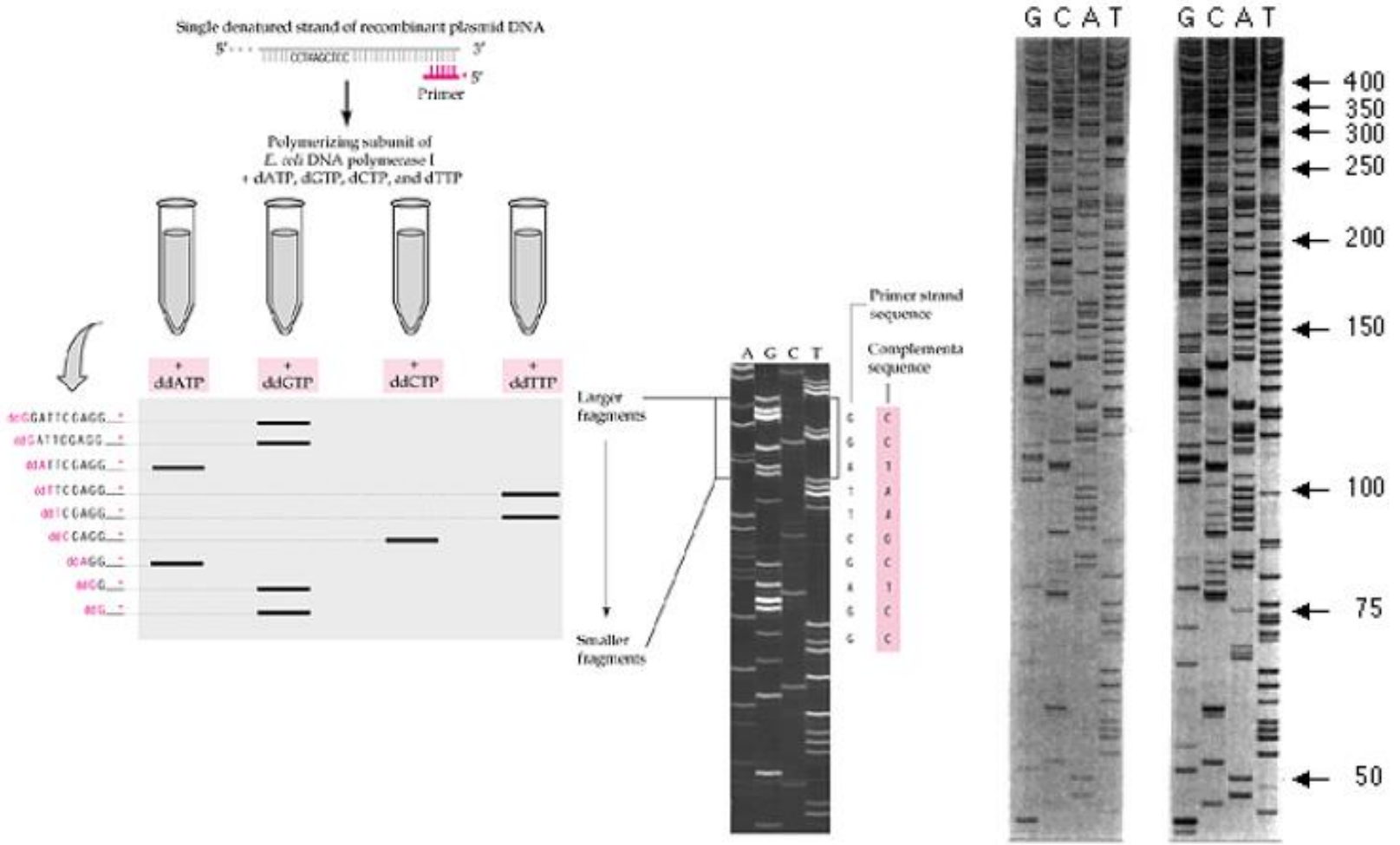
MAFtools

Visualize DNA mutations

Sanger Dideoxy Sequencing

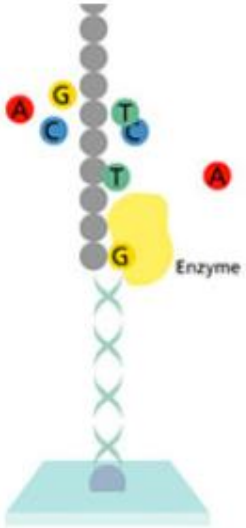


Sanger, F. et al. Nature 24, 687–695 (1977).
Sanger, F., Nicklen, S. & Coulson, A.R. Proc. Natl. Acad. Sci. USA 74, 5463–5467 (1977).

Sequencing length: ~1 kb

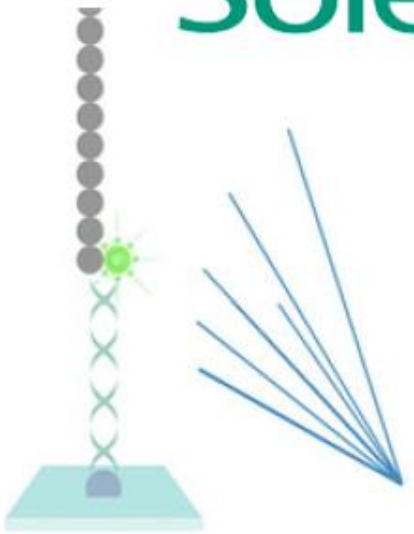
Solexa™

Cycle 1



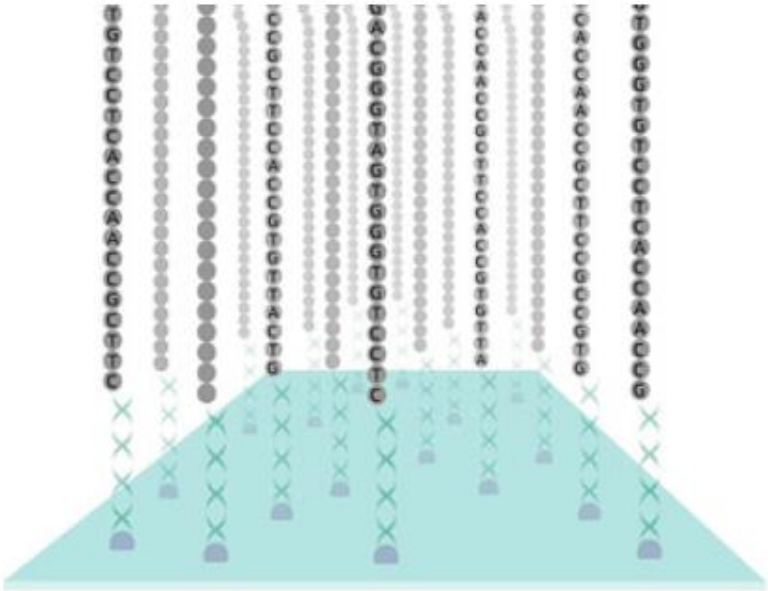
Incorporation

Cycle 1

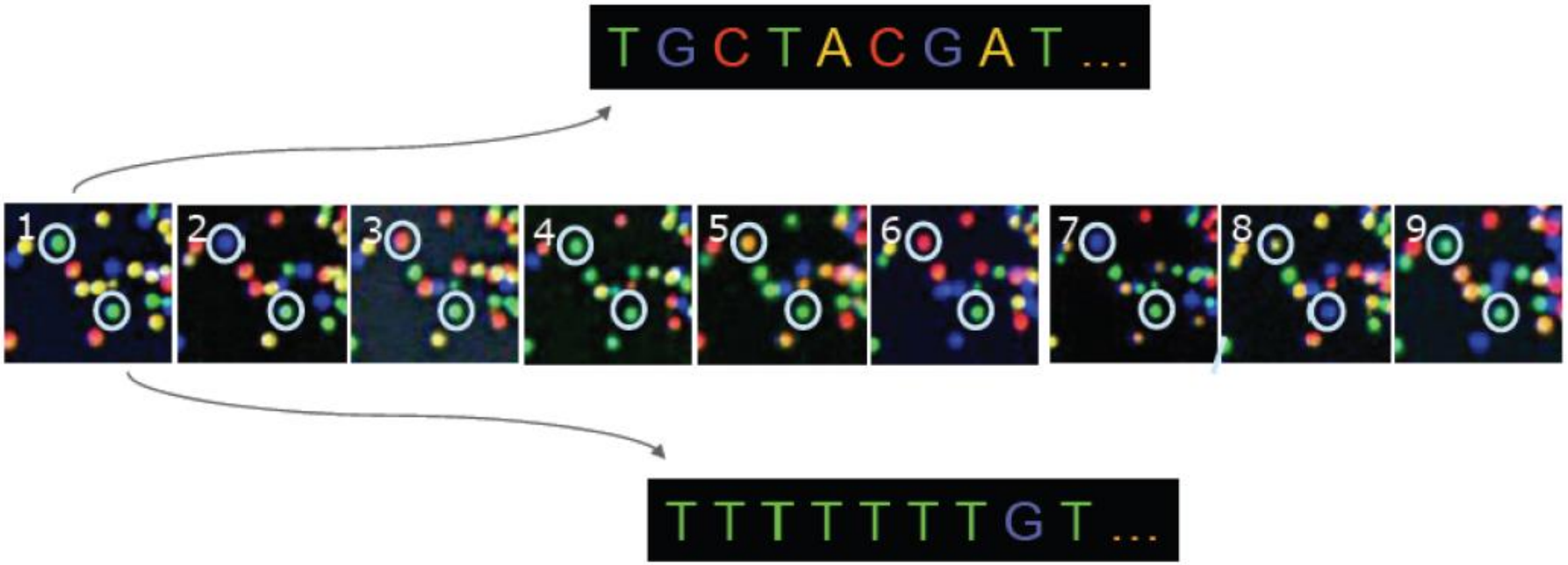


Fluorescent Detection

Sequence Generated At Every Site On The Array



Sequencing-by-synthesis



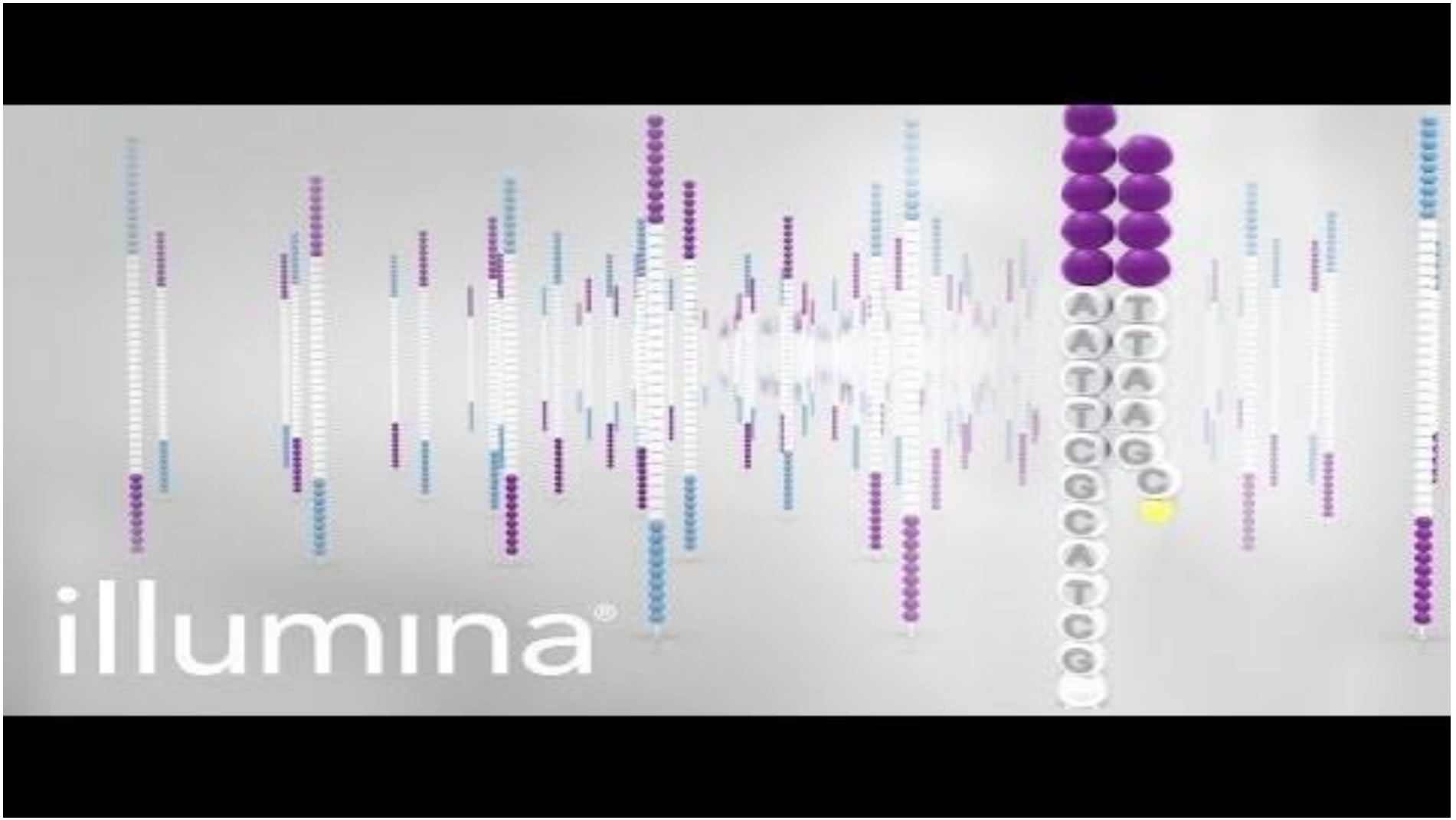
The identity of each base of a cluster is read off from sequential images

From Debbie Nickerson, Department of Genome Sciences, University of Washington. http://tinyurl.com/6zbzh4

Sequencing length: 0.05~0.2 kb


AAACCGTCTTGCGAGAAAACCGTAAGTCA
AAACCGTC
AACCGTCTTGC
ACCGTCTTGCGA
GTCTTGCGAGAA
GAAAACCGTAA
AAACCGTAAGTCA
Read

https://www.youtube.com/watch?v=fCd6B5HRaZ8

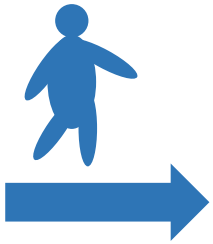



Contig Editor: -693189 SRR006330.210784
Cons 2
Qual 0
Insert
Edit Modes >>
Cutoffs
Undo
Next Search
Commands >>
Settings >>
Quit
Help >>
<< < > >>
70 96780 96790 96800 96810 96820 96830 96840 968
+798891 SRR006330.2238 ACAGGCGGGCACCTTGCTGGGCTG
+798905 SRR006330.3559 ACAGGCGGGCACCTTGCTGGGCTGCAAACAAA
-798957 SRR006330.4248 ACAGGCGGGCACCTTGCTGGGCTGCAAACAAACTGATCTTGTCACTGGC
-798972 SRR006330.3045 ACAGGCGGGCACCTTGCTGGGCTGCAAACAAACTGATCTTGTCACTGGCGTGTGCAATCAAAC*TTTTT
-799004 SRR006330.2510 ACAGGCGGGCACCTTGCTGGGCTGCAAACAAACTGATCTTGTCACTGGCGTGTGCAATCAAAC*TTTTTGCGTGCTTATG
-799008 SRR006330.2564 ACAGGCGGGCACCTTGCTGGGCTGCAAACAAACTGATCTTGTCACTGGCGTGTGCAATCAAACTTTTTTGCGTGCTTATG
-799017 SRR006330.3770 ACAGGCGGGCACCTTGCTGGGCTGCAAACAAACTGATCTTGTCACTGGCGTGTGCAATCAA
+799022 SRR006465.7152 ACAGGCGGGCACCTTGCTGGG
-799034 SRR006330.1830 ACAGGCGGGCACCTTGCTGGGCTGCAAACAAACTGATCTTGTCACTGGCGTGTGCAATCAAACTTTTTTGCGTGCTTATG
-799050 SRR006330.2574 ACAGGCGGGCACCTTGCTGGGCTGCAAACAAACTGATCTTGTCACTGGCGTGTGCAATCAAAC*TTTTTGCGTGCTTATG
-799069 SRR006330.4899 ACAGGCGGGCACCTTGCTGGGCTGCAAACAAACTGATCTTGTCACTGGCGTGTGCAATCAAA
+799078 SRR006330.4461 ACAGGCGGGCACCTTGCTGGGCTGCAAACAAACTGATCTTGTCACTGGCGTGTGCAATCAAACTTTTTTGCGTGCTTATG
+799081 SRR006330.4193 ACAGGCGGGCACCTTGCTGGGCTGCAAACAAACTGATCTTGTCACTGGCGTGTGCAATCAAACTTTTTTGCGTGCTTATG
-799091 SRR006330.4700 ACAGGCGGGCACCTT-CTGGGCTGCAAACAAACTGATCTTGTCACTGGCGTGTGCAA-CAAAC*TTTTTGCGTGCTTATG
+799096 SRR006332.7495 AC
+799097 SRR006332.4165 AC
-799098 SRR006332.4389 AC
-799099 SRR006330.4357 ACAGGCGGGCACCTT-CTGGGCTGCAAACAAACTGATCTTGTCACTGGCGTGTGCAA-CAAACTTTTTTGCGTGCTTATG
-799101 SRR006332.4943 ACAGGC
-799102 SRR006332.2311 ACAGGCGGGC
+799103 SRR006330.4567 ACAGGCGGGCACCTTGCTGGGCTGCAAACAAACTGATCTTGTCACTGGCGTGTGCAATCAAACTTTTTTG
+799104 SRR006332.8325 ACAG
+799105 SRR006330.7286 ACAGGCGGGCACCTTGCTGGGCTGCAAACAAACTGATCTTGTCACTGGCGTGTGCAATCAAACTTTTTTGCGTGCTTATG
-799106 SRR006332.2550 ACAGGCGGGCACCT
-799107 SRR006332.1124 ACAGGCGG
+799108 SRR006332.4026 ACAGGCGGGCACCTT
+799109 SRR006332.2982 ACAGGCG
-799110 SRR006332.5993 ACAGGCGGGCACCTTGC
> CONSENSUS -**- ACAGGCGGGCACCTTGCTGGGCTGCAAACAAACTGATCTTGTCACTGGCGTGTGCAATCAAACTTTTTTGCGTGCTTATG
Tag type:SRMr Direction:- Comment:"Strong Repeat Marker base found by MIRA"

