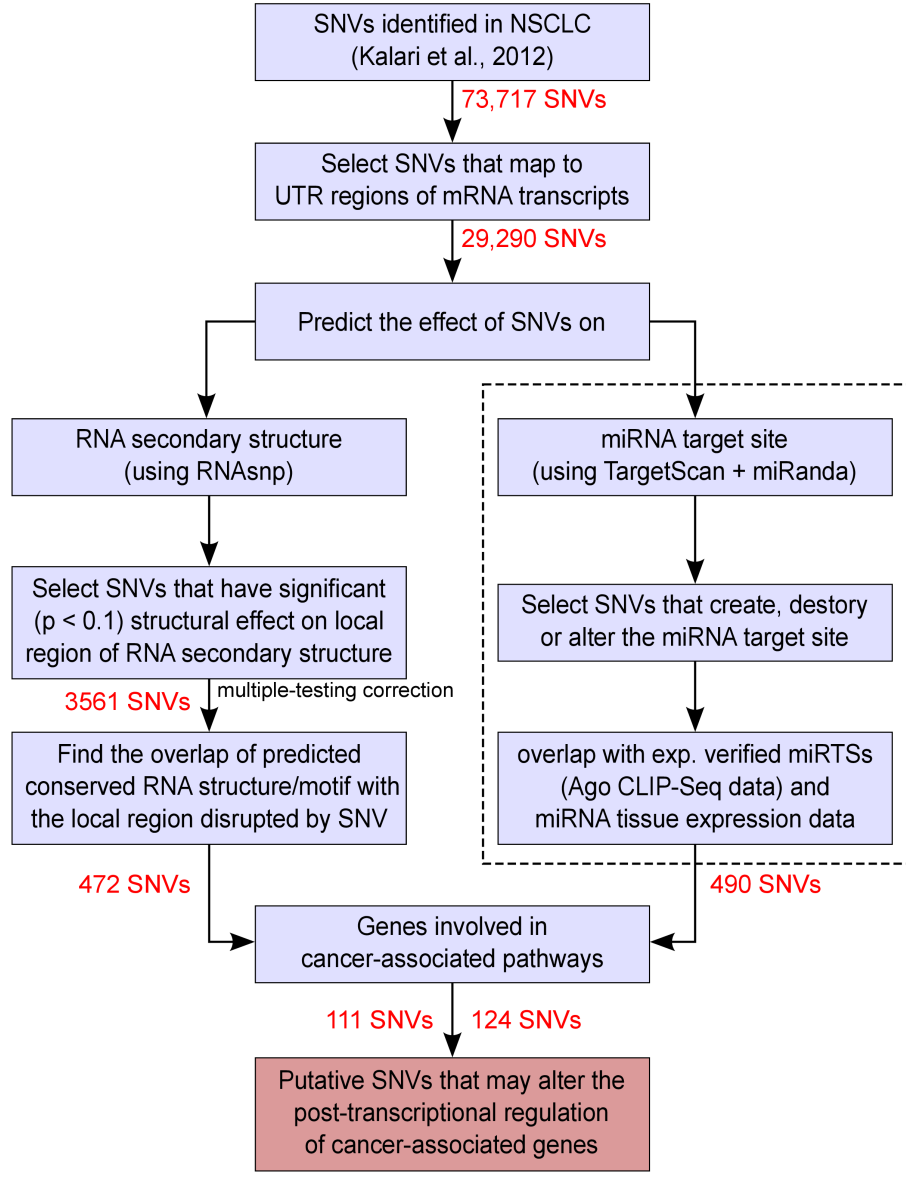
→ About 40% of the total SNVs (73,717) maps to UTRs.

† Kalari KR, *et al.*, *Front. Oncol.*, 2012.

Analysis SNPs in UTRs expressed in lung cancer

Combine[‡]

- `RNAsnp`
- miRNA target prediction
  - `TargetScan`
  - `miRanda`



[‡] Sabarinathan, Wenzel, Novotny, Tang, Kalari, Gorodkin. PLoS One, 2014.

Results

→ Data set contains 29,290 SNVs (in 6462 genes)
  → Of these, 6519 SNVs are in 1347 cancer-related genes‡

Cancer-related genes:

→ 20.8% to begin with.

