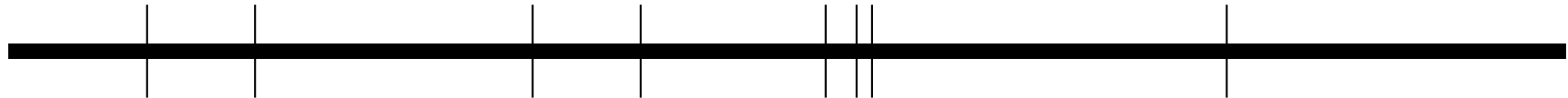


Three types of maps

Genetic map (cM or mu)



topology preserved

Metric:
Mb/mu

Physical map (Mb)



DNA sequence (bp)

actggtcaaacgtacggatacaggtaattgaccacattatatagaccacgattcgatcgcatcac

Sequencing Strategies

Whole genome “shotgun” and assembly

Shotgun sequencing

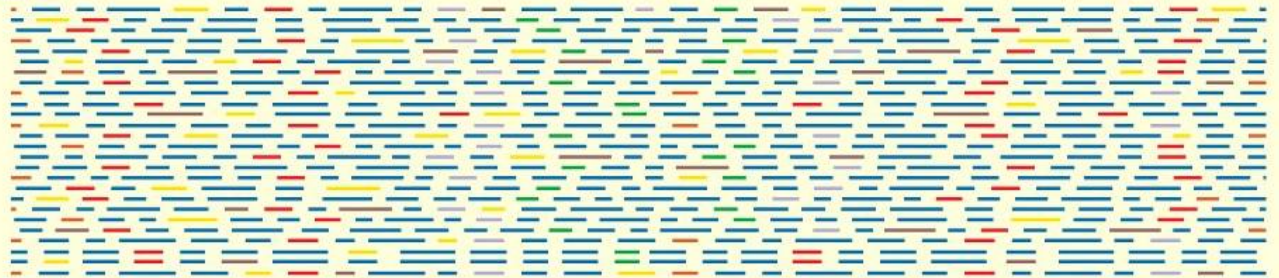
A genome

Genome

1 Cut many genome copies into random fragments.

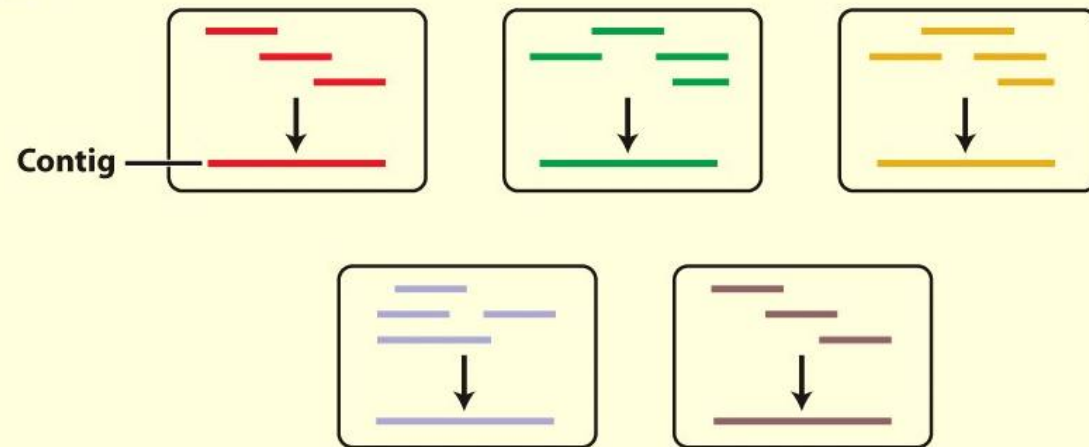


2 Make library of cloned fragments.



3 Sequence each clone.

4 Overlap sequence reads.

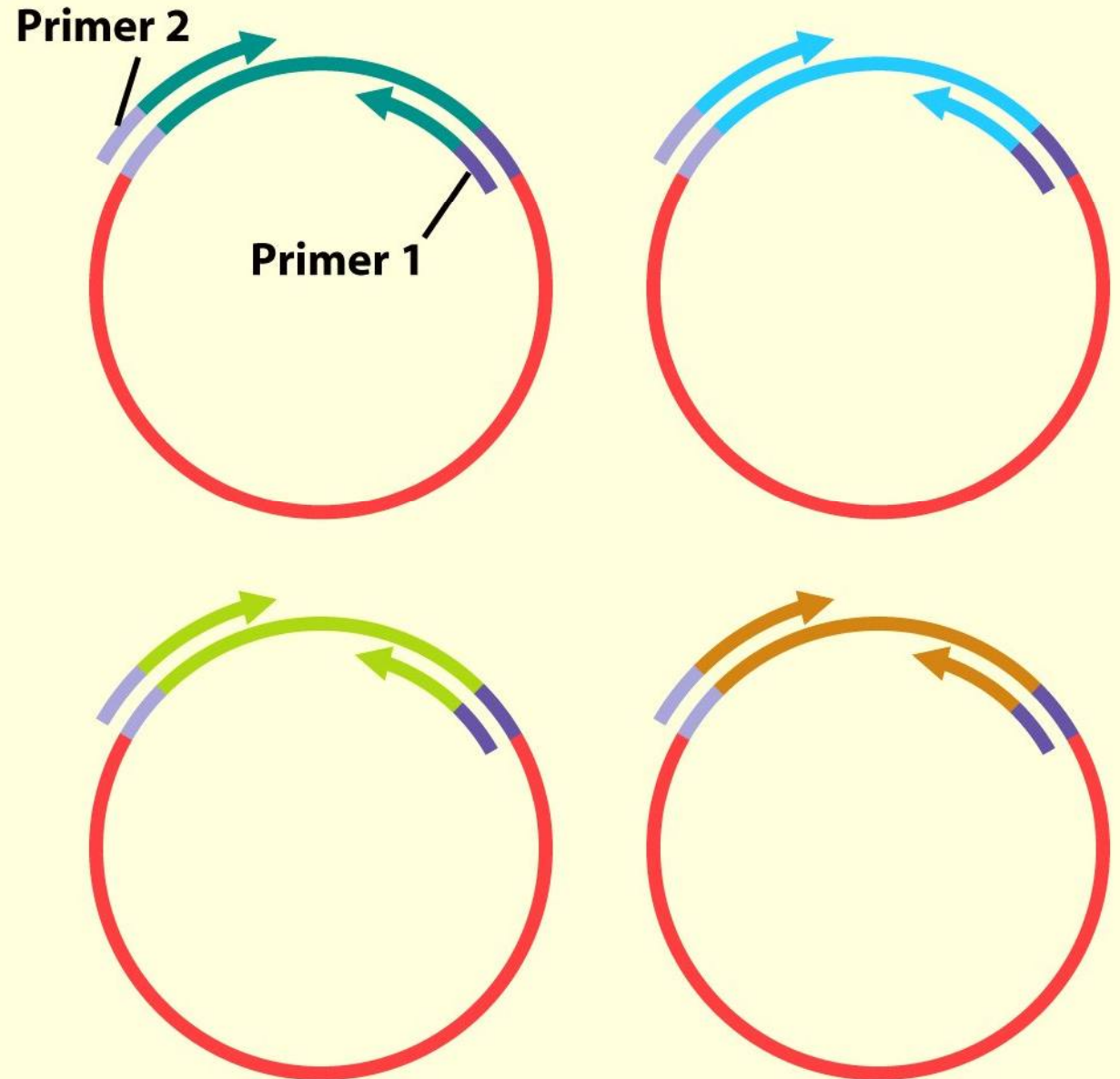


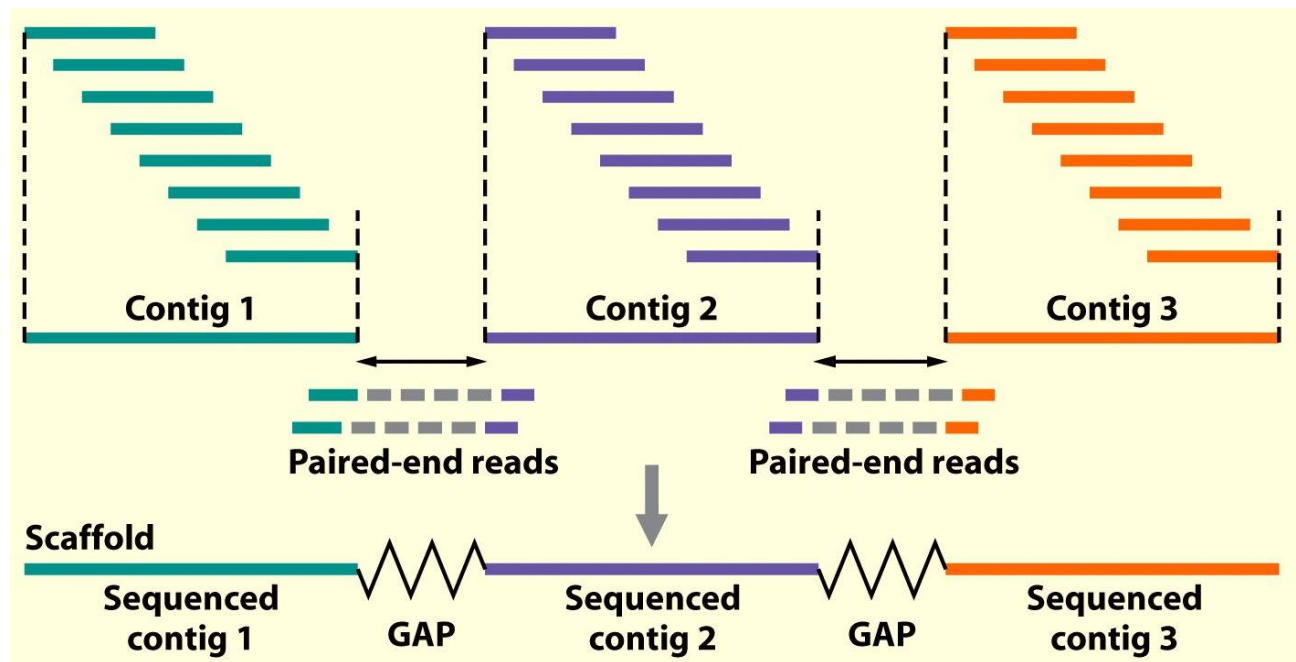
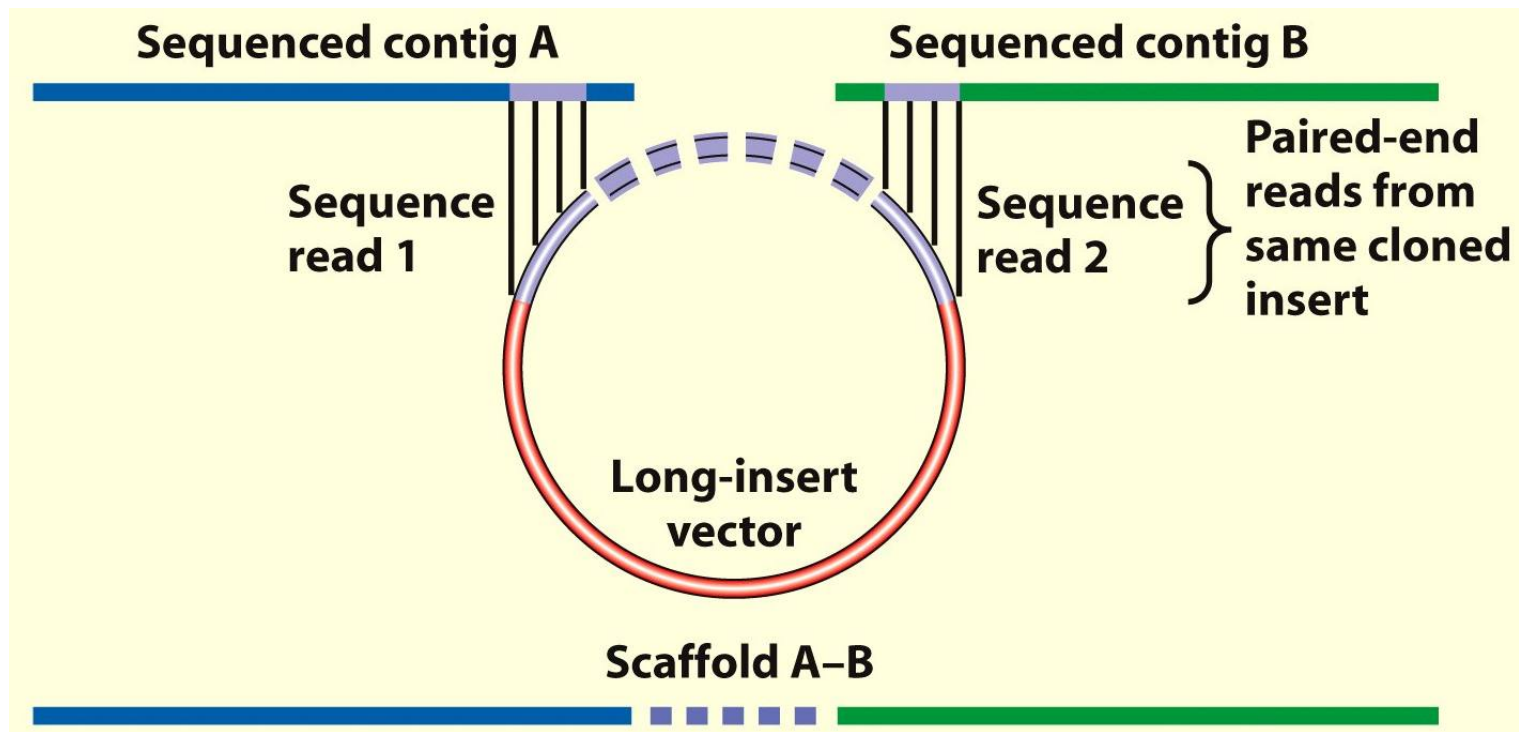
5 Overlap contigs for complete sequence.



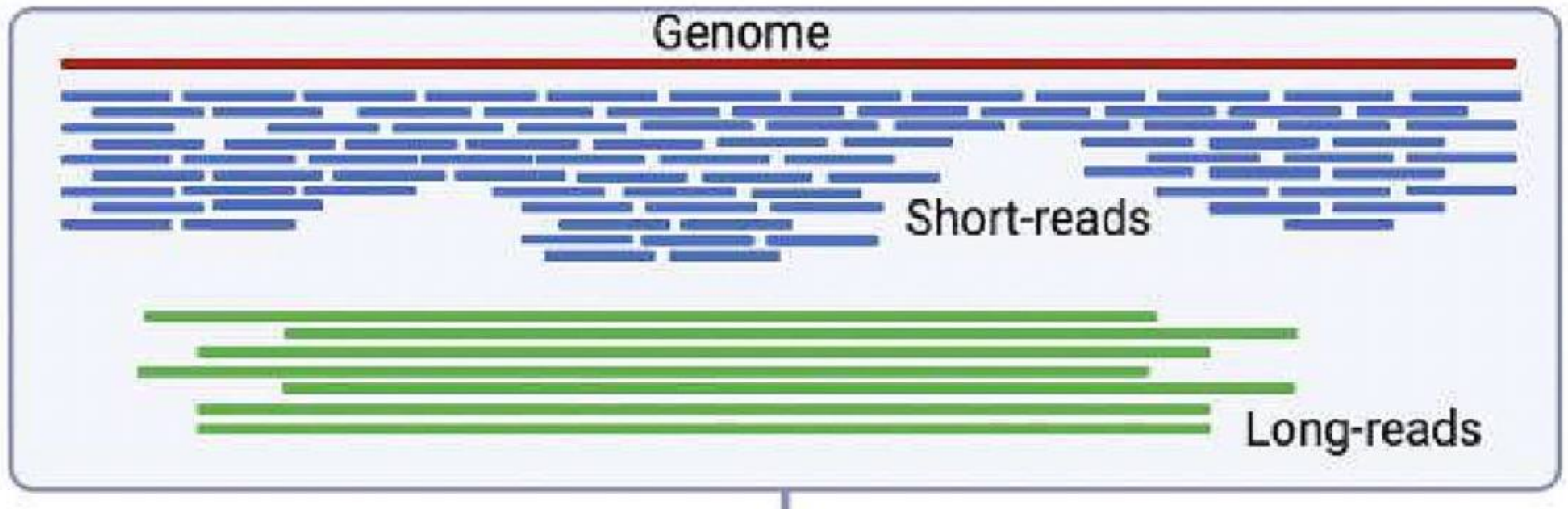
End reads from multiple inserts may be overlapped to produce a contig

Shear the DNA and then clone it into vectors that have primer sites that read into the insert





De Novo Genome Assembly



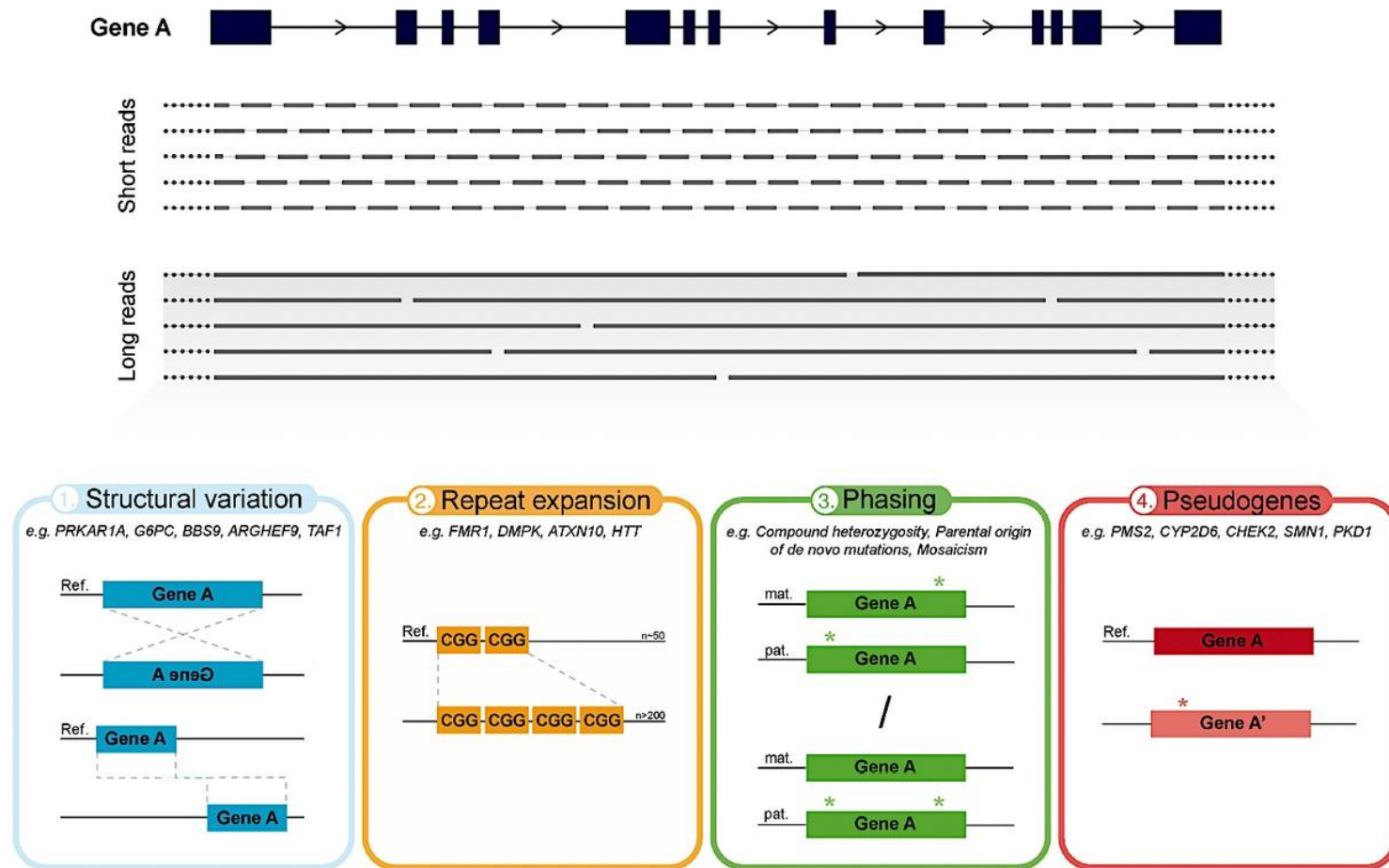


FIGURE 1 | Overview of the main advantages of current long-read sequencing (LRS) approaches in medical genetics. The predominant difference between LRS and the conventional SR-NGS approaches is the significant increase in read length. In contrast to short reads (150–300 bp), LRS has the capacity to sequence on average over 10 kb in one single read, thereby requiring less reads to cover the same gene (illustrated in **top panel**). Hence, aside from reducing alignment and mapping errors, LRS holds various advantages over short-read approaches which can greatly impact medical genetics (**bottom panel**). (1) Improved detection and characterization of large structural variation (SV), due to, e.g., large inversions or translocations. (2) Capacity to directly, and with that more accurately, sequence over long tandem repeat expansions and extreme GC-rich regions. (3) Enhanced phasing, i.e., assignment of genetic variants to the homologous paternal or maternal chromosomes, to determine inheritance patterns, parental origin of *de novo* events, mosaicism, allele specific expression and disease risk haplotypes. (4) Improved discrimination of clinically relevant genes from their pseudogenes.

