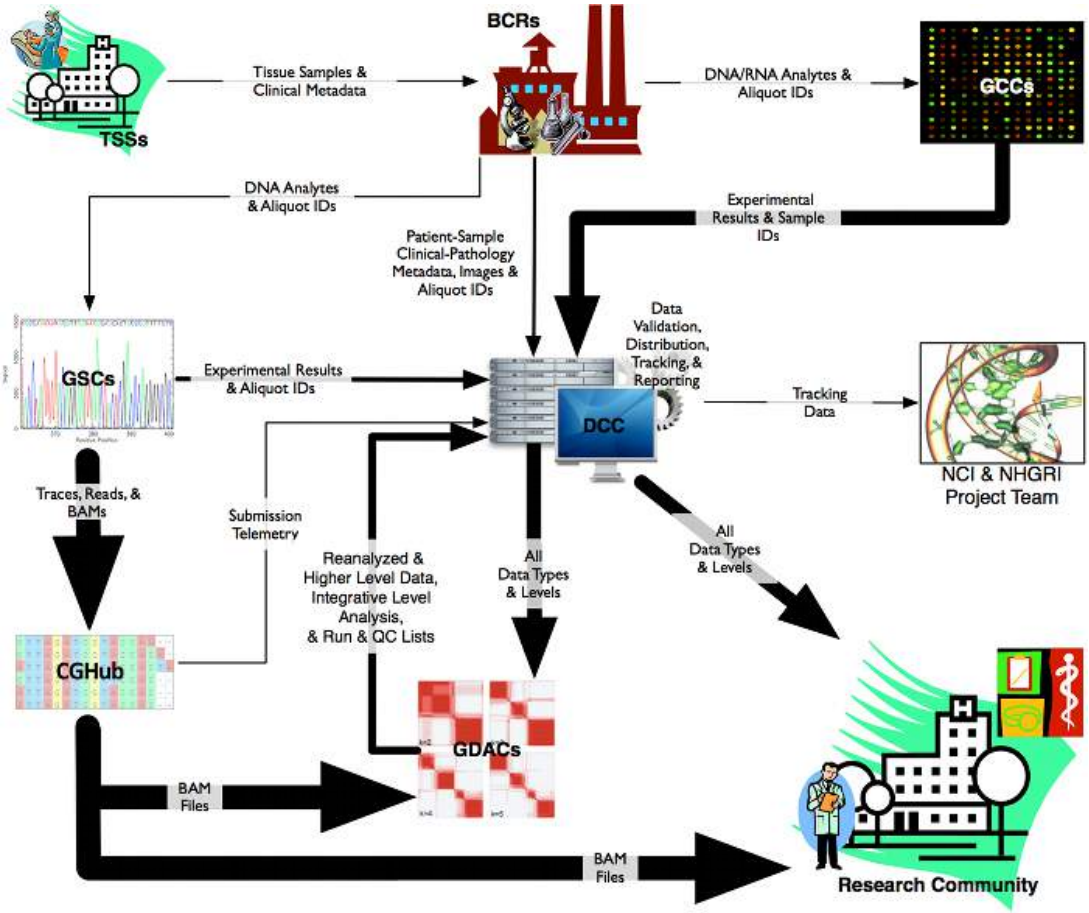
The Cancer Genome Atlas: Product Offering by Station X in GenePool

The Cancer Genome Atlas (TCGA) is a large, multi-center effort to elucidate the molecular basis of cancer through the application of genome analysis technologies, primarily large-scale genome sequencing. The TCGA project, led by a project team composed of individuals from the National Cancer Institute (NCI) and the National Human Genome Research Institute (NHGRI), began as a pilot in 2006 with three cancer types and was expanded in 2009 to more than 30 (see status table at end of document). To date TCGA has received more than $200 million in funding, has completed characterization of more than 11,000 cases (and almost 24,000 samples) and will have released an estimated 2.5 petabytes of data at project completion.