→ 23.4% after pipeline (P=0.032)

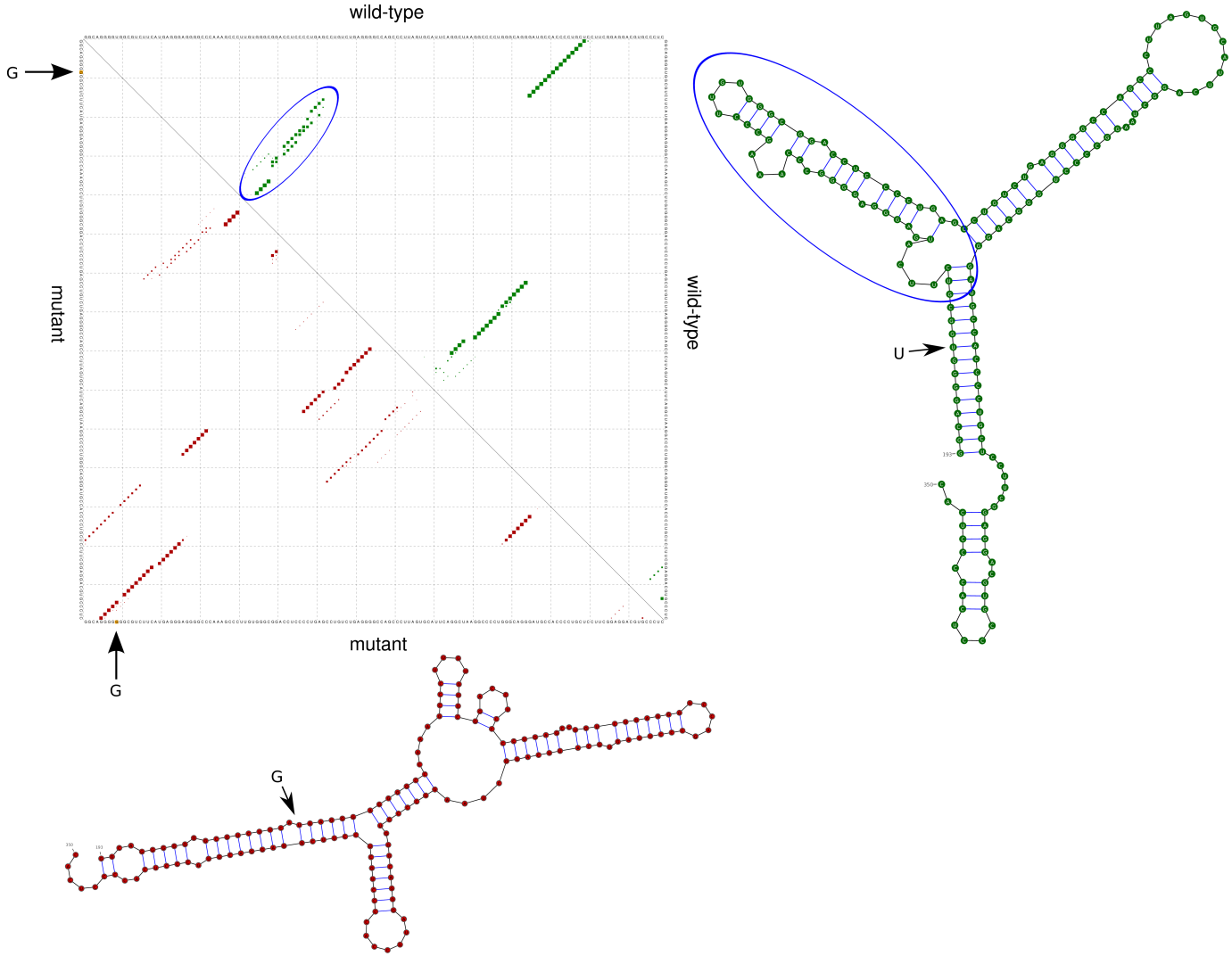
Some details:

	Effect of SNVs on		
gene type	**Sec. Str.** (#SNVs)	**miRNA TS** (#SNVs)	**both** (#SNVs)
All	472 (in 408 genes)	490 (in 447 genes)	48
Cancer-related	111 (in 98 genes)	124 (in 104 genes)	15

‡ obtained from COSMIC & Qiagen data bases

Analysis SNPs in UTRs expressed in lung cancer

Effect of SNVs on RNA secondary structure of *GPX3* mRNA



SNV U1552G predicted to cause significant local secondary structure changes (d_{max} p-value: 0.0474 in 3' UTR of GPX3 mRNA. This local change disrupt the structure of SECIS (blue circle).

Outlook

- RNAsnp tool for analyzing RNA structure disrupting SNPs.
- Taking 3D structure into account.

Webservers, software, data resources: http://rth.dk/resources.

