Bioinformatics I (KF) VU 041035 WS2024

http://icbi.at/mo

Hubert Hackl
Biocenter, Institute of Bioinformatics,
Medical University of Innsbruck,
Innrain 80, 6020 Innsbruck, Austria
Tel: +43-512-9003-71403,
Email: hubert.hackl@i-med.ac.at
URL: http://icbi.at

Contents

Nomenclature of nuclein acids

Base	Symbol	Occurrence
Adenin	A	DNA, RNA
Guanin	G	DNA, RNA
Cytosin	C	DNA, RNA
Thymin	T	DNA
Uracil	U	RNA

Symbol	Meaning	Description
R	A or G	pu**R**ine
Y	C or T	p**Y**rimidine
W	A or T	**W**eak hydrogen bonds
S	G or C	**S**trong hydrogen bonds
M	A or C	a**M**ino groups
K	G or T	**K**eto groups
H	A, C, or T (U)	not G, (**H** follows G)
B	G, C, or T (U)	not A, (**B** follows A)
V	G, A, or C	not T (U), (**V** follows U)
D	G, A, or T (U)	not C, (**D** follows C)
N	G, A, C or T (U)	a**N**y nucleotide

Nomenclature

DNA sequences are always from 5‘ to 3‘

```
+ strand      5´-ACGGTCGCTGTCGGTAGC-3´
- strand      3´-TGCCAGCGACAGCCATCG-5´
```

e.g. in fasta format :

```
>gene sequence|gi12345|chr17|-
GCTACCGACAGCGACCGT
```

Positions in the genome (genome assembly) are chromosome wise

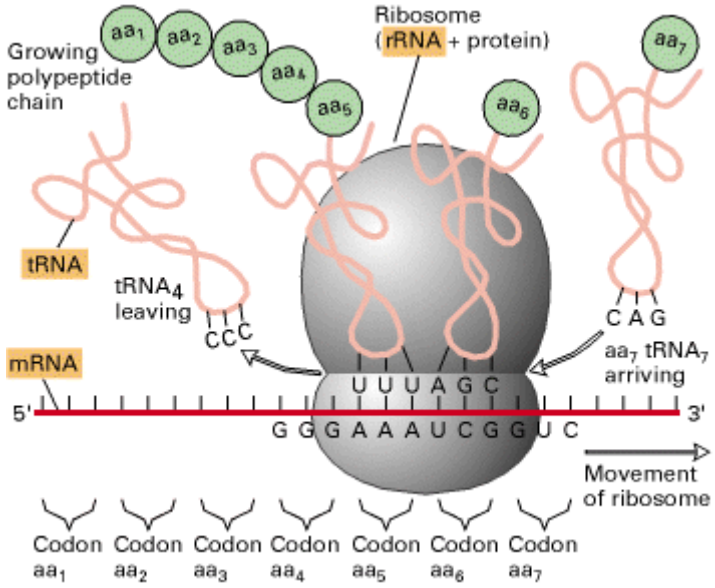
e.g. human GRCh37/hg19

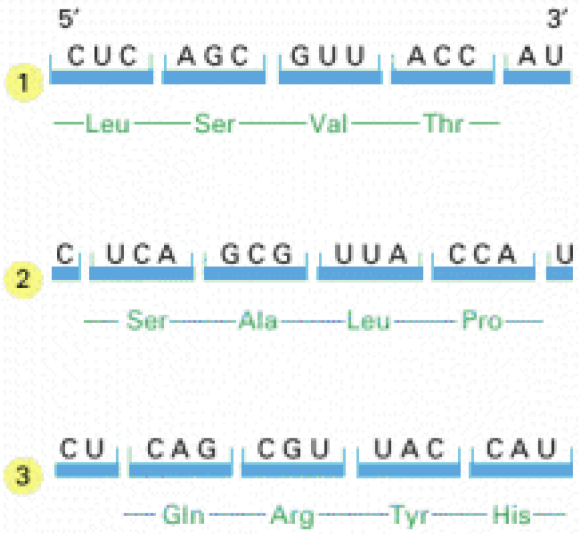
chr11:1-100 *chr11:49,686,777-49,689,777*

Positions in the chromosome start for **both!!** strands from position 1

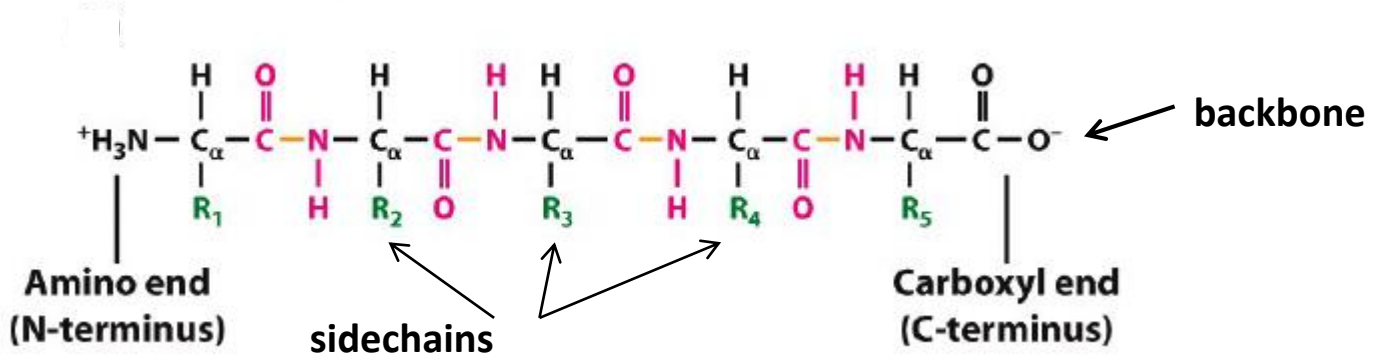
Translation, genetic code and reading frames

GCA GCC GCG GCU	AGA AGG CGA CGC CGG CGU	GAC GAU	AAC AAU	UGC UGU	GAA GAG	CAA CAG	GGA GGC GGG GGU	CAC CAU	AUA AUC AUU	UUA UUG CUA CUC CUG CUU	AAA AAG	AUG	UUC UUU	CCA CCC CCG CCU	AGC AGU UCA UCC UCG UCU	ACA ACC ACG ACU	UGG	UAC UAU	GUA GUC GUG GUU	UAA UAG UGA
Ala	Arg	Asp	Asn	Cys	Glu	Gln	Gly	His	Ile	Leu	Lys	Met	Phe	Pro	Ser	Thr	Trp	Tyr	Val	stop
A	R	D	N	C	E	Q	G	H	I	L	K	M	F	P	S	T	W	Y	V	





Peptid chain, amino acid sequence, proteins

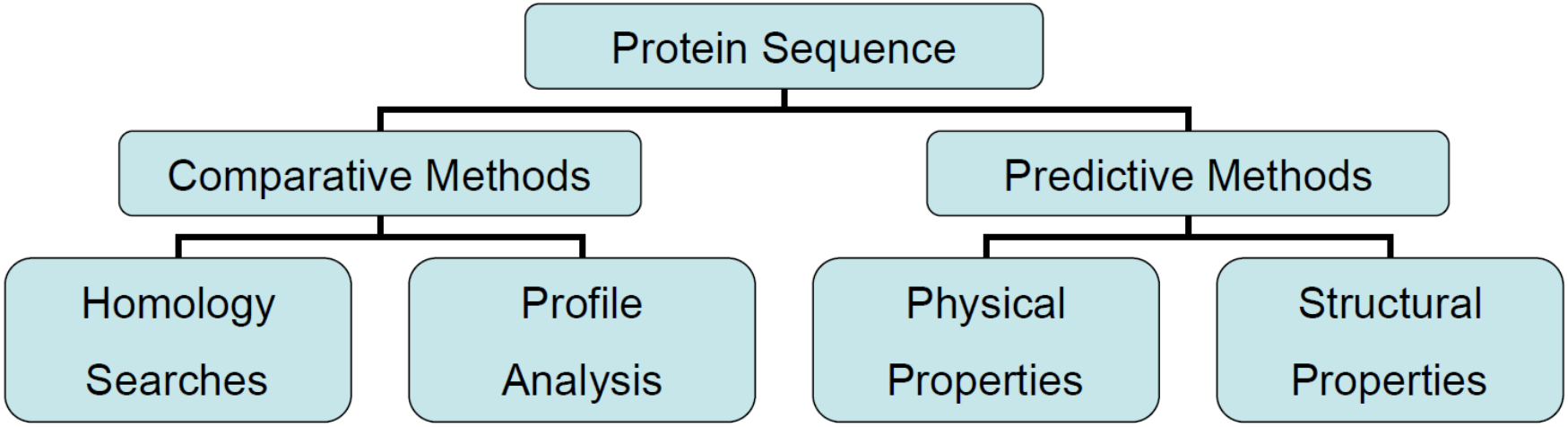


Protein sequences are always form N-terminal end to C-terminal end

E.g.. SCD sequence in fasta format

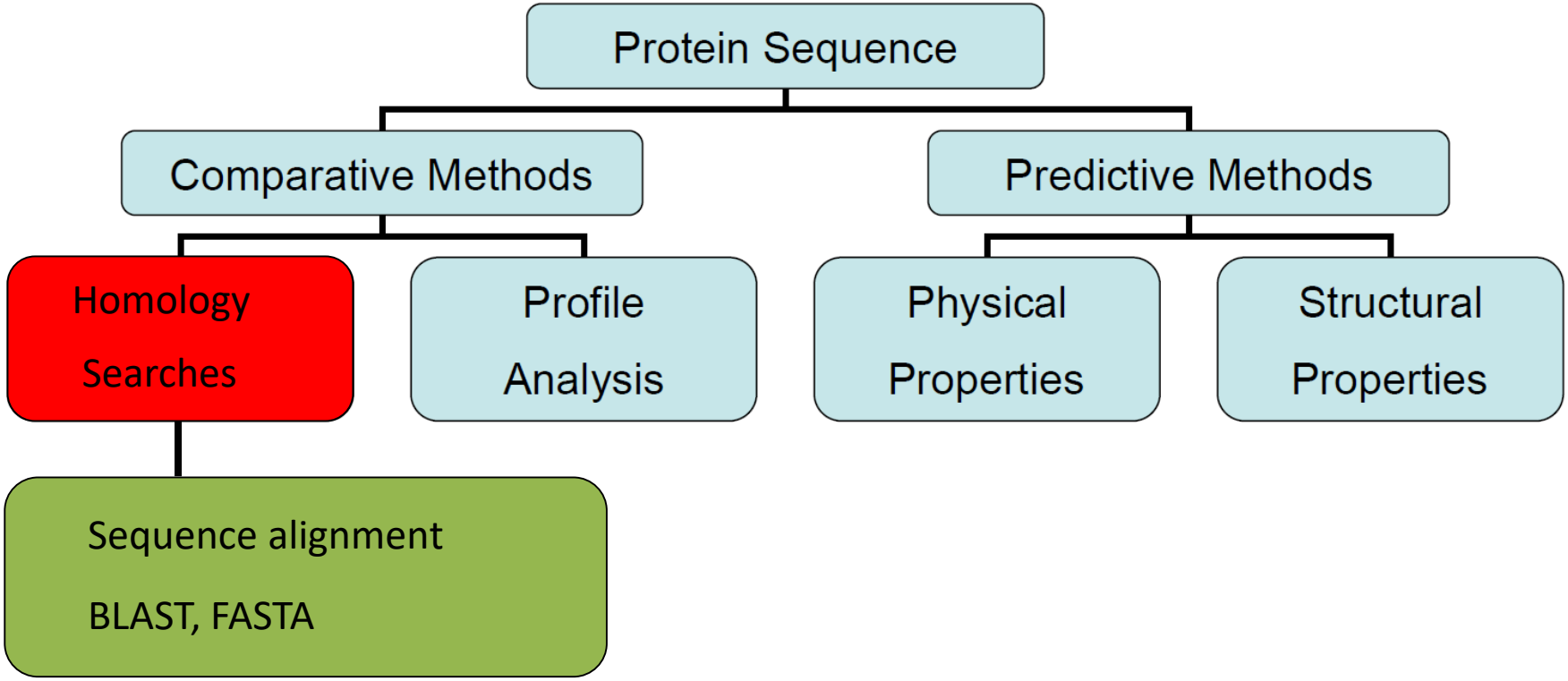
```
>gi|53759151|ref|NP_005054.3| acyl-CoA desaturase [Homo sapiens]
MPAHLLQDDISSSYTTTTTITAPPSRVLQNGGDKLETMPLYLEDDIRPDIKDDIYDPTYKDKEGPSPKVE
YVWRNIILMSLLHLGALYGITLIPTCKFYTWLWGVFYYFVSALGITAGAHRLWSHRSYKARLPLRLFLII
ANTMAFQNDVYEWARDHRAHHKFSETHADPHNSRRGFFFSHVGWLLVRKHPAVKEKGSTLDLSDLEAEKL
VMFQRRYYKPGLLMMCFILPTLVPWYFWGETFQNSVFVATFLRYAVVLNATWLVNSAAHLFGYRPYDKNI
SPRENILVSLGAVGEGFHNYHHSFPYDYSASEYRWHINFTTFFIDCMAALGLAYDRKKVSKAAILARIKR
TGDGNYKSG
```

Protein Sequence Analysis

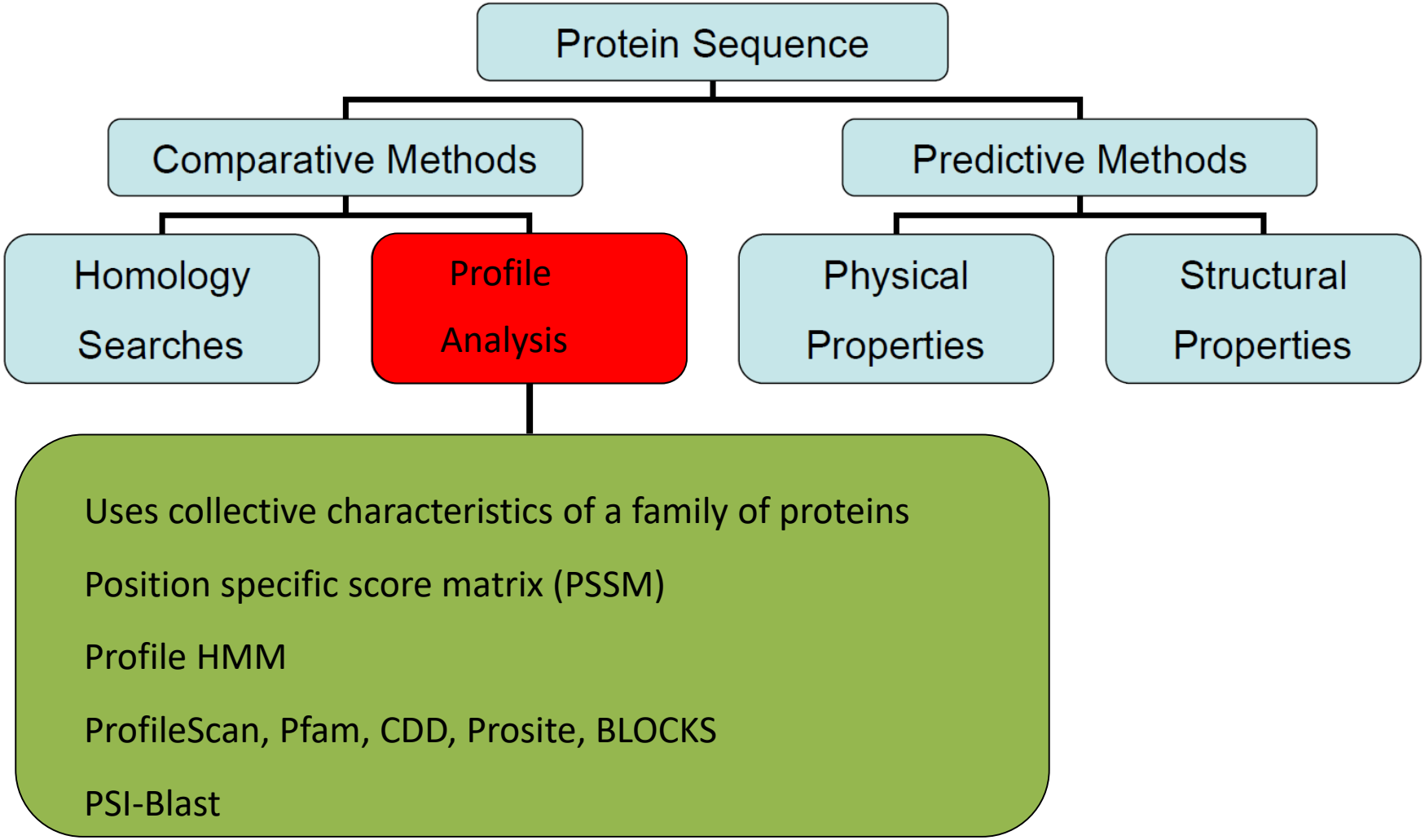


Shared ancestry?
Similar function?
Domain or complete sequence?
Are functional sequences conserved?

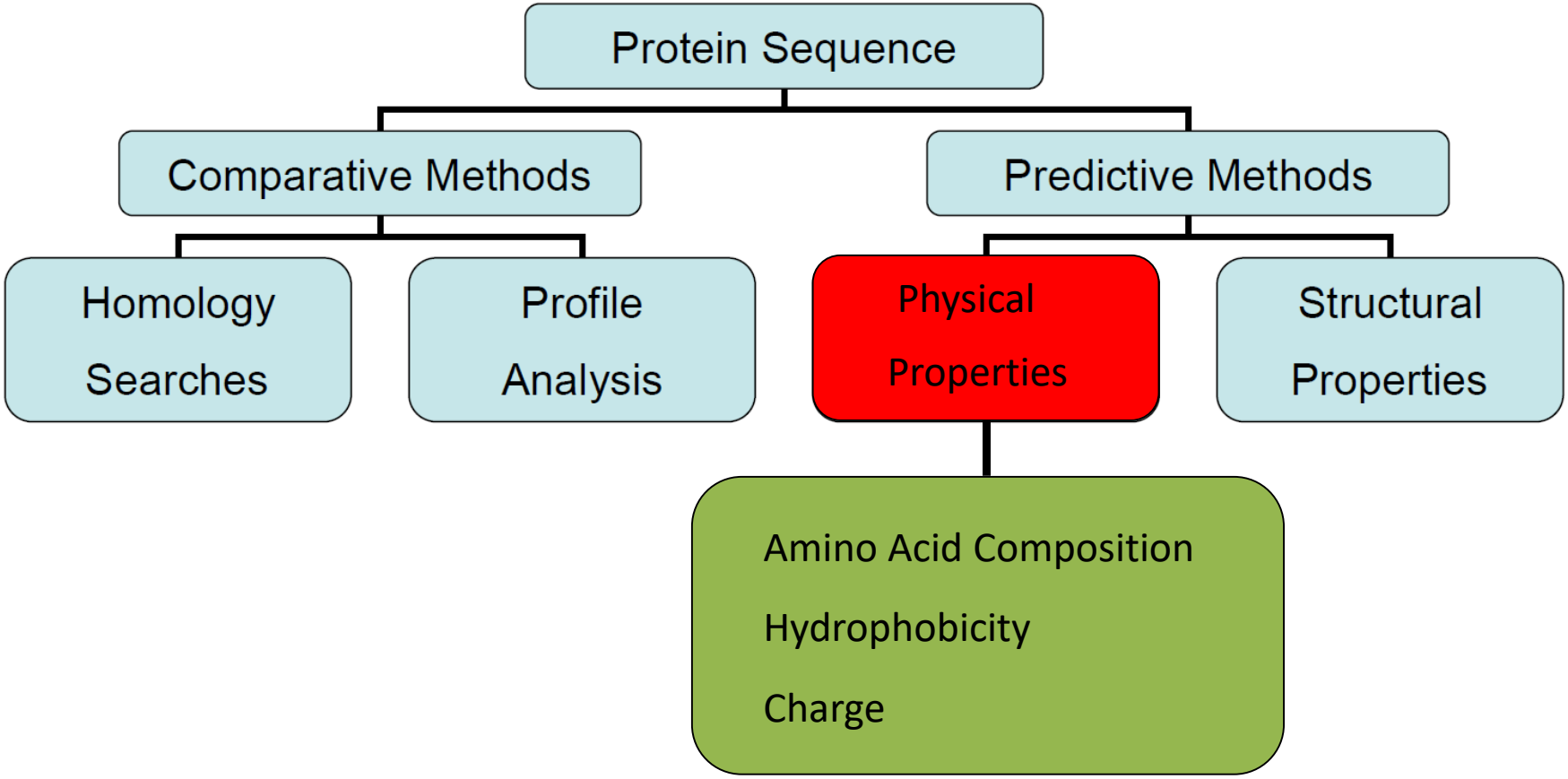
Homology searches



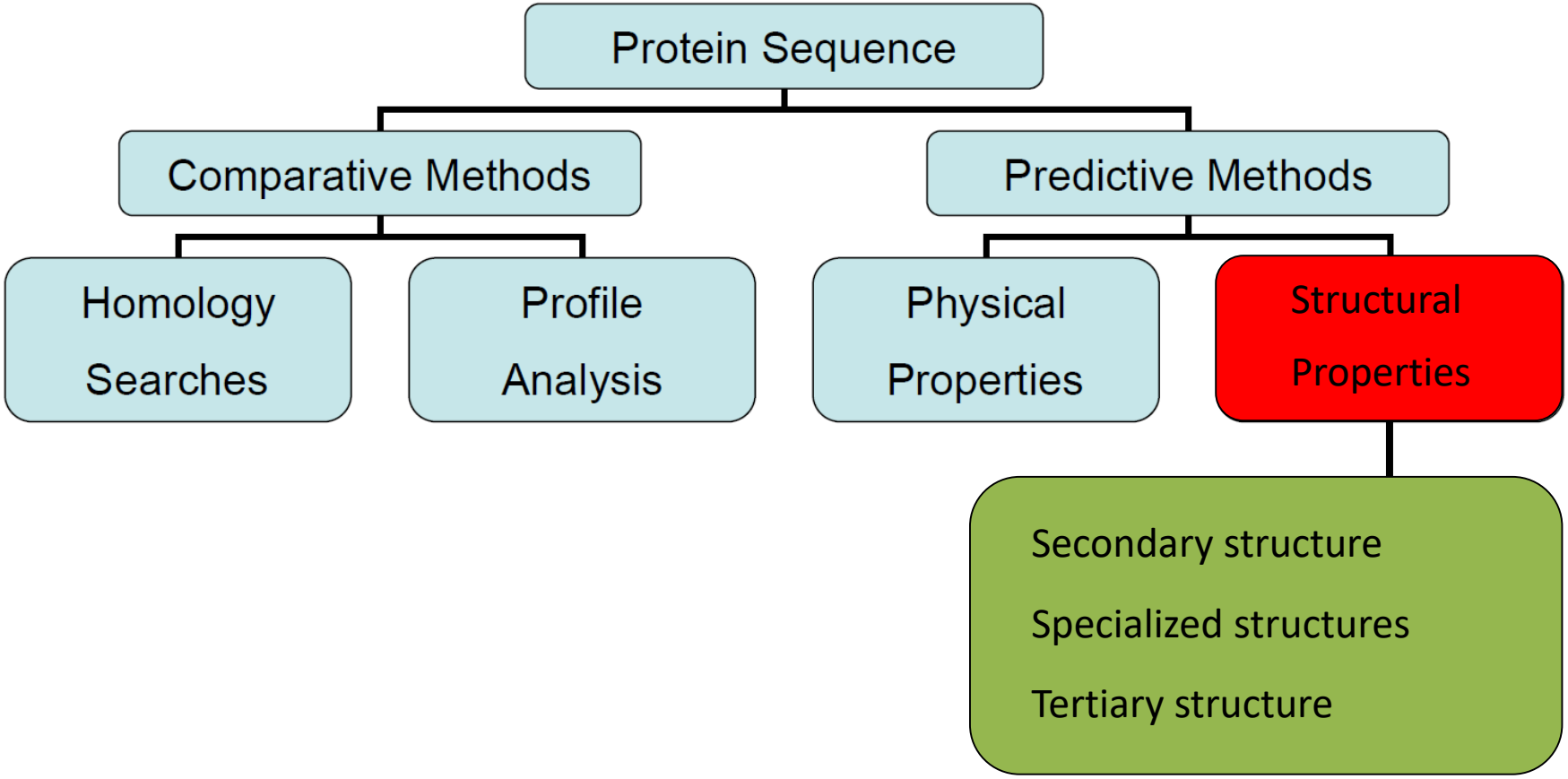
Profile Analysis



Protein Sequence Analysis



Protein Sequence Analysis



Substitutions matrices

- Unrelated or random model assumes that letter a occurs independently with some frequency q_a.

$$P(x,y|R) = \prod q_{xi} \prod q_{xj}$$

- The alternative match model of aligned pairs of residues occurs with a joint probability p_{ab}.

$$P(x,y|M) = \prod p_{xi\ yi}$$

- Odds ratio

$$\frac{P(x,y|M)}{P(x,y|R)} = \frac{\prod p_{xi\ yi}}{\prod q_{xi} \prod q_{yj}} = \prod \frac{p_{xi\ yi}}{q_{xi}\, q_{yj}}$$

BLOSSUM62

	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V	B	Z	X	*
A	4	-1	-2	-2	0	-1	-1	0	-2	-1	-1	-1	-1	-2	-1	1	0	-3	-2	0	-2	-1	0	-4
R	-1	5	0	-2	-3	1	0	-2	0	-3	-2	2	-1	-3	-2	-1	-1	-3	-2	-3	-1	0	-1	-4
N	-2	0	6	1	-3	0	0	0	1	-3	-3	0	-2	-3	-2	1	0	-4	-2	-3	3	0	-1	-4
D	-2	-2	1	6	-3	0	2	-1	-1	-3	-4	-1	-3	-3	-1	0	-1	-4	-3	-3	4	1	-1	-4
C	0	-3	-3	-3	9	-3	-4	-3	-3	-1	-1	-3	-1	-2	-3	-1	-1	-2	-2	-1	-3	-3	-2	-4
Q	-1	1	0	0	-3	5	2	-2	0	-3	-2	1	0	-3	-1	0	-1	-2	-1	-2	0	3	-1	-4
E	-1	0	0	2	-4	2	5	-2	0	-3	-3	1	-2	-3	-1	0	-1	-3	-2	-2	1	4	-1	-4
G	0	-2	0	-1	-3	-2	-2	6	-2	-4	-4	-2	-3	-3	-2	0	-2	-2	-3	-3	-1	-2	-1	-4
H	-2	0	1	-1	-3	0	0	-2	8	-3	-3	-1	-2	-1	-2	-1	-2	-2	2	-3	0	0	-1	-4
I	-1	-3	-3	-3	-1	-3	-3	-4	-3	4	2	-3	1	0	-3	-2	-1	-3	-1	3	-3	-3	-1	-4
L	-1	-2	-3	-4	-1	-2	-3	-4	-3	2	4	-2	2	0	-3	-2	-1	-2	-1	1	-4	-3	-1	-4
K	-1	2	0	-1	-3	1	1	-2	-1	-3	-2	5	-1	-3	-1	0	-1	-3	-2	-2	0	1	-1	-4
M	-1	-1	-2	-3	-1	0	-2	-3	-2	1	2	-1	5	0	-2	-1	-1	-1	-1	1	-3	-1	-1	-4
F	-2	-3	-3	-3	-2	-3	-3	-3	-1	0	0	-3	0	6	-4	-2	-2	1	3	-1	-3	-3	-1	-4
P	-1	-2	-2	-1	-3	-1	-1	-2	-2	-3	-3	-1	-2	-4	7	-1	-1	-4	-3	-2	-2	-1	-2	-4
S	1	-1	1	0	-1	0	0	0	-1	-2	-2	0	-1	-2	-1	4	1	-3	-2	-2	0	0	0	-4
T	0	-1	0	-1	-1	-1	-1	-2	-2	-1	-1	-1	-1	-2	-1	1	5	-2	-2	0	-1	-1	0	-4
W	-3	-3	-4	-4	-2	-2	-3	-2	-2	-3	-2	-3	-1	1	-4	-3	-2	11	2	-3	-4	-3	-2	-4
Y	-2	-2	-2	-3	-2	-1	-2	-3	2	-1	-1	-2	-1	3	-3	-2	-2	2	7	-1	-3	-2	-1	-4
V	0	-3	-3	-3	-1	-2	-2	-3	-3	3	1	-2	1	-1	-2	-2	0	-3	-1	4	-3	-2	-1	-4
B	-2	-1	3	4	-3	0	1	-1	0	-3	-4	0	-3	-3	-2	0	-1	-4	-3	-3	4	1	-1	-4
Z	-1	0	0	1	-3	3	4	-2	0	-3	-3	1	-1	-3	-1	0	-1	-3	-2	-2	1	4	-1	-4
X	0	-1	-1	-1	-2	-1	-1	-1	-1	-1	-1	-1	-1	-1	-2	0	0	-2	-1	-1	-1	-1	-1	-4
*	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	1

Database search

- Database:

  AIKWQPRSTW...

  IKMQRHIKW...

  HDLFWHLWH...

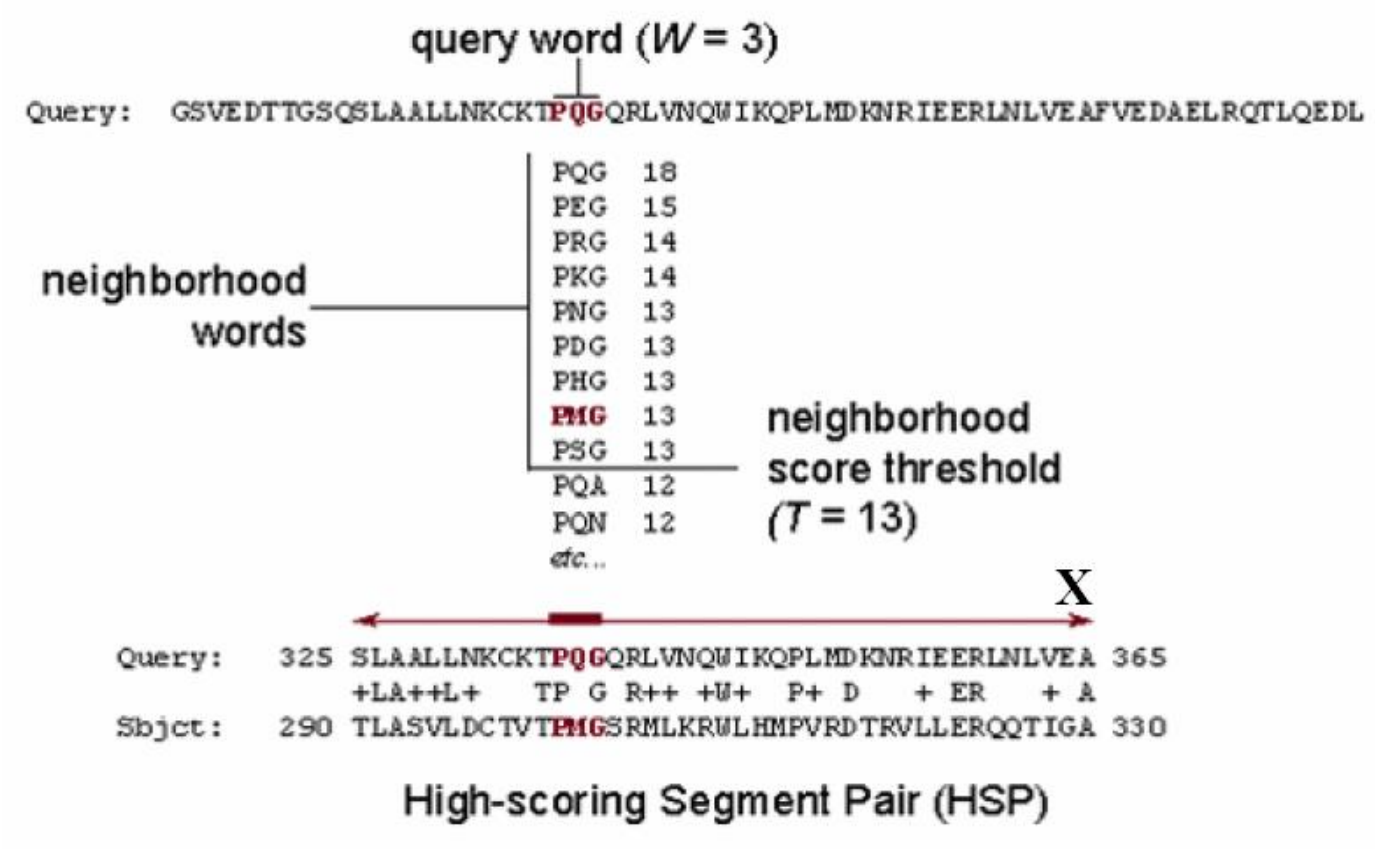
  .....

- Query:

  RGIKW

- Output: sequences *similar* to query

High-scoring segment pairs



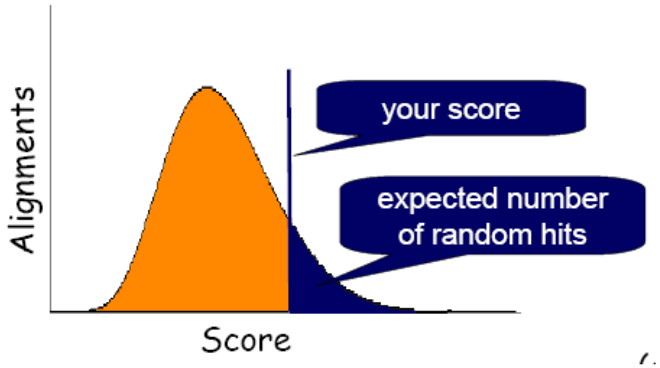
Significance of scores

The number of unrelated matches with score greater than S is approximately Poisson distributed with mean

$$E(S)=Kmne^{-\lambda S}$$

where λ is a scaling factor m and n are the length of the sequences

The probability that there is a match of score greater than S follows a extreme value distribution:



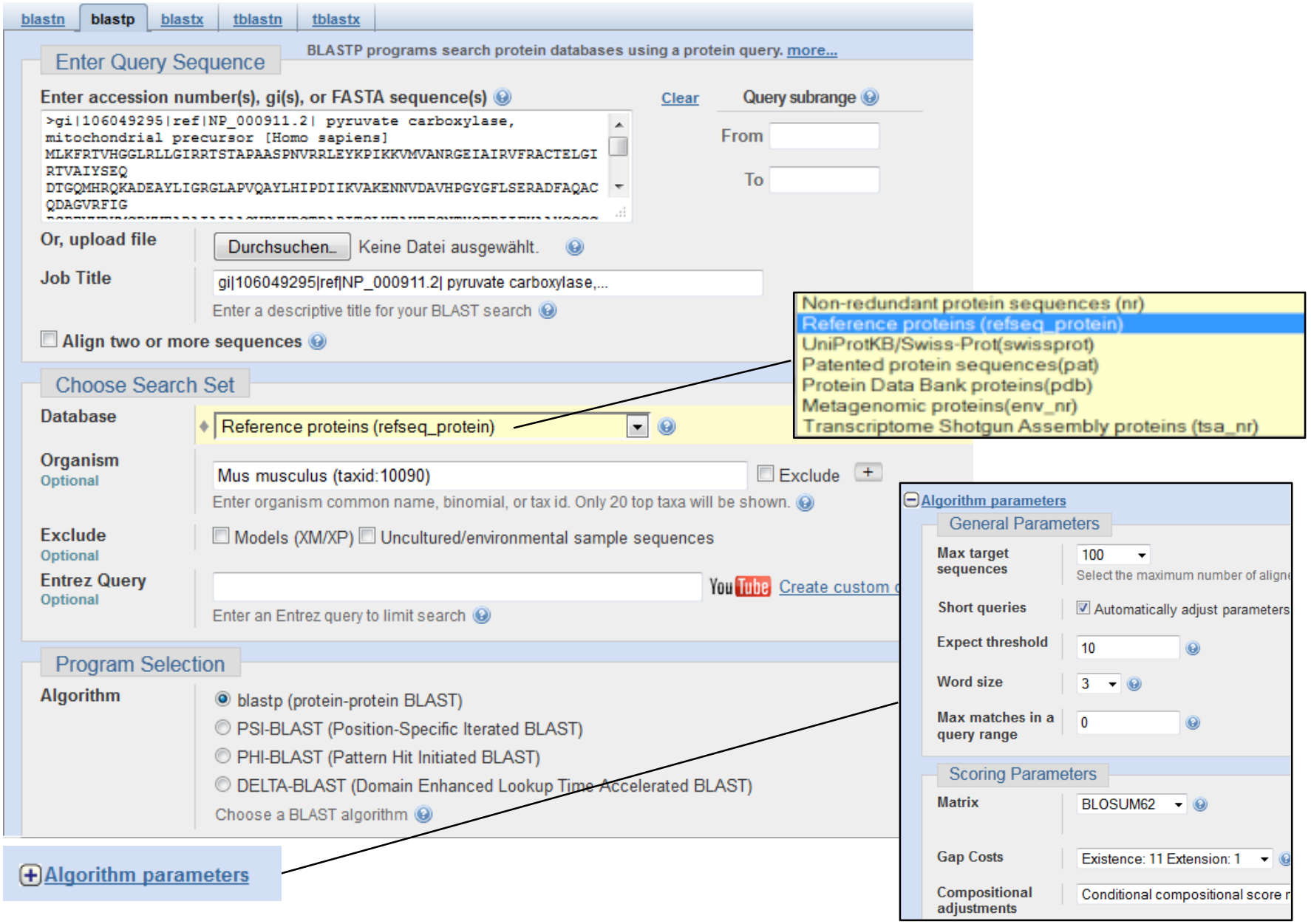
$$P(x>S)=1-e^{-E(S)}$$

Karlin S, Altschul S. *Proc Natl Acad Sci* (1990)

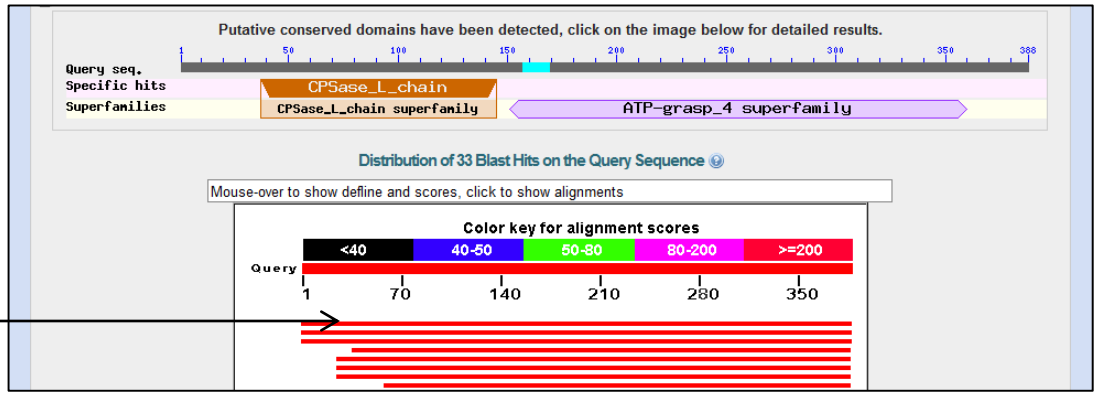
NCBI Blast

Program	*Query sequence*	*Subject sequence*
BLASTN	Nucleotide	Nucleotide
BLASTP	Protein	Protein
BLASTX	Nucleotide six-frame translation	Protein
TBLASTN	Protein	Nucleotide six-frame translation
TBLASTX	Nucleotide six-frame translation	Nucleotide six-frame translation

NCBI Blast Example



Blast Results



conserved domain database (CDD)

graphical visualization

Best hit

Description	Max score	Total score	Query cover	E value	Ident	Accession
pyruvate carboxylase, mitochondrial isoform 1 [Mus musculus]	781	781	100%	0.0	96%	NP_001156418.1
pyruvate carboxylase, mitochondrial isoform 2 [Mus musculus] >ref|XP_006531741.1| PREDICT	780	780	100%	0.0	96%	NP_032823.2
PREDICTED: pyruvate carboxylase, mitochondrial isoform X1 [Mus musculus]	780	780	100%	0.0	96%	XP_006531740.1
methylcrotonoyl-CoA carboxylase subunit alpha, mitochondrial [Mus musculus]	330	330	90%	5e-105	48%	NP_076133.3
PREDICTED: propionyl-CoA carboxylase alpha chain, mitochondrial isoform X4 [Mus musculus]	322	322	93%	1e-103	47%	XP_006518496.1

description

E-value

Score (S)

pyruvate carboxylase, mitochondrial isoform 1 [Mus musculus]

Sequence ID: ref|NP_001156418.1| Length: 1179 Number of Matches: 1

Range 1: 2 to 389 GenPept Graphics Next Match Previous Match

Score	Expect	Method	Identities	Positives	Gaps
781 bits(2016)	0.0	Compositional matrix adjust.	374/388(96%)	384/388(98%)	0/388(0%)

```
Query  1    MLKFRTVHGGLRLLGIRRTSTAPAASPNVRRLEYKPIKKVMVANRGEIAIRVFRACTELG  60
            MLKF+TV GGLRLLG+RR+S+AP ASPNVRRLEYKPIKKVMVANRGEIAIRVFRACTELG
Sbjct  2    MLKFQTVRGGLRLLGVRRSSSAPVASPNVRRLEYKPIKKVMVANRGEIAIRVFRACTELG  61
                                    ...

Query  361  GLRQENIRINGCAIQCRVTTEDPARSFQ  388
            GLRQENIRINGCAIQCRVTTEDPARSFQ
Sbjct  362  GLRQENIRINGCAIQCRVTTEDPARSFQ  389
```

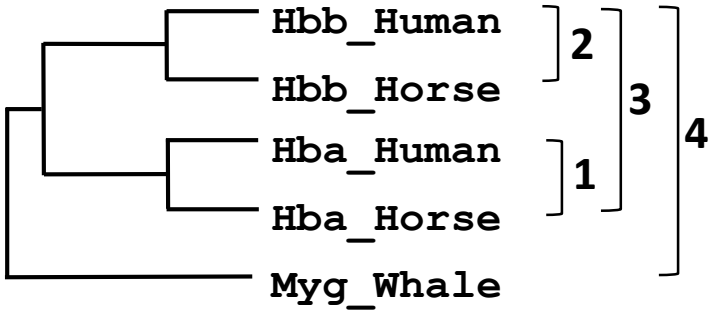
...