TCGA tissue samples along with clinical data have been collected by 43 source sites and sent to the **Biospecimen Core Resources** (BCRs) for processing, quality control and storage. The BCRs submit patient and sample data to the **Data Coordinating Center** (DCC) and specimens for characterization to each of seven **Genome Characterization Centers** (GCCs) and for large-scale molecular analysis, to one of three **Genome Sequencing Centers** (GSCs). The GCCs use multiple platforms to generate exome and genome sequences; mRNA, miRNA and protein expression; copy number alterations, methylation data. Unprocessed and processed data from the GCCs and GSCs is submitted to the DCC and deposited in the **Cancer Genomics Hub** (CGHub). Finally, seven Genome Data Analysis Centers (GDAC) were commissioned to integrate and analyze TCGA data.



TCGA Data Flow from https://wiki.nci.nih.gov/display/TCGA/The+Cancer+Genome+Atlas.

Researchers can access TCGA data through the Data Portal (https://tcga-data.nci.nih.gov/tcga/) for open-access data, or from CGHub (https://cghub.ucsc.edu/) for controlled-access data. Controlled-access data requires application to and approval from dbGaP.

TCGA open-access data contains:

- Available clinical information for each participant (including demographic, treatment, survival data)
- Exome and genome MAF files containing **somatic** (only) mutations for each participant
- RNA-seq calculated expression levels for genes, exons, splices and isoforms
- miRNA-seq calculated expression levels for miRNA
- Limited array-based normalized gene expression levels
- Limited protein expression images and normalized expression levels
- Array-based methylation probe intensities and calculated beta values
- Regions of normalized copy number and purity/ploidy data
- Regions of coverage variation between normal and tumor samples from low-pass sequencing

TCGA controlled-access data contains:

- Pathology reports for a subset of participants
- Exome and genome BAM files for both tumor and normal sample for each participant
- Exome and genome VCF files containing somatic and *germline* mutations for each participant
- RNA-seq BAM files for each participant's tumor sample
- miRNA-seq BAM files for each participant's tumor sample

Station X Strategy

Currently, researchers wanting to incorporate information from TCGA into their workflows need to download data stored at multiple locations, integrate with their own data and then compute over this collection on powerful computational systems. This has become untenable given the enormous growth of TCGA and other large-scale sequencing datasets.

In order to mitigate the daunting challenges brought about by this deluge of genomic data, Station X has built a cloud-based Software as a Service (SaaS) platform called GenePool, which uses TCGA data to satisfy these key aims:

1. Provide a persistent, secure and curated repository for storing, cataloging, and accessing cancer genome sequence, alignment, sequence variation, expression, copy number and structural variation information from the TCGA consortium
2. Deliver an innovative data model and cloud-based compute infrastructure to enable large-scale, multi-genome comparisons with quickly and easily filtering to identify associations of genomic features with clinical outcomes (i.e. biomarkers)
3. Provide an open and robust set of tools to meet the demands of a broad user base, including application developers, bioinfomaticians, computational biologists, molecular biologists, and clinicians.

Station X has focused on the bringing in the data types that are the most comprehensively-available across all cancer types, and will prioritize cancer types by their degree of completion (fraction of samples assayed) and research interest. TCGA content in GenePool will be regularly updated, and may be adjusted due to changes in data availability, progress and updating.

Station X has currently imported open-access data into GenePool from 33 cancer types, listed in the table below, and will update content quarterly based on availability. This data comprises for over 68,000 individual datasets for 6 data types (copy number, methylation, miRNA-seq, protein expression, RNA-seq and somatic mutations) derived from almost 24,000 samples from more than 11,000 patients.

TCGA Data in GenePool

Station X has currently imported seven data types for the 33 cancers shown in the table below.

Copy Number

Copy Number alterations for genes are derived from the TCGA Level 3 SNP Array data. The source data was generated with the Affymetrix Genome-Wide Human SNP Array 6.0 and was processed by the Broad Institute GCC to derive segmented copy–number data using a GenePattern pipeline. Station X converted the copy number log-ratios for genomic regions in the source data to copy number values for genes in those regions. Copy number variants observed in normal samples were not included.

