

Mathematical & Computational Biology Seminar

Organizer: Lior Pachter

Wednesday, 2:00–3:00pm, 939 Evans

Apr. 15 **Wei-Chun Kao**, University of California, Berkeley
BayesCall: A model-based basecalling algorithm for high-throughput short-read sequencing

Recent advances in high-throughput sequencing technology is allowing researchers to generate immense amounts of raw instrument data from ultra high-throughput platforms. Soon, obtaining whole-genome sequences will become a routine assay, but before this compelling vision can be realized, several immediate challenges must be overcome. In particular, we need to develop computational tools that can extract accurate sequence information from raw data. To address this important challenge, we introduce a novel basecalling algorithm for high-throughput short-read sequencing. Our objective is to model the sequencing process to the best of our knowledge, explicitly taking stochasticity into account. We show that our approach, which being founded on the tools of statistical learning, is flexible enough to incorporate various features of the sequencing process. We illustrate this point by incorporating time-dependent parameters into our model. We show that our new approach significantly reduces the error rate in basecalls, particularly in the later cycles of a sequencing run. In this talk, I will give a brief description of our model and illustrate interesting findings regarding Illumina Genome Analyzer I and II platforms. This is a joint work with Kristian Stevens (UC Davis) and Yun Song (UC Berkeley).