Long-read sequencing allows you to recover transcript isoforms without a reference genome to act as a scaffold.

It also allows you to see which exons are linked to which other exons in the same transcript

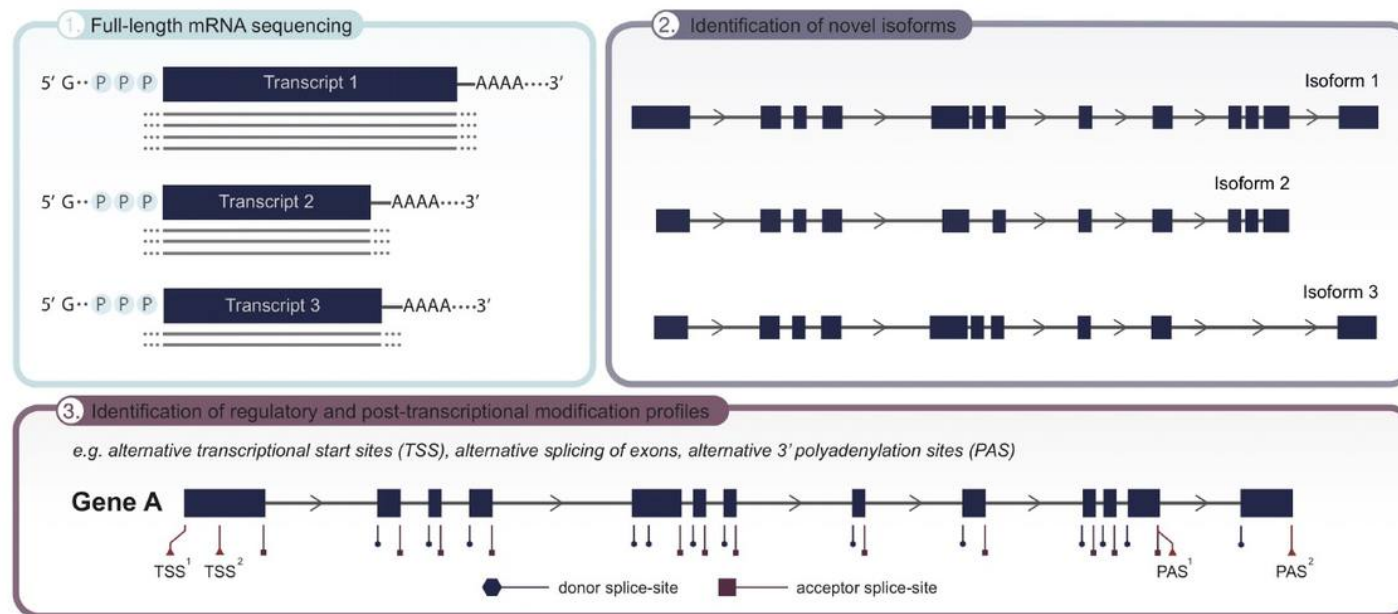


FIGURE 2 | Applicability of long-read sequencing (LRS) to unveil the transcriptome landscape of cells and tissues. Given the significant improvements in read length, employing LRS on RNA level now allows for full-length isoform sequencing, covering the complete mRNA transcript in one single read (**panel 1**). As recent advances have demonstrated the isoform landscape to be more complex than initially thought, LRS holds the potential to identify novel isoforms (**panel 2**), as well as detect transcriptional and post-transcriptional modification sites, e.g., alternative transcriptional start sites (TSS), alternative splicing of exons and alternative transcription termination sites (3'polyadenylation sites; PAS), that underpin the emergence of different isoforms (**panel 3**). Collectively, uncovering the full isoform diversity within cells and tissues.

A more complete reference

The new human genome assembly, T2T-CHM13 from the Telomere-to-Telomere Consortium, includes complex and repetitive regions of chromosomes that had not been included in previous versions of the human genome assembly (GRCh38). Although the Y chromosome remains to be completed, this new reference could be annotated with regulatory regions, variants, and sequence diversity to give a fuller picture of human genomic variation.

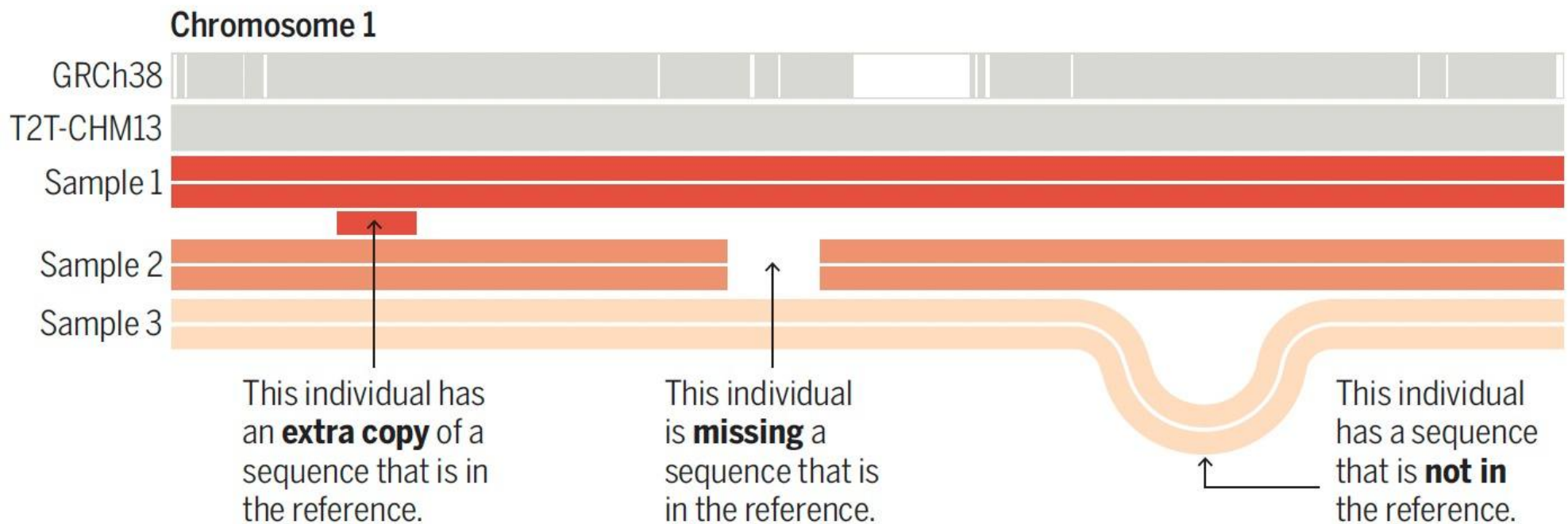


Table 1 Summary of variants in the euchromatic portion of a human genome

Variant type	Average number of sites (thousands) ^a	Average sum of variant length (Mbp) ^b	Percentage of diploid genome ^c
All	5,045.39	44.24	0.763
SNV (including MNPs)	3,992.73	3.99	0.069
Indel	1,021.73	3.63	0.063
SV ^d	30.93	36.62	0.631
STR	2.65	0.19	0.003
VNTR	12.58	1.36	0.023
Other low complexity	2.58	0.13	0.002
SD	0.55	6.25	0.108
Mobile element	6.18	1.91	0.033
LINE1	0.98	0.91	0.016
ERV	0.64	0.27	0.005
Alu	3.49	0.48	0.008
SVA	1.07	0.25	0.004
Inversion	0.15	23.2	0.400
Unclassified/mixed	6.23	3.58	0.062

Abbreviations: ERV, endogenous retrovirus; indel, insertion or deletion; LINE1, long interspersed element 1; MNP, multiple-nucleotide polymorphism; SD, segmental duplication; SINE, short interspersed element; SNV, single-nucleotide variant; STR, short tandem repeat; SV, structural variant; SVA, SINE-VNTR-Alu; VCF, Variant Call Format; VNTR, variable number tandem repeat.

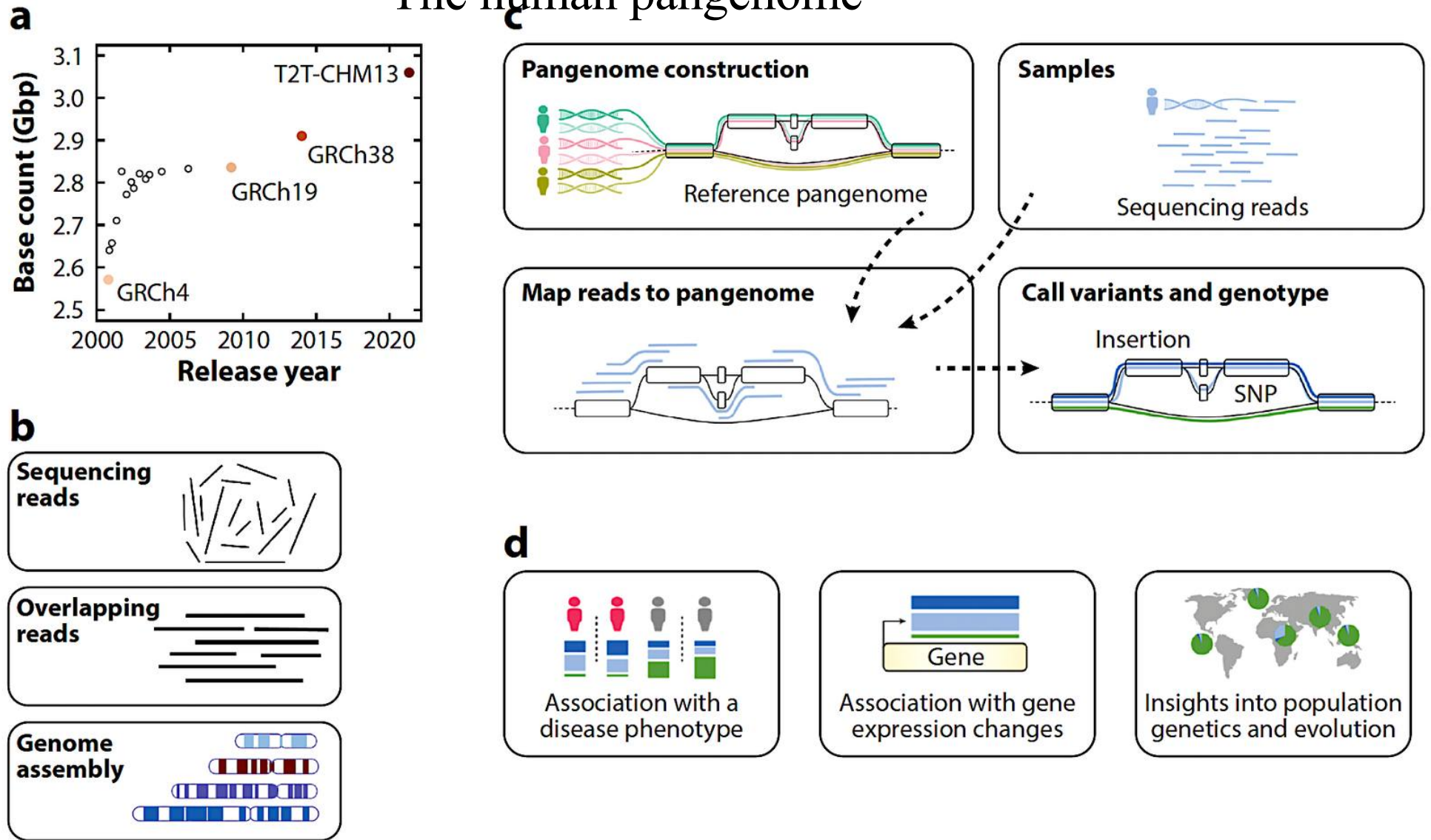
a The average number of sites observed of a given variant type within each genome.

b The average total length of variant sites.

c The percentage of a diploid genome that each variant type represents, assuming a 5.8-Gb diploid euchromatic genome length. The values exclude heterochromatin due to uncertainty around assembly and alignment for all variants except inversions.

d SVs include all structural variants; the remaining rows are SV subclasses. Unclassified/mixed denotes a class of SVs for which reliable annotation could not be given.

The human pangenome



Overview of the process of human genome sequence assembly. (a) Sequence length improvements to the human genome reference sequence over time. (b) Overview of the genome sequence assembly process. First, individual sequencing reads are generated from a sample. Then, the sequencing reads are compared with each other to identify overlaps. Overlapping reads are then merged to generate a genome sequence. (c) Overview of pangenome reference assembly and analysis. First, the pangenome is assembled from multiple individual genome sequences, revealing commonalities and differences among them. Later, sequencing reads generated from other samples can be mapped (or aligned) to the pangenome reference to detect variants and establish genotypes. (d) Applications for human genome analysis. Abbreviations: SNP, single-nucleotide polymorphism; T2T, telomere-to-telomere.

00 00 00
Days Hours Minutes
00
Seconds

FALL SALE - Get \$500 OFF Whole Genome Sequencing with lifetime membership

	Whole Genome Sequencing \$99 \$299 Good choice if you want to start exploring your family history and learn about predisposition to common traits and conditions. Add to cart	Whole Genome Sequencing \$299 \$999 The best choice for most users. It gives deep insight into ancestry and allows you to learn about common as well as rare traits and conditions. Add to cart	Deep Whole Genome Sequencing \$999 \$2999 The most advanced DNA test on the market that decodes your entire genome with ultra-high accuracy enabling most comprehensive and accurate reports. Add to cart
Amount of DNA Decoded	Decodes the most important parts of your DNA	Decodes 100% of your DNA	Decodes 100% of your DNA with ultra high accuracy
DNA sequencing depth	0.4x Generates over 2 gigabytes of DNA data	30x Generates over 100 gigabytes of DNA data	100x Generates over 300 gigabytes of DNA data
Ancestry reports	Basic ancestry reports based on autosomal DNA	Deep ancestry reports based on Y chromosomal and mitochondrial DNA	Deep ancestry reports based on Y chromosomal and mitochondrial DNA

Assembling random DNA sequence reads into sequence contigs

actgggtttaatatgacgatcgatcgtcagc

cgtcagctcgcatcgactagctacg

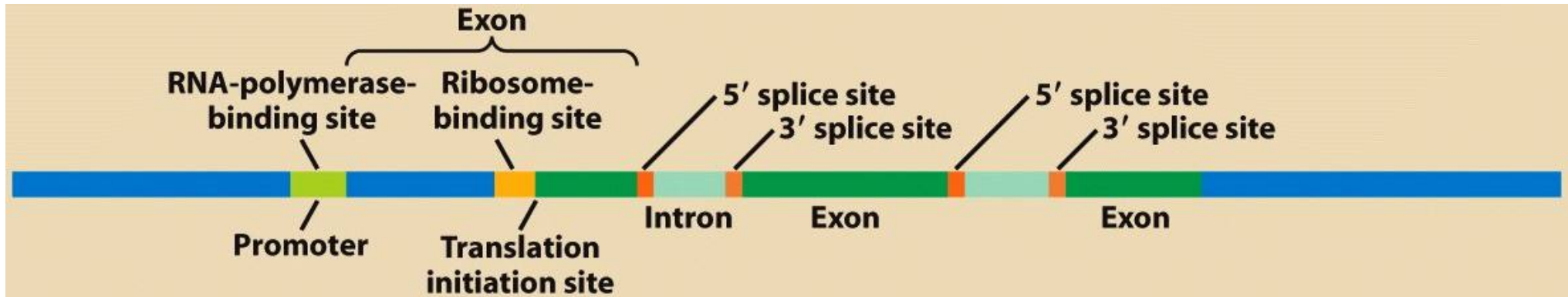
actgggtttaata

tgacgatcgatcgtcagctcgcatcgactagctacg

**With more than 30% of the genome being repeats, this can be
challenging**

actgggtttaatatgacgatcgatcgtcagctcgcatcgactagctacg

>gi|20553358|ref|NT_007422.9|Hs6_7579 H. sapiens chromosome 6 working draft seq. segment



AACCATCTGATCTTTGACAAACGTGACAAAAACAAGCAATGGAGAAAGGATTCCCTATTTAATAAATGGT
GTTGGGAAACTAGTTAGCCATATGCAGAAAACGGAAACTGGACCCCTTCCTTACACCTTATACAAAAAT
TAACTCAAGATGGATTAACGACTTAAACATAAAACATAAAAACATAAAAACCCTAGAAAGAAAACCTAGAC
AATATGATTCAGGACATAGGCATGGGAGAAGACTTTATGACTAAAACACCAAAGCAATAGCAACAAAAG
CCAAAATTGACAAATGGGATCTAATTAAGCTAAAGAGCTTCTGCACAGCAAAGAACTATCATCAGAGT
GAATAGGCAACCTACATAATGGGAGAAAATTTTTGCAATCTATCCATCTGACAAAGGGCTAATATCCAGA
ATCTACAAAGAACTTGAACAAATTTACAAGAAAAAACAACCTCCATCAAAAAGTGGGCAGTAAGAACAGA
CACTTCTCAAAAGAAGACATTTATGCAGCCAATACACATATGAAAAAAGCTCATCATCACTGGTCATTA
GAGAAATGCAAATGAAAATCACAATGAGATACCATCTCATGCCAGTTAGAATGGTGATCCTTAGGAGGTC
AGGAAACAACAGATGTTGAAGAGGATGTGGAGAAATAGGAACACTTTTACACTGTTGGTGGGAGTATCAA
TTAGTTCAACCACTGTGGAAGACAGTGTGGCAATTCCTCAAGGATCTAGAACTAGAAATACCATTTGACC
CAGCAATTCCATTGCTAGGTATATACCCAAAGGATTATAAATCATTCTACTATAAAGACACACACACAG
TATGTTTACTGCGGCACTGTTTACAATAGCAAATACTTGAACCAACCCAAATGCCCATCAGTGATAGAC
TGGATAAAGAAAACATGGCACATATACACCATGGAATACTATGCAGCCATAAAAAGGTTGAGTTCACGTC
CTTTGCAGGGACATGGCTGAAGCTGGAAACCATCATTCTCAGCCAATAACGTAAGAACAGAAAACCAA
CACTGCATATTTTCACTGATAAGCGTGAGTTGAACAATGAGAACACATGGACATGGGGAAGGGAACATCA
CACACTGGGGCCTGTCAGGGGTTGAGGGGTTAGGGGAGGGATAGTATTAGGAGAAATACCTAATGTAGAT
GACAAGTTGATGGGTGCAGCAAACCACCATGGCACATGGATATCTATGTAACAAACCTGCATGTTCTGCA

Finding genes in genome sequence can be hard

As your english teacher always told you,

grammar and punctuation are the key!!!!!!!!!!