alignment

Multiple sequence alignment (Clustal W)

```
Hbb_Human   -
Hbb_Horse  .17   -
Hba_Human  .59  .60   -
Hba_Horse  .59  .59  .13   -
Myg_Whale  .77  .77  .75  .75   -
```

Pairwise alignment
calculate distance matrix



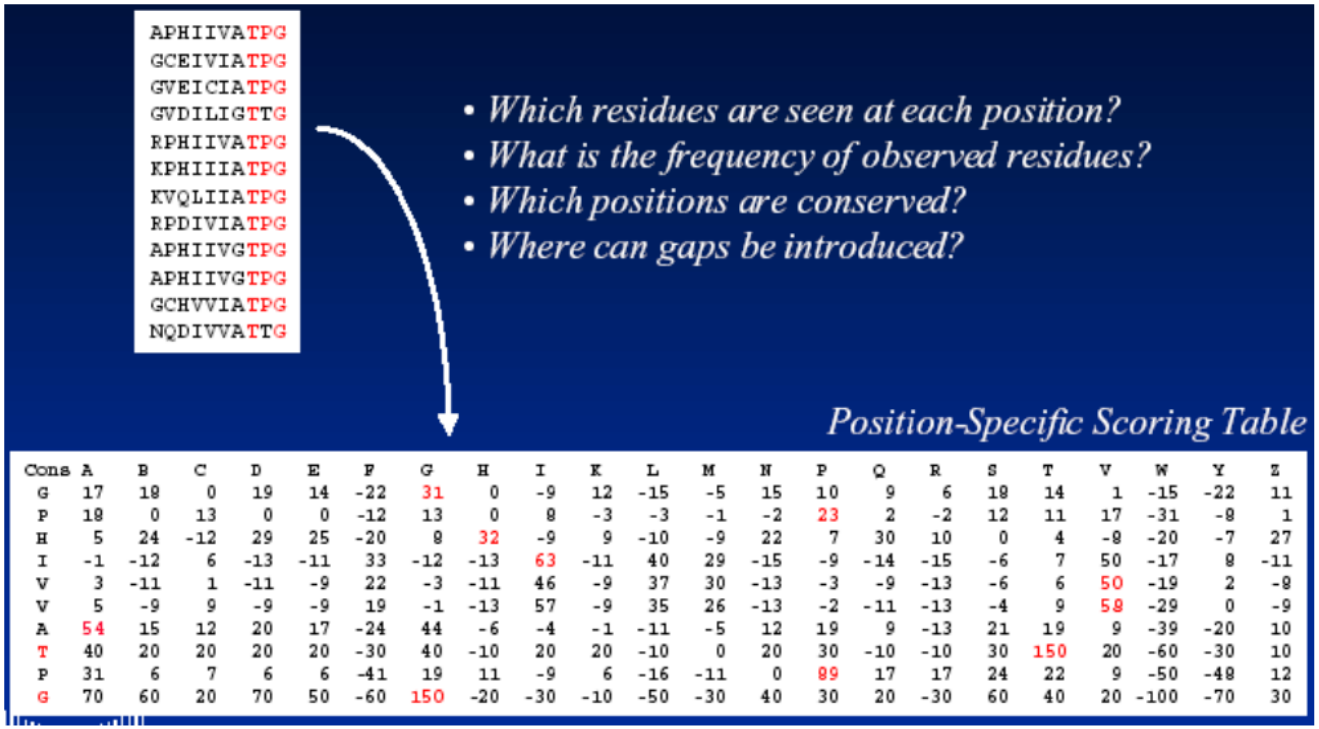
Rooted neighbor-joining tree
(guide tree) and sequence weights

```
1   PEEKSAVTALWGKVN--VDEVGG
2   GEEKAAVLALWDKVN--EEEVGG
3   PADKTNVKAAWGKVGAHAGEYGA
4   AADKTNVKAAWSKVGGHAGEYGA
5   EHEWQLVLHVWAKVEADVAGHGQ
```

2 3 4 1

Progressive alignment
following guide tree

Profile Construction



Cons	A	B	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	Y	Z
G	17	18	0	19	14	-22	31	0	-9	12	-15	-5	15	10	9	6	18	14	1	-15	-22	11
P	18	0	13	0	0	-12	13	0	8	-3	-3	-1	-2	23	2	-2	12	11	17	-31	-8	1
H	5	24	-12	29	25	-20	8	32	-9	9	-10	-9	22	7	30	10	0	4	-8	-20	-7	27
I	-1	-12	6	-13	-11	33	-12	-13	63	-11	40	29	-15	-9	-14	-15	-6	7	50	-17	8	-11
V	3	-11	1	-11	-9	22	-3	-11	46	-9	37	30	-13	-3	-9	-13	-6	6	50	-19	2	-8
V	5	-9	9	-9	-9	19	-1	-13	57	-9	35	26	-13	-2	-11	-13	-4	9	58	-29	0	-9
A	54	15	12	20	17	-24	44	-6	-4	-1	-11	-5	12	19	9	-13	21	19	9	-39	-20	10
T	40	20	20	20	20	-30	40	-10	20	20	-10	0	20	30	-10	-10	30	150	20	-60	-30	10
P	31	6	7	6	6	-41	19	11	-9	6	-16	-11	0	89	17	17	24	22	9	-50	-48	12
G	70	60	20	70	50	-60	150	-20	-30	-10	-50	-30	40	30	20	-30	60	40	20	-100	-70	30

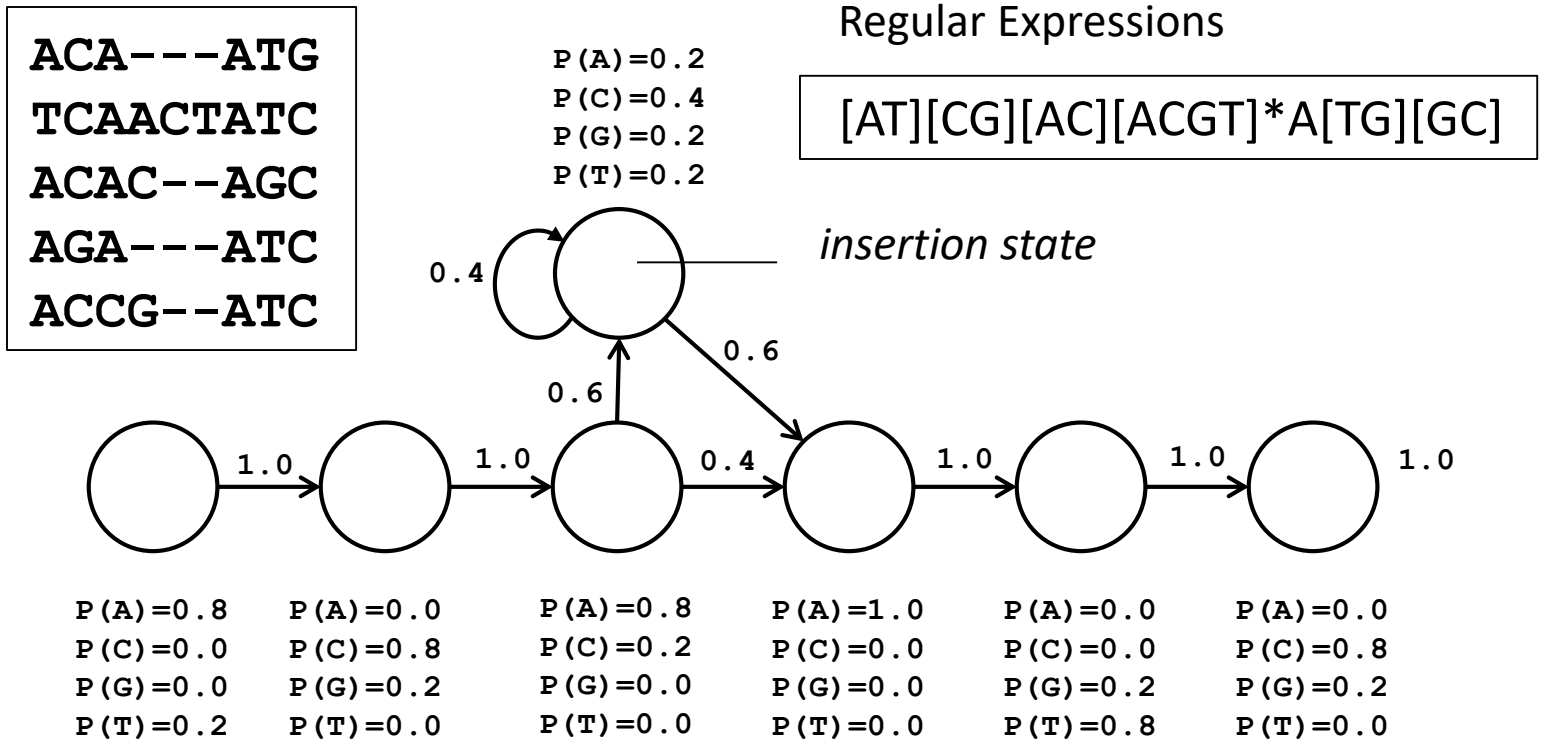
$$PSSM(p,a) = \sum_{b=1}^{20} f(p,b)*s(a,b)$$

f(p,b) = frequency of amino acid b in position p

s(a,b) is the score of (a,b) (from, e.g., BLOSUM or PAM)

Profile Hidden Markov Model

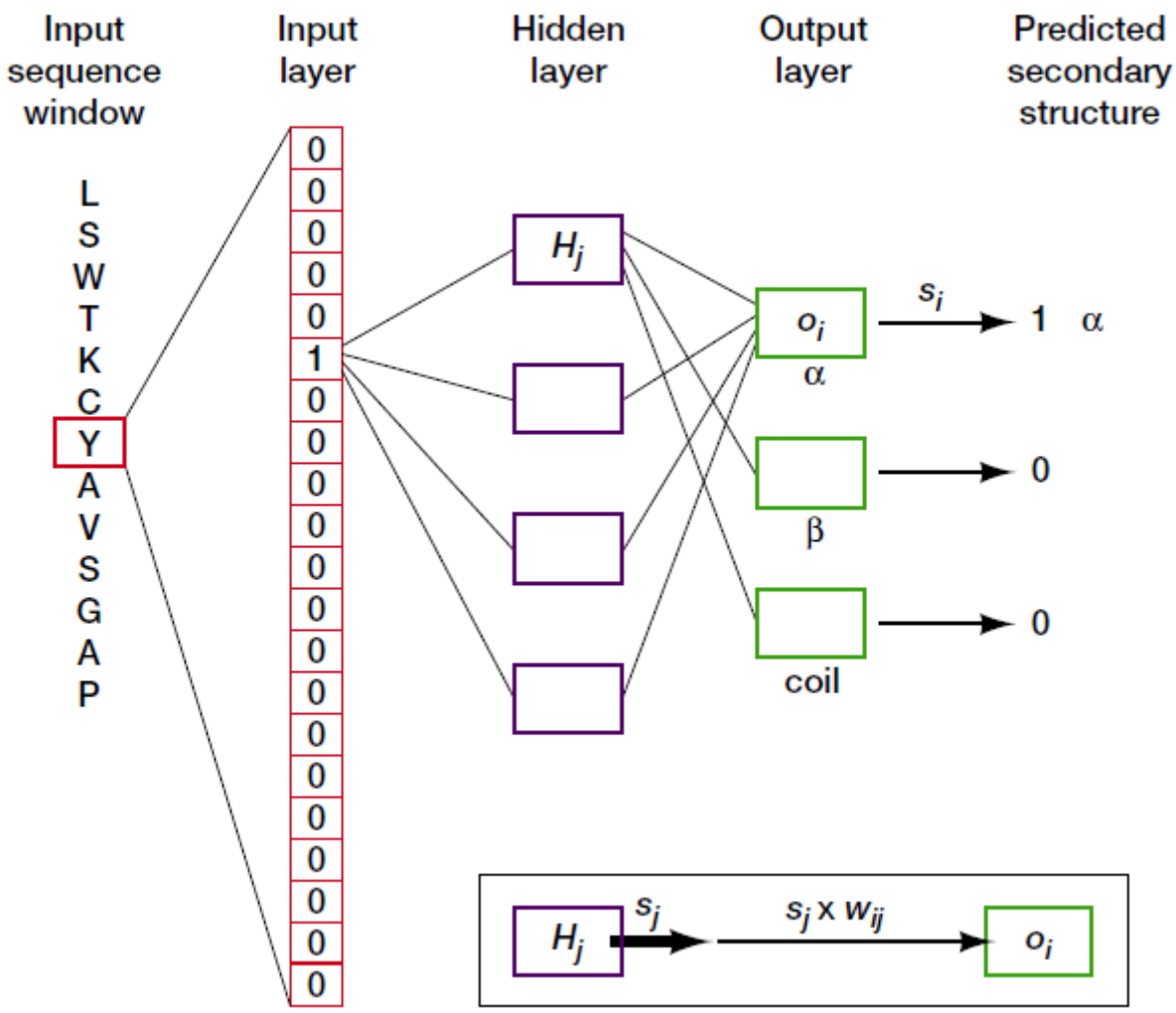
- For multiple alignments (e.g. DNA sequences)



p(*ACACATC*)=0.8*1*0.8*1*0.8*0.6 *0.4*0.6*1*1*0.8*1*0.8=0.047

log-odds=log(p(S)/0.25^L)=log(0.047/0.25^7)

Neuronal network for secondary structure prediction



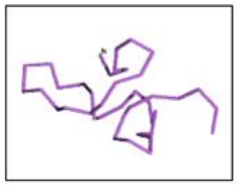
PredictProtein

- Multi-step predictive algorithm (Rost et al., 1994)
  - Protein sequence queried against SWISS-PROT
    MaxHom used to generate iterative, profile-based
    Multiple sequence alignment (Sander and Schneider,1991)
  - Multiple alignment fed into neural network (PHDsec)
- Accuracy: Average > 70%, Best-case > 90%
- http://www.predictprotein.org/

Prediction of protein function

Flexible regions

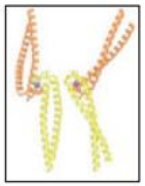
no inherent structure, bias to small/polar AA



Compositional bias
CAST, DisEMBL, GlobPlot
SAPS, SEG, XNU
Targeting signals and PTMs
Big-_
MyPS/NMT, PrePS
PTS1
SignalP, Sigcleave

Fibrillar domains

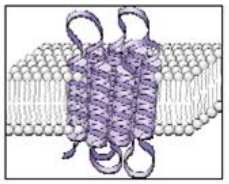
native 2D structure, monotonous hydrophobic pattern



IMP-coil (after Lupas et al.), Predator

Membrane regions

native 2D structure, hydrophobic bias

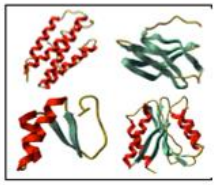


DAS-TMfilter
TMHMM
Phobius
HMMTOP
SAPS,
Toppred

- about 30 analytic methods with several parameter sets
- output (1-100 MB ASCII text)

Globular domains

native tertiary structure



PFAM
SMART
PROSITE
L. Aravind's library
Y. Wolf's library
M. Andrade's repeat library
IMP library

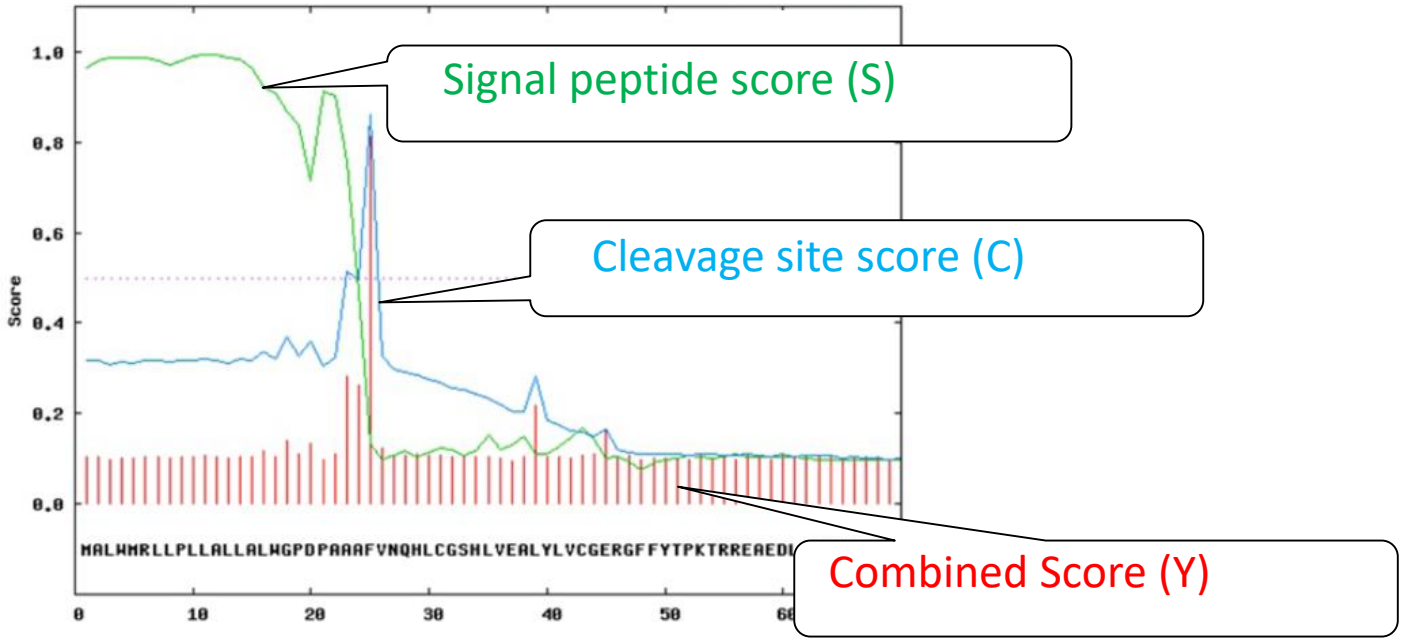
BLAST/PSI-BLAST
RPS-Blast
IMPALA

Annotator

Frank Eisenhaber et al.

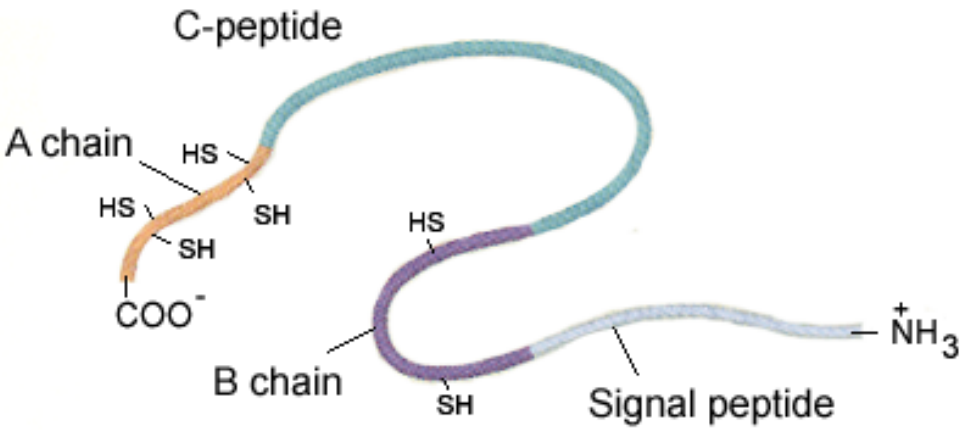
SignalP

- Neural network trained based on phylogeny
  - Gram-negative prokaryotic
  - Gram-positive prokaryotic
  - Eukaryotic
- Predicts secretory signal peptides
- http://www.cbs.dtu.dk/services/SignalP/



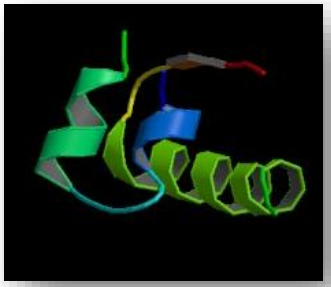
Insulin

1. Preproinsulin

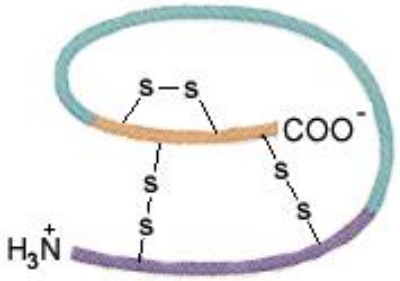


2. Membrane transport

3. Cleavege of signal peptide

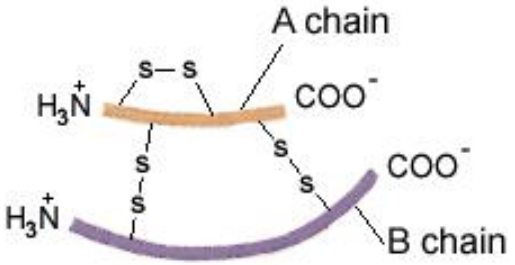


4. Disulfide bonds form

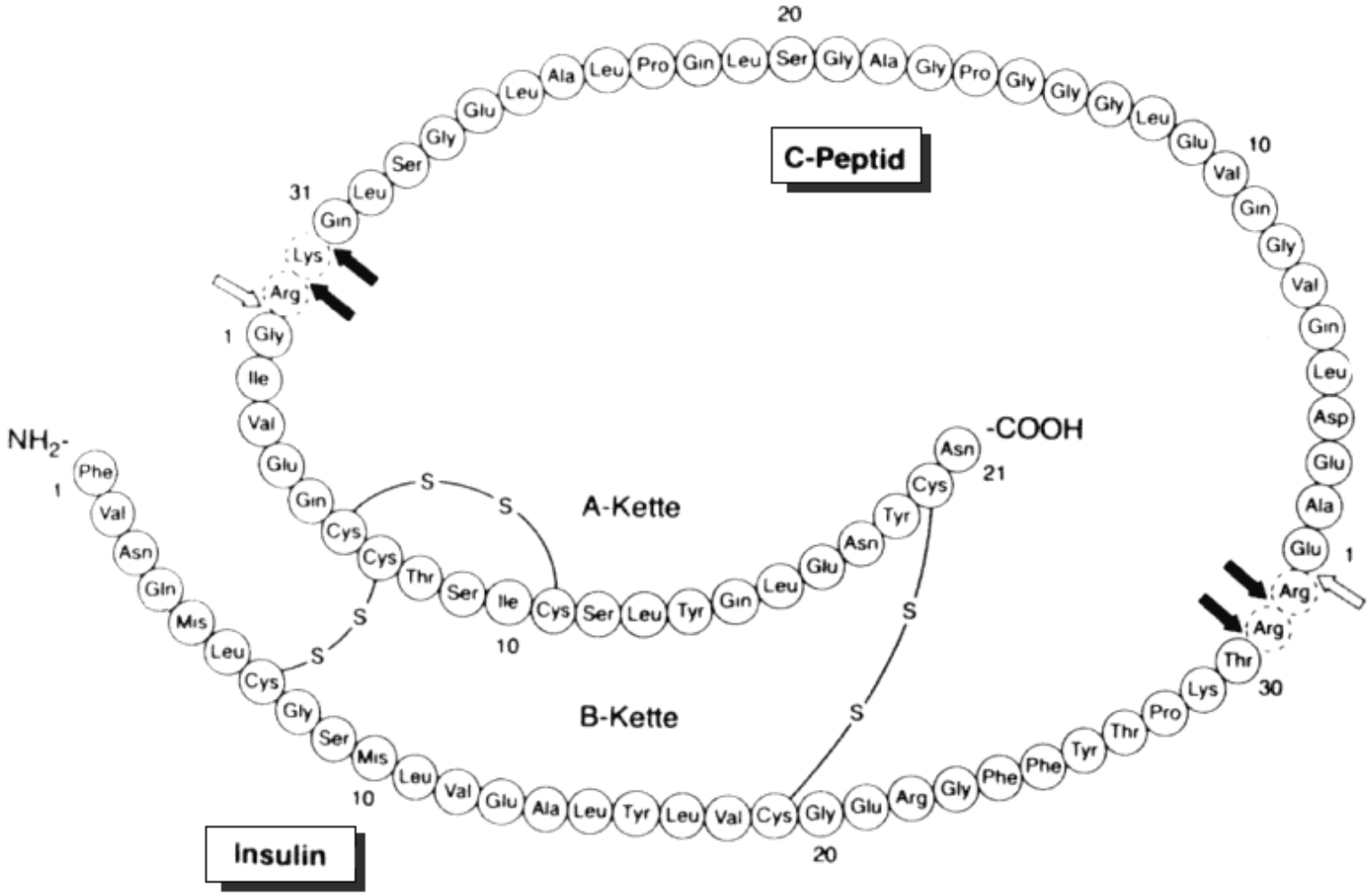


5. C-peptide is cleaved

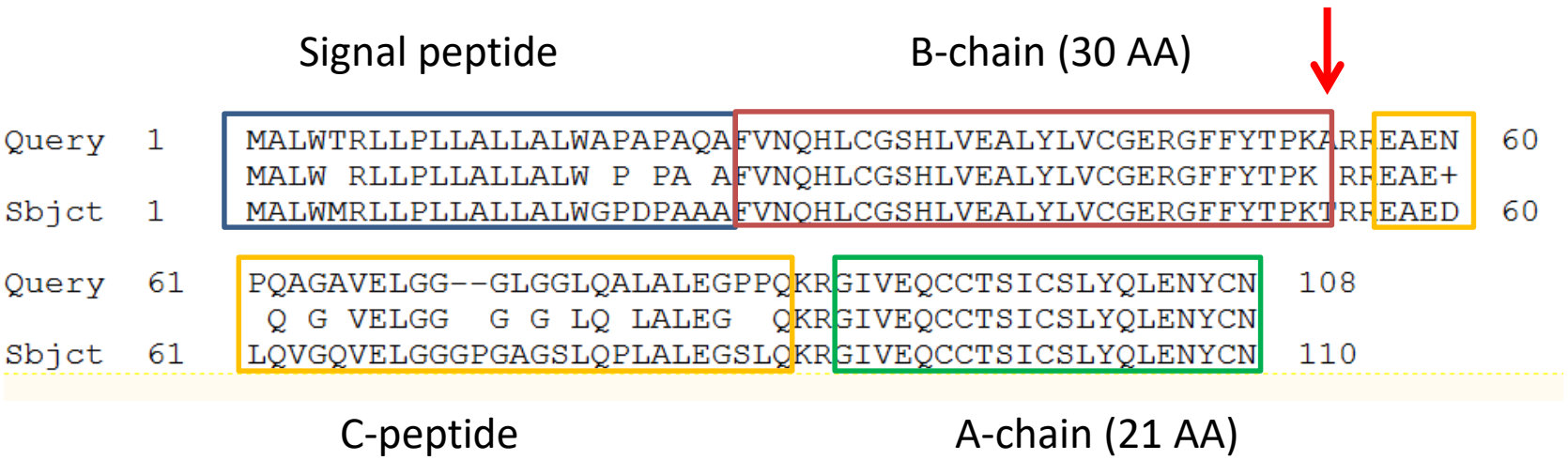
6. Formation of the mature insulin molecule



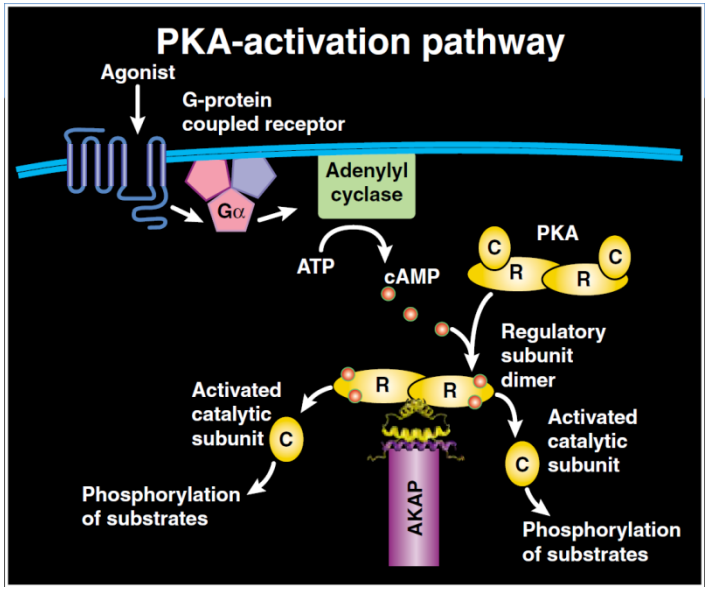
Proinsulin


C-Peptid
A-Kette
B-Kette
Insulin
NH2-
-COOH
Phe
Val
Asn
Gln
His
Leu
Cys
Gly
Ser
His
Leu
Val
Glu
Ala
Leu
Tyr
Leu
Val
Cys
Gly
Glu
Arg
Gly
Phe
Phe
Tyr
Thr
Pro
Lys
Thr
Arg
Arg
Glu
Ala
Glu
Asp
Leu
Gln
Val
Gly
Gln
Val
Glu
Leu
Gly
Gly
Pro
Gly
Ala
Gly
Ser
Leu
Gln
Pro
Leu
Ala
Leu
Glu
Gly
Ser
Leu
Gln
Lys
Arg
Gly
Ile
Val
Glu
Gln
Cys
Cys
Thr
Ser
Ile
Cys
Ser
Leu
Tyr
Gln
Leu
Glu
Asn
Tyr
Cys
Asn
S
S
S
S
S
S
1
10
20
30
31
1
10
20
21

Difference between pig and human insulin = 1AA

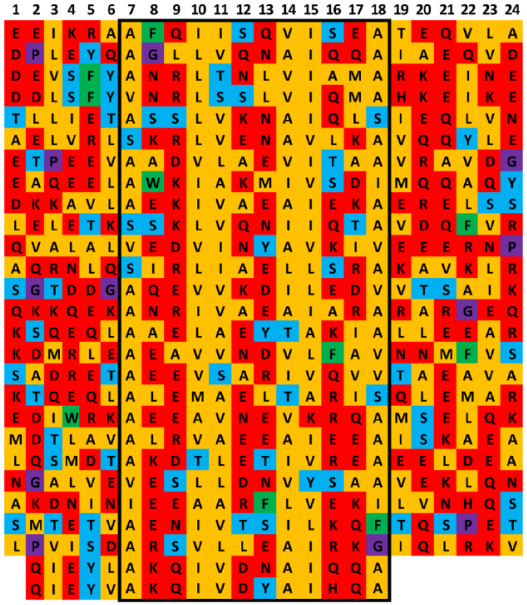


A-kinase anchoring proteins (AKAPs) binding to the regulatory subunit of protein kinase A (PKA)



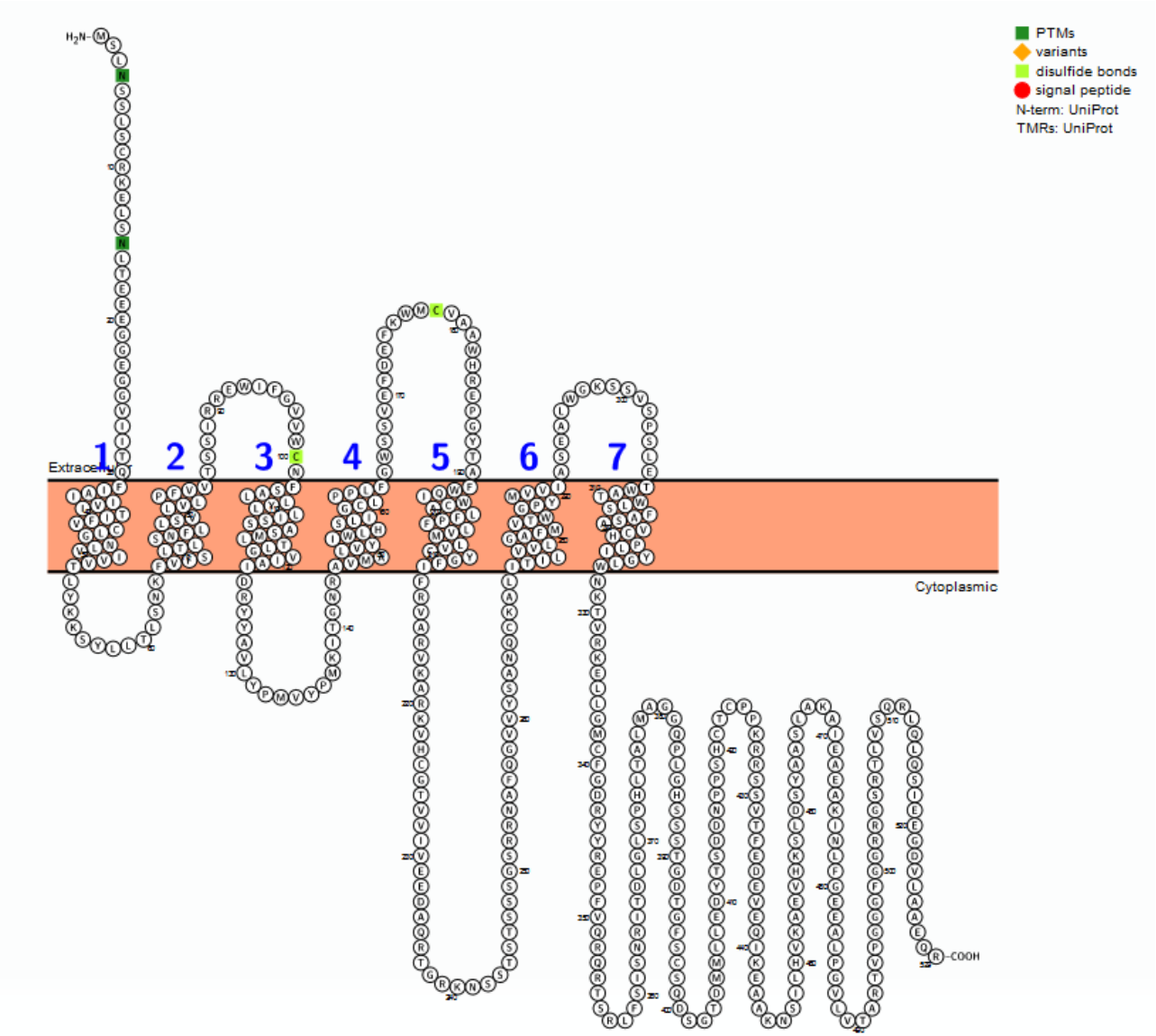
PKA-RII

1 AKAP1 (341-353) - Q92667
2 AKAP2/KL (563-586) - Q9Y2D5
3 AKAP3/AKAP110/FSP95 (121-144) - O75969
4 AKAP4/AKAP82/FSC1 (214-237) - Q5JQC9
5 AKAP5/AKAP79 (389-412) - P24588
6 AKAP7/AKAP15/AKAP18 (293-316) - Q9P0M2
7 AKAP8/AKAP95 (567-590) - O43823
8 AKAP10/dAKAP2 (628-651) - O43572
9 AKAP11/AKAP220 (1644-1667) - Q9UKA4
10 AKAP12/AKAP250/GRAVIN (1539-1562) - Q02952
11 AKAP14/AKAP28 (37-60) - Q86UN6
12 AKAP17A (366-389) - Q02040
13 BIG2 (278-301) - Q9Y6D5
14 Chromodomain-helicase-DNA-Binding (454-477) - Q9HCK8
15 EZRIN (412-435) - P15311
16 GSKIP (30-53) - Q9P0R6
17 MAP2 (88-111) - P11137
18 Moesin (412-435) - P26038
19 MTG16B (398-421) - O75081
20 MyRIP (187-210) - Q8NFW9
21 Myosprin (3629-3652) - Q8N3K9
22 Neurobeachin (1092-1103) - Q8NFP9
23 RAB32 (181-192) - Q13637
24 Synemin (638-649) - O15061
25 WAVE1 (497-508) - Q92558
26 AKAP-IS
27 SUPER-AKAP-IS

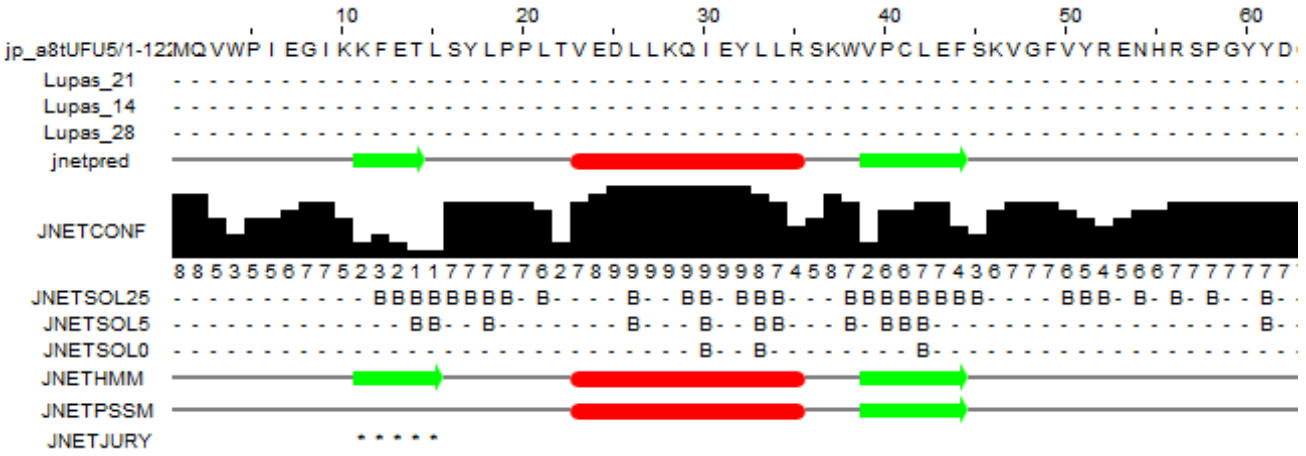


Protter

GPR161



Protein secondary structure prediction (Jpred)



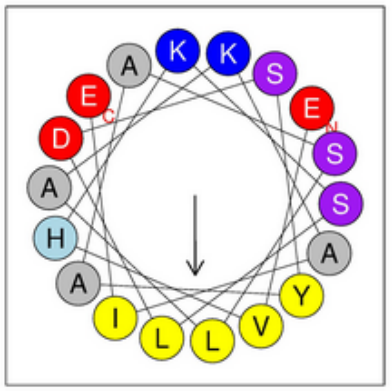
Amphipatic helix

Heliquest

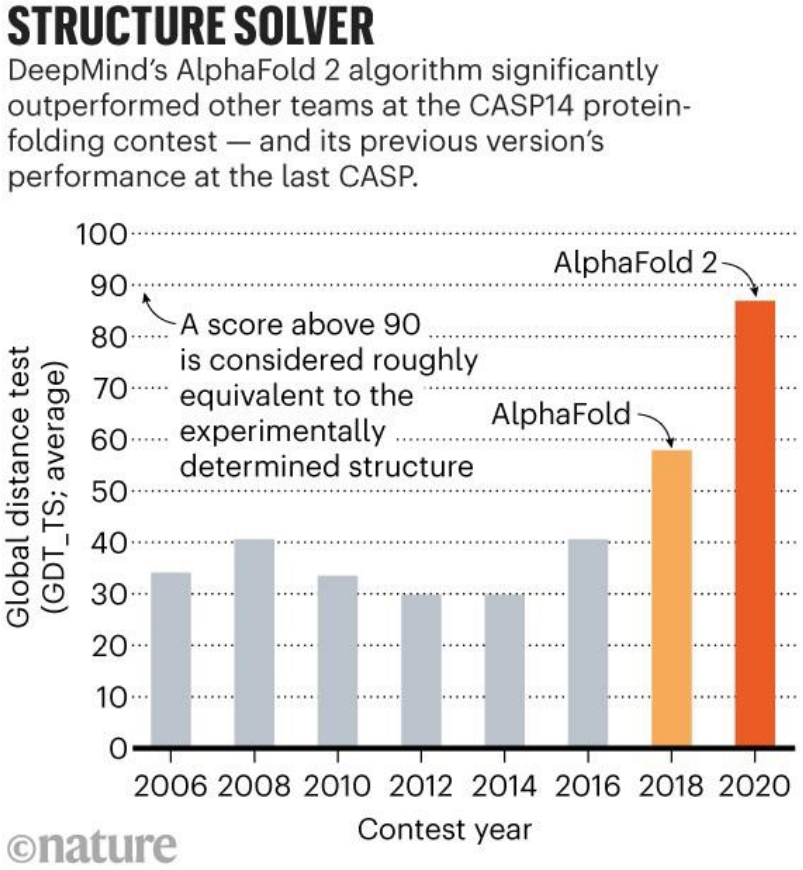
$_{1}$EVHKSLDSYAASLAKAIE$_{18}$

Physico-chemical properties	Polar residues + GLY	Nonpolar residues
Hydrophobicity <H>	Polar residues + GLY (n / %)	Nonpolar residues (n / %)
0.256	9 / 50.00	9 / 50.00
Hydrophobic moment <μH>	Uncharged residues + GLY	Aromatic residues
0.542	HIS 1, SER 3, GLY 0	TYR 1,
Net charge z	Charged residues	Special residues
-1	LYS 2, GLU 2, ASP 1,	CYS 0, PRO 0
Hydrophobic face : A Y V L L I A		
Go to screening	Manual mutation	GA mutation

