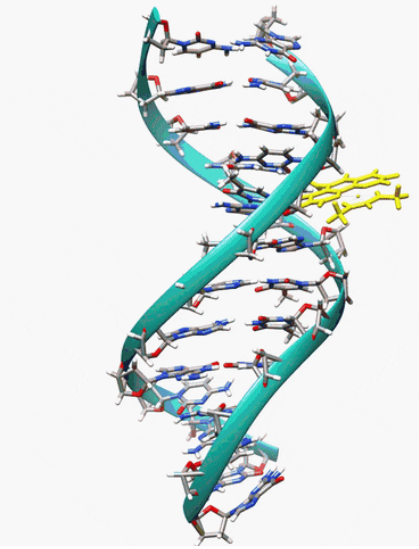
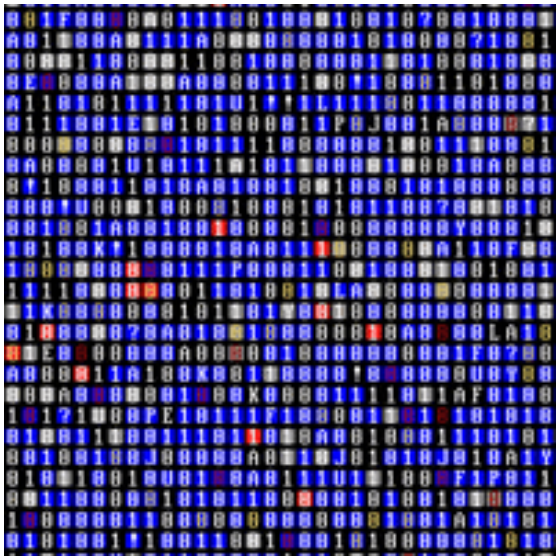
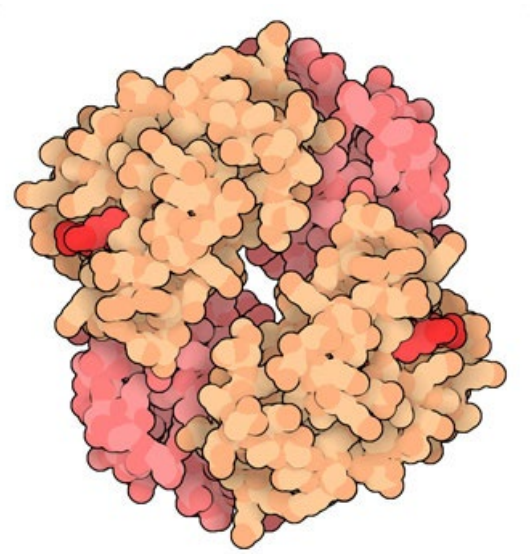
Data Centric View of DNA/RNA Information Flow

F. Alex Feltus, Ph.D.
CU-HackIt 2022
Professor, Clemson Dept. of Genetics & Biochemistry
Co-Founder Praxis AI LLC
ffeltus@clemson.edu

Important point:
The cellular "hard drive" is molecular. Molecular storage devices like DNA & RNA are sequences (aka strings) that are perfect for analysis on computers!





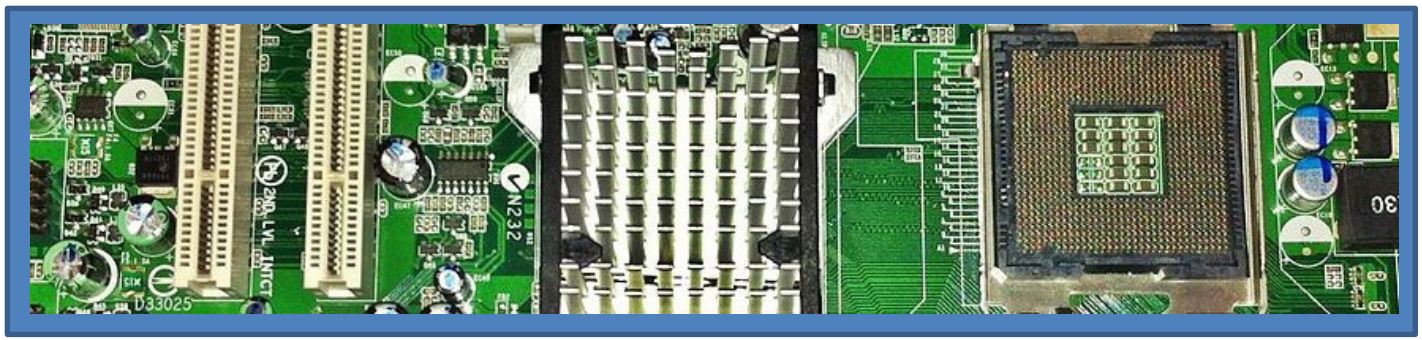


Who is this guy and why should I not be bored?

I Run A Small Interdisciplinary Computational Biology Research Lab

Mammals

Plants

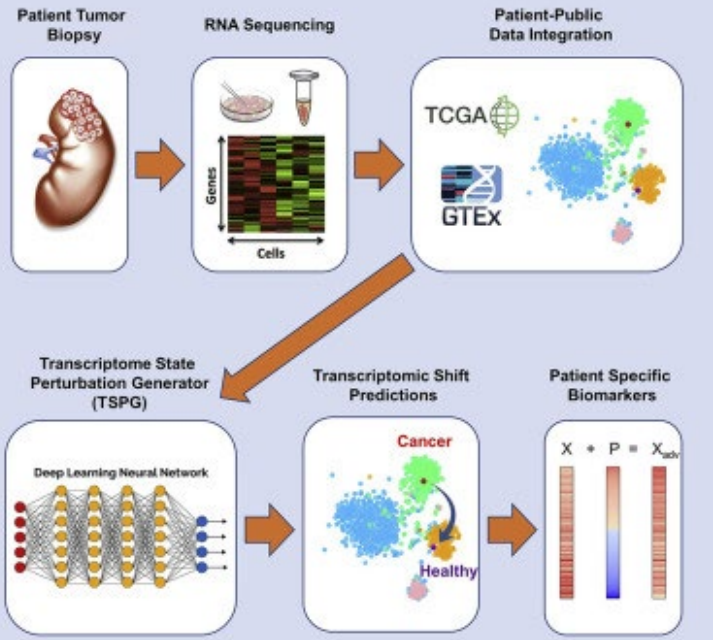


Bioinformatics + Cyberinfrastructure Engineering

Source: Wikipedia; Feltus Lab

Example Interdisciplinary Research in Feltus Lab @ Clemson

Leverage AI for Precision Medicine



Patterns

Available online 17 August 2020, 100087

In Press, Corrected Proof

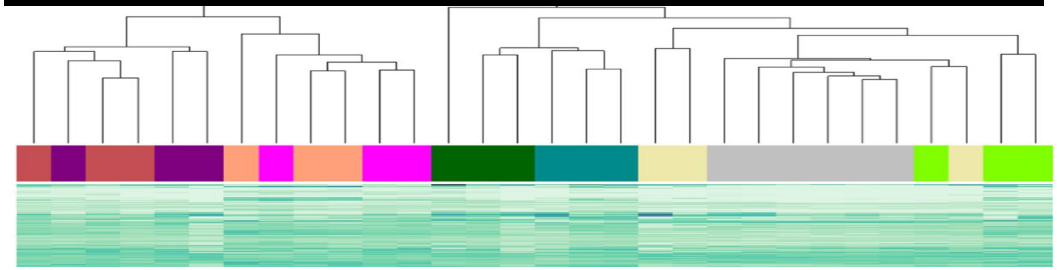
CellPress

Article

Cellular State Transformations Using Deep Learning for Precision Medicine Applications

Colin Targonski [1, 6], M. Reed Bender [2, 6], Benjamin T. Shealy [1], Benafsh Husain [2], Bill Paseman [3], Melissa C. Smith [1], F. Alex Feltus [2, 4, 5, 7]

Find Symbiosis Genes that may Reduce Fertilizer Usage in Agriculture



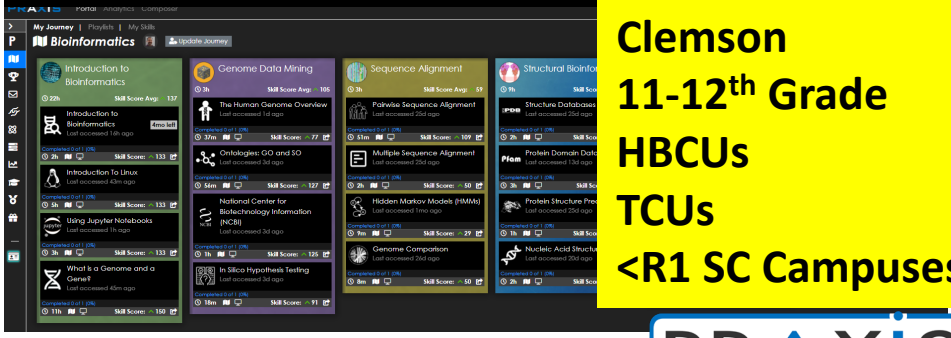
ORIGINAL RESEARCH ARTICLE

Front. Plant Sci., 31 October 2019 | https://doi.org/10.3389/fpls.2019.01409

Identifying Temporally Regulated Root Nodulation Biomarkers Using Time Series Gene Co-Expression Network Analysis

William L. Poehlman†, Elise L. Schnabel, Suchitra A. Chavan, Julia A. Frugoli* and Frank Alex Feltus†

Scalable Computational Biology Training



Clemson
11-12th Grade
HBCUs
TCUs
<R1 SC Campuses

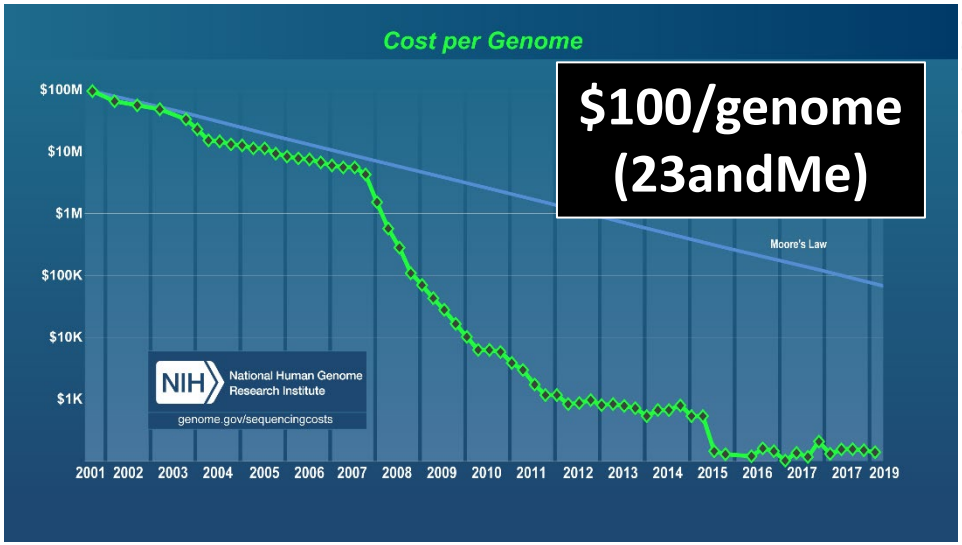
PRAXIS

prxai.com

+Brain and Cancer Genomics!

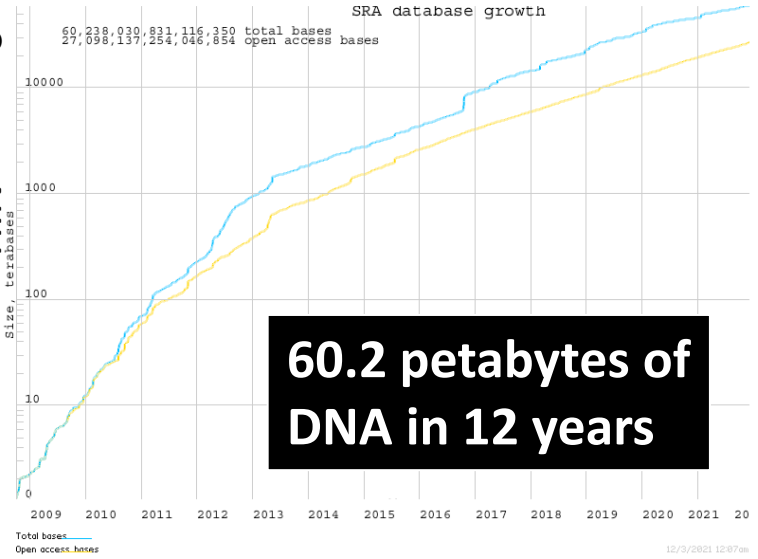
A Lot Has Happened in Biology in the Last Decade

DNA Sequencing Cost is Dropping

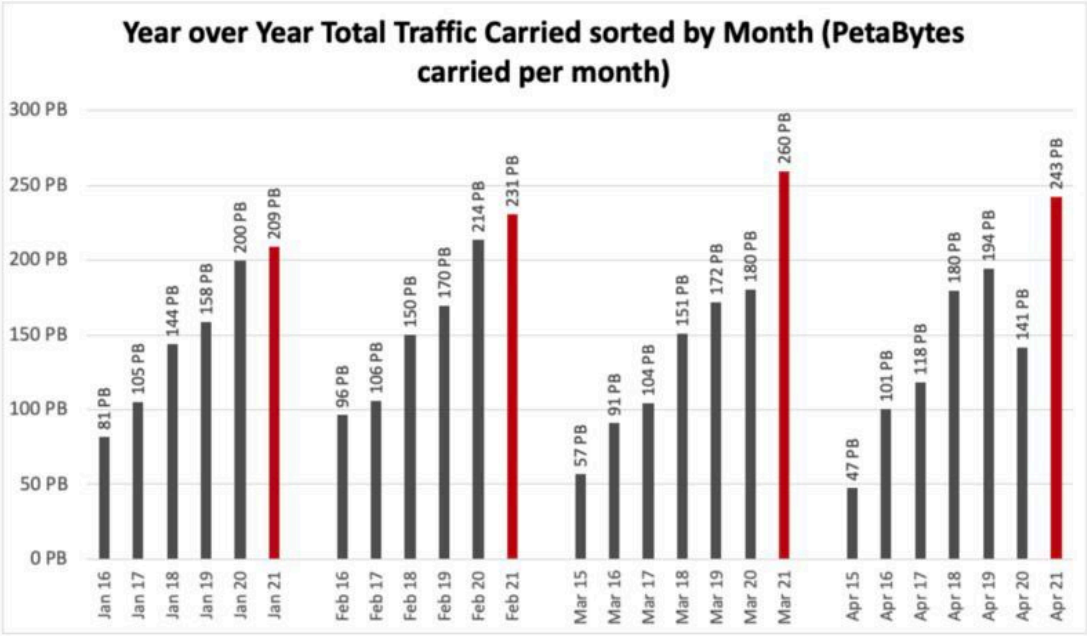


Source: NIH

DNA Databases are Swelling



Source: NCBI SRA

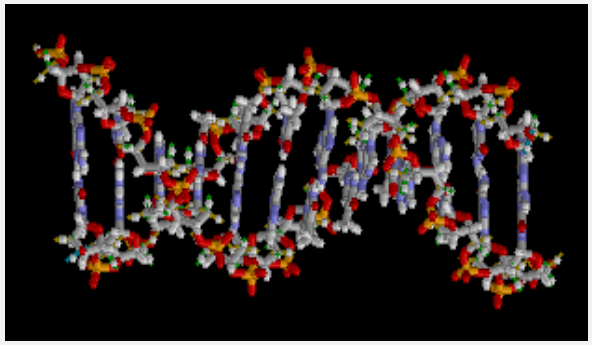


Internet2 Traffic > 1 Exabyte Jan-May 2021

Bioinformatics Requires Advanced Cyberinfrastructure

Cyberinfrastructure is a research environment that supports advanced data acquisition, data storage, data management, data integration, data mining, data visualization and other computing and information processing services distributed over the Internet beyond the scope of a single institution. -wikipedia

Bioinformatics is the technology that uses computers for storage, retrieval, manipulation, and distribution of information related to biological macromolecules (DNA, RNA, proteins, metabolites, and lipids).