ajisufjjulibnbzvxcqplmjueyyetyttb
winsoutherncaliforniawithgraceand
splendorboundwherethe loftymountai
npeakslookouttolandbeyondproudlys
tandsouralmamatergloroustoseewer
aiseourvoicesproudlyhailinghailin
gtheechoesringingwhileweresingin
goverlandandseathehallsoffamereso
undthynamenoblecitjiuienenenebstijz
inzyiplmnhhtgbswbneemnimistiungjzx

ajisufjjulibnbzvxcqplmjueyyetyttb
winsoutherncaliforniawithgraceand
splendorboundwheretheloftymountai
npeakslookouttolandbeyondproudlys
tandsouralmamatergloroustoseewer
aiseourvoicesproudlyhailinghailin
gtheechoesringingwhileweresingin
goverlandandseathehallsoffamereso
undthynamenoblecitjiuienenebstijz
inzyiplmnhtgbswbneemnimistiungjzx

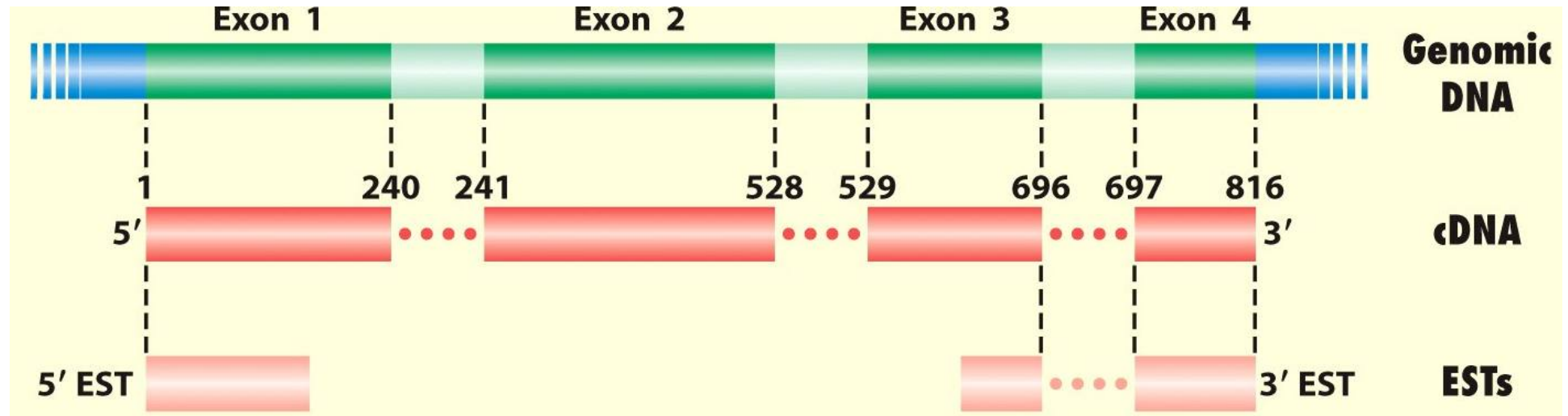
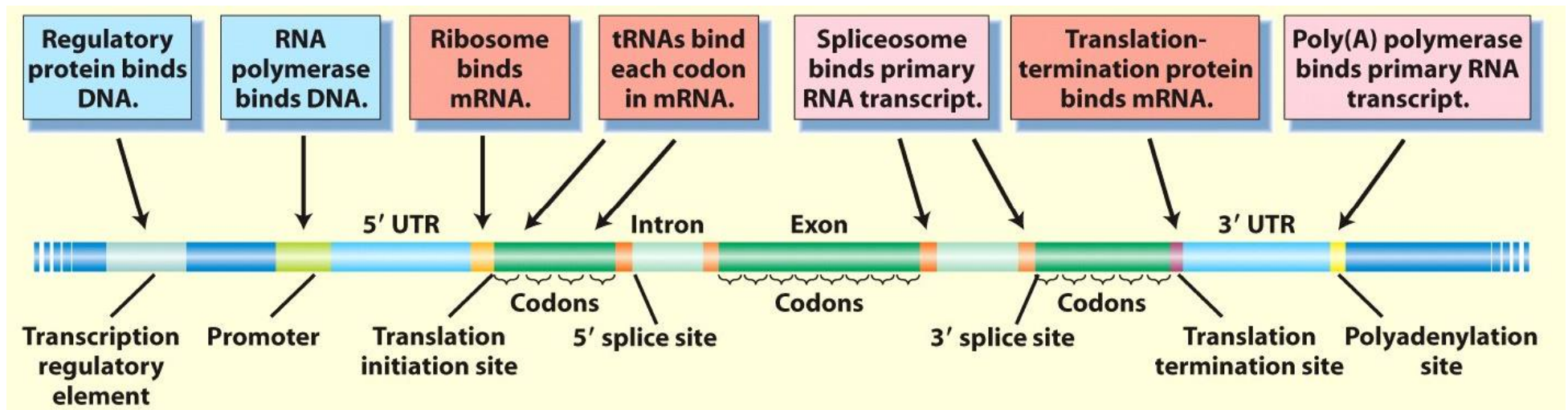
ajisufjjulibnbzvxcqplmjueyyetyttb
w**in**southern**californiawith**graceand
splendorboundwhere**theloftymountai**
npeakslookouttolandbeyondproudlys
tandsouralmamaterglorious**to**seewer
aiseourvoicesproudlyhailinghailin
gtheechoesringingwhileweresingin
goverlandandseathehallsoffamereso
und**thynamenoblecit**jiuienenenestijz
inzyiplmnhtgbswbneemnimestiungjzx

ajisufjjulibnbzvxcqplmjueyyetyttb
wIn southern California with grace and
splendor bound Where the lofty mountai
n peaks lookout to land beyond Proudly s
tands our Alma Mater glorious to see We r
aise our voices proudly hailing hailin
g Thee Echoes ringing while we are singin
g over land and sea The hall of fame reso
und thy name noble CIT jiui eneneb sti jz
inzyiplmnh tgb swbneemnimistiungjzx

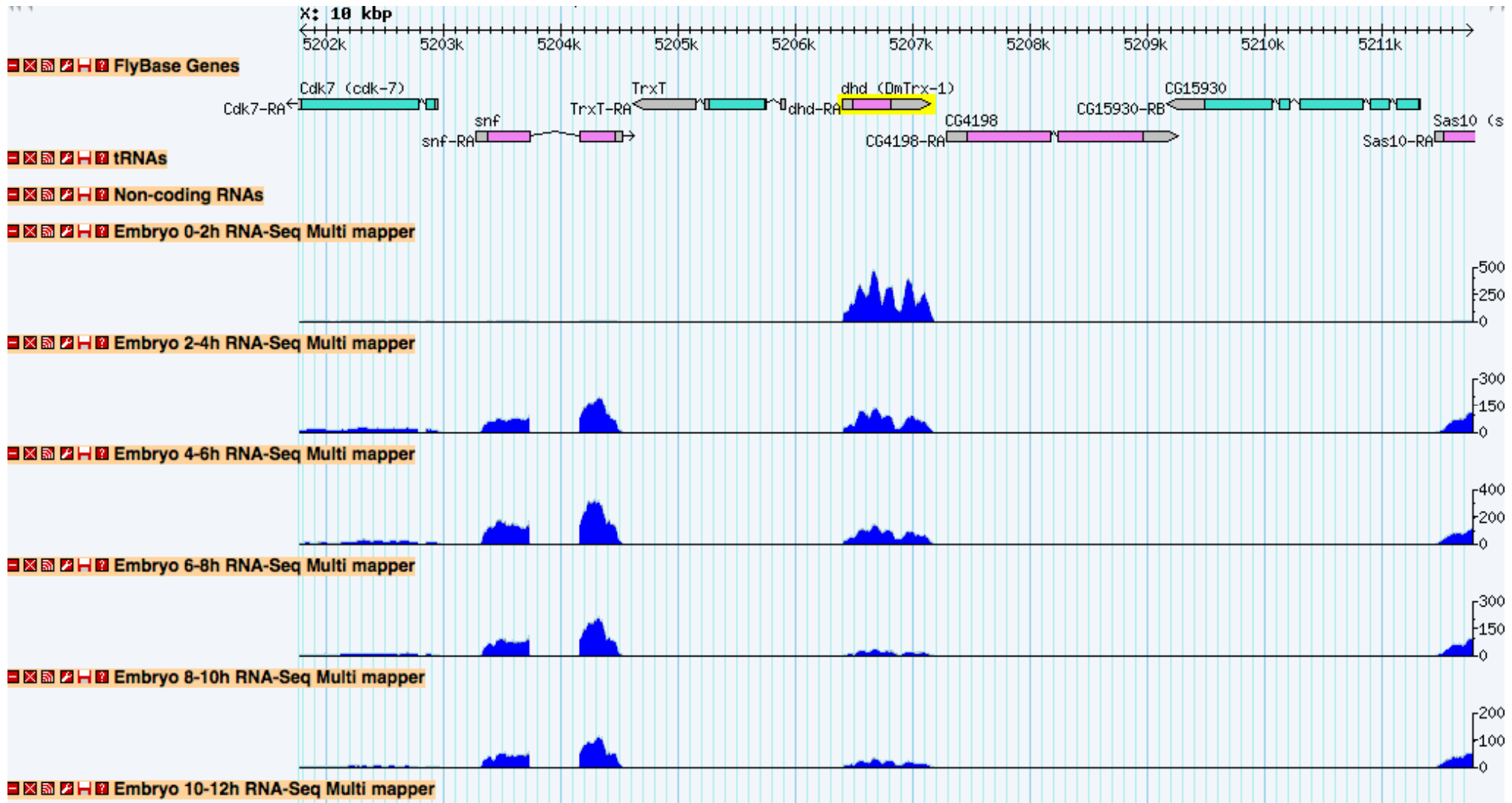
In southern California
with grace and splendor bound
Where the lofty mountain peaks
lookout to land beyond
Proudly stands our Alma Mater
glorious to see
We raise our voices proudly
hailing hailing Thee
Echoes ringing while were singing
over land and sea
The halls of fame resound thy name
noble CIT

In southern California
with grace and splendor bound,
Where the lofty mountain peaks
lookout to land beyond,
Proudly stands our Alma Mater,
glorious to see;
We raise our voices proudly,
Hailing, hailing Thee!
Echoes ringing while we're singing
over land and sea;
The halls of fame resound thy name,
noble CIT!

Manton Barnes, BS '21 EE



<http://www.modencode.org/>



Gene finding

actgggtcaaacgtacggatacaggtaattgaccacattatagaccacgattcgatcgcatcac

Open reading frames

Codon bias

Potential splice sites “gt.....ag”

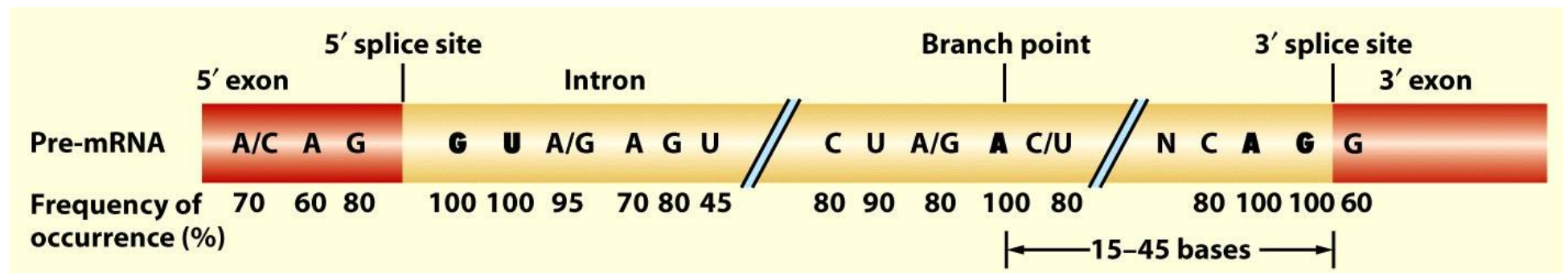
Starts ATG

Stops TGA, TAG, TAA

Poly A addition to 3' UTR AAUAAA

Homology

What about noncoding RNAs?



Gene Models

ced-4 (C35D10.9)

(curated coding gene)

Operons

CEOP3224

Briggsae alignments

cb25.fpc4079:55706..56044

RNAi experiments

JA:C35D10.9

TH:308C6

TH:312D8

Genbank entry

U21324

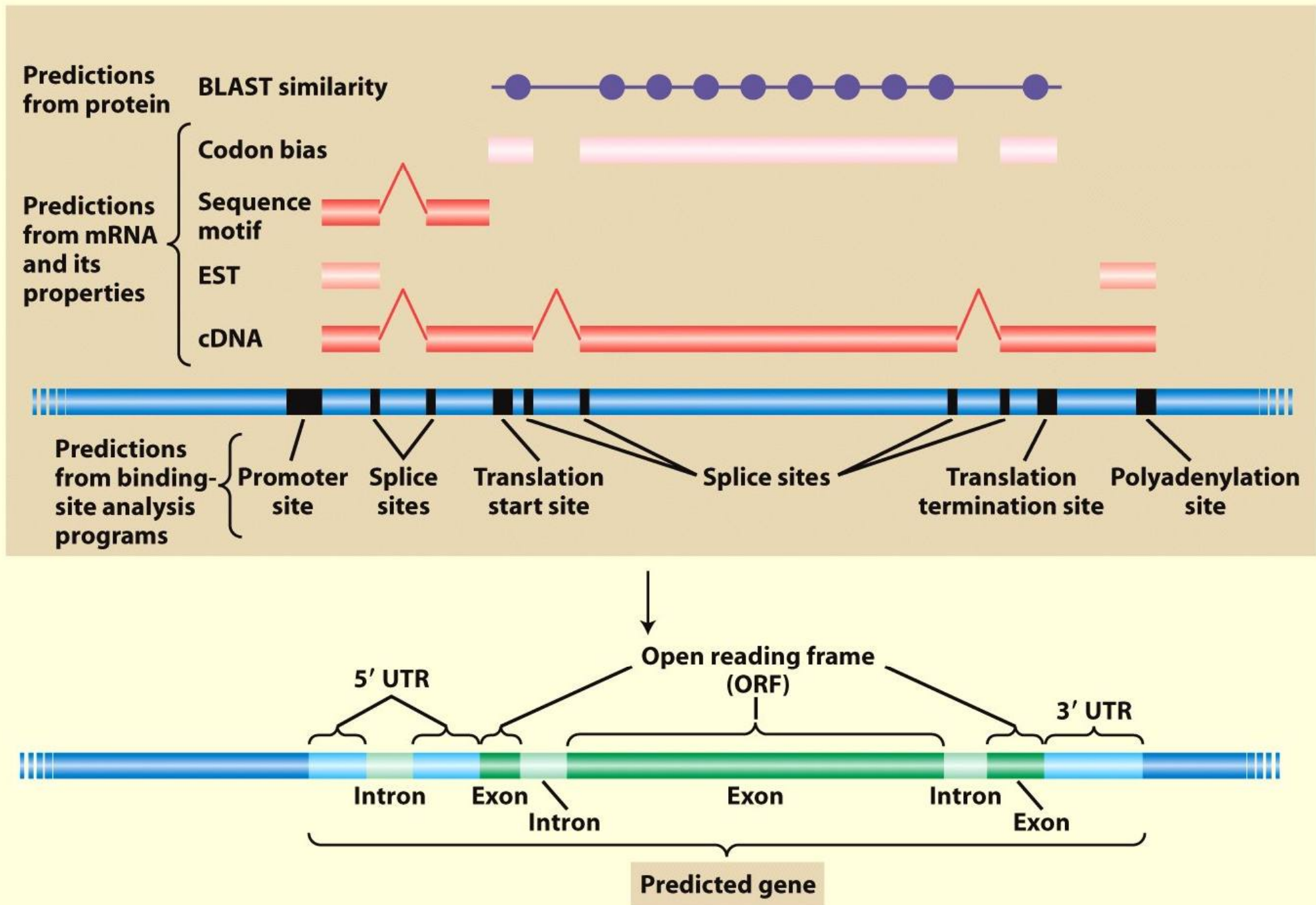
3-frame translation (forward)

T L M M F F K S C E P K T F E K * V G H T N L R L L K * F I L Q *
 R L * C F S S L V N R K H L K S E W D I P I * D F * N N L F Y N
 N A Y D V F Q V L * T E N I * K V S G T Y Q F E T F K I I Y S T I
 aacgcttatgatgtttttcaagtcttgtgaaccgaaaacatttgaaaagtggagtgaggacataccaatttgagactttttaaataattttatttctacaataa
 ttgcgaataactacaaaaagttcagaacacttggccttttgtaaaccttttcactcaccctgtatgggttaaactctgaaaattttattaaataaagatggttatt

*** = stop codons**

XXXgt = splice donor

Many forms of evidence are integrated to make gene predictions



Ribosome profiling: Which RNAs are translated, and which ORFs?

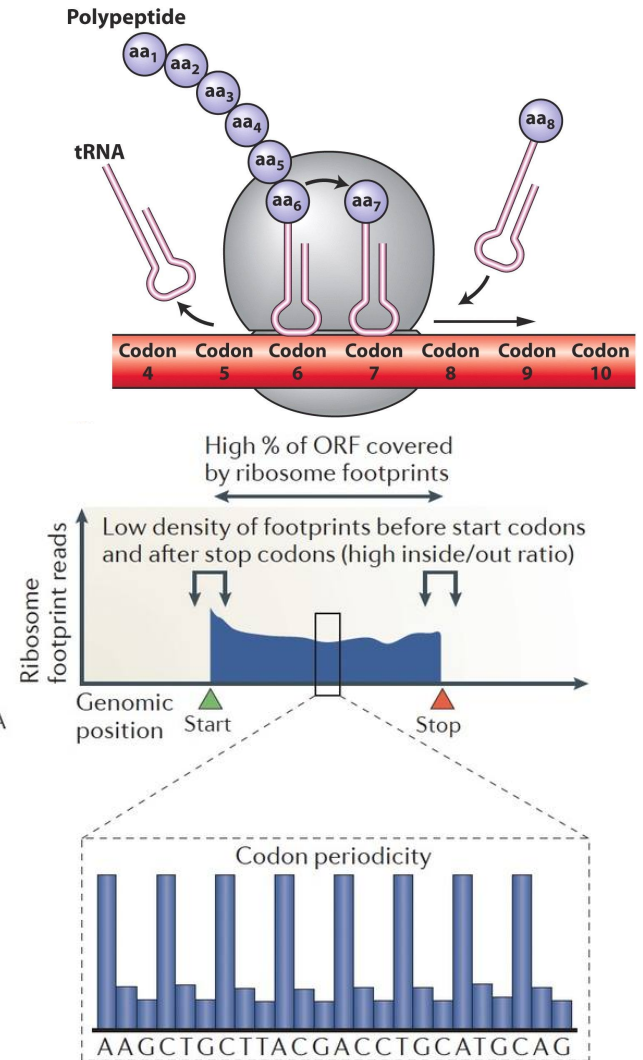
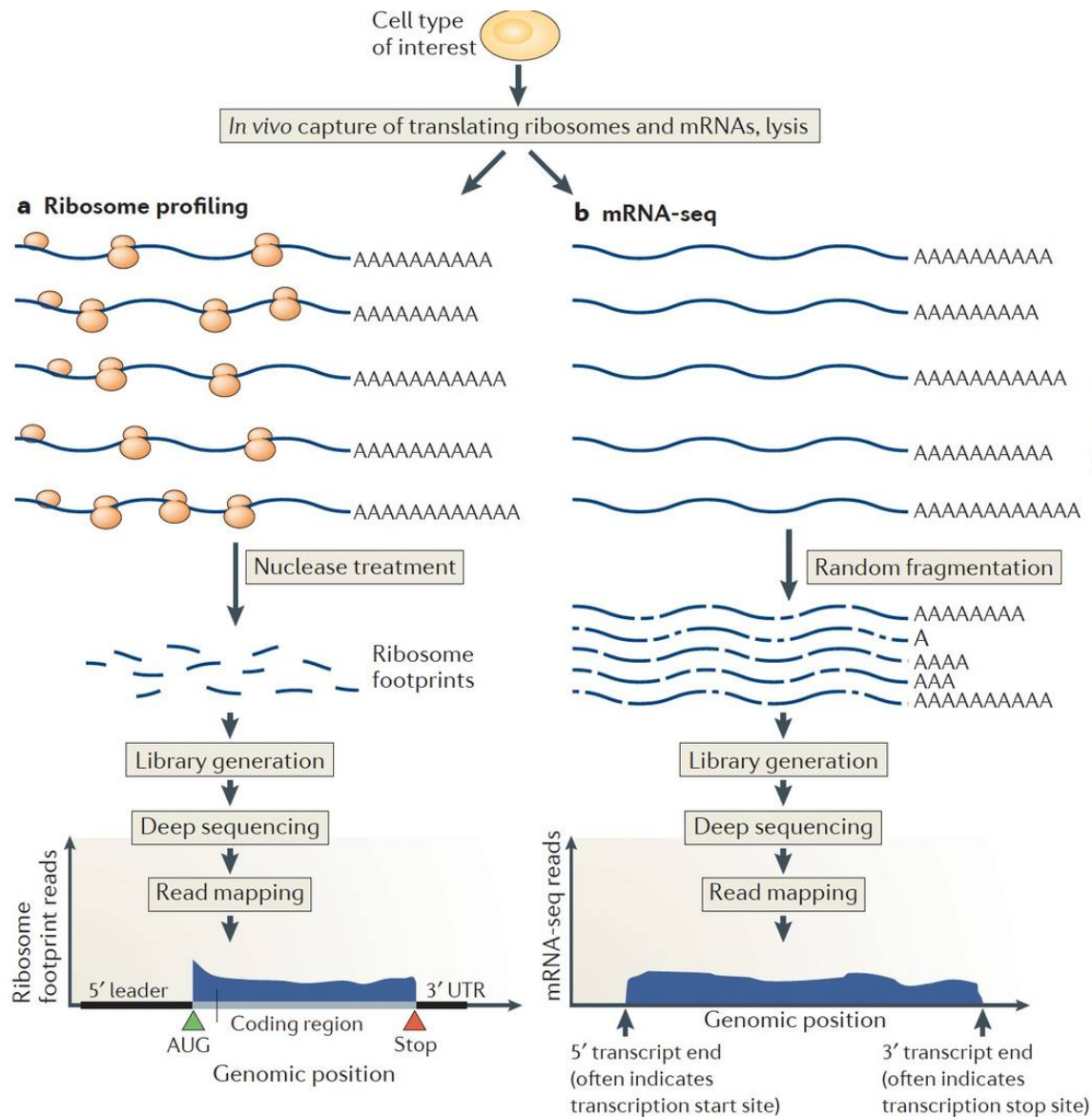