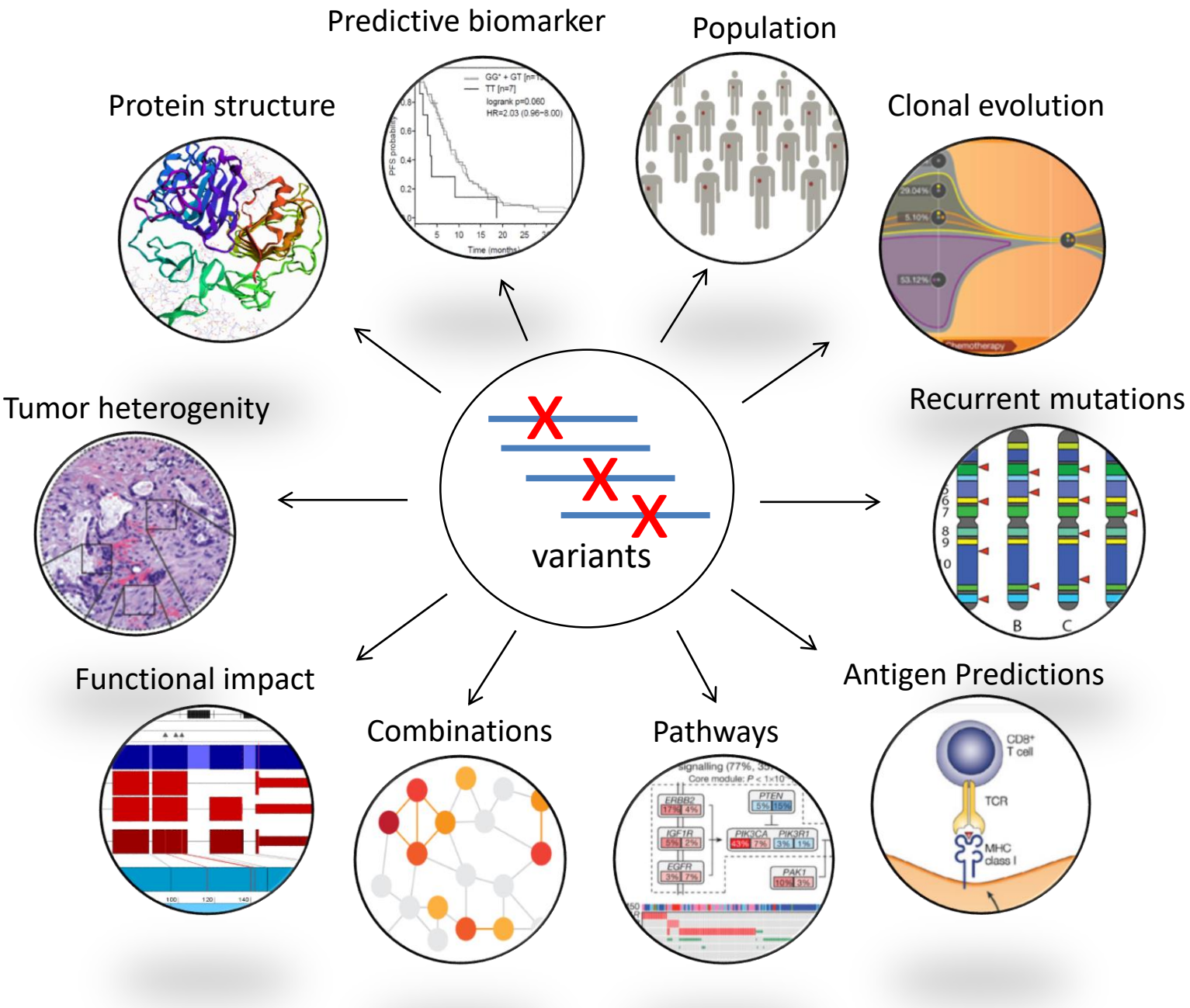
Click to enlarge



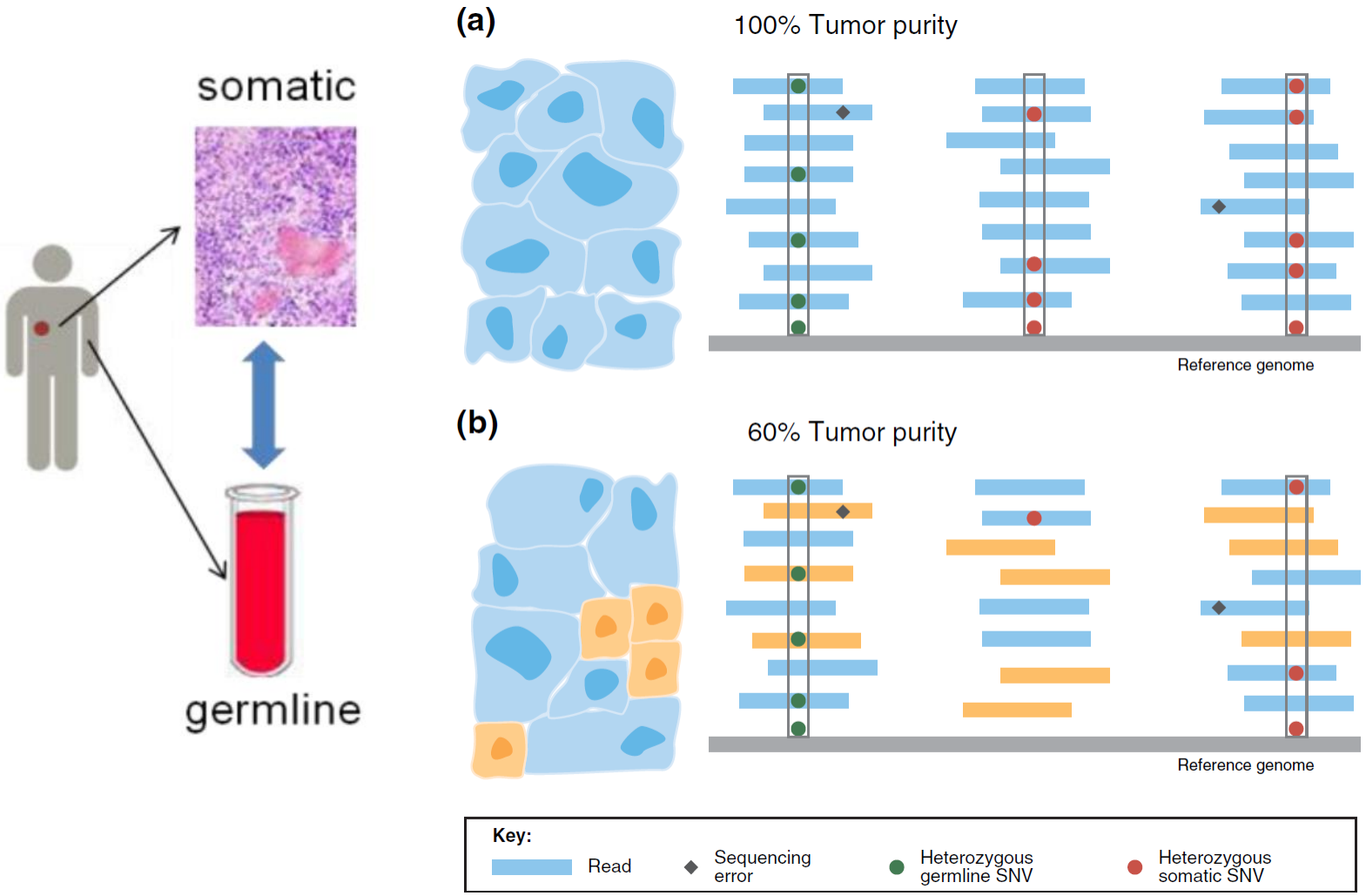
Predicting the 3D structures of proteins from their amino-acid sequences



Analyses and interpretation of DNA variants

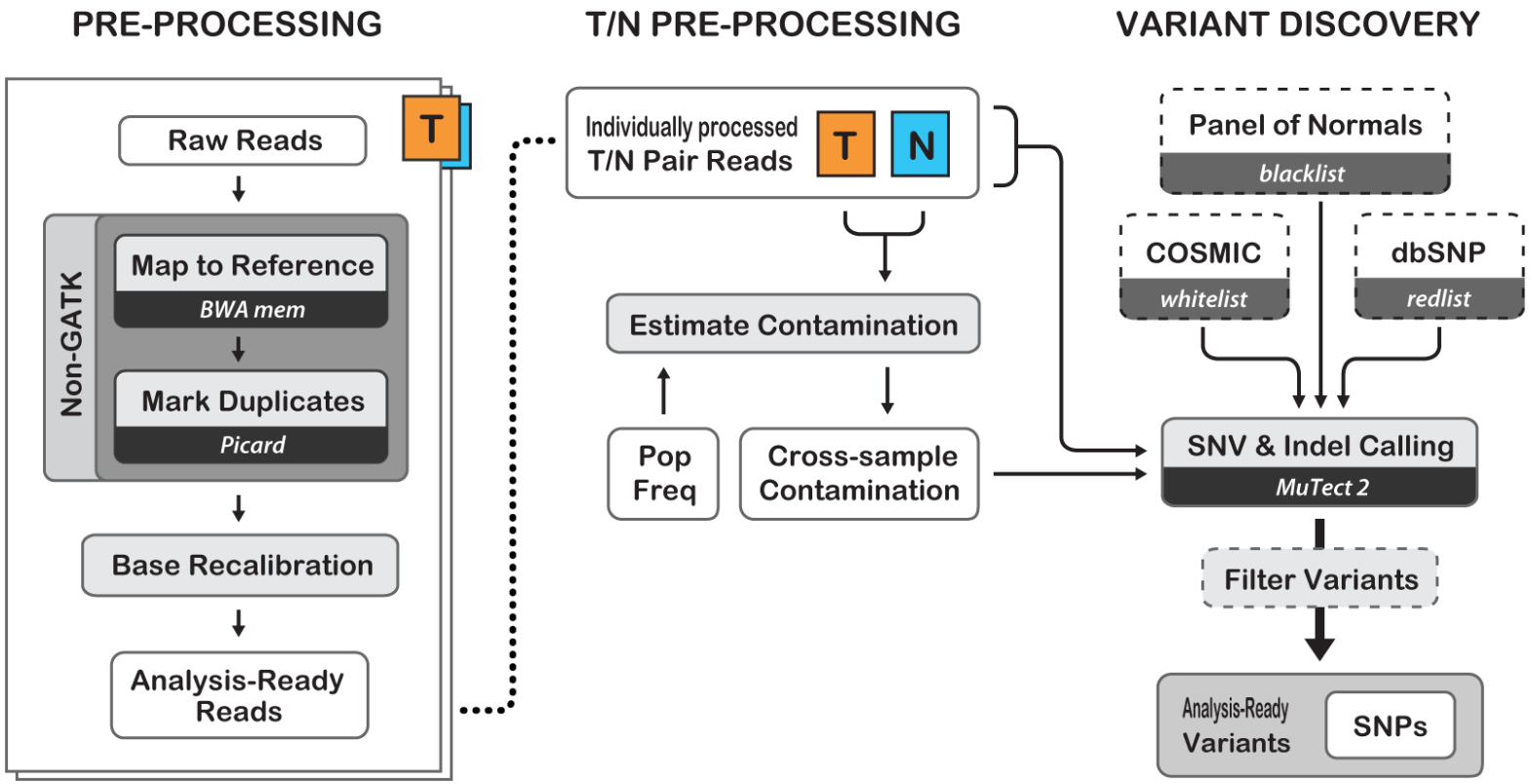


Somatic mutation detection in tumor samples



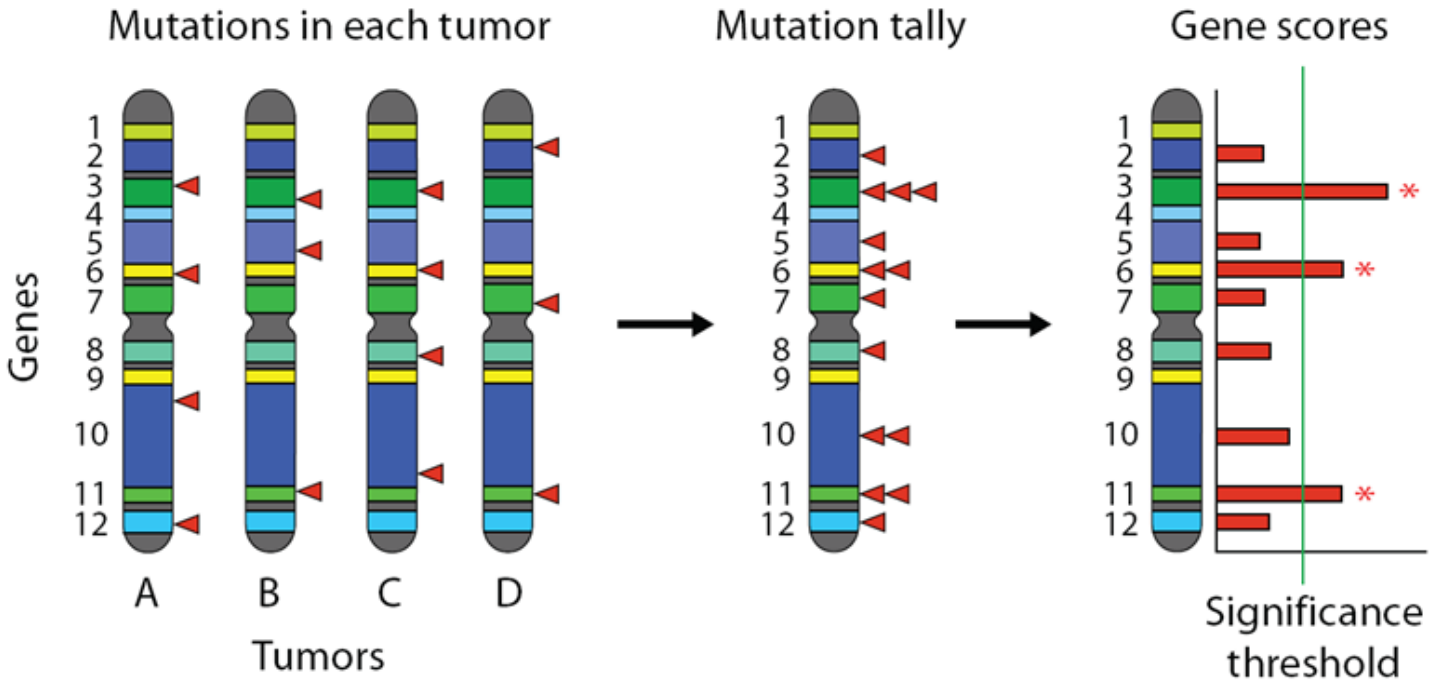
Raphael et al. Genome Med. 2014

MuTect 2 (GATK)



Best Practices for Somatic SNVs and Indels in Whole Genomes and Exomes - BETA

MutSig (MutSigCV)



MutSig builds a model of the background mutation processes (BMR) that were at work during formation of the tumors, and it analyzes the mutations of each gene to identify genes that were mutated more often than expected by chance, given the background model.

MutSigCV (CV for 'covariate') improves the BMR estimation by pooling data from 'neighbor' genes with similar genomic properties such as DNA replication time, chromatin state (open/closed), and general level of transcription activity.

Lawrence, M. et al. Nature 2013

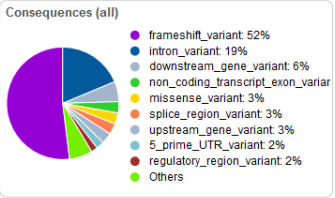
Ensembl Variant Effect Predictor (VEP)

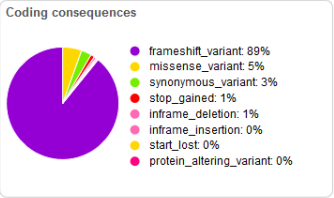
Variant Effect Predictor results

Job details

Summary statistics

Category	Count
Variants processed	544
Variants filtered out	0
Novel / existing variants	475 (87.3) / 69 (12.7)
Overlapped genes	61
Overlapped transcripts	317
Overlapped regulatory features	20





Results preview

Navigation — Page: 1 of 109 | Show: 1 5 10 50 All variants

Filters — Uploaded variant | is | defined | Add

Download — All: VCF VEP TXT; BioMart: Variants Genes

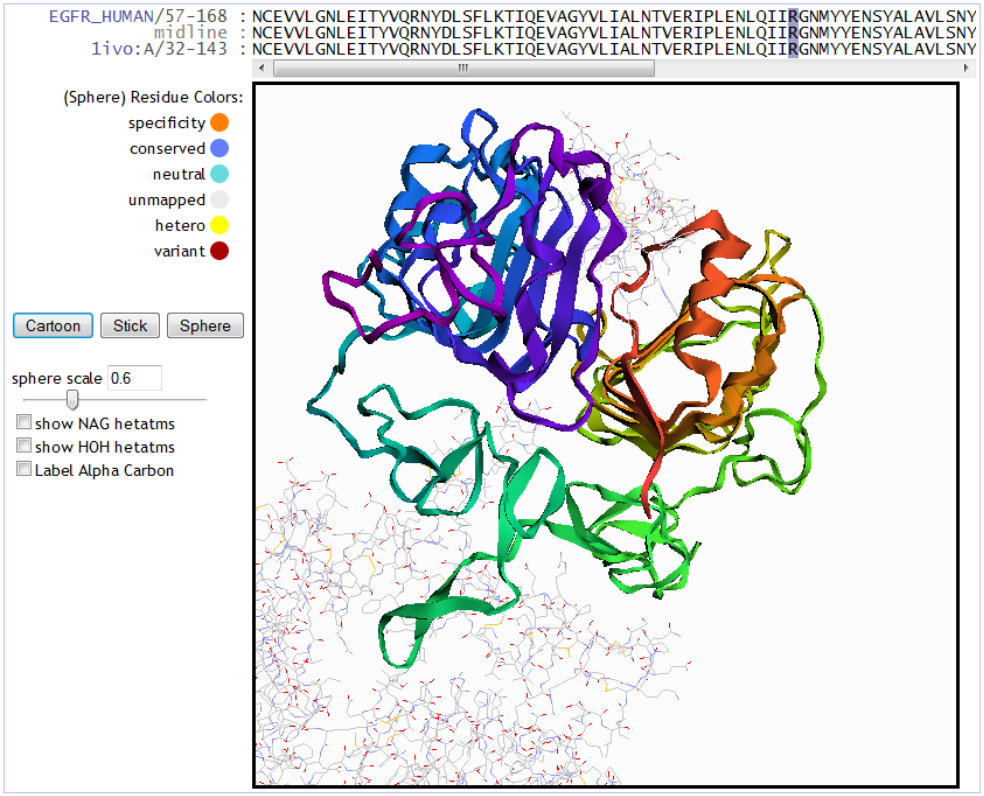
Show/hide columns

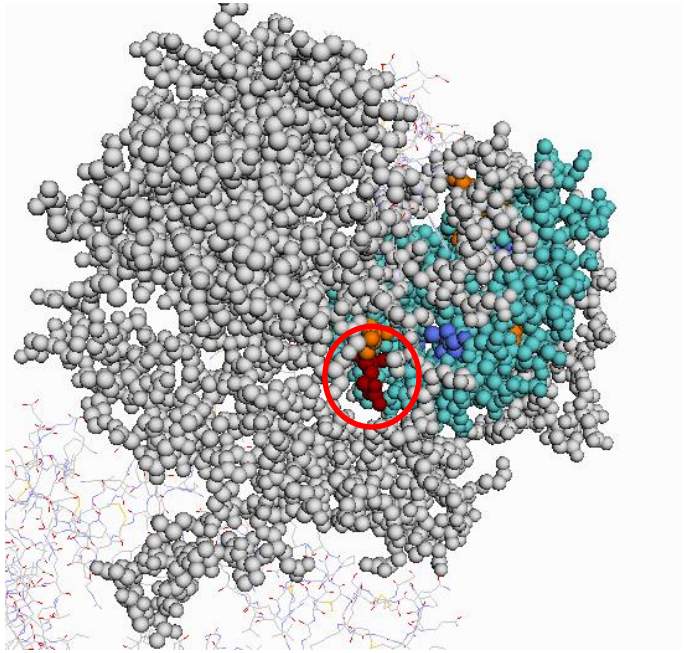
Uploaded variant	Location	Allele	Consequence	Impact	Symbol	Gene	Feature type	Feature	Biotype	Exon	Intron	HGVSc	HGVSp	cDNA position	CDS position	Protein position	Amino acids	Codons	Existing variant
.	1:43815005-43815005	T	frameshift_variant	HIGH	MPL	4352	Transcript	NM_005373.2	protein_coding	10/12	-	-	-	1585-1586	1540-1541	514	R/MX	AGG/ATGG	-
.	1:43815005-43815005	T	frameshift_variant	HIGH	MPL	4352	Transcript	XM_005270874.1	protein_coding	10/12	-	-	-	1519-1520	1519-1520	507	R/MX	AGG/ATGG	-
.	1:43815005-43815005	T	regulatory_region_variant	MODIFIER	-	-	RegulatoryFeature	ENSR00000005582	promoter_flanking_region	-	-	-	-	-	-	-	-	-	-
.	1:43815008-43815009	-	frameshift_variant	HIGH	MPL	4352	Transcript	NM_005373.2	protein_coding	10/12	-	-	-	1589	1544	515	W/X	TGG/TG	COSM142839, COSM18918, COSM3719407
.	1:43815008-43815009	-	frameshift_variant	HIGH	MPL	4352	Transcript	XM_005270874.1	protein_coding	10/12	-	-	-	1523	1523	508	W/X	TGG/TG	COSM142839, COSM18918, COSM3719407
.	1:43815008-43815009	-	regulatory_region_variant	MODIFIER	-	-	RegulatoryFeature	ENSR00000005582	promoter_flanking_region	-	-	-	-	-	-	-	-	-	COSM142839, COSM18918, COSM3719407
.	1:43815013-43815013	T	frameshift_variant	HIGH	MPL	4352	Transcript	NM_005373.2	protein_coding	10/12	-	-	-	1593-1594	1548-1549	516-517	-/X	-/T	-
.	1:43815013-43815013	T	frameshift_variant	HIGH	MPL	4352	Transcript	XM_005270874.1	protein_coding	10/12	-	-	-	1527-1528	1527-1528	509-510	-/X	-/T	-
.	1:43815013-43815013	T	regulatory_region_variant	MODIFIER	-	-	RegulatoryFeature	ENSR00000005582	promoter_flanking_region	-	-	-	-	-	-	-	-	-	-
.	1:43815036-43815036	CC	splice_region_variant, intron_variant	LOW	MPL	4352	Transcript	NM_005373.2	protein_coding	-	10/11	-	-	-	-	-	-	-	-
.	1:43815036-43815036	CC	splice_region_variant, intron_variant	LOW	MPL	4352	Transcript	XM_005270874.1	protein_coding	-	10/11	-	-	-	-	-	-	-	-

MutationAssessor

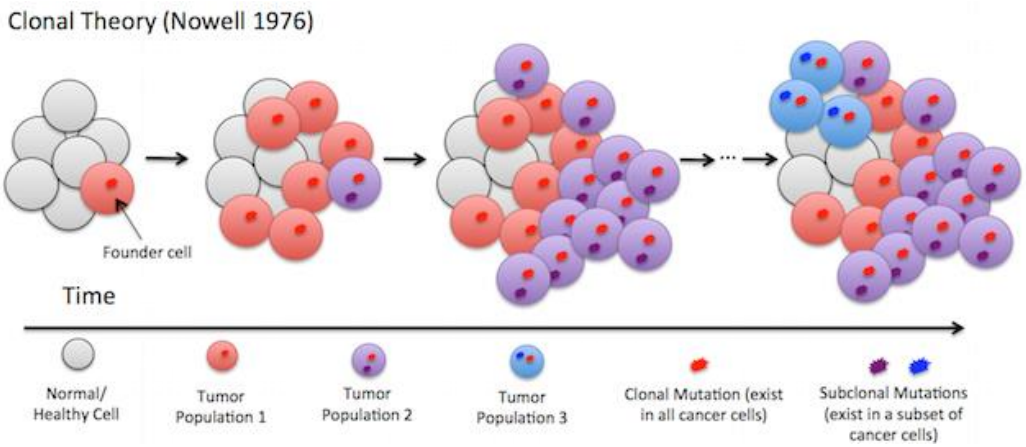
Chr Pos RefAll AltAll

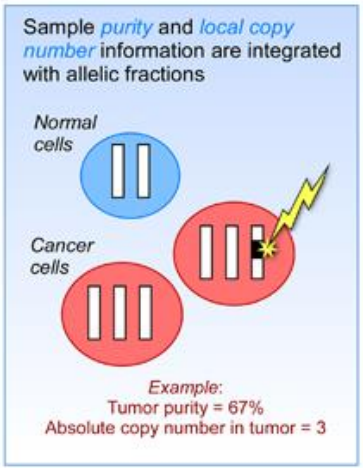
Mutation	AA variant	Gene	MSA	PDB	Func. Impact	FI score	Uniprot	Refseq	MSA height
2,212288939,C,G	G936A	ERBB4	msa	pdb	medium	2.4	ERBB4_HUMAN	XP_005246432	1684

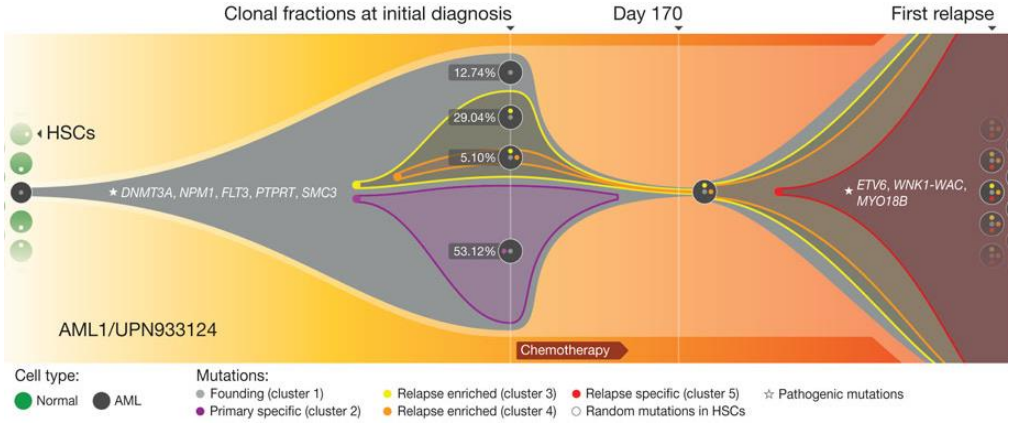




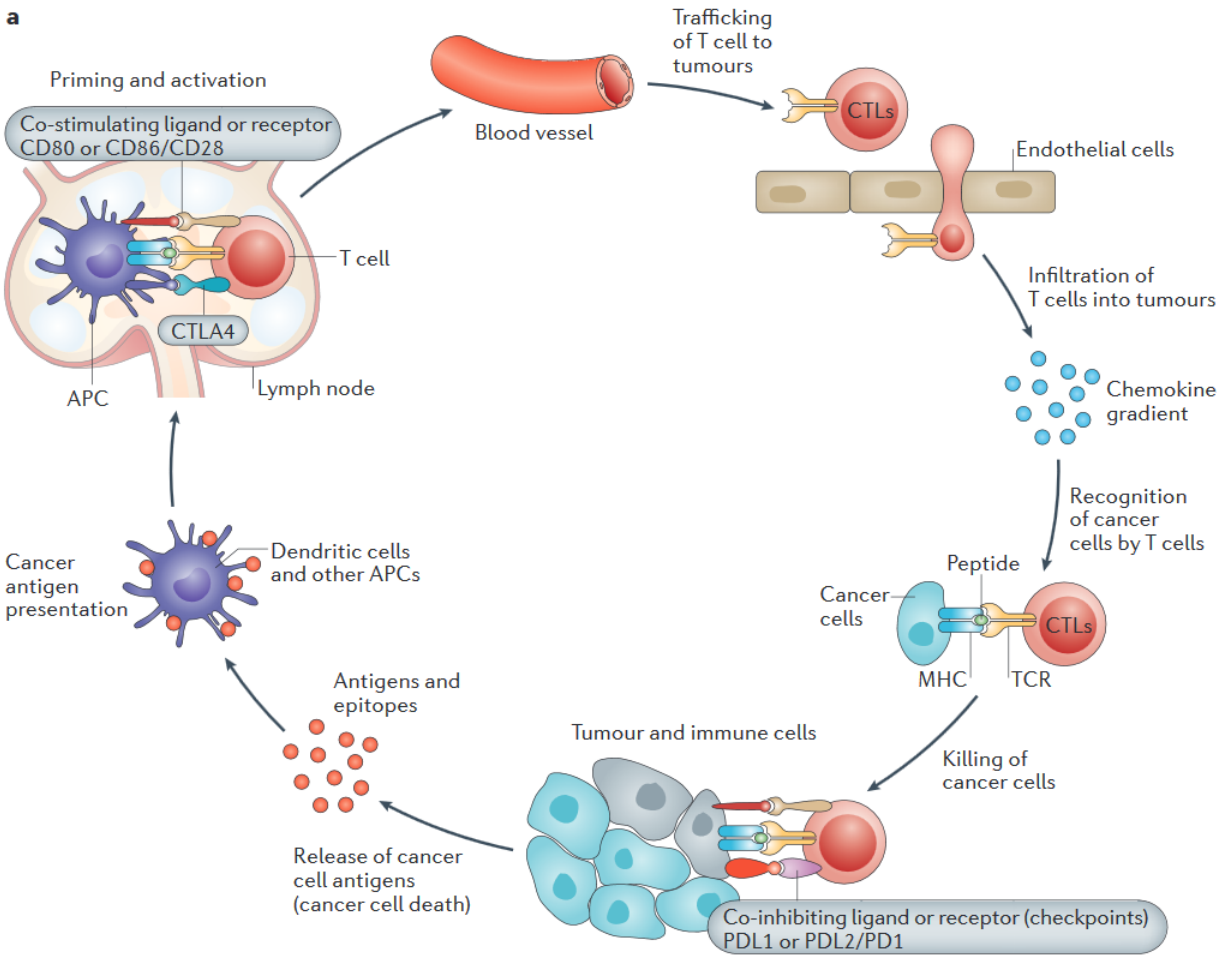
Intratumor heterogenity and clonal evolution





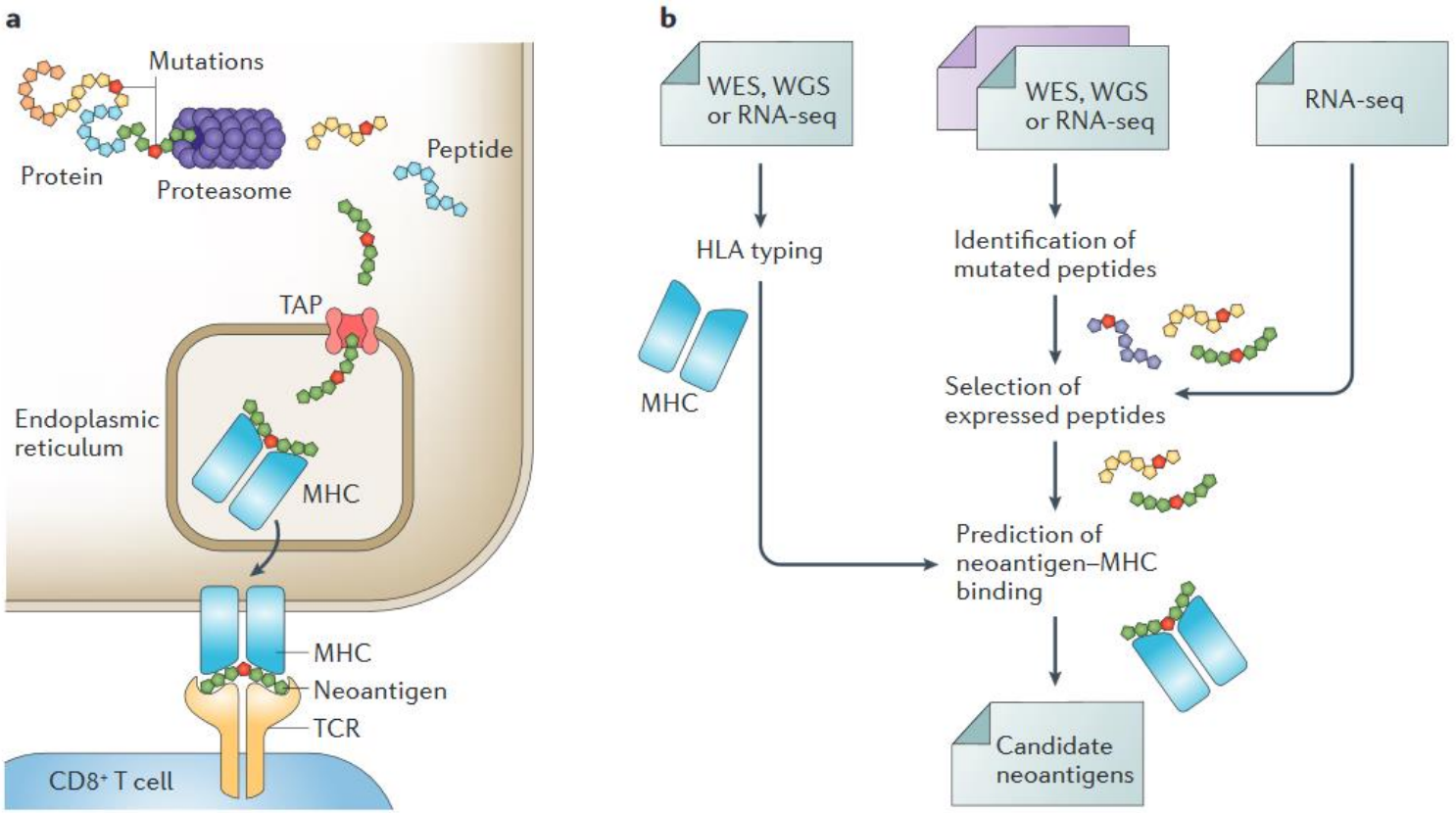


Cancer-Immunity cycle



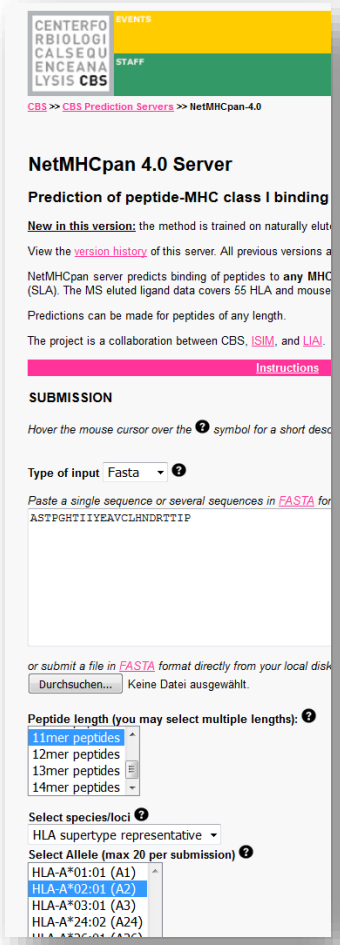
Hackl et al. Nat Rev Genet 17: 441-458 (2016)

Neoantigen prediction



Hackl et al. Nat Rev Genet 17: 441-458 (2016)

NetMHCpan



Fasta input:

>Gag_180_209
TPQDLNTMLNTVGGHQAAMQMLKETINEEA

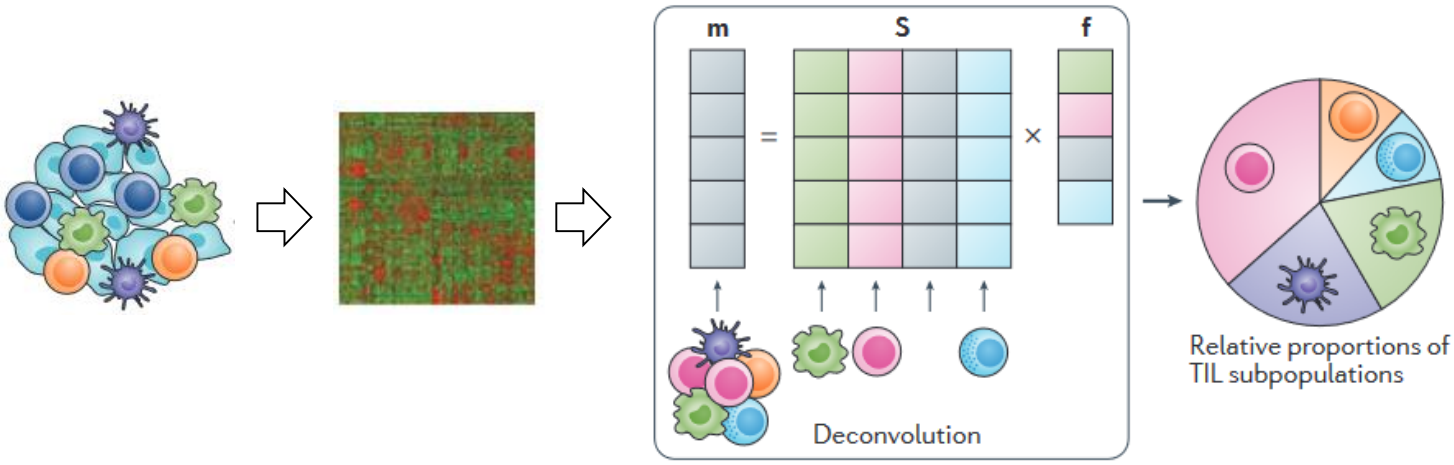
Peptide length: **8, 9, 10, 11, 12**
Allele: **HLA-A*0301**
Toggle **Sort by prediction score**

will return the following predictions:

```
# NetMHCpan version 4.0

# Tmpdir made /usr/opt/www/webface/tmp/server/netmhcpan/59DBCCFF00005A84DAFF1311/netMHCpanVszuD8
# Input is in FSA format

# Peptide length 8,9,10,11,12

# Make Eluted ligand likelihood predictions

HLA-A03:01 : Distance to training data  0.000 (using nearest neighbor HLA-A03:01)

# Rank Threshold for Strong binding peptides   0.500
# Rank Threshold for Weak binding peptides   2.000
-----------------------------------------------------------------------------------
  Pos          HLA         Peptide       Core Of Gp Gl Ip Il        Icore        Identity     Score   %Rank  BindLevel
-----------------------------------------------------------------------------------
   15  HLA-A*03:01       HQAAMQMLK  HQAAMQMLK  0  0  0  0  0    HQAAMQMLK     Gag_180_209 0.5697290  0.2857 <= SB
   14  HLA-A*03:01      GHQAAMQMLK  GQAAMQMLK  0  1  1  0  0   GHQAAMQMLK     Gag_180_209 0.2137130  1.1582 <= WB
    7  HLA-A*03:01       TMLNTVGGH  TMLNTVGGH  0  0  0  0  0    TMLNTVGGH     Gag_180_209 0.0487720  3.0466
    8  HLA-A*03:01       MLNTVGGHQ  MLNTVGGHQ  0  0  0  0  0    MLNTVGGHQ     Gag_180_209 0.0319510  3.7842
   13  HLA-A*03:01     GGHQAAMQMLK  GQAAMQMLK  0  1  2  0  0  GGHQAAMQMLK     Gag_180_209 0.0313010  3.8215
   12  HLA-A*03:01    VGGHQAAMQMLK  VQAAMQMLK  0  1  3  0  0 VGGHQAAMQMLK     Gag_180_209 0.0166440  5.2079
   15  HLA-A*03:01      HQAAMQMLKE  HQAAMQMLK  0  0  0  0  0    HQAAMQMLK     Gag_180_209 0.0124970  5.9719
   16  HLA-A*03:01        QAAMQMLK  QAA-MQMLK  0  0  0  3  1     QAAMQMLK     Gag_180_209 0.0086270  7.1279
   21  HLA-A*03:01       MLKETINEE  MLKETINEE  0  0  0  0  0    MLKETINEE     Gag_180_209 0.0079270  7.4157
--
--
-----------------------------------------------------------------------------------

Protein Gag_180_209. Allele HLA-A*03:01. Number of high binders 1. Number of weak binders 1. Number of peptides 105

Link to Allele Frequencies in Worldwide Populations HLA-A03:01
-----------------------------------------------------------------------------------
```

http://www.cbs.dtu.dk/services/NetMHCpan/

Deconvolution analyses



R package: *immunedeconv*

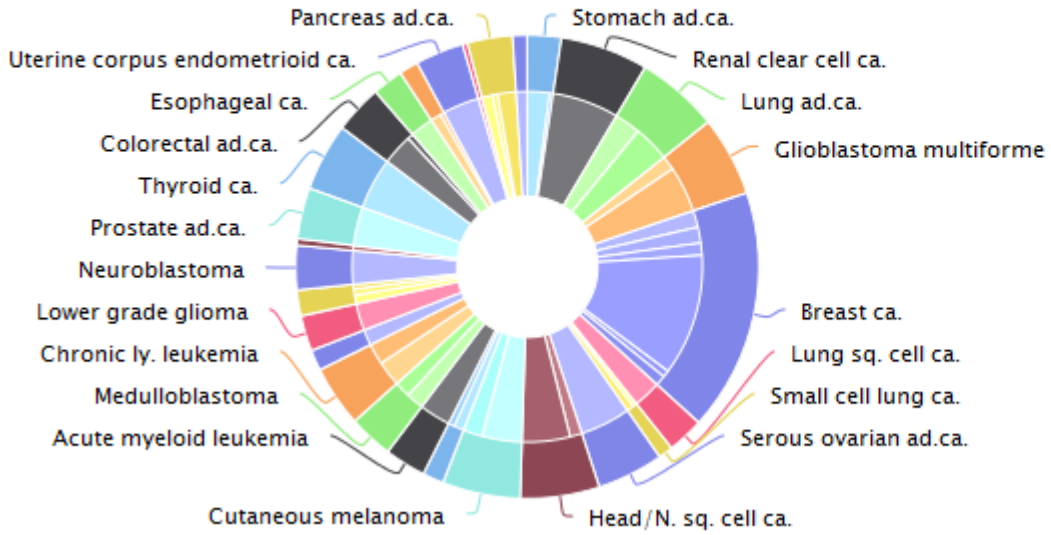
- quanTIseq
- EPIC

Hackl H *et al.* Nat Rev Genet 2017, Finotello F *et al.* Genome Med 2019, Racle J *et al*. Elife 2019, Sturm G *et al.* Bioinformatics 2019

IntOGen

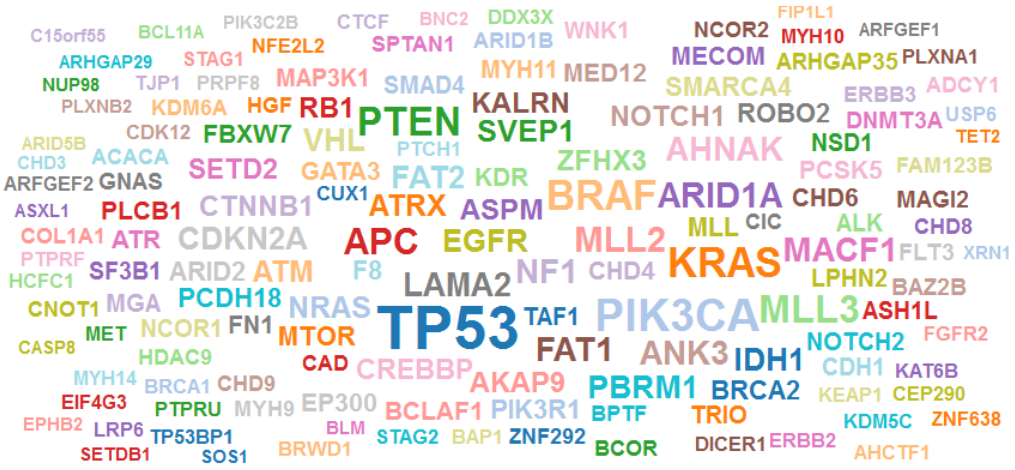
IntOGen Mutations 2014.12

Cancer types and projects chart



Cancer Types	28
Projects	48
Samples	6792
Somatic mutations	1341752
Coding sequence mutations (CSMs)	
in driver genes	21648
in all genes	1341706
Protein affecting mutations (PAMs)	
in driver genes	18649
in all genes	603770

Driver genes



Cancers arise due to alterations in genes that confer growth advantage to the cell . More than 400 such ‘cancer genes’, identified to date are currently annotated in the Cancer Gene Census.