DNA sequencing & analysis



Sequencing Machine



Single nucleotide Variation (SNV)

ACTGAAGCTTTGCG…

Human Reference Genome



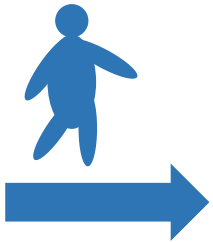
ACTGGAGCTGTGCG…

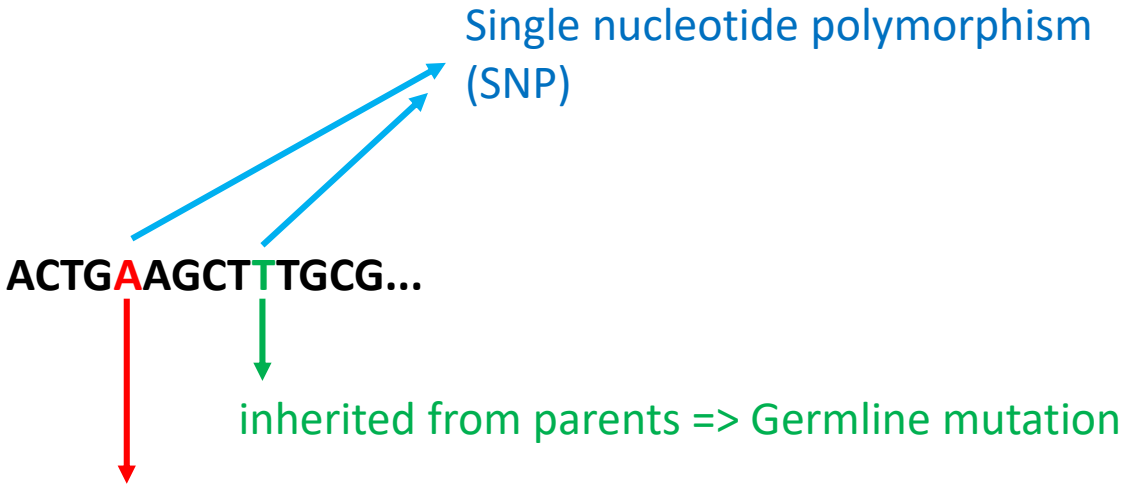
from 13 anonymous volunteers

DNA sequencing & analysis



Sequencing Machine





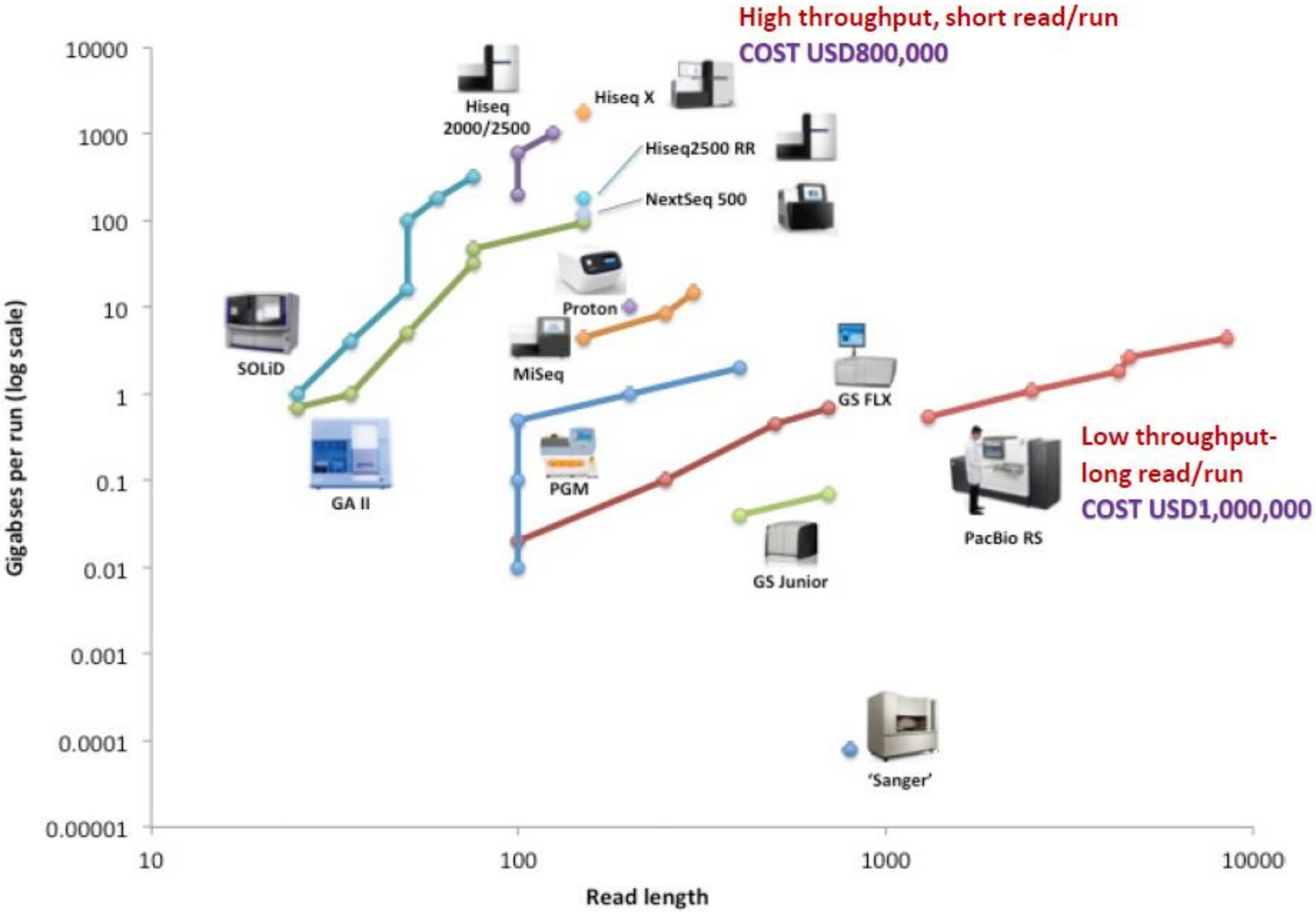
Human R
Genome
(hg19 & h

Whole exome sequencing:

~100 to 1,000 somatic mutations each sample

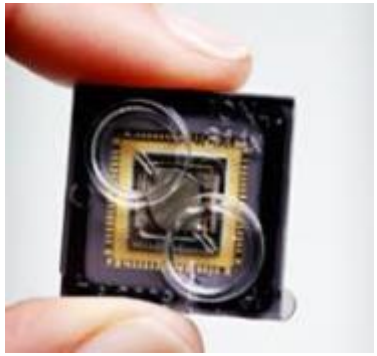
~1,000 to 10,000 germline mutations each sample

Cost of NGS machines



Manufacturers of DNA sequencers

- Roche
- Illumina
- Life Technologies
- Beckman Coulter
- Pacific Biosciences
- Oxford Nanopore

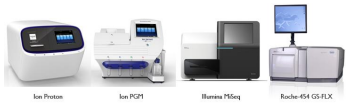




	 MiniSeq System	 MiSeq Series	 NextSeq Series	 HiSeq Series	 HiSeq X Series*
Key Methods	Amplicon, targeted RNA, small RNA, and targeted gene panel sequencing.	Small genome, amplicon, and targeted gene panel sequencing.	Everyday exome, transcriptome, and targeted resequencing.	Production-scale genome, exome, transcriptome sequencing, and more.	Population- and production-scale whole-genome sequencing.
Maximum Output	7.5 Gb	15 Gb	120 Gb	1500 Gb	1800 Gb
Maximum Reads per Run	25 million	25 million†	400 million	5 billion	6 billion
Maximum Read Length	2 × 150 bp	2 × 300 bp	2 × 150 bp	2 × 150 bp	2 × 150 bp
Run Time	4–24 hours	4–55 hours	12–30 hours	<1–3.5 days (HiSeq 3000/HiSeq 4000) 7 hours–6 days (HiSeq 2500)	<3 days

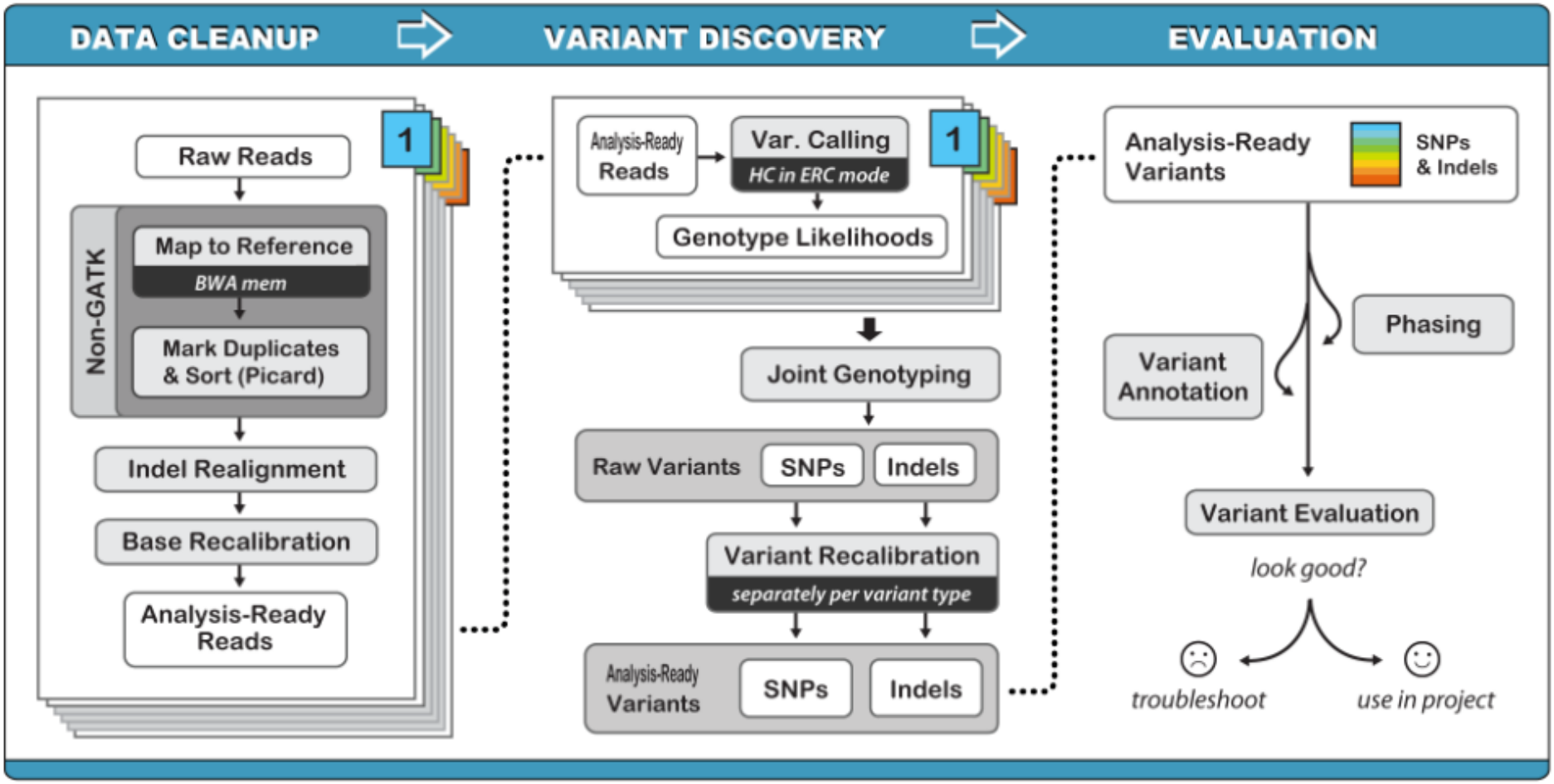
25百萬條
長度為 2x150 bases reads

25,000,000 x 2 x 150 = 7,500,000,000 = 7.5 Gb



Sequencer ----------FASTQ/Sam/Bam/Vcf---------->

Analysis Pipeline



FASTQ/Sam/Bam/Vcf/Maf

@開頭，描述資訊:
- **@SEQ_ID**

序列:
- **GATTTGGGGTTCAAAGCAGTATCGATCAAATAGTAAATCCATTTGTTCAACTCACAGTTT**

+開頭，描述資訊:
- **+**

Quality:
- **!''*((((***+))%%%++)(%%%%).1***-+*''))**55CCF>>>>>>CCCCCCC65**

Quality value 有不同的定義，常使用的 Sanger 分數：

p 代表的是每個base對應的錯誤機率

例如: p=0.01代表錯誤率為百分之一，換算成quality value則為20

$$Q_{\text{sanger}} = -10 \log_{10} p$$

$-10 * \log_{10}0.01 = -10 * (-2) = 20$

FASTQ/Sam/Bam/Vcf/Maf

Sanger/Illumina 1.8 format
can encode a
Phred quality score
from 0 to 93
using ASCII 33 to 126

The ASCII code

American Standard Code for Information Interchange

ASCII control characters

DEC	HEX	Simbolo ASCII	
00	00h	NULL	(carácter nulo)
01	01h	SOH	(inicio encabezado)
02	02h	STX	(inicio texto)
03	03h	ETX	(fin de texto)
04	04h	EOT	(fin transmisión)
05	05h	ENQ	(enquiry)
06	06h	ACK	(acknowledgement)
07	07h	BEL	(timbre)
08	08h	BS	(retroceso)
09	09h	HT	(tab horizontal)
10	0Ah	LF	(salto de linea)
11	0Bh	VT	(tab vertical)
12	0Ch	FF	(form feed)
13	0Dh	CR	(retorno de carro)
14	0Eh	SO	(shift Out)
15	0Fh	SI	(shift In)
16	10h	DLE	(data link escape)
17	11h	DC1	(device control 1)
18	12h	DC2	(device control 2)
19	13h	DC3	(device control 3)
20	14h	DC4	(device control 4)
21	15h	NAK	(negative acknowle.)
22	16h	SYN	(synchronous idle)
23	17h	ETB	(end of trans. block)
24	18h	CAN	(cancel)
25	19h	EM	(end of medium)
26	1Ah	SUB	(substitute)
27	1Bh	ESC	(escape)
28	1Ch	FS	(file separator)
29	1Dh	GS	(group separator)
30	1Eh	RS	(record separator)
31	1Fh	US	(unit separator)
127	20h	DEL	(delete)

