An Integrated Tool to Assess the Functional Impact of SNPs

Peter F. Stadler

Bioinformatics Group, Dept. of Computer Science &
Interdisciplinary Center for Bioinformatics,
University of Leipzig
Max Planck Institute for Mathematics in the Sciences
RNomics Group, Fraunhofer Institute for Cell Therapy and Immunology
Institute for Theoretical Chemistry, Univ. of Vienna (external faculty)
Center for non-coding RNA in Technology and Health, U. Copenhagen
The Santa Fe Institute (external faculty)

Nantes, 29 Aug 2013

Diverse Effects of SNPs

Aim within QUANTOMICS Identify likely **causal** polymorphisms by quantifying molecular effects of substitutions on genome-wide scales. Annotation dependent!

- changes of amino acid sequences
  - only 1.5% of the genome code for proteins
  - well-established tools are available for quantifying amino acid substitutions on protein structures (e.g. SIFT)
- changes in RNA structures
- changes in splice sites can alter gene structures
- changes in promoter and enhancer sites
- changes in microRNA target sites can alter post-transcriptional regulation

Integration into a single pipeline `SNPpredict` ongoing

SNPs and Splicing

- **Idea:** Use experimentally detected splice sites from RNAseq data to identify potentially relevant splice junctions
- Consider all SNPs falling in a ±3 window of the splice sites of these splits
- Score sequence of acceptor and donor splice sites using the MaxEntScan method **before and after** insertion of respective SNP
- High changes of the MaxEnt score (>7.7) indicate putative loss or gain of the splice junction

Extensive processing and comparative analysis of RNAseq data:

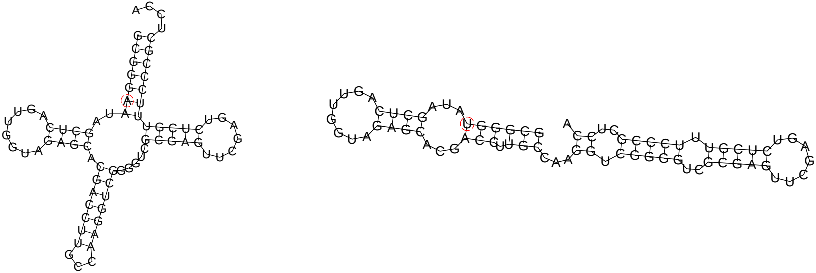
1. Collection of appropriate data sets from the “Avian RNAseq consortium” amd from diverse mammalian sources
2. Mapping, quality control and merging of RNAseq data
3. Use information from split reads to identify splice junctions
    - $\approx$ 320k detected splits after stringend filtering

SNPs in MicroRNA targets

- Collect human microRNA target sites from starBase `http://starbase.sysu.edu.cn/`
- Transfer to cattle coordinates using UCSC liftOver
- Overlap SNPs with potential target sites

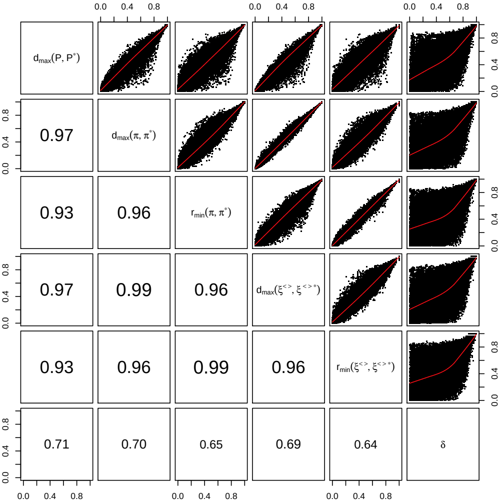
Putative promoter/enhance sites: consider overlap with high local conservation scores in unannotated regions

SNPs and RNA Structure



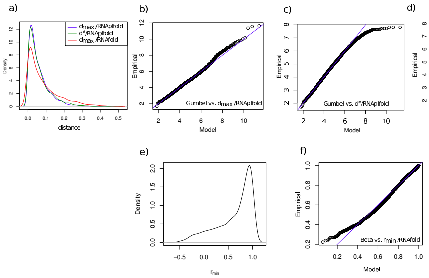
- Effect of point mutations depends strongly on position and substitution
- Quantification of structural difference
  - fraction of different base pairs
  - distance measures between base pairing matrices
  - distances between vectors of probabilities of pairing for individual bases
  - local, regional, and global measures

RNA Structure Distances

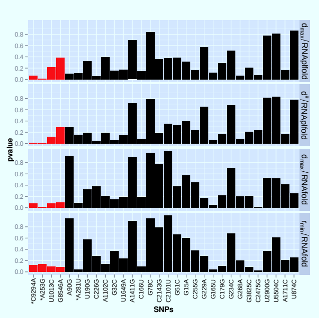


Detect Large Effects

`RNAsnp` computes empirical *p*-values for observed structure changes.



p-values obtained from extensive precomputed tables



only very few SNPs with known structural effects for validation

available as a stand-alone webservice
`http://rth.dk/resources/rnasnp/`

SNP prioritization

Aim: Prioritize SNPs from whole-genome resequencing in the QTLRs

- **Cattle**
  1. coding regions
  2. RNAsnp
     - Selection of RNA structural effects with $p_{d_{max}} < 0.01$
  3. microRNA target sites
- **Chicken**
  1. Variant annotation with ANNOVAR and ENSEMBL (release 71)
  2. Overlap with conserved elements from Phastcons 7way MCE
  3. Sorting Intolerant from Tolerant (SIFT) obtained using the Ensembl Variant Effect Predictor (VEP)
  4. RNAsnp
     - Selection of RNA structural effects with $p_{d_{max}} < 0.008$ and $p_{R_{min}} < 0.1$
  5. Splice site effects

Summary of SNP effect predictions (c

	Finnish Ayrshire	**Brown Swiss**
Total SNPs in QTLRs (filtered)	240,051	166,933
Coding regions (by UMIL, MTT)		
Total SNPs in coding regions	1834	1254
Stop-gain mutations	2	3
Splicing variants	5	5
Nonsynonymous variants	454	316
SIFT deleterious	93	62
Non-coding regions		
RNAsnp	936	863
microRNA target sites	27	22

Summary of SNP effect predictions (c

500 SNPs (from 443,802) were to be selected for further validation.

	Selected SNPs
Coding regions	
Stop-gain mutations	20
Splicing variants	77
SIFT deleterious and consverved	251[1])
Non-coding regions	
RNAsnp effect	149
mature microRNA	4

[1] 251 most conserved (PhastCons score) SNPs were chosen from the 1420 SNPs predicted to be deleterious by the SIFT method.

Acknowledgments

within QUANTOMICS

- ULEI: **Mario Fasold**, Anne Nitsche, Gero Doose, **Hakim Tafer**
- Roslin: Jacqueline Smith, Almas Gheyas, Dave Burt
- MTT: Johanna Vilkki
- UMIL: Marlies Dolezal
- Toine Roozen

outside QUANTOMICS

- Vienna: Ivo Hofacker
- RTH Copenhagen: Radhakrishnan Sabarinathan, Stefan E. Seemann, Jan Gordkin