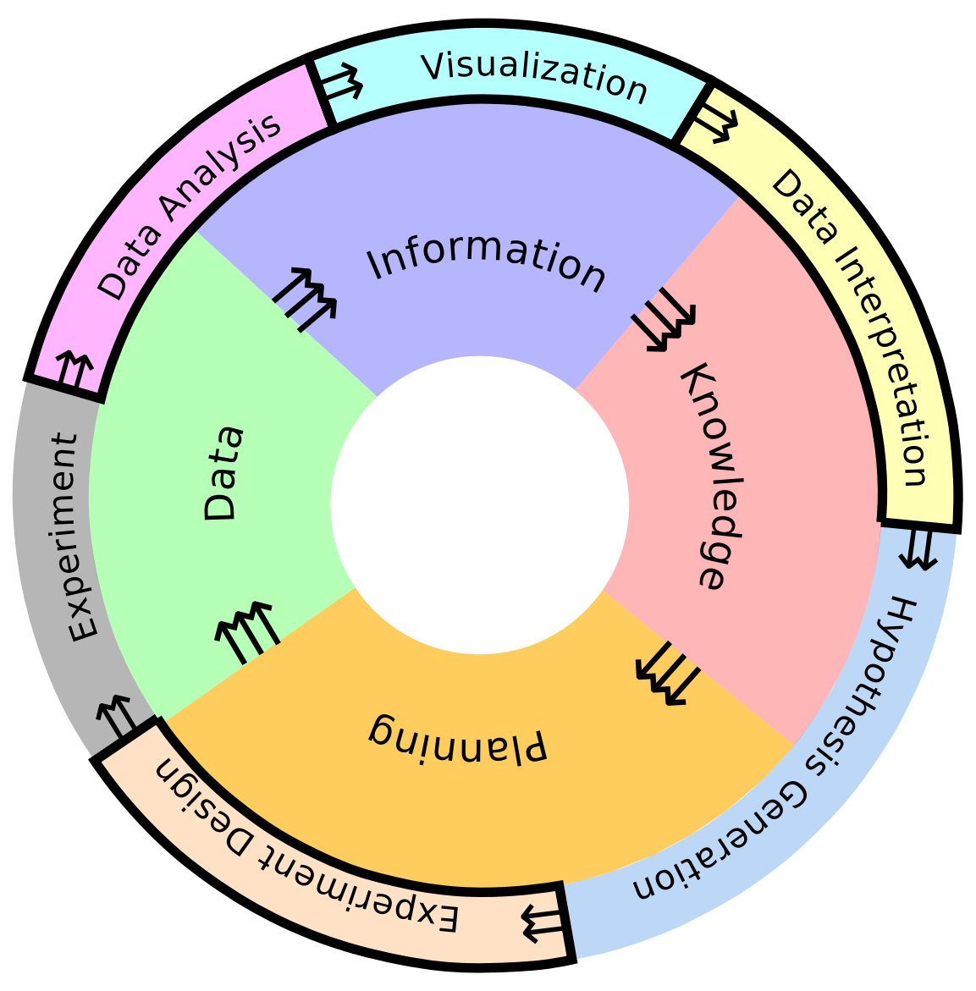
F43-01: Beyond the surface: RNA Regulation Bioinformatics

Michael T. Wolfinger[1,2,3], Fabian Amman[1,4], Arndt von Haeseler[3,5], Ivo L. Hofacker[1,5]

1 Institute for Theoretical Chemistry, University of Vienna, Austria
2 Department of Biochemistry and Molecular Cell Biology, Max F. Perutz Laboratories, Vienna, Austria
3 Center for Integrative Bioinformatics Vienna, Max F. Perutz Laboratories, Vienna, Austria
4 Department of Computer Science and the Interdisciplinary Center for Bioinformatics, University of Leipzig, Germany
5 Bioinformatics and Computational Biology, Faculty of Computer Science, University of Vienna, Austria

The coordination project within the special research program *''RNA regulation of the transcriptome''* (SFB RNA-REG) is responsible for ● **bioinformatics support** and analysis of next-generation sequencing (NGS) data, providing and maintaining a central ● **data repository and hardware infrastructure** as well as ● **capacity building** in applied computational biology aimed at researchers in participating groups. We use state of the art bioinformatics approaches and develop novel methods to provide solutions for a wide range of ● **scientific questions** in computational molecular biology.