ASCII printable characters

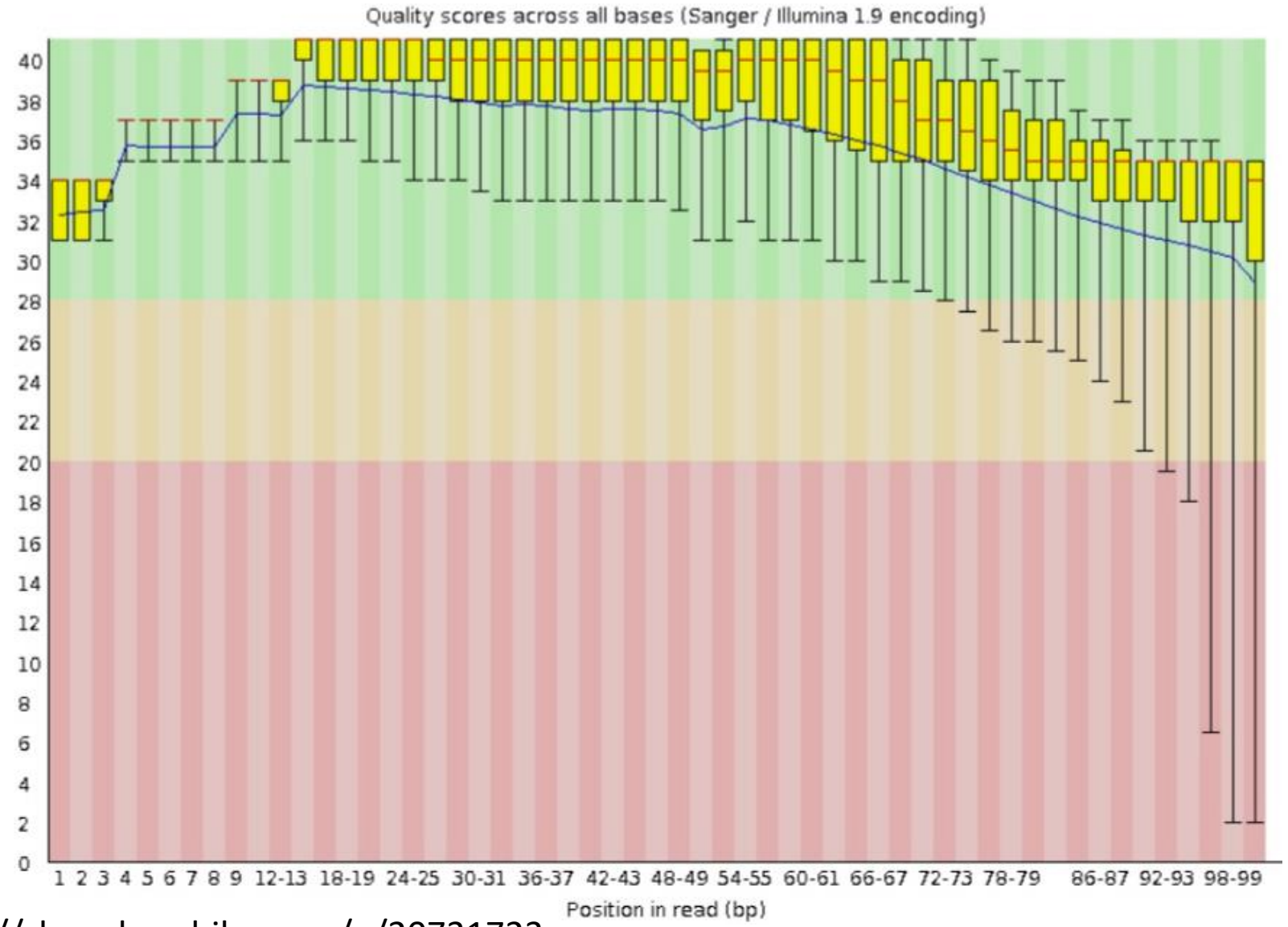
DEC	HEX	Simbolo	DEC	HEX	Simbolo	DEC	HEX	Simbolo
32	20h	espacio	64	40h	@	96	60h	`
33	21h	!	65	41h	A	97	61h	a
34	22h	"	66	42h	B	98	62h	b
35	23h	#	67	43h	C	99	63h	c
36	24h	$	68	44h	D	100	64h	d
37	25h	%	69	45h	E	101	65h	e
38	26h	&	70	46h	F	102	66h	f
39	27h	'	71	47h	G	103	67h	g
40	28h	(	72	48h	H	104	68h	h
41	29h	)	73	49h	I	105	69h	i
42	2Ah	*	74	4Ah	J	106	6Ah	j
43	2Bh	+	75	4Bh	K	107	6Bh	k
44	2Ch	,	76	4Ch	L	108	6Ch	l
45	2Dh	-	77	4Dh	M	109	6Dh	m
46	2Eh	.	78	4Eh	N	110	6Eh	n
47	2Fh	/	79	4Fh	O	111	6Fh	o
48	30h	0	80	50h	P	112	70h	p
49	31h	1	81	51h	Q	113	71h	q
50	32h	2	82	52h	R	114	72h	r
51	33h	3	83	53h	S	115	73h	s
52	34h	4	84	54h	T	116	74h	t
53	35h	5	85	55h	U	117	75h	u
54	36h	6	86	56h	V	118	76h	v
55	37h	7	87	57h	W	119	77h	w
56	38h	8	88	58h	X	120	78h	x
57	39h	9	89	59h	Y	121	79h	y
58	3Ah	:	90	5Ah	Z	122	7Ah	z
59	3Bh	;	91	5Bh	[	123	7Bh	{
60	3Ch	<	92	5Ch	\	124	7Ch	|
61	3Dh	=	93	5Dh	]	125	7Dh	}
62	3Eh	>	94	5Eh	^	126	7Eh	~
63	3Fh	?	95	5Fh	_			

theASCIIcode.com.ar

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 10,000	99.99%
50	1 in 100,000	99.999%
60	1 in 1,000,000	99.9999%

FastQC - quality visualization tool



https://zhuanlan.zhihu.com/p/20731723

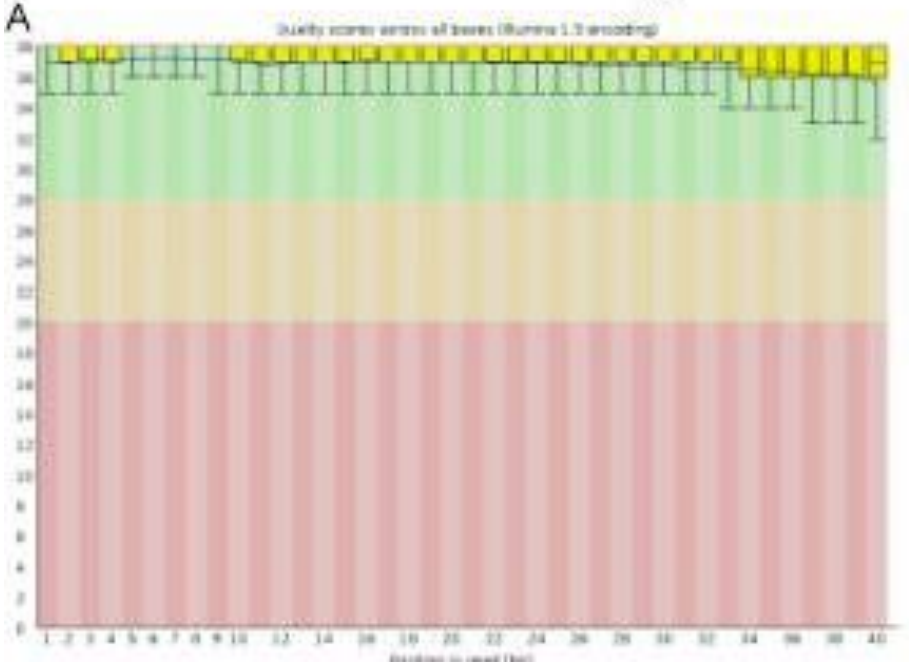
- # X軸 Position in read (bp)
- # Y軸 Q = -10*log10（error P）即20表示1%的錯誤率，30表示0.1%
- # 每一個boxplot，都是該位置的所有序列的測序品質的一個統計，
- 上面的bar是90%分位數，下面的bar是10%分位數，箱子的中間的橫線是50%分位數，箱子的上邊是75%分位數，下邊是25%分位數
- # 圖中藍色的細線是各個位置的平均值的連線

- # 一般要求此圖中，所有位置的10%分位數大於20, 也就是我們常說的Q20過濾
- # 所以上面的這個測序結果，需要把後面的87bp以後的序列切除(trim)，從而保證後續分析的正確性
- # Warning: 如果任何鹼基品質低於10,或者是任何中位數低於25
- # Failure: 如果任何鹼基品質低於5,或者是任何中位數低於20

https://zhuanlan.zhihu.com/p/20731723

Good Quality

A



Bad Quality

B



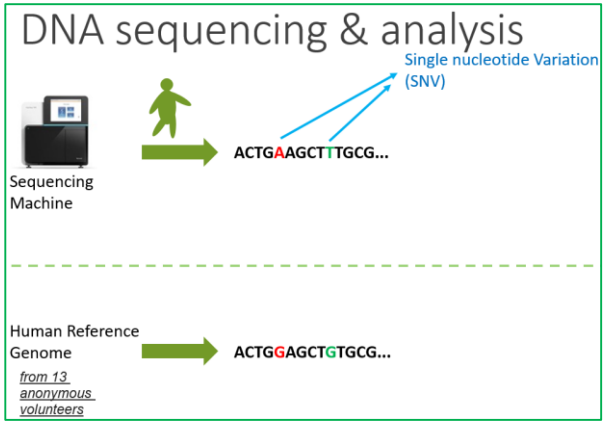
http://pipe-tuxedo.readthedocs.io/en/latest/explain_qc.html

Sequence Alignment/Map Format

FASTQ/Sam/Bam/Vcf/Maf

```
Coor     12345678901234  5678901234567890123456789012345
ref      AGCATGTTAGATAA**GATAGCTGTGCTAGTAGGCAGTCAGCGCCAT

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

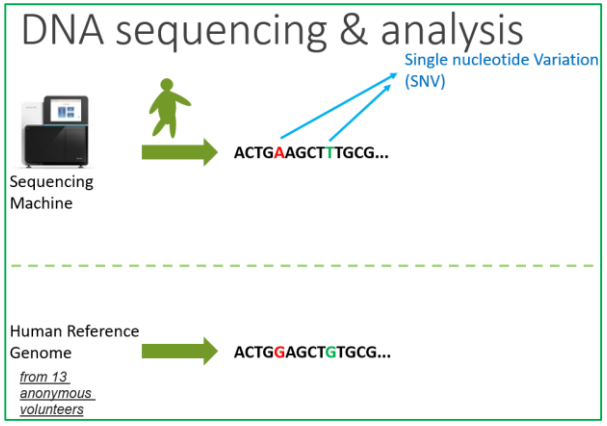


https://samtools.github.io/hts-specs/SAMv1.pdf

Variant Call Format

FASTQ/Sam/Bam/Vcf/Maf

