Leveraging EST data for marker and target discovery

Sebastian Hoersch[1,*], Keith Robison[2], Sudeshna Das[2], Dave Rose[2], Christophe Leroy[1], John Morris[2], Maggie Zhou[2], John Chan[2], Jeremy Ahouse[1], Tom Barnes[3], Barbara Bryant[1], and Chris Sander[1]

[1] Millennium Predictive Medicine, Inc., One Kendall Square, Bldg. 700, Cambridge, MA 02139, USA

[2] Millennium Pharmaceuticals, Inc., One Kendall Square, Bldg. 700, Cambridge, MA 02139, USA

[3] Millennium Biotherapeutics, Inc., 620 Memorial Drive, Cambridge, MA 02139, USA

[*] To whom correspondence should be addressed: hoersch@mpi.com

We report on a modular approach towards maximally utilizing EST data from both public and in-house cDNA libraries. Components for gene finding include commercially available clustering software, integrated with in-house software for sequence similarity searches and maximal functional annotation, as well as software modules custom-tailored to address specialized questions like tissue-specific gene-expression. To account for the need of library prioritization in the light of limited sequencing capacity, software tools were developed to simulate libraries and assess their complexity and novelty by various statistical approaches.