We use the Cloud and Supercomputers to do our work

Clemson Palmetto Cluster

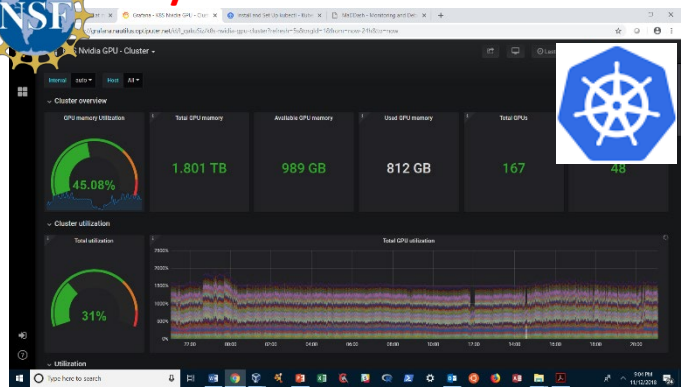
Open Science Grid

CENTER FOR HIGH THROUGHPUT COMPUTING chtc.cs.wisc.edu

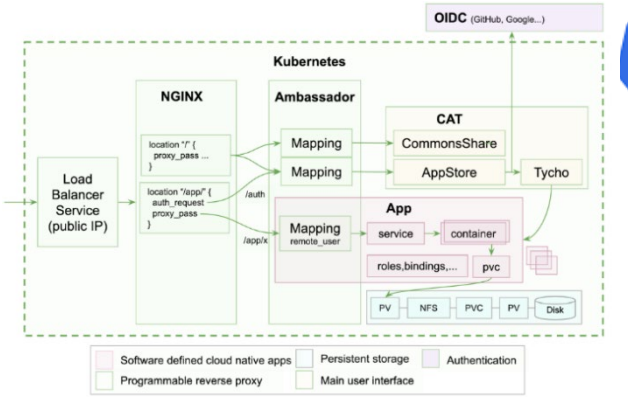
In 2017 on OSG …

8.43 Million Wall Hours
4.50 Million CPU Hours
8.92 Million Jobs
16.6 Million Transfers
4.07 PB

PRP/NRP Kubernetes Cluster



SciDAS (Scientific Data Analysis at Scale): NSF CC*



Google Cloud Platform

Cisco Container Platform

CISCO

Google Cloud

TACC Rodeo Cluster

TACC
TEXAS ADVANCED COMPUTING CENTER

A Lot Will Happen in Biology in the Next 15 Years

The tea leaves say…

In 2022, most funded biology labs are outsourcing DNA sequencing.
Terascale genomics experiments are common now and accessible to all in the US.

In 2032, every university research lab will have a DNA sequencer.
R1 research labs will move to the peta-/exascale in this PhD generation.

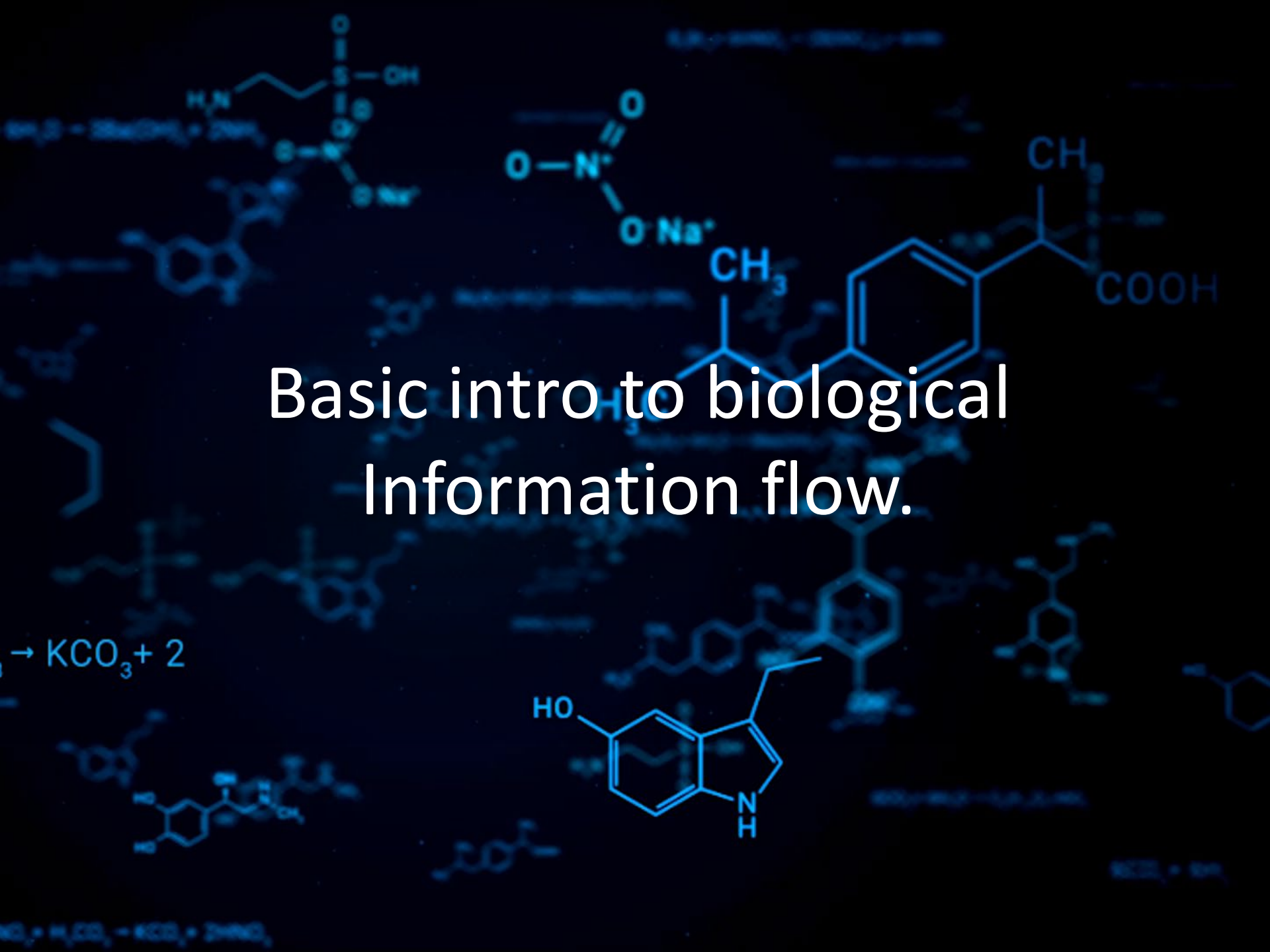
In 2037, all pharmacies, subways, hospitals, police stations, etc will have DNA sequencers.
These IoT DNA "sensors" will generate exabytes of data in aggregate each week.

We need more technical workers and start ups!

I am only talking DNA Sequencers…not CryoEM, Simulations, Medical Imaging

amazon web services™ INTERNET 2 CVS pharmacy® 23andMe

Basic intro to biological Information flow.



Organisms* are mostly 6 chemical elements (CHNOPS)

CARBON HYDROGEN NITROGEN OXYGEN PHOSPHOROUS SULFUR







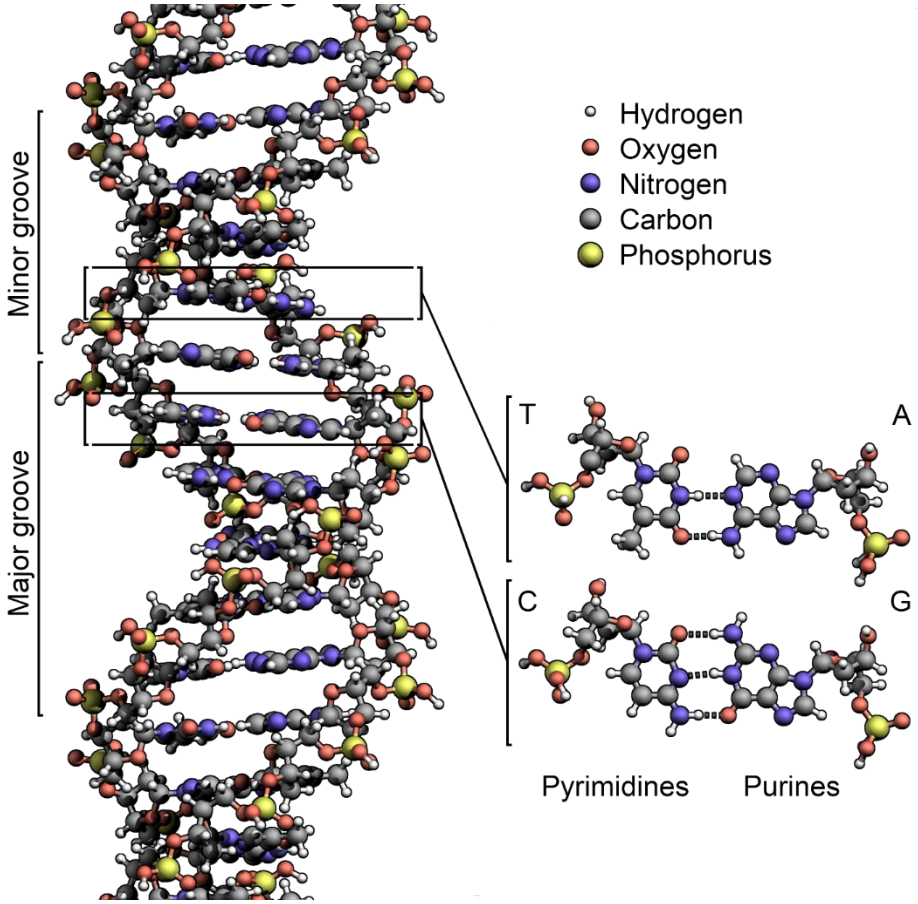
Group	1	2		3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
Period																			
1	1 H																		2 He
2	3 Li	4 Be												5 B	6 C	7 N	8 O	9 F	10 Ne
3	11 Na	12 Mg												13 Al	14 Si	15 P	16 S	17 Cl	18 Ar
4	19 K	20 Ca		21 Sc	22 Ti	23 V	24 Cr	25 Mn	26 Fe	27 Co	28 Ni	29 Cu	30 Zn	31 Ga	32 Ge	33 As	34 Se	35 Br	36 Kr
5	37 Rb	38 Sr		39 Y	40 Zr	41 Nb	42 Mo	43 Tc	44 Ru	45 Rh	46 Pd	47 Ag	48 Cd	49 In	50 Sn	51 Sb	52 Te	53 I	54 Xe
6	55 Cs	56 Ba	*	71 Lu	72 Hf	73 Ta	74 W	75 Re	76 Os	77 Ir	78 Pt	79 Au	80 Hg	81 Tl	82 Pb	83 Bi	84 Po	85 At	86 Rn
7	87 Fr	88 Ra	**	103 Lr	104 Rf	105 Db	106 Sg	107 Bh	108 Hs	109 Mt	110 Ds	111 Rg	112 Cn	113 Uut	114 Fl	115 Uup	116 Lv	117 Uus	118 Uuo

*Lanthanoids	*	57 La	58 Ce	59 Pr	60 Nd	61 Pm	62 Sm	63 Eu	64 Gd	65 Tb	66 Dy	67 Ho	68 Er	69 Tm	70 Yb
**Actinoids	**	89 Ac	90 Th	91 Pa	92 U	93 Np	94 Pu	95 Am	96 Cm	97 Bk	98 Cf	99 Es	100 Fm	101 Md	102 No

*CHNOPS only applies to organisms from this region of the Milky Way.

Wikipedia; giphy.com

DNA Is the Blueprint of Life and Built from CHNOP Elements



The alphabet of DNA are 4 molecules:

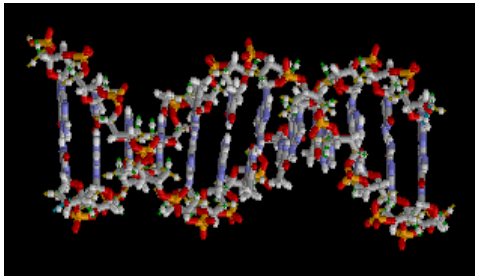
Adenine

Thymine

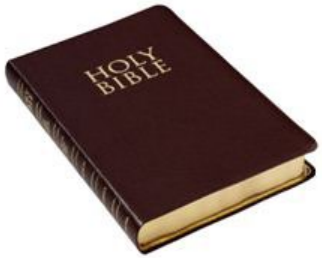
Guanine

Cytosine

The sequence of the alphabet makes words called genes that make the sentences, paragraphs, and chapters of an organism.



Genetics: Information in Human Book vs Human Genome

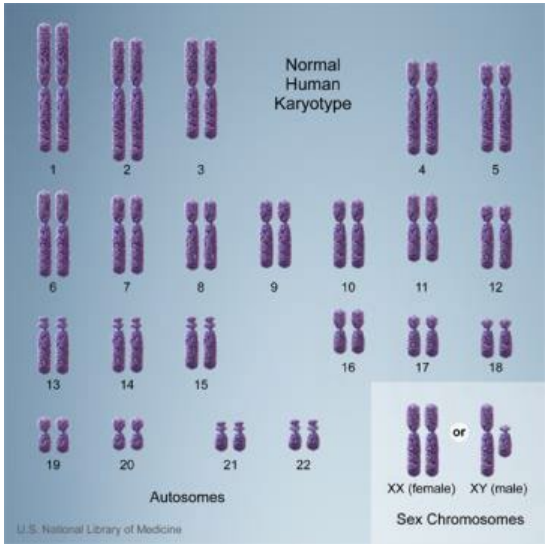


774,746 Words
26 letter alphabet

John 3:16: For God so loved the world that he gave his one and only Son, that whoever believes in him shall not perish but have eternal life.

John 3:16: For God so loved the world that he gave his one and only Sun, that whoever believes in him shall not perish but have eternal life.

One letter change = big difference!



60,651 Words (Genes)
4 letter alphabet (ATGC)
6.4 billion letters

AATGGAGCCACATAACACATTCAAACTTACTTGCAAAATAT

AATGGAGCCACATAACACATGCAAACTTACTTGCAAAATAT

One letter change = higher risk for breast and ovarian cancer!

Change DNA and Change The Person: Coffee Example

TGTGGGCACAGGAC

TGTGGGCACAGGAC

The sequence of DNA contains information on how an organism responds to the world.

TGTGGGCCCAGGAC

TGTGGGCACAGGAC

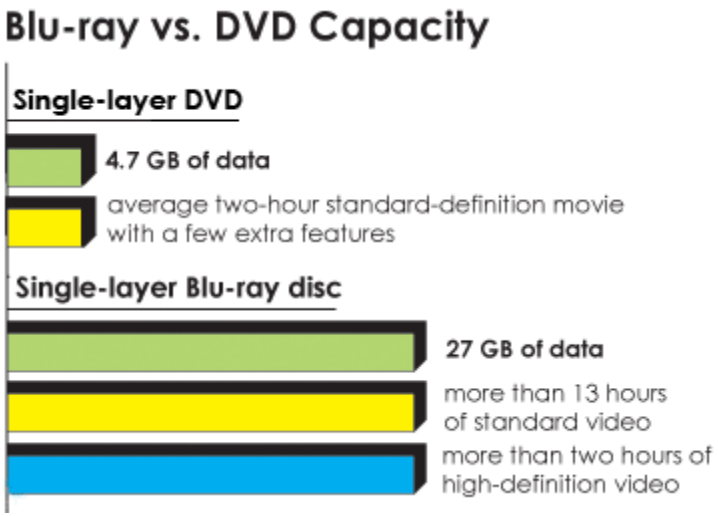
TGTGGGCCCAGGAC

TGTGGGCCCAGGAC

	Genotype	What It Means
	AA	Fast caffeine metabolizer: drinking coffee didn't increase subjects' heart attack risk
	AC	Slow caffeine metabolizer: drinking coffee increased subjects' heart attack risk.
	CC	Slow caffeine metabolizer: drinking coffee increased subjects' heart attack risk.