Figure 1 | **An overview of ribosome profiling.** **a** | Ribosome-bound mRNAs are isolated by size and treated with a nonspecific nuclease (typically RNase I or micrococcal nuclease), resulting in protected mRNA fragments termed 'footprints'. These ribosome footprints are isolated and converted to a library for deep sequencing. The ribosome footprints typically show precise positioning between the start and the stop codon of a gene, which facilitates global and experimental identification of genomic coding regions. **b** | By comparison, mRNA sequencing (mRNA-seq) captures random fragments covering the entire mRNA transcript. The positional information determined by standard mRNA-seq

allows approximate determination of transcript boundaries, but it is less precise than that collected by ribosome profiling, owing to the loss of 5' and 3' ends during the fragment generation method that is typically used. **c** | Translated open reading frames (ORFs) contain a stereotyped organization of ribosome footprints. Ribosome footprint density over ORFs begins sharply at the start codon, ends sharply at the stop codon and shows evidence of codon periodicity. True translated regions tend to show ribosome footprint coverage over the majority of the ORF and not typically in the regions before the putative start codon and after the putative stop codon. UTR, untranslated region.

5'3' Frame 1

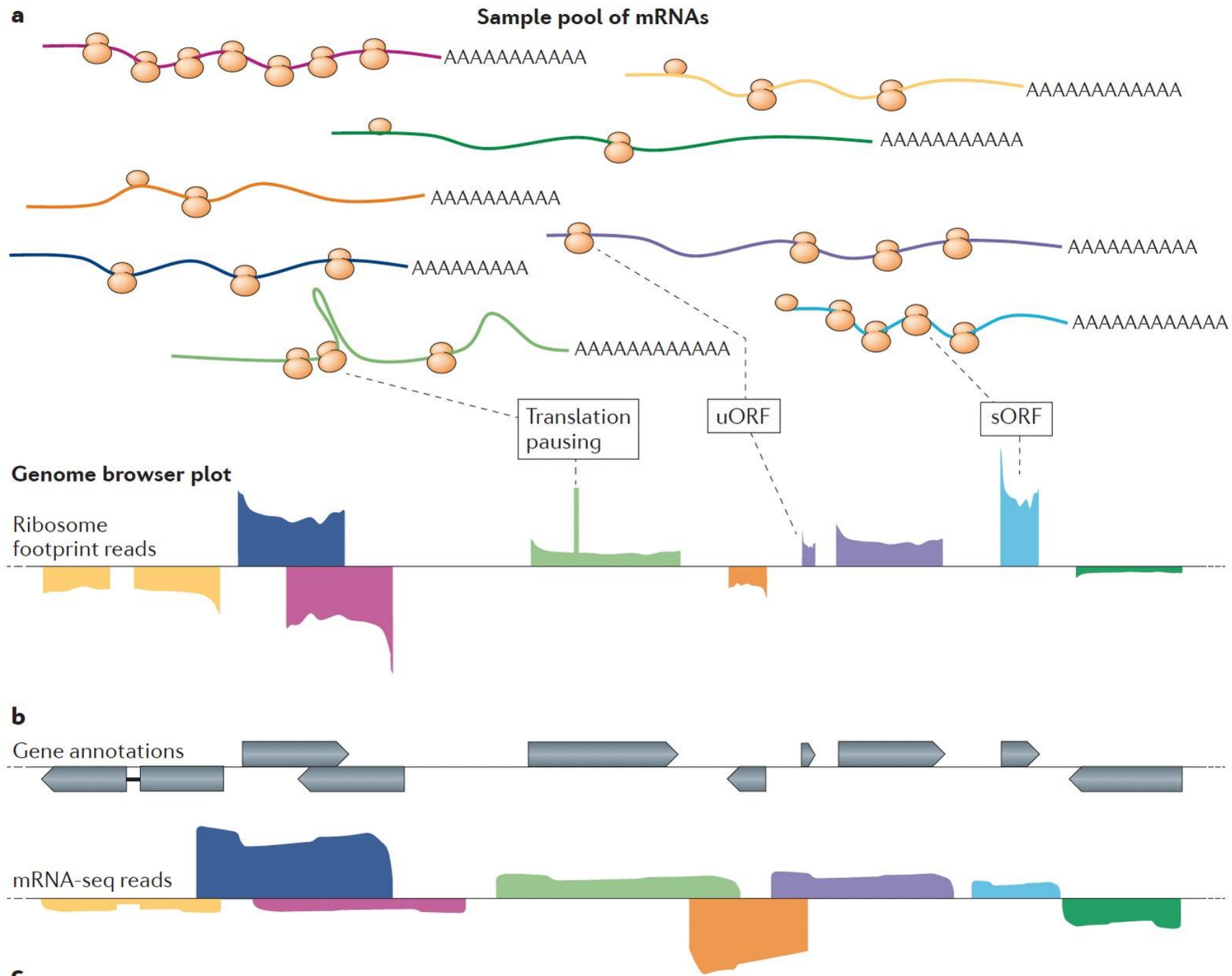
-NYDFVC-QLFVFEILIIYSEI-IACKIVIN-T-KSFKKNIPV-RTAGIANESADIPPPKIIAHLRGKKLFRPFRPASSLSFPRQHPLAH-KAKQNKTKRERRTDKNTYRKVSYIGT-LWRSRCTAFKPTSRRHSCVNILYSKTYFK-FYNLIVVR-EVNENN-RFGGRKKALKKPMIN-YVITV-NVSGK-I-ICVFNLPAVYRKNKMGLLYYT-IMLHTNRRRLSTFLSPTRLFSFMSLPPHKQGHDPYSLSHSLSLPLSPYLALRSLAFSSDF-NNLEAEIHMYIHT-AYVRIPSYLIRSCASLSF-GMGWNEICLSHMFLPGY-LRLYLLSPPMLSSFASHFIHDEINTTPMLQNPPIRGTRLLARIFQFLITAQCHFHLQRCVGSYD-WGFAYSISCLPHLFY-NEIHISLHSLVLRHSFVRYLFKTTVNWQV-SIPSNGD-MVSQEIRSL

5'3' Frame 2

KTTTSCADSCFLKF-YIQYK-LAK-LLIKLKKVLKKTQFRDAQLE-PTKVRTSHQK-LRIAWAAKNYFDRSAPLVLSLSPASTLSLTRKPSKTKQENAEPTKQIPIGK-AI-AHSFGVVVARHSNPPVDVSTVV-TFYTAKLILSDFIILSWCDKR-TKTINDLVAEKRH-KNQ--TDTLSLFKMSLVSKYKYVCSIFPRFIEKTKWVYFTILELCCIQIGAACLPSFLQHAFFHSCPYPHINRDTTIIPPSLTLSLSPFLPISLSALWRFLVISKTIWRPRYTCTYILRHMSGRLILFALVLACHFEGWVGMSVCRTCFCLDIN-DCTFYLRRC-AHSHRISFMMKSTPPPCCKTHPFVERVY-LEFSNF-LPLNVISIYSAA-VP MIDGVLPASVAYLTFTRTKFIYPRCILFCVIHS-DICLKQLSTGKSNLYLVTVTRWFHRKLGH

5'3' Frame 3

KLRLRVLTVVCF-NFNIFRNINSLQNSY-LNLKKF-KKHSSLETHSWNSQRKCGHPTKNNCASLGRQKTISTVPPR-FSLFPPPAPSR SLESQAKQNKTTRTQNRQNKYL-ESELYRHIALA-SLHGIQTHQST-AQLCKHFIQQNLV-VIL-SYRGAIRGERKQLTIIWWPKKGIKKTNDKLIRYHCLKCLW-VNINMCVQSSRGL-KKQNGFTLLYLNYAAYK-APPVYLSLSNTPFFIHVLTPT-TGTRPLSLPLSLSLSPFFSLSRSPLSGVF--FLKQSGGRDTHVHTYLGICQDPVLSYSLLC-LVILRDGLE-NLSVAHVS AWILTKIVPFISADAELIRIAFHS--NQHH PHAAKPTHSWNAFTS-NFPISNYRSMSFPFTALRRFL-LMGFCI-HQLLTSLILLERN SYILAAFCFASFIREIFV-NNCQLASLIYT--R-LDGFTGN-VT



Comparative Sequence Alignment: two versions of Caltech Alma Mater song.

In southern California with grace and splendor bound,
In Smoggy California Where nerds and geeks abound,

Where the lofty mountain peaks lookout to land beyond,
Where the lofty mountain peaks Are seldom ever found,

Proudly stands our Alma Mater, glorious to see;
Proudly stands our torture chamber, Horrible to see.

We raise our voices proudly, Hailing, hailing Thee!
We raise our voices loudly, Screaming painfully.

Echoes ringing while we're singing over land and sea;
Hammers ringing, Anvils singing, Chains for you and me.

The halls of fame resound thy name, noble CIT!
The halls of flame resound thy name, Noble CIT.

Comparative Sequence Alignment: two versions of Caltech Alma Mater song.

In southern California with grace and splendor bound,
In Smoggy California Where nerds and geeks abound,

Where the lofty mountain peaks lookout to land beyond,
Where the lofty mountain peaks Are seldom ever found,

Proudly stands our Alma Mater, glorious to see;
Proudly stands our torture chamber, Horrible to see.

We raise our voices proudly, Hailing, hailing Thee!
We raise our voices loudly, Screaming painfully.

Echoes ringing while we're singing over land and sea
Hammers ringing, Anvils singing, Chains for you and me.

The halls of fame resound thy name, noble CIT!
The halls of flame resound thy name, Noble CIT.

Proteins are modular: compare domains and architecture

% amino acid identity



20%

50%

75%



20%

50%



All Databases



Search



COVID-19 is an emerging, rapidly evolving situation.
Get the latest public health information from CDC: <https://www.coronavirus.gov>.
Get the latest research from NIH: <https://www.nih.gov/coronavirus>.
Find NCBI SARS-CoV-2 literature, sequence, and clinical content: <https://www.ncbi.nlm.nih.gov/sars-cov-2/>.

NCBI Home

Resource List (A-Z)

All Resources

Chemicals & Bioassays

Data & Software

DNA & RNA

Domains & Structures

Genes & Expression

Genetics & Medicine

Genomes & Maps

Homology

Literature

Proteins

Welcome to NCBI

The National Center for Biotechnology Information advances science and health by providing access to biomedical and genomic information.

[About the NCBI](#) | [Mission](#) | [Organization](#) | [NCBI News & Blog](#)

Submit

Deposit data or
manuscripts into NCBI
databases



Download

Transfer NCBI data to
your computer



Learn

Find help documents,
attend a class or watch
a tutorial



Popular Resources

PubMed

Bookshelf

PubMed Central

BLAST

Nucleotide

Genome

SNP

Gene

Protein

PubChem

NCBI News & Blog



COVID-19 is an emerging, rapidly evolving situation.

Get the latest public health information from CDC: <https://www.coronavirus.gov>.

Get the latest research from NIH: <https://www.nih.gov/coronavirus>.

Find NCBI SARS-CoV-2 literature, sequence, and clinical content: <https://www.ncbi.nlm.nih.gov/sars-cov-2/>.

Basic Local Alignment Search Tool

BLAST finds regions of similarity between biological sequences. The program compares nucleotide or protein sequences to sequence databases and calculates the statistical significance. [Learn more](#)

NEWS

New Columns added to the Description Table.

A user driven enhancement to improve the BLAST solution..

Tue, 17 Nov 2020 12:00:00 EST

[More BLAST news...](#)

Web BLAST



Nucleotide BLAST
nucleotide ► nucleotide

blastx
translated nucleotide ► protein

tblastn
protein ► translated nucleotide



Protein BLAST
protein ► protein

BLAST: Basic Local Alignment Search Tool

Designed to be fast for searching large databases



Find local match

Extend out

Calculate P value

Display best hits by

P-value

maximum score

Score = 227 bits (573), Expect = 1e-57

Identities = 122/264 (46%), Positives = 165/264 (62%), Gaps = 1/264 (0%)

Frame = +1

Query: 206 SIVPATPHTLYKKQCA-NCGEGRAYKNISKDPALDGFTPKYYDELEKNEEHFIEIEDLLQ 264
SIVP + + K+ E AY+ I K+P P Y+ E +HFIE++DLL
Sbjct: 78475 SIVPTSTGIVRKRIISGLEDESEVHAYRLICKEPQTAQIVPAYFGIQEMQSQHFIELQDLLA 78654

Query: 265 QFHDPTKTAIMDIKIGTRTFLESEVSNTKKRADLYEKMVAIDNDEPTEEERKCGAITKLR 324
F DP +MDIK+G+RTFLESEVSN R DLY+KM+A+D PT E + AITKLR
Sbjct: 78655 GFRDP---CVMDIKMGSRTFLESEVSNATLRPDLYQKMIAVDAGAPTPAEHEARAITKLR 78825

Query: 325 YMQFRERESSTAQLGFRIEAAKRLEGALEKNFKKVRTVEDVTTTFMDFFGTQSRVRQQL 384
YM FRE SS+ GFRIEA + K+ K R+ E + T F +RS V+++L
Sbjct: 78826 YMTFRESLSSSHSKGFRIEALRLRGRPPVKDLKTCRSSEQIAQTIEQFLAARRS-VQKEL 79002

Query: 385 IERLKSMRKAIEHSSFFNSHEVVGSSILIVFDTEKVGCMIDFAKSSPVPNGRTLNRHTT 444
++RLK MR IE S+FF SHE++GSSI IV+D ++VG W+IDFAK +P ++HR+
Sbjct: 79003 LKRLKHMRLVIEQSTFFASHEIIGSSIFIVYDDDRVGVWLIDFAKCRELPPHVRVDHRSA 79182

Query: 445 WIPGNNEGDGYLIGIDNLVKILEEL 468
W PGN E+G L G+D L++ EE+
Sbjct: 79183 WAPGNREEGLLRGMDELIRSFEEV 79254

Nucleotide sequences that are similar are likely to encode similar functions

Score = 600 bits (664), Expect = 1e-167
Identities = 676/910 (74%), Gaps = 45/910 (4%)
Strand=Plus/Plus

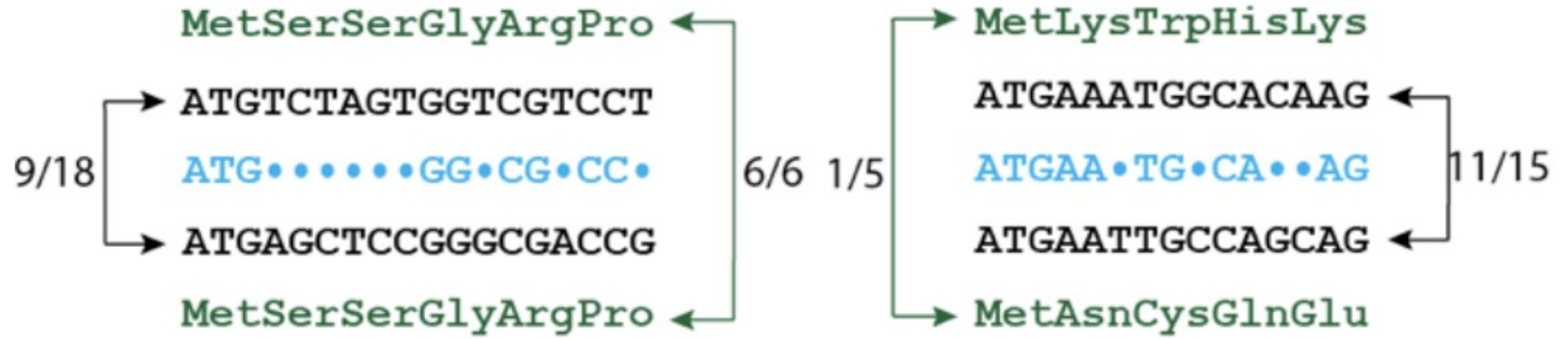
Probability of sequence alignment
to this extent by chance

Query	479	GATCAGGTGGATAACAACACGAACGCGACCCAGCTATTCAAAAATAATATAAACAAAACC	538
Sbjct	73	GACCAGGTGGACAACAATACGAATGCTACGCAGCTGTTCAAGAACAACATCTACAACATC	132
Query	539	AGAATGAACGATTTTAAACCGCGAGGAGACGCGATTAAAGACCTTCACCGACTGGCCGCTA	598
Sbjct	133	A---TGAACGAGTACAATCGGGAGGAAGCGCGCCTGAAAACCTTCGCCAATTGGCCTCTG	189
Query	599	GATTGGCTGGATAAACGCCAATTGGCCCAAACCGGCATGTACTTCACACACGCCGCGAC	658
Sbjct	190	GCGTGGCTCGACAAGCATCAGTTGGCCCGAACGGGCATGTTCTATACGAACGATAACGAT	249
Query	659	AAAGTTAAATGCTTTTTCTGCGGCGTGGAATCGGTTGCTGGGAGCAGGAGGATCAGCCC	718
Sbjct	250	AAGGTCAAGTGCTATTTCTGCGAGGTGGAGATCGGGCGCTGGGATCTCGATGATCAACCT	309
Query	719	GTGCCGGAACATCAGCGATGGTCGCCCAACTGTCCACTGTTGCGCCGGCGCACTACCAAC	778
Sbjct	310	GTACCTGAGCACCTGCGCTGGTCGCCCAACTGTCCGCTCCTGCGCCGTCGCACCACCAAC	369
Query	779	AATGTGCCGATCAATGCCGAAGCATTAGATCGCATCCTGCCGCCAATAAGCTACGATATC	838
Sbjct	370	AATGTGCCGCTGAACAGCGAGGCCCTCGACCGCATCCTGCCCCCATCAGCTACGACATT	429
Query	839	TGCGGCGCCAACGACTCGACGCTAGAGATGAGGGAGCACGCCTACGCAGAAGGCGTCATA	898
Sbjct	430	TGCGGCGCCAACGACTCGACGGTCGAGCTGCGCGAGAGCGCCTATGCCGAGGGCCGCATA	489