Cancer driver mutations (cancer drivers) versus passenger mutations can be identified based on:

- Functional Impact
- Recurrence

IntOgen (Integrative Onco Genomics)

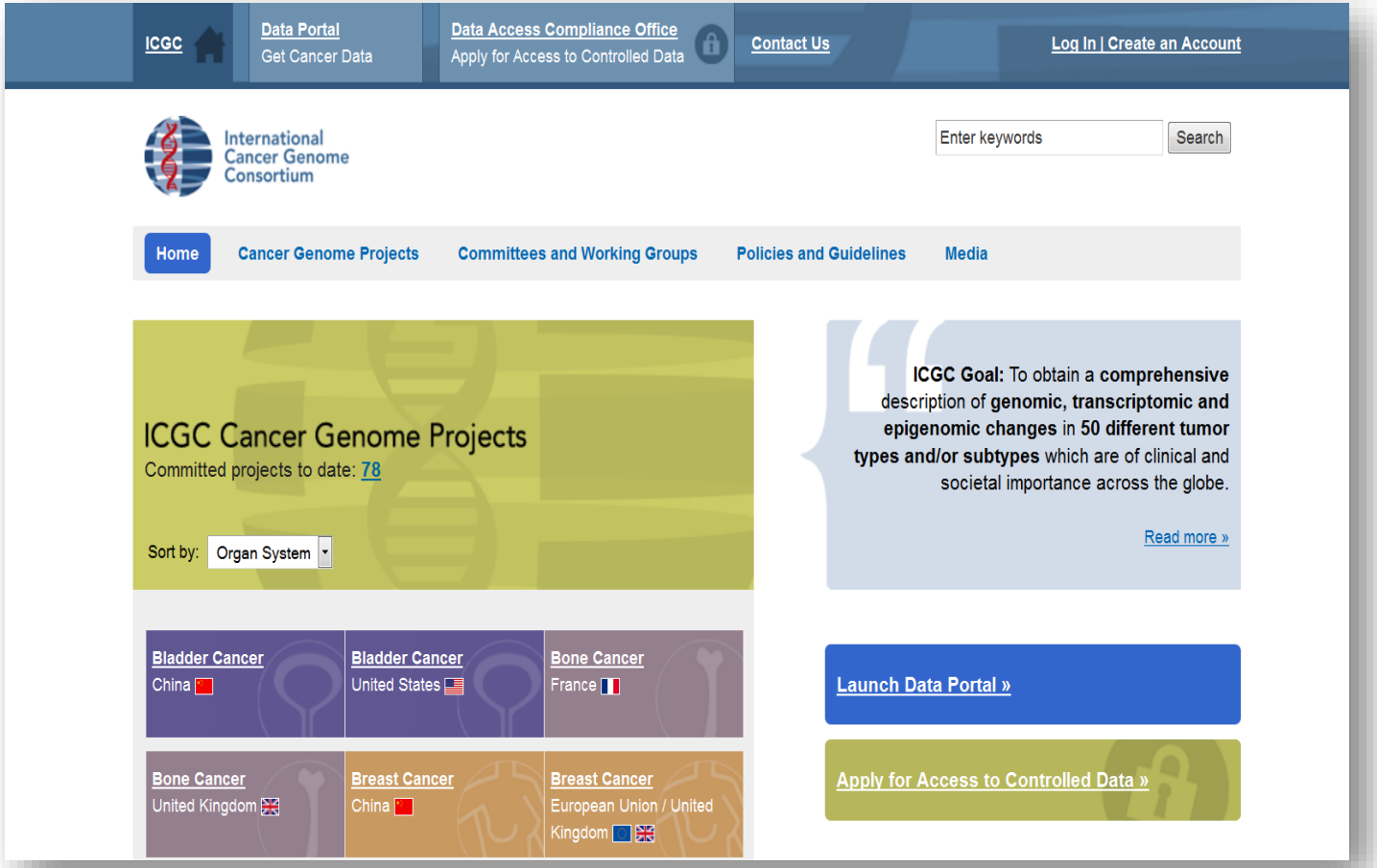
- Driver signals
  - Clustered mutations (OncodriveCLUST)
  - Functional Mutations (OncodriveFM)
  - Recurrent Mutations (MutSigCV)
- Mutation frequency per cancertype
  - No. of mutated samples in cancer type
  - No. of protein affecting mutated samples in cancer type (PAM)
- Mutation distribution along protein sequence
  - Protein domains
  - No. and position of mutationof protein affecting mutated samples in cancer type (PAM)
  - Different transcripts

Examples

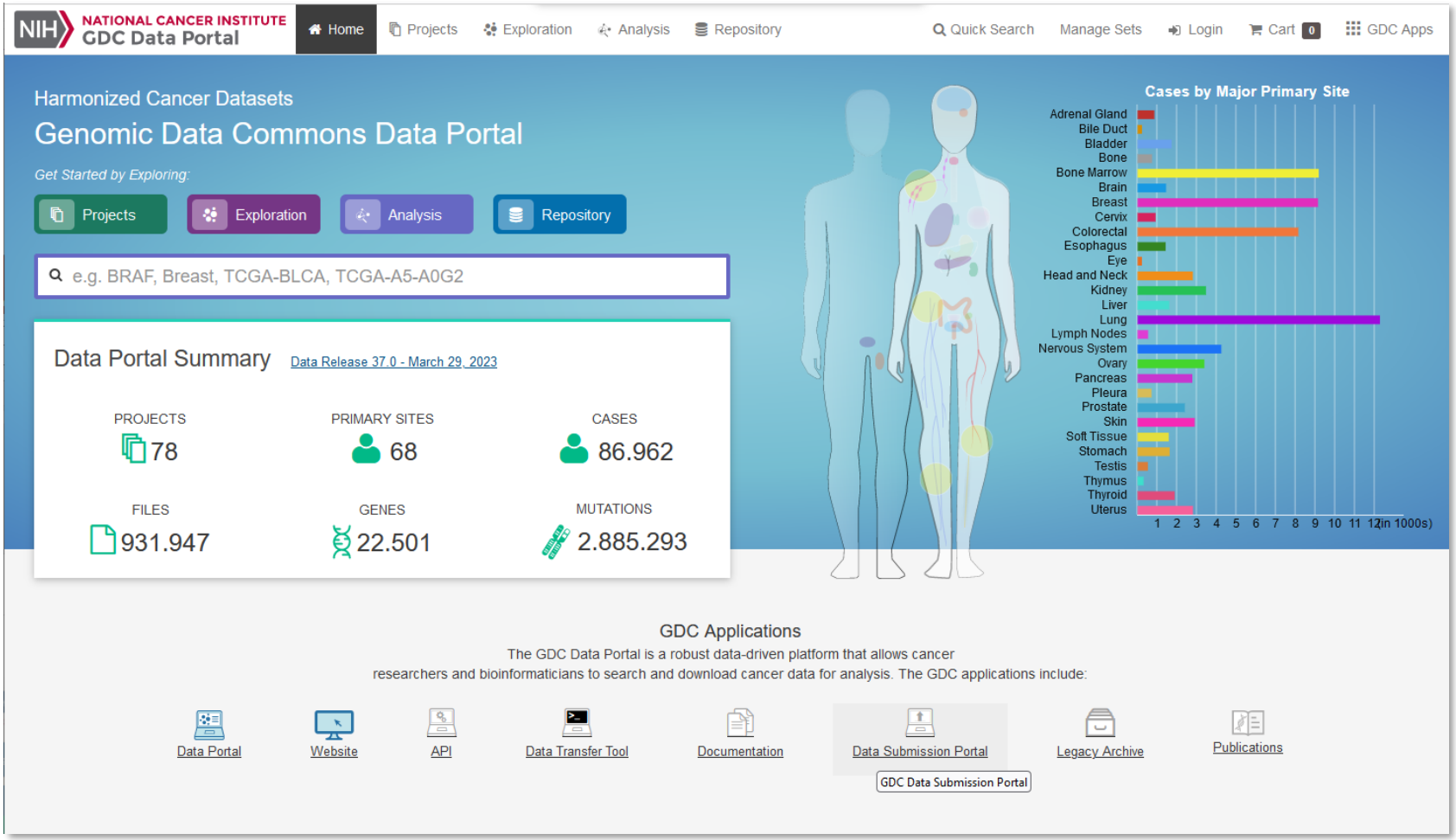
- What is the most common BRAF mutation
- In which cancer types IDH1 is a cancer driver and in which cancer type mutation of IDH1 is most frequent
- Most common drivers in breast carcinoma
- Mutation frequency of VHL

TCGA

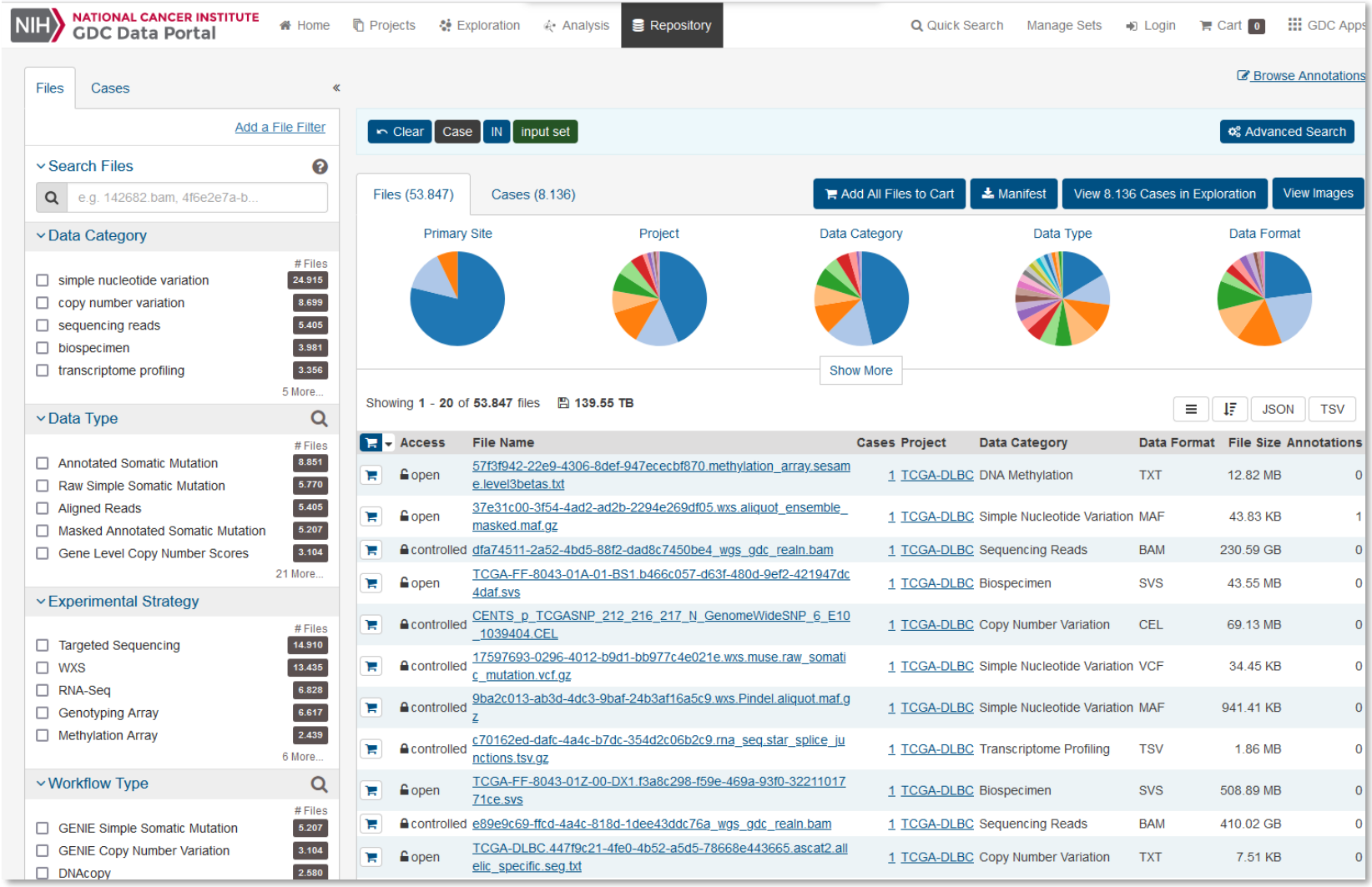
International Cancer Genome Consortium (ICGC)



TCGA

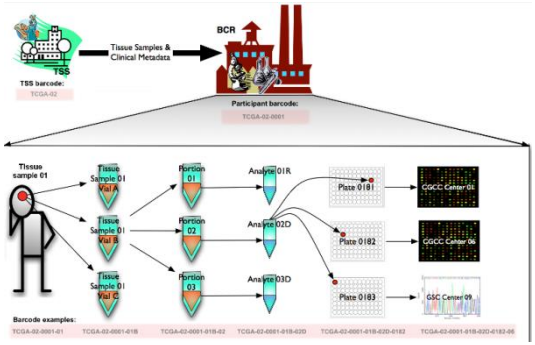

NIH NATIONAL CANCER INSTITUTE GDC Data Portal
Home
Projects
Exploration
Analysis
Repository
Quick Search
Manage Sets
Login
Cart 0
GDC Apps
Harmonized Cancer Datasets
Genomic Data Commons Data Portal
Get Started by Exploring:
Projects
Exploration
Analysis
Repository
e.g. BRAF, Breast, TCGA-BLCA, TCGA-A5-A0G2
Data Portal Summary
Data Release 37.0 - March 29, 2023
PROJECTS 78
PRIMARY SITES 68
CASES 86.962
FILES 931.947
GENES 22.501
MUTATIONS 2.885.293
Cases by Major Primary Site
Adrenal Gland
Bile Duct
Bladder
Bone
Bone Marrow
Brain
Breast
Cervix
Colorectal
Esophagus
Eye
Head and Neck
Kidney
Liver
Lung
Lymph Nodes
Nervous System
Ovary
Pancreas
Pleura
Prostate
Skin
Soft Tissue
Stomach
Testis
Thymus
Thyroid
Uterus
1 2 3 4 5 6 7 8 9 10 11 12 (in 1000s)
GDC Applications
The GDC Data Portal is a robust data-driven platform that allows cancer researchers and bioinformaticians to search and download cancer data for analysis. The GDC applications include:
Data Portal
Website
API
Data Transfer Tool
Documentation
Data Submission Portal
GDC Data Submission Portal
Legacy Archive
Publications

Genomic Data Commons (TCGA)

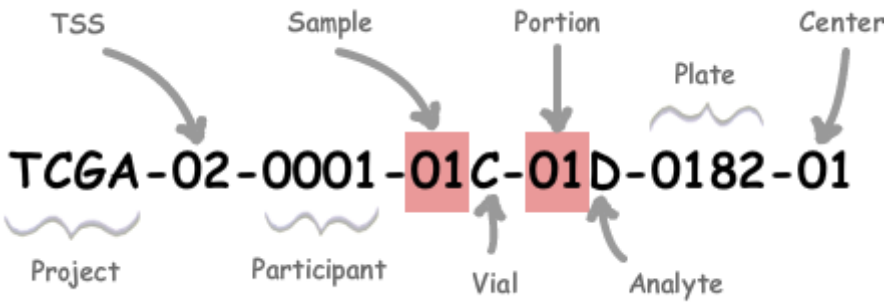


TCGA barcodes

Tissue source site



Biospecimen core resource



Sample

01	Primary solid Tumor
02	Recurrent Solid Tumor
03	Primary Blood Derived Cancer - Peripheral Blood
04	Recurrent Blood Derived Cancer - Bone Marrow
05	Additional - New Primary
06	Metastatic
07	Additional Metastatic
08	Human Tumor Original Cells
09	Primary Blood Derived Cancer - Bone Marrow
10	Blood Derived Normal
11	Solid Tissue Normal
12	Buccal Cell Normal
13	EBV Immortalized Normal
14	Bone Marrow Normal
20	Control Analyte
40	Recurrent Blood Derived Cancer - Peripheral Blood
50	Cell Lines

Universally Unique Identifiers (UUIDs) have replaced TCGA barcodes

TCGA data levels

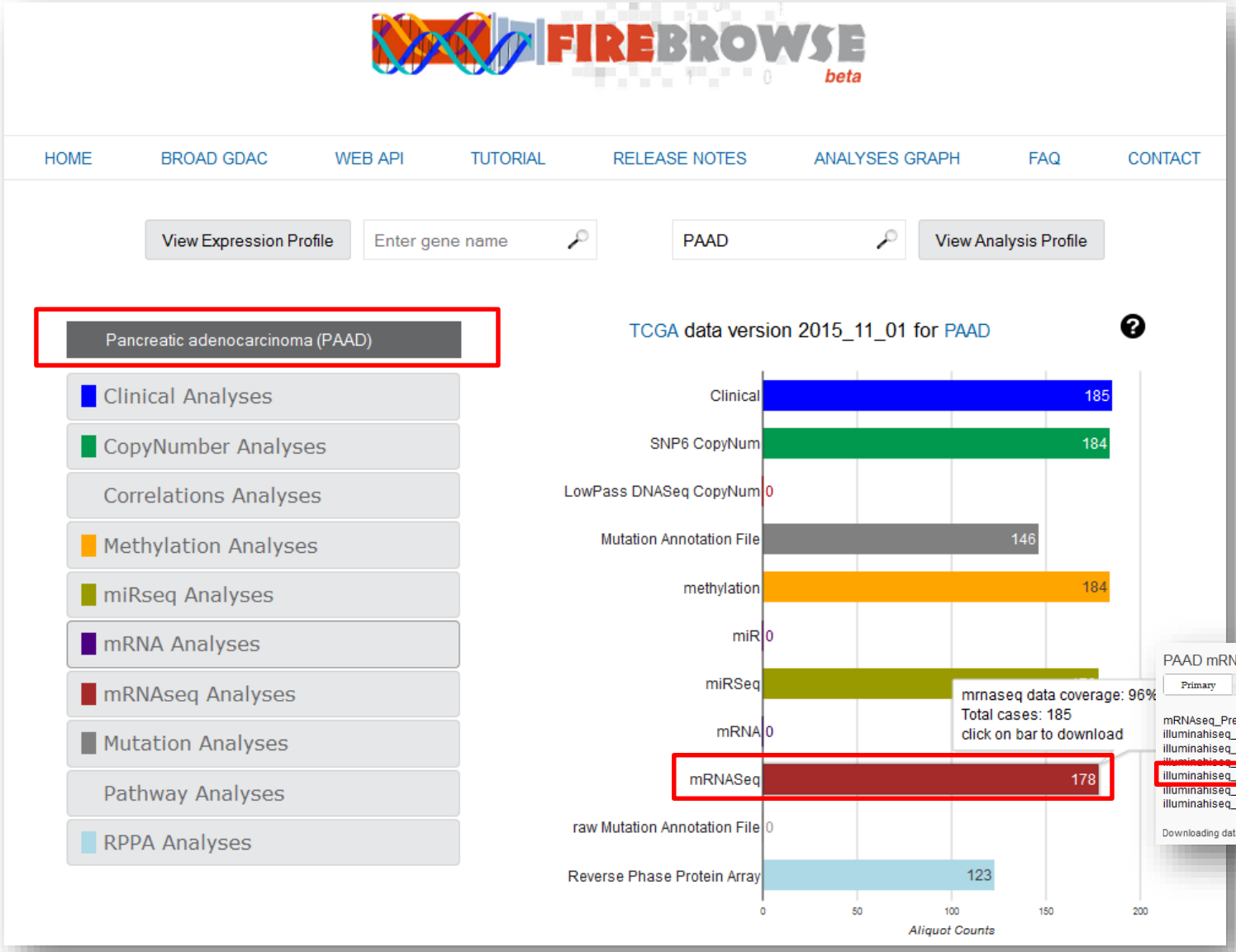
Data Level	Level Type	Description	Example
1	Raw	• Low-level data for single sample • Not normalized	• Sequence trace file • Affymetrix CEL file[1] • BAM file
2	Processed	• Normalized single sample data • Interpreted for presence or absence of specific molecular abnormalities	• Putative mutation call for a single sample • Probed locus amplification/deletion/Loss of Heterozygosity (LOH) calls in a sample • Signal of a probe or probe set for a sample
3	Segmented/Interpreted	• Aggregate of processed data from single sample • Grouped by probed loci to form larger contiguous regions (in some cases)	• Validated mutation call for a single sample • Amplification/deletion/Loss of Heterozygosity (LOH) calls for a sample region • Expression signal of a gene for a sample • Genomic copy-number data

4	Summary/Regions of Interest (ROI)	• Quantified association across classes of samples
		• Associations based on two or more
		  • Molecular abnormalities
		  • Sample characteristics
		  • Clinical variables
			• Discovery that a genomic region is amplified in 10% of TCGA glioma samples.

Sequencing data (fastq, BAM) (data level 1) is control accessed at Cancer Genomics Hub (CGHub)

Firebrowse

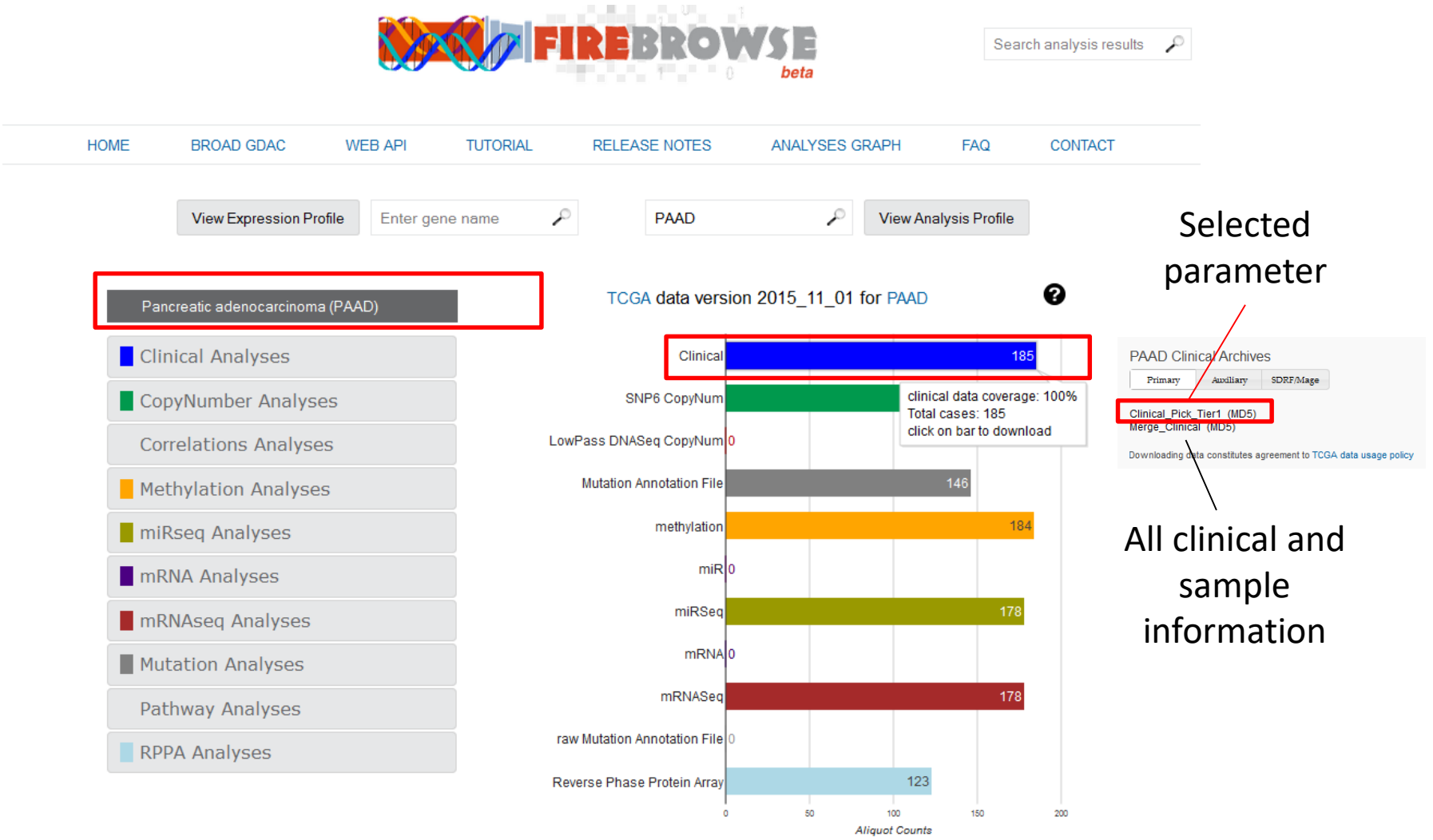
Download RNAseqV2 with Firebrowse



~_Level_3__RSEM_genes__data.data.txt

Hybridization REF	TCGA-2J-AAB1-01A-11R-A41B-07	TCGA-2J-AAB1-01A-11R-A41B-07	TCGA-2J-AAB1-01A-11R-A41B-07	TCGA-2J-AAB4-01A-12R-A41B-07	TCGA-2J-AAB4-01A-12R-A41B-07	TCGA-2J-AAB4-01A-12R-A41B-07	TCGA-2J-AAB6-01A-11R-A41B-07
gene_id	raw_count	scaled_estimate	transcript_id	raw_count	scaled_estin	transcript_id	raw_count
A1BG|1	167.92	3.43E-06	2qsd.3,uc002	134.85	2.46E-06	uc002qsd.3,u	141.16
A1CF|29974	52	9.63E-07	uc001jjk.1,uc0	127	2.03E-06	uc001jjh.2,uc	14
A2BP1|54715	1	8.82E-09	2cyx.2,uc002c	5	4.07E-08	uc002cyr.1,u	0
A2LD1|87769	370.02	8.87E-06	1,uc001vor.2,u	263.92	6.07E-06	uc001voq.1,u	278.94
A2ML1|144568	176	1.49E-06	lqva.1,uc001c	0	0	uc001quz.3,u	3105
A2M|2	40392.8	0.000548528	,uc001qvk.1,u	37630.67	0.00050451	uc001qvj.1,u	14564.83
A4GALT|53947	3160	5.56E-05	3bdb.2,uc010	2744	4.47E-05	uc003bdb.2,u	1917
A4GNT|51146	893	1.89E-05	uc003ers.2	113	2.21E-06	uc003ers.2	2
AAA1|404744	4	1.44E-07	uc010kwp.1,u	1	5.74E-08	uc003tdz.2,u	2
AAAS|8086	1402	2.85E-05	01scr.3,uc001s	1268	2.38E-05	uc001scr.3,uc	1427
AACSL|729522	1	1.69E-08	2,uc011dgk.1,	2	3.13E-08	uc003mjk.2,u	0
AACS|65985	2445	2.89E-05	2,uc009zyg.2,	2915	3.28E-05	uc001uhc.2,u	994

Download clinical data with Firebrowse



Selected parameter

All clinical and sample information

Download clinical data with Firebrowse

Patients →

Selected clinical parameter ↓

Hybridization REF		tcga-2j-aab1		tcga-2j-aab4		tcga-2j-aab6		tcga-2j-aab8		tcga-2j-aab9		tcga-2j-aaba	
Composite Element REF		value	value	value	value	value	value	value	value	value	value	value	value
years_to_birth	65	48	75	71	70	55	73	73	61	55	71	43	58
vital_status	1	0	1	0	1	1	0	1	0	0	0	0	0
days_to_death	66	NA	293	NA	627	607	NA	691	NA	NA	NA	NA	NA
days_to_last_followup		NA	729	NA	80	NA	NA	676	NA	1287	969	484	440
tumor_tissue_site		pancreas		pancreas		pancreas		pancreas		pancreas		pancreas	
pathologic_stage		stage iib		stage iib		stage iia		stage iib		stage iib		stage iib	
pathology_T_stage		t3	t2	t3	t3	t3	t3	t3	t3	t3	t3	t3	t3
pathology_N_stage		n1	n1	n0	n1	n1	n1	n0	n1	n0	n0	n1	n1
pathology_M_stage		m0	m0	m0	m0	m0	m0	m0	m0	m0	m0	m0	m0
gender	male	male	male	male	female	male	male	male	male	female	male	male	female	female
date_of_initial_pathologic_diagnosis		2012	2012	2012	2012	2012	2012	2012	2012	2012	2012	2011	2011
radiation_therapy		no	no	no	NA	yes	no	yes	yes	no	yes	yes	yes
histological_type		pancreas-adenocarcinoma-other subtype				pancreas-adenocarcinoma-other subtype						pancreas-adenoc	
number_pack_years_smoked		25	NA	NA	NA	NA	51	2.2	7.5	NA	42	NA	
year_of_tobacco_smoking_onset		1962	NA	NA	NA	NA	NA	1960	1960	NA	1973	NA	
residual_tumor	r0	r0	r0	r1	r0	r1	r0	r1	r0	r0	r1	r0	r0
number_of_lymph_nodes		7	0	0	3	6	2	0	4	0	0	2	8
race	white	white	white	white	white	white	white	white	white	white	white	white	white
ethnicity		NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA

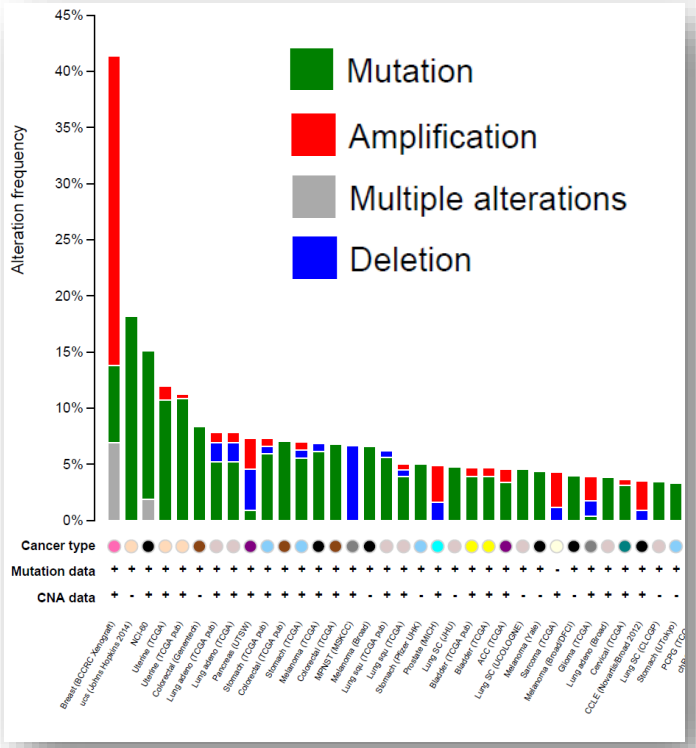
cBioPortal

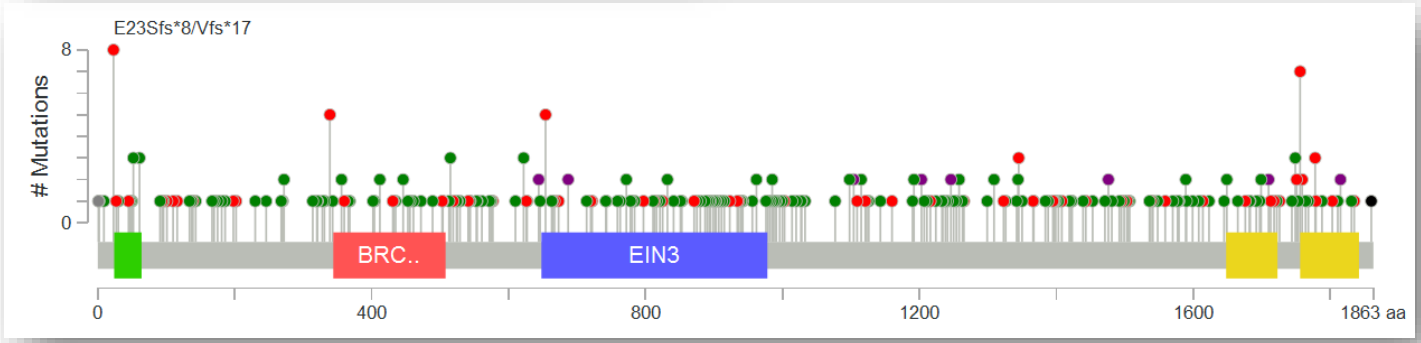
cBioPortal for Cancer Genomics

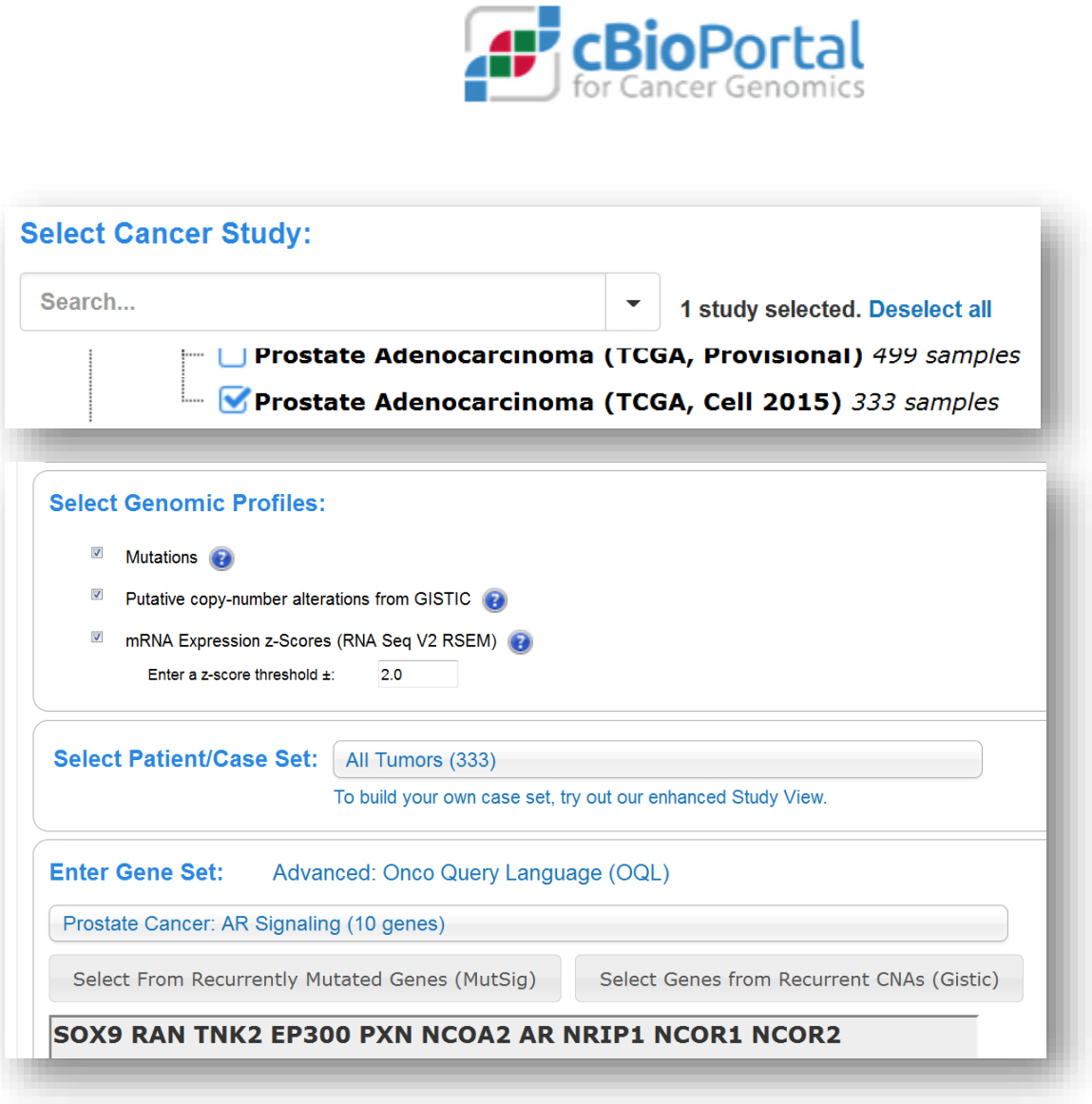
- data from 105 cancer genomics studies
- TCGA and other studies
- Query and download
- Many different analyses options

Memorial Sloan Kettering Cancer Center

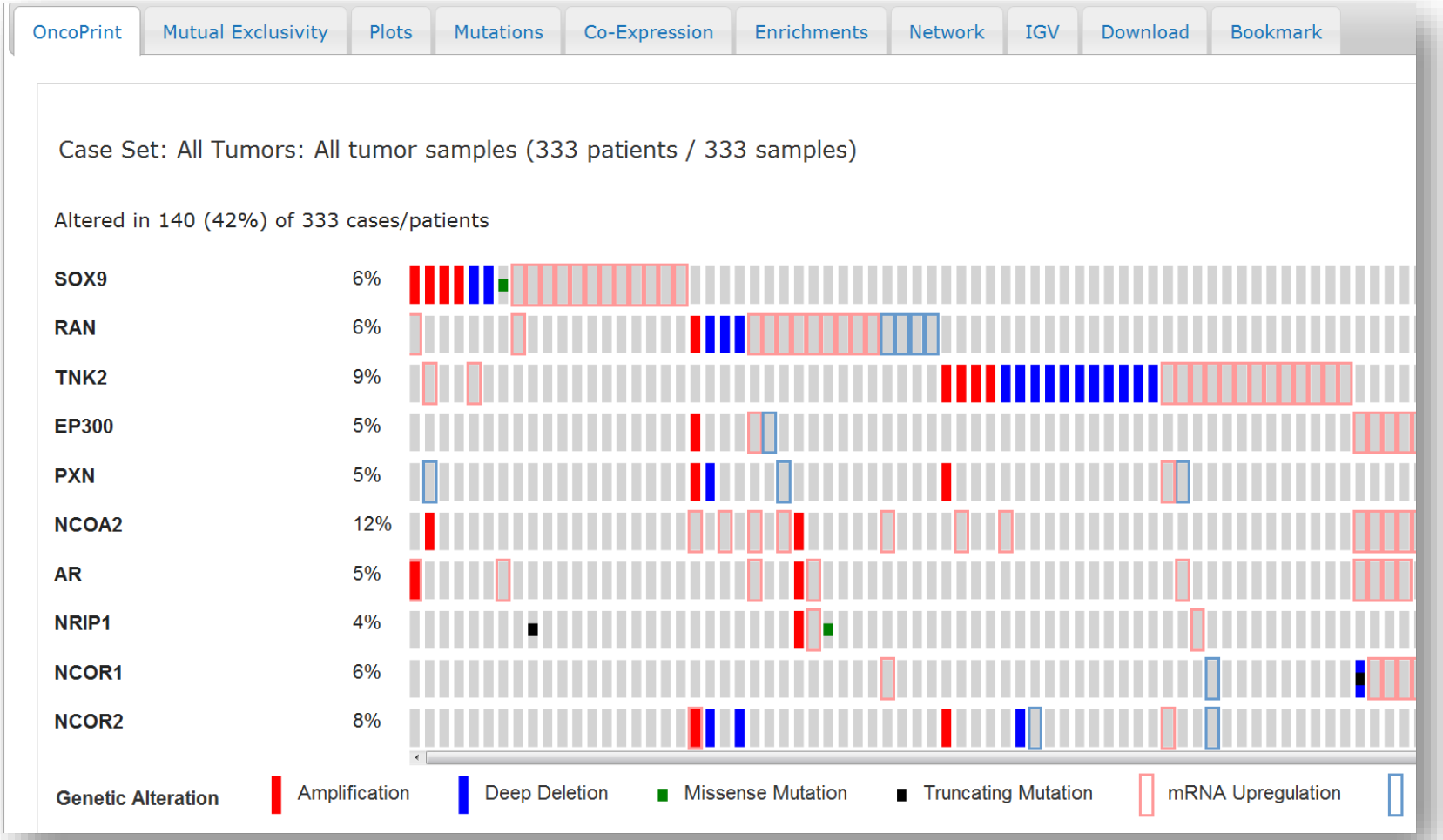
cBioPortal for Cancer Genomics






cBioPortal
for Cancer Genomics
Select Cancer Study:
Search...
1 study selected. Deselect all
Prostate Adenocarcinoma (TCGA, Provisional) 499 samples
Prostate Adenocarcinoma (TCGA, Cell 2015) 333 samples
Select Genomic Profiles:
Mutations
Putative copy-number alterations from GISTIC
mRNA Expression z-Scores (RNA Seq V2 RSEM)
Enter a z-score threshold ±:
2.0
Select Patient/Case Set:
All Tumors (333)
To build your own case set, try out our enhanced Study View.
Enter Gene Set:
Advanced: Onco Query Language (OQL)
Prostate Cancer: AR Signaling (10 genes)
Select From Recurrently Mutated Genes (MutSig)
Select Genes from Recurrent CNAs (Gistic)
SOX9 RAN TNK2 EP300 PXN NCOA2 AR NRIP1 NCOR1 NCOR2

cBioPortal for Cancer Genomics



cBioPortal for Cancer Genomics

Horizontal Axis

Genetic Profile Clinical Attribute

Clinical Attribute

Subtype

Vertical Axis

Genetic Profile Clinical Attribute

Gene EP300

Profile Type mRNA

Profile Name mRNA expression (RNA Sec

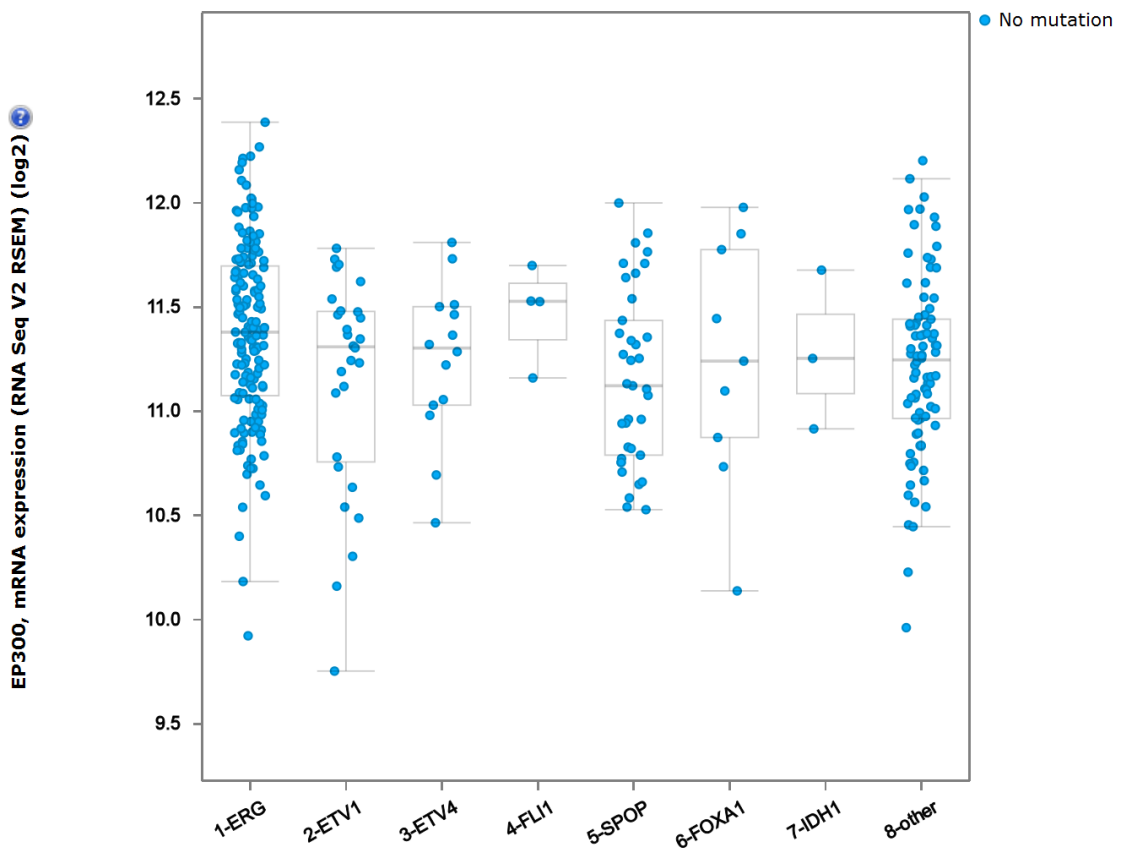
Apply Log Scale

Utilities

Search Case(s) Case ID..