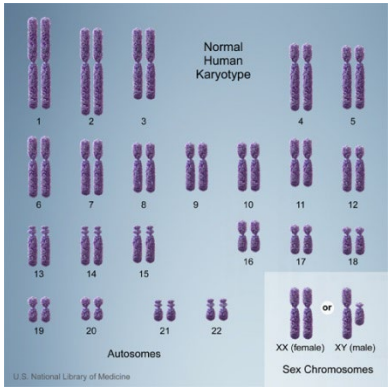
https://www.23andme.com/health/Caffeine-Metabolism/

Information is Encoded as A,T,G,C;
The Chromosome is The Storage Device



01010111 01101001 01101011
01101001 01110000 01100101
01100100 01101001 01100001

Check out:
http://www.ascii-code.com/



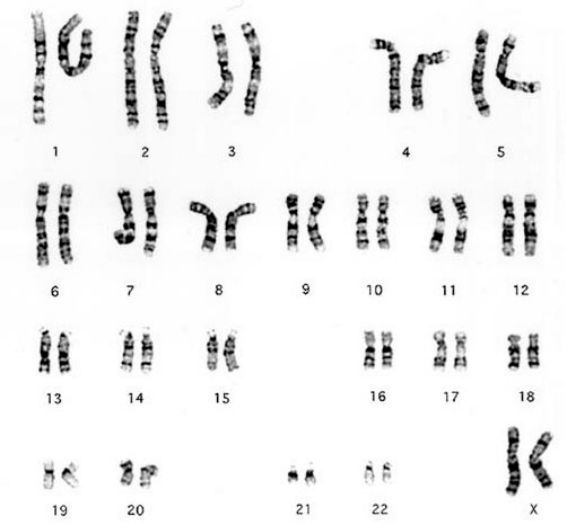
GGTTTCTATCTTACTAAAGGATATTTCAGAAAATCTATATTCACTGAG
GAGGATGATAATTGGGTCTACTAACATTGAGACTGAACTGAGGCCC
AGCAATAATTTAAACTTATTATCCTTTGAAGATTCAACTACTGGGGG
AGTACAACAGAAACAAATTAGAGAACATGAAGTTTTAATTCACGTT
GAAGATGAAACATGGGACCCAACACTTGATCATTTAG

All the information to make the unique version of us in encoded in 23 chromosome pairs spanning ~6,400,000,000 A,T,G,C base pairs (bytes)

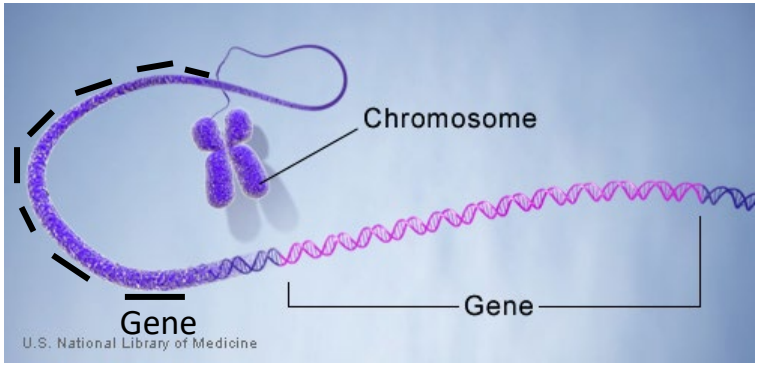
What is a genome? What is a gene?

Human Genome "Facts"

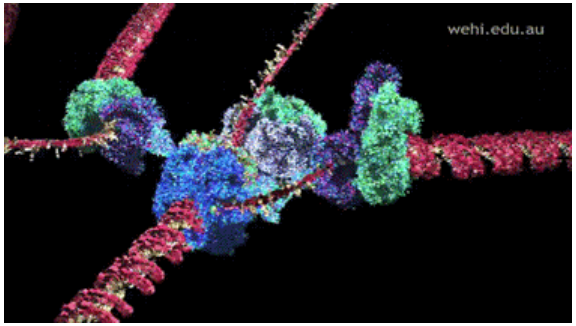
Initiation instructions for development
(3.2 billion*2) ATGCN
46 total chromosomes
24 unique chromosomes
(Chr1 = 250 million ATGCN)
60,651 genes



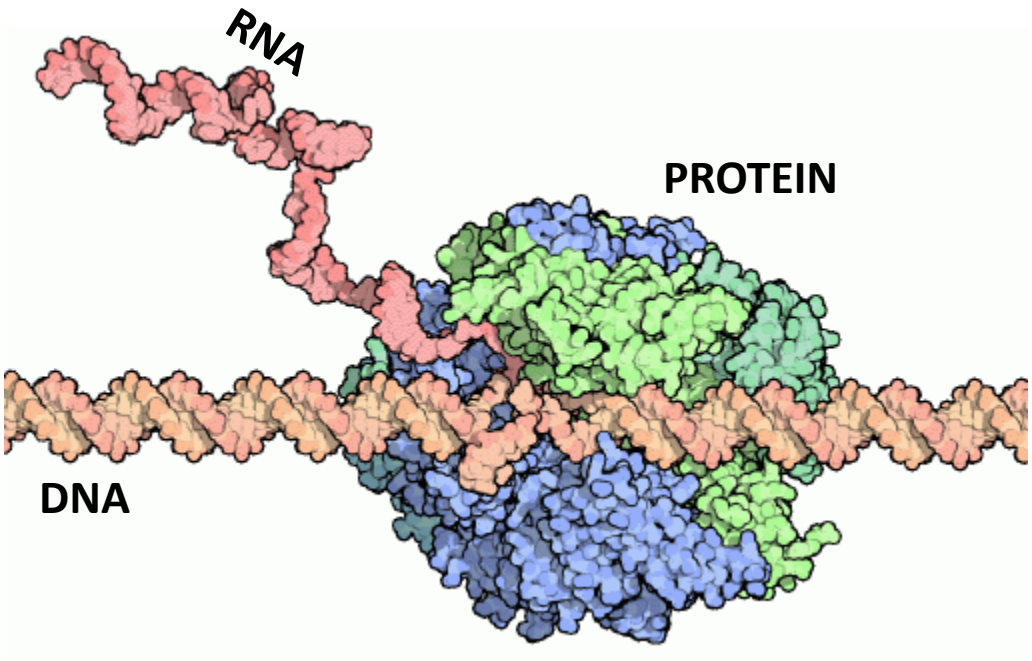
Courtesy of Dr. K. Phelan, Greenwood Genetic Center.
Noncommercial, educational use only.



- A **gene** is a **DNA interval** on the chromosome that gets copied (transcribe) into RNA.
- A **genome** is the set of **all chromosomes** and gene intervals.



Chromosome Information is Accessed with Enzymes; Blu Ray Information is Accessed with Lasers

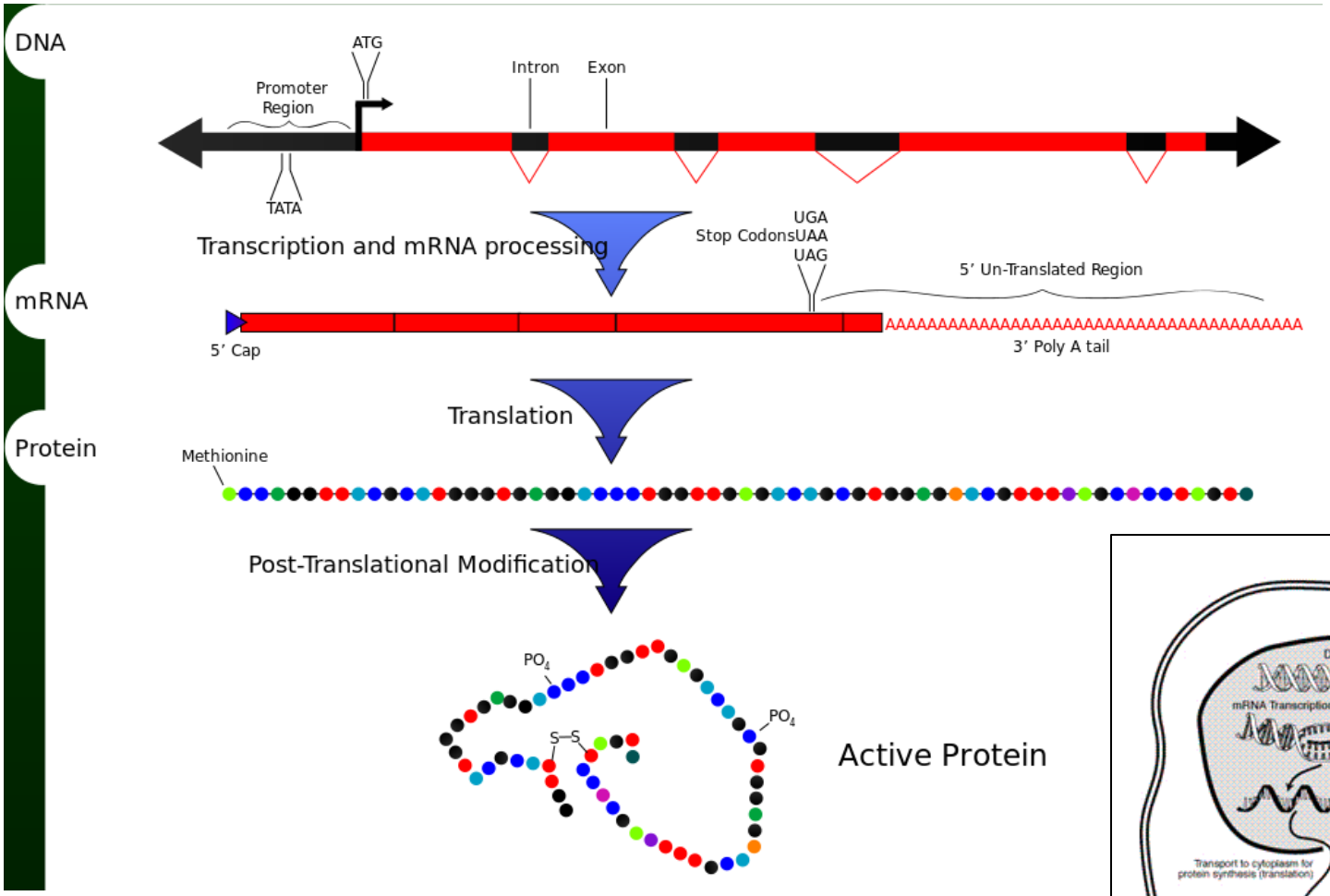


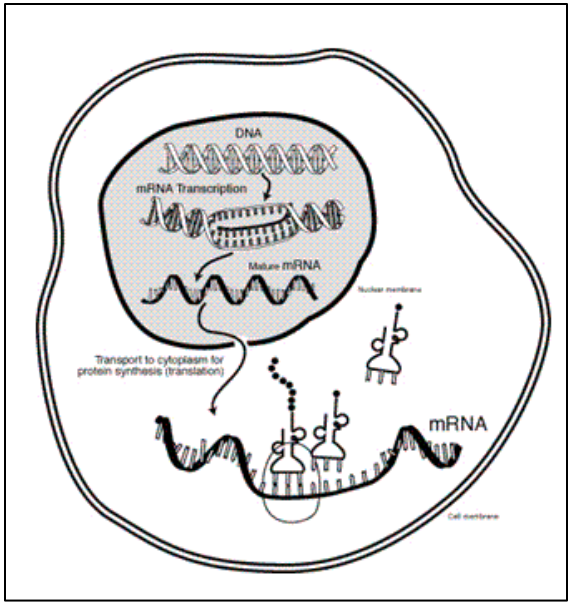
~The DNA molecule can be copied (Replication), transcribed (Transcription), and converted into protein (Translation)

http://www.rcsb.org/pdb/101/motm.do?momID=40

http://www.bigbrownboxblog.com.au

Molecular Information Flow

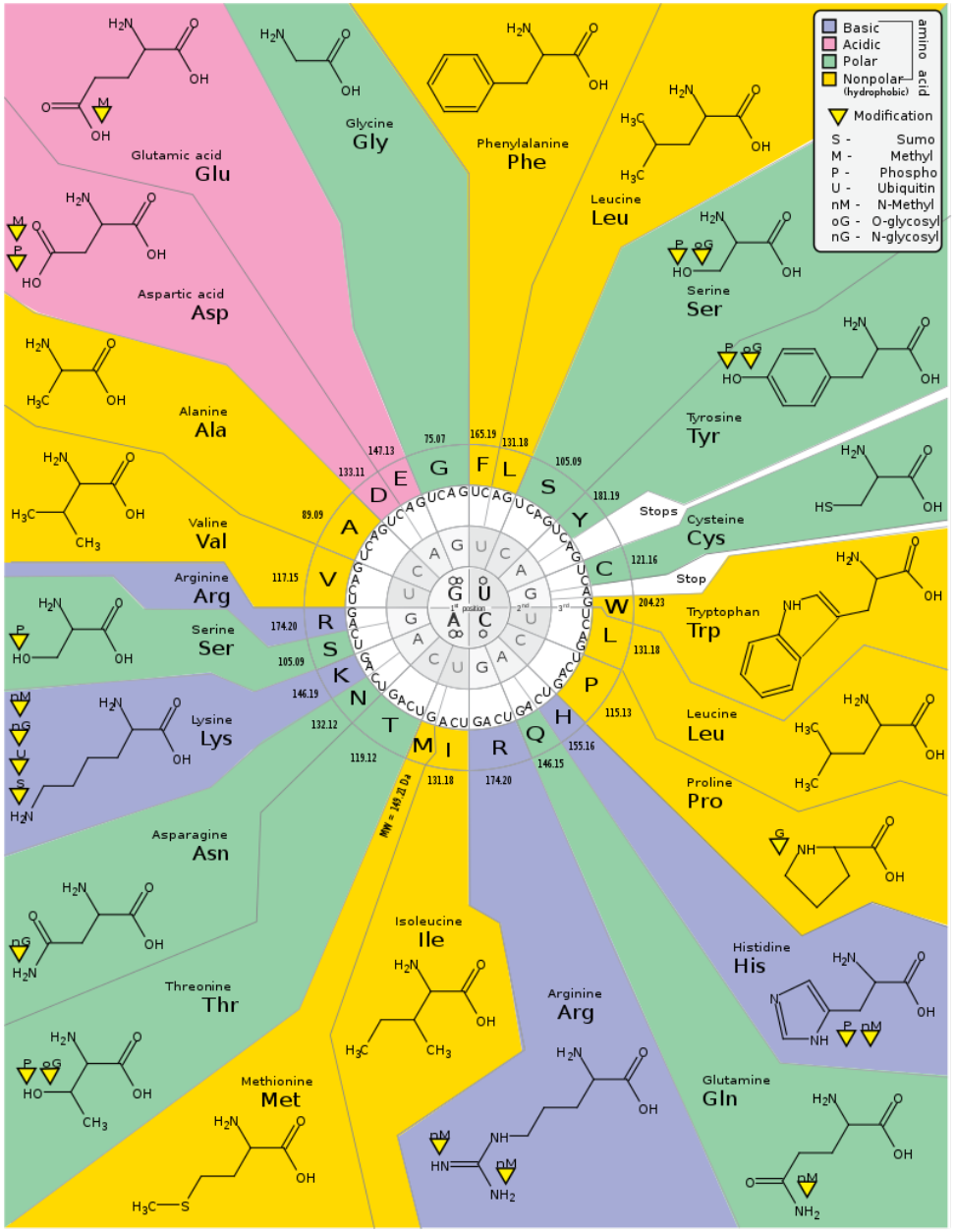




wikipedia

Cell = Sacks in a sack

Molecular Information Flow Starts with the Genetic Code



>NBR1 Gene Fragment DNA (String of Nucleotides)
ATGGCAGTTAAACAGGGAAACCAACTGCAGTGCAAG
TCCACGAAGGGCACCATGTCGTTGATGAAGCCCCAC

?

>NBR1 Gene Product RNA (String of Nucleotides)
AUGGCAGUUAAACAGGGAAACCAACUGCAGUGCAAG
UCCACGAAGGGCACCAUGUCGUUGAUGAAGCCCCAC

?