Methylation

Methylation levels for genes are derived from the TCGA Level 3 array-based data. The source data was generated with the Illumina Infinium HumanMethylation450 BeadChip and was processed by the Johns Hopkins GCC to derive beta values for CpG sites and their association with gene regions using various R routines including methylumi. Station X converted the CpG beta values to average beta methylation levels for genes.

miRNA-seq

miRNA expression counts are derived from the TCGA Level 3 miRNA sequencing data. The source data was generated using the Illumina HiSeq platform and was processed by Canada's Michael Smith Genome Sciences Centre GSS to derive counts of reads mapped to each miRNA.

Protein Expression

Protein expression levels are derived from the TCGA Level 3 normalized expression data. The source data was generated using reverse phase protein array (RPPA) and was processed by the MDAnderson Cancer Center GCC. Expression values are median-centered normalized and log2-scaled

RNA-seq

mRNA expression counts are derived from the TCGA Level 3 RNAseqV2 expression data. The source data was generated using the Illumina HiSeq 2000 (and GenomeAnalyzer for 3 cancers) and was processed by the University of North Carolina GCC to produce counts using MapSplice for alignment and RSEM for quantification. Station X currently provides gene and isoform read counts, and will soon include exons and junctions mapped to their respective genes.

Somatic Mutations

Somatic mutations are derived from the TCGA Level 2 MAF files. The source data was generated by tumor-specific Analysis Working Groups (AWGs) which take the auto-generated variant calls from the Genome Sequencing Centers (GSCs), and remove false-positive variants, or recover those missed by the GSCs. This is either done manually by experts, or by automated scripts or filters based on the extensive domain knowledge. The most complete somatic mutation call sets were imported from the Broad, HGSC (Baylor College of Medicine) and Washington University GSCs.

GenePool

GenePool TCGA Open-Access Dashboard

Code	Copy Number	Methylation	miRNA-seq	Protein Expression	RNA-seq	Somatic Mutations	Description
ACC	180	80	80	46	79	**91**	Adrenocorticol Carcinoma
BLCA	803	440	436	**344**	427	412	Urothelial Bladder Carcinoma
BRCA	2,218	**930**	869	**410**	1,218	1,044	Breast Ductal and Lobular Invasive Carcinoma
CESC	588	312	313	173	310	194	Cervical Cancer
CHOL	85	45	45	30	45	72	Cholangiocarcinoma
COAD	978	354	**280**	**364**	329	221	Colon Adnocarcinoma
DLBC	98	48	47	33	48	**48**	Lymphoid Neoplasm Diffuse Large B-cell Lymphoma
ESCA	373	191	200	126	196	190	Esophageal Cancer
GBM	1,140	155	5	244	174	291	Glioblastoma Multiforme
HNSC	1,091	580	532	212	566	**571**	Head and Neck Squamous Cell Carcinoma
KICH	132	66	91	63	91	**66**	Chromophobe Renal Cell Carcinoma
KIRC	1,113	485	332	**478**	606	494	Clear Cell Kidney Carcinoma
KIRP	603	321	326	216	323	168	Papillary Kidney Carcinoma
LAML	380	194	188	—	173	197	Acute Myeloid Leukemia
LGG	1,019	534	530	435	534	530	Lower Grade Glioma
LIHC	762	430	426	184	424	230	Liver Hepatocellular Carcinoma
LUAD	1,141	507	504	**365**	576	648	Lung Adenocarcinoma
LUSC	1,052	415	388	**328**	554	178	Lung Squamous Cell Carcinoma
MESO	172	87	87	63	87	**83**	Mesothelioma
OV	1,190	10	495	**443**	**309**	142	Ovarian Serous Cystadenocarcinoma
PAAD	365	195	183	123	183	**147**	Pancreatic Ductal Adenocarcinoma
PCPG	346	187	187	82	187	**187**	Paraganglioma & Pheochromocytoma
PRAD	1,029	553	551	352	550	473	Prostate Adenocarcinoma
READ	320	106	97	**131**	105	81	Rectal Adnocarcinoma
SARC	516	269	263	227	265	263	Sarcoma
SKCM	939	476	453	**356**	474	371	Cutaneous Melanoma
STAD	973	398	446	392	**450**	289	Stomach Adenocarcinoma
TGCT	304	156	156	122	156	156	Testicular Germ Cell Tumors
THCA	1,026	571	573	376	572	428	Papillary Thyroid Carcinoma
THYM	248	126	126	90	122	125	Thymoma
UCEC	1,100	485	450	**442**	201	248	Uterine Corpus Endometrial Carcinoma
UCS	111	57	56	48	57	**57**	Uterine Carcinosarcoma
UVM	161	80	80	12	80	**80**	Uveal Melanoma
Total	**22,556**	**9,843**	**9,795**	**7,310**	**10,471**	**8,775**	**All 33 cancer types**

