NGS Data Storage Requirements

There are two common modes of DNA sequencing: whole genome sequencing and exome sequencing. Exome sequencing methods sequence just the exonic regions which typically comprise 1-2% of the whole genome. Whole genome sequencing methods of course sequence the whole genome. Reads coming from the sequencer are then aligned to the reference genome and the resulting BAM file is imported into Strand NGS. For storage size computation, all data upstream of this BAM file can be treated as transient, so only storage for BAM files and subsequent analyses needs to be planned.

The size of a BAM file depends on coverage (the average number of times each base is read) and read length. A few examples are provided in Table 1 below. Please note that sizes in Strand NGS have an overhead. This arises from storage of extra information, which enables fast access and visualization later.

	Coverage	No. of Reads	Read Length	BAM File Size	Strand NGS Size
Whole Genome	37.7x	975,000,000	115	82 GB	104 GB
Whole Genome	38.4x	3,200,000,000	36	138 GB	193 GB
Exome	40x	110,000,000	75	5.7 GB	7.1 GB

Table 1: Overview of storage requirements depending on coverage, no. of reads and read length

Allowing for some extra analysis results storage and assuming whole genome samples are done at read lengths of 75 or above, the size of each whole genome sample can be rounded off to about 150 GB and the size of each exome sample to about 8 GB. Space for backups also needs to be taken into consideration. With these assumptions, the total storage requirement for a few scenarios is illustrated in Table 2 below.

Whole Genome Samples	Exome Samples	Space	Space including Backup
0	200	1.6 TB	3.2 TB
0	1000	8.0 TB	16 TB
100	0	15 TB	30 TB
1000	0	150 TB	300 TB
100	1000	23 TB	46 TB

Table 2: Storage requirements based on number of samples and allowing for backup

2 TB hard drives are available off the shelf; two of these should more than suffice for running 250 exome sequencing samples. Strand NGS can be configured to add storage incrementally, so you can start with a 2*2 TB hard disk and add further disks on demand if needed. If you

need to plan for more than 10 TB of storage we recommend a network storage solution as opposed to adding disks to a single machine.

Computation Speeds

Computation speeds for various tasks for Strand NGS v2.8.1 are given below. These are generated on a 16-core machine, but these analyses can even be run on a standard laptop with 4 GB of RAM at proportionately reduced speeds. A minimum of 8 GB of RAM is recommended for alignment tasks in case of large genomes.

Machine details 16 cores @ 2.7GHz, 32 GB RAM	
Sample details DNA reads of a human (NA12878) sample Size of the fastq.gz files: 92 GB; #Reads: 1.16 billion paired-end reads Read length: 150bp	
Task	Time Taken
Alignment of DNA reads	6 hr 26 min (~11.5 million reads/hour/core)
Import of the aligned reads (includes computation of QC statistics)	5 hr 59 min
Local realignment (includes recomputation of QC statistics)	9 hr 31 min
Base quality recalibration (includes recomputation of QC statistics)	8 hr 54 min
Read Filters (includes recomputation of QC statistics)	10 hr 41 min
SNP detection (includes annotating with dbSNP 146)	5 hr 47 min

Table 3: Computational times for specific tasks in Strand NGS

Additional Information

If you require more information please contact our **Support Team**. Go to **www.strand-ngs.com** to get a free evaluation license, giving you access to a fully functional version of Strand NGS for 20 days.