Multiple Sequence Alignments

YGL201C	ICC	IDEFDKM	DISDQVAI	HEAMEQQ	TISIAK	AGIHATLN	ARTSIL	AAANP	VGGRYN	RKLS	694
ZK632.1	VCC	IDEFDKM	DLKDQVAI	HEAMEQQ	TISITK	AGVKATLN	ARASIL	AAANP	VNGRYD	RSRP	516
YLR274W	VVC	IDEFDKM	RDEDRVAI	HEAMEQQ	TISIAK	AGITTVLNS	RTSVLA	AAANP	IYGRYD	DLKS	535
R10E4.4	VVC	IDEFDKM	REDDRVAI	HEAMEQQ	TISIAK	AGITTTLNS	RCSVL	AAANS	VYGRWDE	SRG	499
F32D1.10	ICC	IDEFDKM	MDHDRTAI	HEVMEQQ	TISIAK	AGIMTTLN	ARTAI	AAANP	PAYGRYN	PNRS	509
YBR202W	ICC	IDEFDKM	DESDRTAI	HEVMEQQ	TISISK	AVINTNPG	ARTSIL	AAANP	LYGRIN	PNRLS	579
Y39G10A_246.g	VCC	IDEFDKM	NESARSVL	HEVMEQQ	TLSIAK	AGIICQLN	ARASVL	AAANP	VDSKWN	RNKT	588
YPR019W	VCC	IDEFDKM	SDSTRSVL	HEVMEQQ	TISIAK	AGIITTLN	ARSSIL	ASANP	IGSRYN	PNLP	687
YBL023C	VCL	IDEFDKM	NDQDRTSI	HEAMEQQ	SSISISK	AGIVTTLQ	ARCSII	AAANP	NGGRYN	STLP	662
Y17G7B.5	VCL	IDEFDKM	SDQDRTSI	HEAMEQQ	SSISISK	AGIVTSLH	ARCTVIA	ASNPI	IGGRYN	PTRT	617
YEL032W	VVC	IDEFDKM	TDVDRVAI	HEVMEQQ	TVTIAK	AGIHTTLN	ARCSVI	AAANP	VFGQYD	VNRD	528
C25D7.6	VVC	IDEFDKM	SDIDRTAI	HEVMEQQ	GRVTISK	AGIHAKLN	ARCSVL	AAANP	VYGRYN	PFKS	467
Symbols	:	*****	:	:**.*	:*:*	:	:*	:*:*	:	:	

YGL201C	LRG	NLNMTAP	IMSRFDL	FFVILD	DDCNEKID	TELASHI	VDLHM	KRDEA	IEPP	-----	745
ZK632.1	LKY	NVQMSAP	IMSRFDL	FFVLVD	DECNEVTD	YAIARRI	LDNHR	AISEH	TERDSV	-----	569
YLR274W	PGD	NIDFQTTIL	SRFDMIF	IVKDDH	NEERDISI	ANHVINI	HTGNAN	AMQN	-----	585	
R10E4.4	D-DN	IDFMPTIL	SRFDMII	IVKDDH	VLKDATL	AKHVIEV	HVNAS	AAKER	DIAGVPKT	--	556
F32D1.10	IEQ	NVDLPAALL	SRFDLILL	MQDKAD	RENDKIL	AEHITYV	HQHGCH	PNREK	-----	560	
YBR202W	PLD	NINLPAALL	SRFDILFL	MLDIPSR	DDDEKLA	EHVTVYV	HMHNKQ	PDLDFTP	-----	632	
Y39G10A_246.g	IVEN	ITLPHTLL	SRFDLIFLI	VDAQD	EMQDRRL	GNHLVSL	YFENDG	NQEKTEH	-----	641	
YPR019W	VTEN	IDLPPPLL	SRFDLVYL	VLDKVDE	KNDRELAK	HLTNLYL	EDKPEH	ISQDD	-----	740	
YBL023C	LAQ	NVSLTEPIL	SRFDILCV	VRDLVDEE	ADERLAT	FVVD	SHVRSH	PENDED	-----	713	
Y17G7B.5	FAEN	VDLTEPIL	SRFDVLCV	IRDSVDS	VEDERLAK	FVVG	NHRTHHP	DAKKIVKEG	DEL--	675	
YEL032W	PHQ	NIALPDSLL	SRFDLLFV	YTDINE	IRDRSISE	HVLRTH	RYLPPG	YLEGEP	VRERLNL	588	
C25D7.6	PMEN	IGMQDSSL	SRFDLIFV	LLDEHD	ADKIDANV	AHV	VLKLHT	YRTQGE	ADGTVLPMGGG	-	526
Symbols	*	:	:	*****	:	*	:	*	:	:	



Which is protein-coding?

GlyArgProLysTyr

GGTCGTCCTAAGTAC

GG•CG•CC•AA•TA•

GGGCGACCGAAATAT

GlyArgProLysTyr

species 1

species 2

LysTrpHisLysPro

AAATGGCACAAGCCC

•AAT•GCA•A•GC•C

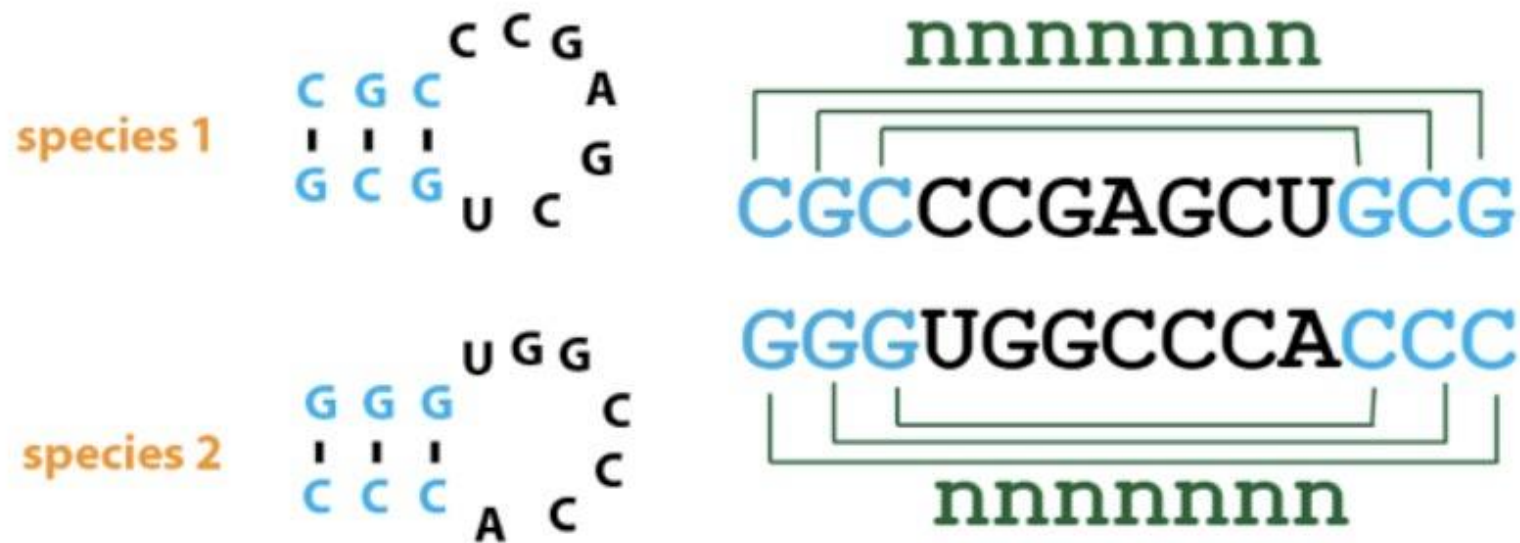
GAATCGCAAATGCGC

AsnCysGlnGluArg

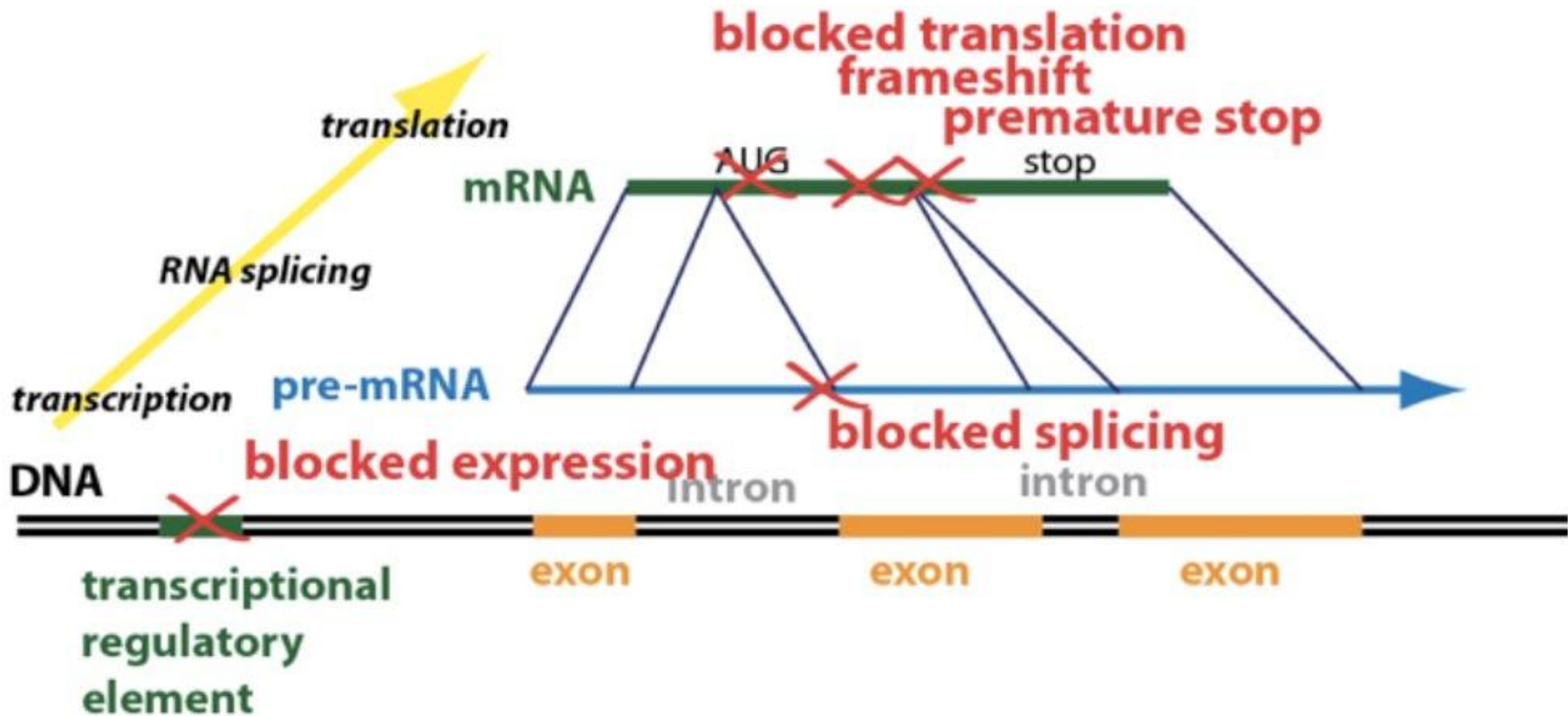
**5/5 substitutions in
wobble position**

**1/5 substitutions in
wobble position**

Sequence constraints are different for non-coding RNA



Pseudogenes



Can still have function. For example, a (now) non-coding transcript can still act as a sink for RNA-binding proteins (or other RNAs) that bind and regulate function of the wildtype transcript

Functional Complementation

genotype	DNA	Phenotype
+	-	wild type
<i>A(lf)</i>	-	mutant
<i>A(lf)</i>	A	wild type
+	A	wild type
<i>A(lf)</i>	species2-A	wild type
+	species2-A	wild type

Genome sizes and gene numbers

Mb	organism	year sequenced	~genes
0.6	endosymbionts (<i>Buchnera</i>)	2000	800
4.6	<i>E. coli</i> (bacteria)		4,300
15	budding yeast	1996	6,300
102.2	<i>Caenorhabditis elegans</i>	1998	20,000
104	<i>Caenorhabditis briggsae</i>	2003	20,000
120	<i>Arabidopsis thaliana</i>	2000	25,000
120+	<i>Drosophila melanogaster</i>	2000	14,000
278	<i>Anopheles gambiae</i>	2002	14,000
~400	puffer fish (<i>Fugu</i>)	2001	39,000
3300	human	2000-2004	23,000
2700	mouse	2000+	23,000
	puffer fish (<i>Tetraodon</i>)	2004	~23,000
	chimpanzee	2005	

Ambitious Project to Sequence Genomes of 1.5 Million Species Kicks Off

The Earth BioGenome Project promises to revolutionize biology



(Mirhee Lee)

By [Jason Daley](#)
[smithsonianmag.com](#)
November 5, 2018

Last week, a global consortium of scientists officially launched the [Earth BioGenome Project](#). As Kate Kelland at [Reuters](#) reports, the backers are calling the extensive initiative the next “moonshot for biology.” Projected to cost \$4.7 billion, it aims to sequence the DNA of the 1.5 million known species of eukaryotic, or complex species of life on Earth. Having a DNA map of so many species, researchers say, will reframe what we know about biology, ecology and biodiversity.

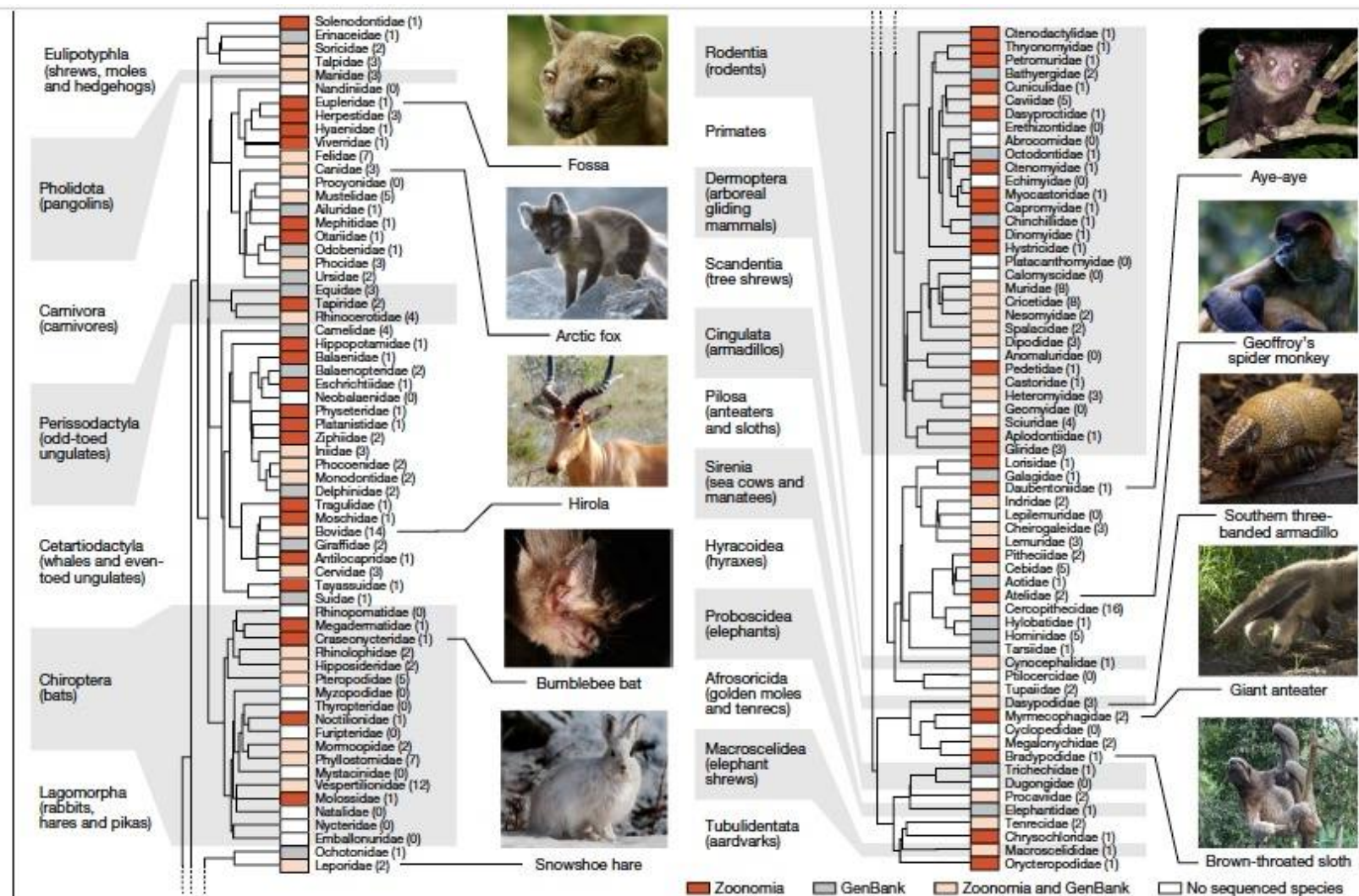


Fig. 1 | The Zoonomia Project brings the fraction of eutherian families that are represented by at least one assembly to 83%. Phylogenetic tree of the mammalian families in the Zoonomia Project alignment, including both our new assemblies and all other high-quality mammalian genomes publicly available in GenBank when we started the alignment (March 2018) (Supplementary Table 2). Tree topology is based on data from TimeTree (www.timetree.org)⁴⁷. Existing taxonomic classifications recognize a total of 127 extant families of eutherian mammal⁴⁸, including 43 families that were not previously represented in GenBank (red boxes) and 41 families with additional representative genome assemblies (pink boxes). Of the remaining families, 21 had GenBank genome assemblies but no Zoonomia Project assembly

(grey boxes) and 22 had no representative genome assembly (white boxes). Parenthetical numbers indicate the number of species with genome assemblies in a given family. Image credits: fossa, Bertal/Wikimedia (CC BY-SA); Arctic fox, Michael Haferkamp/Wikimedia (CC BY-SA); hiroia, JRProbert/Wikimedia (CC BY-SA); bumblebee bat, Sébastien J. Puechmaile (CC BY-SA); snowshoe hare, Denali National Park and Preserve/Wikimedia (public domain); aye-aye, Tom June/Wikimedia (CC BY-SA); Geoffroy's spider monkey, Patrick Gijsbers/Wikimedia (CC BY-SA); southern three-banded armadillo, Hedwig Storch/Wikimedia (CC BY-SA); giant anteater, Graham Hughes/Wikimedia (CC BY-SA); brown-throated sloth, Dick Culbert from Gibsons, B.C., Canada/Wikimedia (CC BY).