>NBR1 Gene Product Protein (String of Amino Acids)
MAVKQGNQLQCKSTKGTMSLMKPH

Second letter

First letter	U	C	A	G	Third letter
U	UUU, UUC: Phe; UUA, UUG: Leu	UCU, UCC, UCA, UCG: Ser	UAU, UAC: Tyr; **UAA Stop**, **UAG Stop**	UGU, UGC: Cys; **UGA Stop**; UGG: Trp	U C A G
C	CUU, CUC, CUA, CUG: Leu	CCU, CCC, CCA, CCG: Pro	CAU, CAC: His; CAA, CAG: Gln	CGU, CGC, CGA, CGG: Arg	U C A G
A	AUU, AUC, AUA: Ile; **AUG Met**	ACU, ACC, ACA, ACG: Thr	AAU, AAC: Asn; AAA, AAG: Lys	AGU, AGC: Ser; AGA, AGG: Arg	U C A G
G	GUU, GUC, GUA, GUG: Val	GCU, GCC, GCA, GCG: Ala	GAU, GAC: Asp; GAA, GAG: Glu	GGU, GGC, GGA, GGG: Gly	U C A G

wikipedia

GGTTTCTATCTTACTAAAGGATATTTCAGAAAATCTATATTCACTGAGGAGGATGATAATTGGGTCTACTAACATTGAGACT
GAACTGAGGCCCAGCAATAATTTAAACTTATTATCCTTTGAAGATTCAACTACTGGGGGAGTACAACAGAAACAAATTAGA
GAACATGAAGTTTTAATTCACGTTGAAGATGAAACATGGGACCCAACACTTGATCATTTAGCTAAACATGATGGAGAAGAT
GTACTTGGAAATAAAGTGGAACGAAAAGAAGATGGATTTGAAGATGGAGTAGAAGACAACAAATTGAAAGAGAATATGG
AAAGAGCTTGTTTGATGTCGTTAGATATTACAGAACATGAACTCCAAATTTTGGAACAGCAGTCTCAGGAAGAATATCTTAG
TGATATTGCTTATAAATCTACTGAGGTACTAAATAAAGAGGAAGCACATTTTTAGTTATTAGTAGGTTCTGGCAGACTTTATT
CCCGTAAAGAGACAGATAGTAAATATTTTAGGCTTTGTGGGACATACAGTCTCTGTTGCAGTAACTCAGTTCTGTTATTGTA
GTGTGAAAGCAGCCACAGACAAATGTAAACAAATGTGCCTGTCTTTCAATAAAACTTTATCTACAAATACAGATAGTGGGC
CACATTTGGCTTGCGAGCTGTAGTTTGCTGTTCCTGACTTAGTATGATATGCAGCATTGTGTAGTGGAAAGATTTTATAGTA
TCAACAGACATTTGTAGGAGTTTGCTATTTTATAATTTCGTATGCTTTTTTCATAATAGATGTCTCTGCTTTTTGTGGATTCTCT
TCATATCTACTAGTGTTTACTCTTCTGACTTTGATTTTTGGGGGTGTATTTTATTTCTTTATACAGAATTTTTTTCTTTTTTCTTT
CTTTTTTTGAGACAAGGTTTTGCTGTGaTCACCCCAGCTGGAGTGCAGTGGTGGGATCTTGGCCCCCTGCGGTCTTCAACTC
CTGGGCTTAAGCGATCCTGTCACCTCAGCCTCCCTAGTAGCTGGGACCACAGGCCTGCACCACTACACTCAGCTAATTTTTG
TATTTTTTGTAGAGATGGGGTTTCGCTCTGTTGCCCAGACTGGTCTCCAACTCCTGGACTCAAGTGATCCACCTGCCTTGGCC
TCCCAAAGTGTTGGAATTATAGGTGTGAGTCACTGTGCCTGGCCCAGAATTTCTTTACTTGCTCATGTGAGAGAGTTTCTAA
TAAACTCCTGATGCTTTTGCTTGACACTATAGTTTTACTGTCATCTCAGAGTCAGATTTTTTCCTTAATTTCTTCATTAAATTGA
TTTCTTAATTTCTCATTTGTTTTGCATTAGTTTACAAAATTACTTAAGATGCCTATATTCTCTGAAATTTTATATTTTATTTCACG
ATAAATAAGTTTAAAAAATTAGATTTGACCATTTAATTAGAGTTATTTCCCCATCATTTTTCTAATGGCAACAATAATTTTAGC
ATTTGAGCAATTCTCTGTATTTTGGTAACTATTAGTAATCACTTTTATCAGTATAAACATGCATTTCATTTTAAAATGGTTTTC
AAAATATTCTCATACCAACCATTCATAAA
ATTTTCTAGGAATTTCTTATTAGGTATAC
TTTTTTTTTGAGATGGAATCTATCTCTGT(
CCGGTTCAAGTGATTCTCCTACCTCAGCC
ATTTTTAGTAGAGAAAGTGCTGGGATTA
TACACCAATTATTGCCATTTTTTTATTGG(

GENE
DNA="ATGC"
(information)

ATATCAGTTTTTGCTAACATCTGCATTTCCCTTTG
ATAATATTATAGGATTATCAGGATTTCATGATTT
GGCGCAGTCTTGGCTCACTACAACCTCTGCCTCC
AGGTGCCCGCCACCACACCCGGCTAATTTTTGT
TGGCCTTGGATTTCGTGATTTGACTGACATCTTT
TCTTTATTCGTTAAATAAGAAGTAAAAGAACTT
TTAAAAAATAAAATAACTTTTTTTCCTAGTACAAAGCTTATTTTCAATTTTTTTTATTGCTTAAATTTTGCAGGTTAATTAAGGC
TGTACTTCTTTTTCCTACTTCCTGTTTAAATATTTGGCATTTTTTGAGGGTTTTCTTTCTATGGATTTTTGTGGGGGATCCCATC
TCGTTGGACACTTGGTTCCTGGAGATTGCCAAATTATTGAAAAGTTTCCTCATACCCAAACATTGGCAATAGACTGAATAAA
TAACATCTCTGACATTATTTTTCAATTATATAGAGAATGCTTTTTGTAAAATACTGATTTATTTAAAGATATTACAATCATAGT
AGTTTTTTGCAATCTTTTTTAAATCAGTGGATGTCAAGGAACAACTATTGTTCTTTGTACGGCATTTTTCAACATTGTGAAGA
GTTCTTATACTCAAAAAGTTTGGGAGACACAGTGTGACATTGTTTAACCTGAGTTGAACTTGTCATTTTGTATTCTTGTTTAG
AGCAGGGATTTTATTTTGTCTCTTTATCTAACTTTGTATTCCTAACTATTTCTTTCTTTCTCTTTGGTTCTTTGTTTCTTTGTTCAT
TCTTCTCTCTCTCCCTTTTTTATCTATCATATAAATATGTAAATATATATTATATATCCATTAGGGAAGAGGAAGTTGTCTAAA
GATATCTAGTATATAGGAGCTTTGTTCTGCAGAAAAGAGGATATGAAGTCAATTATATTGGAAATTAATGCTTAATACTTTT
TTTAAAGCATTTATCTCCCAATGATAATGAAAACGATACGTCCTATGTAATTGAGAGTGATGAAGATTTAGAAATGGAGATG
CTTAAGGTATGTTTACAATTATAAAAATATTACTTCAAGTTCTTTCCAAAGGACATTTAATTAAGTAAAATATTAACTAATTCT
AAACTAGGTTCTACCACAATGAAATTGCTACTAATTATGTAACATTAGATTTCACATTTTCCAATTCATGTTTCTTTCATGTAG
TCTATAAATAATGGGTTAGAGGTAATTTACTAATTTTAAATGTGCTTTCCTTGTTCCTTCTTATTTTTTATTTTGAAGACAGGG
TCTCACTCTATCACCCAGGCTGGAGTGCAGTTGCTCCATCTCGGCTCACTGCAACCTTCACCTCCTGGGCTCAAGTGATCCTC
CTGCATAAGCCTCCCGAGTAGCTGGGATTATGGGCGTGCACCACCATGCCCGGCTAATTTTTATAGTTTTAGTAGAGACAG
GGTCTCACCATGTTGTCCTTGCTGGTCTCGAACTCTTTACCTCAAGTAATCCACCCGCCTTGGCCTCCCAAAGTGCTGGGATT
ACAGGTGTGAGCCACCACACCTGGCCTTGGATTTCATGATTTGACTGACATCTTTTACACAAATTATTGCCATTTTTTTATTG
GGTACCAATTATTAGTATTGTTTTCTTTATTCGTTAAATAAGAAGTAAAATAACTTTTTTTTAAAAAAAGAACTTTTTTCCTAG
TAAAAAGCTTATTTTCAATTTTTTTATTGCTTAAATTTTGCAGGTTAATTAAGGCTGTGCTTCCTTTTCCTACTTTCTGTTTAAA
TGTTTGGCATTTTTTGAGGGTTTTCTTTCTATGGATTTTTGTGGGGGATCCCATCTTGTTGGACACTTGGTTCCTGGAGATTG
CCAAATTATTGAAAAGTTTCCTCATACCCAAACATTGGCAGTAGACTGAATAAATAACATCTCTGACACTATTTTTCAATTAT
ATAGAGAATGCTATTTGTAAAATACTGATTTATTTAAAGATATTACAGTCAGTAGTTTTTTGCAATCTTTTTAAAATCAGTGG
ATGTCAAAAACAGCTATTGTTCTTTGTACAGCATTTTTCAACATTGTGAAGAGTTCTTATACTCAAAAAGTTTGGGAGACACA
GTTTGATATTGTTTAACCTGAGTTGAACTTGTCATTTTGTATTCTTGTTTAGAGCAGGGATTTTATTTTGTCTCTTTATCTAACT
TTGTATTCCTATTTCTTTCTTTCTCTTTGGTTCTTTGTTTCTTTGTTCATTCTTCTCTCTCTCCCTTTTTTATCTATCATATAAATAT
GTAAATATATATTATATATCCATTAGGGAAGAGGAAGCTGTCTAAAGATATCTAGTATATAGGAGCTTTGTTCTGCAGAAAA
GAGGATATGAAGTCAATTATATTGGAAATTAATGCTTAATACTTTTTTTTAAAGCATTTATCTCCCAATGATAATGAAAACGA
TACGTCCTATGTAATTGAGAGTGATGAAGATTTAGAAATGGAGATGCTTAAG

?

GGUUUCUAUCUUACUAAAGGAUAUUUCAGAAAAUCUAUAUU
CACUGAGGAGGAUGAUAAUUGGGUCUACUAACAUUGAGACU
GAACUGAGGCCCAGCAAUAAUUUAAACUUAUUAUCCUUUGA
AGAUUCAACUACUGGGGGAGUACAACAGAAACAAAUUAGAG
AACAUGAAGUUU
CCAACACUUGAUC
ACUUGGAAAUAAA
AUGGAGUAGAAGA
GCUUGUUUGAUG

GENE PRODUCT
RNA="AUGC"
(information)

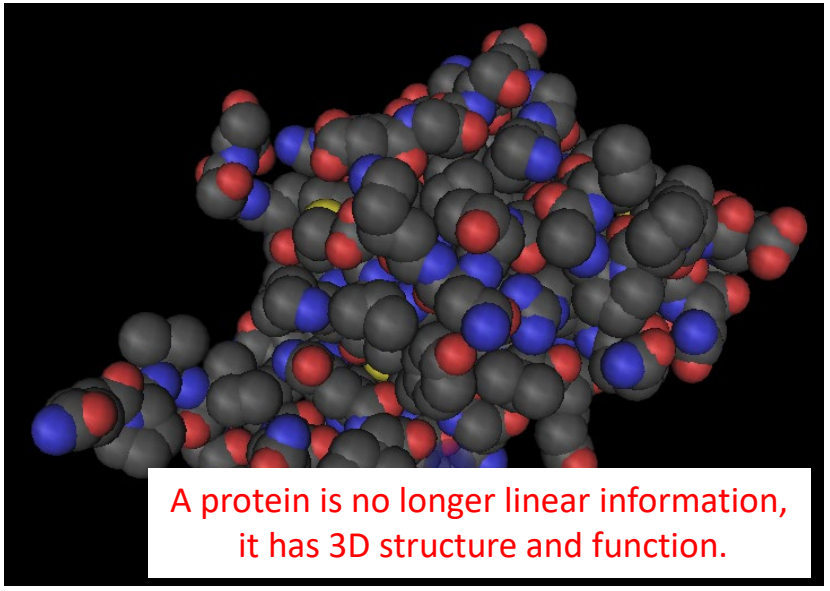
GAAACAUGGGAC
JGGAGAAGAUGU
AUGGAUUUGAAG
AAUAUGGAAAGA
AACAUGAACUCCA
AAUUUUGGAACAGCAGUCUCAGGAAGAAUAUCUUAGUGAUA
UUGCUUAUAAAUCUACUGAGCAUUUAUCUCCCAAUGAUAAU
GAAAACGAUACGUCCUAUGUAAUUGAGAGUGAUGAAGAUUU
AGAAAUGGAGAUGCUUAAG

?

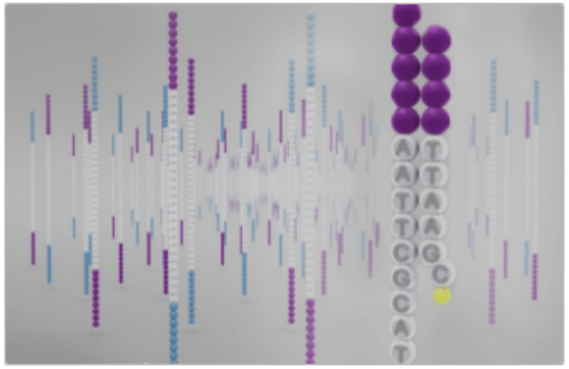
MIIGSTNIETELRPSNNLNLLSFEDSTTGGVQQKQIRHEVLIHVEDETWDP
TLDHLAKHDGEDVLGNKVERKEDGFEDGVEDNKLKENMERACLMSLDI
TEHELC
LK

EM

GENE PRODUCT
PROTEIN = "~20 unique amino acids letters"



How does one obtain DNA sequences from a sample?

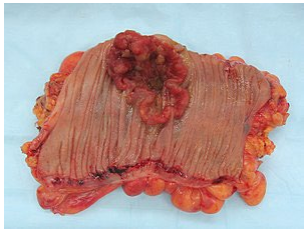


How Does Illumina NGS Work?

Illumina sequencing utilizes a fundamentally different approach from the classic Sanger chain-termination method. It leverages sequencing by synthesis (SBS) technology – tracking the addition of labeled nucleotides as the DNA chain is copied – in a massively parallel fashion.

Next-generation sequencing generates masses of DNA sequencing data, and is both less expensive and less time-consuming than traditional Sanger sequencing.[2] Illumina sequencing systems can deliver data output ranging from 300 kilobases up to multiple terabases in a single run, depending on instrument type and configuration.

Sample → **Purify DNA** → **DNA Sequencer** → **Computational Analysis**



DNA SEQUENCE/QUALITY FASTQ FILE

(10 – 1000 Million "DNA Reads"; $slen=100-300 characters)

```
@SRR001666.1 071112_SLXA-EAS1_s_7:5:1:817:345 length=36
GGGTGATGGCCGCTGCCGATGGCGTCAAATCCCACC
+SRR001666.1 071112_SLXA-EAS1_s_7:5:1:817:345 length=36
IIIIIIIIIIIIIIIIIIIIIIIIIIIIII9IG9IC
```

SEQUENCE

QUALITY

http://en.wikipedia.org/wiki/FASTQ_format

Dec	Hx	Oct	Char	Dec	Hx	Oct	Html	Chr	Dec	Hx	Oct	Html	Chr	Dec	Hx	Oct	Html	Chr
0	0	000	NUL (null)	32	20	040		Space	64	40	100	@	@	96	60	140	`	`
1	1	001	SOH (start of heading)	33	21	041	!	!	65	41	101	A	A	97	61	141	a	a
2	2	002	STX (start of text)	34	22	042	"	"	66	42	102	B	B	98	62	142	b	b
3	3	003	ETX (end of text)	35	23	043	#	#	67	43	103	C	C	99	63	143	c	c
4	4	004	EOT (end of transmission)	36	24	044	$	$	68	44	104	D	D	100	64	144	d	d
5	5	005	ENQ (enquiry)	37	25	045	%	%	69	45	105	E	E	101	65	145	e	e
6	6	006	ACK (acknowledge)	38	26	046	&	&	70	46	106	F	F	102	66	146	f	f
7	7	007	BEL (bell)	39	27	047	'	'	71	47	107	G	G	103	67	147	g	g
8	8	010	BS (backspace)	40	28	050	(	(	72	48	110	H	H	104	68	150	h	h
9	9	011	TAB (horizontal tab)	41	29	051	)	)	73	49	111	I	I	105	69	151	i	i
10	A	012	LF (NL line feed, new line)	42	2A	052	*	*	74	4A	112	J	J	106	6A	152	j	j
11	B	013	VT (vertical tab)	43	2B	053	+	+	75	4B	113	K	K	107	6B	153	k	k
12	C	014	FF (NP form feed, new page)	44	2C	054	,	,	76	4C	114	L	L	108	6C	154	l	l
13	D	015	CR (carriage return)	45	2D	055	-	-	77	4D	115	M	M	109	6D	155	m	m
14	E	016	SO (shift out)	46	2E	056	.	.	78	4E	116	N	N	110	6E	156	n	n
15	F	017	SI (shift in)	47	2F	057	/	/	79	4F	117	O	O	111	6F	157	o	o
16	10	020	DLE (data link escape)	48	30	060	0	0	80	50	120	P	P	112	70	160	p	p
17	11	021	DC1 (device control 1)	49	31	061	1	1	81	51	121	Q	Q	113	71	161	q	q
18	12	022	DC2 (device control 2)	50	32	062	2	2	82	52	122	R	R	114	72	162	r	r
19	13	023	DC3 (device control 3)	51	33	063	3	3	83	53	123	S	S	115	73	163	s	s
20	14	024	DC4 (device control 4)	52	34	064	4	4	84	54	124	T	T	116	74	164	t	t
21	15	025	NAK (negative acknowledge)	53	35	065	5	5	85	55	125	U	U	117	75	165	u	u
22	16	026	SYN (synchronous idle)	54	36	066	6	6	86	56	126	V	V	118	76	166	v	v
23	17	027	ETB (end of trans. block)	55	37	067	7	7	87	57	127	W	W	119	77	167	w	w
24	18	030	CAN (cancel)	56	38	070	8	8	88	58	130	X	X	120	78	170	x	x
25	19	031	EM (end of medium)	57	39	071	9	9	89	59	131	Y	Y	121	79	171	y	y
26	1A	032	SUB (substitute)	58	3A	072	:	:	90	5A	132	Z	Z	122	7A	172	z	z
27	1B	033	ESC (escape)	59	3B	073	;	;	91	5B	133	[	[	123	7B	173	{	{
28	1C	034	FS (file separator)	60	3C	074	<	<	92	5C	134	\	\	124	7C	174			|
29	1D	035	GS (group separator)	61	3D	075	=	=	93	5D	135	]	]	125	7D	175	}	}
30	1E	036	RS (record separator)	62	3E	076	>	>	94	5E	136	^	^	126	7E	176	~	~
31	1F	037	US (unit separator)	63	3F	077	?	?	95	5F	137	_	_	127	7F	177	&#127;	DEL

Source: www.LookupTables.com

http://en.wikipedia.org/wiki/Phred_quality_score
http://www.asciitable.com/

Base Call Quality (Q) score is ASCII Decimal# minus 33 (usually).
Q(I)=40:::99.999% Accuracy

$$Q = -10 \log_{10} P$$

or

$$P = 10^{\frac{-Q}{10}}$$

Phred quality scores are logarithmically linked to error probabilities

Phred Quality Score	Probability of incorrect base call	Base call accuracy
10	1 in 10	90%
20	1 in 100	99%
30	1 in 1000	99.9%
40	1 in 10000	99.99%
50	1 in 100000	99.999%

What do we do with DNA sequences?

