Genomic Information Database

Table taken from ISO/IEC JTC 1/SC 29/WG 11 N15092 (Geneva, CH - February 2015)

ID	Sequencing method	Size	File type	Coverage	Origin	Comments	Path (prepend /data/genome/ within the TNT cluster)
Homo Sapiens							
ERP001775	Illumina HiSeq	~2 TB	FastQ	200x	http://www.ebi.ac.uk/ena/data/view/ERP001775	This is the largest dataset. Can be transcoded from gzip to 7zip if needed.	human/illumina/ERP001775
ERP001960	Illumina HiSeq	~120 GB each genome	BAM	30x	http://www.ebi.ac.uk/ena/data/view/ERP001960	3 genomes selected: SAMEA1573614 SAMEA1573618 SAMEA1573617	human/illumina/ERP001960/tmp/NA12878_S1.sam human/illumina/ERP001960/tmp/NA12879_S1.sam human/illumina/ERP001960/tmp/NA12890_S1.sam
ERP002490	Illumina HiSeq	265 GB	BAM	30-40x	http://www.ebi.ac.uk/ena/data/view/ERP002490	The insert size (distance between the pair of reads) is much larger than usual (about 2K bases instead of 300).	human/illumina/ERP002490/tmp/NA12877_S1.sam human/illumina/ERP002490/tmp/NA12878_S1.sam human/illumina/ERP002490/tmp/NA12882_S1.sam
Low coverage ERR317482 WGS	Illumina HiSeq 2000	6.1 GB	BAM	1.9x	http://www.ebi.ac.uk/ena/data/view/ERR317482 ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/technical/reference/phase2_reference_assembly_sequence/hs37d5.fa.gz	Also used as a low coverage test in the Scramble paper.	human/illumina/ERR317482WGS/tmp/9827_2#49.sam
Low coverage NA21144.chrom11	Illumina HiSeq 2000	1 GB	BAM	7.5x	ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/data/NA21144/alignment/NA21144.chrom11.ILLUMINA.bwa.GIH.low_coverage.20130415.bam ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/technical/reference/phase2_reference_assembly_sequence/hs37d5.fa.gz	Used in Scramble paper. Processed through the GATK pipeline, which makes the auxiliary data bulkier. (Can be stripped off easily if desired.)	human/illumina/NA21144.chrom11/tmp/NA21144.chrom11.ILLUMINA.bwa.GIH.low_coverage.20130415.sam
PacBio_CHM1htert_54x	Pacific Biosciences SMRT	150 GB	FastQ	54x	http://datasets.pacb.com/2014/Human54x/fast.html	Available as zipped FastQ.	human/pacbio
IonTorrent	Ion Torrent	1.3 GB each file	BAM		http://www.ebi.ac.uk/ena/data/view/ERX276880 http://www.ebi.ac.uk/ena/data/view/ERX276881 http://www.ebi.ac.uk/ena/data/view/ERX276882	http://www.ebi.ac.uk/ena/data/warehouse/search?query=%22instrument_platform=%22ION_TORRENT%22%20%20AND%20submitted_format=%22BAM%22%22&domain=read	human/IonTorrent/tmp/sample-2-10_sorted.sam human/IonTorrent/tmp/sample-2-11_sorted.sam human/IonTorrent/tmp/sample-2-12_sorted.sam
RNAseq		16 GB	BAM		http://www.ebi.ac.uk/arrayexpress/files/E-MTAB-1728/K562_cytosol_LID8465_TopHat_v2.bam		human/RNASeq/tmp/K562_cytosol_LID8465_TopHat_v2.sam

Bacteria							
ERX593919 (E. Coli)	Oxford Nanopore	60 GB	Gzipped Fast5		http://www.ebi.ac.uk/ena/data/view/ERX593919	To be converted from Fast5.	--
ERX593921 (E. Coli)	Oxford Nanopore	46 GB	Gzipped Fast5		http://www.ebi.ac.uk/ena/data/view/ERX593921	To be converted from Fast5.	bacteria/oxfordnano
DH10B (E.Coli)	Illumina	1.3 GB	BAM		ftp://webdata:webdata@ussd-ftp.illumina.com/Data/SequencingRuns/DH10B/MiSeq_Ecoli_DH10B_110721_PF.bam https://raw.githubusercontent.com/allanroscoche/PathTree/master/data/DH10B_WithDup_FinalEdit_validated.fasta	Used in the Deez paper.	bacteria/DH10B/tmp/MiSeq_Ecoli_DH10B_110721_PF.sam
ERA269036 9799_7#3.bam (E.Coli)	Illumina	2.3 GB	BAM		ftp://ftp.sra.ebi.ac.uk/vol1/ERA269/ERA269036/bam/ http://www.ncbi.nlm.nih.gov/nuccore/NC_000913.2?report=fasta&format=text	Used in the Scramble paper.	bacteria/ERA269036/tmp/9799_7#3.sam
Metagenomic							
Human gut	Illumina Genome Analyzer II	10 GB	FastQ		http://www.ebi.ac.uk/ena/data/view/ERA000116	3 samples picked: SAMEA728920 SAMEA728635 SAMEA728854	metagenomic/humangut
Cancer cell lines							
Mutation/Variation Calling Benchmark 4 at CGHub		255 GB	BAM	60x	https://cghub.ucsc.edu/datasets/benchmark_download.html	TCGA BENCHMARK CELL LINE: HCC1143 NORMAL 60x f0eaa94b-f622-49b9-8eac-e4eac6762598	cancercells/f0eaa94b-f622-49b9-8eac-e4eac6762598/tmp/G15511.HCC1143_BL.1.sam
Mutation/Variation Calling Benchmark 4 at CGHub		305 GB	BAM	50x	https://cghub.ucsc.edu/datasets/benchmark_download.html	TCGA BENCHMARK CELL LINE: HCC1143 TUMOR 50x ad3d4757-f358-40a3-9d92-742463a95e88	cancercells/ad3d4757-f358-40a3-9d92-742463a95e88/tmp/G15511.HCC1143.1.sam
Mutation/Variation Calling Benchmark 4 at CGHub		130 GB	BAM		https://cghub.ucsc.edu/datasets/benchmark_download.html	UCSC ARTIFICIAL MIXED SAMPLE: 80% HCC1954BL 20% HCC1954 360b4736-6c5e-48df-af58-c1cf51609350	cancercells/360b4736-6c5e-48df-af58-c1cf51609350/tmp/HCC1954.mix1.n80t20.sam
Plants							
T. Cacao	Illumina	8.2 GB	FastQ	10x	http://www.ncbi.nlm.nih.gov/sra/SRX288435		--