```
##fileformat=VCFv4.0
##reference=1000GenomesPilot-NCBI37
#CHROM  POS             ID      REF     ALT   QUAL    FILTER    INFO    FORMAT
chr1    45805566        .       G       C
chr2    47703379        .       C       T
chr2    48010488        .       G       A
chr2    48030838        .       A       T
chr2    48032875        .       CTAT    -
chr2    48032937        .       T       C
chr2    48033551        .       C       G
chr2    48033910        .       A       T
chr2    215632048       .       G       T
chr2    215632125       .       TT      -
chr2    215632155       .       T       C
chr2    215632192       .       G       A
chr2    215632255       .       CA      TG
chr2    215634055       .       C       T
```



http://www.internationalgenome.org/wiki/Analysis/vcf4.0/

Mutation Annotation Format

FASTQ/Sam/Bam/Vcf/Maf

- Mutation Annotation Format (MAF) is a tab-delimited text file with aggregated mutation information from VCF Files and are generated on a project-level.
- MAF files are produced through the Somatic Aggregation Workflow
- MAFs are produced by aggregating the GDC annotated VCF files generated from one pipeline for one project.
- Somatic MAFs are publicly available and can be freely distributed within the boundaries of the GDC Data Access Policies.

https://docs.gdc.cancer.gov/Data/File_Formats/MAF_Format/

Column	Description
1 - Hugo_Symbol	HUGO symbol for the gene (HUGO symbols are always in all caps). "Unknown" is used for regions that not correspond to a gene
2 - Entrez_Gene_Id	Entrez gene ID (an integer). "0" is used for regions that do not correspond to a gene region or Ensembl
3 - Center	One or more genome sequencing center reporting the variant
4 - NCBI_Build	The reference genome used for the alignment (GRCh38)
5 - Chromosome	The affected chromosome (chr1)
6 - Start_Position	Lowest numeric position of the reported variant on the genomic reference sequence. Mutation start coordina
7 - End_Position	Highest numeric genomic position of the reported variant on the genomic reference sequence. Mutation end coordinate
8 - Strand	Genomic strand of the reported allele. Currently, all variants will report the positive strand: '+'
9 - Variant_Classification	Translational effect of variant allele
10 - Variant_Type	Type of mutation. TNP (tri-nucleotide polymorphism) is analogous to DNP (di-nucleotide polymorphism) but for three consecutive nucleotides. ONP (oligo-nucleotide polymorphism) is analogous to TNP but for consecutive runs of four or more (SNP, DNP, TNP, ONP, INS, DEL, or Consolidated)
11 - Reference_Allele	The plus strand reference allele at this position. Includes the deleted sequence for a deletion or "-" for an insertion
12 - Tumor_Seq_Allele1	Primary data genotype for tumor sequencing (discovery) allele 1. A "-" symbol for a deletion represents a variant. A "-" symbol for an insertion represents wild-type allele. Novel inserted sequence for insertion does not include flanking reference bases

https://docs.gdc.cancer.gov/Data/File_Formats/MAF_Format/



Home | Cancer Genome Projects | Committees and Working Groups

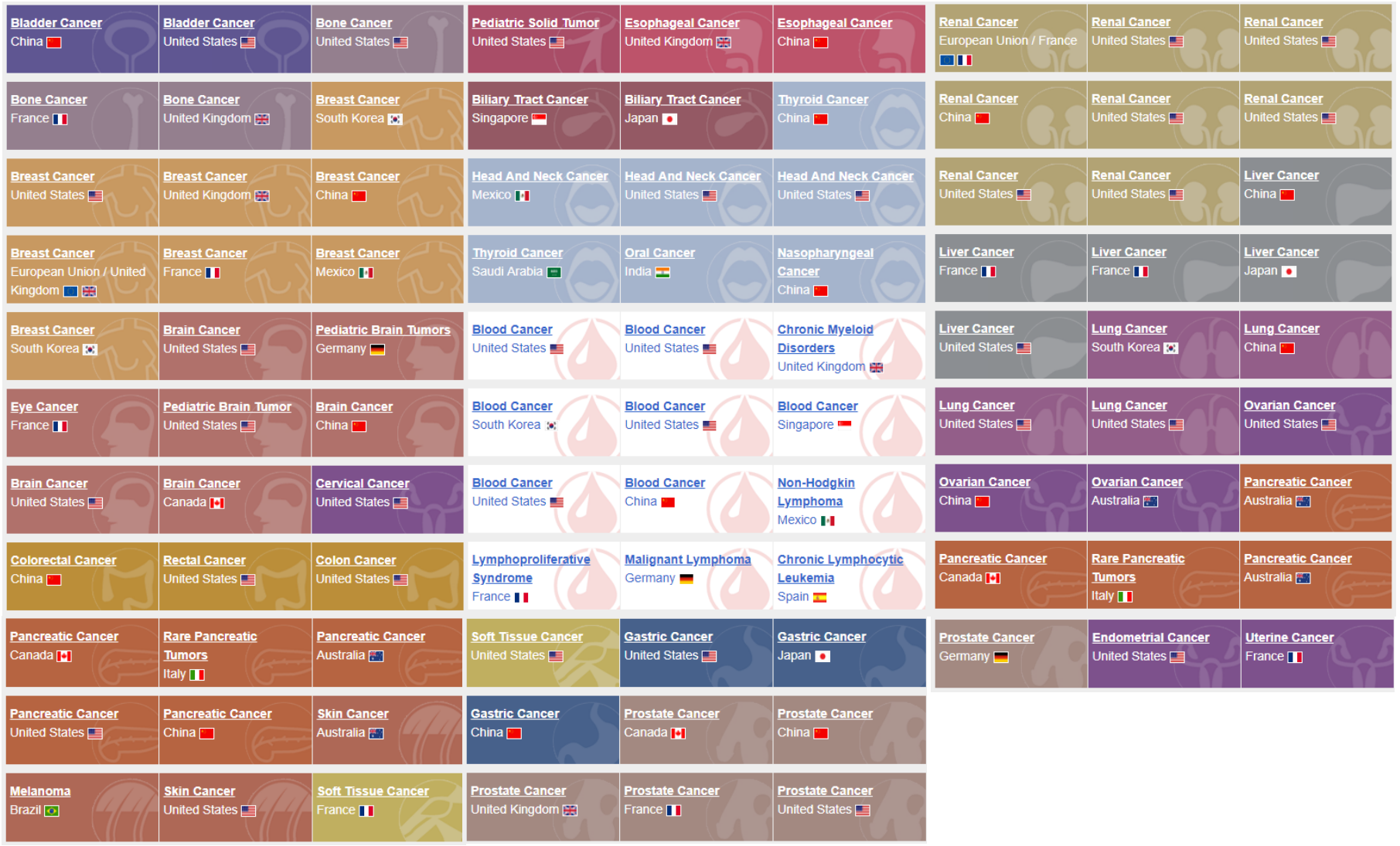
ICGC Cancer Genome Projects

Committed projects to date: 90

ICGC Goal: To obtain a **comprehensive** description of **genomic, transcriptomic and epigenomic changes** in **50 different tumor types and/or subtypes** which are of clinical and societal importance across the globe.

Read more »

http://icgc.org/


Bladder Cancer China
Bladder Cancer United States
Bone Cancer United States
Pediatric Solid Tumor United States
Esophageal Cancer United Kingdom
Esophageal Cancer China
Renal Cancer European Union / France
Renal Cancer United States
Renal Cancer United States
Bone Cancer France
Bone Cancer United Kingdom
Breast Cancer South Korea
Biliary Tract Cancer Singapore
Biliary Tract Cancer Japan
Thyroid Cancer China
Renal Cancer China
Renal Cancer United States
Renal Cancer United States
Breast Cancer United States
Breast Cancer United Kingdom
Breast Cancer China
Head And Neck Cancer Mexico
Head And Neck Cancer United States
Head And Neck Cancer United States
Renal Cancer United States
Renal Cancer United States
Liver Cancer China
Breast Cancer European Union / United Kingdom
Breast Cancer France
Breast Cancer Mexico
Thyroid Cancer Saudi Arabia
Oral Cancer India
Nasopharyngeal Cancer China
Liver Cancer France
Liver Cancer France
Liver Cancer Japan
Breast Cancer South Korea
Brain Cancer United States
Pediatric Brain Tumors Germany
Blood Cancer United States
Blood Cancer United States
Chronic Myeloid Disorders United Kingdom
Liver Cancer United States
Lung Cancer South Korea
Lung Cancer China
Eye Cancer France
Pediatric Brain Tumor United States
Brain Cancer China
Blood Cancer South Korea
Blood Cancer United States
Blood Cancer Singapore
Lung Cancer United States
Lung Cancer United States
Ovarian Cancer United States
Brain Cancer United States
Brain Cancer Canada
Cervical Cancer United States
Blood Cancer United States
Blood Cancer China
Non-Hodgkin Lymphoma Mexico
Ovarian Cancer China
Ovarian Cancer Australia
Pancreatic Cancer Australia
Colorectal Cancer China
Rectal Cancer United States
Colon Cancer United States
Lymphoproliferative Syndrome France
Malignant Lymphoma Germany
Chronic Lymphocytic Leukemia Spain
Pancreatic Cancer Canada
Rare Pancreatic Tumors Italy
Pancreatic Cancer Australia
Pancreatic Cancer Canada
Rare Pancreatic Tumors Italy
Pancreatic Cancer Australia
Soft Tissue Cancer United States
Gastric Cancer United States
Gastric Cancer Japan
Prostate Cancer Germany
Endometrial Cancer United States
Uterine Cancer France
Pancreatic Cancer United States
Pancreatic Cancer China
Skin Cancer Australia
Gastric Cancer China
Prostate Cancer Canada
Prostate Cancer China
Melanoma Brazil
Skin Cancer United States
Soft Tissue Cancer France
Prostate Cancer United Kingdom
Prostate Cancer France
Prostate Cancer United States

DCC DATA RELEASES

DCC / release_26 / Projects / HNSC-US / Filter by file name...

Name	File Size	Date
copy_number_somatic_mutation.HNSC-US.tsv.gz	8.38 MB	Dec 13, 2017
donor.HNSC-US.tsv.gz	54.05 KB	Dec 13, 2017
exp_seq.HNSC-US.tsv.gz	167.67 MB	Dec 13, 2017
meth_array.HNSC-US.tsv.gz	3.04 GB	Dec 13, 2017
mirna_seq.HNSC-US.tsv.gz	7.46 MB	Dec 13, 2017
protein_expression.HNSC-US.tsv.gz	760.37 KB	Dec 13, 2017
sample.HNSC-US.tsv.gz	103.52 KB	Dec 13, 2017
simple_somatic_mutation.open.HNSC-US.tsv.gz	907.43 KB	Dec 13, 2017
specimen.HNSC-US.tsv.gz	116.11 KB	Dec 13, 2017

MAF file (notepad)

File Edit View Settings ?

	Hugo_Symbol	Entrez_Gene_Id	Center	NCBI_Build	Chromosome	Start_Position	End_Position	Strand	Variant_Classification	Variant_Type	Reference_Allele	Tumor_Seq_Allele1
2	ABCA10	10349	genome.wustl.edu	37	17	67170917	67170917	+	Splice_Site	SNP	T	T C TCGA-AB-2988 p.K960R 45.66 NM_080282.3
3	ABCA4	24	genome.wustl.edu	37	1	94490594	94490594	+	Missense_Mutation	SNP	C	C T TCGA-AB-2869 p.R1517H 38.12 NM_000350.2
4	ABCB11	8647	genome.wustl.edu	37	2	169780250	169780250	+	Missense_Mutation	SNP	G	G A TCGA-AB-3009 p.A1283V 46.9721767594108 NM_003742.2
5	ABCC3	8714	genome.wustl.edu	37	17	48760974	48760974	+	Missense_Mutation	SNP	C	C T TCGA-AB-2887 p.P1271S 56.41 NM_003786.1
6	ABCF1	23	genome.wustl.edu	37	6	30554429	30554429	+	Missense_Mutation	SNP	G	G A TCGA-AB-2920 p.G658S 40.95 NM_001025091.1
7	ABCG4	64137	genome.wustl.edu	37	11	119031351	119031351	+	Missense_Mutation	SNP	A	A G TCGA-AB-2934 p.Y567C 32.84 NM_022169.1
8	ABCG8	64241	genome.wustl.edu	37	2	44079555	44079555	+	Missense_Mutation	SNP	G	G A TCGA-AB-2905 p.M208I 37.06 NM_022437.2
9	ABL1	25	genome.wustl.edu	37	9	133760430	133760430	+	Missense_Mutation	SNP	C	C T TCGA-AB-2999 p.P918L 40.75 NM_007313.3
10	ACOXL	55289	genome.wustl.edu	37	2	111542370	111542370	+	Missense_Mutation	SNP	C	C T TCGA-AB-2950 p.A46V 42.94 NM_001105516.2
11	ACPL2	92370	genome.wustl.edu	37	3	141011154	141011154	+	Missense_Mutation	SNP	T	T G TCGA-AB-2923 p.Y184D 38.46 NM_001037172.1
12	ACSM2A	123876	genome.wustl.edu	37	16	20482538	20482538	+	Splice_Site	SNP	G	G A TCGA-AB-2847 p.G247D 37.36 NM_001010845.2
13	ACSS3	79611	genome.wustl.edu	37	12	81536902	81536902	+	Missense_Mutation	SNP	C	C T TCGA-AB-3009 p.A266V 47.3251028806584 NM_024560.2
14	ACTBL2	345651	genome.wustl.edu	37	5	56777739	56777739	+	Missense_Mutation	SNP	A	A G TCGA-AB-2931 p.S266P 49.22 NM_001017992.2
15	ACTL7A	10881	genome.wustl.edu	37	9	111625878	111625878	+	Missense_Mutation	SNP	G	G C TCGA-AB-2920 p.G426R 52.34 NM_006687.2
16	ACVR2B	93	genome.wustl.edu	37	3	38519671	38519671	+	Missense_Mutation	SNP	C	C T TCGA-AB-2973 p.T137M 43.26 NM_001106.3
17	ADAM11	4185	genome.wustl.edu	37	17	42855399	42855399	+	Missense_Mutation	SNP	C	C A TCGA-AB-2854 p.T717K 33.33 NM_002390.4
18	ADAM11	4185	genome.wustl.edu	37	17	42837242	42837242	+	Missense_Mutation	SNP	C	C T TCGA-AB-2906 p.R72C 41.41 NM_002390.4
19	ADAMTS1	9510	genome.wustl.edu	37	21	28214747	28214747	+	Missense_Mutation	SNP	G	G A TCGA-AB-2894 p.R330W 33.68 NM_006988.3
20	ADAMTS12	81792	genome.wustl.edu	37	5	33637780	33637780	+	Missense_Mutation	SNP	C	C T TCGA-AB-2945 p.R597H 30.5 NM_030955.2
21	ADAMTS16	170690	genome.wustl.edu	37	5	5239824	5239824	+	Missense_Mutation	SNP	G	G A TCGA-AB-2847 p.G770E 32.73 NM_139056.2
22	ADAMTS17	170691	genome.wustl.edu	37	15	100589063	100589063	+	Splice_Site	SNP	G	G A TCGA-AB-2913 p.R864W 38.38 NM_139057.2
23	ADAMTSL1	92949	genome.wustl.edu	37	9	18639381	18639381	+	Missense_Mutation	SNP	G	G T TCGA-AB-2996 p.G269V 47.85 NM_001040272.4
24	ADCY10	55811	genome.wustl.edu	37	1	167873180	167873180	+	Missense_Mutation	SNP	T	T A TCGA-AB-2952 p.R66S 40.07 NM_018417.3
25	ADCY3	109	genome.wustl.edu	37	2	25050814	25050814	+	Missense_Mutation	SNP	G	G A TCGA-AB-2805 p.P797S 47.57 NM_004036.3
26	ADCY5	111	genome.wustl.edu	37	3	123071313	123071313	+	Missense_Mutation	SNP	G	G A TCGA-AB-2833 p.A417V 42.67 NM_183357.1
27	ADNP	23394	genome.wustl.edu	37	20	49509105	49509105	+	Missense_Mutation	SNP	T	T G TCGA-AB-2890 p.K716Q 37.17 NM_015339.2
28	ADRBK1	156	genome.wustl.edu	37	11	67049291	67049291	+	Splice_Site	SNP	A	A G TCGA-AB-2911 17.5 NM_001619.3
29	ADSS	159	genome.wustl.edu	37	1	244580929	244580929	+	Splice_Site	SNP	C	C T TCGA-AB-2912 32.03 NM_001126.2
30	AFF2	2334	genome.wustl.edu	37	23	147891443	147891443	+	Splice_Site	SNP	G	G A TCGA-AB-3001 p.R362Q 23.82 NM_002025.2

MAF file (Excel)

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q
1	Hugo_Symbol	Entrez_Gene_Id	Center	NCBI_Build	Chromosome	Start_Position	End_Position	Strand	Variant_Classification	Variant_Type	Reference_Allele	Tumor_Seq_Allele1	Tumor_Seq_Allele2	Tumor_Sample_Barcode	Protein_Change	i_TumorVAF_WU	i_transcript_name
2	ABCA10	10349	genome.wustl.edu	37	17	67170917	67170917	+	Splice_Site	SNP	T	T	C	TCGA-AB-2988	p.K960R	45.66	NM_080282.3
3	ABCA4	24	genome.wustl.edu	37	1	94490594	94490594	+	Missense_Mutation	SNP	C	C	T	TCGA-AB-2869	p.R1517H	38.12	NM_000350.2
4	ABCB11	8647	genome.wustl.edu	37	2	169780250	169780250	+	Missense_Mutation	SNP	G	G	A	TCGA-AB-3009	p.A1283V	46.9721768	NM_003742.2
5	ABCC3	8714	genome.wustl.edu	37	17	48760974	48760974	+	Missense_Mutation	SNP	C	C	T	TCGA-AB-2887	p.P1271S	56.41	NM_003786.1
6	ABCF1	23	genome.wustl.edu	37	6	30554429	30554429	+	Missense_Mutation	SNP	G	G	A	TCGA-AB-2920	p.G658S	40.95	NM_001025091.1
7	ABCG4	64137	genome.wustl.edu	37	11	119031351	119031351	+	Missense_Mutation	SNP	A	A	G	TCGA-AB-2934	p.Y567C	32.84	NM_022169.1
8	ABCG8	64241	genome.wustl.edu	37	2	44079555	44079555	+	Missense_Mutation	SNP	G	G	A	TCGA-AB-2905	p.M208I	37.06	NM_022437.2
9	ABL1	25	genome.wustl.edu	37	9	133760430	133760430	+	Missense_Mutation	SNP	C	C	T	TCGA-AB-2999	p.P918L	40.75	NM_007313.3
10	ACOXL	55289	genome.wustl.edu	37	2	111542370	111542370	+	Missense_Mutation	SNP	C	C	T	TCGA-AB-2950	p.A46V	42.94	NM_001105516.2
11	ACPL2	92370	genome.wustl.edu	37	3	141011154	141011154	+	Missense_Mutation	SNP	T	T	G	TCGA-AB-2923	p.Y184D	38.46	NM_001037172.1
12	ACSM2A	123876	genome.wustl.edu	37	16	20482538	20482538	+	Splice_Site	SNP	G	G	A	TCGA-AB-2847	p.G247D	37.36	NM_001010845.2
13	ACSS3	79611	genome.wustl.edu	37	12	81536902	81536902	+	Missense_Mutation	SNP	C	C	T	TCGA-AB-3009	p.A266V	47.3251029	NM_024560.2
14	ACTBL2	345651	genome.wustl.edu	37	5	56777739	56777739	+	Missense_Mutation	SNP	A	A	G	TCGA-AB-2931	p.S266P	49.22	NM_001017992.2
15	ACTL7A	10881	genome.wustl.edu	37	9	111625878	111625878	+	Missense_Mutation	SNP	G	G	C	TCGA-AB-2920	p.G426R	52.34	NM_006687.2
16	ACVR2B	93	genome.wustl.edu	37	3	38519671	38519671	+	Missense_Mutation	SNP	C	C	T	TCGA-AB-2973	p.T137M	43.26	NM_001106.3
17	ADAM11	4185	genome.wustl.edu	37	17	42855399	42855399	+	Missense_Mutation	SNP	C	C	A	TCGA-AB-2854	p.T717K	33.33	NM_002390.4
18	ADAM11	4185	genome.wustl.edu	37	17	42837242	42837242	+	Missense_Mutation	SNP	C	C	T	TCGA-AB-2906	p.R72C	41.41	NM_002390.4
19	ADAMTS1	9510	genome.wustl.edu	37	21	28214747	28214747	+	Missense_Mutation	SNP	G	G	A	TCGA-AB-2894	p.R330W	33.68	NM_006988.3
20	ADAMTS12	81792	genome.wustl.edu	37	5	33637780	33637780	+	Missense_Mutation	SNP	C	C	T	TCGA-AB-2945	p.R597H	30.5	NM_030955.2
21	ADAMTS16	170690	genome.wustl.edu	37	5	5239824	5239824	+	Missense_Mutation	SNP	G	G	A	TCGA-AB-2847	p.G770E	32.73	NM_139056.2
22	ADAMTS17	170691	genome.wustl.edu	37	15	100589063	100589063	+	Splice_Site	SNP	G	G	A	TCGA-AB-2913	p.R864W	38.38	NM_139057.2
23	ADAMTSL1	92949	genome.wustl.edu	37	9	18639381	18639381	+	Missense_Mutation	SNP	G	G	T	TCGA-AB-2996	p.G269V	47.85	NM_001040272.4
24	ADCY10	55811	genome.wustl.edu	37	1	167873180	167873180	+	Missense_Mutation	SNP	T	T	A	TCGA-AB-2952	p.R66S	40.07	NM_018417.3

視覺化!

MAF files

Install MAFtools

install R

http://cran.csie.ntu.edu.tw/

http://cran.csie.ntu.edu.tw/bin/windows/base/R-3.5.3-win.exe

install R-studio

https://www.rstudio.com/products/rstudio/download/

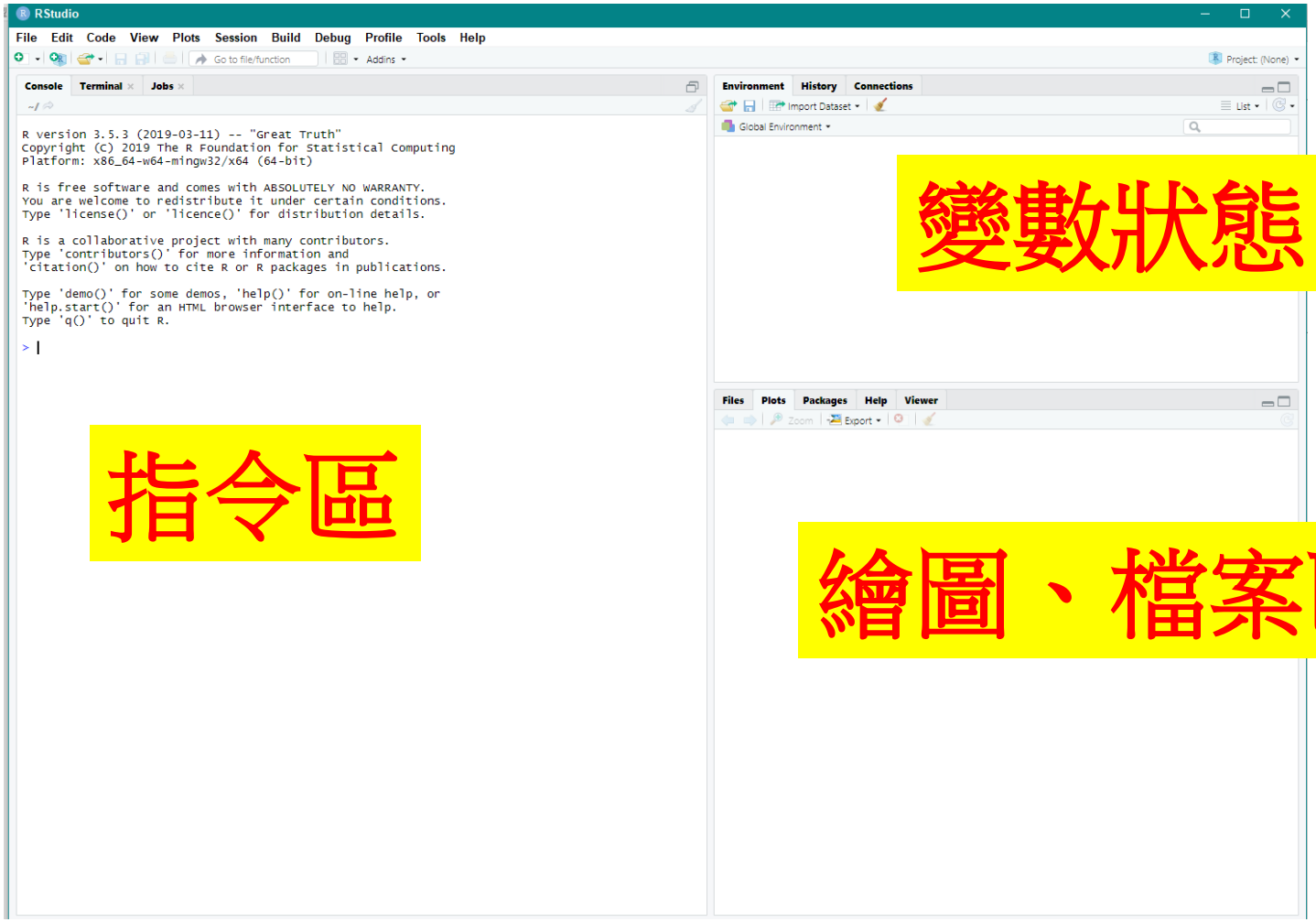
在R-studio下輸入安裝指令:

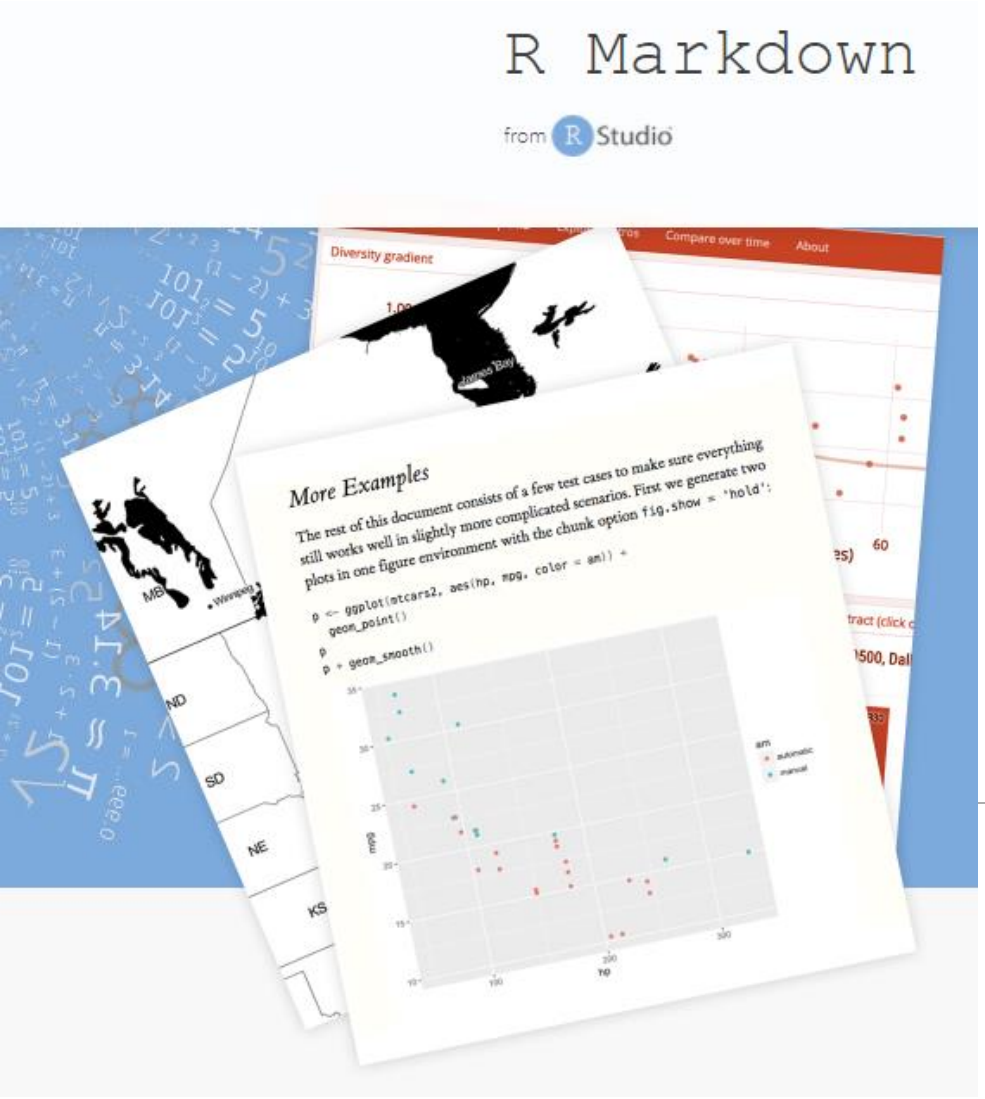
```
BiocManager::install(c("GenomicFeatures", "AnnotationDbi"))