In **cooperation** with SFB groups we analyzed more than 200 samples from 26 NGS projects in various organisms, including *E. coli*, *P. aeruginosa*, *A. thaliana*, *D. melangoaster* and *H. sapiens*, resulting in more than 25 TB of data. We implemented more than 15,000 lines of source code to build a versatile pipeline that incorporates state of the art NGS analysis tools. **Custom processing** was required for more than 140 samples, depending on and related to scientific question and experimental setup. Besides NGS support, we addressed 10 complementary bioinformatics projects as follow-up or side tasks of ongoing research.



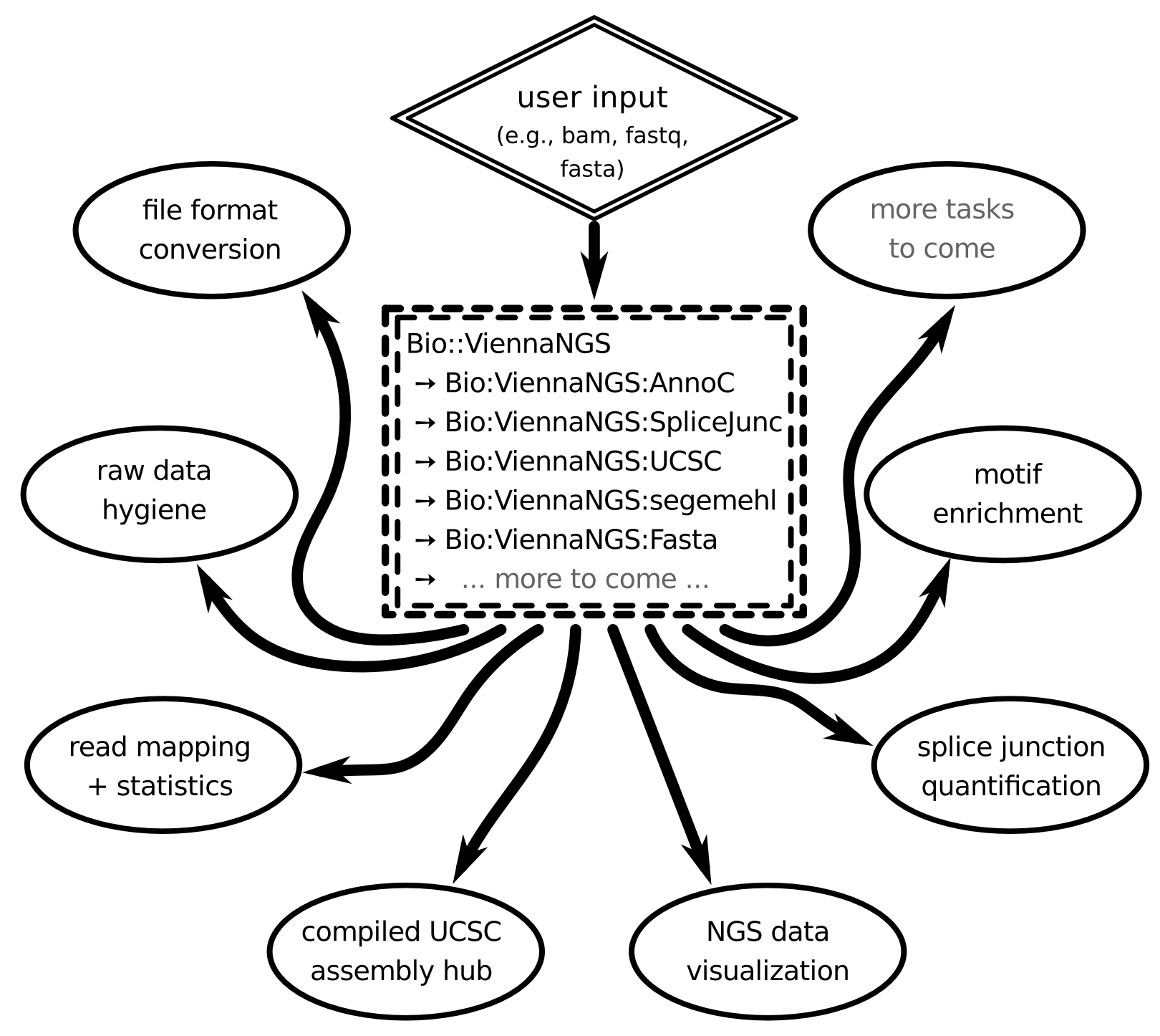
We serve as a **central hub for scientific computing** within the SFB consortium and contribute to every stage, from experimental design, to data analysis, to interpretation (framed fields in figure).