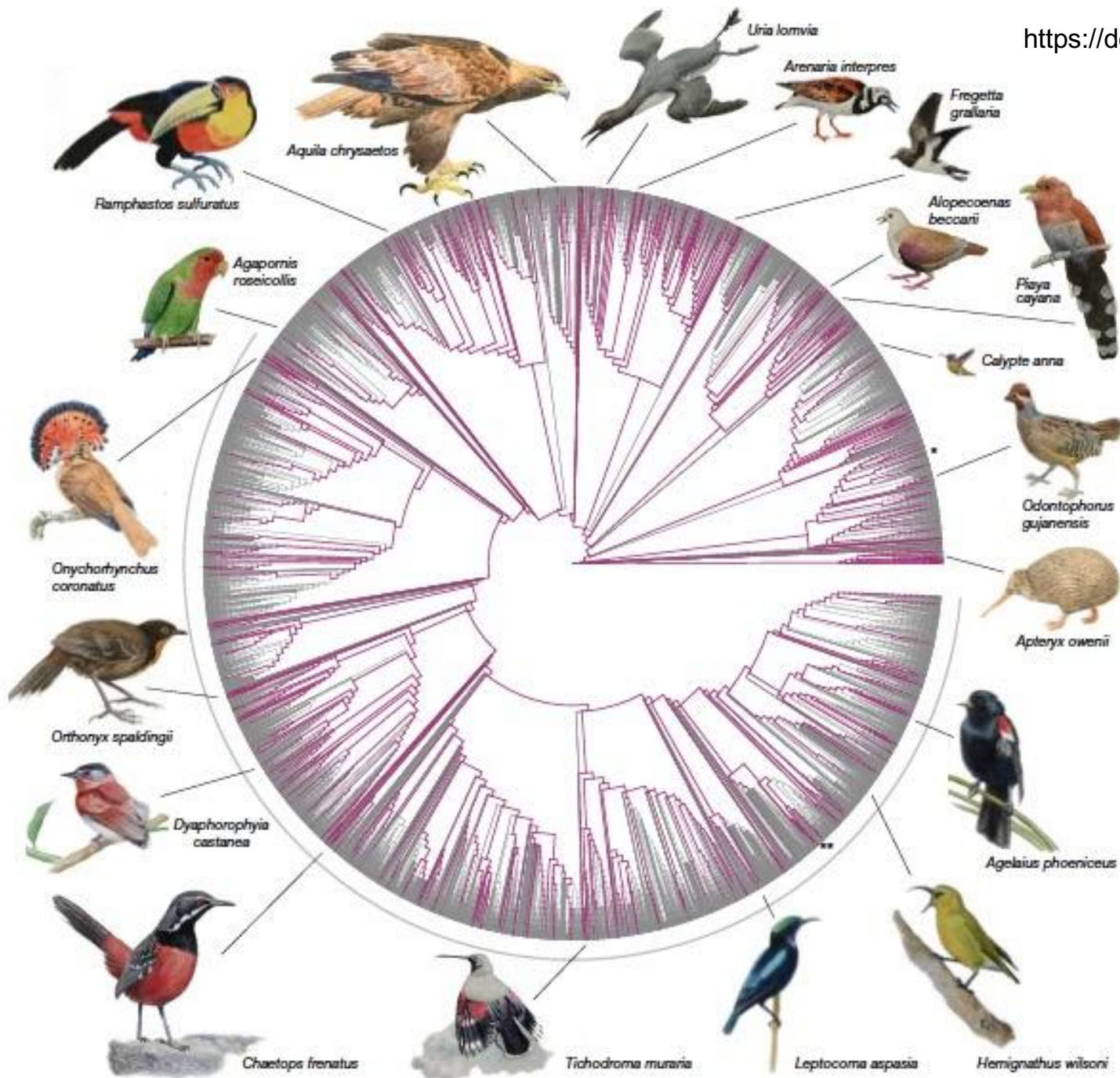


Fig. 1 | Newly sequenced genomes densely cover the bird tree of life. The 10,135 bird species^{11,12} are shown on a draft phylogeny that synthesizes taxonomic and phylogenetic information³⁴ (Supplementary Data). In total, 363 species, covering 92.4% of all families, now have at least 1 genome assembly

per sequenced family (purple branches). The grey arc marks the diverse Passeriformes radiation, with 6,063 species, of which 173 species have genome assemblies now. Chicken (*) and zebra finch (**) are marked for orientation. Paintings illustrate examples of sequenced species.

How do organisms increase in complexity (create novelty)
if the gene number is relatively constant?

Genome sizes and gene numbers

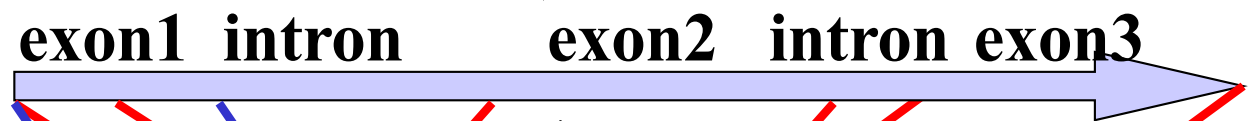
Mb	organism	year sequenced	~genes
0.6	endosymbionts (<i>Buchnera</i>)	2000	800
4.6	<i>E. coli</i> (bacteria)		4,300
15	budding yeast	1996	6,300
102.2	<i>Caenorhabditis elegans</i>	1998	20,000
104	<i>Caenorhabditis briggsae</i>	2003	20,000
120	<i>Arabidopsis thaliana</i>	2000	25,000
120+	<i>Drosophila melanogaster</i>	2000	14,000
278	<i>Anopheles gambiae</i>	2002	14,000
~400	puffer fish (<i>Fugu</i>)	2001	39,000
3300	human	2000-2004	23,000
2700	mouse	2000+	23,000
	puffer fish (<i>Tetraodon</i>)	2004	~23,000
	chimpanzee	2005	

Complex genes

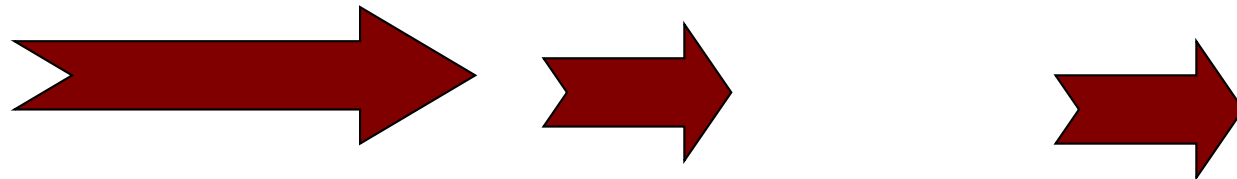
A. More Regulatory Sites



B. More Transcripts

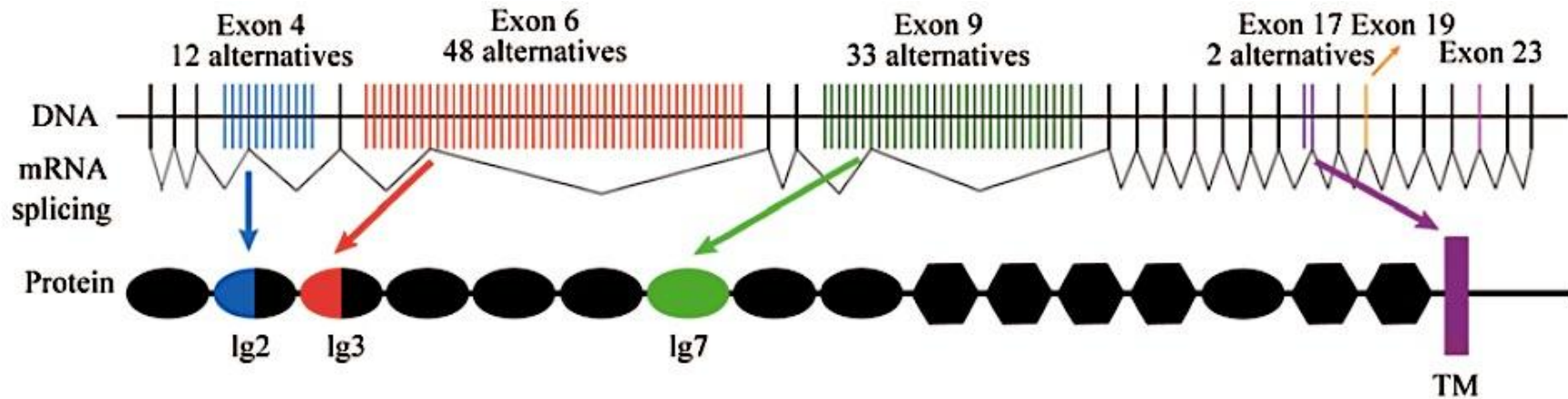


C. Alternative Splicing



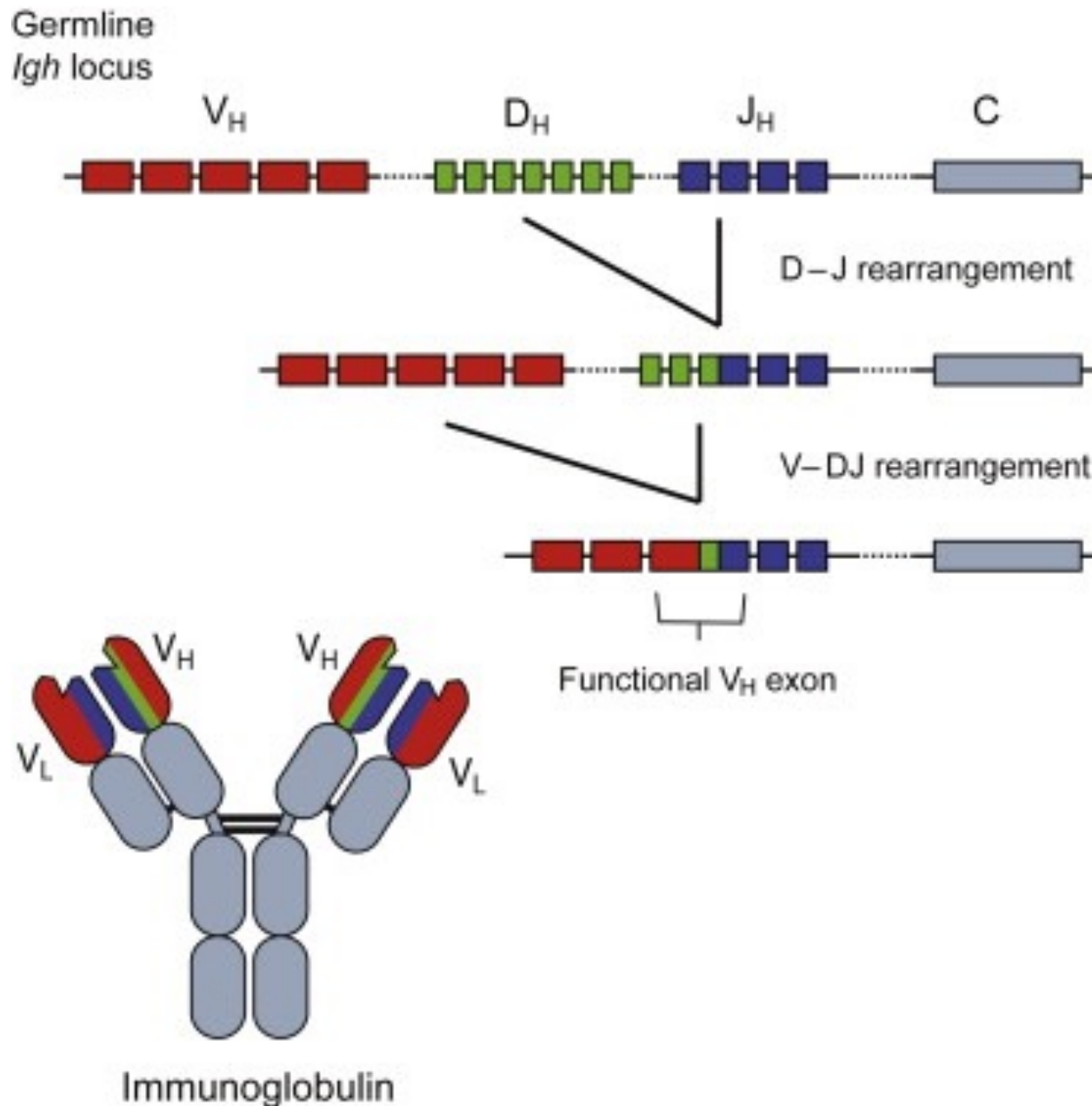
**MULTIPLE PROTEINS MADE IN DIFFERENT AMOUNTS
AND PLACES**

38,016 possible splice variants !!!

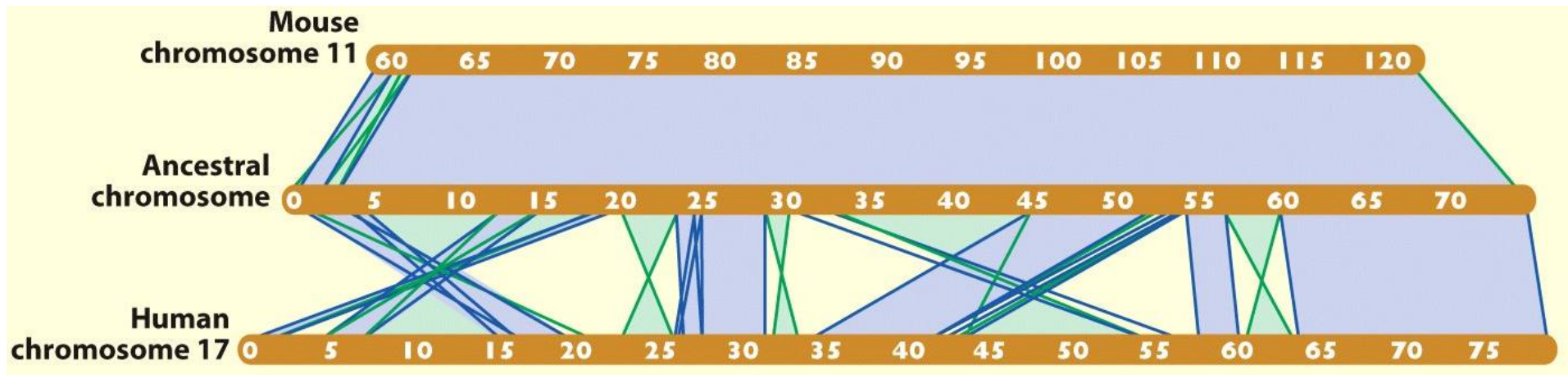


Generation of diverse isoforms via alternative splicing (adopted from Hattori D *et al.*, 2008)^[46]. There are 4 alternative-spliced exon clusters in the *Drosophila Dscam* gene: the exon 4 contains 12 alternatives, the exon 6 contains 48 alternatives, the exon 9 contains 33 alternatives and the exon 17 contains 2 alternatives. Each transcript contains one alternative exon from each exon cluster. The full-length of endodomain is largely restricted in embryogenesis stage, while most Dscams have no exon 19 and exon 23 in postembryonic stages. The extracellular domain of Dscam contains 10 immunoglobulin (Ig) domains (ellipse), 6 fibronectin type III repeats (hexagons) and a short transmembrane domain (TM). Alternative splicing generates diversity in the Ig2, Ig3, Ig7 and TM domains.

Cell-type specific rearrangements of DNA to create novel protein coding regions



The mouse and human genome have large syntenic blocks of genes in common



>gi|19773481|dbi|AB072322.1|AB072322S1

Pan troglodytes COX8 gene for cytochrome oxidase subunit VIII,
exon 1

ATGTCCGTCTGACGCCGCTGCTGCTGCGGGGCTTGACAGGCTCGGCCCGGGCGGCTCCCAAGTGCCGCGCG
CCAAGATCCATTCTGTTGCCGCCGAGGAGAAGCTTGGGATCATG

>gi|11436791|ref|XM_006132.1|

Homo sapiens cytochrome c oxidase subunit VIII (COX8), mRNA
Length = 473

Score = 218 bits (110), Expect = 8e-55
Identities = 113/114 (99%)
Strand = Plus / Plus

Chimp
Human

Query: 1 atgtccgtctgacgccgctgctgctgccccgttgacaggctcgccccgagcgtccca 60

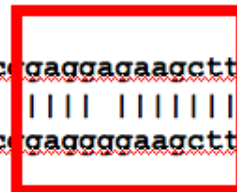
|||||

Sbjct: 42 atgtccgtctgacgccgctgctgctgccccgttgacaggctcgccccgagcgtccca 101

Query: 61 gtgccgcgcgccaagatccattcattgccgcg gaggagaagcttggatcatg 114

|||||

Sbjct: 102 gtgccgcgcgccaagatccattcattgccgcg gaggagaagcttggatcatg 155



The Age of Databases:

**Information about an organism can be organized
around the Genome sequence**

- **The sequence and its variation,**
- **The genes, mRNAs and proteins**
- **The regulatory elements,**
- **Where and when each gene is expressed,**
- **How genes and proteins interact with each other to specify
the properties of each organism.**

There are Model Organism Genome Databases such as

WormBase (www.wormbase.org)

FlyBase (www.flybase.org)

Yeast (yeastgenome.org)

Knowledgebases are part of our lives

WormBase Version: WS257

My WormBase (0 ★) | [Login](#) | [For Developers](#) | [Contact Us](#)

Search directory... for a gene

[About](#) [Directory](#) [Tools](#) [Downloads](#) [Community](#) [Support](#)

[Submit Data](#)

[ParaSite](#)



Species: *C. elegans* (Genome assembly: WBcel235)

Gene » *let-23*



Page Content

[Overview](#)

[Expression](#)

[External Links](#)

[Gene Ontology](#)

[Genetics](#)

[History](#)

[Homology](#)

[Human Diseases](#)

[Interactions](#)

[Location](#)

[Mapping data](#)

[Phenotype Graph](#)

[Phenotypes](#)

[Reagent](#)

References

[Sequence features](#)

[Sequences](#)

Tools

[Tree Display](#)

[Genetic Map](#)

[Nucl. Aligner](#)

Overview

let-23 (LEThal)

let-23 encodes an EGF-receptor-family transmembrane tyrosine kinase that affects viability, inductive signaling during development of the vulva, male spicule formation, posterior development of the epidermis, ovulation and stress induced sleep; *let-23* is genetically upstream of the *let-60/RAS* pathway with respect to viability and vulval development, is upstream of phospholipase C gamma with respect to fertility and sleep, and is expressed in vulval precursor cells and the ALA neuron.

Date last updated: 2015-5-29

Person evidence: Paul Sternberg

Curator: Erich Schwarz, Ranjana Kishore

Paper evidence: Hill, Mansfield, Lopez, Raizen, & Van Buskirk, 2014

evidence

► Details

► From *C. elegans* I and II

Species: *Caenorhabditis elegans*

Sequence: ZK1067.1

Other names: CELE_ZK1067.1, *kin-7*

Type: protein coding

Gene class: *let*

Status: Live

Clone: [ZK1067](#)

[T08E2](#)

Parent seq: [ZK1067](#)

Gene name: Ann Rose

evidence:

WormBase ID: WBGene00002299

Sequences

Location

Homology

Genetics

Interactions

Expression

[Need help or have feedback?](#)

ZFIN ID: ZDB-GENE-030918-1

Your Input Welcome

Gene Name: *epidermal growth factor receptor a (erythroblastic leukemia viral (v-erb-b) oncogene homolog, avian)*

Gene Symbol: *egfra*

Sequence Ontology ID : [SO:0000704](#)

Previous Names: *egfr, erbb1a* (1)

Location: [Chr: 2 Mapping Details/Browsers](#)

[Nomenclature History](#)

GENE EXPRESSION

All Expression Data: [8 figures](#) from 7 publications

Wild-type Stages, Structures: [Gastrula:50%-epiboly \(5.25h-5.66h\)](#) to [Adult \(90d-730d, breeding adult\)](#)

[brain](#) ☐, [fin](#) ☐, [gill](#) ☐, [intestine](#) ☐ (all 12) [▶](#)

MUTATIONS AND SEQUENCE TARGETING REAGENTS

Allele	Type	Localization	Consequence	Mutagen	Suppliers
la023699Tg	Transgenic Insertion	Unknown	Unknown	DNA	Zebrafish International Resource Center (ZIRC) (order this)
sa1683	Point Mutation	Unknown	Premature Stop	ENU	Zebrafish International Resource Center (ZIRC) (order this)
sa11905	Point Mutation	Unknown	Premature Stop	ENU	European Zebrafish Resource Center (EZRC) (order this) Zebrafish International Resource Center (ZIRC) (order this)
sa14931	Point Mutation	Unknown	Premature Stop	ENU	European Zebrafish Resource Center (EZRC) (order this) Zebrafish International Resource Center (ZIRC) (order this)
sa15365	Point Mutation	Unknown	Premature Stop	ENU	European Zebrafish Resource Center (EZRC) (order this)



FB2022_05, released September 29, 2022

A Database of *Drosophila* Genes & Genomes

Home Tools Downloads Links Community Species About Help Archives

J2G Jump to Gene

D.melanogaster
D.virilis
A.mellifera
BLAST

JBrowse

ON
OFF
RNA-Seq

GO
PHENOTYPE
ANATOMY
DISEASE
MORE
Vocabularies

ID
ID
ID Validator

FIELD
DATA
sequence
Batch Download

Fast-Track

Tweets from @FlyBaseDotOrg

Follow



External Resources

- News and Outreach
- Drosophila Journals and Protocols
- Tool/Reagent Resources (wiki)
- Meetings/Courses/Jobs
- Drosophila in Disease Research
- Inter-species
- Fly Boards

Please **PAY** your yearly website access fee! (annual renewal is automatic)

The NHGRI has reduced the funding of FlyBase by 50%

Please help with a FlyBase website access fee of \$150.00 PER person / PER year
Questions? See this [FAQ](#) or [email us](#). Optional general tax-deductible contributions [here](#).