(Align $little.string to $big.string)

Sequencing Output Per Flow Cell*

NovaSeq 6000 System

Flow Cell Type	SP	S1	S2	S4
1 × 35 bp	N/A	N/A	N/A	280-350 Gb
2 × 50 bp	65–80 Gb	134–167 Gb	333–417 Gb	N/A
2 × 100 bp	134–167 Gb	266–333 Gb	667–833 Gb	1600–2000 Gb
2 × 150 bp	200–250 Gb	400–500 Gb	1000–1250 Gb	2400–3000 Gb
2 x 250 bp	325-400 Gb	N/A	N/A	N/A

*Specifications based on Illumina PhiX control library at supported cluster densities.

```
for (my $i=0; $i <= $dna.read.count; $i++) {
        for (my $k=0; $k <= $ chromosome.count; $k++) {
                return_genome_coordinate ($DNA[$i], $chromosome[$k])
                if ($coordinate overlaps $gene) {$gene.$count++}
                else {evidence.for.new.gene}
        }
}
```

chr	total length
1	249,250,621
2	243,199,373
3	198,022,430
4	191,154,276
5	180,915,260
6	171,115,067
7	159,138,663
8	146,364,022
9	141,213,431
10	135,534,747
11	135,006,516
12	133,851,895
13	115,169,878
14	107,349,540
15	102,531,392
16	90,354,753
17	81,195,210
18	78,077,248
19	59,128,983
20	63,025,520
21	48,129,895
22	51,304,566
X	155,270,560
Y	59,373,566

Example Alignment of 1 out of $2*10^8$ Sequences

Primary Assembly

chr	total length
1	249,250,621
2	243,199,373
3	198,022,430
4	191,154,276
5	180,915,260
6	171,115,067
7	159,138,663
8	146,364,022
9	141,213,431
10	135,534,747
11	135,006,516
12	133,851,895
13	115,169,878
14	107,349,540
15	102,531,392
16	90,354,753
17	81,195,210
18	78,077,248
19	59,128,983
20	63,025,520
21	48,129,895
22	51,304,566
X	155,270,560
Y	59,373,566

Length of $chromosome.01

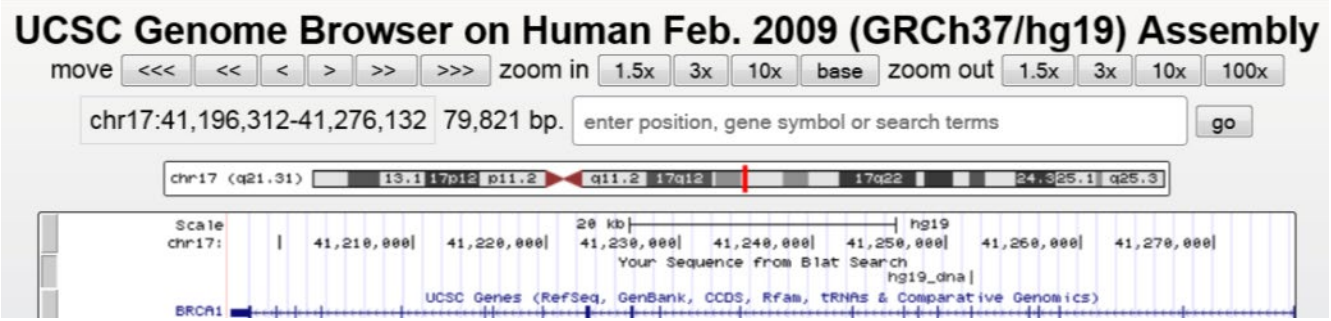
Read01 out of 200,000,000:

TCAATGTAGA CAGACGTCTT TTGAGGTTGT ATCCGCTGC

Search ALL chromosomes for alignment to Read01:
Hit! on Chromosome 17 @ 41,251,797-41,251,835

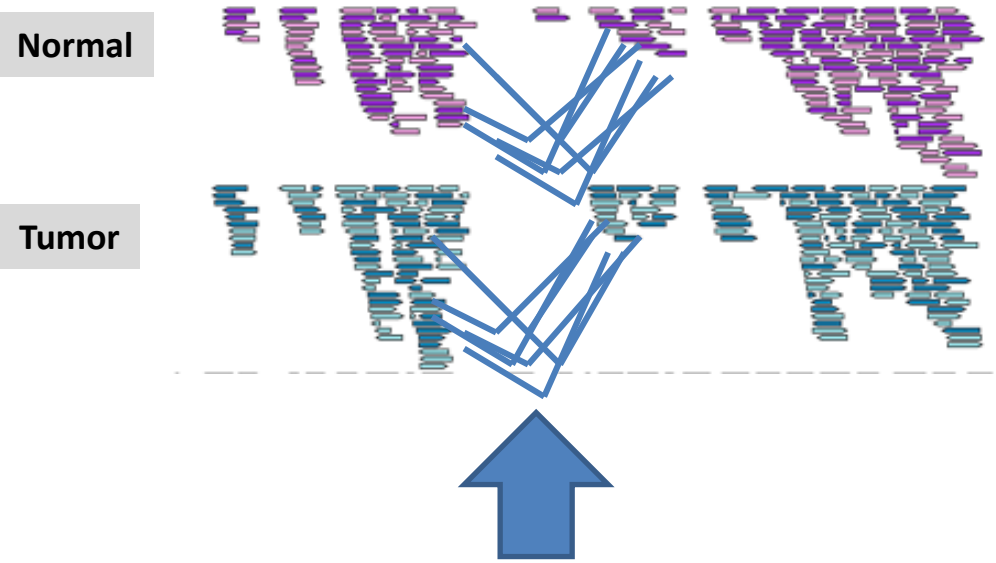
```
00000001 tcaatgtagacagacgtcttttgaggttgtatccgctgc 00000039
>>>>>>>> |||||||||||||||||||||||||||||||||||||||| >>>>>>>>
41251797 tcaatgtagacagacgtcttttgaggttgtatccgctgc 41251835
```

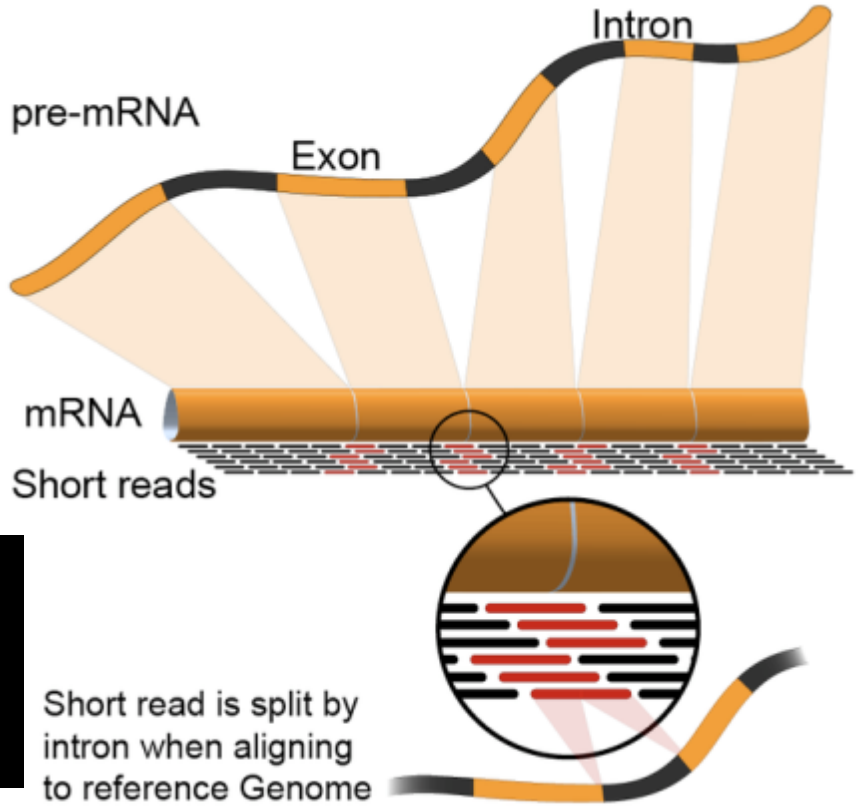
Add one to number of BRCA1 RNA transcripts expressed
Since this gene is encoded in the DNA at that coordinate.



Application: Counting RNA Molecules for Detection of Differential Gene Expression (DEG)

Are molecule counts significantly different at any given gene between experimental conditions?





RNA molecules are typically sheared to 300 base pairs long and sequenced at both ends. If the ends are longer than 300bp when aligned to DNA, then that is evidence for an intron-exon boundary

http://www.data2bio.com/services/rnaseq/

http://en.wikipedia.org/wiki/RNA-Seq

OUTPUT: Sample X Gene (Gene Expression Matrix: GEM)

Samples →

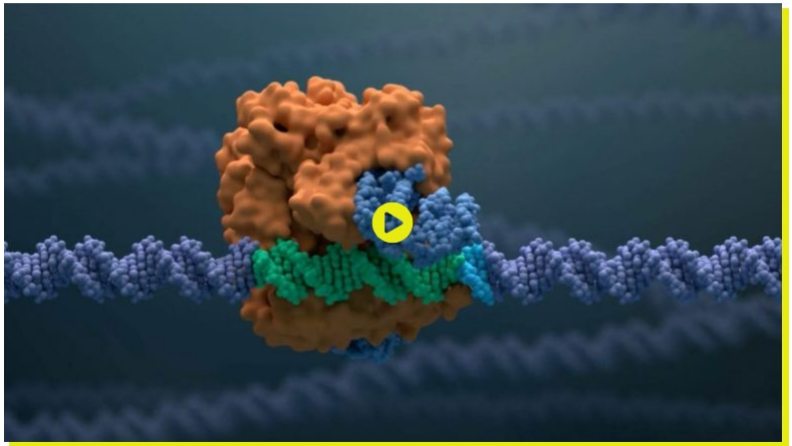
Genes ↓

	ATLAS	C.AMBER	GRASSL	LEOTI	PI_267573	PI_506030	PI_506069	PI_510757	RIO	WRAY	PI_229841	PI_297130	PI_329311
Sobic.001G000100.1	43	80	59	57	63	55	141	81	98	64	39	23	34
Sobic.001G000200.1	5019	4612	4159	4997	5025	5037	5688	5192	4721	4790	5003	5318	4080
Sobic.001G000300.1	13	211	363	197	816	16	22	25	609	33	936	1008	812
Sobic.001G000400.1	96	247	155	225	173	181	176	122	21	88	148	142	111
Sobic.001G000400.2	18214	16957	20109	22401	22026	14420	15878	17051	20284	21036	22790	24456	21542
Sobic.001G000500.1	0	2	0	0	0	0	0	2	0	0	0	0	0
Sobic.001G000600.1	158	895	1053	70	1405	1247	1264	1036	1194	35	70	176	1931
Sobic.001G000700.1	362	175	533	145	924	133	199	224	436	256	343	301	598
Sobic.001G000700.2	9223	7622	7228	8348	8610	10394	9611	10517	10360	8733	11743	11149	13108
Sobic.001G000800.1	13066	13519	12412	13375	14575	14003	15294	12427	14094	14013	13180	13968	12237
Sobic.001G000900.1	35	50	44	28	27	32	21	12	22	16	50	18	11
Sobic.001G001000.1	268	279	100	222	256	424	490	202	222	241	131	103	14
Sobic.001G001100.1	32	38	14	28	24	89	89	52	41	55	44	29	8
Sobic.001G001200.1	128	146	183	165	144	159	284	223	178	190	282	248	343
Sobic.001G001300.1	709	670	1114	1124	971	889	1092	826	809	909	926	395	753
Sobic.001G001400.1	12837	11848	10055	12620	13128	13101	12425	11465	12352	12017	12296	13688	8615
Sobic.001G001500.1	7002	6330	5217	6856	6767	6375	7224	5941	6374	6656	6639	8265	5347
Sobic.001G001600.1	0	0	0	0	0	17	0	9	0	0	0	29	0
Sobic.001G001700.1	83	0	11	49	43	76	0	89	0	103	95	211	0
Sobic.001G001700.2	33	1	17	11	0	0	0	1	8	13	27	0	0
Sobic.001G001800.1	3	0	0	3	6	0	0	6	0	0	10	29	0
Sobic.001G001900.1	5550	5800	4852	4827	4740	5997	4939	4807	4943	4973	5813	5272	5730
Sobic.001G002000.1	9862	11281	9520	10405	12690	11908	12670	10716	10612	10418	11096	12666	9160
Sobic.001G002100.1	1253	984	1615	1517	1244	1064	973	1064	1626	1675	746	1237	914
Sobic.001G002200.1	0	0	0	0	0	0	0	0	0	0	0	0	0
Sobic.001G002300.1	31	44	58	26	24	53	24	13	24	25	10	31	0
Sobic.001G002400.1	77	24	66	11	15	44	26	12	115	44	2	99	0
Sobic.001G002500.1	38	17	11	48	57	6	42	41	98	22	16	37	11
Sobic.001G002600.1	2189	1300	3272	3372	3617	4483	3582	3795	3332	2515	3798	3413	2815

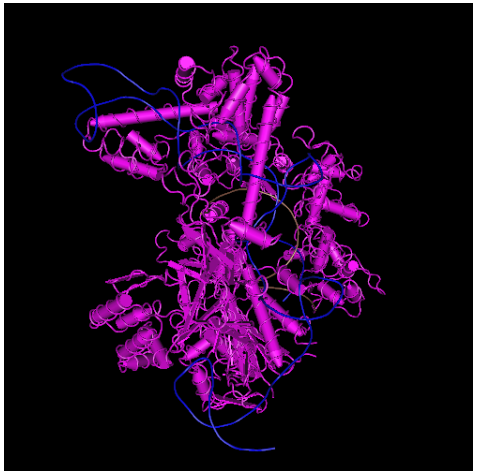
The values are molecules counted at each gene interval.

Physically editing DNA sequences with CRISPR/CAS9

CRISPR/Cas9



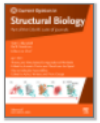
STRUCTURE:



ELSEVIER

Current Opinion in Structural Biology

Volume 69, August 2021, Pages 86-98



Engineering Cas9 for human genome editing

Ian M. Slaymaker, Nicole M. Gaudelli

Show more

https://doi.org/10.1016/j.sbi.2021.03.004

Questions:

How do you know which genes to edit?

How do you get the CAS9-RNA complex to cells?