Capacity building in bioinformatics methods is an asset for the coordination project. We offer workshops to familiarize participants with **NGS data analysis** and provide hands-on computational experience using **best-practice approaches** for data quality assessment, sequence alignment, RNA-seq data processing and statistical analysis incorporating the UNIX command line.

In this line, we have organized a 5-day **workshop** for 20 PhD students on "*Statistical analysis of NGS and transcriptomics data with R*" , and are currently preparing a more technical workshop on "*Bioinformatics methods in NGS analysis*" for bench scientists with little bioinformatics experience who are currently using NGS technologies in their research.

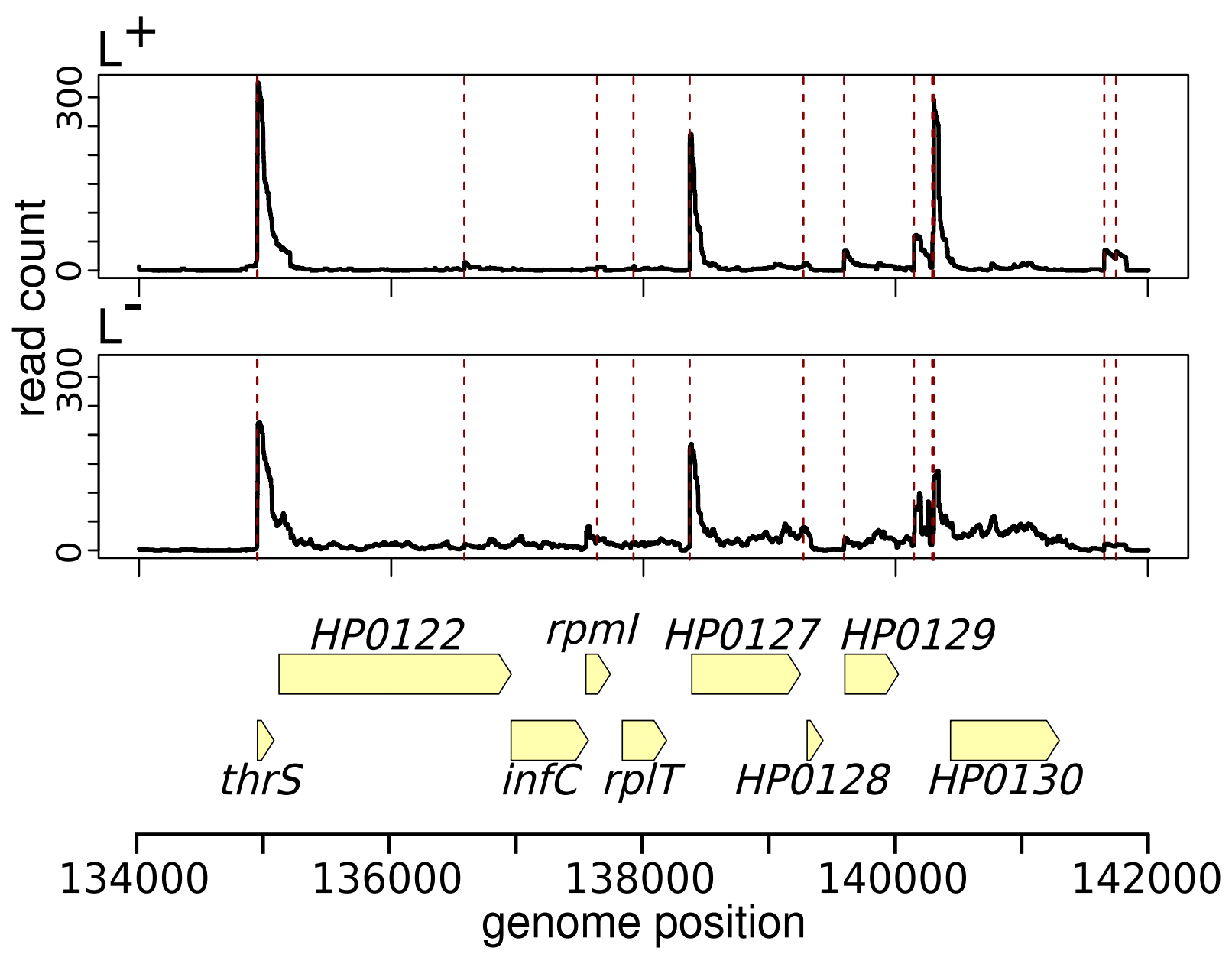
Reusable software components form the foundation of modern software architecture. Present NGS technologies allow researchers to address specific questions by means of highly customized experimental setups, thus demanding non-standard analysis tools.