QuickSearch

Human Disease Protein Domains Gene Groups Pathways GO Data Class
Search FlyBase Homologs GAL4 etc Expression Phenotype References

Everything

Click [here](#) to submit multiple IDs/symbols.

Note: Wild cards (*) can be added to your search

Commentary see all Commentaries

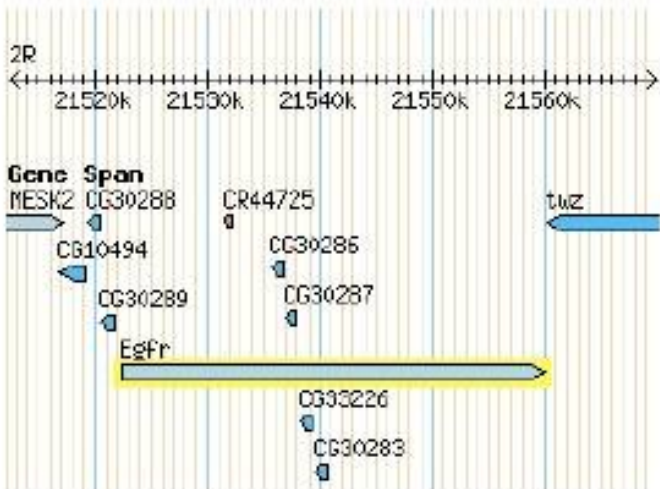


FB2017_01, released February 14, 2017

Gene Dmel\Egfr

[Home](#)
[Tools](#)
[Downloads](#)
[Links](#)
[Community](#)
[Species](#)
[About](#)
[Help](#)
[Archives](#)
[Jump to Gene](#)
[Go](#)

[FlyGene Wiki](#)
[Help](#)
[Open All](#)
[Close All](#)

General Information			
Symbol	DmelEgfr	Species	D. melanogaster
Name	Epidermal growth factor receptor	Annotation symbol	CG10079
Feature type	protein_coding_gene	FlyBase ID	FBgn0003731
Gene Model Status	Current	Stock availability	53 publicly available
Also Known As	DER, top, fib, Elp, dEGFR, Egf-r, c-erbB		
Gene Snapshot	Epidermal growth factor receptor (Egfr) is the transmembrane tyrosine kinase receptor for signaling ligands in the TGF α family (<i>grk</i> , <i>spl</i> , <i>vn</i> , and <i>Krn</i>), which utilises the intracellular MAP kinase pathway. Egfr roles include growth regulation, cell survival and developmental patterning. [Date last reviewed: 2016-10-06]		
Genomic Location			
Cytogenetic map	57E9-57F1	Sequence location	2R:21,522,420..21,559,977 [+]
Recombination map	2-95		
Genomic Maps	GBrowse JBrowse  Decorated FASTA Get genome region Gene region Get FASTA		
Other Genome Views	The following external sites may use different assemblies or annotations than FlyBase.		

www.alliancegenome.org

one-stop shopping for genes

The screenshot displays the Alliance of Genome Resources website. At the top left is the logo with the text "ALLIANCE of GENOME RESOURCES". Below it is a navigation bar with links for "HOME", "ABOUT US", and "WORK PRODUCTS". A search bar on the right contains the text "egfr" and a magnifying glass icon. A dropdown menu lists search results for "egfr", each preceded by a small icon and followed by the word "Gene". The results include: Egfr (Mmu), Egfr (Dme), EGFR (Hsa), Egfr (Rno), EGFR-AS1 (Hsa), Egfr::Tl (Dme), Egfrs (Mmu), egfra (Dre), Egfr::Rat\Neu (Dme), and Egfr::Hsap)EGFR (Dme). The background features a large graphic with the text "ALLIANCE of GENOME RESOURCES" and "The Genome Six Mo joined of Gen".

ALLIANCE
of GENOME RESOURCES

HOME ABOUT US WORK PRODUCTS

egfr

Egfr (Mmu) Gene
Egfr (Dme) Gene
EGFR (Hsa) Gene
Egfr (Rno) Gene
EGFR-AS1 (Hsa) Gene
Egfr::Tl (Dme) Gene
Egfrs (Mmu) Gene
egfra (Dre) Gene
Egfr::Rat\Neu (Dme) Gene
Egfr::Hsap)EGFR (Dme) Gene

ALLIANCE
of GENOME RESOURCES

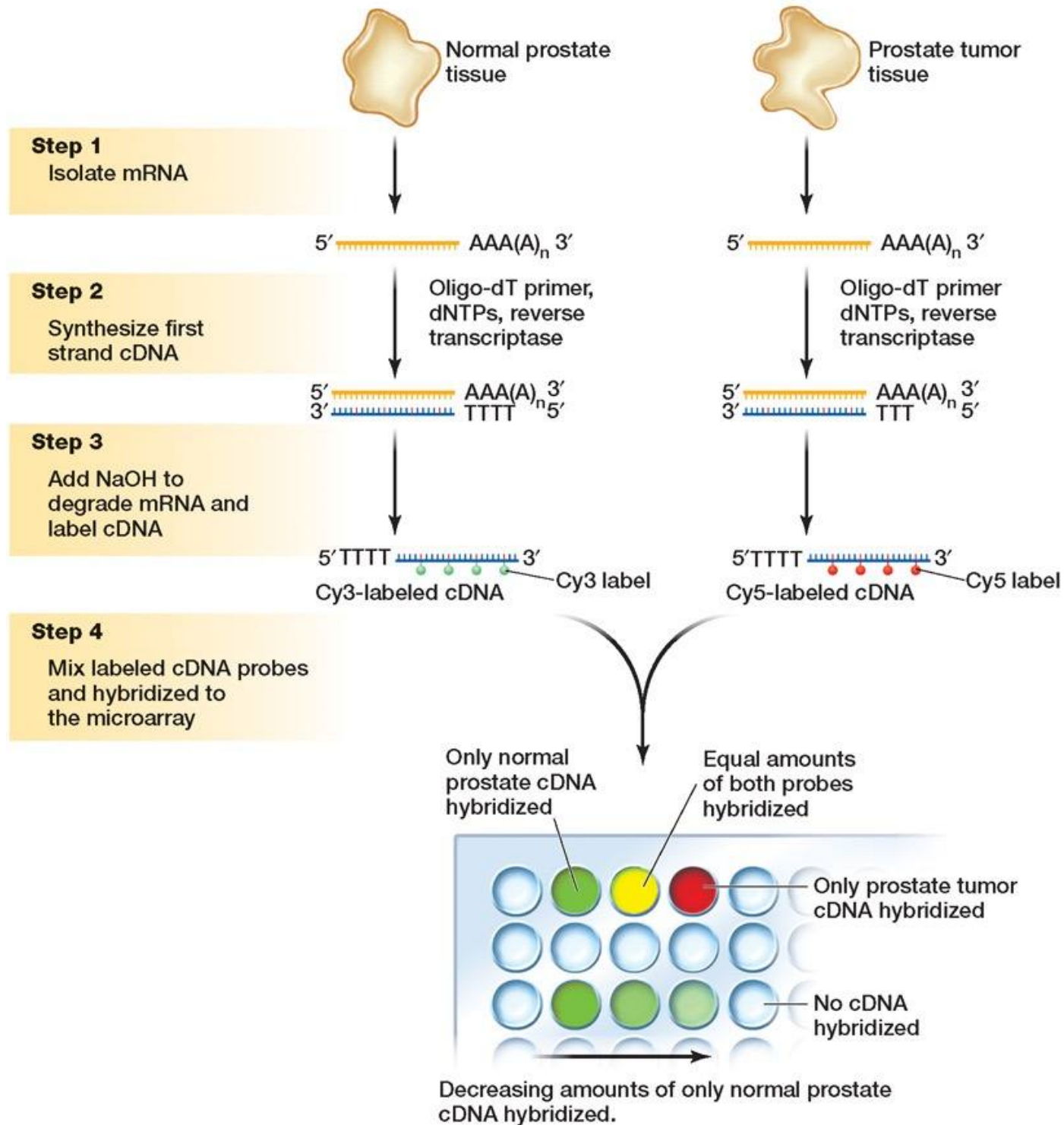
The Genome Six Mo joined of Gen



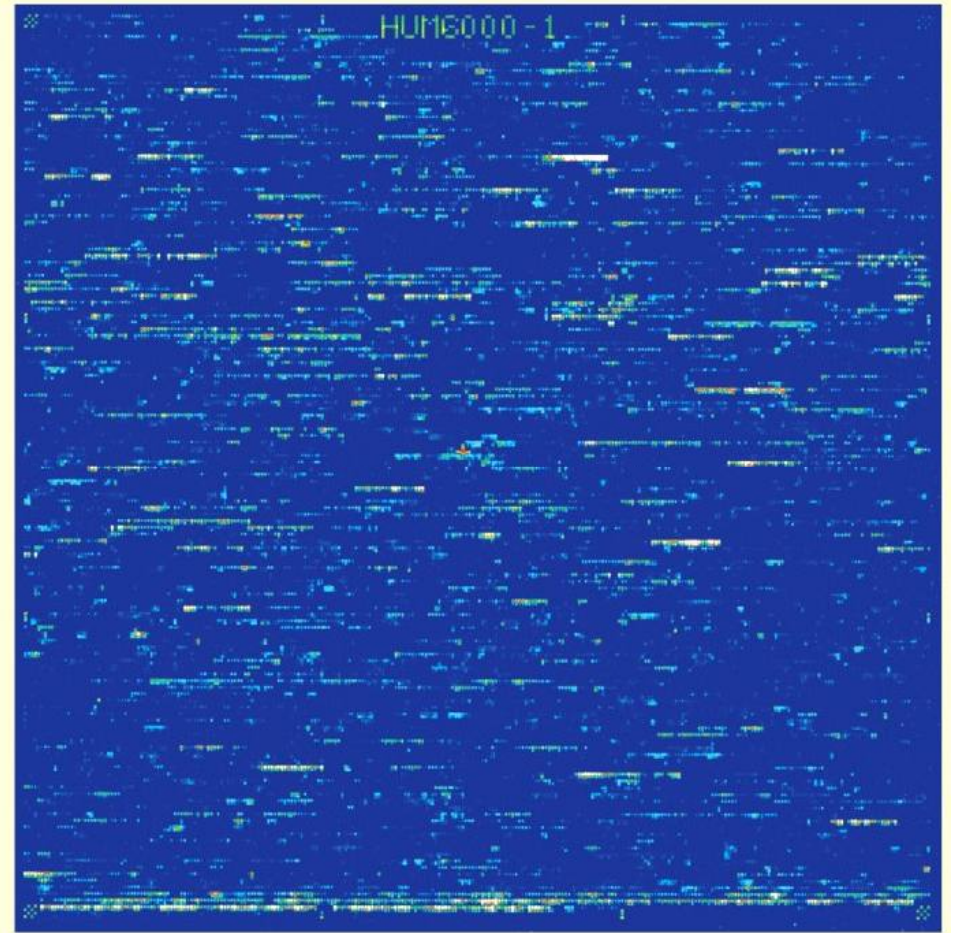
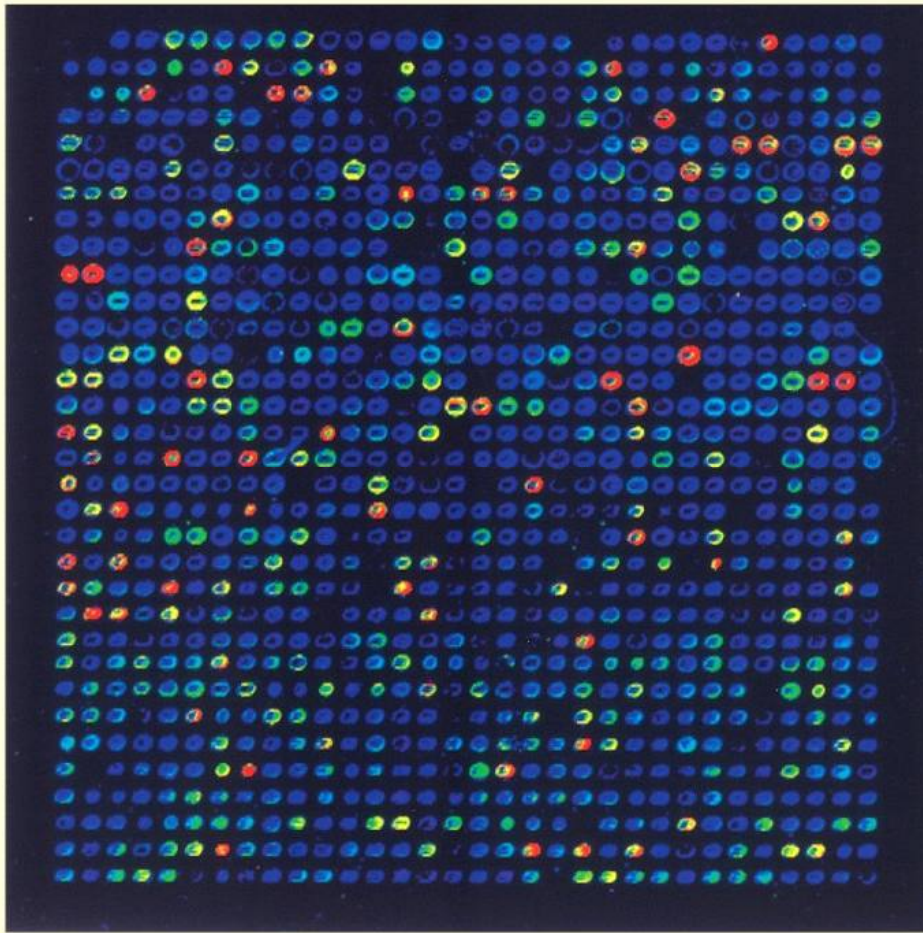
National Human Genome
Research Institute

Genome-scale screening

Many different ways

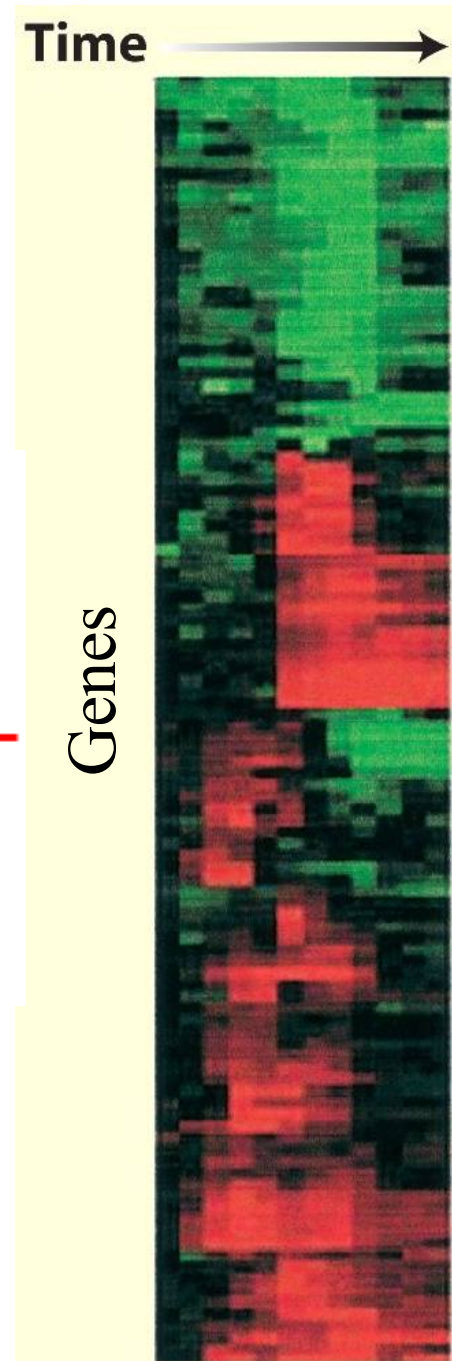
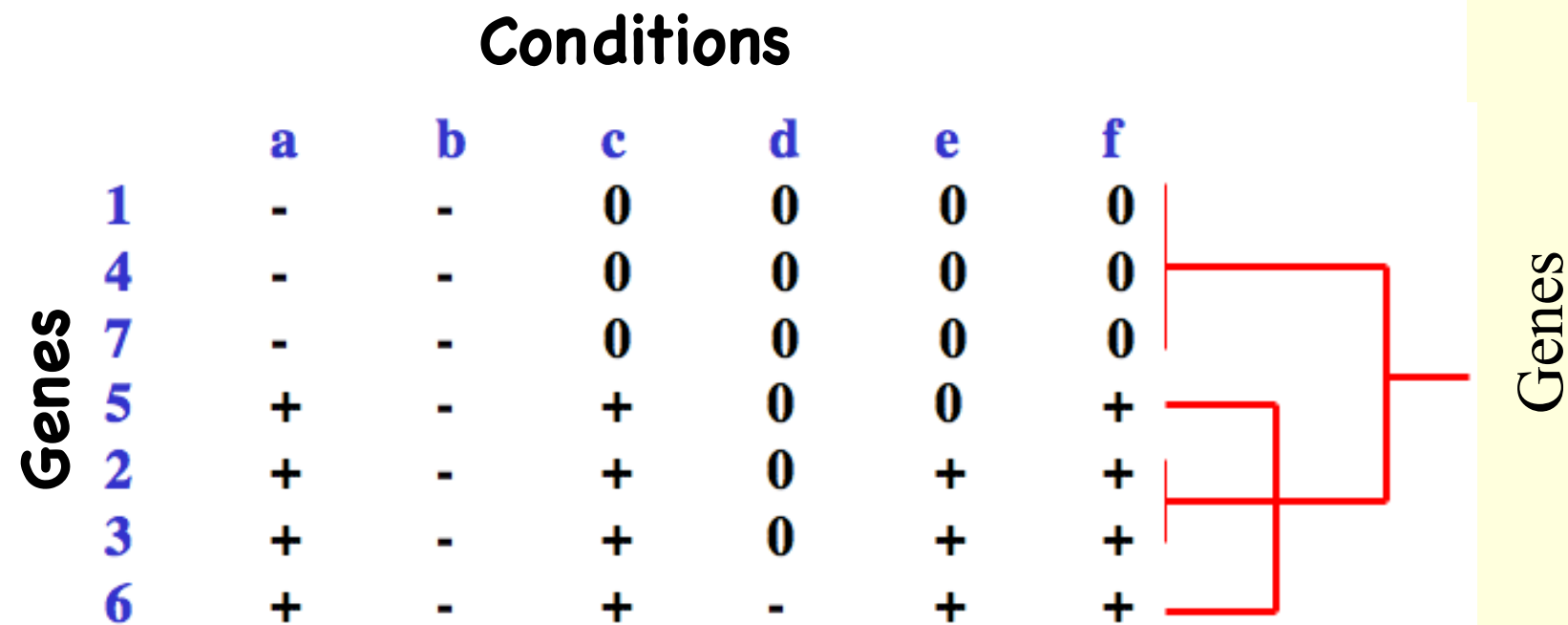


**Two different sample RNAs. One is labeled with a red dye.
The other with a green dye. Spot color provides an indication
Of how the expression of many genes changes with conditions**



Transcriptional profiling

Genes that show similar expression patterns are likely to be Regulated through similar mechanisms (binding sites, Transcription factors, etc).