Search Mutation(s) Protein Change..

Download SVG PDF Data

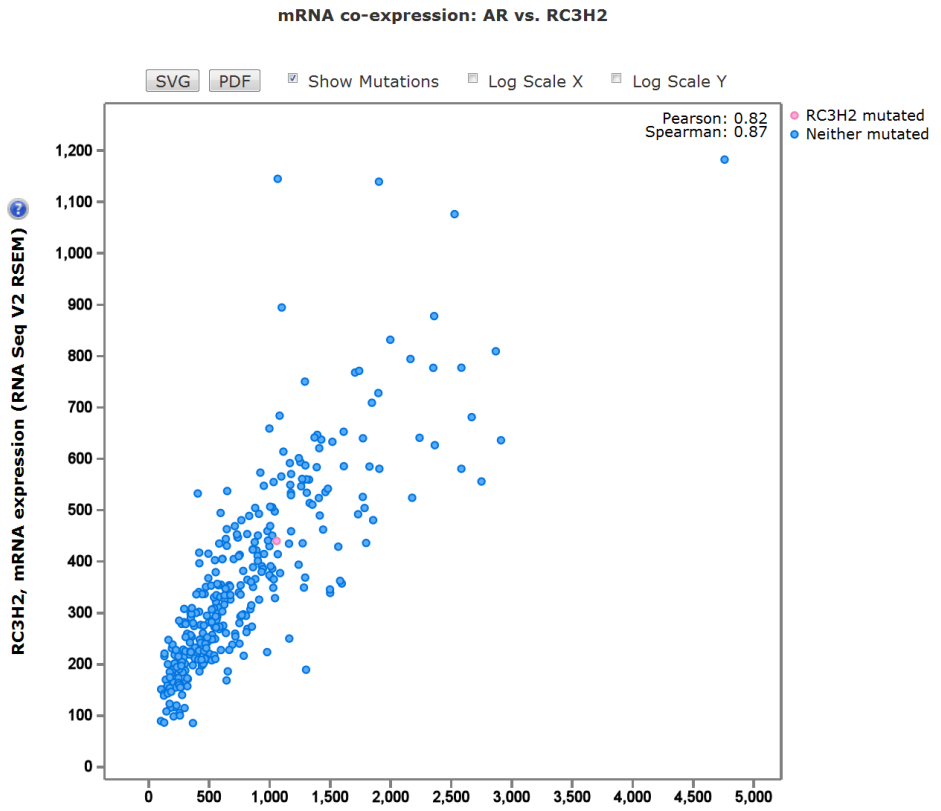


cBioPortal for Cancer Genomics

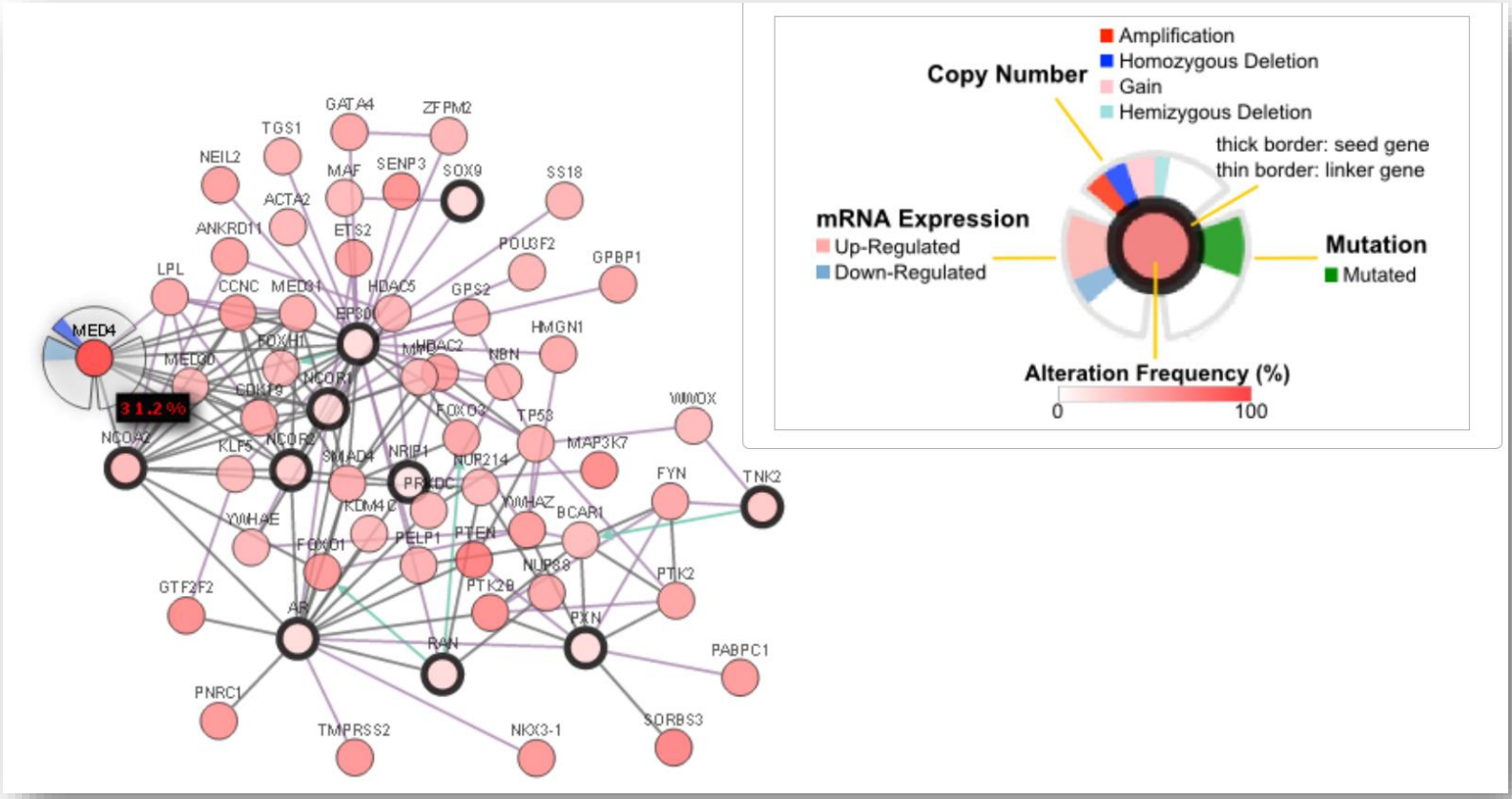
Search Gene

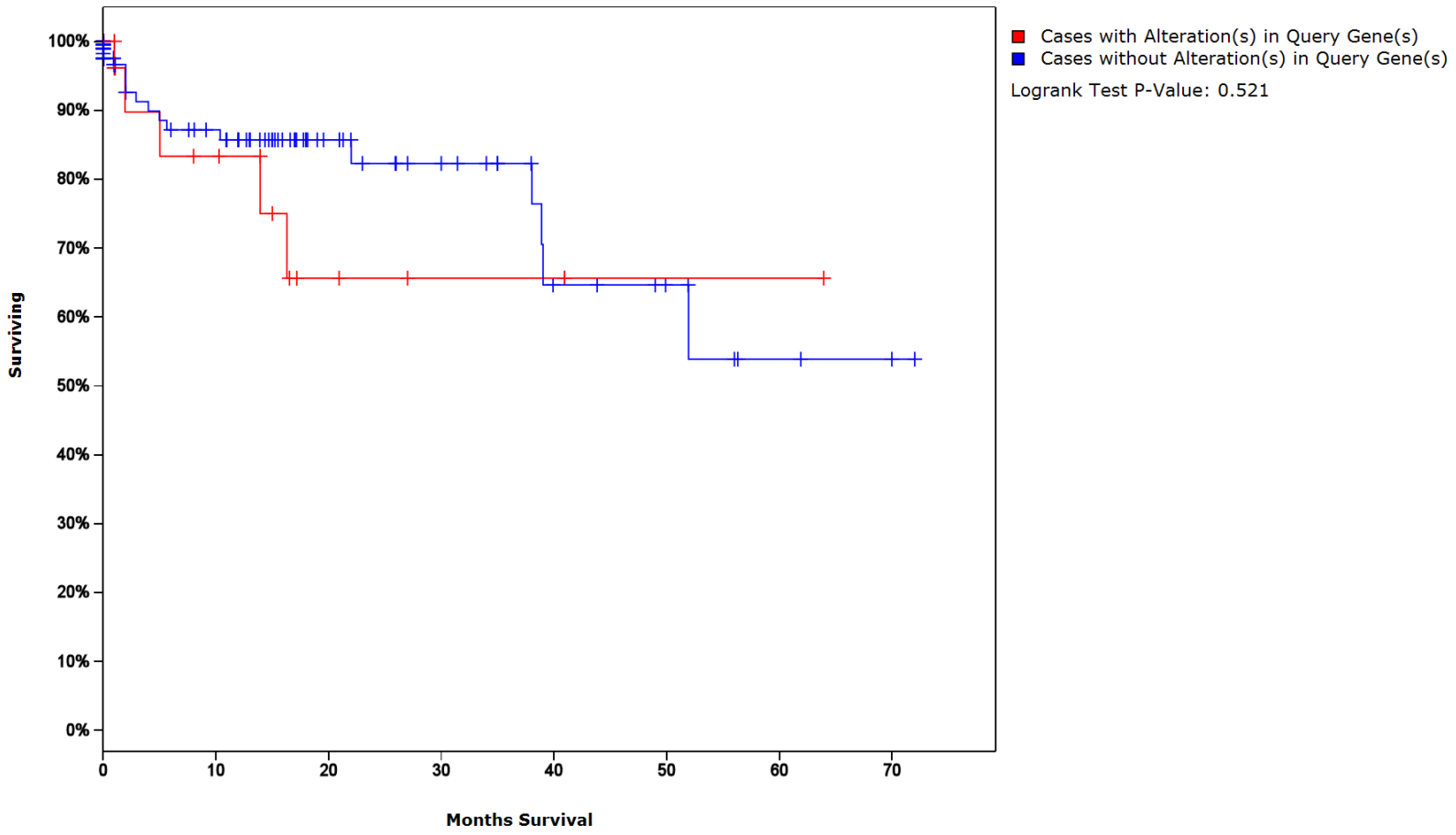
Show All

Correlated Gene	Pearson's Correlation	Spearman's Correlation
RC3H2	0.82	0.87
ERCC6L2	0.80	0.81
C9ORF129	0.79	0.85
BRWD3	0.79	0.79
UHMK1	0.79	0.86
CCNT1	0.78	0.86
GTF2A1	0.78	0.87
FAM168A	0.78	0.89
UBXN7	0.78	0.85
TOR1AIP2	0.77	0.79
TAOK1	0.77	0.85
BPTF	0.76	0.80
NCOA2	0.76	0.88
KLHL11	0.76	0.84
APOOL	0.75	0.86
ARID1A	0.75	0.80
HUWE1	0.75	0.83
GTF3C4	0.75	0.83
CLOCK	0.75	0.87
CEP97	0.75	0.82
UHRF1BP1	0.75	0.81
WDFY3	0.75	0.84
DDI2	0.75	0.88
BIRC6	0.75	0.84
LMBRD2	0.75	0.85
REST	0.74	0.84
ZNF426	0.74	0.83
NUP155	0.74	0.79



cBioPortal for Cancer Genomics




Surviving
100%
90%
80%
70%
60%
50%
40%
30%
20%
10%
0%
0
10
20
30
40
50
60
70
Months Survival
Cases with Alteration(s) in Query Gene(s)
Cases without Alteration(s) in Query Gene(s)
Logrank Test P-Value: 0.521

TCIA

TCIA

The Cancer Immunome Atlas

Home | About | Contact | Tools

Paste a list of Patient IDs

Reset Filter | Explore

ALL datasources

ALL diseases

All genders

Age >= | Age =<

ALL stages

ALL T stages

ALL N stages

ALL M stages

CRC/STAD/UCEC specific filters

ALL MSI states

ALL methylation subtypes

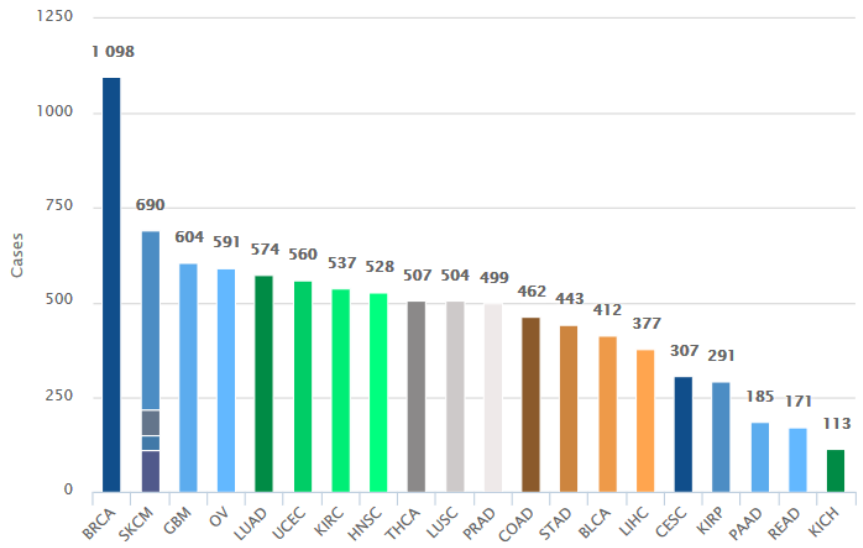
ALL hypermutation states

BRCA specific filters

ALL HER2 states

ALL tnbc states

The Cancer Immunome Database **(TCIA)** provides results of comprehensive immunogenomic analyses of next generation sequencing data (NGS) data for **20 solid cancers** from The Cancer Genome Atlas **(TCGA)** and other datasources.

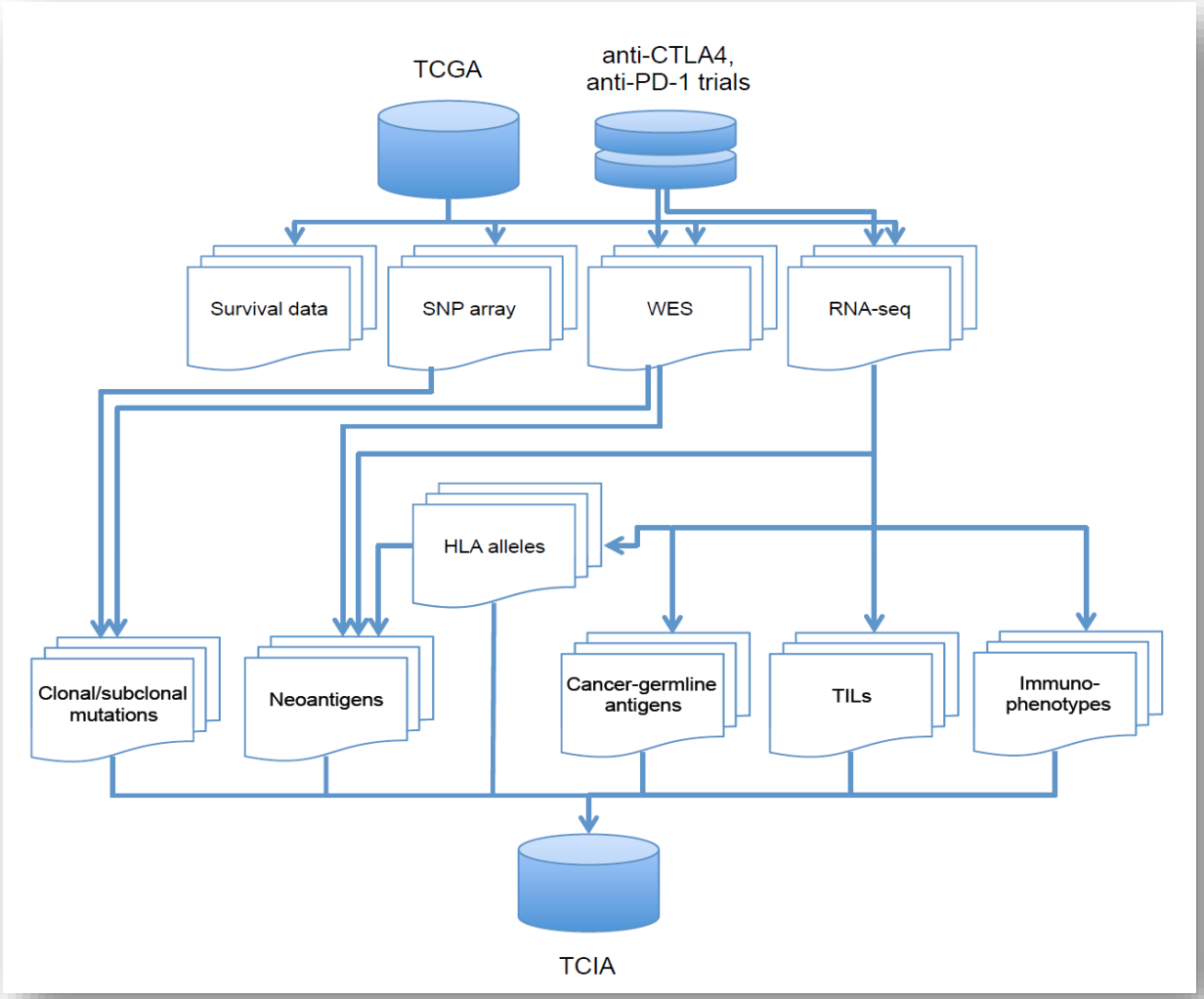


Projects 23

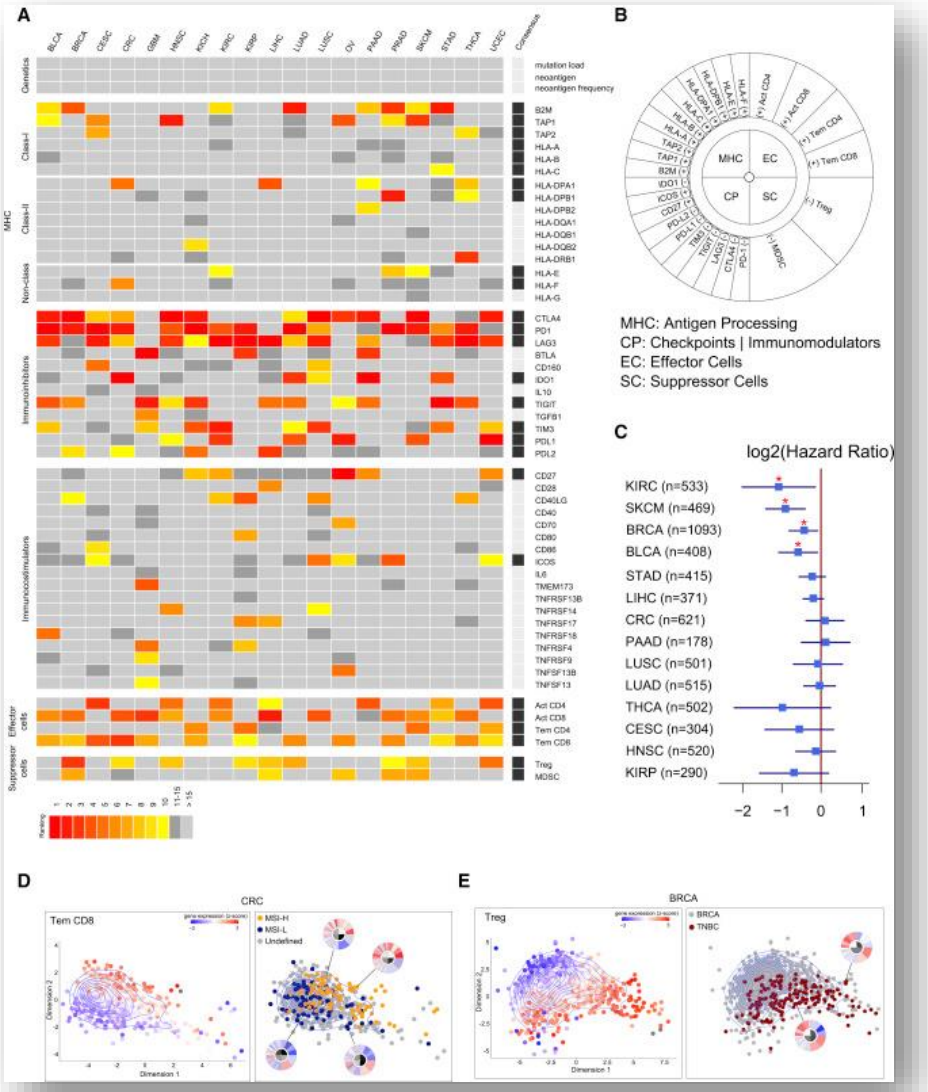
Diseases 20

Cases 9,453

The Cancer Immunome Atlas



Charoentong et al. Cell Rep. 2017. 18:248-262

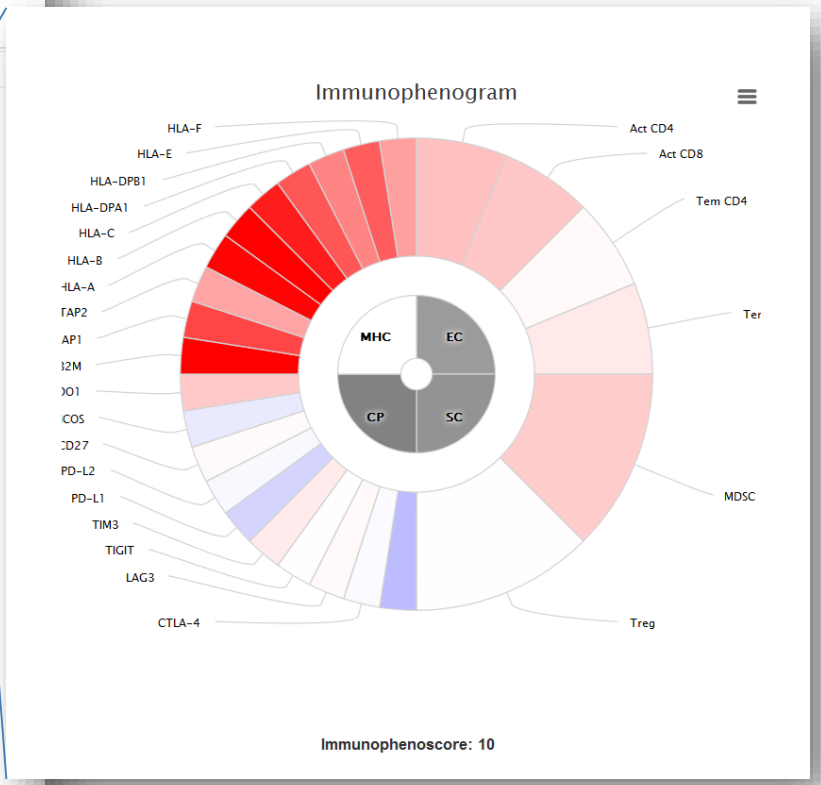


Charoentong et al. Cell Rep. 2017. 18:248-262

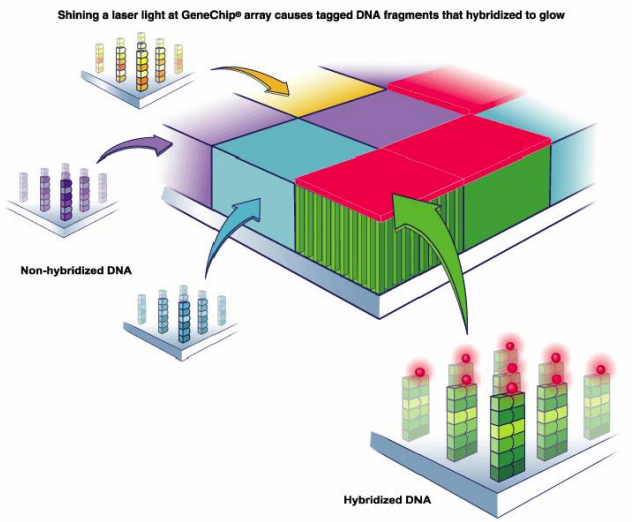
TCIA (Immunophenogram)

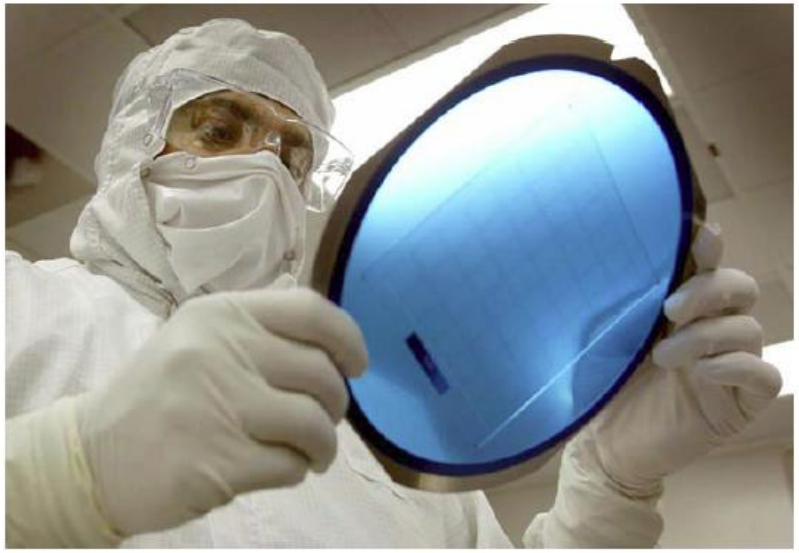
Patient	Disease	Gender	Age (years)	IPG	Study
search for patient	All	All	min >=< max		
TCGA-BH-A0B2	BRCA				
TCGA-E2-A14N	BRCA	female	37		

Clinical parameter	value
ER+	0
HER2+	0
PAM50MRNA	Basal
PR+	0
TNBC	1
date of initial pathologic diagnosis	2007
days to death	NA
days to last followup	1434
days to last known alive	NA
ethnicity	not hispanic or latino
gender	female
histological type	infiltrating ductal carcinoma
number of lymph nodes	1
pathologic stage	stage iib
pathology M stage	m0
pathology N stage	n1
pathology T stage	t2
race	white
radiation therapy	yes
tumor tissue site	breast
vital status	0
years to birth	37

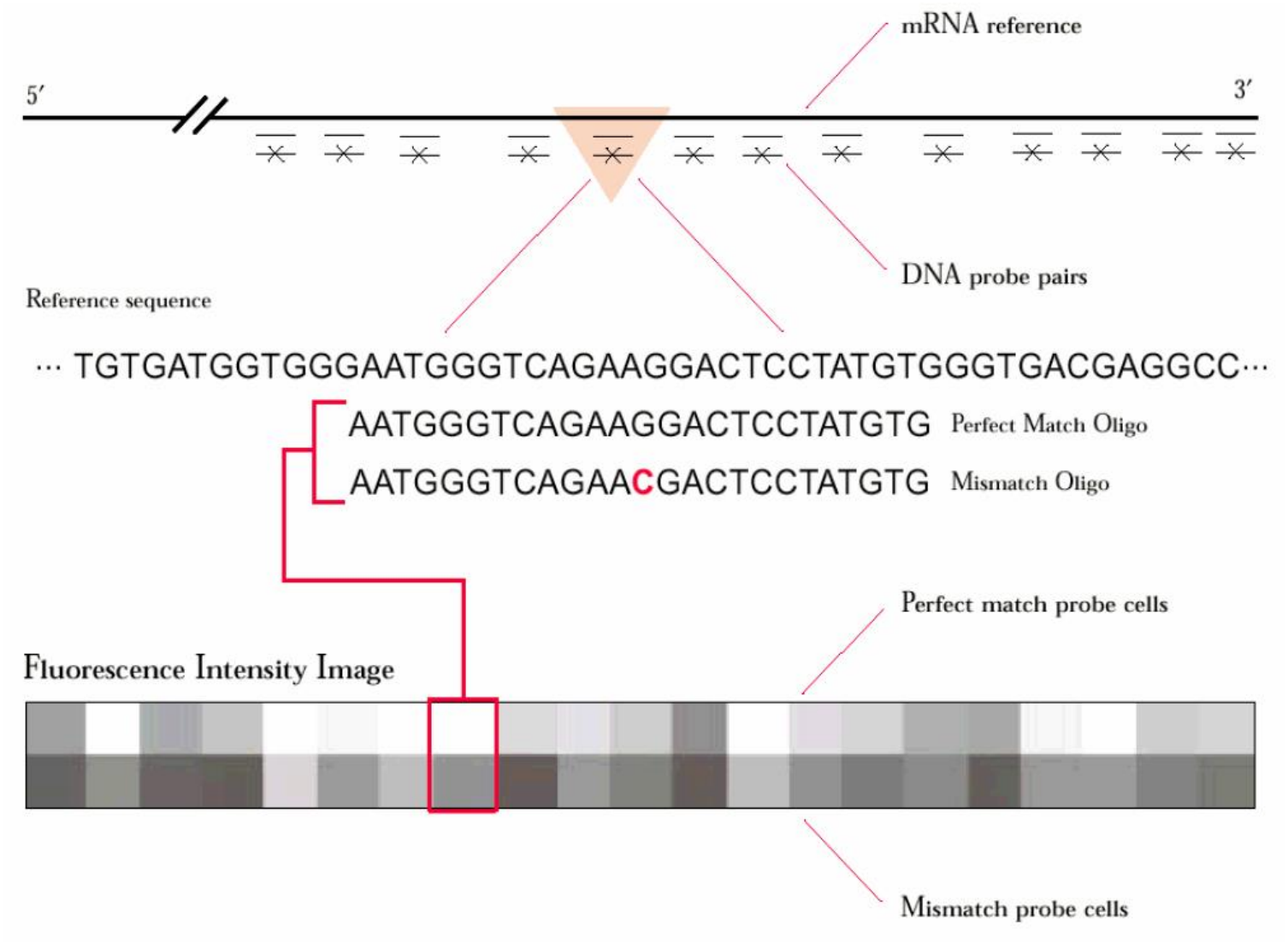


One color microarrays (Affymetrix)





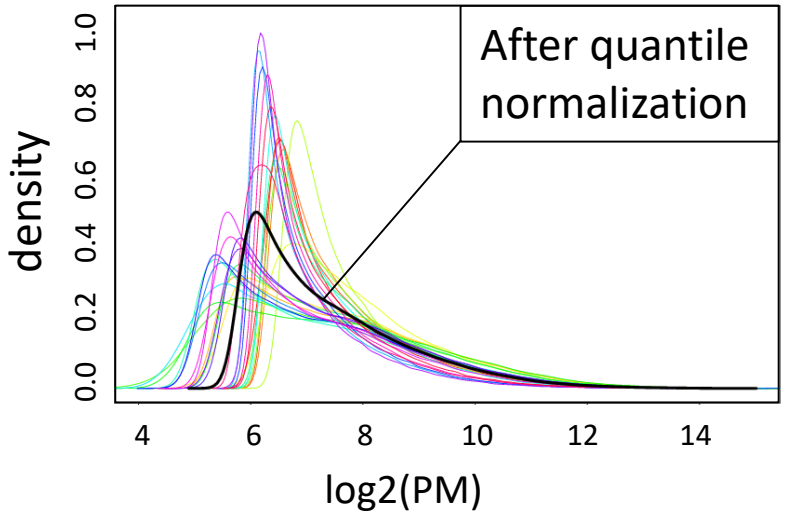
Affymetrix chips



Processing of Affymetrix chips

Robust Microarray Averaging (R/Bioconductor pkg. RMA)

- Background modeling (PM vs. MM)
- Quantile normalization across all arrays



- Probe summarization (median polish)
- Log2-transformation (log2-intensities)

Differentially expressed genes

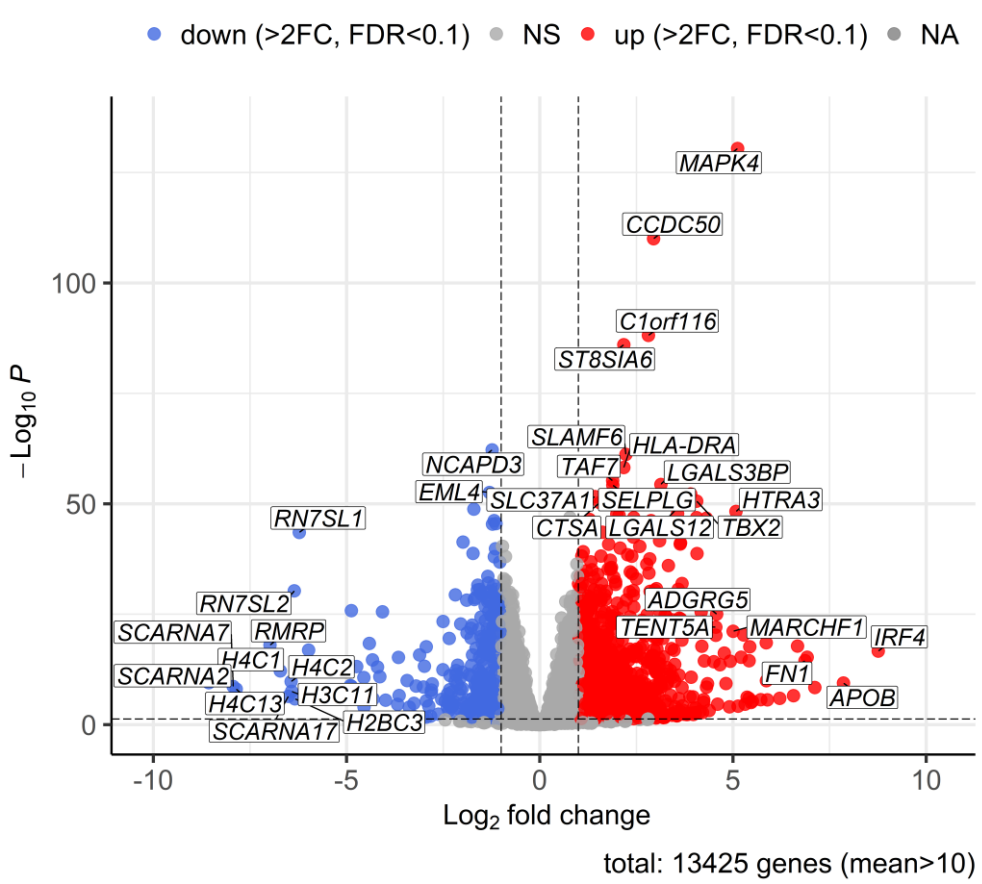
16134 probesets

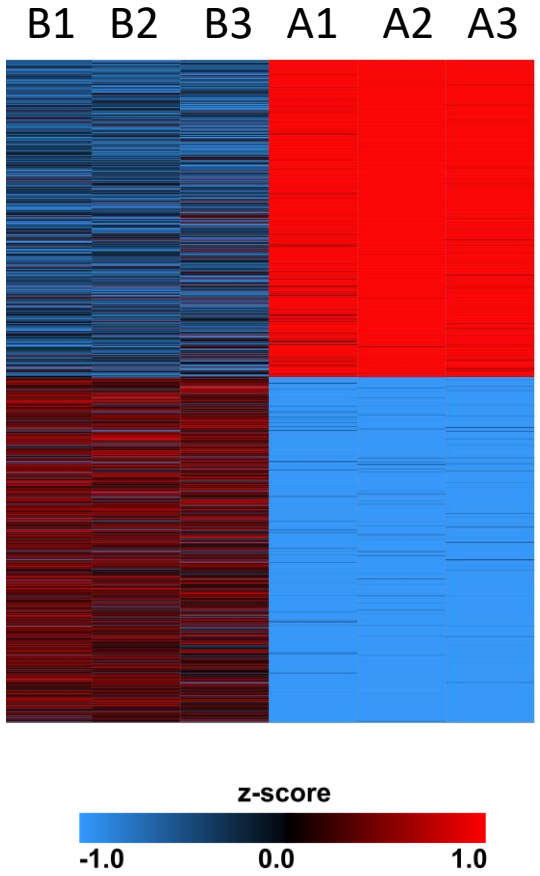
ID	GENE	KO1	KO2	KO3	WT1	WT2	WT3	logFC	AveExpr	t	P.Value	adj.P.Val
10386473	Srebf1	5.72	5.58	6.06	4.91	4.88	5.09	0.83	5.33	7.66	3.7E-09	4.6E-05
10463355	Scd2	6.63	6.26	6.92	5.13	4.77	5.01	1.64	5.59	7.52	5.6E-09	4.6E-05
10548105	Ccnd2	5.56	5.48	5.49	5.05	5.11	5.02	0.45	5.23	5.21	7.3E-06	3.9E-02
10587284	Elovl5	5.81	5.67	5.97	5.05	5.06	5.35	0.66	5.44	4.87	2.1E-05	8.4E-02
10540122	Slc6a6	7.27	7.16	7.35	6.75	6.81	6.71	0.50	7.04	4.80	2.6E-05	8.5E-02
10605437	Pls3	5.50	5.63	5.41	4.88	4.93	4.87	0.62	5.20	4.63	4.3E-05	9.7E-02
10543791	Podxl	7.30	7.03	7.08	6.31	6.52	6.33	0.75	6.59	4.61	4.6E-05	9.7E-02
10356084	Irs1	8.30	8.76	7.61	6.62	7.33	7.19	1.18	7.60	4.57	5.2E-05	9.7E-02
10346164	Sdpr	5.68	5.37	5.43	5.00	5.03	4.95	0.50	5.17	4.54	5.7E-05	9.7E-02
10387625	Chrnb1	6.31	6.08	6.06	5.73	5.59	5.81	0.44	6.01	4.52	6.0E-05	9.7E-02
10407390	Ptbp1	4.84	5.26	5.07	4.22	3.98	4.64	0.77	4.88	4.43	8.0E-05	1.1E-01
10507539	Elovl1	5.08	4.58	4.89	4.33	4.34	4.55	0.44	4.61	4.40	8.7E-05	1.1E-01
10585988	Myo9a	4.05	4.00	4.01	3.50	3.64	3.79	0.38	3.93	4.39	9.1E-05	1.1E-01
10371959	Elk3	5.94	5.85	5.78	5.28	5.44	5.46	0.47	5.66	4.38	9.3E-05	1.1E-01

condition KO vs. condition WT

Differentially expressed genes

Condition A vs. B





Differentially expressed genes

Moderated t-test (R/Bioconductor package *limma*)

$$t = \frac{\bar{M}}{(a+s)/\sqrt{n}}$$

=> p-value

estimated from all genes

- At a significance level of 0.05 in the case of 10000 tests 500 might be wrong.
- Account for this by correction for multiple hypothesis testing
  - Bonferroni correction (multiply p with number of tests)
  - Benjamini-Hochberg correction (based on the FDR)
- adjusted p-value<0.05 (<0.1) significantly differentially expressed

Methods to correct p-values for multiple testing

Ranked p	Bonferroni	Benjamini-Hochberg (FDR)
$p_{(1)}$	$p_{(1)} * n$	$p_{(1)} * n$
$p_{(2)}$	$p_{(2)} * n$	$p_{(2)} * n/2$
..	..	..
$p_{(i)}$	$p_{(i)} * n$	$p_{(i)} * n/i$
..	..	..
$p_{(n-1)}$	$p_{(n-1)} * n$	$p_{(n-1)} * n/(n-1)$
$p_{(n)}$	$p_{(n)} * n$	$p_{(n)}$

smallest p →

largest p →

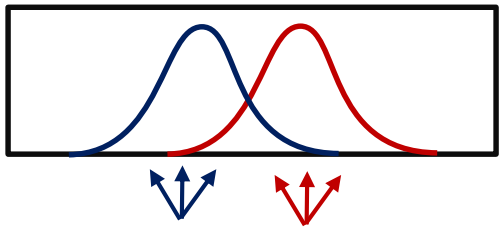
..