if (!require("BiocManager"))
    install.packages("BiocManager")
BiocManager::install("maftools")
```

R notebook

R-studio 操作


RStudio
File Edit Code View Plots Session Build Debug Profile Tools Help
Console Terminal Jobs
R version 3.5.3 (2019-03-11) -- "Great Truth"
Copyright (C) 2019 The R Foundation for Statistical Computing
Platform: x86_64-w64-mingw32/x64 (64-bit)
R is free software and comes with ABSOLUTELY NO WARRANTY.
You are welcome to redistribute it under certain conditions.
Type 'license()' or 'licence()' for distribution details.
R is a collaborative project with many contributors.
Type 'contributors()' for more information and
'citation()' on how to cite R or R packages in publications.
Type 'demo()' for some demos, 'help()' for on-line help, or
'help.start()' for an HTML browser interface to help.
Type 'q()' to quit R.
>
Environment History Connections
Import Dataset
Global Environment
變數狀態
Files Plots Packages Help Viewer
Zoom Export
指令區
繪圖、檔案區



R Markdown documents are fully reproducible. Use a productive notebook interface to weave together narrative text and code to produce elegantly formatted output.

http://svmiller.com/blog/2016/03/svm-r-markdown-cv/

William Sealy Gosset

Chief Brewer, Arthur Guinness & Son
Guinness Brewery · Park Royal · London NW10 7RR, UK
guinness@consumer-care.net +353 1 408 4800 guinness.com

Employment

Arthur Guinness & Son

Chief Brewer (Park Royal)	1935-1937
Researcher (Dublin)	1899-1935

Education

New College, Oxford University, B.A. Chemistry 1899

Publications

Books

Pearson, E.S. and John Wishart (eds.). 1943. *"Student's" Collected Papers* London: Biometrika Office.

Journal Articles

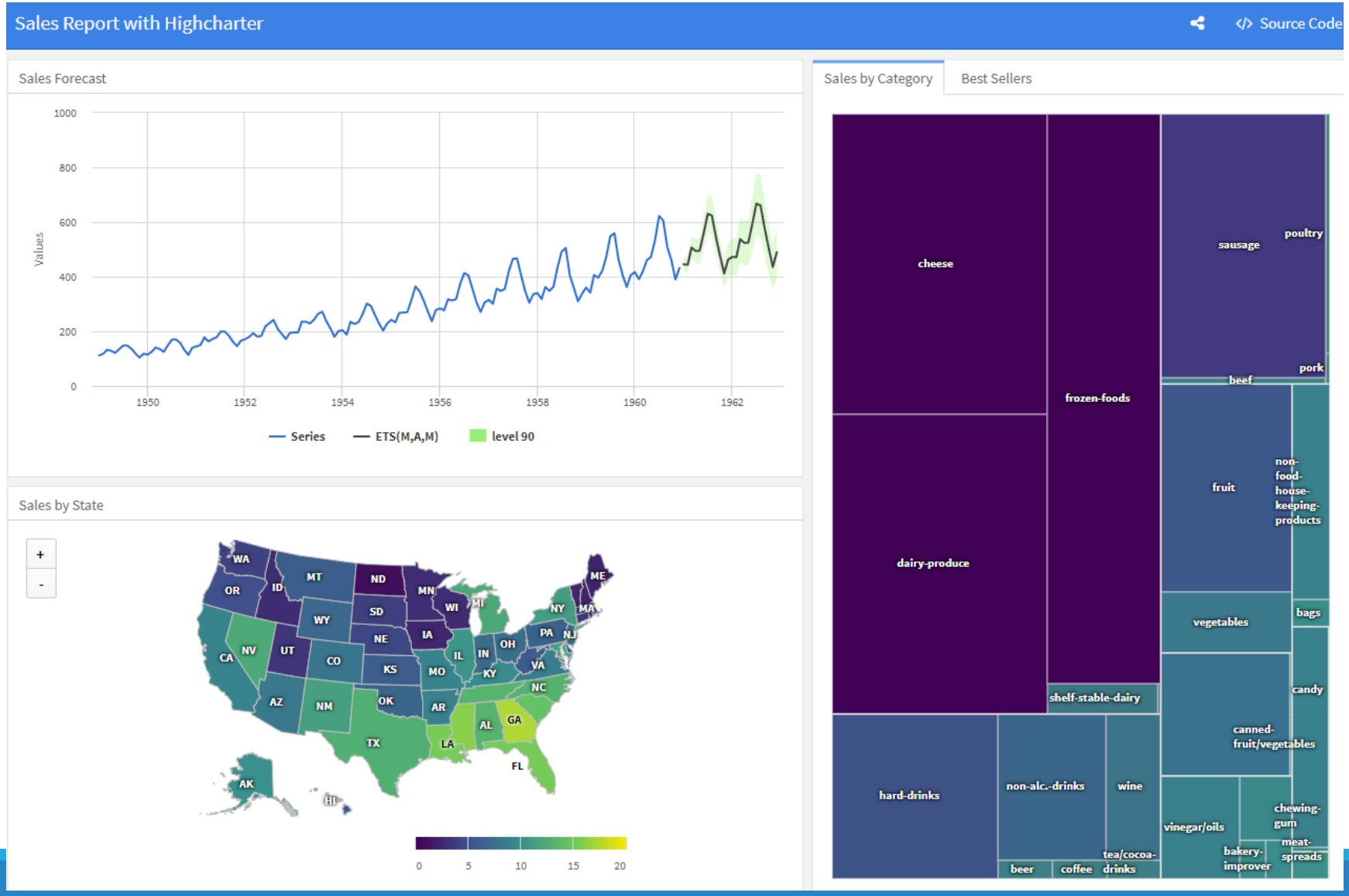
Student. 1938. "Comparison between Balanced and Random Arrangements on Field Plots." *Biometrika* 29(3-4): 363–378.

Student. 1936. "Co-operation in Large Scale Experiments." *Supplement to Journal of the Royal Statistical Society* 3(2): 115–136.

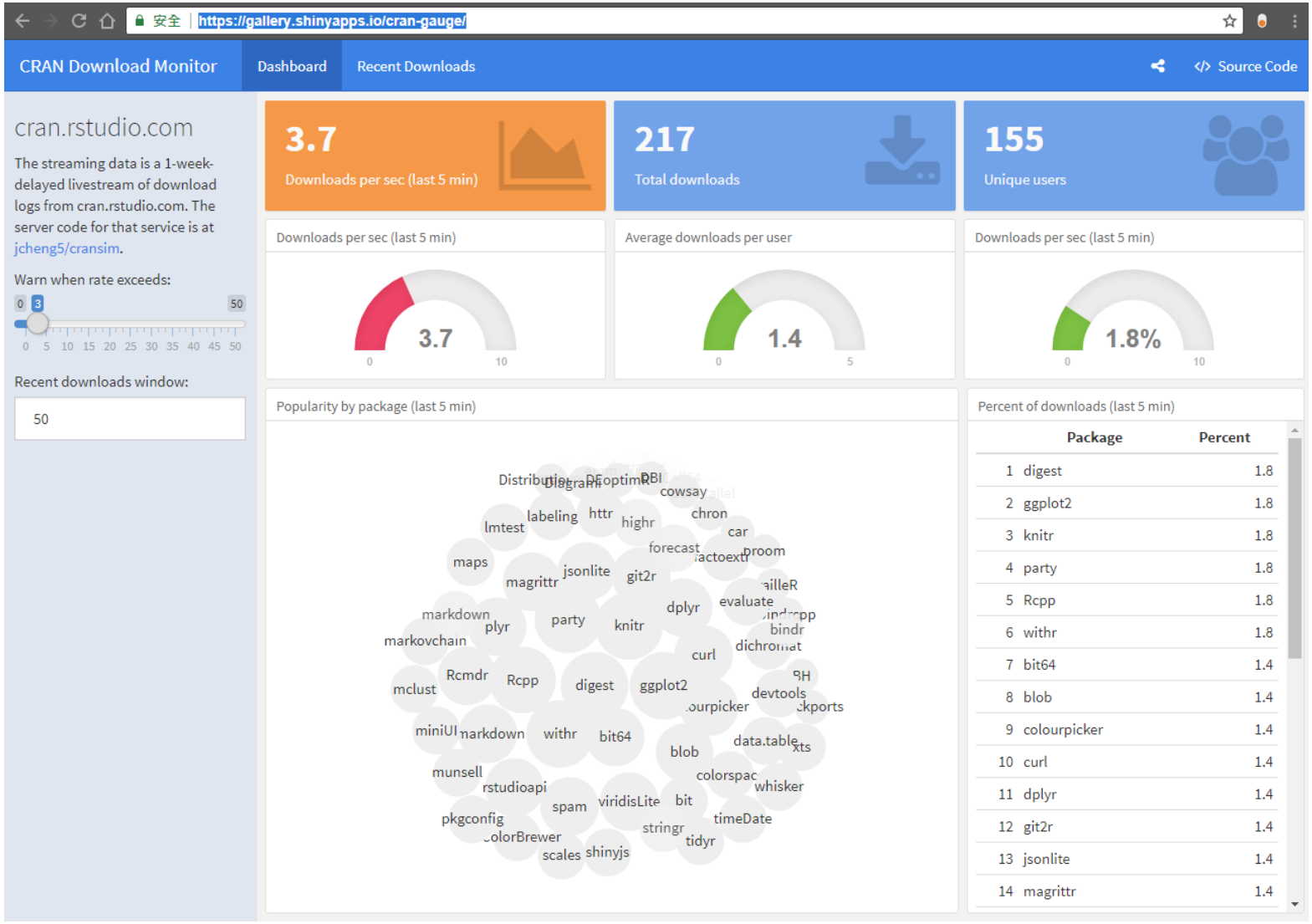
Student. 1931. "On the z Test." *Biometrika* 23(3-4): 407-408.

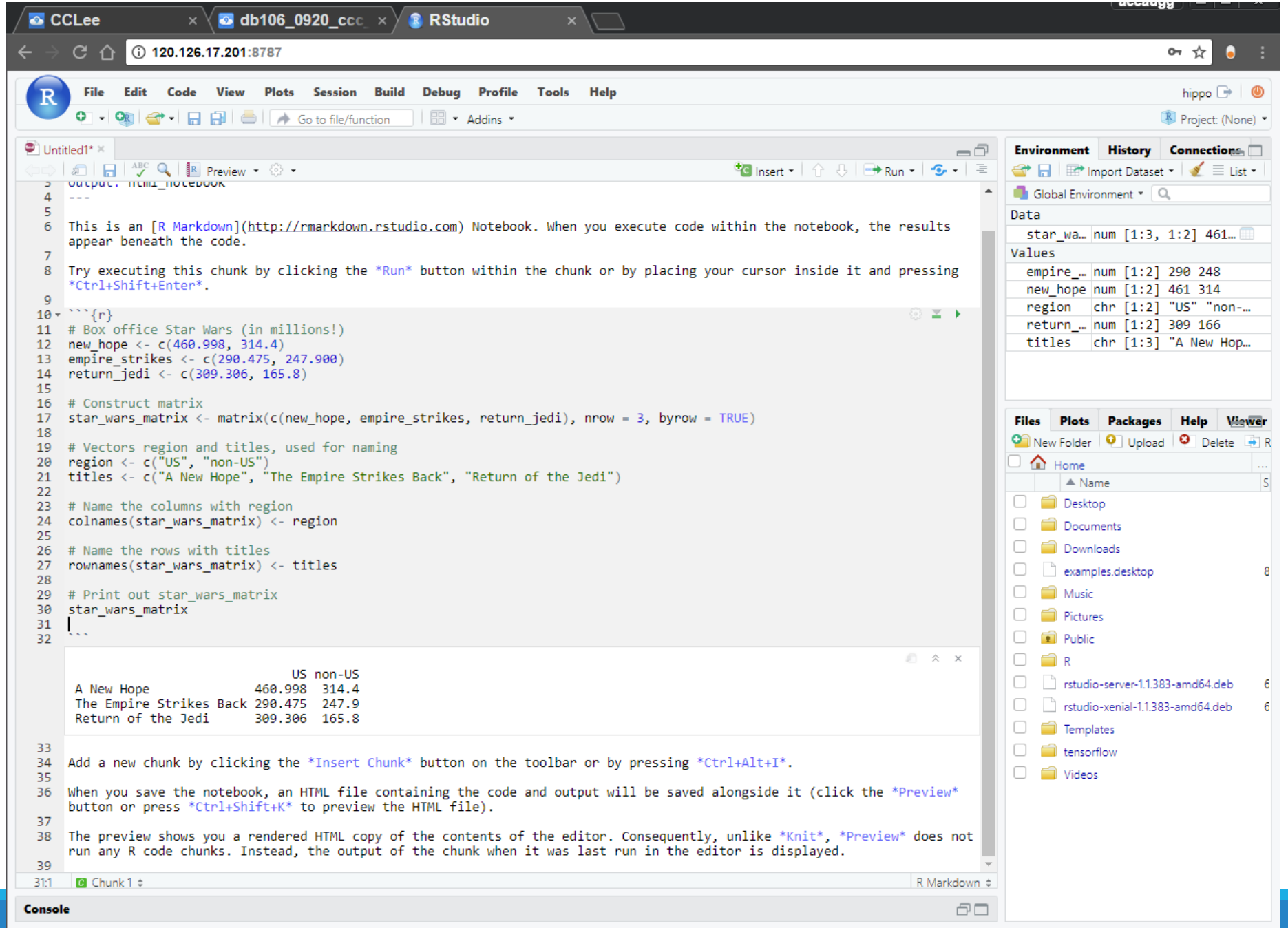
Student. 1931. "The Lanarkshire Milk Experiment." *Biometrika* 23(3-4): 398–406.

https://beta.rstudioconnect.com/jjallaire/htmlwidgets-highcharter/htmlwidgets-highcharter.html#sales-by-category

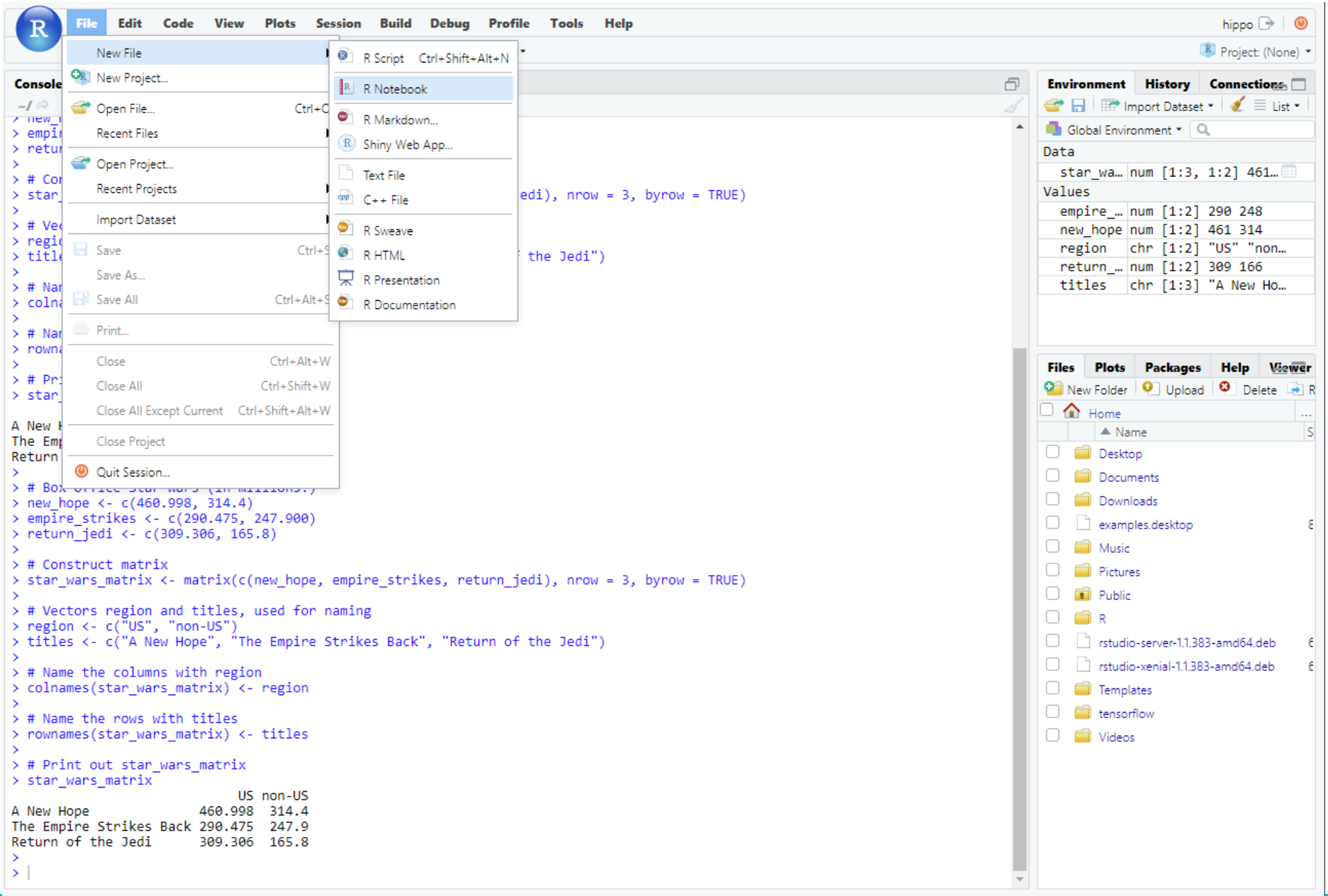


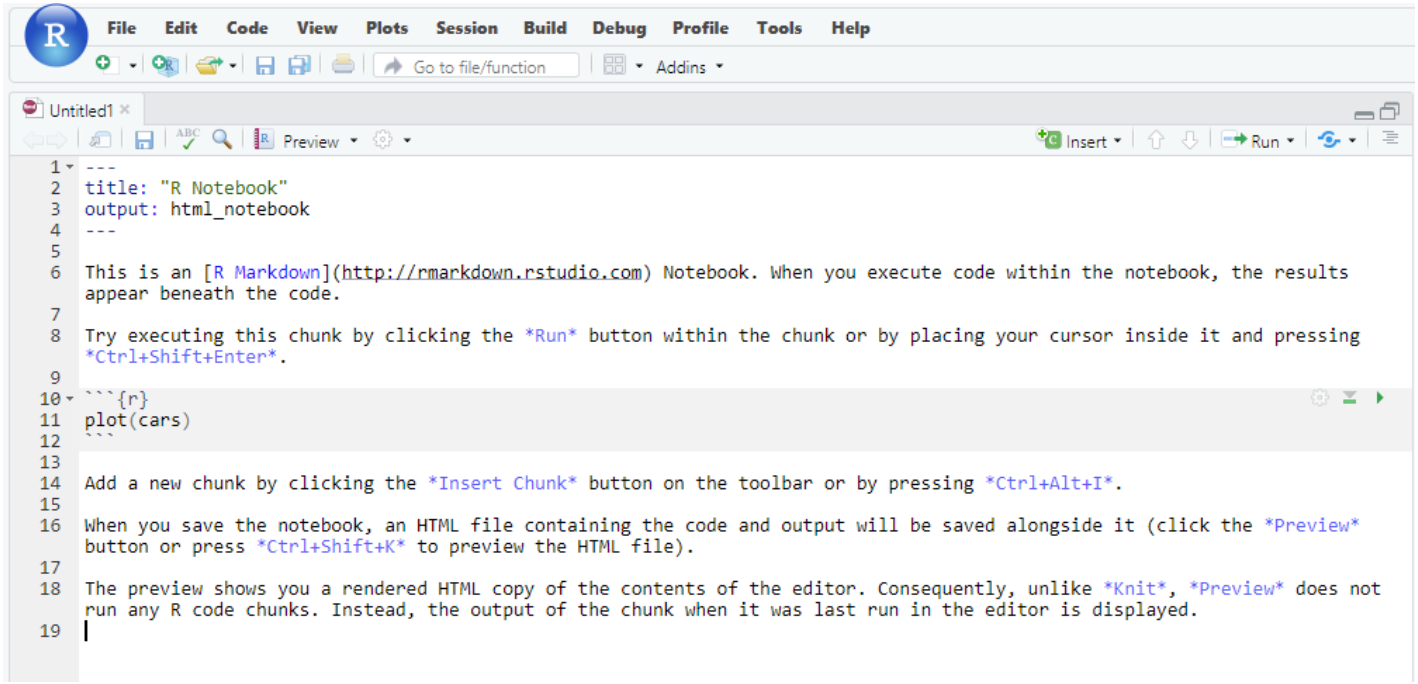
https://gallery.shinyapps.io/cran-gauge/




CCLee
db106_0920_ccc
RStudio
120.126.17.201:8787
File Edit Code View Plots Session Build Debug Profile Tools Help
hippo
Go to file/function
Addins
Project: (None)
Untitled1*
Preview
Insert
Run
output: html_notebook

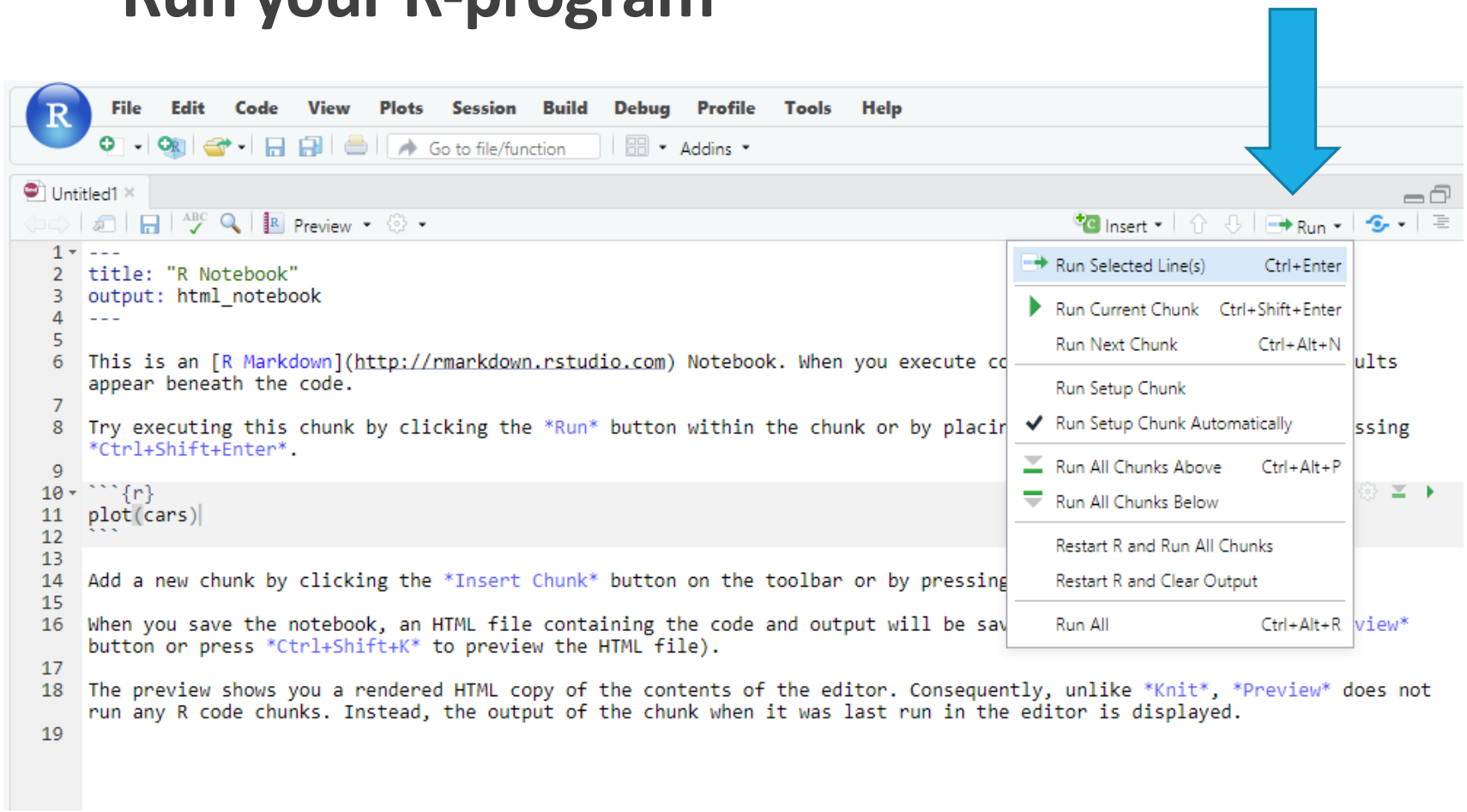
This is an [R Markdown](http://rmarkdown.rstudio.com) Notebook. When you execute code within the notebook, the results appear beneath the code.
Try executing this chunk by clicking the *Run* button within the chunk or by placing your cursor inside it and pressing *Ctrl+Shift+Enter*.
```{r}
# Box office Star Wars (in millions!)
new_hope <- c(460.998, 314.4)
empire_strikes <- c(290.475, 247.900)
return_jedi <- c(309.306, 165.8)
# Construct matrix
star_wars_matrix <- matrix(c(new_hope, empire_strikes, return_jedi), nrow = 3, byrow = TRUE)
# Vectors region and titles, used for naming
region <- c("US", "non-US")
titles <- c("A New Hope", "The Empire Strikes Back", "Return of the Jedi")
# Name the columns with region
colnames(star_wars_matrix) <- region
# Name the rows with titles
rownames(star_wars_matrix) <- titles
# Print out star_wars_matrix
star_wars_matrix
```
US non-US
A New Hope 460.998 314.4
The Empire Strikes Back 290.475 247.9
Return of the Jedi 309.306 165.8
Add a new chunk by clicking the *Insert Chunk* button on the toolbar or by pressing *Ctrl+Alt+I*.
When you save the notebook, an HTML file containing the code and output will be saved alongside it (click the *Preview* button or press *Ctrl+Shift+K* to preview the HTML file).
The preview shows you a rendered HTML copy of the contents of the editor. Consequently, unlike *Knit*, *Preview* does not run any R code chunks. Instead, the output of the chunk when it was last run in the editor is displayed.
