# Finding mutations in gene regulatory DNA using funMotifs

Supervisors:
Professors Jan Komorowski, ICM <jan.komorowski@icm.uu.se>
and
Claes Wadelius, IGP <claes.wadelius@igp.uu.se>

**Background:**
Identification of transcription factor (TF) motifs with the potential regulatory impact on gene expression is problematic due to a lack of functional characterization of TF motifs.
We created a funMotifs framework[1] (https://github.com/komorowskilab/funMotifs, http://bioinf.icm.uu.se:3838/funmotifs/) that enables to identify and analyze significant regulatory TF motifs in noncoding regions of the human genome. The funMotifs framework determines regulatory mutations that are tissue-specific based on the number of annotation tracks. The framework incorporates data from large-scale genomics platforms including ENCODE[2,3], RoadMap Epigenomics[4], and FANTOM[5]. Recently, funMotifs was used to determine regulatory mutations and significantly mutated regulatory elements in various types of cancer based on the Pan-Cancer Analysis of Whole Genomes consortium data[6]. The study contains data from over 2,500 cancer genomes samples in 37 types of cancer. By incorporating the funMotifs framework, we were able to identify 5,749 mutated regulatory elements containing 11,962 candidate regulatory mutations. Additionally, we identified a number of genes nearby the mutated regulatory elements that were significantly dysregulated in the mutated samples. Furthermore, enrichment of cancer-related pathways was observed for the genes associated with the mutated regulatory elements. An interesting aspect of mutations that disturb TF motifs is the creation of *de novo* motifs. It has been shown that the presence of mutation within a *TERT* promoter, on some specific positions, may lead to the creation of a *de novo* ETS motif[7,8]. The current design of funMotifs enables to discover only pre-annotated TF motifs. The creation of a *de novo* ETS motif was linked with up-regulating *TERT* expression[9]. Incidentally, the *TERT* promoter was reported as the most recurrent element in the Pan-Cancer data. We collaborate with Prof Tobias Sjöblom who has recently sequenced colon cancer and constitutive genome in 1 000 cases of this cancer. The project involves finding the regulatory mutations in large and unique cohort.

**Project aims:**
The project aims to finalize an update of funMotifs to incorporate the latest genome annotation based on close to 1 million regulatory elements. Another aim is to develop a module for the identification of *de novo* TF motifs and merge the module into the funMotifs framework. The module can use an external tool designed for *de novo* motif identification, for instance, TrawlerWeb[10], MMARGE[11] or MAGGIE[12], or other tools suggested by the student.

**Tasks:**
- Optimize the last steps in funMotifs based on current genome annotation;
- Run an application to show the module performance;
- Use the pipeline to analyze mutations in a unique cohort of 1 000 colon cancers;
- Visualization of the results;
- Interpretation of the results.

**Requirements:**

- Basic knowledge of Unix systems and running bash scripts using Slurm;
- The framework is implemented in Python and uses the PostgreSQL system to construct a database. Consequently, experience in programming using Python is required. The project requires modification of the existing code and implementation of novel modules;
- Experience in working with GitHub;
- Statistical validation and visualization can be done using Python or R.

**References**


1. Umer, H. M. *et al.* funMotifs: Tissue-specific transcription factor motifs. *bioRxiv* 683722 (2019) doi:10.1101/683722.
2. Davis, C. A. *et al.* The Encyclopedia of DNA elements (ENCODE): data portal update. *Nucleic Acids Research* **46**, D794–D801 (2018).
3. An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**, 57–74 (2012).
4. Kundaje, A. *et al.* Integrative analysis of 111 reference human epigenomes. *Nature 2015 518:7539* **518**, 317–330 (2015).
5. Lizio, M. *et al.* Gateways to the FANTOM5 promoter level mammalian expression atlas. *Genome Biology 2015 16:1* **16**, 1–14 (2015).
6. HM, U., K, S., J, K. & C, W. Functional annotation of noncoding mutations in cancer. *Life science alliance* **4**, (2021).
7. Panebianco, F., Nikitski, A. v., Nikiforova, M. N. & Nikiforov, Y. E. Spectrum of TERT promoter mutations and mechanisms of activation in thyroid cancer. *Cancer Medicine* **8**, 5831 (2019).
8. Yuan, X., Larsson, C. & Xu, D. Mechanisms underlying the activation of TERT transcription and telomerase activity in human cancer: old actors and new players. *Oncogene 2019 38:34* **38**, 6172–6183 (2019).
9. JI, P. *et al.* Role of early growth response 1 in arteriogenesis: impact on vascular cell proliferation and leukocyte recruitment in vivo. *Thrombosis and haemostasis* **107**, 562–574 (2012).
10. Dang, L. T. *et al.* TrawlerWeb: an online de novo motif discovery tool for next-generation sequencing datasets. *BMC Genomics 2018 19:1* **19**, 1–9 (2018).
11. Link, V. M., Romanoski, C. E., Metzler, D. & Glass, C. K. MMARGE: Motif Mutation Analysis for Regulatory Genomic Elements. *Nucleic Acids Research* **46**, 7006 (2018).
12. Shen, Z., Hoeksema, M. A., Ouyang, Z., Benner, C. & Glass, C. K. MAGGIE: leveraging genetic variation to identify DNA sequence motifs mediating transcription factor binding and function. *Bioinformatics* **36**, i84–i92 (2020).