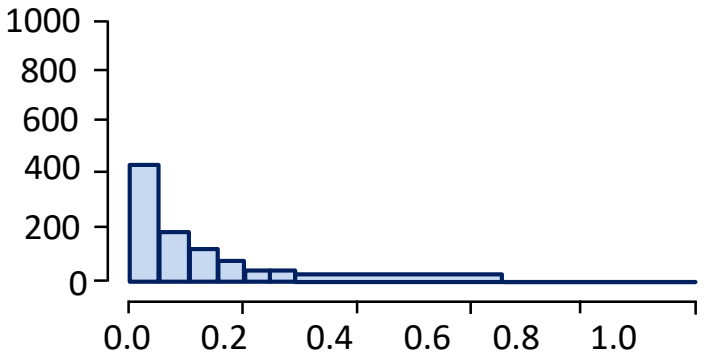
keep smaller one

$$p_{(i)}^{BH} = \min \{ \min_{j \geq i} \{ p_{(j)} * n/j \}, 1 \}$$

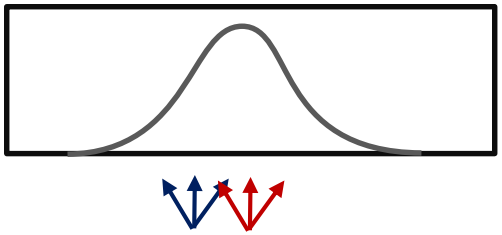
P-value distribution

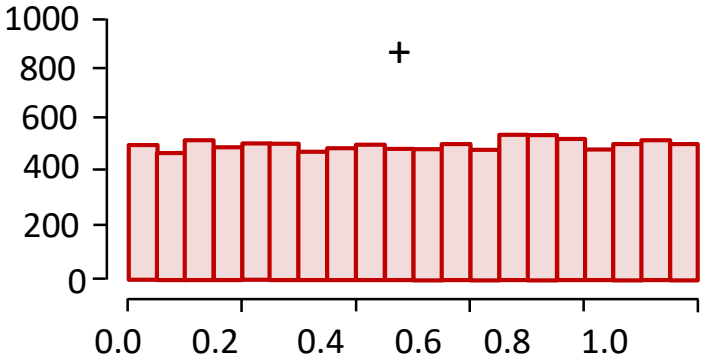
1000 genes affected by treatment
=> measurem. come from 2 different distributions





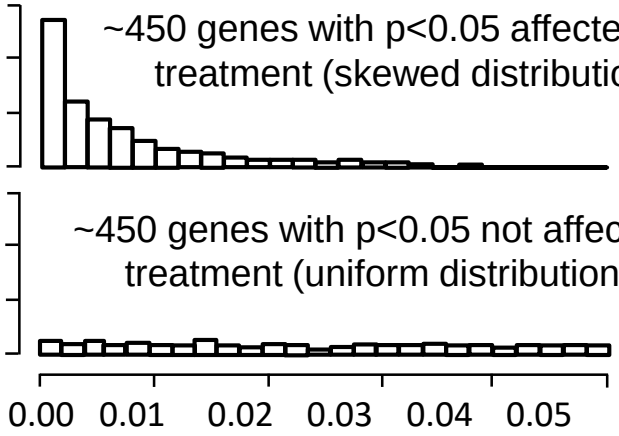
9000 remaining genes not affected by treatment
=> measurem. come from the same distribution

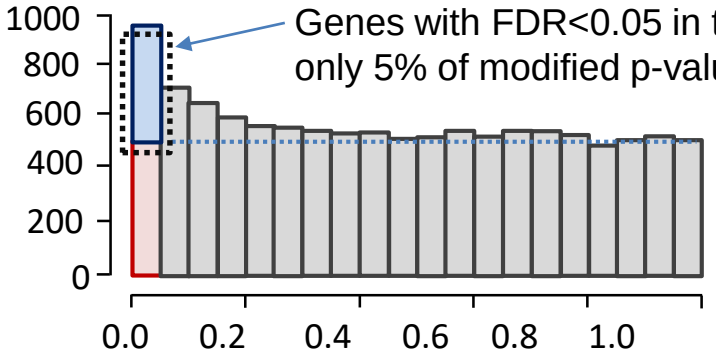




+

=

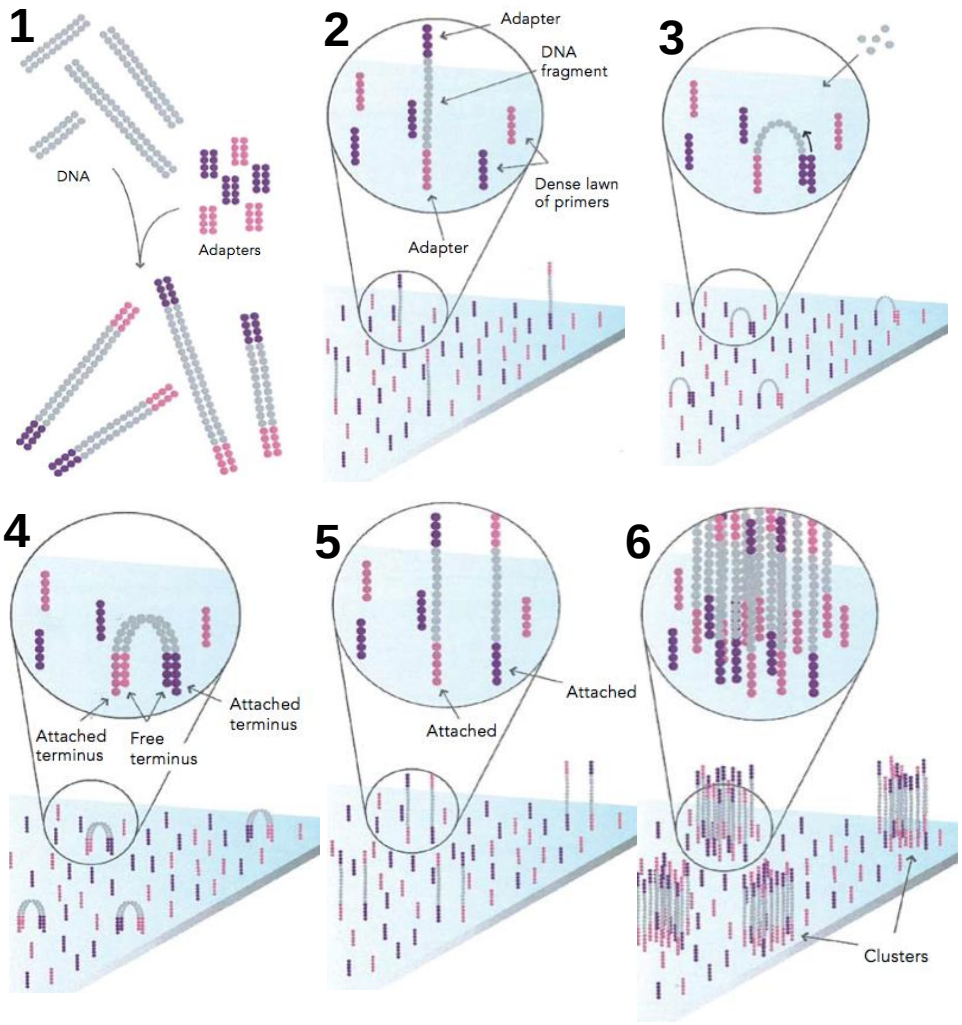




Genes with FDR<0.05 in the box
only 5% of modified p-values are FP

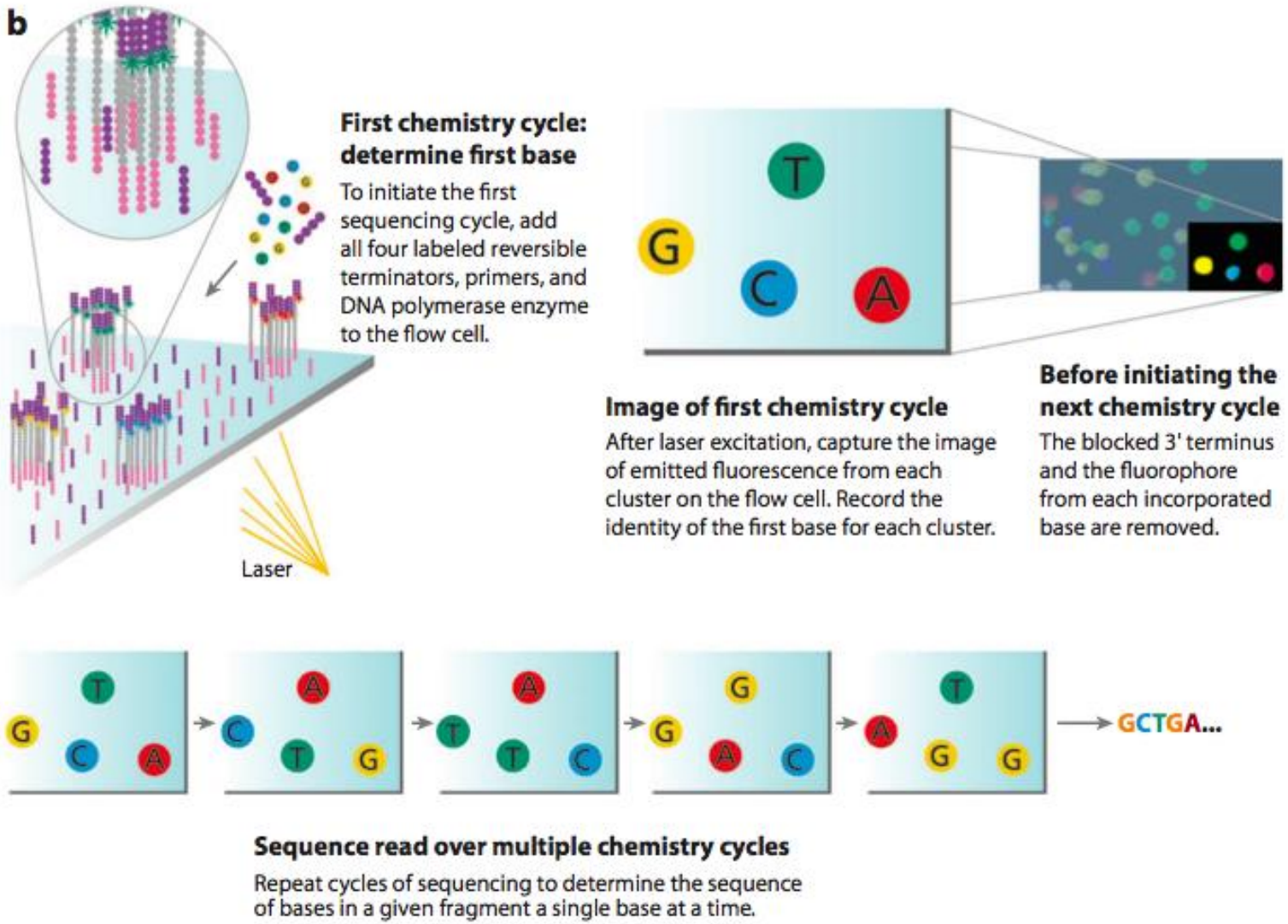
Josh Starmer (StatQuest)

Solexa (Illumina)

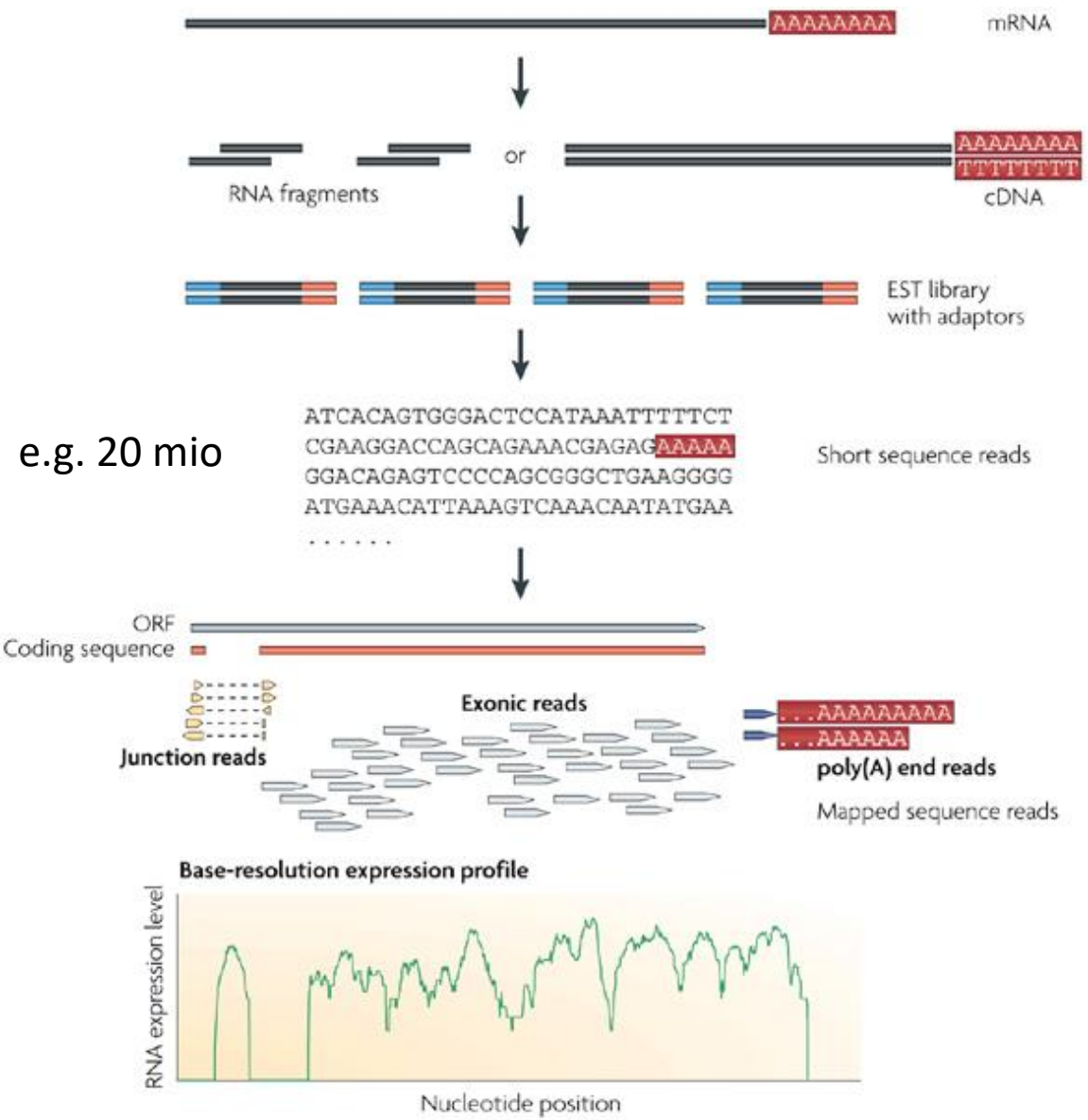


1. Prepare genomic DNA sample
2. Attach DNA to surface
3. Bridge amplification
4. Fragments become double stranded
5. Denature double stranded DNA
6. Complete amplification

Solexa (Illumina)



RNAseq


AAAAAAAA
mRNA
or
RNA fragments
AAAAAAAA
TTTTTTTT
cDNA
EST library
with adaptors
ATCACAGTGGGACTCCATAAATTTTTCT
CGAAGGACCAGCAGAAACGAGAGAAAAA
GGACAGAGTCCCCAGCGGGCTGAAGGGG
ATGAAACATTAAAGTCAAACAATATGAA
. . . . . .
Short sequence reads
e.g. 20 mio
ORF
Coding sequence
Junction reads
Exonic reads
. . .AAAAAAAAA
. . .AAAAAA
poly(A) end reads
Mapped sequence reads
Base-resolution expression profile
RNA expression level
Nucleotide position

Phred Quality Score

$Q=-10^* \log P$

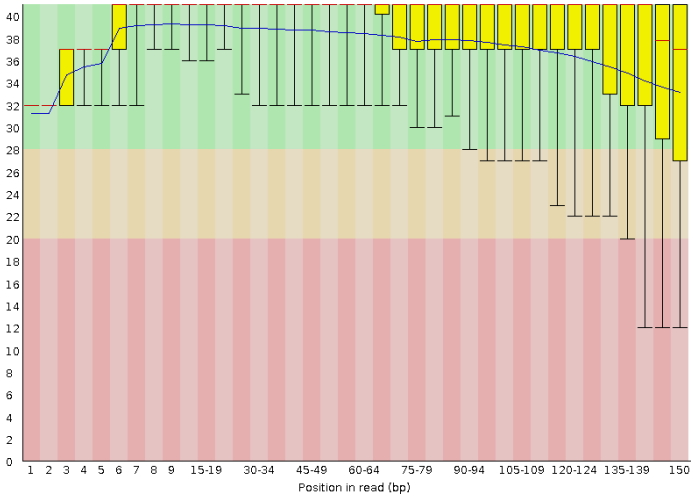
Phred Quality Score	Probability of incorrect base call	Base call accuracy
10	1 in 10	90%
20	1 in 100	99%
30	1 in 1000	99.9%
40	1 in 10,000	99.99%
50	1 in 100,000	99.999%
60	1 in 1,000,000	99.9999%

fastq format

```
@EAS54_6_R1_2_1_413_324
CCCTTCTTGTCTTCAGCGTTTCTCC
+
;;3;;;;;;;;;;;;7;;;;;;;88
```

← Q in ASCII

Quality of Sequencing (FASTQC)



Analysis steps

0. Image analysis and base calling (Phred quality score)

=> FastQ files (sequence and corresponding quality levels)

1. Trimming adaptors and low quality reads
2. Read mapping (Spliced alignment) (STAR)

=> SAM/BAM files

3. Transcriptome reconstruction (reference transcriptome, GTF file)
4. Expression quantification (transcript isoforms) (HTseq, featureCounts)

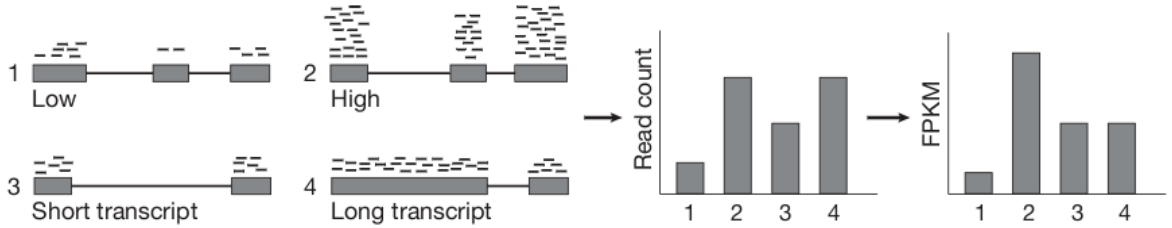
   -count reads

   -normalization

4. Differential expression analysis (negative-binomial test) (DESeq2, edgeR)

Normalization

- Reads per kilobase per million reads (RPKM)
- Fragments per kilobase per million (FPKM) for paired-end seq.



- TPM (transcripts per million)
- Quantile normalization (upper quantile normalization)
- TMM (trimmed mean of M values) (edgeR) => cpm
- Relative log expression (RLE) (DESeq2)

  ⇒ log2 (norm_counts+1)

  ⇒ regularized log (rlog)

  ⇒ variance stabilisation transform (vst)

RPKM (FPKM)

GENE	S1	S2	S3
A (2kb)	10	12	30
B (4kb)	20	25	60
C (1kb)	5	8	15
D (10kb)	0	0	1

Tens(Mio)	3.5	4.5	10.6

1. Divide by millions of reads

RPM

A (2kb)	2.86	2.61	2.83
B (4kb)	5.71	5.43	5.66
C (1kb)	1.43	1.96	1.42
D (10kb)	0.00	0.00	0.09

2. Divide by gene length in kb

RPKM

A (2kb)	1.43	1.30	1.42
B (3kb)	1.43	1.36	1.42
C (1kb)	1.43	1.96	1.42
D (10kb)	0.00	0.00	0.01

TPM

GENE	S1	S2	S3
A (2kb)	10	12	30
B (4kb)	20	25	60
C (1kb)	5	8	15
D (10kb)	0	0	1

1. Divide by gene length in kb

RPK

A (2kb)	5	6	15
B (4kb)	5	6.25	15
C (1kb)	5	8	15
D (10kb)	0	0	0.1

Tens(Mio)	1.5	2.025	4.51

2. Divide by millions of RPK

TPM

A (2kb)	3.33	2.96	3.326
B (3kb)	3.33	3.09	3.326
C (1kb)	3.33	3.95	3.326
D (10kb)	0	0	0.02

TCGA

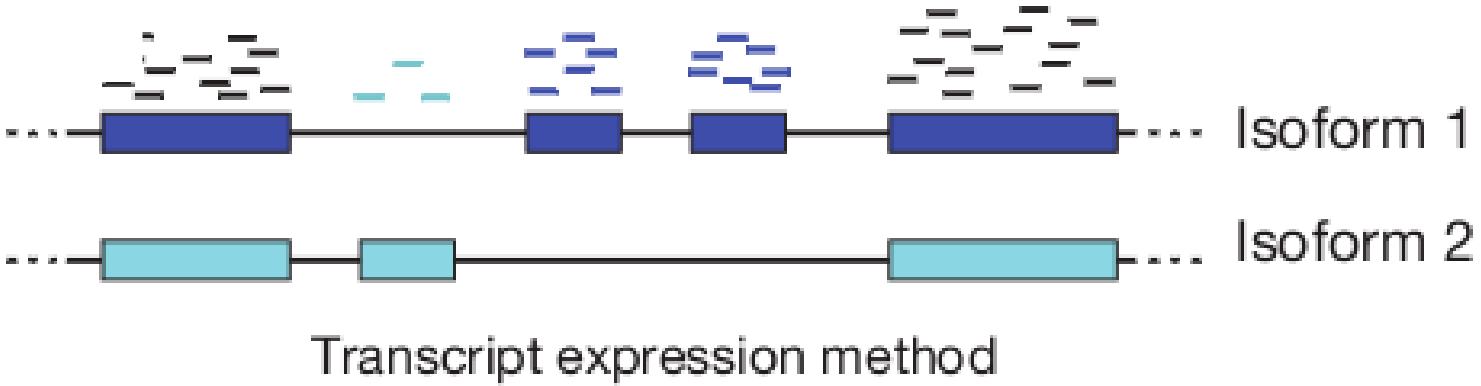
RNAseqV2 analysis ***MapSplice*** is used to do the alignment and ***RSEM*** to perform the quantitation.

raw_count … for EBseq introduce directly
for DESeq2, EdgeR use integers

scaled_estimate … transcript per million TPM=scaled estimate*10^6

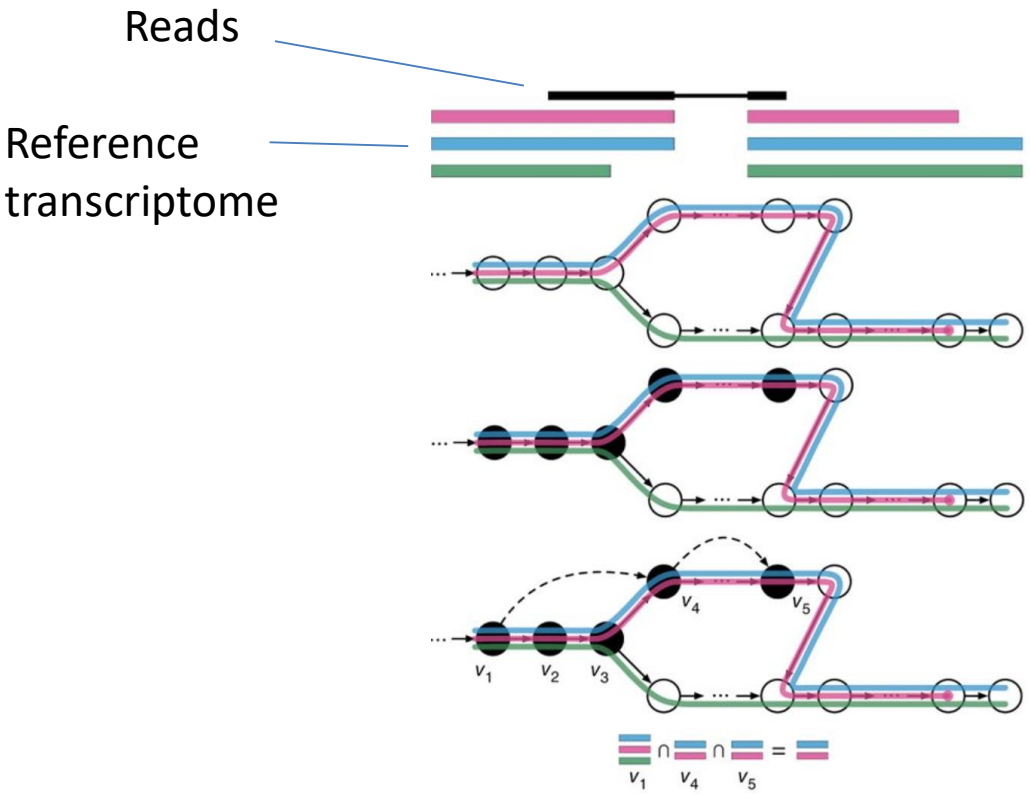
normalized_count …. upper quartile normalized RSEM count estimates

Isoform quantification



- Uncertainy in assigning reads to isoforms
- Paired-end sequencing
- Spliced alignment
- Alternative splicing (statistical significant?)

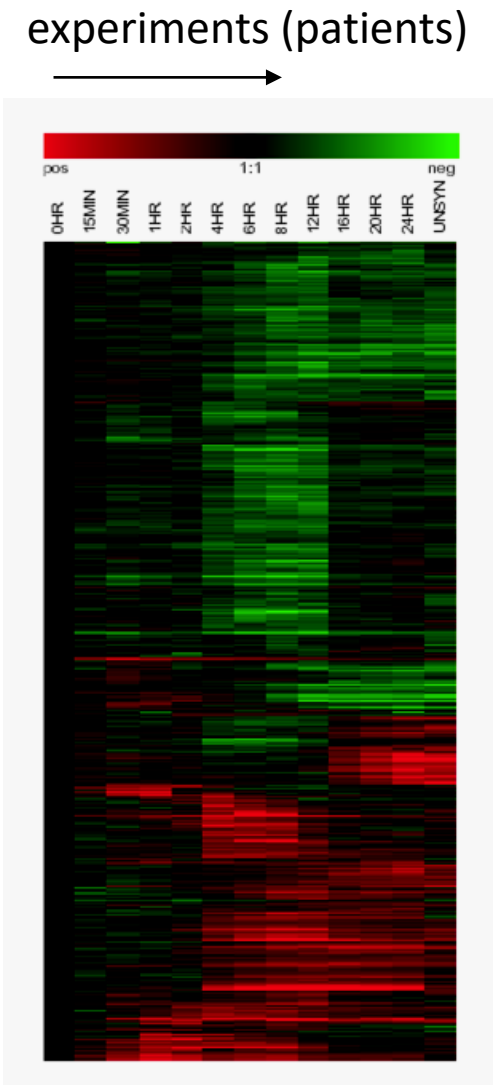
RNA seq quantification using pseudoalignment (kallisto)



Transcriptome de Bruijn Graph (T-DBG) where nodes (v1, v2, v3, ...) are *k*-mers

Bray et al. Nature Biotechnology 2016

Representation of gene expression



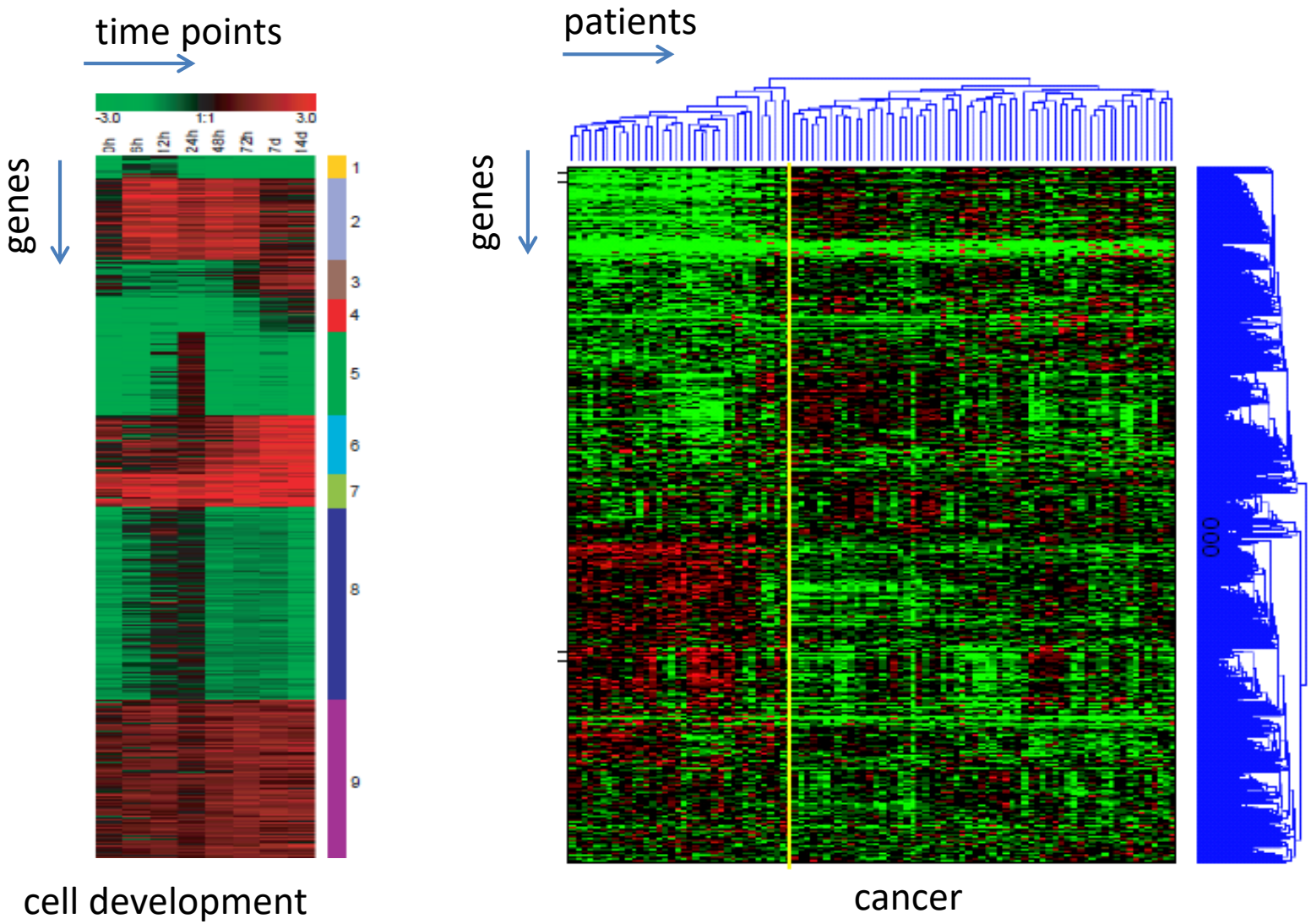
heatmap

- *n* x *m* matrix with *n* genes and *m* experiments (conditions, patient samples)
- Representation as heatmap (e.g. *red* upregulated genes, green down regulated genes, *black* no change)
- For experiments in reference design:
  - $\log_2$-fold change ($\log_2$FC, $\log_2$(A/B), $\log_2$ ratio)
- For patient samples and no reference:
  - mean centered $\log_2$-levels for each gene

    $\log_2$-intensities for one-color arrays
    $\log_2$-RPKM for RNAseq
  - z-score of $\log_2$-levels

    $Z = (X-m)/s$ X…$\log_2$-levels, m…mean, s…standard deviation

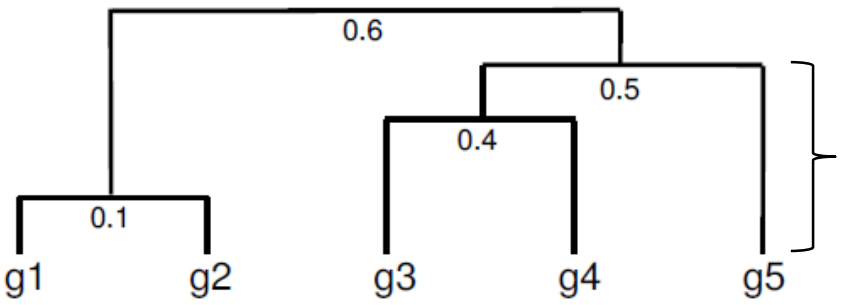
Gene expression profiling



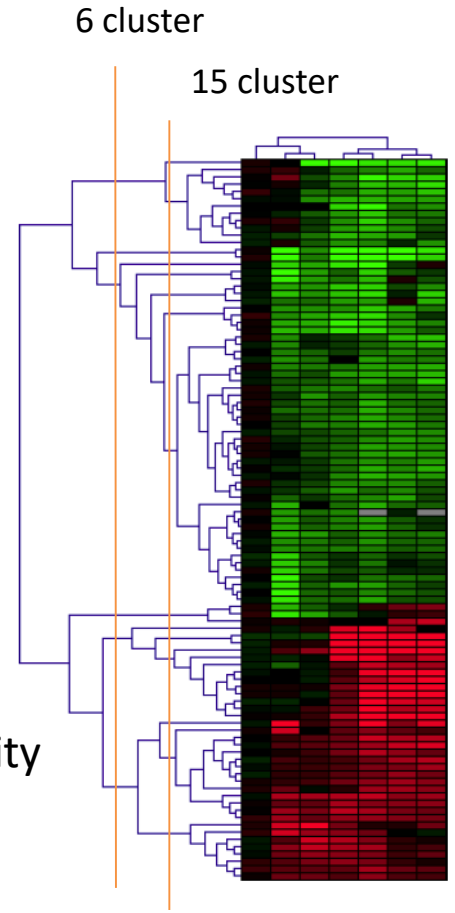
Hierarchical clustering

- Agglomerative (bottom up), unsupervized
- Cluster genes or samples (or both= biclustering)
- Distances are encoded in dendogram (tree)
- Cut tree to get clusters
- Pearson correlation, Euclidean distance
- Computational intensive (correlation matrix)

1. Identify clusters (items) with closest distance
2. Join to new clusters
3. Compute distance between clusters (items) (see linkage
4. Return to step 1

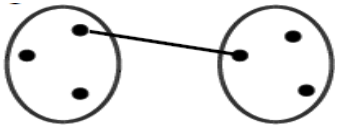


Dendrogram

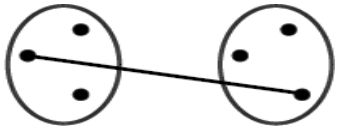


Linkage

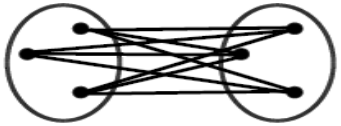
Single-linkage clustering
Minimal distance



Complete-linkage clustering
Maximal distance



Average-linkage clustering
Calculated using average distance (UPGMA)
Average from distances not! expression values



Weighted pair-group average
Like UPGMA but weighted according cluster size

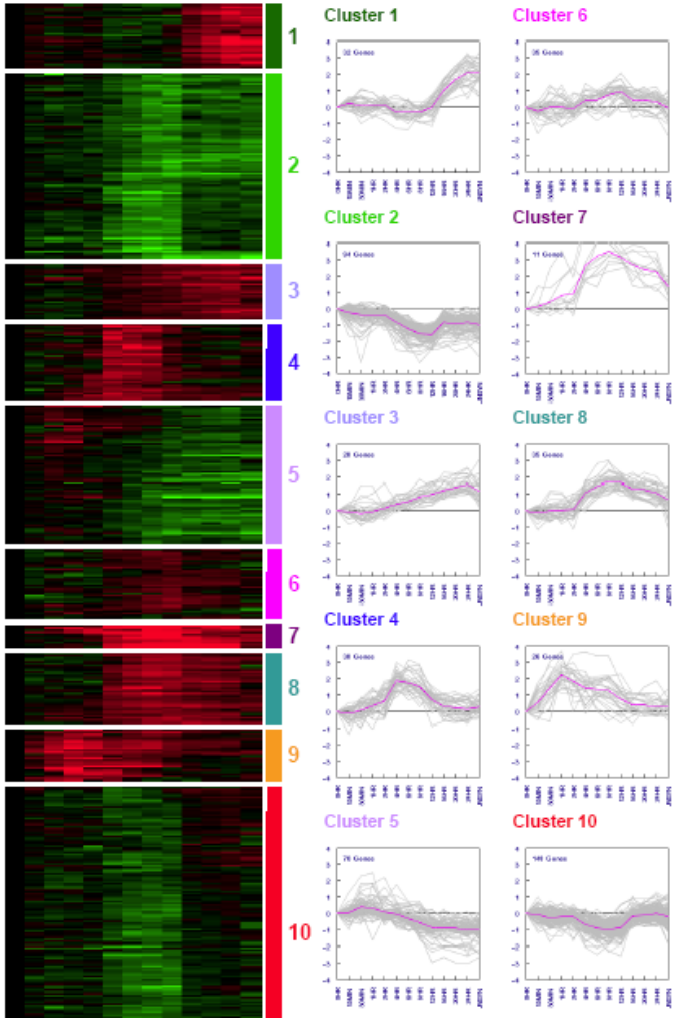
Within-groups clustering
Average of merged cluster is used instead of cluster elements

Ward's method
Smallest possible increase in the sum of squared errors

K-means clustering

- partition *n* genes into *k* clusters, where *k* has to be predetermined
- k-means clustering minimizes the variability within and maximize between clusters
- Moderate memory and time consumption

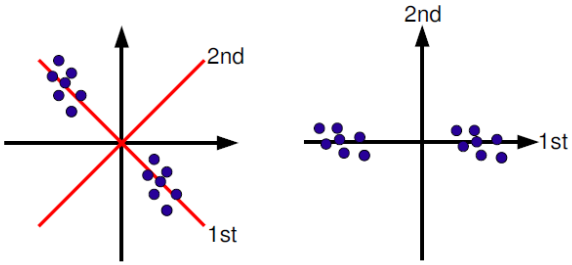
1. Generate random points ("cluster centers") in n dimensions (results are depending on these seeds).
2. Compute distance of each data point to each of the cluster centers.
3. Assign each data point to the closest cluster center.
4. Compute new cluster center position as average of points assigned.
5. Loop to (2), stop when cluster centers do not move very much.



Principal component analysis (PCA)

PCA is a data reduction technique that allows to simplify multidimensional data sets into smaller number of dimensions (r<n).

Variables are summarized by a linear combination to the principal components. The origin of coordinate system is centered to the center of the data (mean centering) . The coordinate system is then rotated to a maximum of the variance in the first axis.

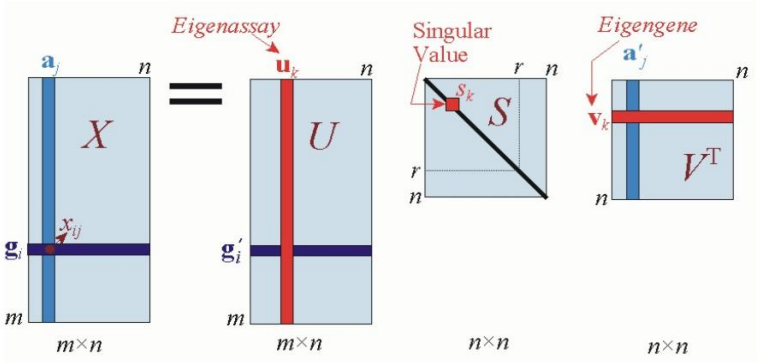


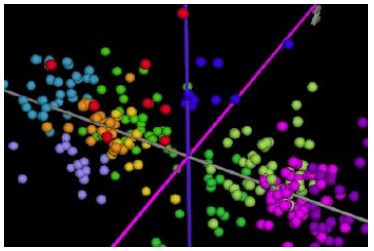
Subsequent principal components are orthogonal to the 1st PC. With the first 2 PCs usually 80-90% of the variance can already be explained.

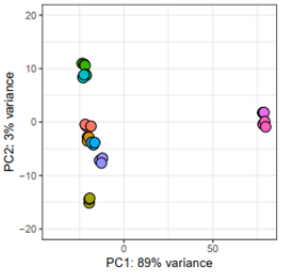
This analysis can be done by a special matrix decomposition (singular value decomposition SVD).

Singular value decomposition (SVD)

$$X = USV^T \text{ with } UU^T = V^TV = VV^T = I$$







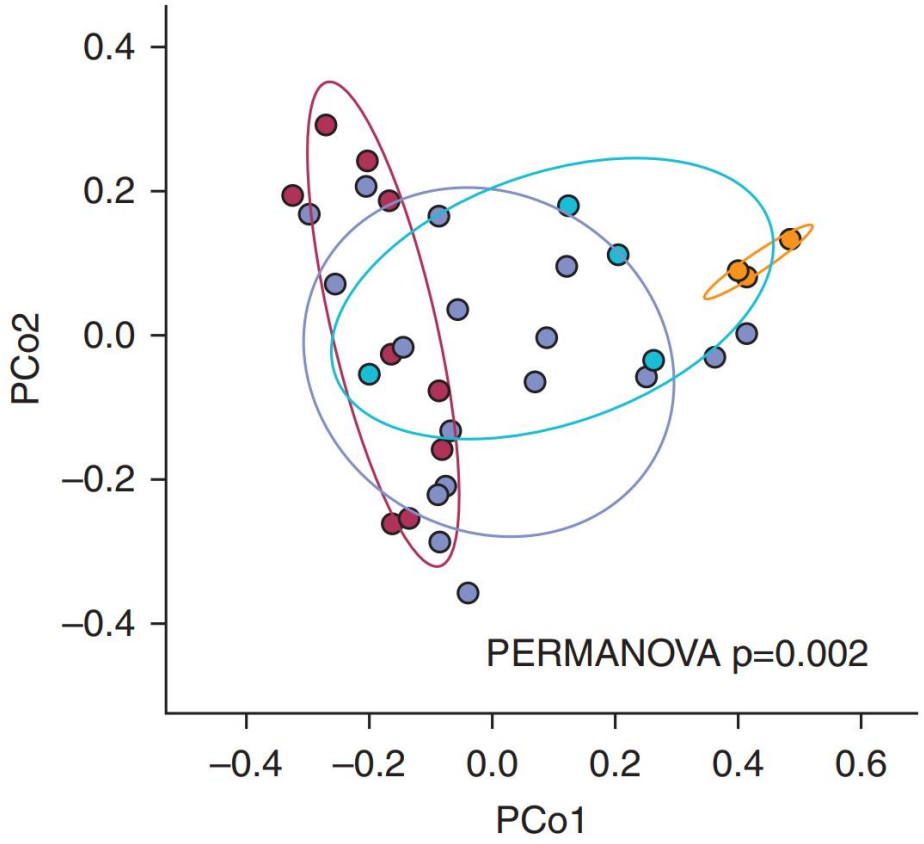
For mean centered data the Covariance matrix C can be calculated by XX^T. U are eigenvectors of XX^T and the eigenvalues are in the diagonal of S defined by the characteristic equation $|C - \lambda I| = 0$.