31:1 Chunk 1 R Markdown
Console
Environment History Connections
Import Dataset List
Global Environment
Data
star_wa… num [1:3, 1:2] 461…
Values
empire_… num [1:2] 290 248
new_hope num [1:2] 461 314
region chr [1:2] "US" "non-…
return_… num [1:2] 309 166
titles chr [1:3] "A New Hop…
Files Plots Packages Help Viewer
New Folder Upload Delete
Home
Name
Desktop
Documents
Downloads
examples.desktop
Music
Pictures
Public
R
rstudio-server-1.1.383-amd64.deb
rstudio-xenial-1.1.383-amd64.deb
Templates
tensorflow
Videos


File Edit Code View Plots Session Build Debug Profile Tools Help
hippo
Project: (None)
New File
New Project...
Open File... Ctrl+O
Recent Files
Open Project...
Recent Projects
Import Dataset
Save Ctrl+S
Save As...
Save All Ctrl+Alt+S
Print...
Close Ctrl+Alt+W
Close All Ctrl+Shift+W
Close All Except Current Ctrl+Shift+Alt+W
Close Project
Quit Session...
R Script Ctrl+Shift+Alt+N
R Notebook
R Markdown...
Shiny Web App...
Text File
C++ File
R Sweave
R HTML
R Presentation
R Documentation
Console
> # Box Office Star Wars (in millions)
> new_hope <- c(460.998, 314.4)
> empire_strikes <- c(290.475, 247.900)
> return_jedi <- c(309.306, 165.8)
>
> # Construct matrix
> star_wars_matrix <- matrix(c(new_hope, empire_strikes, return_jedi), nrow = 3, byrow = TRUE)
>
> # Vectors region and titles, used for naming
> region <- c("US", "non-US")
> titles <- c("A New Hope", "The Empire Strikes Back", "Return of the Jedi")
>
> # Name the columns with region
> colnames(star_wars_matrix) <- region
>
> # Name the rows with titles
> rownames(star_wars_matrix) <- titles
>
> # Print out star_wars_matrix
> star_wars_matrix
US non-US
A New Hope 460.998 314.4
The Empire Strikes Back 290.475 247.9
Return of the Jedi 309.306 165.8
>
Environment History Connections
Import Dataset List
Global Environment
Data
star_wa… num [1:3, 1:2] 461…
Values
empire_… num [1:2] 290 248
new_hope num [1:2] 461 314
region chr [1:2] "US" "non…
return_… num [1:2] 309 166
titles chr [1:3] "A New Ho…
Files Plots Packages Help Viewer
New Folder Upload Delete
Home
Name
Desktop
Documents
Downloads
examples.desktop
Music
Pictures
Public
R
rstudio-server-1.1.383-amd64.deb
rstudio-xenial-1.1.383-amd64.deb
Templates
tensorflow
Videos

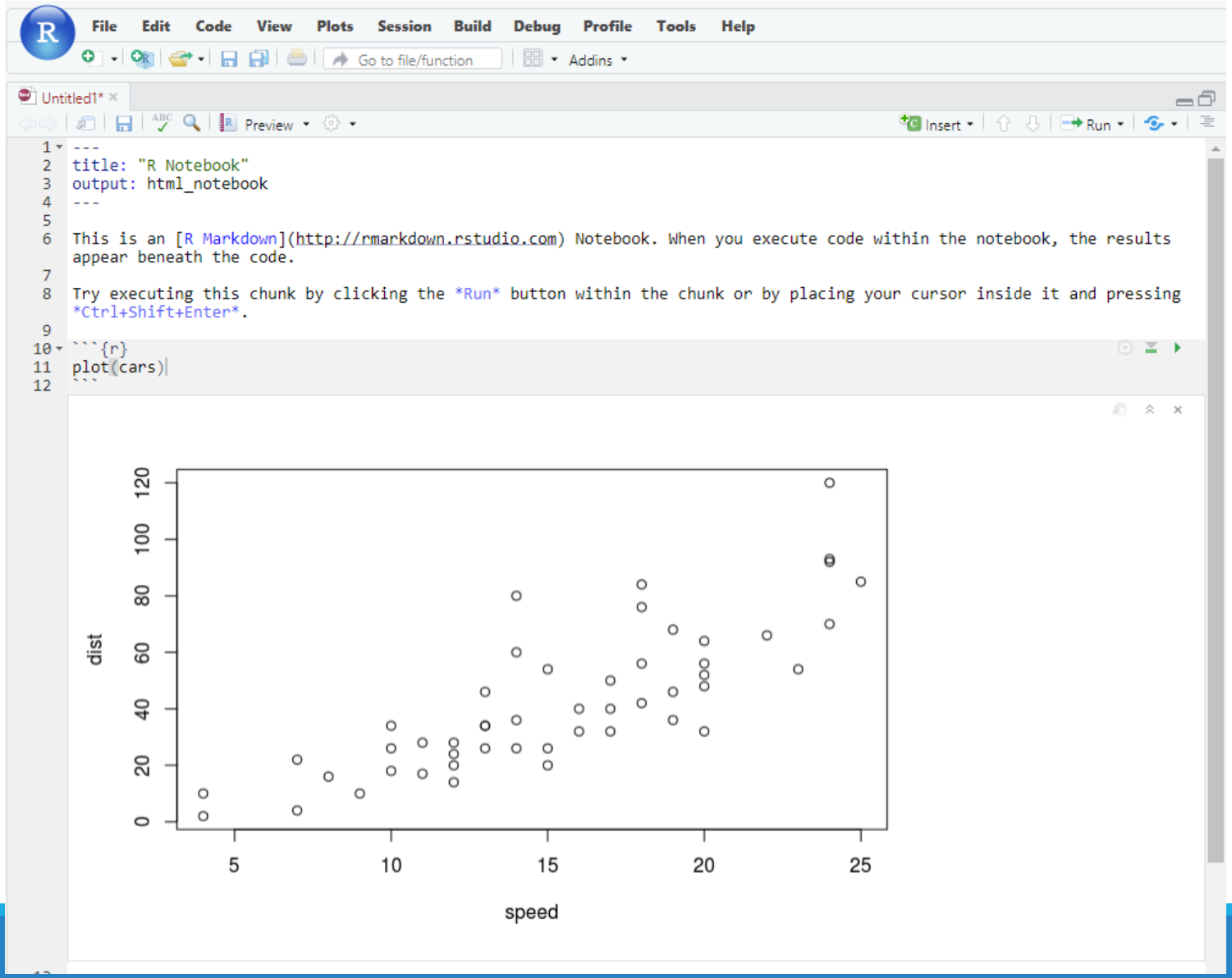

File Edit Code View Plots Session Build Debug Profile Tools Help
Go to file/function
Addins
Untitled1
Preview
Insert
Run
---
title: "R Notebook"
output: html_notebook

This is an [R Markdown](http://rmarkdown.rstudio.com) Notebook. When you execute code within the notebook, the results appear beneath the code.
Try executing this chunk by clicking the *Run* button within the chunk or by placing your cursor inside it and pressing *Ctrl+Shift+Enter*.
```{r}
plot(cars)
```
Add a new chunk by clicking the *Insert Chunk* button on the toolbar or by pressing *Ctrl+Alt+I*.
When you save the notebook, an HTML file containing the code and output will be saved alongside it (click the *Preview* button or press *Ctrl+Shift+K* to preview the HTML file).
The preview shows you a rendered HTML copy of the contents of the editor. Consequently, unlike *Knit*, *Preview* does not run any R code chunks. Instead, the output of the chunk when it was last run in the editor is displayed.

Run your R-program



結果直接顯示在block下方

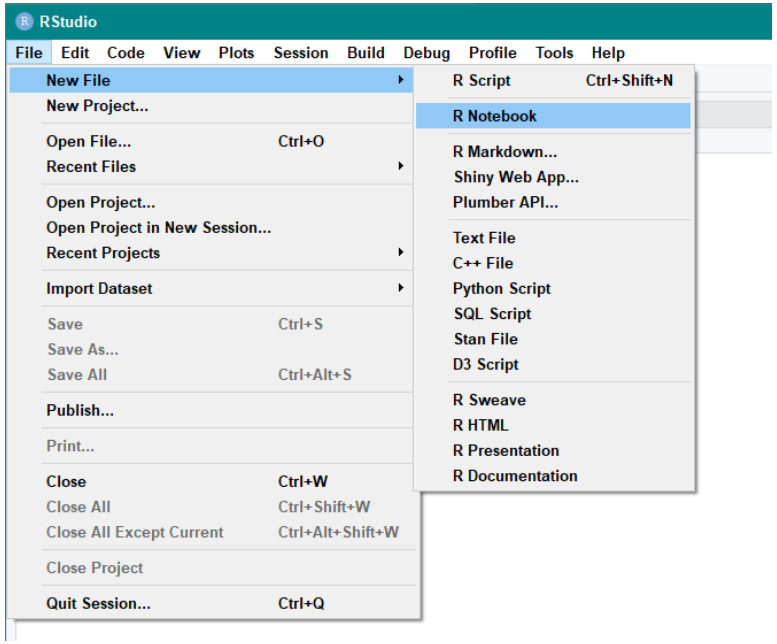


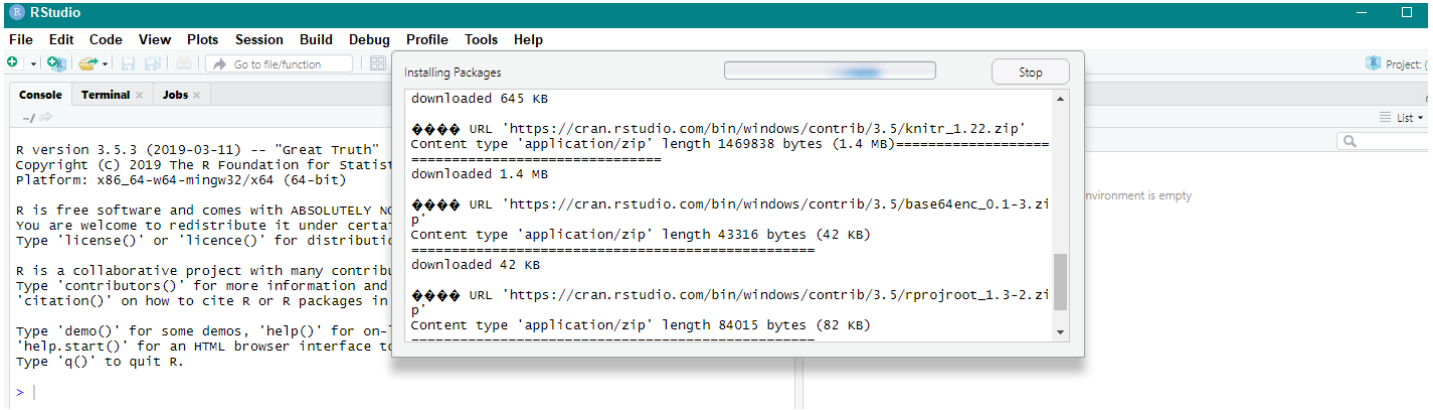
Try it!!

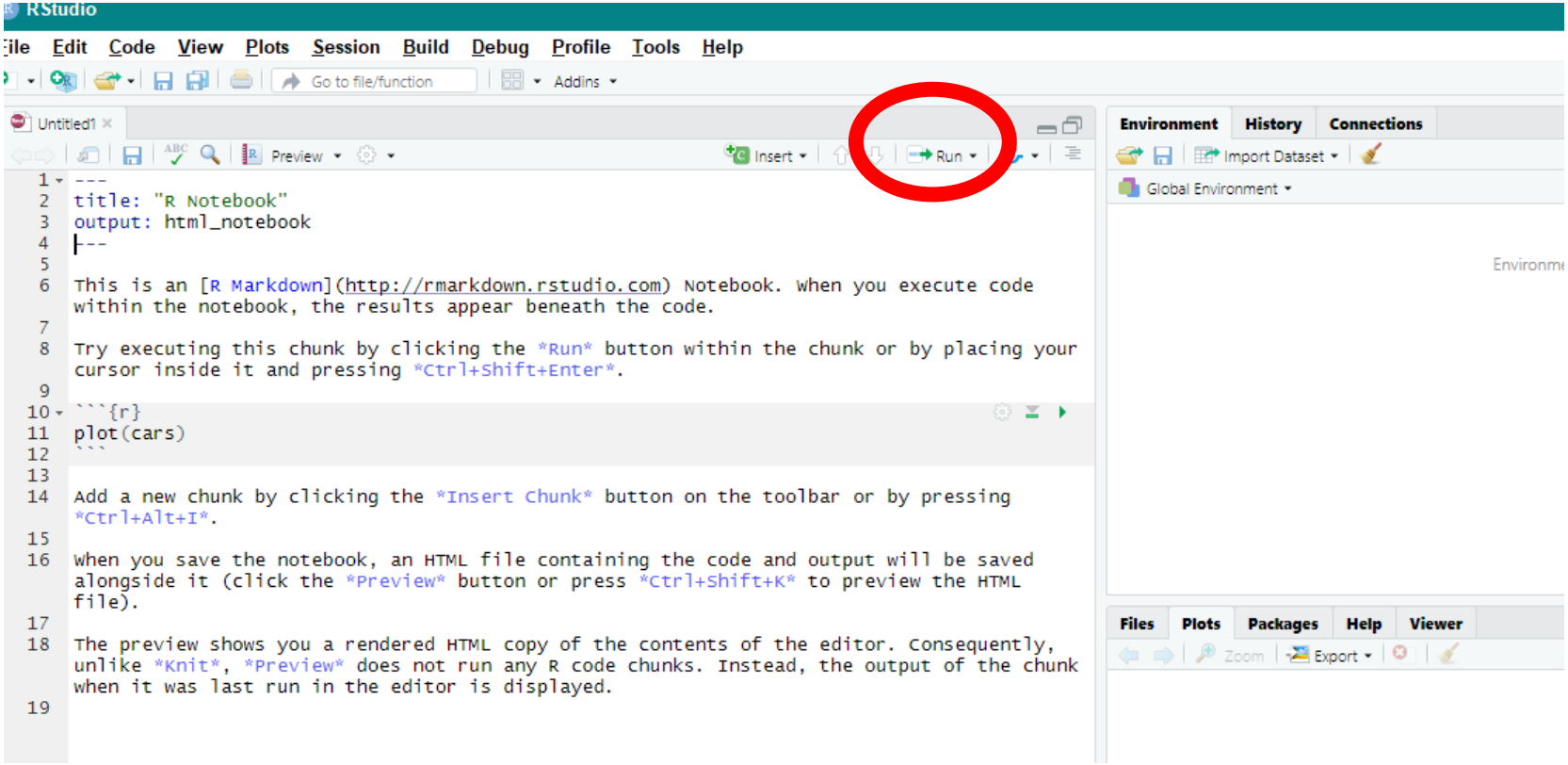
使用 R notebook

執行 1+2+3+4+5+6+7+8+9+10

R Notebook

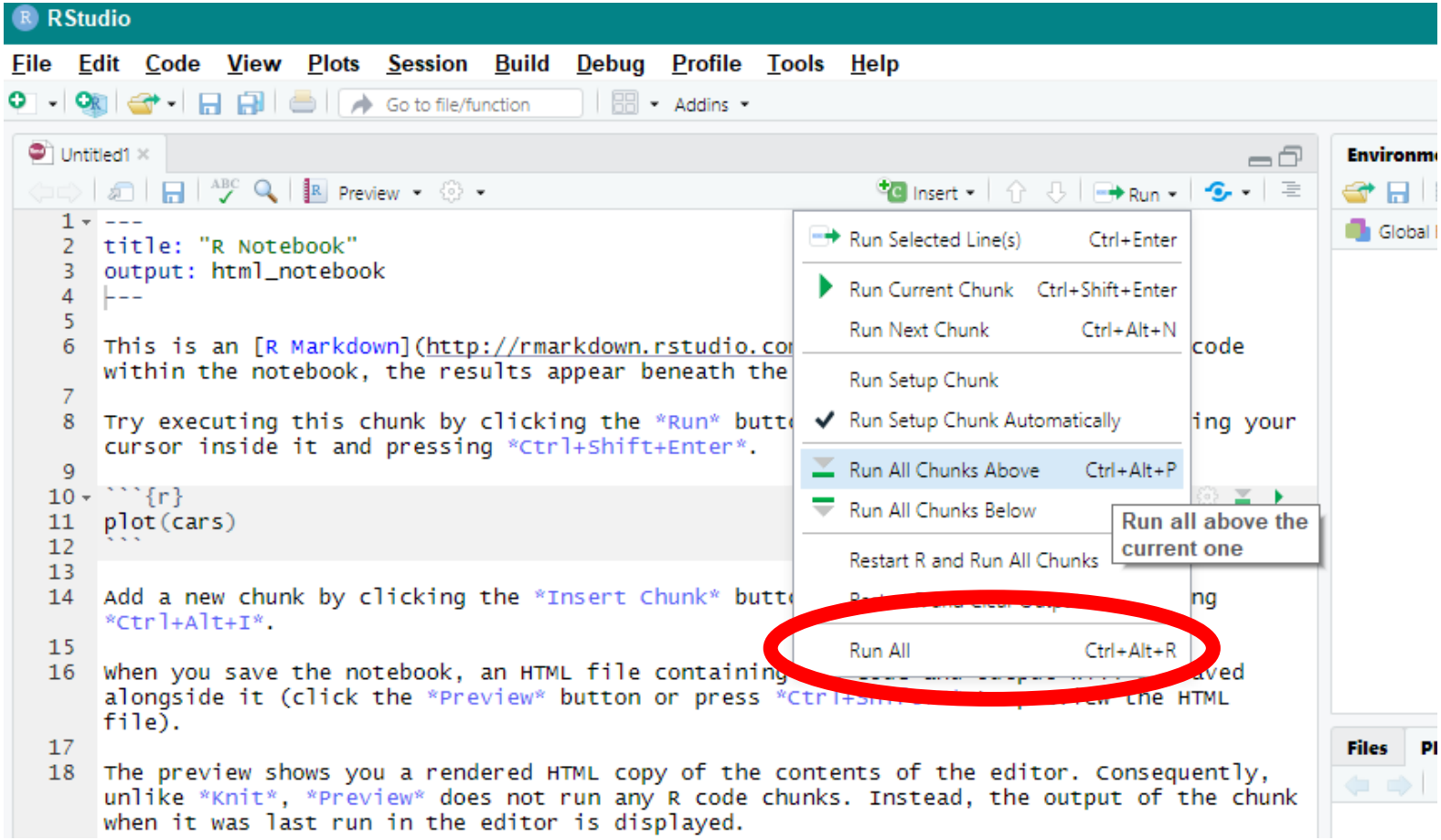

RStudio
File Edit Code View Plots Session Build Debug Profile Tools Help
New File
New Project...
Open File... Ctrl+O
Recent Files
Open Project...
Open Project in New Session...
Recent Projects
Import Dataset
Save Ctrl+S
Save As...
Save All Ctrl+Alt+S
Publish...
Print...
Close Ctrl+W
Close All Ctrl+Shift+W
Close All Except Current Ctrl+Alt+Shift+W
Close Project
Quit Session... Ctrl+Q
R Script Ctrl+Shift+N
R Notebook
R Markdown...
Shiny Web App...
Plumber API...
Text File
C++ File
Python Script
SQL Script
Stan File
D3 Script
R Sweave
R HTML
R Presentation
R Documentation


RStudio
