NCBI SRA Toolkit Technology for Next Generation Sequence Data

Stephen Sherry, Chunlin Xiao, Kenneth Durbrow, Michael Kimelman, Kurt Rodarmer, Martin Shumway, and Eugene Yaschenko

National Center for Biotechnology Information (NCBI), National Library of Medicine, National Institutes of Health (NIH), Bethesda MD, USA.

NIH has directed NCBI's Sequence Read Archive (SRA) to continue to serve as the agency's central repository for sequence data. Using fully indexed columnar database design, the SRA toolkit has reduced lossless compression of sequence, qualities, and alignment properties from 32 bits per base in 2008 to under 5 bits per base for 1000 genomes phase 1 data. The cSRA technology provides a single toolkit interface that supports efficient compression, slices of data on read, indexed retreival, serialization of data for pipeline processing, retention of optional BAM tags, lossy compression.and support for non-standard reference sequences. The SRA toolkit including cSRA is freely distributed under multiple platforms and in unrestricted source code form, see http://www.ncbi.nlm.nih.gov/Traces/sra/sra.cgi?view=software . The toolkit technology is engineered for indexing, slicing, streaming, and real-time compression and encryption of datasets for optimal retrieval by SRA users. The toolkit can also be deployed at local sites to realize the same services and storage efficiencies for a project's internal NGS data.