Acknowledgements

Uni CPH / RTH:

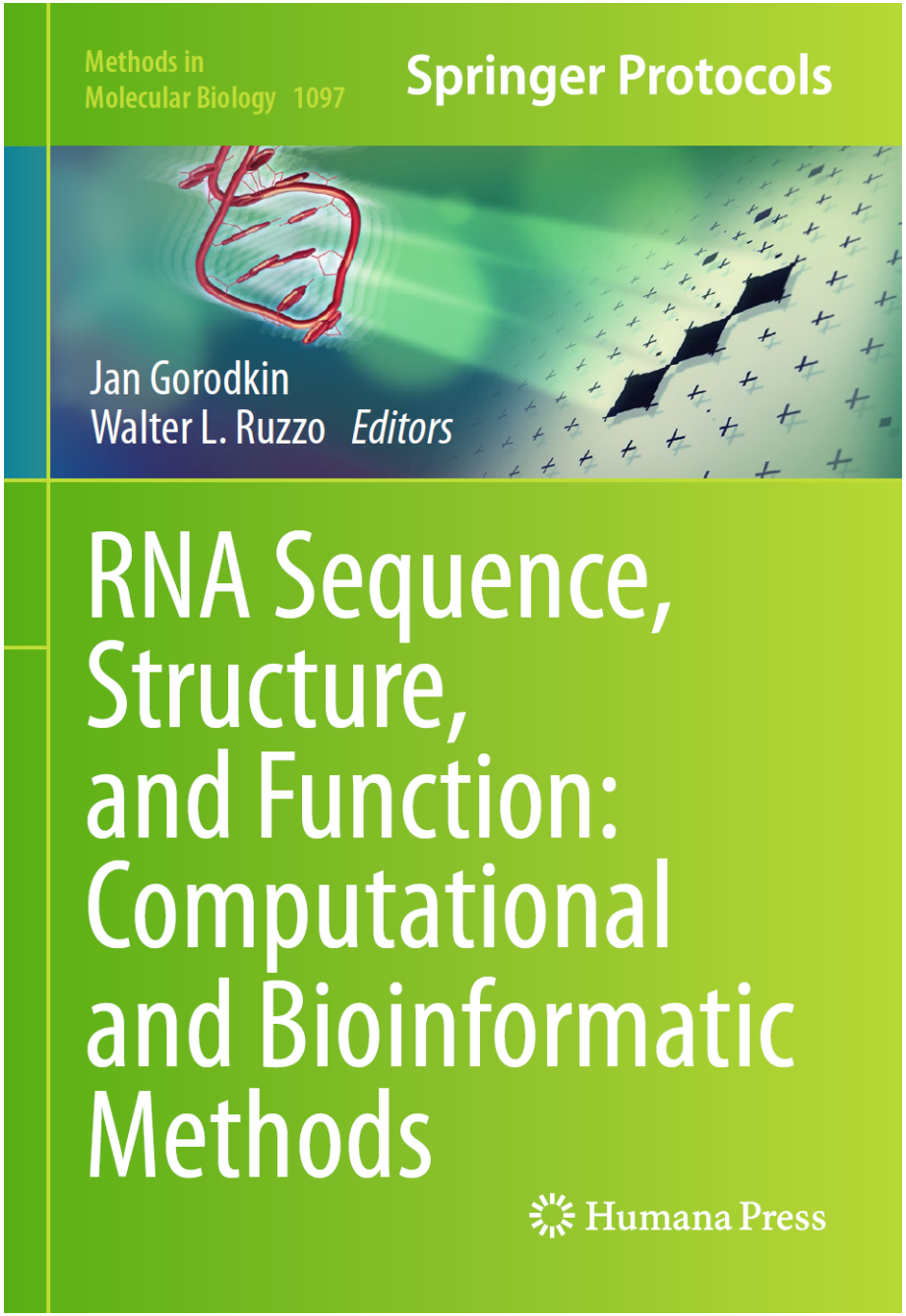
- Sabarinathan Radhakrishnan (Alumni)
- Stefan E. Seemann
- Jakob H. Havgaard
- Christian Anthon
- Anne Wenzel
- Peter Novotny
- Ferhat Alkan
- Nikolai Hecker
- Xiaoyong Pan
- Rebecca Kirsch
- Corinna Theis
- Victor Carmelo
- Alexander Junge
- Daniel Sundfeld
- Shiqi Zhang

External collaborators:

- Walter L. Ruzzo, Washington University, Seattle
- Peter Stadler, University of Leipzig
- Hakim Tafer, University of Leipzig
- Steve Hoffmann, University of Leipzig
- Rolf Backofen, University of Freibrug
- Ivo Hofacker, University of Vienna
- Krishna R. Kalari, Mayo Clinic
- Xiaojia Tang, Mayo Clinic

Funding:

- Danish Strategic Research Council
- Innovation Fund Denmark
- Danish Center for Scientific Computing
- The Lundbeck Foundation



http://rth.dk/rnabook

Upcoming Elixir position in RNA tools infrastructure (gorodkin@rth.dk)