OUTPUT: SAM/BAM Alignment File

1.1 An example

Suppose we have the following alignment with bases in lower cases clipped from the alignment. Read r001/1 and r001/2 constitute a read pair; r003 is a chimeric read; r004 represents a split alignment.

```
Coor     12345678901234  56789012345678901234567890123456789012345
ref      AGCATGTTAGATAA**GATAGCTGTGCTAGTAGGCAGTCAGCGCCAT

+r001/1        TTAGATAAAGGATA*CTG
+r002         aaaAGATAA*GGATA
+r003       gcctaAGCTAA
+r004                      ATAGCT..............TCAGC
-r003                            ttagctTAGGC
-r001/2                                        CAGCGGCAT
```

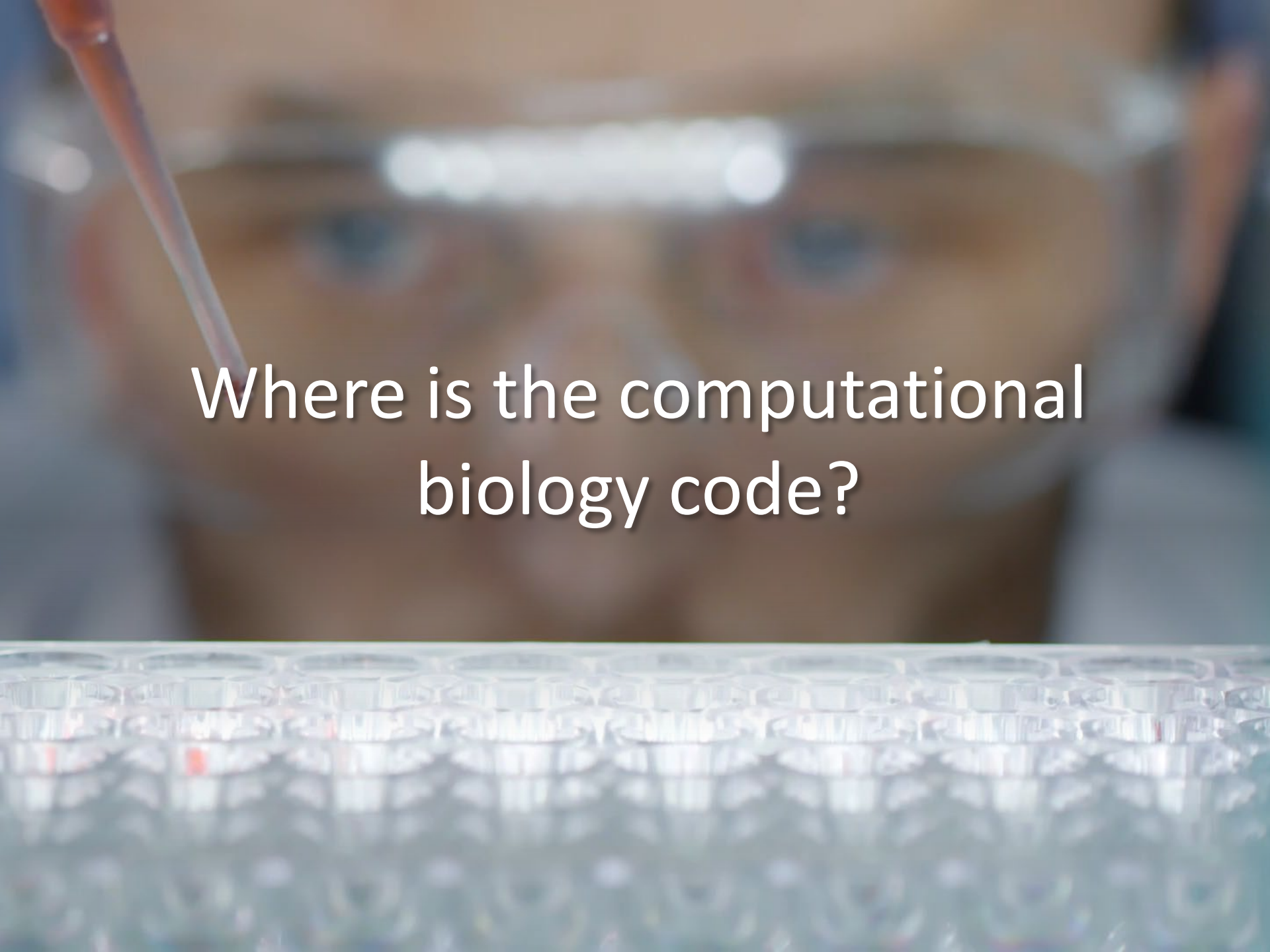
The corresponding SAM format is:

```
@HD VN:1.5 SO:coordinate
@SQ SN:ref LN:45
r001   99 ref  7 30 8M2I4M1D3M = 37  39 TTAGATAAAGGATACTG *
r002    0 ref  9 30 3S6M1P1I4M *  0   0 AAAAGATAAGGATA    *
r003    0 ref  9 30 5S6M       *  0   0 GCCTAAGCTAA       * SA:Z:ref,29,-,6H5M,17,0;
r004    0 ref 16 30 6M14N5M    *  0   0 ATAGCTTCAGC       *
r003 2064 ref 29 17 6H5M       *  0   0 TAGGC             * SA:Z:ref,9,+,5S6M,30,1;
r001  147 ref 37 30 9M         =  7 -39 CAGCGGCAT         * NM:i:1
```

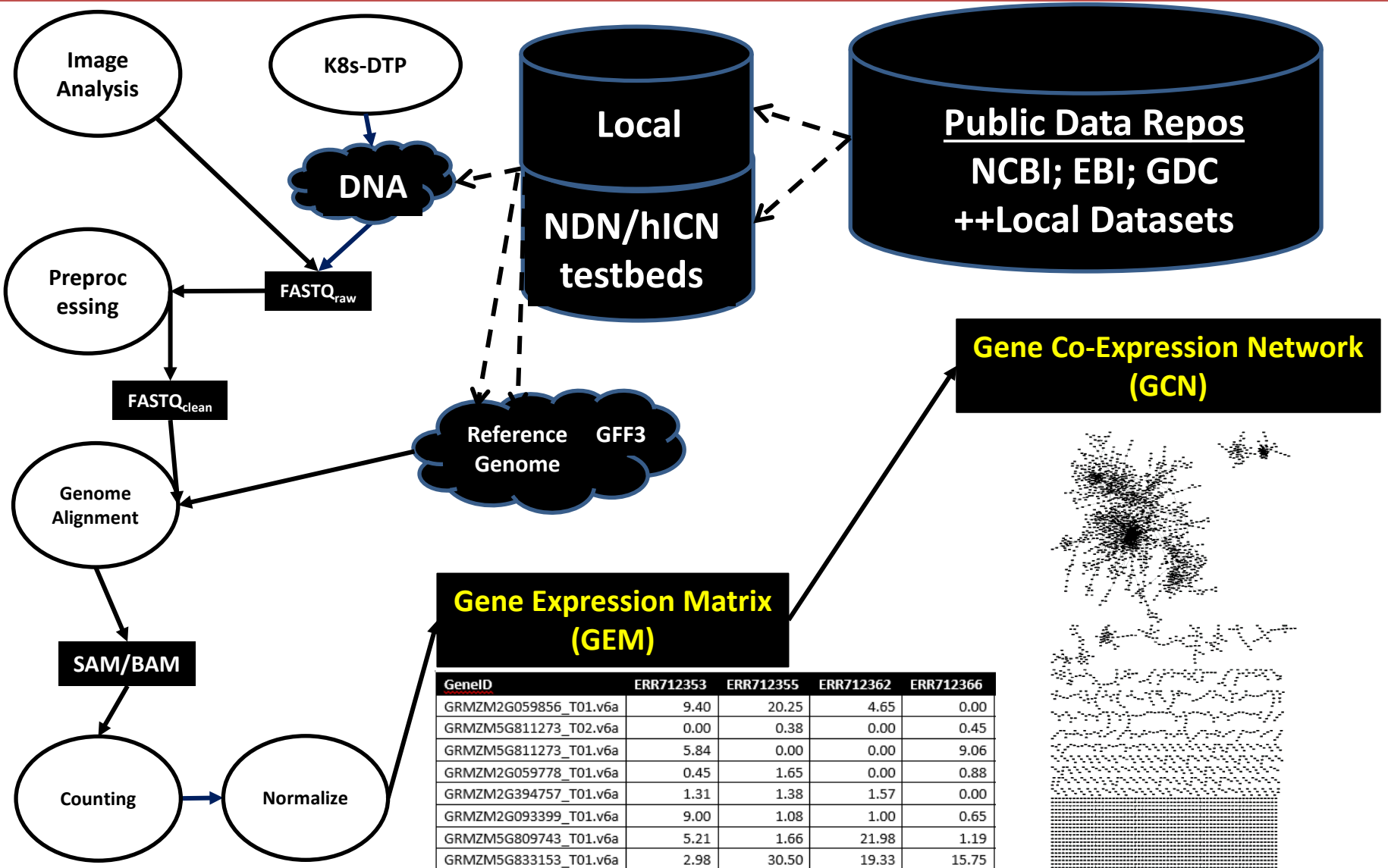
MILLIONS OF ALIGNMENTS!

SAM = ASCII
BAM = Binary

SPECIFICATION::: http://samtools.github.io/hts-specs/SAMv1.pdf


Where is the computational biology code?

Computational Workflows Are Part of the Modern Biology Lab

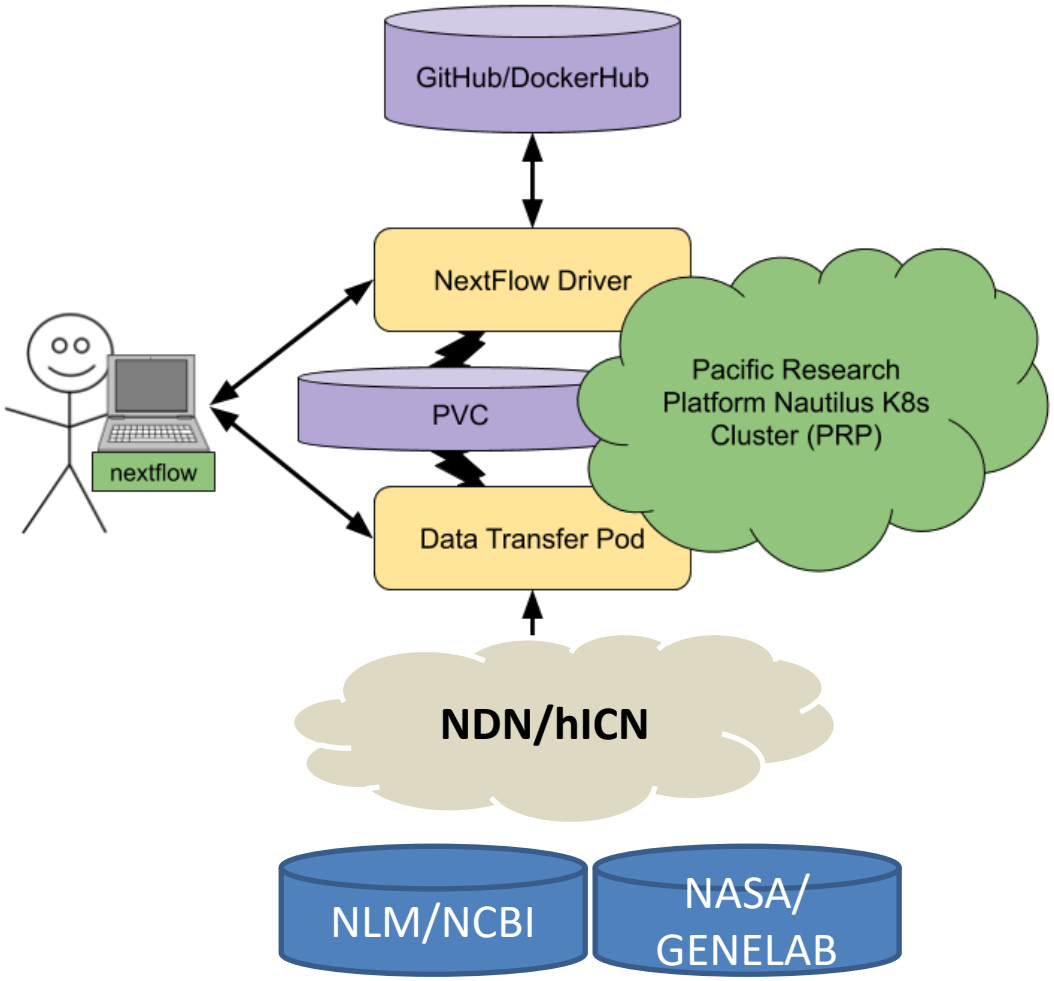


GeneID	ERR712353	ERR712355	ERR712362	ERR712366
GRMZM2G059856_T01.v6a	9.40	20.25	4.65	0.00
GRMZM5G811273_T02.v6a	0.00	0.38	0.00	0.45
GRMZM5G811273_T01.v6a	5.84	0.00	0.00	9.06
GRMZM2G059778_T01.v6a	0.45	1.65	0.00	0.88
GRMZM2G394757_T01.v6a	1.31	1.38	1.57	0.00
GRMZM2G093399_T01.v6a	9.00	1.08	1.00	0.65
GRMZM5G809743_T01.v6a	5.21	1.66	21.98	1.19
GRMZM5G833153_T01.v6a	2.98	30.50	19.33	15.75
AC177838.2_FGT012.v6a	0.29	0.24	0.00	0.00

nextflow Galaxy Pegasus KNC GEMmaker

We Wrap Our Containerized Genomic Workflows in NextFlow

Nextflow is a workflow manager that enables the streamlined execution of multiple steps in a scientific workflow. Users can simply change a configuration file to modify a workflow.



Workflows:

- SystemsGenetics/KINC-nf
- SystemsGenetics/GEMmaker
- SystemsGenetics/GEMprep
- SystemsGenetics/gene-oracle

Additional Tools:

- SystemsGenetics/kube-runner
- SciDAS/nextflow-api

Kubernetes Platforms:

- **Pacific Research Platform (PRP)**
- **Cisco Container Platform (CCP)**
- **TACC Rodeo**
- **Google Cloud Platform (GCP)**

Data-intensive compute (DIC) workflow development is impossible in the commercial cloud for small labs. Democratized DIC systems are CRITICAL.

nextflow https://www.nextflow.io/

GEM Construction with GEMmaker Containerized Workflow

https://github.com/SystemsGenetics/GEMmaker

GEMmaker is a Nextflow workflow for large-scale gene expression sample processing, expression-level quantification and Gene Expression Matrix (GEM) construction. Results from GEMmaker are useful for differential gene expression (DGE) and gene co-expression network (GCN) analyses. The GEMmaker workflow currently supports Illumina RNA-seq datasets.

GEMmaker uses the following tools:

- python3 v3.5.1
- nextflow v0.32
- sratoolkit v2.9.2
- fastQC v0.11.7
- trimmomatic v0.38
- hisat2 v2.1.0
- samtools v1.3.1
- stringTie v1.3.4d
- MultiQC v1.5

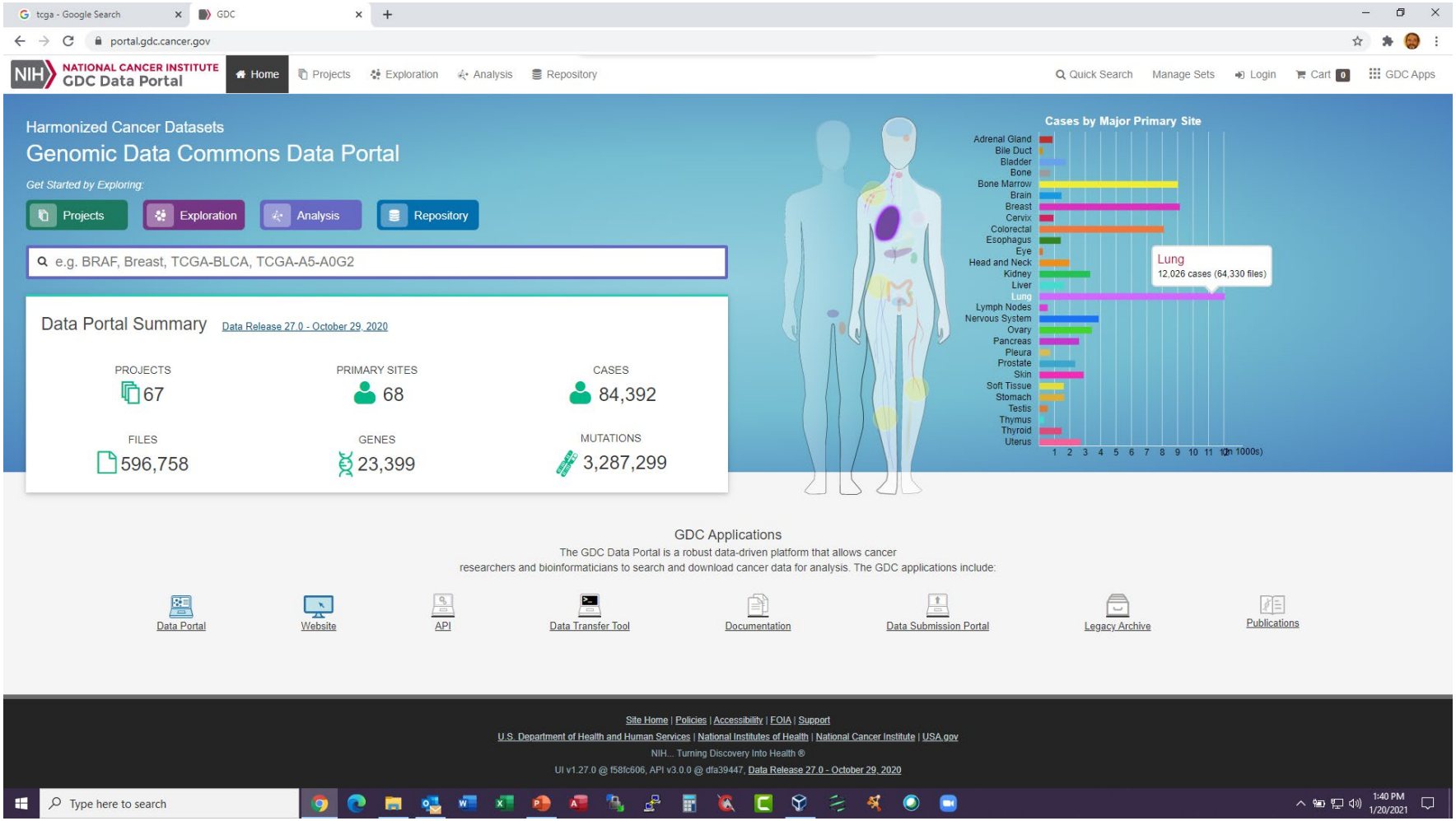
GEMmaker is a collaborative project of the Ficklin and Feltus programs at Washington State University and Clemson University respectively with guidance from RENCI.