We started to develop ***ViennaNGS*** [3], a collection of Perl modules that can be used to build efficient NGS analysis pipelines. *ViennaNGS* provides functionality for many standard and non-standard bioinformatics tasks, including but not limited to quality assessment, data visualization, motif discovery and splice junction characterization. *ViennaNGS* makes available **modular and reusable code for state-of-the-art NGS** processing in a popular scripting language, whose components can readily be included in custom scripts. *ViennaNGS* is actively being developed and shared with the scientific community through CPAN [4].

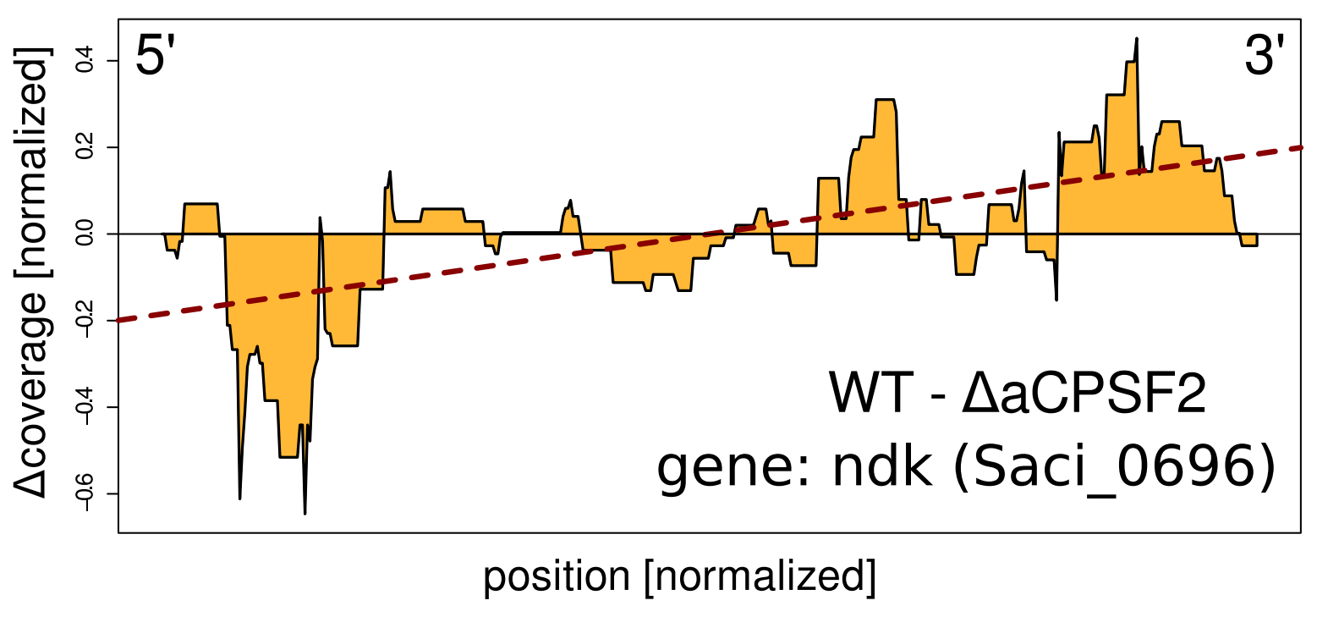


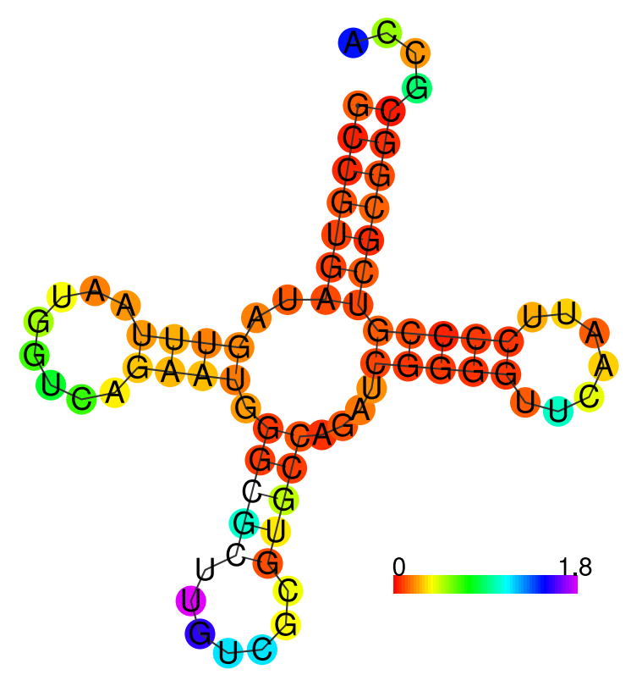
A **central data repository** has been established for storing NGS data produced by the SFB, thereby making available more than 25 TB of raw, intermediate and processed data to all participating groups. We maintain powerful virtualization infrastructure, offer **high-performance computing resources** and run dedicated Web and database servers, hosting custom instances of the **UCSC and GBrowse genome browsers** to allow for data integrity, rapid data accessibility and customizability.

The development of **statistical analysis** still lags behind the rapid development of novel RNA-seq methods. One example is differential RNA-seq (dRNA-seq), which aims at **annotating primary transcription start sites** (TSS). Here, the exonuclease TEX is used to selectively degrade non-primary 5' RNA ends in one library (L$^+$), whereas a control library is not TEX treated (L$^-$). Since the enzymatic digestion is not perfectly specific and special positions are also enriched in the L$^-$ library, the remaining challenge is to call position which are significantly enriched in the L$^+$ relative to L$^-$. Therefore, we developed ***TSSAR*** [1], a software tool which models read start counts in each library by a zero inflated Poisson distribution (thus accounting for over-dispersion and un-transcribed regions) and calculates the likelihood to observe a certain difference between L$^+$ and L$^-$ by a Skellam distribution, respecting local transcription activity.

