Deep sequencing of *in vitro* nucleic acid selection experiments: High-throughput functional characterization and rare variant discovery

Experiments that isolate new functional nucleic acid molecules from random sequence pools tend to recover the smallest and simplest motifs with the required activity. Thus, the 14- and 16-nucleotide motifs of the 8-17 and 10-23 deoxyribozyme families dominate DNA pools selected for RNA-cleaving activity and hammerhead ribozymes dominate RNA pools selected for self-cleavage. These common motifs may obscure rarer and more complex motifs with potentially higher activities that exist in the same random library, and this potential shortcoming of *in vitro* selection has been termed a "Tragedy of the Commons". Indeed, there is evidence that within families of GTP aptamers only longer motifs with higher information content can access higher binding affinities. This trend is taken to an extreme when comparing the simple structure and relatively weak affinity of the best FMN-binding aptamer isolated by *in vitro* selection to the ornate FMN aptamer with exquisite specificity that controls gene expression in bacteria.

Next-generation sequencing platforms such as the Roche GS-FLX and the Illumina Genome Analyzer are able to sequence billions of bases in a single run with individual read lengths long enough to span typical *in vitro* selection randomized regions (50-100 nt). We will use deep sequencing to search for extremely rare functional variants in random nucleic acid pools and to determine the activities of many sequences in parallel by following their frequencies over time. These tasks require the development of new computational tools capable of classifying millions of reads into families of functional nucleic acid folds while taking into account sequencing errors and PCR amplification biases. We will use a well-characterized selection procedure for RNA-cleaving deoxyribozymes in proof-of-concept experiments. Unlike selections for functional RNAs, this design does not require a reverse transcription step, thus simplifying the generation of DNA samples for sequencing and eliminating an additional source of error. It also makes it possible to utilize a non-radioactive design for convenient kinetic characterization by incorporating fluorescent molecules into the oligonucleotides used to amplify the pool.

Specific Aim 1. Develop computational tools to classify functional nucleic acids and to estimate the relative activities of individual sequences in a random pool from deep sequencing data. Test on model datasets constructed from reported aptamer sequences.

Specific Aim 2. Perform *in vitro* selection monitored by deep sequencing to search for new classes of RNA-cleaving deoxyribozymes with more ornate structures in completely random and structurally constrained sequence libraries.

Specific Aim 3. Optimize the function and characterize the information content of diverse families with re-selection experiments that explore adjacent to an archetypal sequence. Characterize by structure probing and determining the pH dependence of cleavage rates.

The PI has relevant experience with *in vitro* selection, comparative nucleic acid secondary structure prediction, and using next-generation sequencing data to estimate the frequencies of mutations in mixed bacterial populations from an evolution experiment. Deoxyribozymes, ribozymes, and aptamers are finding diverse applications in gene therapy and biotechnology, but many of the known folds are not able to function adequately, especially under *in vivo* conditions. In addition to gaining a richer understanding of how functional nucleic acids are distributed in sequence space, it is possible that this approach will be generally useful for discovering new folds with useful functional parameters comparable to biological molecules. In addition, the basic paradigm of creating a mixture of sequences, subjecting it to a selection step for a desired function, and examining how the distribution of recovered sequences differs from the input can be used to probe many other functional biological sequences, such as promoters, DNA-binding sites, and even entire genes, as long as an *in vitro* or *in vivo* selection exists.