STRAND LIFE SCIENCES WHITE PAPER

strandngs

Variant Allele List Subsetting Utility

Analyze | Visualize | Annotate | Discover

strand
New Generation Healthcare

The advent of NGS and advancements in SNP genotyping technologies has caused a surge in the number of SNPs in dbSNP database. The dbSNP 146 version has around 150 million SNPs and this number seems poised to increase (Table 1). One of the key steps in variant analysis is the annotation of identified SNPs using dbSNP. Strand NGS does this by scanning through the entire dbSNP database only once and annotating the variants detected in all the samples simultaneously. This process is quite efficient since the time taken is independent of the number of samples in the experiment and is dependent only on the number of SNP records in the database. A majority of current DNA-Seq studies involve targeted panels, thus negating the need to scan through all the SNPs scattered throughout the genome. Our new 'Create targeted VAL' feature is a Variant Allele List (VAL) subsetting utility available from Strand NGS v2.7 onwards. It is aimed at reducing SNP annotation time by first creating a subset of the dbSNP database which would contain only those SNPs located in the target regions of interest.

This simple and easy to use feature involves the following steps:

1 Run the dbSNP subsetting step on any dbSNP version to obtain the subset relevant to your targets (The 'Create targeted VAL' feature can be applied not just to dbSNP but any VAL list to create a subset specific to your target regions). Note: This step needs to be performed only once per targeted panel.

2 Use the created subset for every subsequent SNP detection.

dbSNP release of human	Release date	Number of RefSNP clusters (rs#'s)
142	14th Oct 2014	112,743,739
144	08th Jun 2015	149,735,377
146	24th Nov 2015	150,482,731

Table 1: Number of SNPs in each version of dbSNP release

To evaluate the efficacy of this feature, we performed SNP annotation with and without the dbSNP subset option. Table 2 briefly describes the datasets used in this study. Dataset-1 was sequenced on the TruSeq Exome Panel. It consists of two sequenced samples from the Lung Cancer Sequencing Project Exome available at ENA (study accession ID: ERP001575). Dataset-2 is the Horizon control sample, HD200, sequenced in-house using the TruSeq Amplicon - Cancer Panel on an Illumina MiSeq sequencer. All the computations in this study were carried out on a Linux machine with 4 cores and 8GB RAM.

Data	No of Samples	Read Length	Read Length
Dataset-1	2	100 bp, Paired end	259,798,204
Dataset-2	1	150 bp, Paired end	5,986,870

Table 2: Brief description of datasets used in this study

Using the VAL subsetting feature, we created a subset of dbSNP 146 for the two target panels (TruSeq Exome Panel and TruSeq Amplicon - Cancer Panel). The SNP subset creation time for each panel was approximately 50 minutes. This time is proportional to the size of dbSNP version and is independent of the target panel.

Following SNP detection, annotation of SNPs was carried out in both settings and the results are shown in Table 3. When compared to the older method of scanning the whole database, using the 'Create Targeted VAL' subsetting feature leads to a time reduction of 94.86% and 99.77% for Dataset-1 and Dataset-2 respectively. This reduction in the annotation time is constant irrespective of the number of samples processed in a batch and is dependent only on the size of the SNP subset database. In large-scale research studies or clinical settings where hundreds of samples are processed on a monthly basis, this new VAL subsetting feature in Strand NGS can easily save you several hours.

Sample used	Target Panel	No of Target regions	Size of Target panel file	No of SNPs in dbSNP 146 subset	%SNPs in dbSNP 146 subset	No of SNPs detected	SNP Annotations Time		% reduction time
							StrandNGS without dbSNP subset	StrandNGS with dbSNP subset	
Dataset-1	TruSeq Exome Panel	201,071	12 Mb	7,702,573	5.12	42,404	20 min 46 sec	1 min 4 sec	94.86
Dataset-2	TruSeq Amplicon-Cancer Panel	181	18 Kb	5,785	0.003	65	16 min 18 sec	2.24 sec	99.77

Table 2: Brief description of datasets used in this study

For more information or to avail free online training, please write to us at sales@strandngs.com

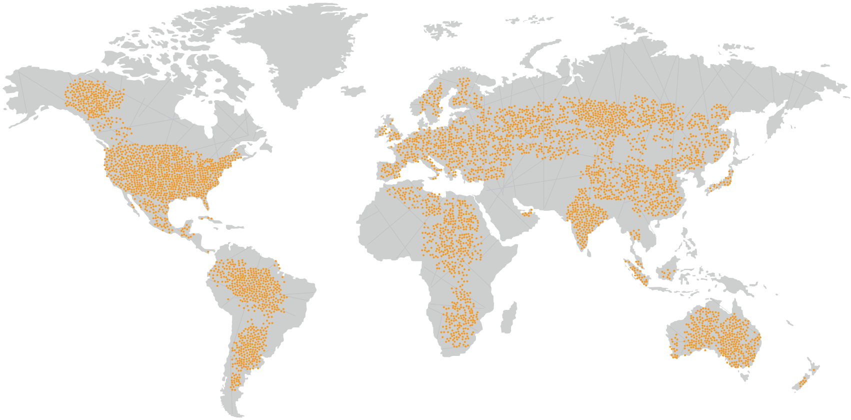
strand
New Generation Healthcare

About Strand

A History of Innovative Genomic Research

Strand Life Sciences is a global genomic profiling company and leader in precision medicine diagnostics, aimed at empowering cancer care and genetic testing for inherited diseases. Strand works with physicians and hospitals to enable faster clinical decision support for accurate molecular diagnosis, prognosis, therapy recommendations, and clinical trials. The Strand Center for Genomics & Personalized Medicine is India's 1st and only CAP & NABL accredited NGS laboratory.

www.strandls.com



A Trusted Partner to Companies Worldwide

For 15 years, our genomics products and solutions have facilitated the work of leading researchers and medical geneticists in over 2,000 laboratories and 100 hospitals around the world.

Strand Life Sciences Pvt. Ltd
5th Floor, Kirloskar Business Park, Bellary Road, Hebbal, Bangalore 560024
Phone:+91-80-40 (787263) Fax: +91-80-4078-7299