GenePool

TCGA Case Dashboard

Disease	% Complete	Genome	Exome	Mutation	RNASeq	miRNASeq	Methylation	SNP/CN
LUAD	499/500	50/50	582/500	546/500	517/500	513/500	581/500	518/500
HNSC	491/500	64/50	527/500	526/500	521/500	524/500	530/500	526/500
THCA	479/500	50/50	502/500	441/500	505/500	506/500	509/500	505/500
COAD	402/424	49/50	459/424	269/424	458/424	444/424	461/424	460/424
UCEC	472/500	51/50	555/500	248/500	558/500	550/500	561/500	559/500
LUSC	462/500	50/50	502/500	178/500	501/500	478/500	505/500	505/500
PRAD	460/500	20/50	498/500	425/500	497/500	494/500	500/500	498/500
SKCM	454/500	40/50	470/500	369/500	469/500	448/500	472/500	470/500
READ	151/169	18/50	171/169	116/169	167/169	161/169	168/169	167/169
LGG	444/500	40/50	516/500	516/500	516/500	512/500	518/500	515/500
GBM	441/500	54/20	430/350	410/350	184/189	N/A	600/500	604/500
OV	440/500	69/20	546/350	463/350	423/490	489/490	603/500	604/500
STAD	431/500	40/50	443/500	397/500	419/500	436/500	445/500	443/500
BRCA	836/1000	99/50	1084/1000	990/1000	1095/1000	1079/1000	1099/1000	1098/1000
KIRC	405/500	41/50	525/500	451/500	533/500	516/500	537/500	534/500
BLCA	405/500	23/50	412/500	412/500	408/500	409/500	414/500	412/500
LIHC	318/500	54/50	376/500	375/500	371/500	373/500	379/500	377/500
CESC	269/500	20/50	305/500	198/500	304/500	307/500	309/500	304/500
KIRP	266/500	38/50	290/500	283/500	290/500	291/500	293/500	290/500
SARC	225/500	40/50	255/500	258/500	259/500	259/500	263/500	261/500
LAML	207/500	50/50	151/450	197/500	179/500	188/500	194/500	200/500
PAAD	176/500	140/50	185/500	173/500	178/500	178/500	186/500	185/500
ESCA	133/500	19/500	184/500	185/500	184/500	184/500	187/500	185/500
PCPG	119/500	0/500	179/500	179/500	179/500	179/500	180/500	179/500
TGCT	100/500	0/500	150/500	150/500	150/500	150/500	151/500	150/500
THYM	82/500	0/500	123/500	123/500	120/500	124/500	125/500	124/500
UVM	62/500	0/500	80/500	80/500	80/500	80/500	81/500	80/500
ACC	56/500	0/500	92/500	92/500	79/500	80/500	81/500	92/500
MESO	50/500	0/500	0/500	83/500	87/500	87/500	88/500	87/500
KICH	50/500	66/500	66/500	66/500	66/500	66/500	67/500	66/500
UCS	42/500	53/500	57/500	57/500	57/500	57/500	58/500	57/500
DLBC	41/500	7/50	48/500	48/500	48/500	47/500	50/500	50/500
CHOL	25/500	0/500	51/500	36/500	36/500	36/500	37/500	36/500
FPPP	5/500	0/500	0/500	38/500	0/500	0/500	0/500	0/500
TOTALS	9498/17093	1245/6590	10814/1674	9378/16793	10438/1677	10245/16583	11232/17093	11141/1709