File Edit Code View Plots Session Build Debug Profile Tools Help
Console Terminal Jobs
R version 3.5.3 (2019-03-11) -- "Great Truth"
Copyright (C) 2019 The R Foundation for Statist
Platform: x86_64-w64-mingw32/x64 (64-bit)
R is free software and comes with ABSOLUTELY NO
You are welcome to redistribute it under certa
Type 'license()' or 'licence()' for distributio
R is a collaborative project with many contribu
Type 'contributors()' for more information and
'citation()' on how to cite R or R packages in
Type 'demo()' for some demos, 'help()' for on-
'help.start()' for an HTML browser interface to
Type 'q()' to quit R.
Installing Packages
Stop
downloaded 645 KB
URL 'https://cran.rstudio.com/bin/windows/contrib/3.5/knitr_1.22.zip'
Content type 'application/zip' length 1469838 bytes (1.4 MB)
downloaded 1.4 MB
URL 'https://cran.rstudio.com/bin/windows/contrib/3.5/base64enc_0.1-3.zip'
Content type 'application/zip' length 43316 bytes (42 KB)
downloaded 42 KB
URL 'https://cran.rstudio.com/bin/windows/contrib/3.5/rprojroot_1.3-2.zip'
Content type 'application/zip' length 84015 bytes (82 KB)

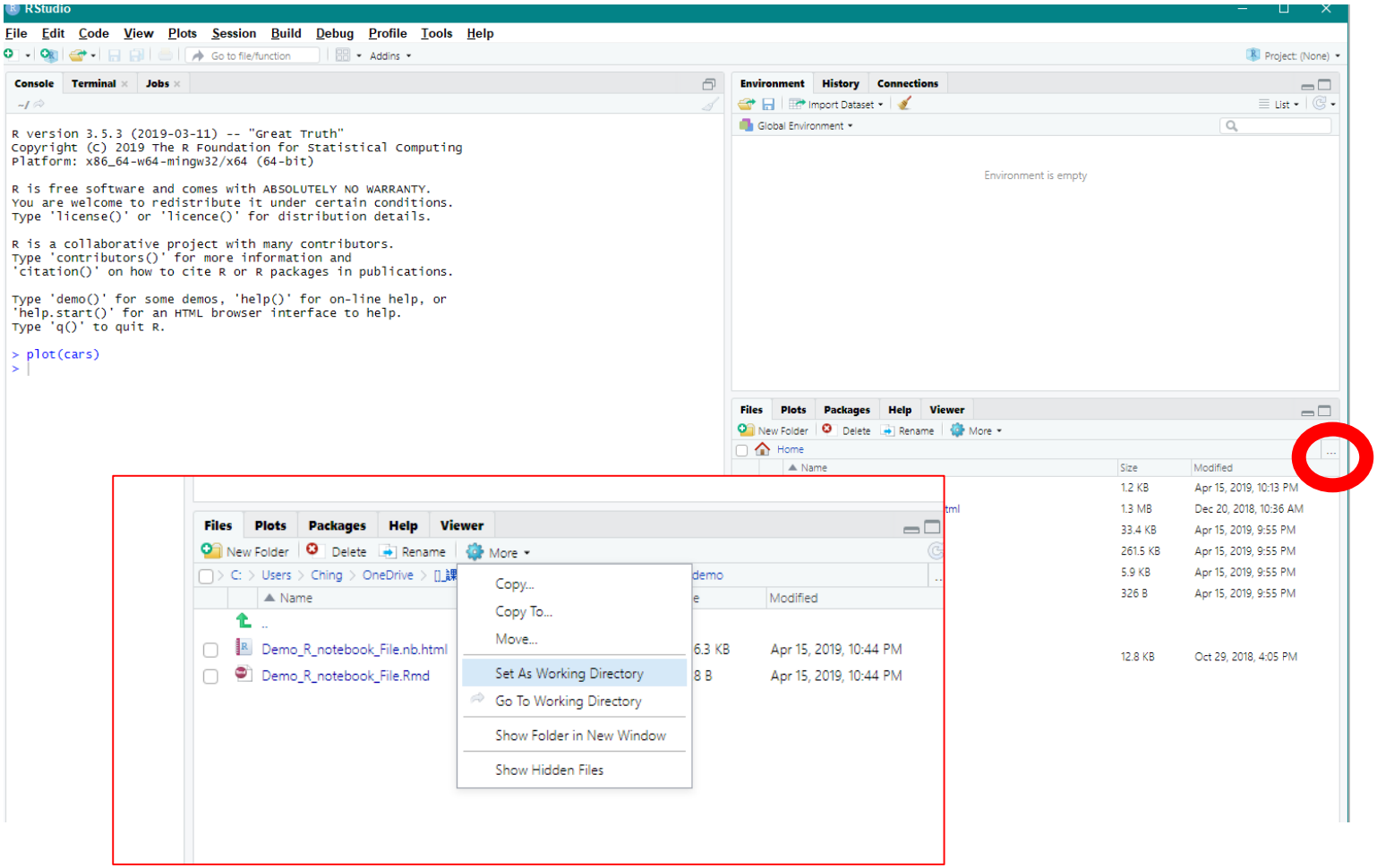

RStudio
File Edit Code View Plots Session Build Debug Profile Tools Help
Go to file/function
Addins
Untitled1
Preview
Insert
Run
---
title: "R Notebook"
output: html_notebook

This is an [R Markdown](http://rmarkdown.rstudio.com) Notebook. When you execute code within the notebook, the results appear beneath the code.
Try executing this chunk by clicking the *Run* button within the chunk or by placing your cursor inside it and pressing *Ctrl+Shift+Enter*.
```{r}
plot(cars)
```
Add a new chunk by clicking the *Insert Chunk* button on the toolbar or by pressing *Ctrl+Alt+I*.
When you save the notebook, an HTML file containing the code and output will be saved alongside it (click the *Preview* button or press *Ctrl+Shift+K* to preview the HTML file).
The preview shows you a rendered HTML copy of the contents of the editor. Consequently, unlike *Knit*, *Preview* does not run any R code chunks. Instead, the output of the chunk when it was last run in the editor is displayed.
Environment History Connections
Import Dataset
Global Environment
Files Plots Packages Help Viewer
Zoom
Export

執行 Run All


RStudio
File Edit Code View Plots Session Build Debug Profile Tools Help
Go to file/function
Addins
Untitled1
Preview
Insert
Run
Run Selected Line(s) Ctrl+Enter
Run Current Chunk Ctrl+Shift+Enter
Run Next Chunk Ctrl+Alt+N
Run Setup Chunk
Run Setup Chunk Automatically
Run All Chunks Above Ctrl+Alt+P
Run All Chunks Below
Restart R and Run All Chunks
Run All Ctrl+Alt+R
Run all above the current one
Files

指定工作目錄(資料夾)


RStudio
File Edit Code View Plots Session Build Debug Profile Tools Help
Console Terminal Jobs
R version 3.5.3 (2019-03-11) -- "Great Truth"
Copyright (C) 2019 The R Foundation for Statistical Computing
Platform: x86_64-w64-mingw32/x64 (64-bit)
R is free software and comes with ABSOLUTELY NO WARRANTY.
You are welcome to redistribute it under certain conditions.
Type 'license()' or 'licence()' for distribution details.
R is a collaborative project with many contributors.
Type 'contributors()' for more information and
'citation()' on how to cite R or R packages in publications.
Type 'demo()' for some demos, 'help()' for on-line help, or
'help.start()' for an HTML browser interface to help.
Type 'q()' to quit R.
> plot(cars)
Environment History Connections
Import Dataset
Global Environment
Environment is empty
Files Plots Packages Help Viewer
New Folder Delete Rename More
Home
Name Size Modified
1.2 KB Apr 15, 2019, 10:13 PM
1.3 MB Dec 20, 2018, 10:36 AM
33.4 KB Apr 15, 2019, 9:55 PM
261.5 KB Apr 15, 2019, 9:55 PM
5.9 KB Apr 15, 2019, 9:55 PM
326 B Apr 15, 2019, 9:55 PM
12.8 KB Oct 29, 2018, 4:05 PM
Files Plots Packages Help Viewer
New Folder Delete Rename More
C: Users Ching OneDrive demo
Name Modified
Demo_R_notebook_File.nb.html 6.3 KB Apr 15, 2019, 10:44 PM
Demo_R_notebook_File.Rmd 8 B Apr 15, 2019, 10:44 PM
Copy...
Copy To...
Move...
Set As Working Directory
Go To Working Directory
Show Folder in New Window
Show Hidden Files

執行結果

