Reconstructions of intron evolution in the past several years have reached a consensus, viewing spliceosomal introns as neutral or slightly deleterious elements that rapidly spread throughout eukaryotic genomes during the very early days of eukaryogenesis. Consequently, most eukaryotic lineages harbored numerous introns. These introns, being non-coding segments at a close proximity to the coding exons, form a major evolutionary playground. Intronic mutations can accumulate at almost neutral rate, and just by chance some would be advantageous and would acquire function. Today, after way more than a billion years of eukaryotic evolution, many introns fulfill important cellular functions and are essential to the organism (we published a review about it, see [Igor B. Rogozin, Liran Carmel, Miklos Csuros and Eugene V. Koonin, Origin and evolution of spliceosomal introns, *Biology Direct* **7** (2012) 11]). However, since intron roles are regulatory and diverse, and sometimes even independent on their sequence, there is no good way to tell which intron is functional.

The fundamental hypothesis underlying this project is that functionally important introns should be characterized by distinct evolutionary trajectories. For example, we them to display decreases rates of loss, to be present in more closely related species, etc. Evolution of introns can be traced back in time by analyzing the conservation of their position with respect to the exonic mRNA sequence, or the *intron positional conservation*. The project is designed to prove the link between intron positional conservation and intron function, and to identify the evolutionary trends that are unique to functional introns. Specifically, we have suggested three specific aims.

Aim 1: Building a gene architecture database, that will keep architectonic properties of genes in many eukaryotes.

Aim 2: Testing the hypothesis that intron positional conservation is indicative of functionality.

Aim 3: Characterizing the positional distribution of introns with high level of positional conservation.

Our work during the project duration (2009-2013) culminated, as we hoped, in the first characterization of functional introns, and a classifier that reliably discriminates functional introns from non-functional ones [Michal Chorev and Liran Carmel, Computational identification of functional introns: high positional conservation of introns that harbor RNA genes, *Nucleic Acids Research* **41** (2013) 5604-5613]. This is the first work to suggest a way to predict whether a particular intron is functional or not, and was therefore selected by the Nucleic Acids Research editorial board as a *featured article* (to 5% of articles in the journal). See an example of several properties that characterize functional introns in Figure 1 below.

In order to get to this achievement, we built a gene architecture database (Aim 1), which contains information on more than 70 eukaryotic species, along with information on the known isoforms of each gene. It also contains assignment of transcripts from various species into groups of orthologs, which is crucial for subsequent comparative analysis. Moreover, we have completed the development of the algorithm, EREM, that allows evolutionary reconstruction of introns [Liran Carmel, Yuri I. Wolf, Igor B. Rogozin and Eugene V. Koonin, EREM: Parameter estimation and ancestral reconstruction by expectation-maximization algorithm for a

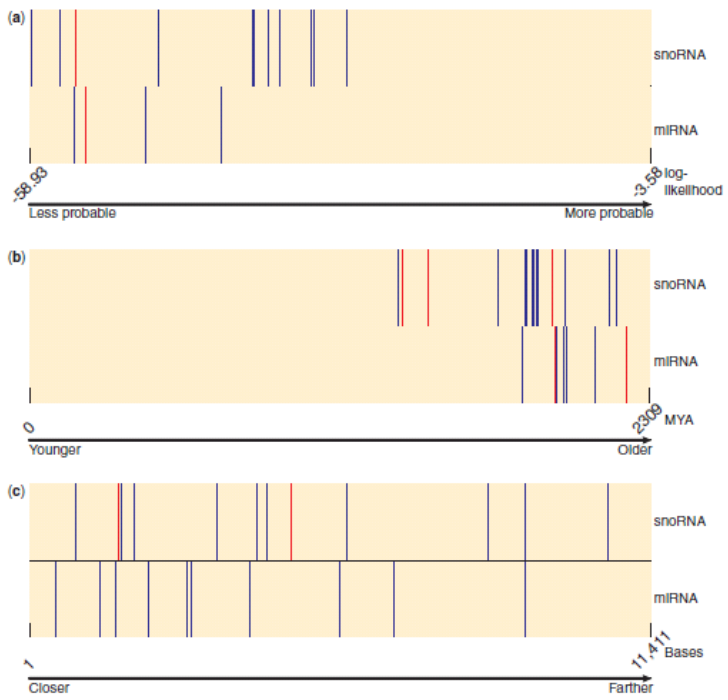

Figure 1: Functional introns are indicated by red or blue bars (indicating different types of functions), non-functional introns are depicted by beige bars (background). The introns are ranked by (a) their log-likelihood along the phylogenetic tree (using EREM); (b) their antiquity; and (c) their distance from the coding sequence start.

probabilistic model of genomic binary characters evolution, *Advances in Bioinformatics* **2010** (2010) Article ID 167408]. EREM is designed to be able to directly read output produced by our database, and it uses a comprehensive probabilistic model to describe gain and loss of introns, in a way that depends on both the specific gene and the specific branch in the tree. Combining EREM with our gene architecture database, we were able to ask more questions about intron properties. First, we studied in depth possible mechanisms for intron gain and loss, in order to understand the links between intron gain and loss processes [Noa E. Cohen, Roy Shen and Liran Carmel, The role of reverse-transcriptase in intron gain and loss mechanisms, *Molecular Biology and Evolution* **29** (2012) 179-186.]. In addition, we are using whole-exome sequencing in order to find the effects of splicing disruptions on human diseases. To this end, we are developing an algorithm to predict the effect of any mutation on the normal splicing pattern of nearby splicing junction. We are working on many diseases, and so far published one paper [David Zangen, Yotam Kaufman, Sharon Zeligson, Shira Perlberg, Hila Fridman, Moein Kanaan, Maha Abdulhadi-Atwan, Abdulsalam Abu Libdeh, Ayal Gussow, Irit Kisslov, Liran Carmel, Paul Renbaum and Ephrat Levy-Lahad, XX Ovarian Dysgenesis Is Caused by a PSMC3IP/HOP2 Mutation that Abolishes Coactivation of Estrogen-Driven Transcription, *The American Journal of Human Genetics* **89** (2011) 572-579], albeit eventually this particular disease didn't involve splicing disruptions.

In summary, I believe that this project was very successful in that it proved convincingly the fundamental premise that underlies our research. It resulted in a total of five papers that incrementally strived at the main result – showing that we can discriminate between functional and non-functional introns based on the evolutionary properties of their positional conservation.