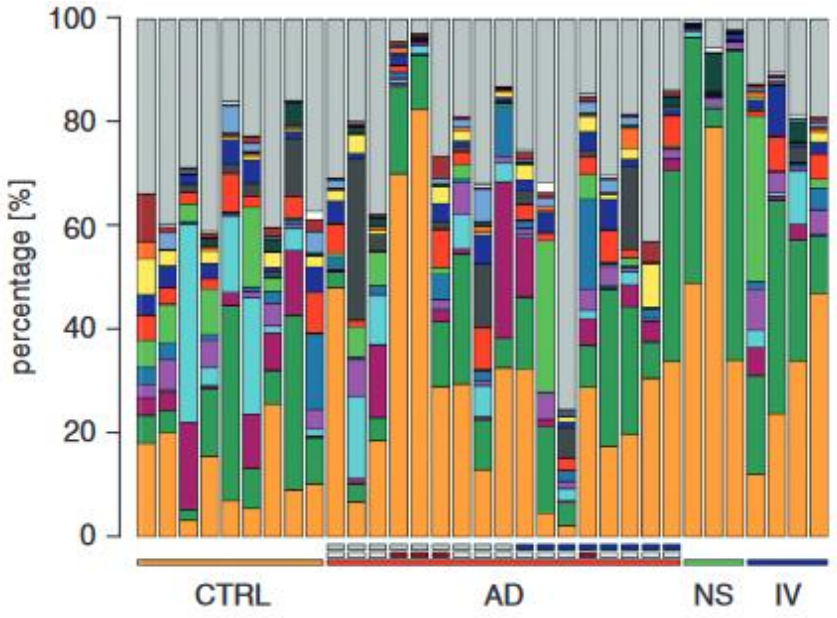
Transformation of the input vectors into the principal component space can be described by $Y = XU$ where the projection of sample i along the axis is defined by the j-th PC:

$$y_{ij} = \sum_{t=1}^{m} x_{it} u_{tj}$$

Metric multidimensional scaling (MDS) = Principal Coordinate Analysis (PCoA)

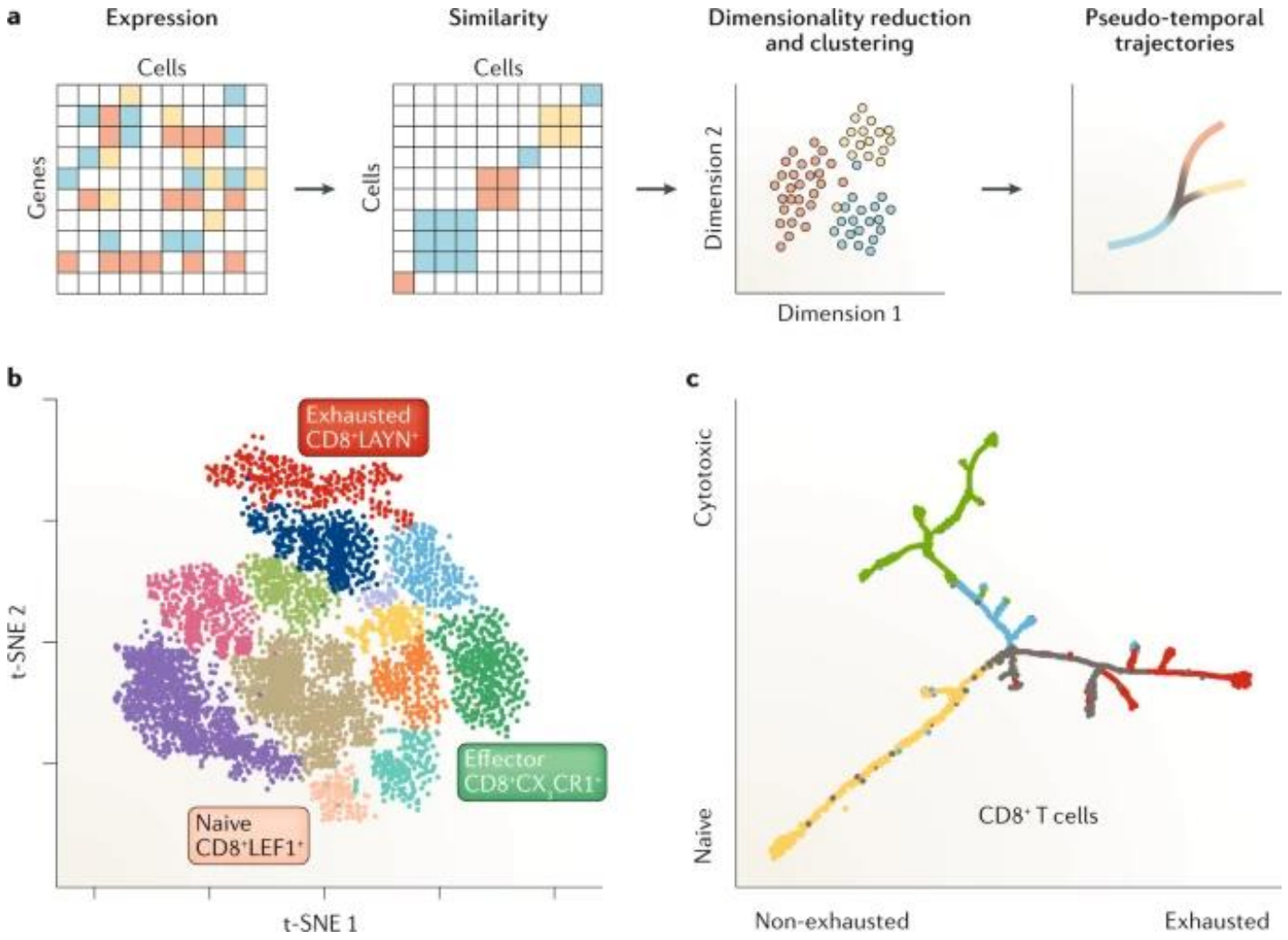
e.g. Microbiome based on Bray-Curtis dissimilarity index





Moosbruger-Martinz et al. J Invest Derm 2020

Single-cell RNAseq analysis



Finotello F et al. Nat Rev Genet 2019

Single cell RNAseq

- Cell barcodes are random sequences to tag single cells used for multiplex sequencing
- UMIs are random sequences specific for each molecule to avoid amplification bias.

Illumina paired-end RNA-seq adapters P5-P7

Read 1:

- Barcode: cell-specific (16 bp)
- UMI: transcript-specific (12 bp)

Read 2:

- cDNA fragment of interest (only 3' tag)

  - Microwell based
  - Droplet based

Single cell RNAseq analyses

- Seurat (R based, Butler 2018), Scanpy (Python based, Wolf 2018)
- QC filtering based on number of counts, number of detected genes, counts from mitochondrial genes to filter out dying cells (low counts but high mitochondrial fraction) or multiplets (high counts, high number of detected genes) but could be confounded by big cells, high mRNA content , quiescent cells, activated cellular respiration porcesses
- Gene filtering if only expressed in a few cells (advanced algorithms)
- Global scaling normalization (for each cell divide by the total expression, multiplies this by a scale factor (10,000 by default), and log-transforms
- Batch effect removal
- Regression out biological effects (e.g. cell cycle processes based on marker genes)
- Scaling (mean=0, variance=1) and linear dimension reduction (PCA)
- Cluster cells by graph embeded methods such as KNN graph and modularity optimization techniques such as the Louvain algorithm
- Nonlinear dimension reduction (tSNE, UMAP)
- Differentially expressed genes between clusters
- Identify marker genes and assign cell type identy to clusters

t-distributed stochastic neighbor embedding (t-SNE)

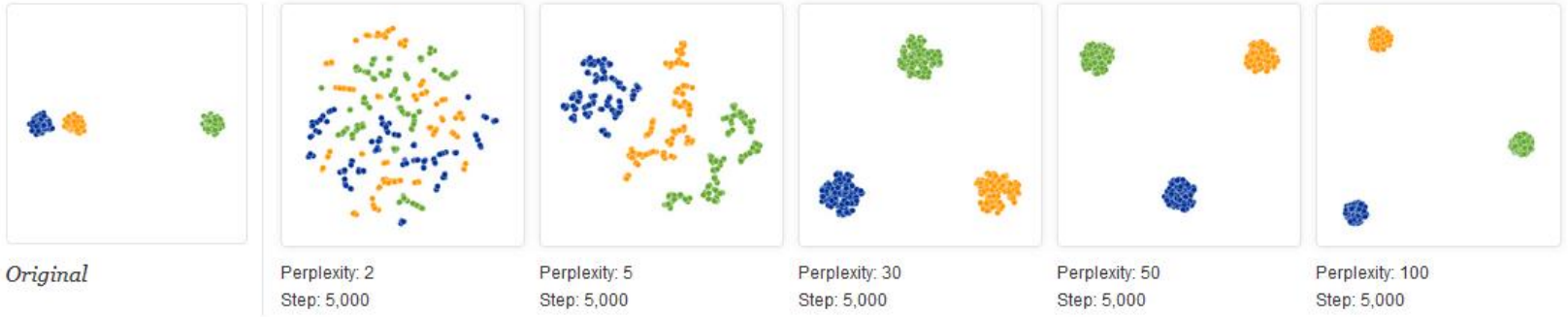
Non-linear dimension reduction method

Converting the high-dimensional Euclidean distances between data points into **conditional probabilities** (based on Students T distribution) that represent similarities.

Minimization of the sum of difference of conditional probability t-SNE minimizes the sum of Kullback-Leibler divergence of overall data points using a gradient descent method.

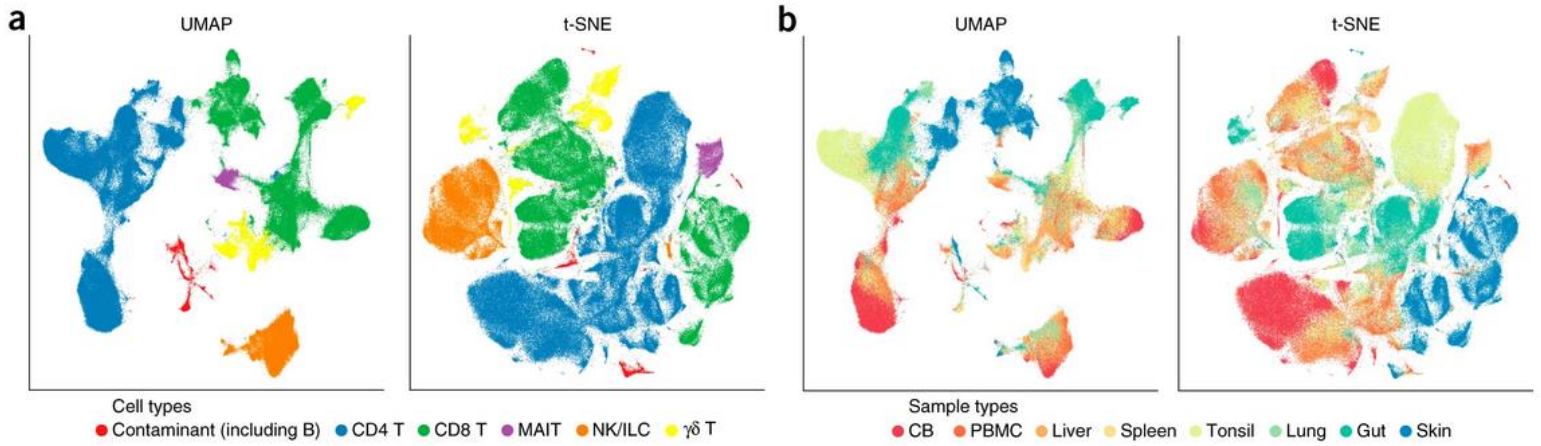
Hyperparamter: Perplexity (*Pi*)= $2^{H(Pi)}$ (5-50) with *Pi* is a probability distribution over all of the other data points explained by variance *σi,* learning rate ε (5), number of steps (e.g. 5000).

Distances between clusters might not mean anything



van der Maaten and Hinton. Journal of Machine Learning Research 2008

Uniform Manifold Approximation and Projection for Dimension Reduction (UMAP)

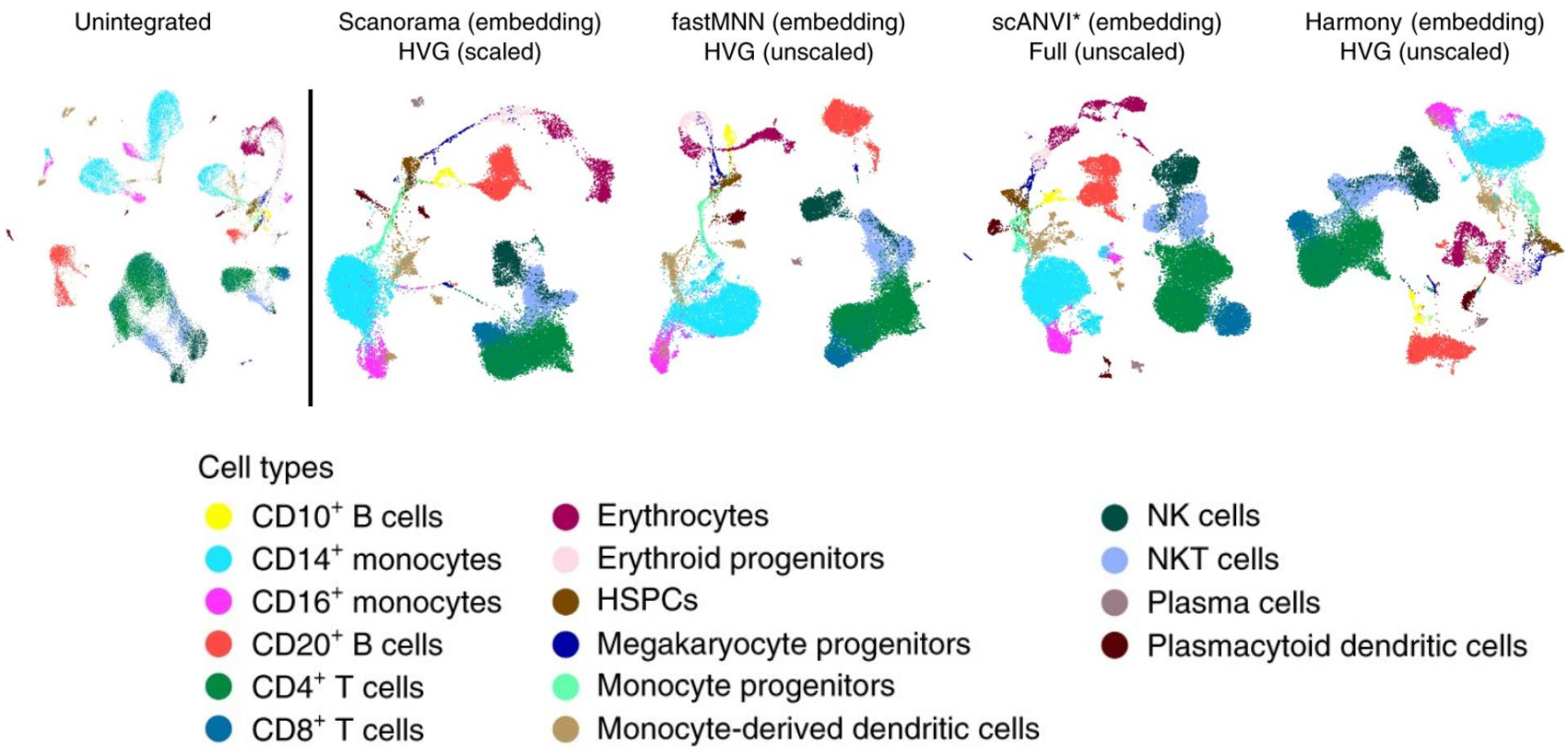


Construction of a weighted k-neighbor graph.

In practice UMAP uses a force directed graph layout algorithm in low dimensional space

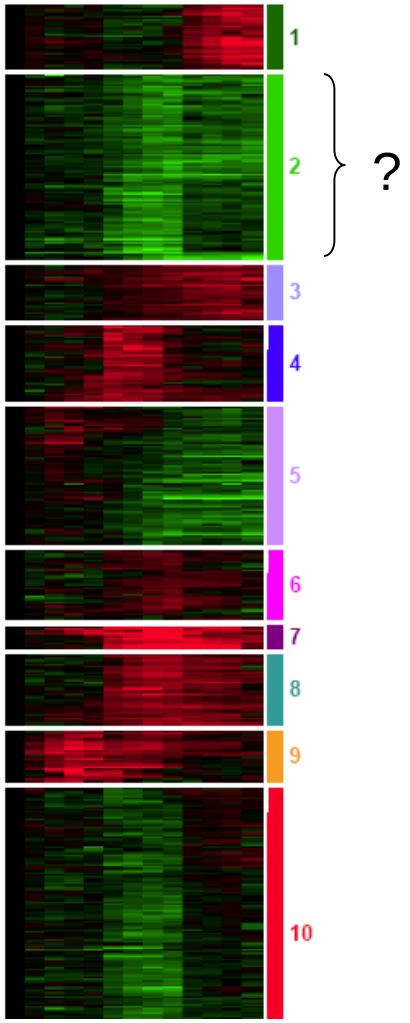
Becht E et al. Nat Biotechnol 2018

Single-cell data integration



Luecken et al. Nat Methods 2022

Biological meaning of the gene sets



- Gene ontology terms
- Pathway mapping
- Linking to Pubmed abstracts or associated MESH terms
- Regulation by the same transcription factor (module)
- Protein families and domains
- Gene set enrichment analysis
- Over representation analysis

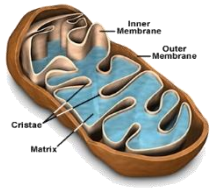
Gene Ontology (GO)

The Gene Ontology project (http://geneontology.org) provides a **controlled vocabulary** to describe gene and gene product attributes in any organism.

The three organizing principles (categories) of GO are

- cellular component

mitochondrium

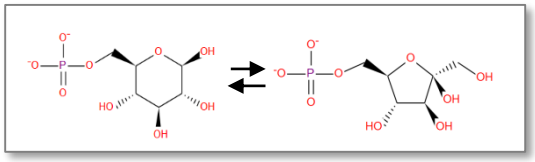


- biological process

cell cycle

- molecular function

isomerase activity



What's in a GO term?

- **Term**

  transcription initiation

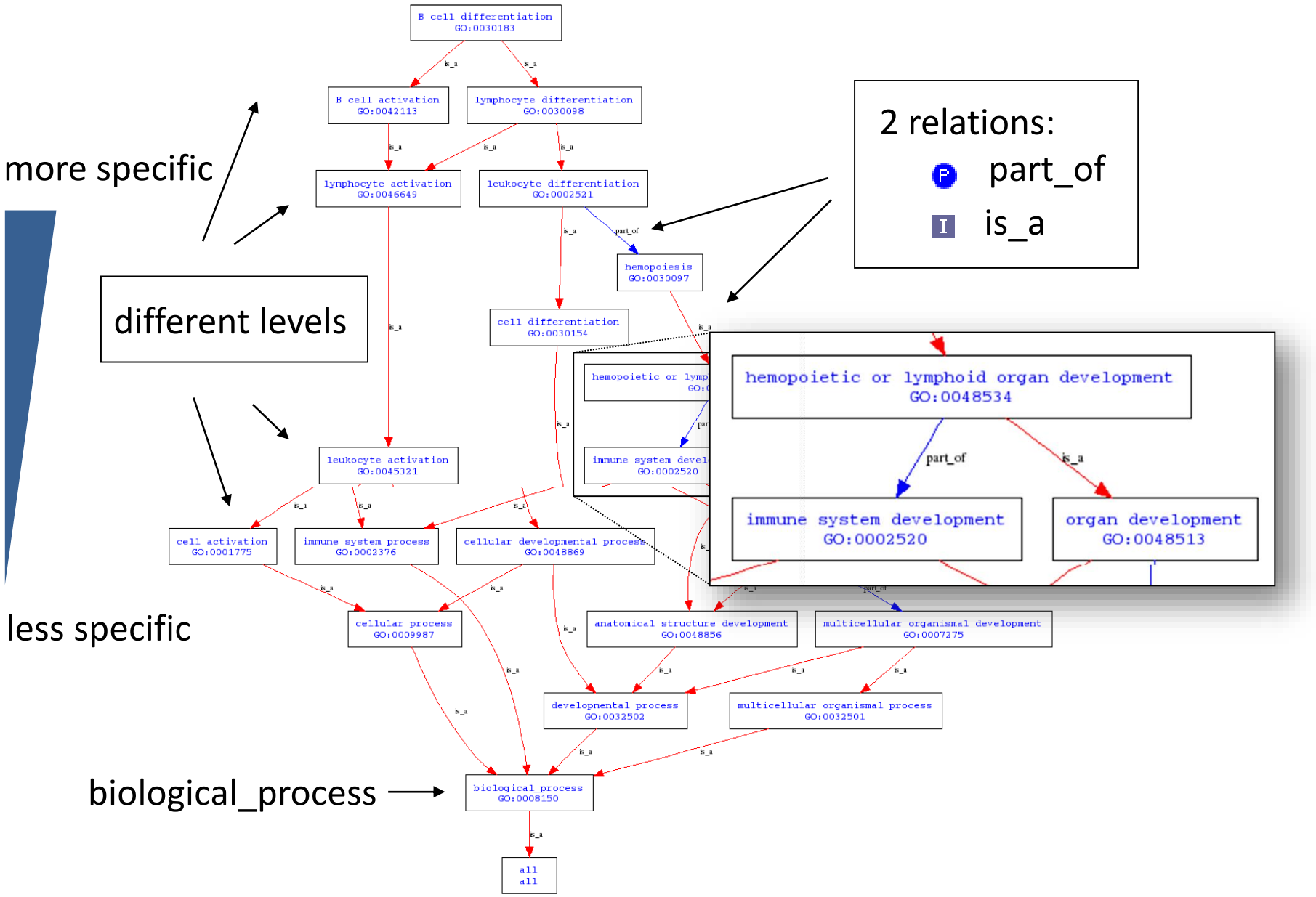
- **ID**

  GO:0006352

- **Definition**

  Processes involved in starting transcription, where transcription is the synthesis of RNA by RNA polymerases using a DNA template.

Parent /child relation in directed acyclic graph (DAG)



Evidence code for GO annotations

ISS	Inferred from Sequence Similarity
IEP	Inferred from Expression Pattern
IMP	Inferred from Mutant Phenotype
IGI	Inferred from Genetic Interaction
IPI	Inferred from Physical Interaction
IDA	Inferred from Direct Assay
RCA	Inferred from Reviewed Computational Analysis
TAS	Traceable Author Statement
NAS	Non-traceable Author Statement
IC	Inferred by Curator
ND	No biological Data available

Pathways

Definition:

A **biological pathway** is a series of actions among molecules in a cell that leads to a **certain product** or a **change in a cell**. Such a pathway can trigger the assembly of new molecules, such as a fat or protein. Pathways can also turn genes on and off, or spur a cell to move (genome.gov/27530687).

Types of biological pathways:

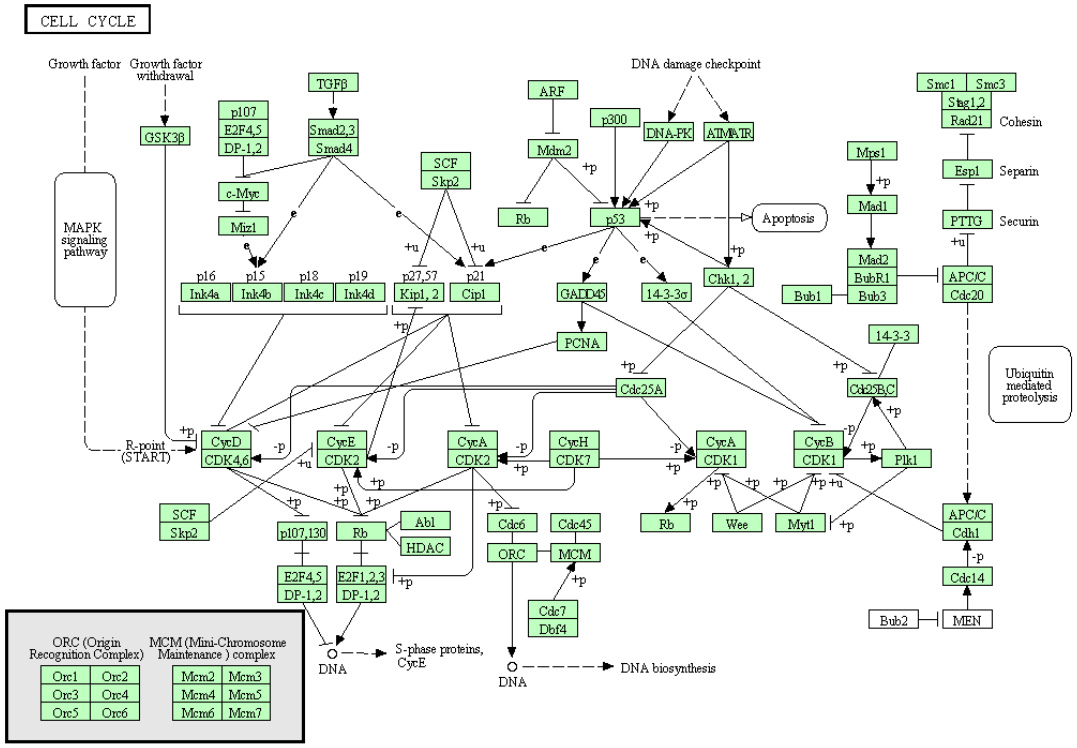
- metabolic pathways
- signaling pathways
- gene regulation pathways

Canonical Pathways:

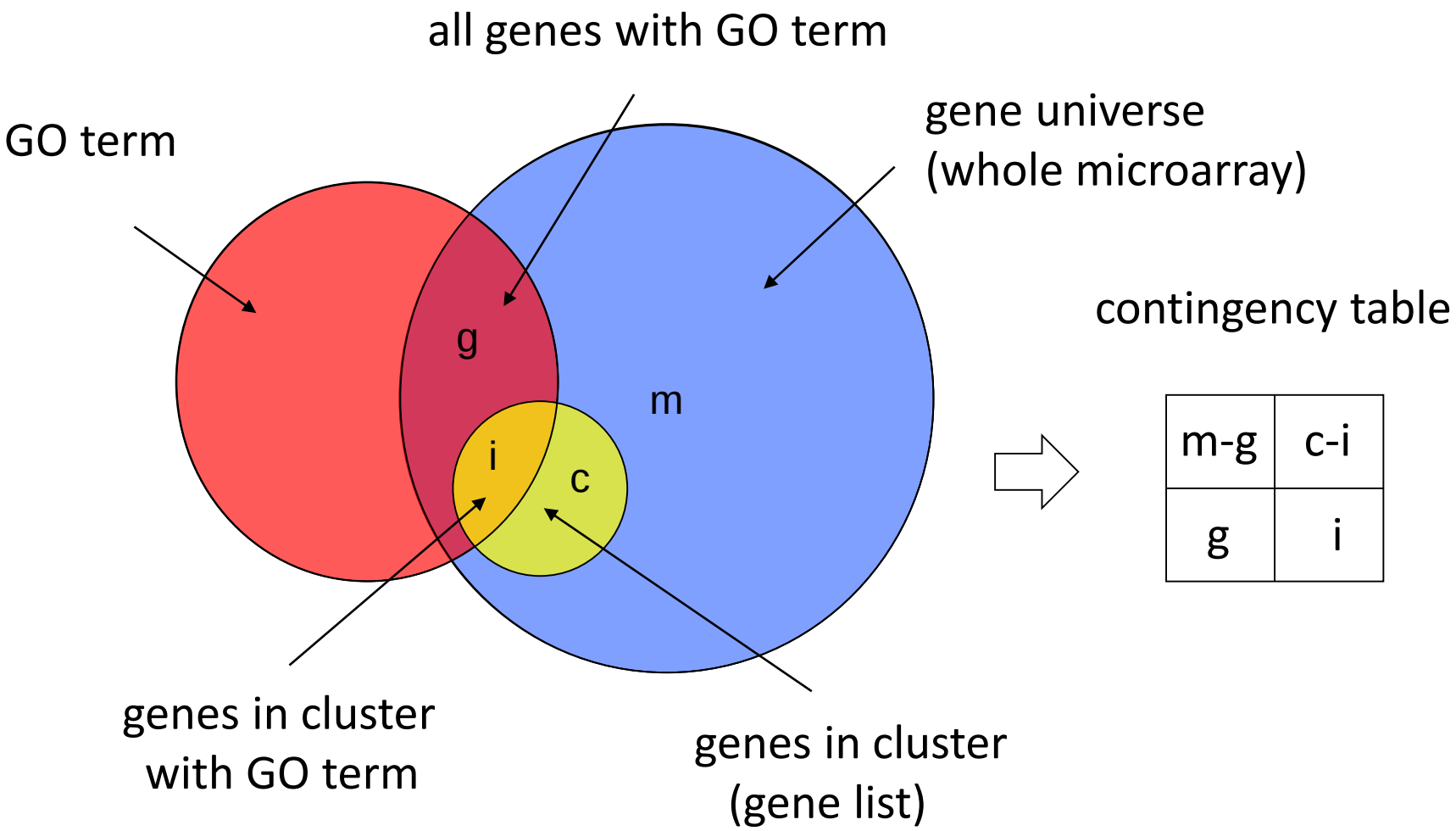
Idealized or generalized pathways that represent common properties of a particular signaling module or pathway

Pathways

- Kyoto Encylopedia of Genes and Genomes (KEGG)
- Reactome
- Wiki Pathways
- BioCyc
- Biocarta
- PANTHER



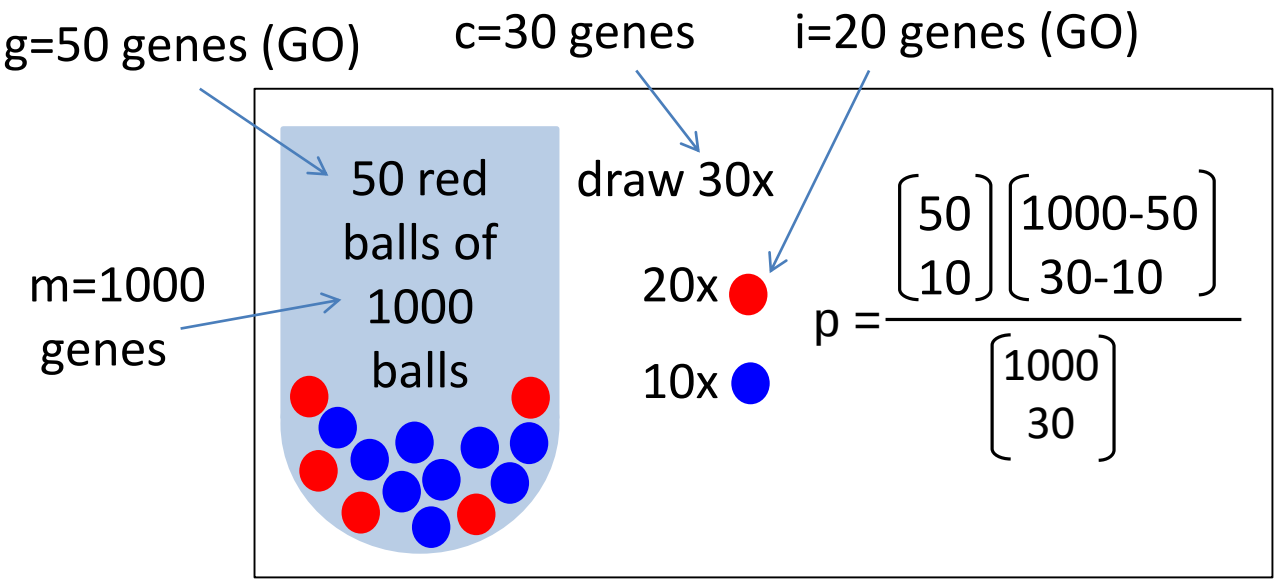
Over representation analysis



Over representation analysis

- Fisher exact test for contingency table
- Hypergeometric distribution

m-g	c-i
g	i



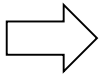
- Multiple hypothesis testing => adjust p-value
- Not only for GO Terms also for TFBS, pathways,..

DAVID

- Database for Annotation, Visualization and Integrated Discovery
- https://david.ncifcrf.gov
- Functional annotation tool (over representation analysis)

1019 mouse gene symbols

Dnajb1
Wnt11
Sorbs3
D230025D16Rik
Sfxn3
Hspa5
Golga3
Hgs
Npc1
Mta2
Cnn2
Spg20
Zpr1
…



DAVID Bioinformatics Resources 6.7
National Institute of Allergy and Infectious Diseases (NIAID), NIH

Functional Annotation Chart

Help and Manual

Current Gene List: List_1
Current Background: Mus musculus
962 DAVID IDs
⊞ **Options**

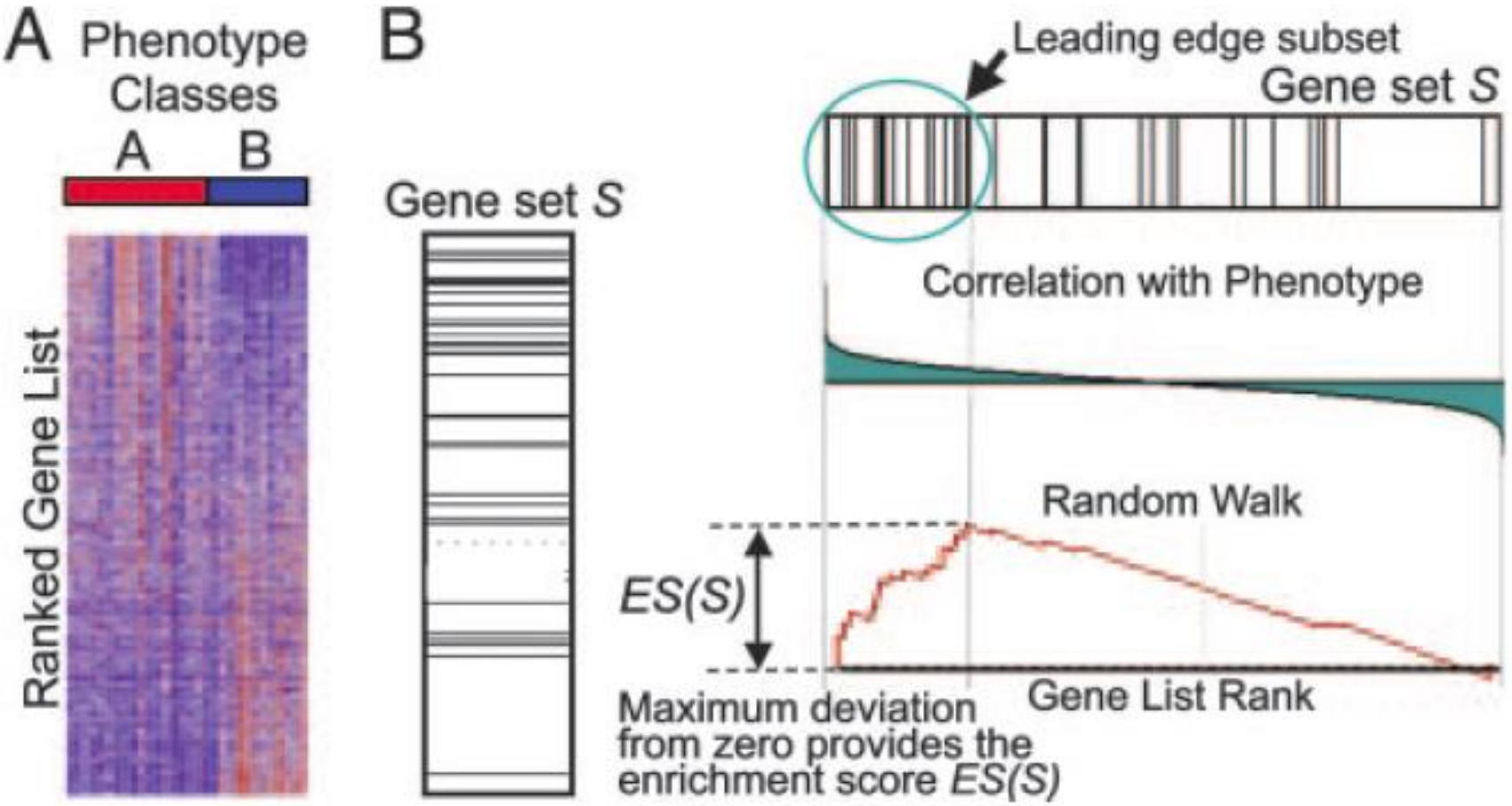
Rerun Using Options | Create Sublist

363 chart records — Download File

Sublist	Category	Term	RT	Genes	Count	%	P-Value	Benjamini
	GOTERM_BP_ALL	cellular process	RT		597	62,1	1,0E-18	2,7E-15
	GOTERM_BP_ALL	cellular metabolic process	RT		407	42,3	2,5E-13	3,4E-10
	GOTERM_BP_ALL	regulation of cellular metabolic process	RT		227	23,6	1,1E-12	1,0E-9
	GOTERM_BP_ALL	regulation of metabolic process	RT		236	24,5	1,7E-12	1,1E-9
	GOTERM_BP_ALL	regulation of gene expression	RT		202	21,0	6,1E-12	3,3E-9
	GOTERM_BP_ALL	regulation of macromolecule biosynthetic process	RT		198	20,6	1,1E-11	4,9E-9
	GOTERM_BP_ALL	regulation of cellular biosynthetic process	RT		203	21,1	1,4E-11	5,4E-9
	GOTERM_BP_ALL	regulation of biosynthetic process	RT		203	21,1	2,0E-11	6,6E-9
	GOTERM_BP_ALL	regulation of macromolecule metabolic process	RT		215	22,3	4,0E-11	1,2E-8
	GOTERM_BP_ALL	cellular macromolecule metabolic process	RT		324	33,7	6,1E-11	1,6E-8
	GOTERM_BP_ALL	regulation of primary metabolic process	RT		212	22,0	1,3E-10	3,2E-8
	GOTERM_BP_ALL	regulation of transcription	RT		183	19,0	2,6E-10	5,8E-8
	GOTERM_BP_ALL	positive regulation of cellular biosynthetic process	RT		67	7,0	3,4E-10	7,1E-8

…

Gene set enrichment analysis



Subramanian A et al. Proc Natl Acad Sci (2005)

Gene set enrichment analysis

1. Given an *a priori* defined set of genes.

2. Rank genes (e.g. by t-value between 2 groups of microarray samples) → ranked gene list ***L***.

3. Calculation of an enrichment score (*ES*) that reflects the degree to which a gene set ***S*** is overrepresented at the extremes (top or bottom) of the entire ranked list ***L***.

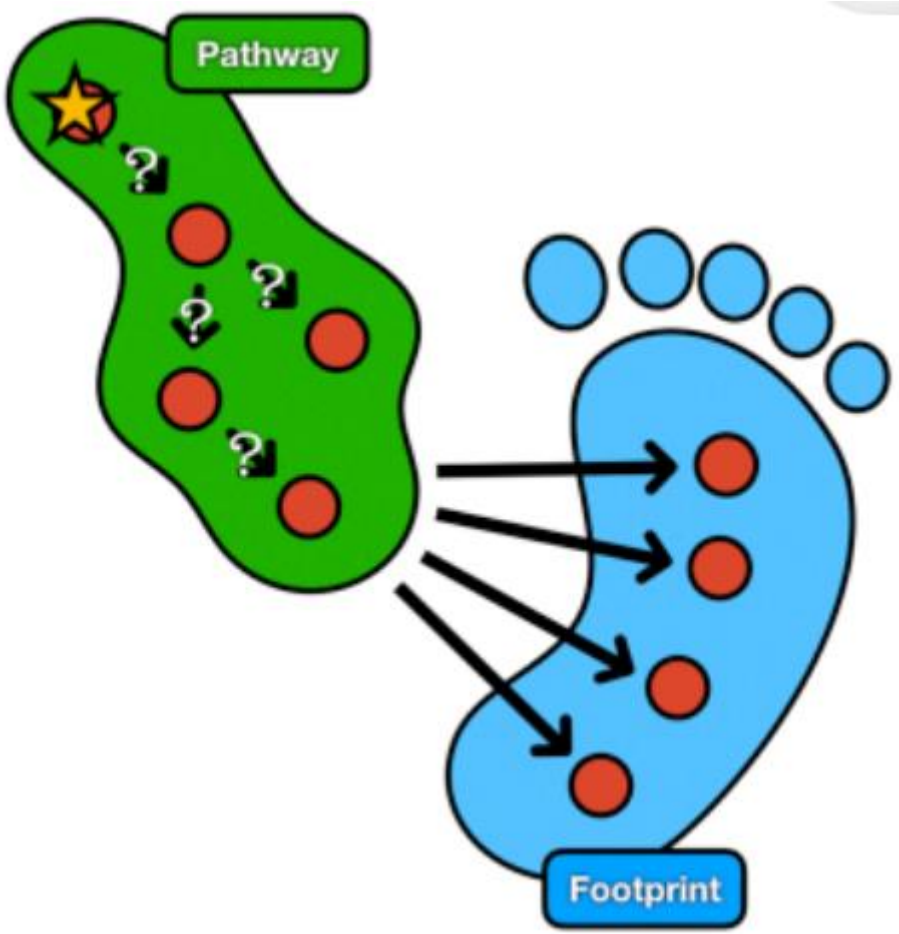
4. Estimation the statistical significance (nominal *P* value) of the *ES* by using an empirical phenotype-based permutation test procedure.