```
---
title: "R Notebook"
output: html_notebook
---


```{r}
require(maftools)

laml.maf = system.file('extdata', 'tcga_laml.maf.gz', package = 'maftools') #path to TCGA LAML MAF file
laml.clin = system.file('extdata', 'tcga_laml_annot.tsv', package = 'maftools') #clinical information containing survival information and histology. This is optional

laml = read.maf(maf = laml.maf, clinicalData = laml.clin)
```
```

R Console | data.table 6 x 2 | data.table 1241 x 11

Hugo_Symbol <chr>	Frame_Shift_Del <int>	Frame_Shift_Ins <int>	In_Frame_Del <int>	In_Frame_Ins <int>
FLT3	0	0	1	33
DNMT3A	4	0	0	0
NPM1	0	33	0	0
IDH2	0	0	0	0
IDH1	0	0	0	0
TET2	10	4	0	0
RUNX1	1	3	1	0
TP53	2	2	0	0
NRAS	0	0	0	0
CEBPA	6	2	2	6

1-10 of 1,241 rows | 1-5 of 11 columns Previous 1 2 3 4 5 6 ... 100 Next

Global Environment

Data

laml	Large MAF

Values

laml.clin	"C:/Users/
laml.maf	"C:/Users/

Files | Plots | Packages | Help | Viewer

New Folder | Delete | Rename | Mo

C: > Users > Ching > OneDrive > [] 課程演講

Name

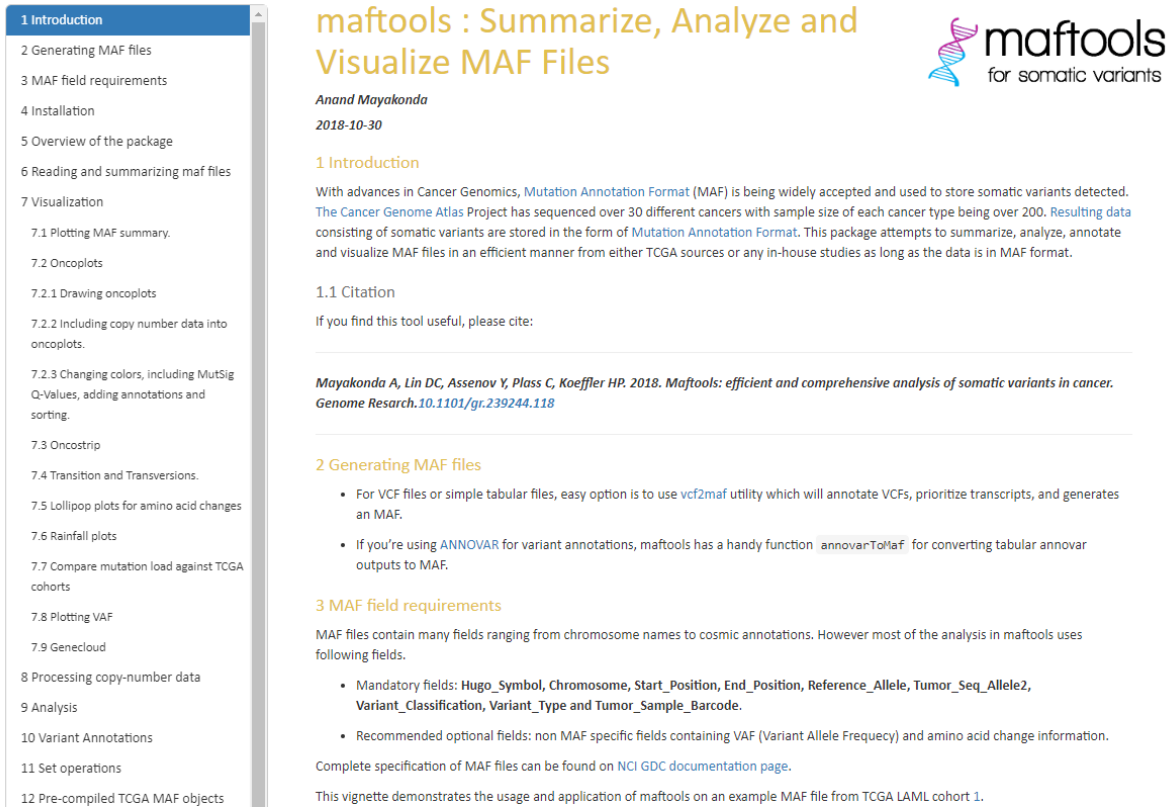
..

Demo_R_notebook_File.nb.html

Demo_R_notebook_File.Rmd

說明文件:

https://bioconductor.org/packages/release/bioc/vignettes/maftools/inst/doc/maftools.html



1 Introduction
2 Generating MAF files
3 MAF field requirements
4 Installation
5 Overview of the package
6 Reading and summarizing maf files
7 Visualization
7.1 Plotting MAF summary.
7.2 Oncoplots
7.2.1 Drawing oncoplots
7.2.2 Including copy number data into oncoplots.
7.2.3 Changing colors, including MutSig Q-Values, adding annotations and sorting.
7.3 Oncostrip
7.4 Transition and Transversions.
7.5 Lollipop plots for amino acid changes
7.6 Rainfall plots
7.7 Compare mutation load against TCGA cohorts
7.8 Plotting VAF
7.9 Genecloud
8 Processing copy-number data
9 Analysis
10 Variant Annotations
11 Set operations
12 Pre-compiled TCGA MAF objects

maftools : Summarize, Analyze and Visualize MAF Files

maftools for somatic variants

Anand Mayakonda

2018-10-30

1 Introduction

With advances in Cancer Genomics, Mutation Annotation Format (MAF) is being widely accepted and used to store somatic variants detected. The Cancer Genome Atlas Project has sequenced over 30 different cancers with sample size of each cancer type being over 200. Resulting data consisting of somatic variants are stored in the form of Mutation Annotation Format. This package attempts to summarize, analyze, annotate and visualize MAF files in an efficient manner from either TCGA sources or any in-house studies as long as the data is in MAF format.

1.1 Citation

If you find this tool useful, please cite:

Mayakonda A, Lin DC, Assenov Y, Plass C, Koeffler HP. 2018. Maftools: efficient and comprehensive analysis of somatic variants in cancer. Genome Resarch.10.1101/gr.239244.118

2 Generating MAF files

- For VCF files or simple tabular files, easy option is to use vcf2maf utility which will annotate VCFs, prioritize transcripts, and generates an MAF.
- If you're using ANNOVAR for variant annotations, maftools has a handy function `annovarToMaf` for converting tabular annovar outputs to MAF.

3 MAF field requirements

MAF files contain many fields ranging from chromosome names to cosmic annotations. However most of the analysis in maftools uses following fields.

- Mandatory fields: **Hugo_Symbol, Chromosome, Start_Position, End_Position, Reference_Allele, Tumor_Seq_Allele2, Variant_Classification, Variant_Type and Tumor_Sample_Barcode**.
- Recommended optional fields: non MAF specific fields containing VAF (Variant Allele Frequecy) and amino acid change information.

Complete specification of MAF files can be found on NCI GDC documentation page.

This vignette demonstrates the usage and application of maftools on an example MAF file from TCGA LAML cohort 1.

Overview of the package

