

Abstract 27

Developments in dbSNP for 2003

S.T. Sherry

National Center for Biotechnology Information (NCBI), National Library of Medicine, Bethesda, MD, U.S.A.

NCBI's dbSNP database of genome sequence variation offers data for population structure and haplotype analysis, association studies, and functional analysis. The database serves the community in dual roles; both as an author-driven archive and curated resource for whole genome annotation. The complete contents of dbSNP are available freely to the public.

dbSNP currently contains submissions for 11.45 million sequence variations observed in 19 species. Organized by class of sequence variation the database contains: 11.1M single nucleotide polymorphisms SNPs; 331K deletion/insertion polymorphisms (DIPs), 331 polymorphic retroposons and 5K short tandem repeats (STRs). The high levels of redundant submissions require active curation and clustering by the dbSNP staff. Identical, independent submissions are currently grouped into 7.04M RefSNP clusters. These clusters provide a stable identifier space, with accessions provided by dbSNP anchoring the higher dimensional data of linkage structure, haplotype diversity and ethnic differentiation to the reagents and final deliverables of the genome project. dbSNP currently catalogs extensive variation in humans and other model organism genomes: mosquito, mouse, chimpanzee, and rat. These resources can provide tremendous utility in theoretical, experimental and clinical contexts.

NIH is supporting this potential through a substantial investment in reference reagents such as clone repositories of mammalian genes and immortalized cell lines developed from large samples of contemporary world populations. When complementary reagents are unified by genotypes and systematic measures of linkage disequilibrium across the genome, a framework is established that connects theoretical, experimental and clinical research programs. Stable reference genome sequence is an organizing template for many annotation efforts. Polymorphism annotation currently includes large-scale polymorphism detection results; functional variants in coding regions; individual genotypes in out-bred populations and strain-specific haplotypes in model organisms. Population measures like genotype and allele frequencies are now being supplemented with local coverage-corrected measures of sequence diversity. dbSNP is structured to integrate these basic properties of variation and serve them to the research community through database queries, annotation, distribution of content and network interfaces.