GEMmaker is funded by the NSF SciDAS project, award #1659300


Where is the data?

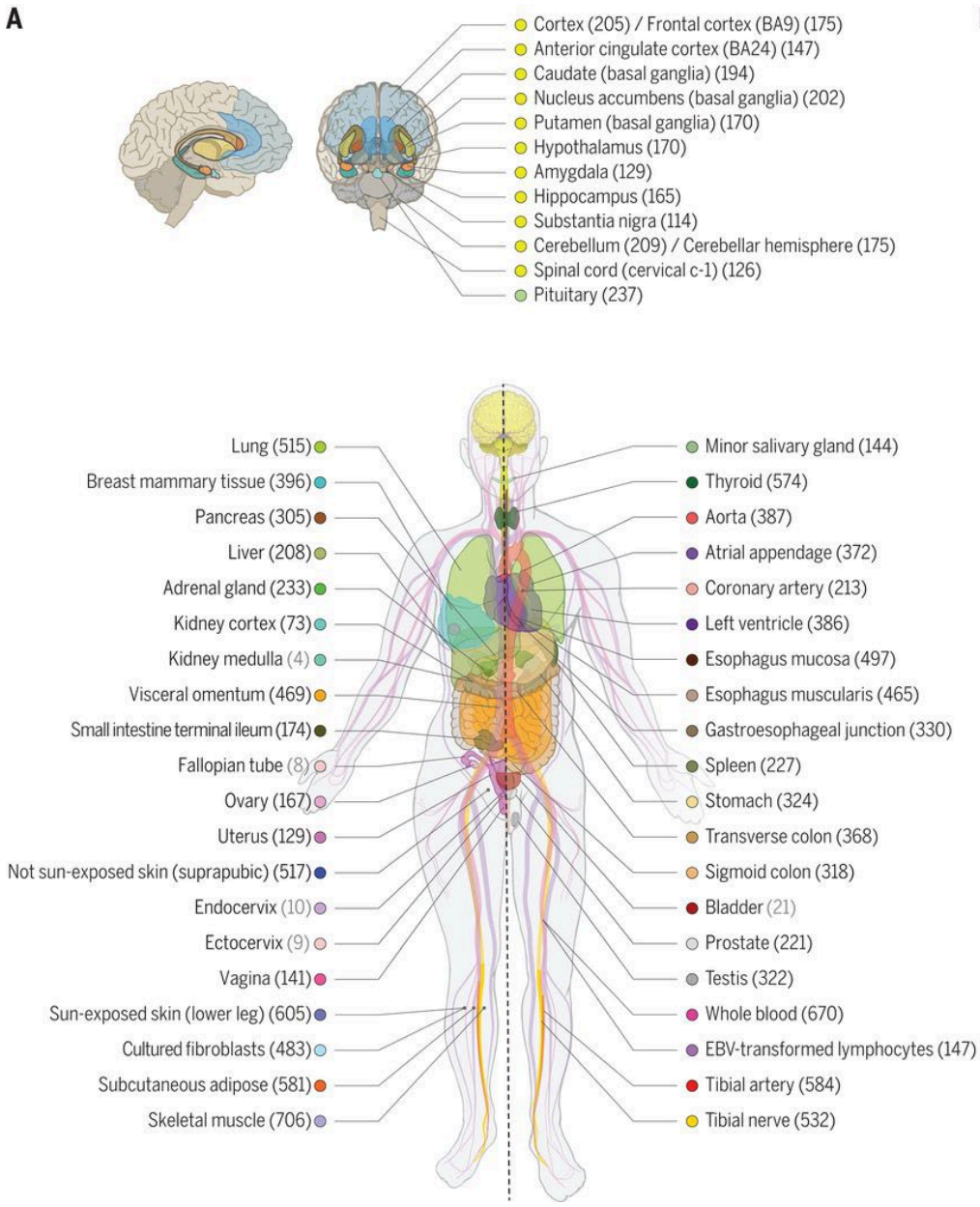
Public Data Source: NCI Genome Data Commons

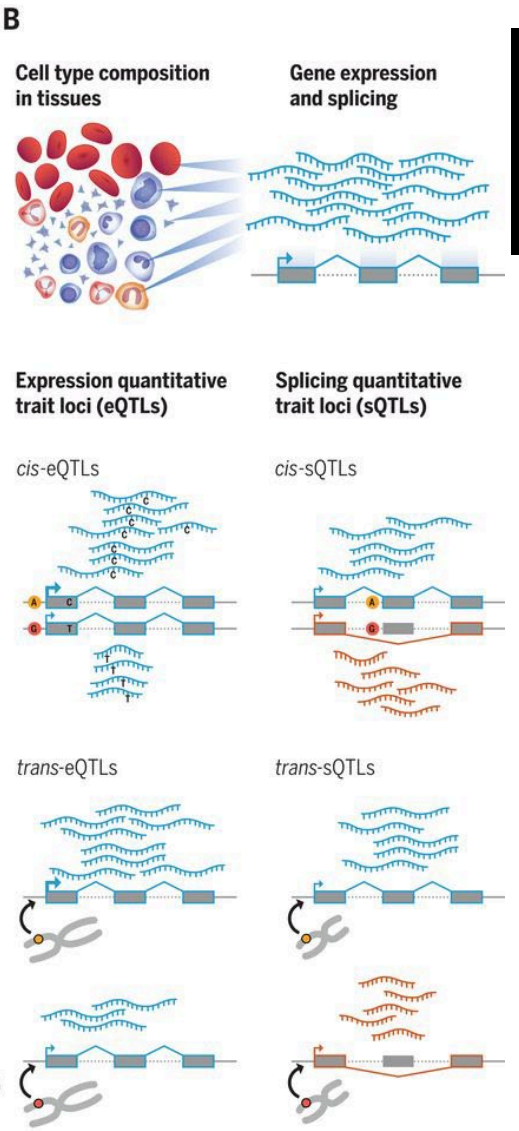
11,315 Samples; 33 Tumor Sub-Types + “Solid Tissue Normal”; 10,610 Pubmed Hits



Source: https://portal.gdc.cancer.gov/

Public Data Source: Genotype-Tissue Expression (GTEx) Project





Release V8
838 donors
49 tissues
15,201 RNAseq

The GTEx Consortium atlas of genetic regulatory effects across human tissues

The GTEx Consortium*
+ See all authors and affiliations

Science 11 Sep 2020:
Vol. 369, Issue 6509, pp. 1318-1330
DOI: 10.1126/science.aaz1776

Source: https://gtexportal.org/


What do you do
with the data?

Example Publication

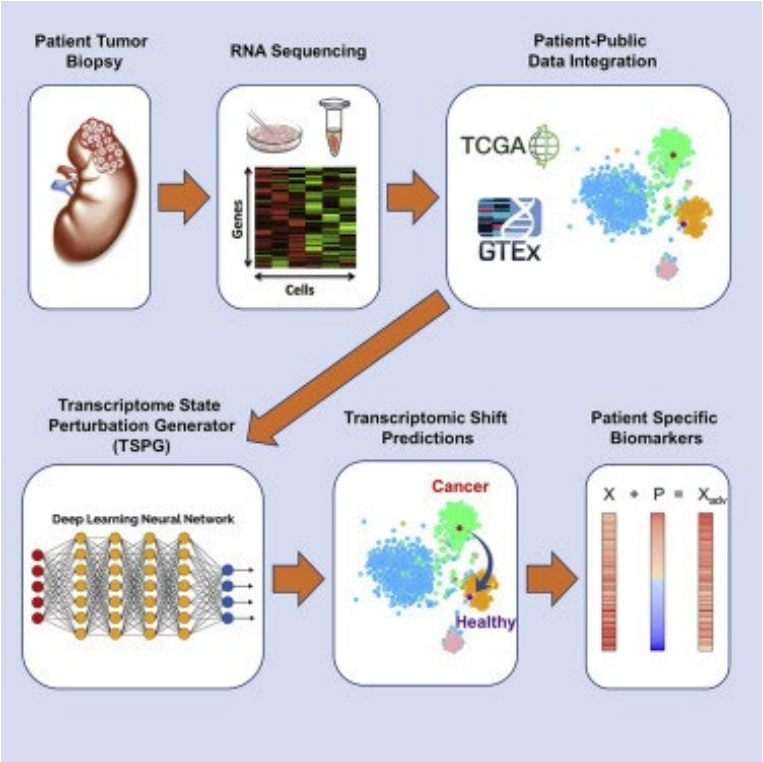
Patterns

Volume 1, Issue 6, 11 September 2020, 100087

Article

Cellular State Transformations Using Deep Learning for Precision Medicine Applications

Colin Targonski [1, 6], M. Reed Bender [2, 6], Benjamin T. Shealy [1], Benafsh Husain [2], Bill Paseman [3], Melissa C. Smith [1], F. Alex Feltus [2, 4, 5, 7]

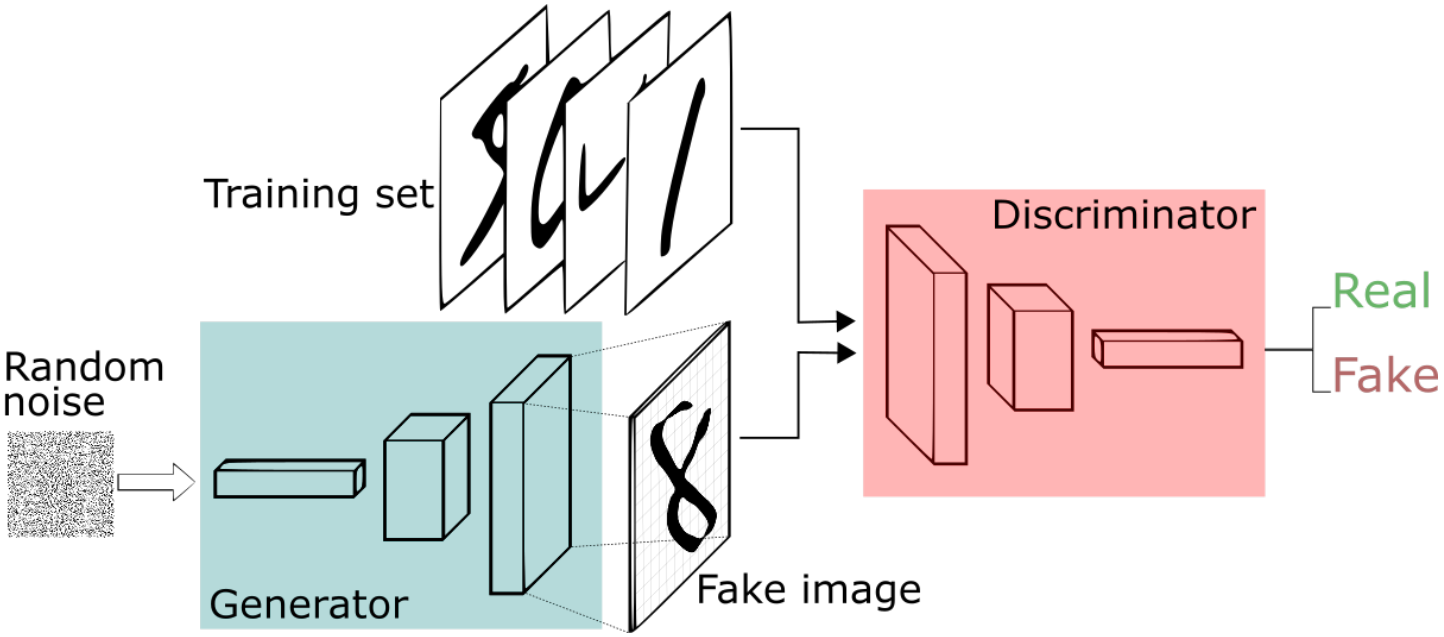


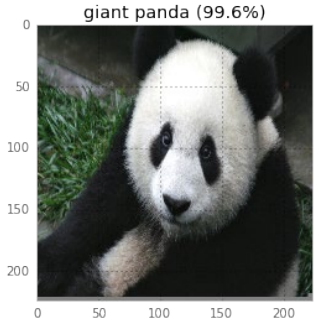
https://github.com/ctargon/TSPG

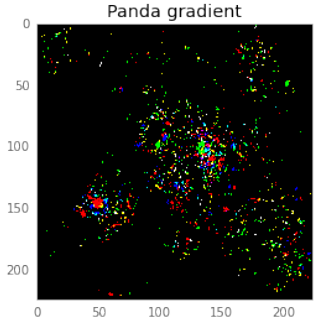
Generative Adversarial Networks: Computer Science Meets Biology

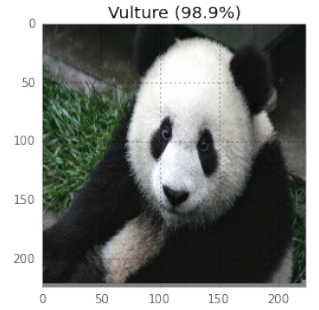
Generative adversarial networks (GANs) are deep neural net architectures comprised of two nets, pitting one against the other (thus the "adversarial").

"GANs' potential is huge, because they can learn to mimic any distribution of data."







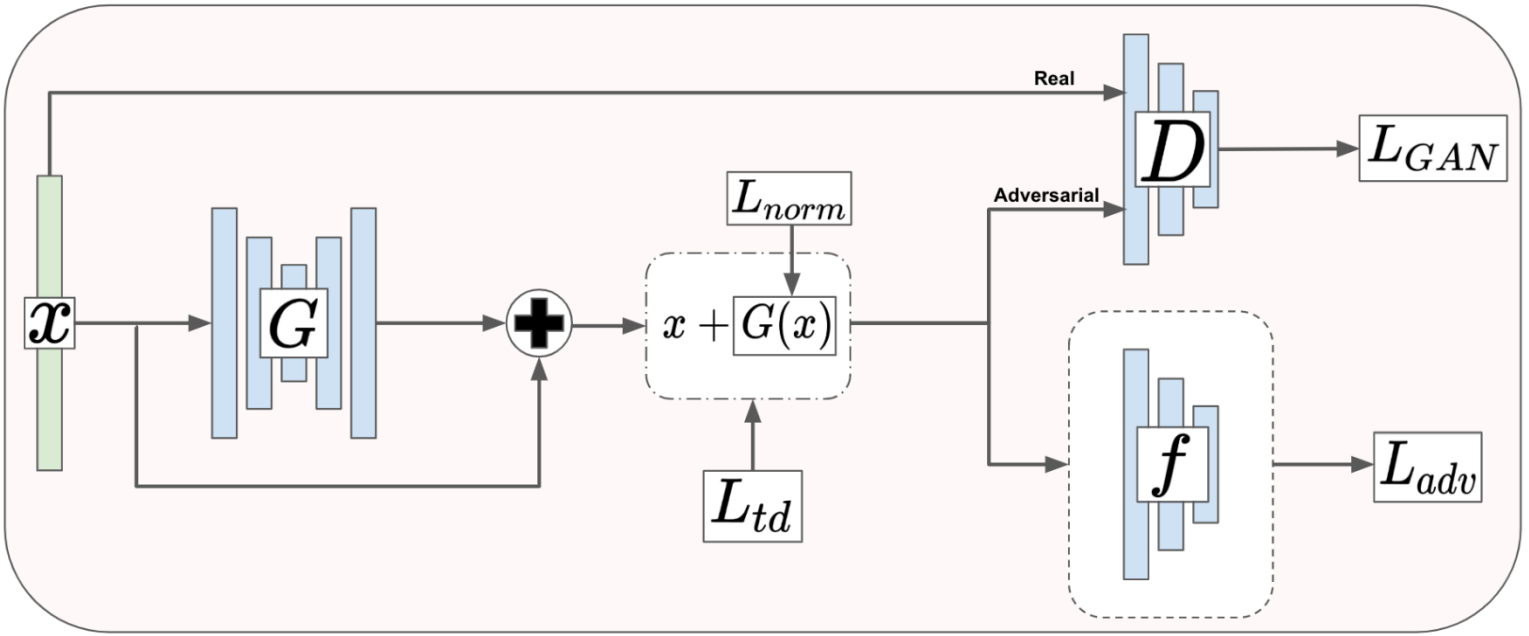


https://codewords.recurse.com/issues/five/why-do-neural-networks-think-a-panda-is-a-vulture

https://skymind.ai/wiki/generative-adversarial-network-gan

The Transcriptome State Perturbation Generator (TSPG)

Goal: Learn to produce x_{adv} classified as $f(x_{adv}) = t$ where t is a target class (e.g. tissue A) different from ground truth label y (e.g. tissue B).