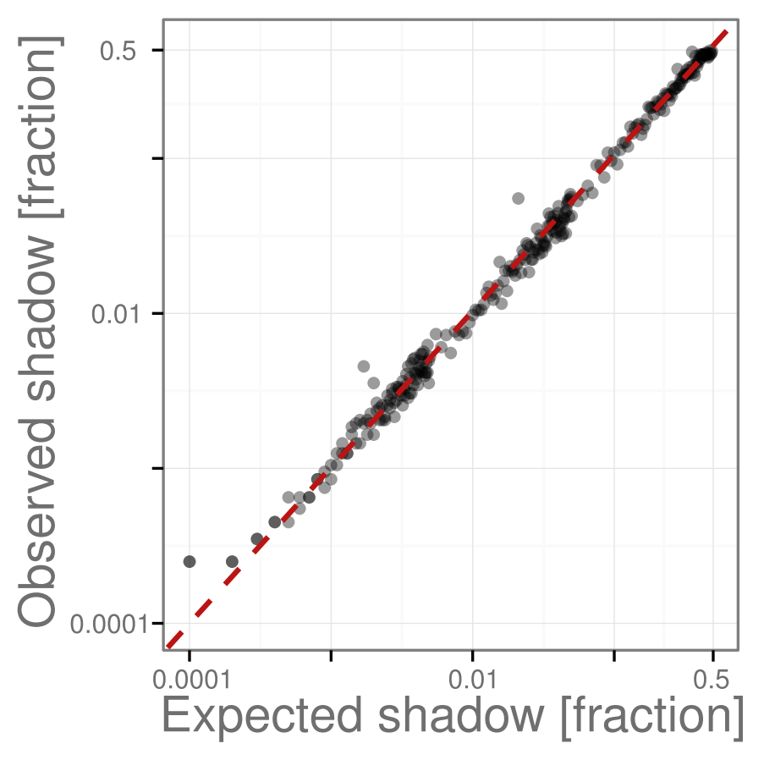


Custom analysis methods are often required to make use of the full **information wealth of NGS data**. As such, we analyzed, in cooperation with the Bläsi and Hofacker groups, RNA-seq data from *S. acidocaldarius* wild type and mutants lacking the 5' to 3' exoribonuclease aCPSF2, with the aim of discovering potential **degradation targets**. To this end, we integrated information from **differential gene expression analysis and information on differential read coverage** across single transcripts between wild type and mutant. The latter was achieved by normalizing read coverage per position with respect to gene length and expression abundance. The difference between normalized wild type and mutant signal was approximated with a regression line and the distribution of the regression line's slope was tested for outliers. This analysis resulted in 27 potential degradation targets, two of which were successfully tested by Northern blot [2].

