Bioinformatics Applications of the SSAHA2 Package.

Adam Spargo and Zemin Ning.
Wellcome Trust Sanger Institute, Hinxton, Cambridge, United Kingdom.
aws@sanger.ac.uk

We will give a brief overview of sequence alignment methods, then describe the SSAHA hashing algorithm in some detail and its implementation in the SSAHA2 package; driving the algorithm via user-options will be explained. Following the theory, we will give a demonstration of how to perform alignment tasks on your local machine and how to run jobs in parallel on a cluster. For the most part we will concentrate on bioinformatics applications, stepping through how to use the various adaptations of the package:

- ssaha2Server can be run on a remote machine using a prebuilt hashtable.
- TraceSearch is a distributed version of the SSAHA2 algorithm, an online service to search the ensembl trace archive at WTSI is demonstrated; applications range from sequence validation and finishing to adding depth for analysis of local features, even across species.
- cross_genome breaks pairs of genomes into chunks and uses SSAHA2 to align them, gathering the results to form a cross-genome alignment. Recent results on the comparison of malaria strains will be presented.
- ssahaEST aligns EST or CDNA data, handling splice-sites correctly.
- ssahaSNP calls SNPs and short in-dels by aligning reads against a reference sequence. Examples from the human genome and a comparison between mouse strains will be used.
- ssahaSV identifies long in-dels and structural variations by splitting reads against a reference sequence. Results using the HapMap and Celera reads against the human reference as well as chimp reads against human are shown.

Users are invited to present particular problems at the end of the demonstration or to send email in advance.