5. Adjustment for multiple hypothesis testing by controlling the false discovery rate (FDR).

http://www.broadinstitute.org/gsea
http://www.broadinstitute.org/cancer/software/gsea/wiki/

Footprint methods to infer functional activity

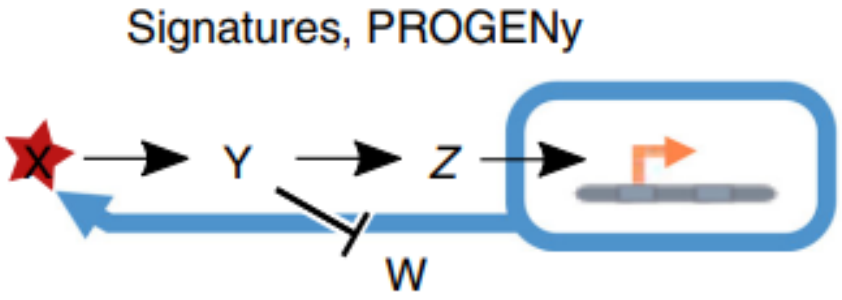
Signalling
Metabolism
Gene expression

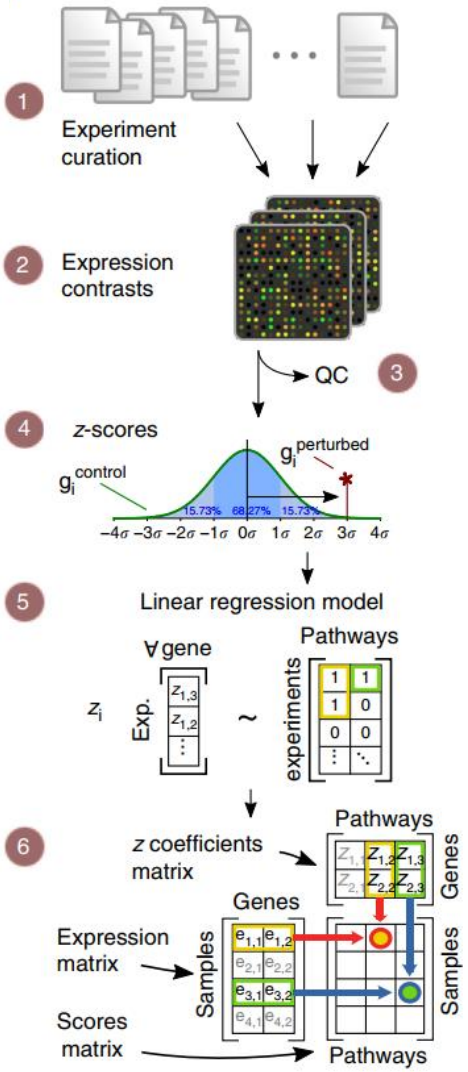


Dugourd and Saez-Rodriguez, Curr Opin Syst Biol, 2019

Footprint methods to infer functional activity

Compendium of perturbation experiments of signaling pathways to infer pathway activity from gene expression readout (14 pathways)

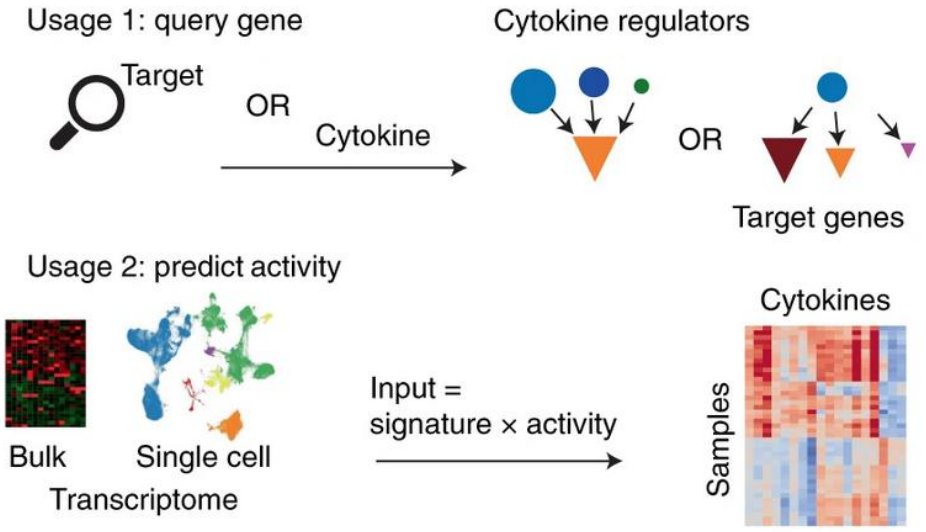




Schubert et al. Nat Commun 2018

CytoSig

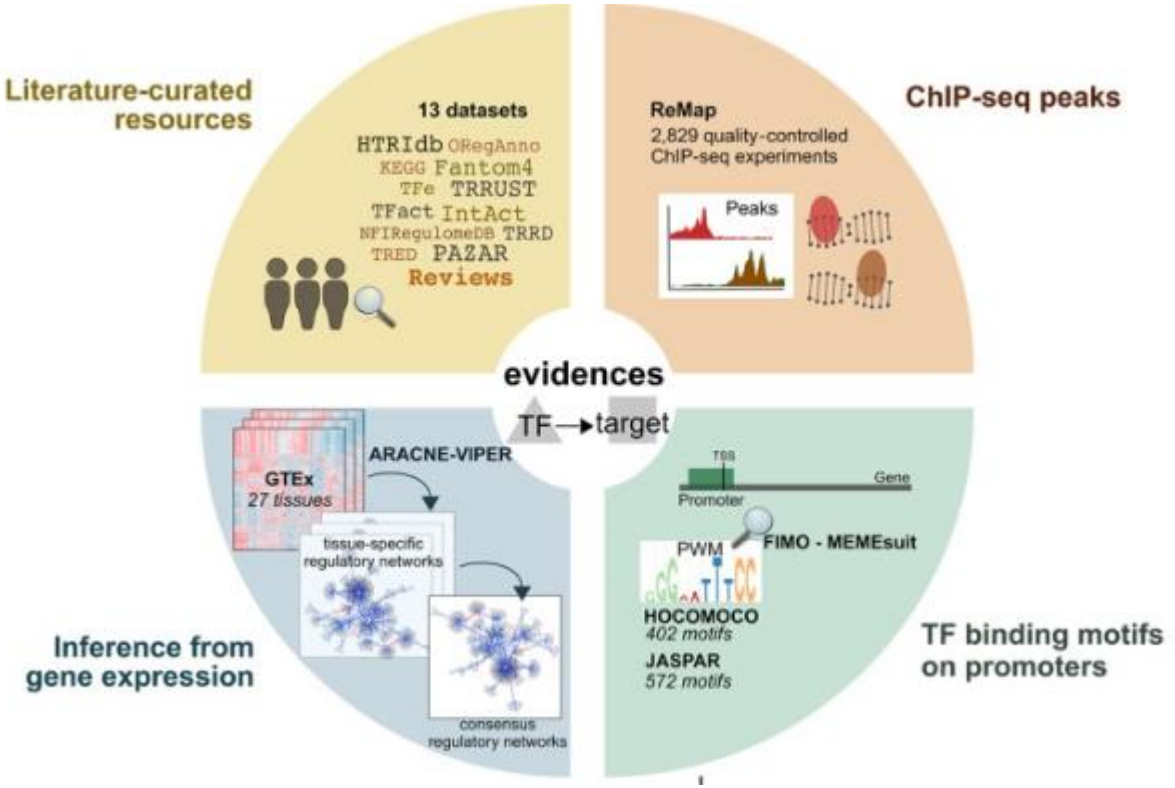
- Collection of >20k transcriptome profiles for human cytokine, chemokine and growth factor responses.
- Prediction of signaling activities in distinct cell populations in infectious diseases, chronic inflammation and cancer using bulk and single-cell transcriptomic data.



Jiang et al. Nat Methods 2021

DoRothEA

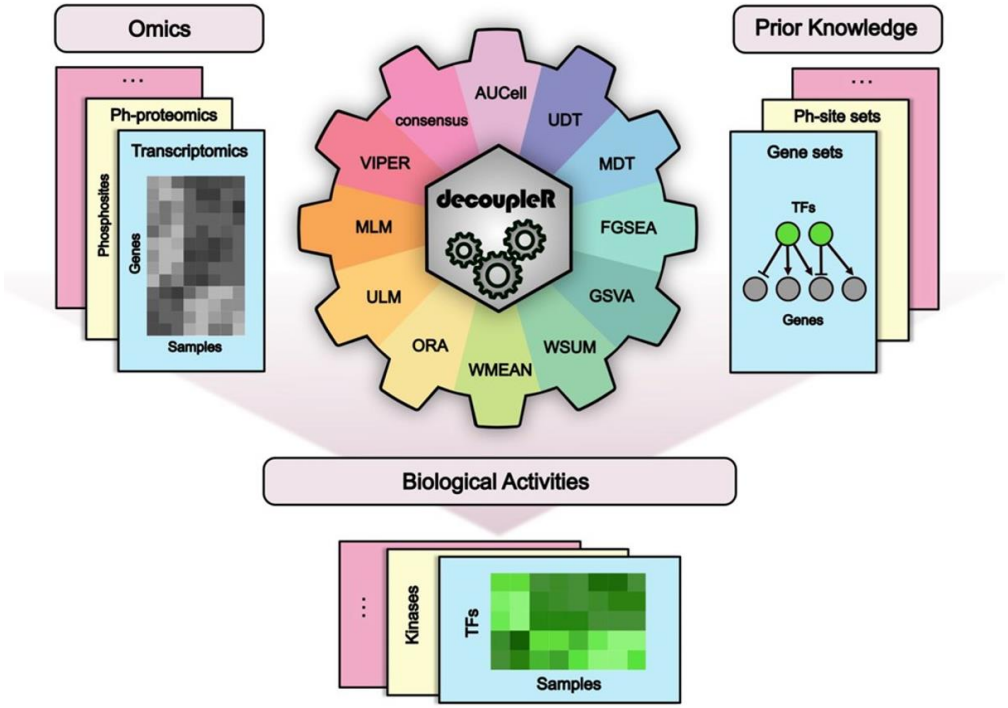
– The prediction of transcription factor (TF) activities from the gene expression of their targets (i.e., TF regulon)



Garcia-Alonso *et al.* Genome Res. 2019

DecoupleR

- R interface to statistical methods to infer biological activities/extract biological signatures integrating omics data (e.g. RNA-seq) with prior knowledge.



Badia-i-Mompel et al. Bioinformatics Adv 2022

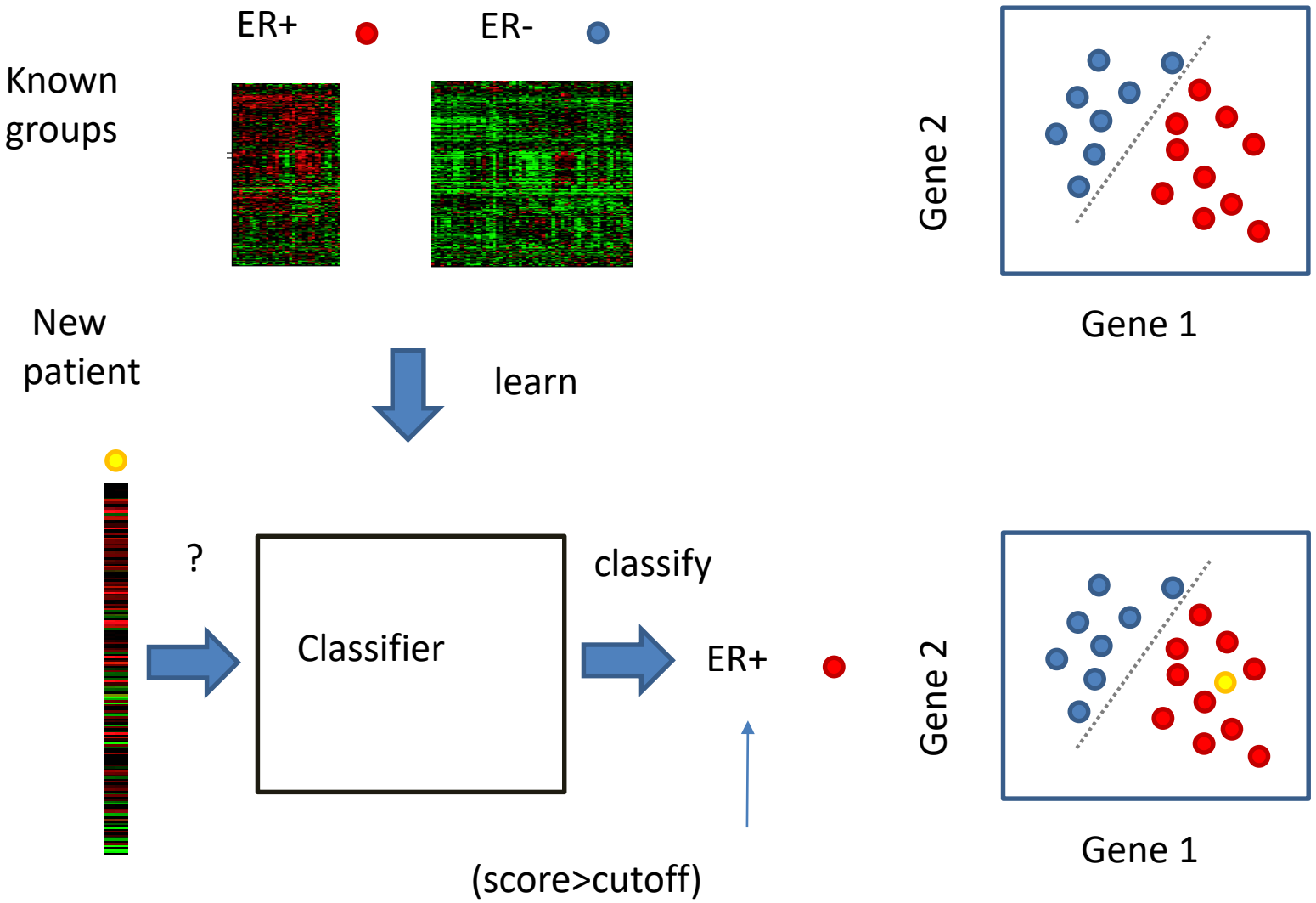
Pathway (and network) analysis

Pathway analysis
- over representation
- mapping gene expression to pathway

Interaction network (direct or indirect interactions)
- from protein-protein interaction databases (HPRD, IntAct)
- known and predicted functional association (STRING)

De novo network construction
- co-expression network
- reversed engineering

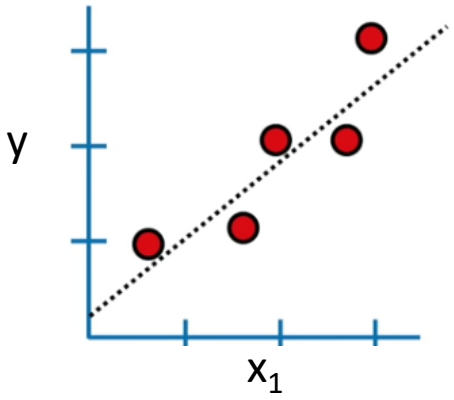
Classification



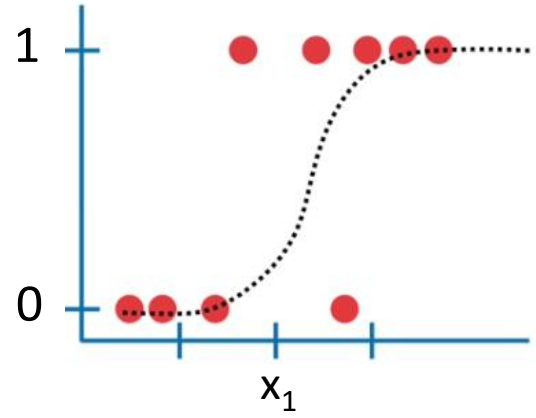
Methods for classification

- K-nearest neighbors
- Linear Models
- Discriminant analysis
- Logistic Regression
- Naïve Bayes
- Decision Trees
- Random Forests
- Support Vector Machines

Linear and logistic regression



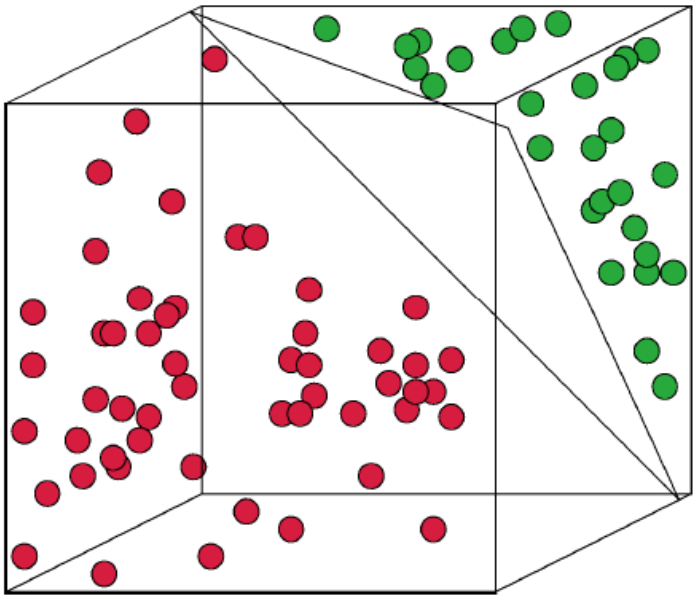
$y=b_0+b_1*x_1+b_2*x_2+\ldots$

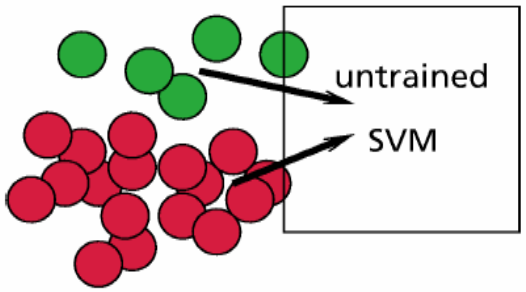


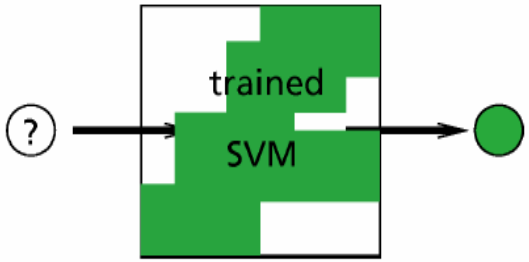
$\ln (P/(1-P))=b_0+b_1*x_1+b_2*x_2+\ldots$

$$L_{LASSO}(\beta) = \frac{1}{n}\sum_{i=1}^{n}|y_i - \boldsymbol{\beta}^\top \boldsymbol{x}_i|^2 + \lambda\sum_{j=1}^{J}|\beta_j|.$$

Support vector machines (SVM)







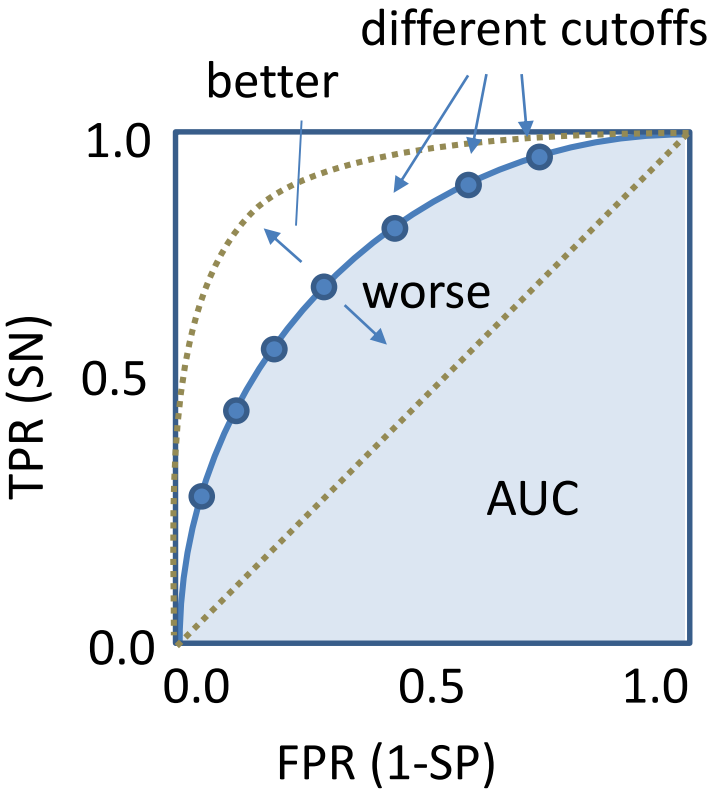
A SVM tries to find an optimal hyperplane that separates all training samples correctly and maximaizes the margins. If this is not possible in the input space (e.g. 2 dimensions) a hyperplane can be found in the higher dimensional features space (e.g. 3 dimensions).

Receiver operator characteristics (ROC)

Classified (> cutoff)	truely ER+	truely ER-
ER+	TP	FP
ER-	FN	TN

Sensitivity
SN=TP/(TP+FN)

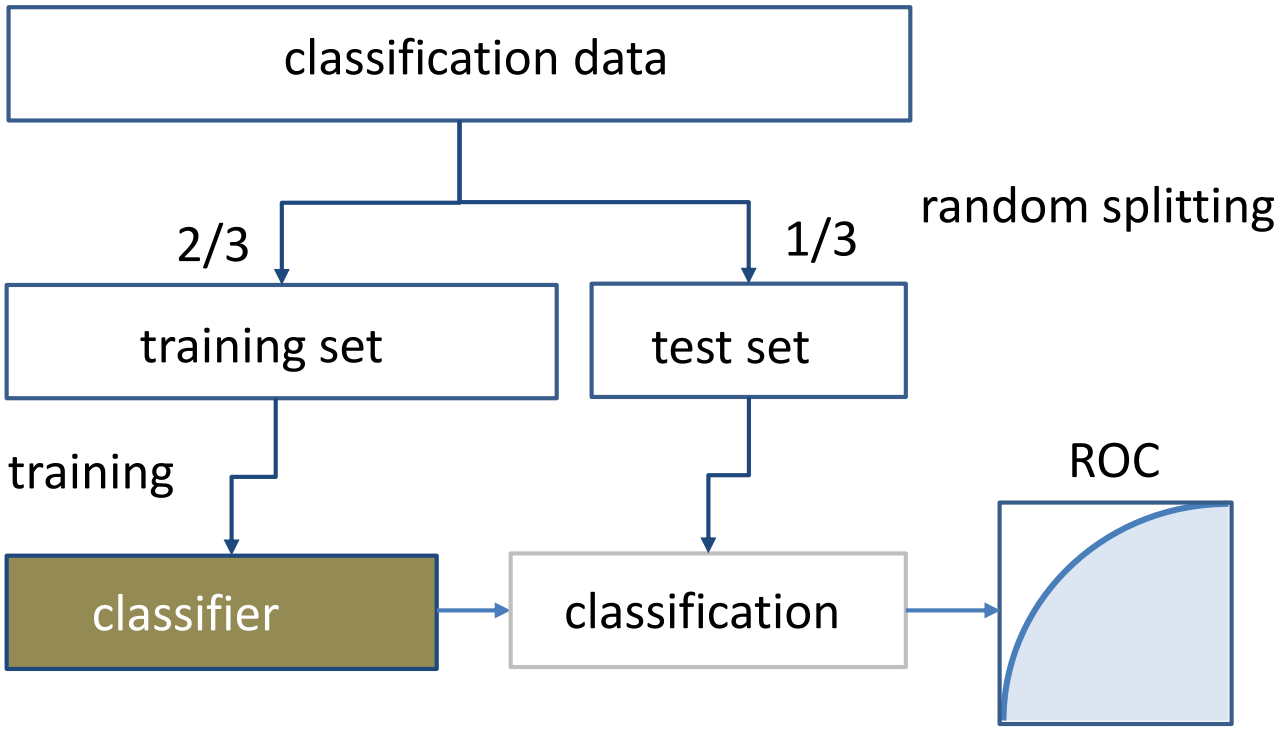
Specificty
SN=TN/(TN+FP)



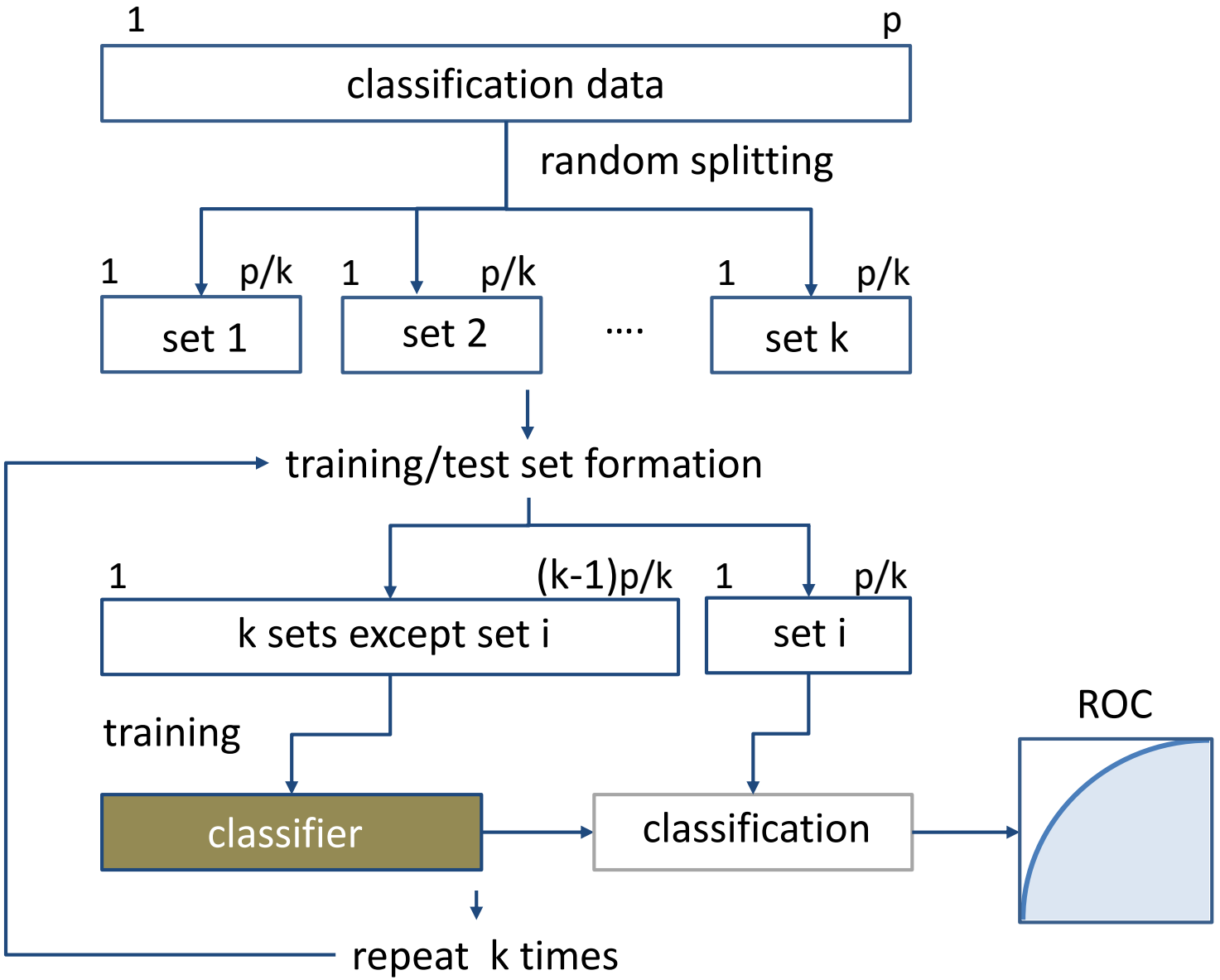
Area under curve (AUC)
AUC=1.0 optimal
AUC=0.5 random

Holdback cross validation

To avoid overfitting data should be splitted into training and test set



K-fold cross validation



Survival analysis

Survival analysis involves the modelling of time to event data, which is in the context of biostatistics time to death or other events (time to relapse, time to re-hospitalization).

In other disciplines this type of analysis is also known as reliability analysis (engineering) or duration analysis (economics).

The aim is to statistically describe survival times and compare survival times of several groups (the longer the survival times the better the therapy).

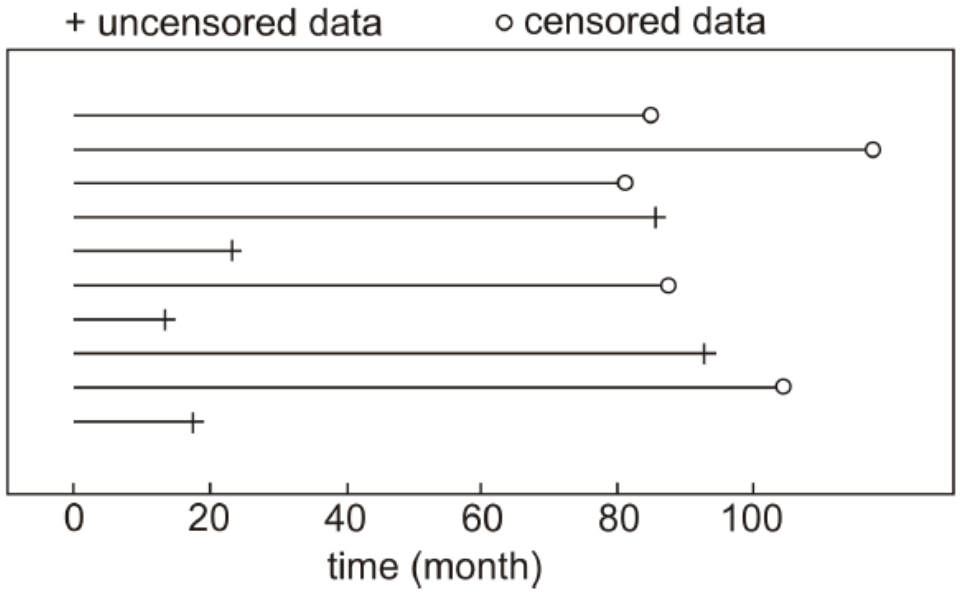
It is also sometimes important to find relations between survival times and other explaining variables (age, type of therapy, severity of disease,...).

Censored data

Censored data (incomplete follow up) arises when a study is finished before all patients died (withdrawn alive).

Another case is when patients have to be excluded from the study due to other reasons (emigration, accidental death).

In general patients are recruited to the study at different time points (e.g. time point of surgery, indicated here as time=0).

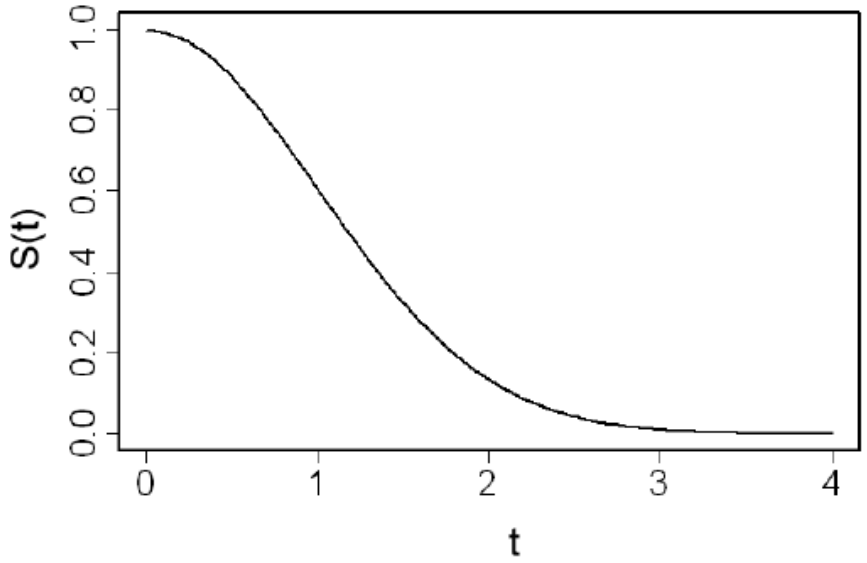


Survival function

If X is a continuous random variable with a cumulative distribution function $F(t)$ of survival times the survival function is defined as:

$$S(t) = P(T > t) = \int_t^{\infty} f(u)du = 1 - F(t)$$

The survival function $S(t)$ shows the proportion of patients (probability), which survived a specified time interval t.

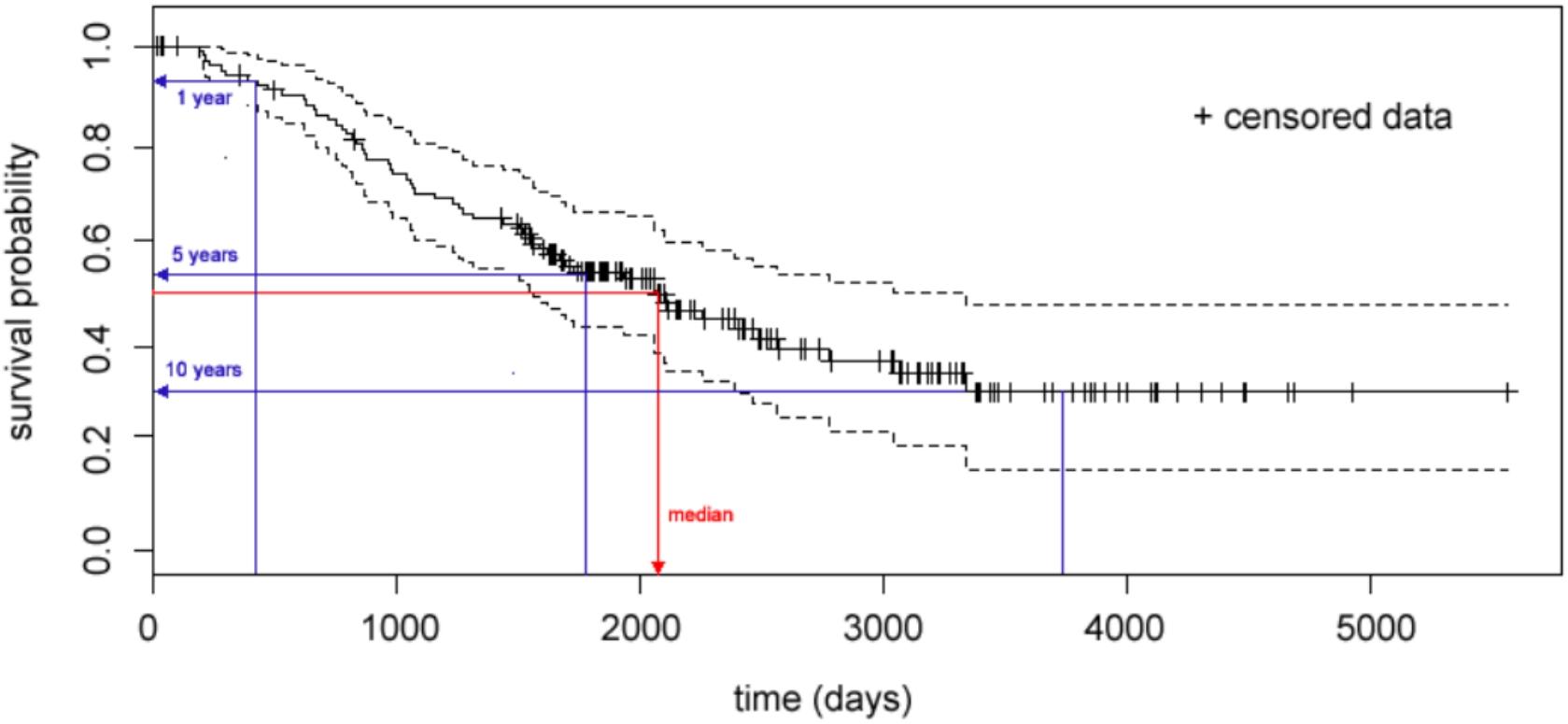


The survival function follows often a Weibull $e^{-(t/\lambda)^k}$ or Exponential $(e^{-(t/\lambda)})$ distribution.

Kaplan-Meier survival curves

The survival function can be estimated by the Kaplan-Meier curves (Kaplan-Meier estimator)

Each event(death) is indicated by a step function and censored data are indicated by (+).



Kaplan-Meier estimator

The calculation of the Kaplan-Meier estimator is using the conditional probability.

The probability of surviving 100 days would than be
$p = p_1 \times p_2 \times ... \times p_{100}$.

In general:

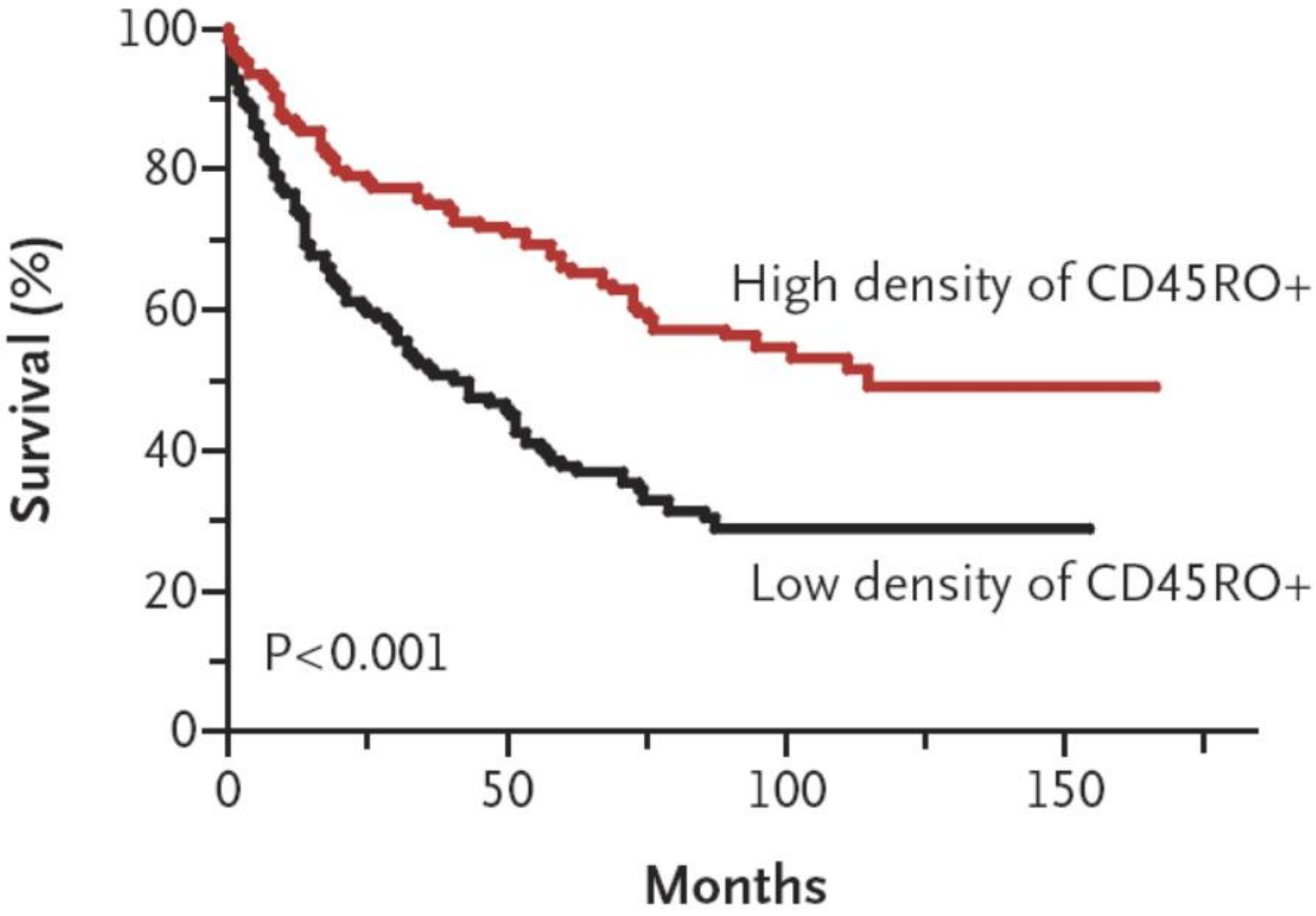
$$p_k = p_{k-1}\frac{r_k - f_k}{r_k} \Rightarrow \hat{S}(t) = \prod_{t_k \leq t}(1 - \frac{f_k}{r_k})$$

where r_k is the number of subjects still at risk (still being followed up) immediately before the kth day, and f_k is the number of observed events on day k.

The standard error of the survival proportion (not for small and very large sample size) can be calculated:

$SE(p_k) = p_k\sqrt{(1-p_k)/r_k}$ and 95% CI: $p_k \pm 1.96SE(p_k)$

Comparison of Kaplan-Meier curves



Log-rank test

The most common (non-parametric) method of comparing independent groups of survival times is the logrank test.

The null hypothesis here is that the groups come from the same population.

The survival times of both groups were ranked together and time intervals were defined between the survival times including the time of one (or more) event(s) as the upper limit of the intervals.

For each time interval we have a 2×2 table:

	group 1	*group 2*	*total*
events	f_1	f_2	f
no events	$r_1 - f_1$	$r_2 - f_2$	$r - f$
total	r_1	r_2	r

Log-rank test

	group 1	*group 2*	*total*
events	f_1	f_2	f
no events	$r_1 - f_1$	$r_2 - f_2$	$r - f$
total	r_1	r_2	r

For the number of observed and expected events we get

$O_i = \sum_{j=1}^{k} f_i^j$ and $E_i = \sum_{j=1}^{k} r_i^j f^j / r^j$ for k time intervals,

and the test statistic $X^2 = \sum_{i=1}^{m} \frac{(O_i - E_i)^2}{E_i}$ for m groups.

Under the null hypothesis the statistic X^2 has a χ^2 distribution with $df = m - 1$.

Hazard ratio (HR)

The logrank test is solely a hypothesis test, comparing survival in two or more groups.

Relative survival in two groups can be measured by comparing the observed number of events with the expected numbers.

The hazard ratio is defined as

$$R = \frac{O_1/E_1}{O_2/E_2}$$

and gives an estimate of relative event rates in the two groups.

$K = \frac{O_1 - E_1}{V}$ is an estimate of the log hazard ratio ($\ln R$).

$SE \approx \frac{1}{\sqrt{V}}$ and $95\% CI : K \pm 1.96/\sqrt{V}$.

Relative risk and proportional hazard model

When a population is divided into 2 subpopulations exposed (E) and non-exposed ($\bar{E}$) by presence or absence of a certain characteristic (an exposure such as smoking), each subpopulation corresponds to a hazard function and the relative risk can be assigned to

$$RR = \frac{h(t, E)}{h(t, \bar{E})}$$

If $RR(t) = c$ we have a proportional hazards model:

$$h(t, E) = c \times h(t, \bar{E})$$

Cox regression

Since we have a multiplicative model (exposure raises the risk by a multiplicative constant) it can be also expressed as

$h(t) = h_0(t)e^{\beta x}$ with $h_0(t) = h(t, \bar{E})$ and
the covariate $x = 1$ for exposed and $x = 0$ for unexposed population.

The Cox regression model is considering several independent variables of interest $(X_1, ..., X_p)$:

$$h(t) = h_0(t)e^{\beta_1 x_1 + ... \beta_p x_p}$$

Adding all the hazards up to time t to get the risk of dying between time 0 and time t gives the cumulative hazard

$$H(t) = H_0(t)e^{\beta_1 x_1 + ... \beta_p x_p}$$

Cox regression

The survival probability can be estimated for any individual with specific values of the variables in the model

$$S(t) = e^{-H(t)}$$

A positive sign of the regression coefficient means that the hazard is higher and thus the prognosis worse for subjects with higher values of this variable.
Interpretation of an individual regression coefficient for two different values of the covariate x by the hazard ratio:

$$\frac{h_1(t)}{h_2(t)} = \frac{h_0(t)e^{\beta x_1}}{h_0(t)e^{\beta x_2}} = e^{\beta x_1 - \beta x_2} = e^{b(x_1 - x_2)}$$

In the special case of a binary variable the hazard ratio is e^{β}.

A prognostic index can be defined as previously:

$$PI = \beta_1 x_1 + \ldots \beta_p x_p \Rightarrow S(t) = e^{-H_0(t)e^{PI}} = S_0(t)^{e^{PI}}$$