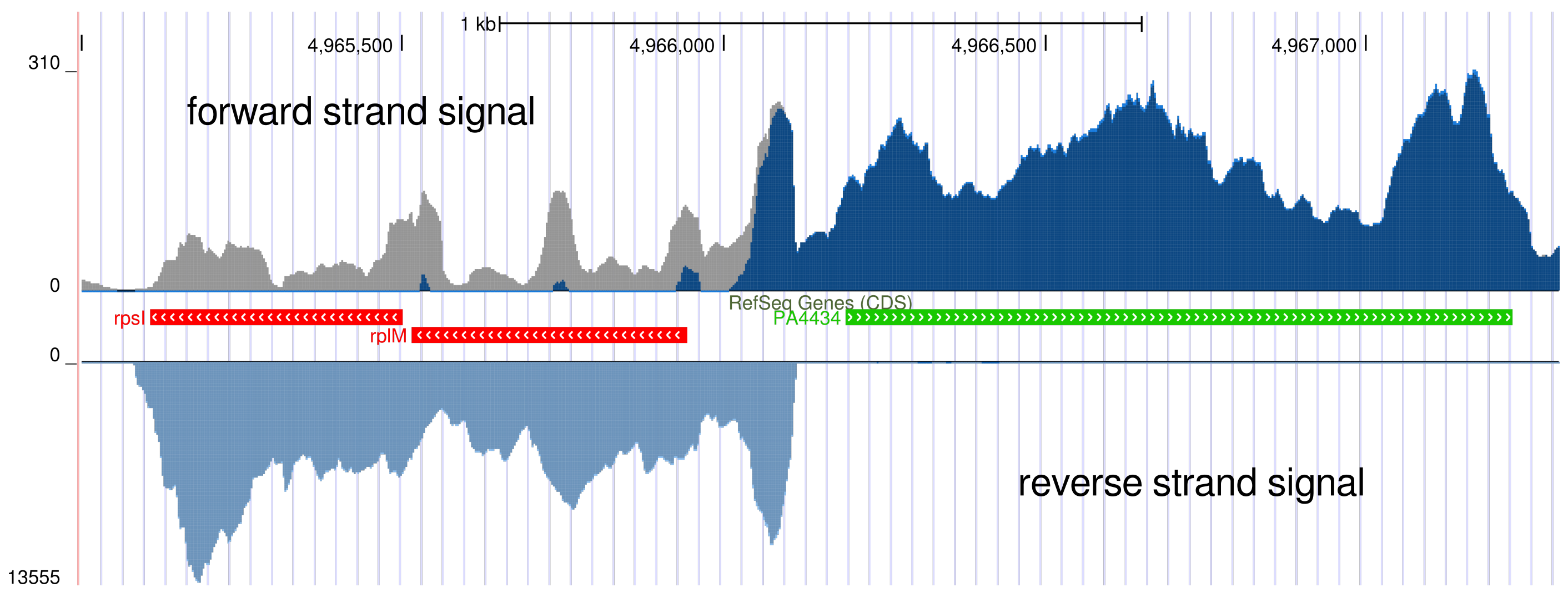




Chemical probing experiments allow for nucleotide resolution assessment of RNA structure. We have developed a method for **improving RNA secondary structure prediction by integrating chemical probing data** through soft constraints, thereby modifying the underlying thermodynamics-based energy model by adding pseudo energies and guiding the folding algorithm towards the experimentally supported structure [5]. In this context we have implemented three published approaches for incorporating experimentally determined **SHAPE reactivities** into the folding algorithms of the ***ViennaRNA Package*** [6]. Benchmarking against a set of RNAs with known reference structure has shown that prediction quality can be substantially improved by incorporating probing data. The picture shows a tRNA structure which has been folded with additional probing information, and whose SHAPE reactivity is encoded in colors from red to violet.



Strand-specificity imposes a problem in current NGS protocols, often resulting in 0.5-10% of erroneously mapped reads. This might feign anti-sense transcripts in the course of *de novo* transcript annotation. To overcome this issue we developed a simple yet powerful approach to detect the fraction of **anti-sense shadow in NGS data and correct the corresponding coverage profiles**. We sampled strand specific and non strand specific sequencing libraries from *S. cerevisiae* in a controlled manner to confirm reliability and could successfully determine the expected shadow (see scatter plot l.h.s.). The grey coverage profile in the genome browser image (r.h.s) corresponds to unmodified RNA-seq signal for an exemplary region in *P. aeruginosa*, whereas the blue curve results from correction with our method.



References:
[1] Amman, F., Wolfinger, M.T., Lorenz, R., Hofacker, I.L., Stadler, P.F., and Findeiß, S. (2014). **TSSAR: TSS annotation regime for dRNA-seq data.** BMC Bioinformatics, 15(1), 89.
[2] Märtens, B., Amman, F., Manoharadas, S., Zeichen, L., Orell, A., Albers, S.V., Hofacker, I.L., and Bläsi, U. (2013). **Alterations of the Transcriptome of Sulfolobus acidocaldarius by Exoribonuclease aCPSF2.** PloS one, 8(10), e76569.
[3] Wolfinger, M.T., Eggenhofer, F. and Fallmann, J., and Amman,F. (2014). **ViennaNGS: A toolbox for next-generation sequencing analysis.** *In preparation.*
[4] ViennaNGS at the Comprehensive Perl Archive Network: **http://search.cpan.org/~mtw/**
[5] Luntzer, D., Lorenz, R., Hofacker, I.L., Stadler, P.F., and Wolfinger M.T. (2014). **SHAPE directed RNA folding.** *In preparation.*
[6] Lorenz, R., Bernhart, S.H., zu Siederdissen, C.H., Tafer, H., Flamm, C., Stadler, P.F., and Hofacker, I.L. (2011). **ViennaRNA Package 2.0.** Algorithms for Molecular Biology, 6(1), 26.