G=generator network
D=discriminator
X=features (genes)
f=pretrained target network
G(x)=feature perturbation
$X_{adv} = X + G(x)$ = adversarial example
$L = L_{GAN} + L_{adv} + L_{norm} + L_{td=}$ Loss for G

Accuracy of Target Network on Perturbed GTEx Datasets

Gene Set	Genes	Target Class	f Accuracy (%)
Hallmark Hedgehog Signaling	36	nerve-tibial	100
Hallmark Peroxisome	107	brain-spinal cord	100
Hallmark Apoptosis	161	lung	99.9
Hallmark E2F Targets	200	artery-coronary	99.8
Hallmark All	4386	thyroid	100
Hallmark All	4386	heart-left ventricle	100

Tricking the Hippocampus to be a Tibial Nerve

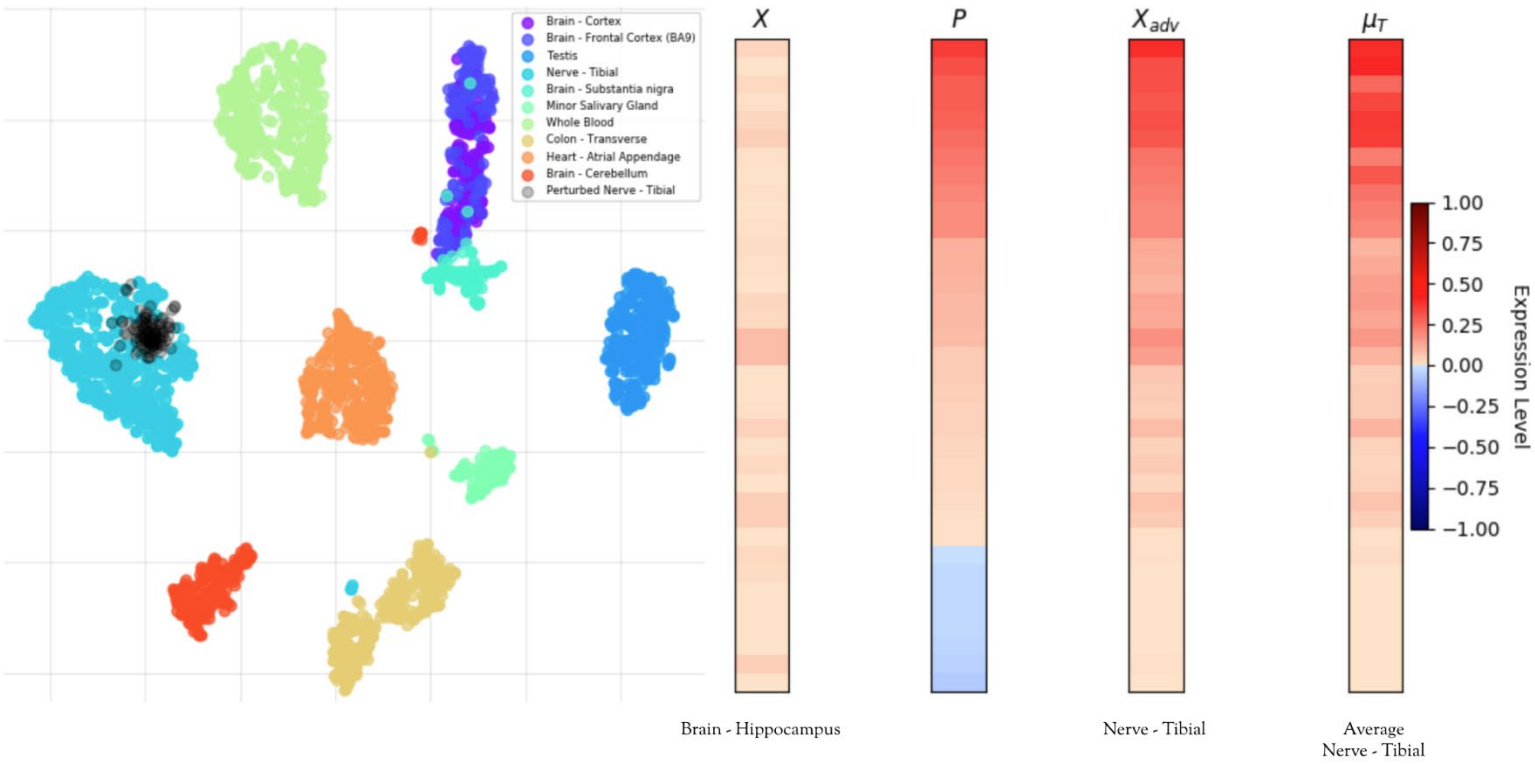


Figure 1: **Adversarial generation for Nerve-Tibial target using the Hallmark Hedgehog Signaling gene set.** t-SNE plot of original and perturbed samples using the Hallmark Hedgehog Signaling gene set (left). Heatmap of cellular transformation from Brain-Hippocampus to Nerve-Tibial (right). Perturbation (P) ranges from $[-1, 1]$, which is added to original sample (X), then adversarial example (x_{adv}) is clipped to $[0, 1]$. The mean expression vector (μ_T) of the target class (Nerve-Tibial) is shown.

Patient-Specific Tumor Analysis

CellPress
OPEN ACCESS

Patterns
Article

Figure 4. Computed Tomography Image of Patient BP's Renal Cell Carcinoma
Imaged in February of 2014, this thoracic computed tomogram with intravenous contrast shows a 5.7 × 4.8-cm mass in the left kidney.

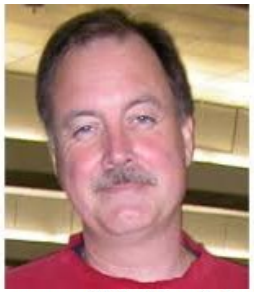


Table 3. Sample Counts Included in the Comprehensive Kidney Cancer GEM

Tissue Type	Dataset of Origin	Count
KIRC tumor	TCGA	475
KIRC normal	TCGA	72
KIRP tumor	TCGA	236
KIRP normal	TCGA	29
KICH tumor	TCGA	60
KICH normal	TCGA	25
Kidney normal	GTEx	32
BP tumor	–	1
BP normal	–	1

Patterns
Volume 1, Issue 6, 11 September 2020, 100087

CellPress

Article

Cellular State Transformations Using Deep Learning for Precision Medicine Applications

Colin Targonski [1, 6], M. Reed Bender [2, 6], Benjamin T. Shealy [1], Benafsh Husain [1], Bill Paseman [3], Melissa C. Smith [1], F. Alex Feltus [2, 4, 5, 7]

RareKidneyCancer.org

TSPG Discovers Kidney Cancer Biomarkers

Patterns
Article

CellPress
OPEN ACCESS

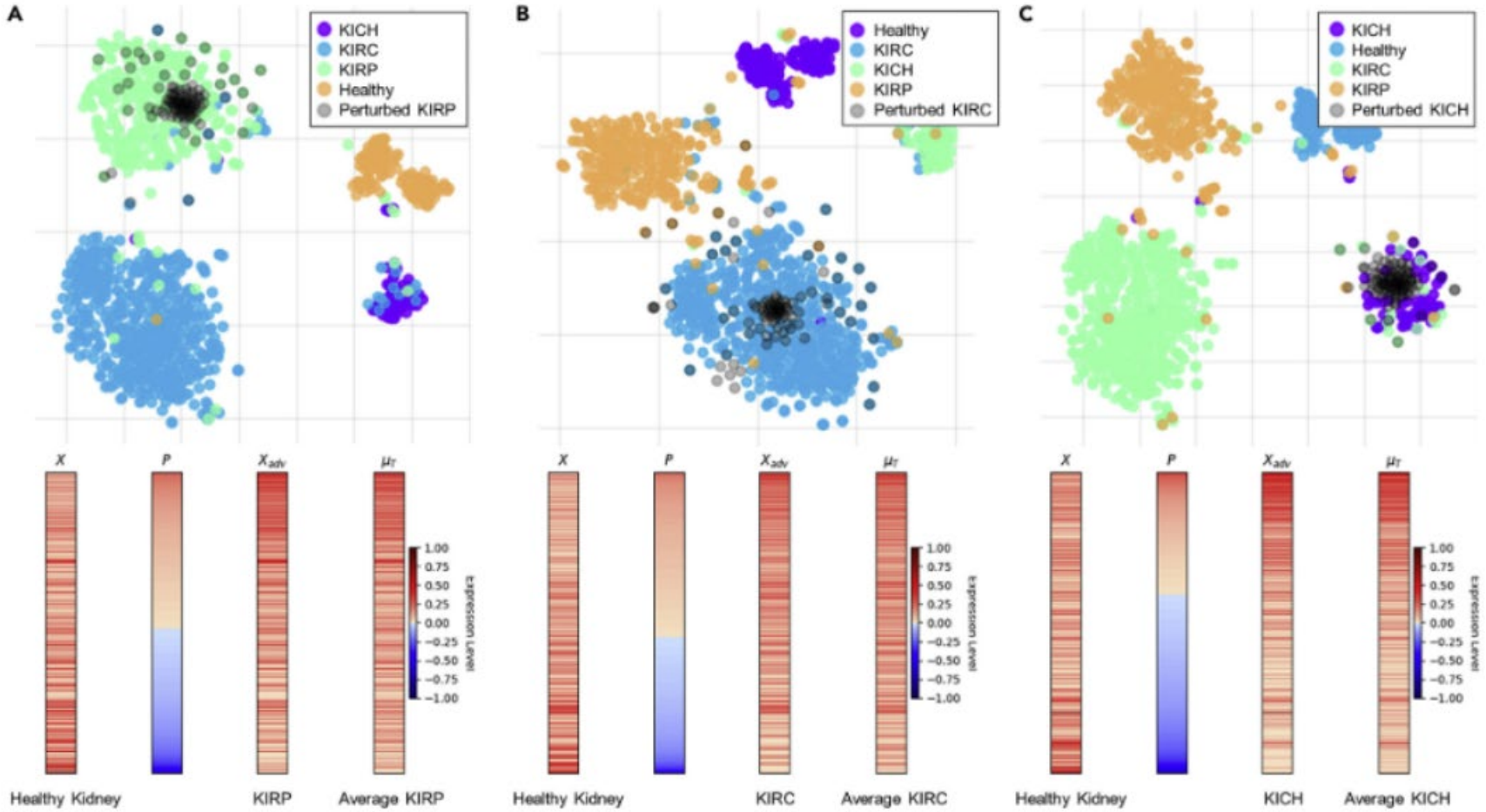


Figure 3. Adversarial Generation for Three Subtypes of Kidney Cancer Using All Hallmark Genes as the Input Gene Set

t-SNE plot and corresponding heatmap of cellular transformation from healthy to KIRP (A), healthy to KIRC (B), and healthy to KICH (C). Perturbations (P) range from $[-1, 1]$, which is added to original sample (X), then adversarial sample (X_{adv}) is clipped to $[0, 1]$. The mean expression vector (μ_T) of the target class is shown.

Geographically Distributed Interdisciplinary Science is Super Fun!

Feltus Lab
Yueyao Gao (<PhD, G&B)
Cole McKnight (Cloud Architect)
Rini Pauly (<PhD, G&B)
Kaitlyn Gmitro (<BSc, G&B)
Meghan Mobley (<BSc, G&B)
Brad Selee (<BSc, ECE)
Devin Keane (<MSc, G&B)

Recent alumni
Nicole Nelligan (BSc, G&B)
Ben Shealy (PhD, ECE)
Allison Hickman (PhD, G&B)
Yuqing Hang (PhD, G&B)
Reed Bender (MSc, BioE)
Ethan Bensman (BSc, CS)
Cameron Ogle (BSc, CS)
Benafsh Husain (PhD, BDSI)
Will Poehlman (PhD, G&B)
Jordan Little (BSc, G&B)
Rachel Eimen (Bsc, ECE)
Courtney Shearer (BSc, CS)
Melissa Judge (BSc, BioE)
Colin Targonski (MSc, , ECE)
Leland Dunwoodie (<MD, G&B)
Olivia Feltus (<BSc, Intern)
Nick Watts (Software Engineer)
Zach Gerstner (MS, BDSI)
Cole Younginer (<MSc, ECE)

@ Clemson
Melissa Smith (ECE)
KC Wang (ECE)
Walt Ligon (ECE)
Nick Mills (ECE)
Jon Calhoun (ECE)
Brian Dean (CS)
Marc Birtwistle (ChemE)
Brian Booth (BioE)
Julia Frugoli (G&B)
Suchitra Chavan (G&B)
Elise Schnabel (G&B)
Susan Duckett (AVS)
Corey Ferrier (CCIT)
Jim Pepin (CCIT)
Xizhou Feng (CCIT)

@ Earth
Stephen Ficklin (WSU)
Josh Burns (WSU)
Tyler Biggs (WSU)
Dorrie Main (WSU)
Sook Jung (WSU)
Joe Breen (Utah)
Jill Wegrzyn (UCONN)
Meg Staton (UTK)
Jim Bottum (Internet2)
Dana Brunson (Internet2)
John Hicks (Internet2)
Marvin Weinstein(QI) Ken Matusow (Synergity)
Karan Sapra (nVidia)
David Clarke (Toolwire)
Jacob Loftis (Cisco)
Mike Shepherd (Cisco)
Mike Kowal (Cisco)
Jordan Auge (Cisco)

@ Earth (cont.)
Mats Rynge (USC-OSG)
Bala Desinghu (U Chicago-OSG)
Andrew Paterson (UGA)
Claris Castillo (RENCI)
Steve Cox (RENCI)
Ray Idaszak (RENCI)
Paul Ruth (RENCI)
Michael Stealy (RENCI)
Isma Gilani (RENCI)
Fan Jiang (RENCI)
Mert Cevik (RENCI)
Ananya Mukherghee (RENCI)
Emily Casanova (USC-GHS)
Manual Casanova (USC-GHS)
Alex Bowers (Columbia U.)
Josh Vandenbrink (Louisiana Tech)
Ann Loraine (UNCC)
Colleen Doherty (NCSU)
John Graham (UCSD)
Wallace Chase (REANNZ)
Christos Papadopoulos (Memphis)
Susmit Shannigrahi (Tennessee Tech)
Chengyu Fan (CSU)

Many many more!

Thank You Funding Agencies!!!!!

- ***"CC* Integration-Large: Prototyping a Secure Distributed Storage Infrastructure for Accelerating Big Science." [2126148] (S. Shannigrahi PI)***
- ***"CC*Data: National Cyberinfrastructure for Scientific Data Analysis at Scale (SciDAS) Source: NSF-CC* [1659300] (A. Feltus PI)***
- ***"RCN: Advancing Research and Education Through a National Network of Campus Research Computing Infrastructures - The CaRC Consortium" Source: NSF [1620695] (J Bottum PI > A. Feltus PI)***
- ***"MRI: Acquisition of a Cyberinstrument for Interdisciplinary Computational Science and Engineering." Source: NSF-MRI [1725573] (A. Apon PI)***

<u>Historical</u>

- *"MCA-PGR: Spatial and Temporal Resolution of mRNA Profiles During Early Nodule Development." Source: NSF-PGRP [1444461] (J. Frugoli PI)*
- *"Tripal Gateway: Platform for Next-Generation Data Analysis and Sharing."*

Source: NSF-DIBBS [1443040] (S. Ficklin, PI)

- *"BIGDATA: F: DKM: Collaborative Research: PXFS: ParalleX Based Transformative I/O System for Big Data"*

Source: NSF-BIGDATA [1447771] (W. Ligon PI)

- *"Genomic and Breeding Foundations for Bioenergy Sorghum Hybrids."*

Source: Plant Feedstock Genomics for Bioenergy [DE-FOA-000041] (S Kresovich, PI).

- *"Big Data Visualization REU".*

Source: National Science Foundation [1359223](V Byrd, PI)

- *"MRI: Acquisition of a High Performance Computing Instrument for Collaborative Data-Enabled Science."* Source: National Science Foundation [1228312] (A Apon, PI)
- *"CC-NIE Integration: Clemson-NextNet"*

Source: National Science Foundation [1245936] (KC Wang, PI)

- *"Building non-model species genome curation communities."*

Source: National Evolutionary Synthesis Center (NESCent) (A Papanicolaou, PI)

- *"Big Data Analysis Tools for Agricultural Genomics."*

Source: Clemson University Experiment Station (USDA Hatch Project) [SC-1700492] (A. Feltus PI).