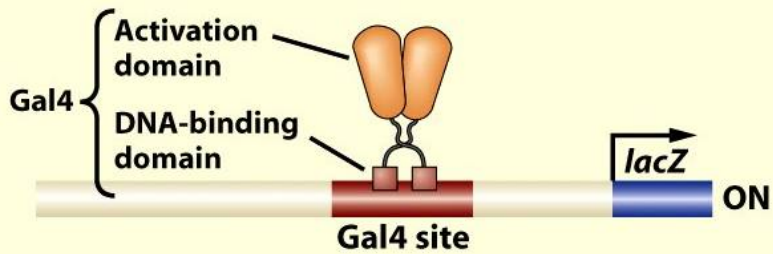


Gene expression pattern clustering algorithm

Compare normal and disease states

Identifying protein-protein interactions using the yeast two-hybrid system

(a) The complete Gal4 dimer



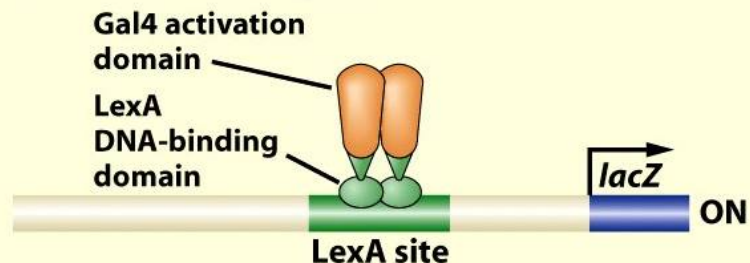
(b) Gal4 lacking the activation domain



(c) LexA lacking the activation domain

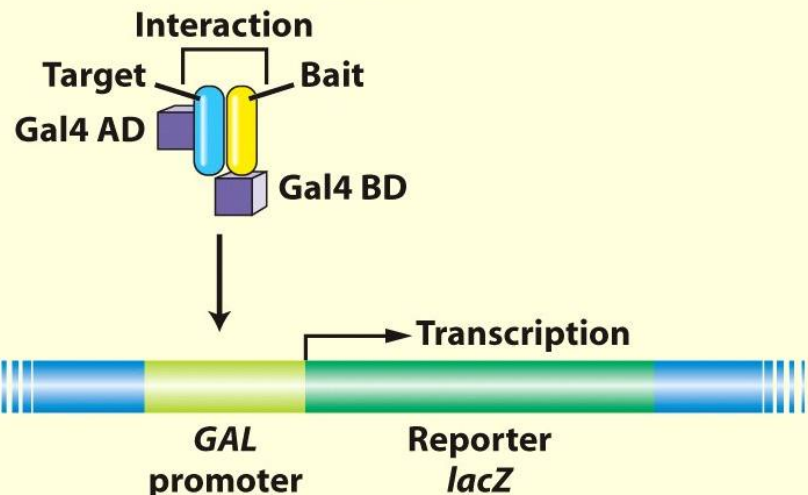
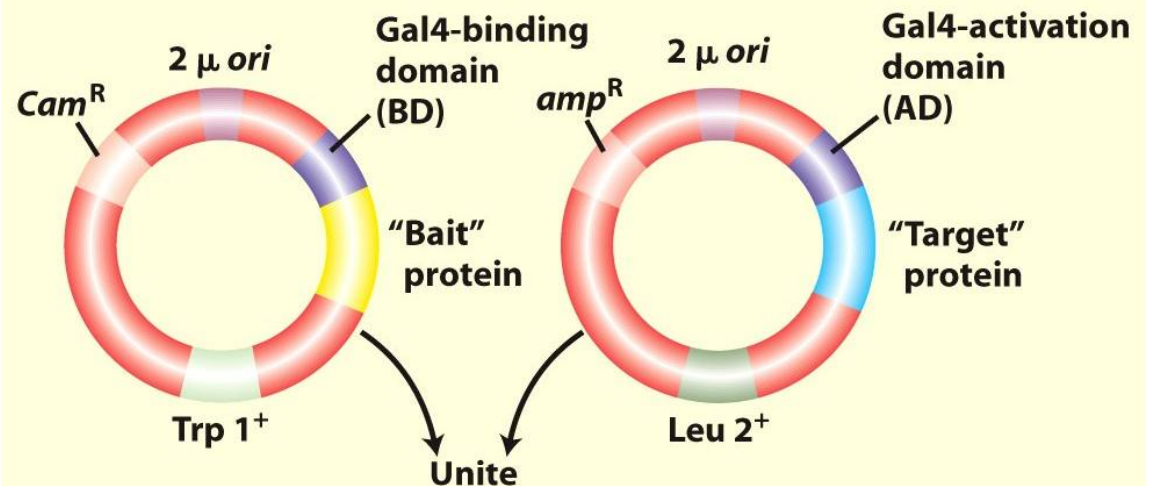


(d) Gal4-LexA hybrid

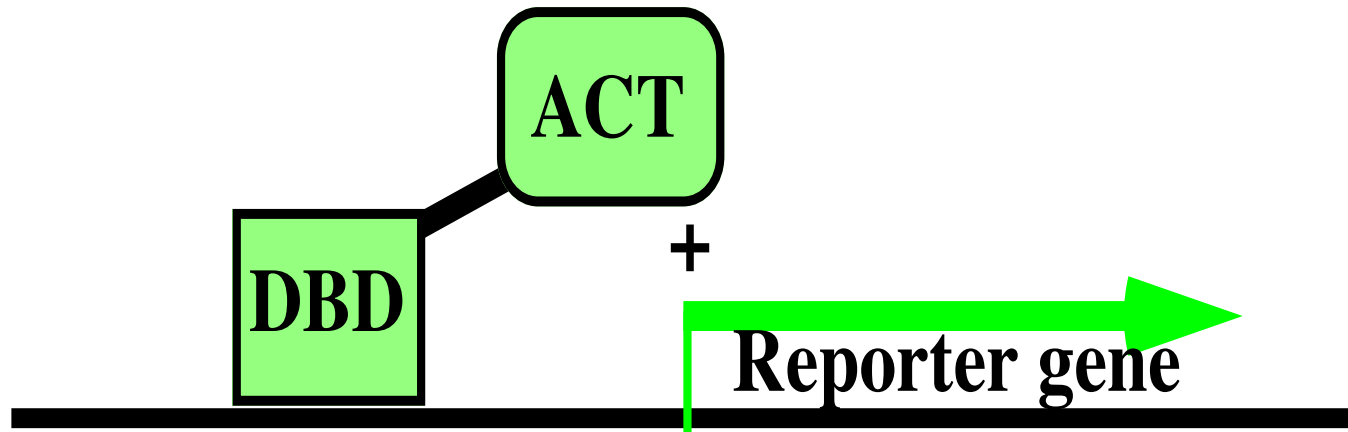


Transform plasmids into cells that are *trp*⁻ and *leu*⁻,
On minimal media. Only plasmid bearing cells survive

Yeast two-hybrid vectors



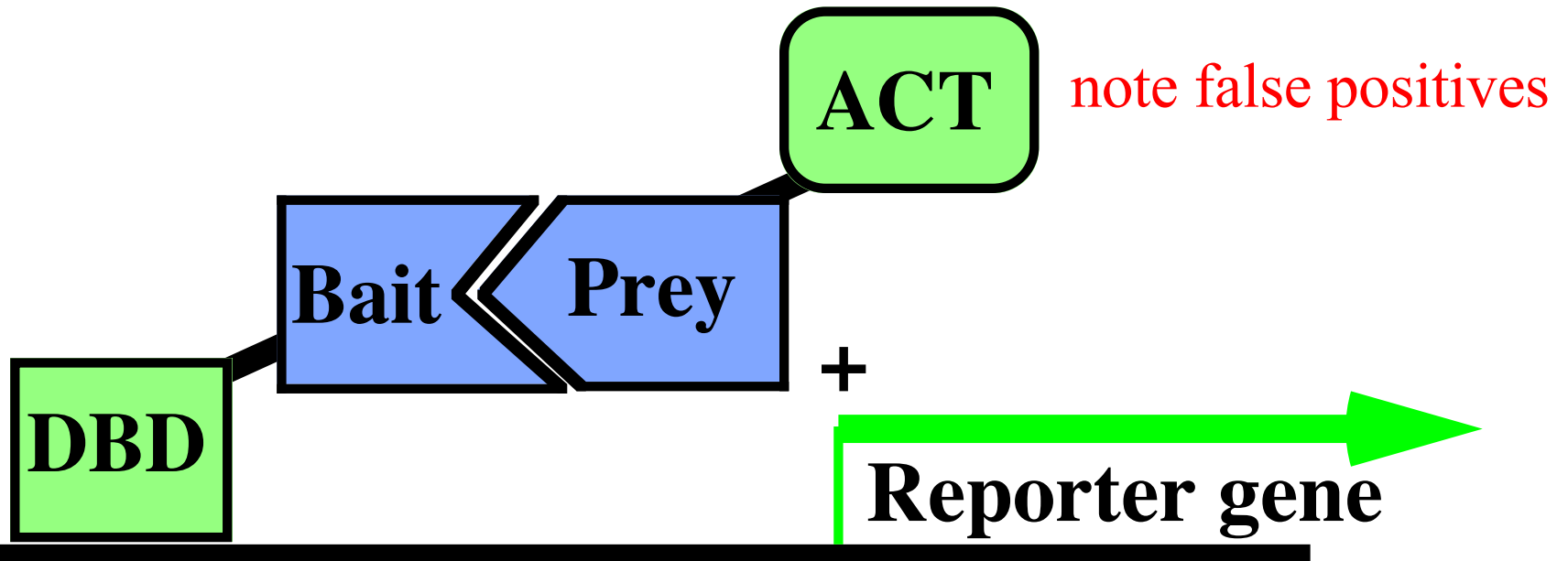
Bi-partite transcriptional activators

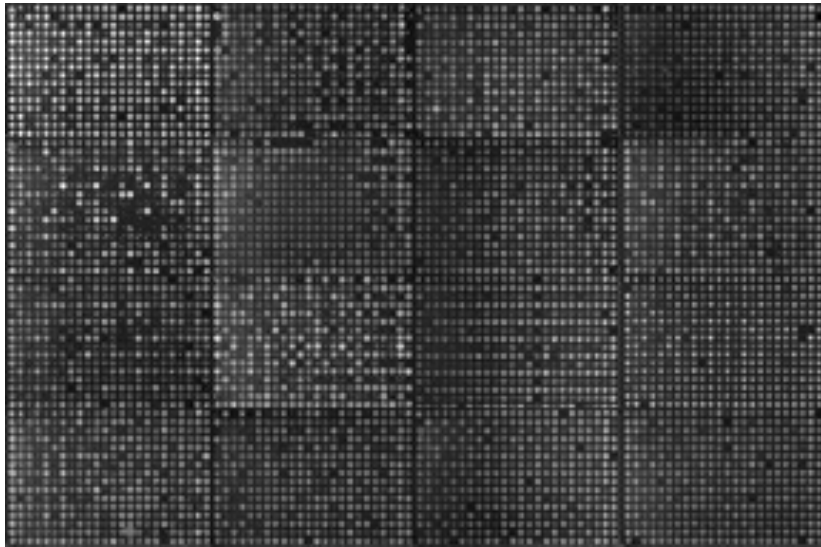


2-Hybrid Assay in yeast

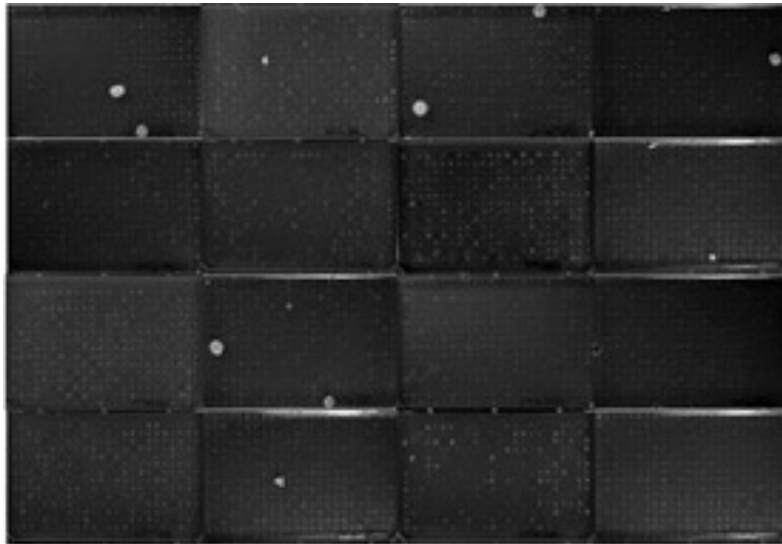
DBD-Bait

Prey-ACT

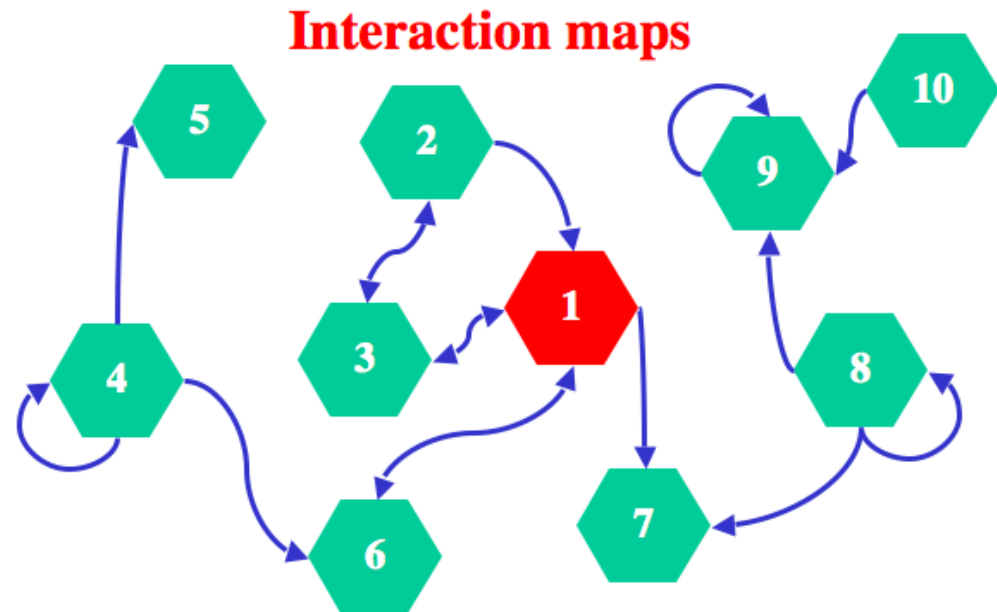




This is the activation domain array consisting of 6144 yeast colonies each expressing a different fusion of the GAL activation domain and one of the yeast ORFs.



This is a set of selective plates (-Leu -Trp-His -Ade) on which only colonies grow that contain RPC19-interacting proteins:





2006

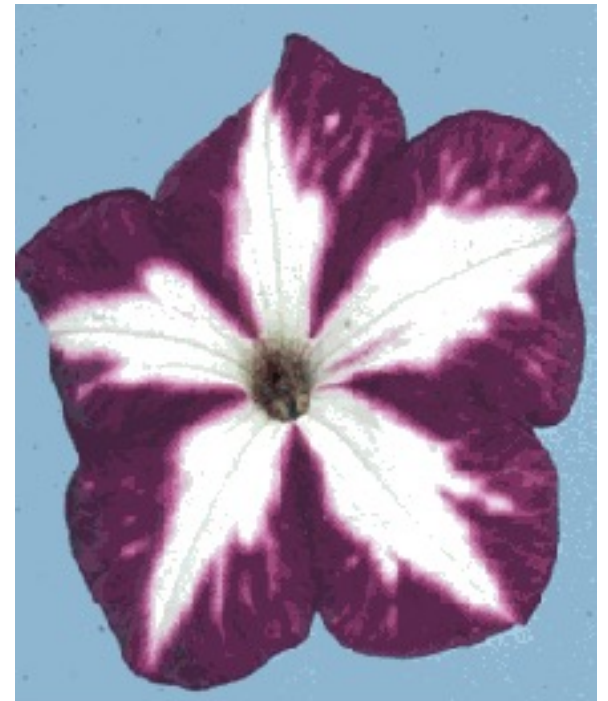


RNA interference

Phenomena first observed in petunia

**Attempted to overexpress chalcone synthase
(anthrocyanin pigment gene) in petunia.
(trying to darken flower color)**

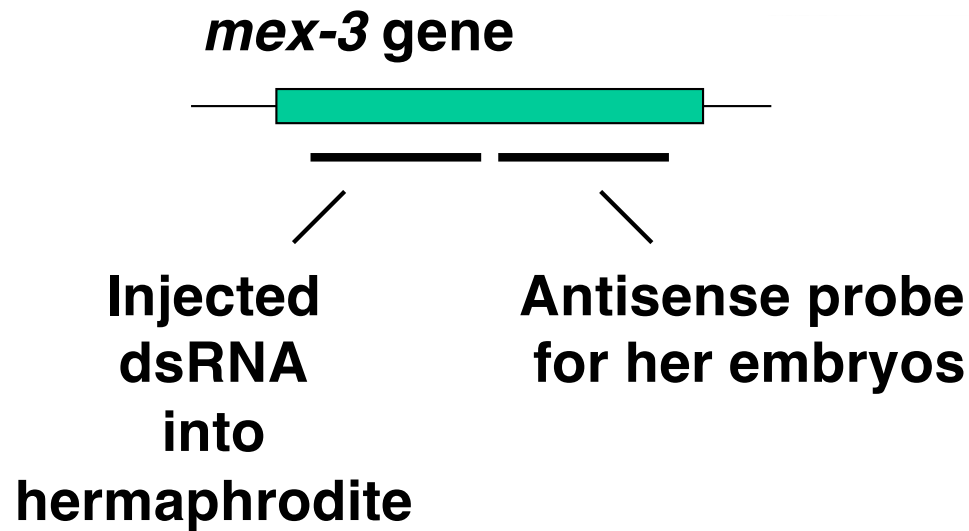
Caused the loss of pigment.



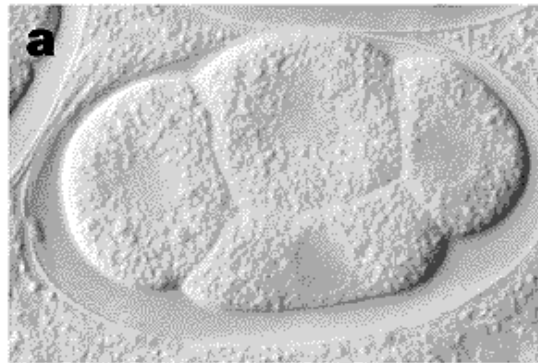
**Called co-suppression
because suppressed
expression of both
endogenous gene and
transgene.**

Injection of dsRNA corresponding to a gene of interest results in the degradation of that transcript

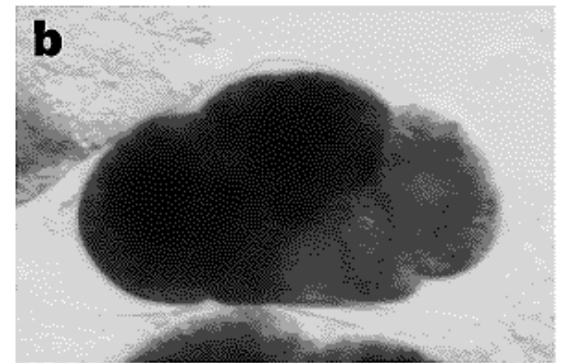
Embryogenesis in the worm *C. elegans*



No probe



mex-3 probe: wild type



mex-3 probe: dsRNA treatment

General mechanism of RNAi

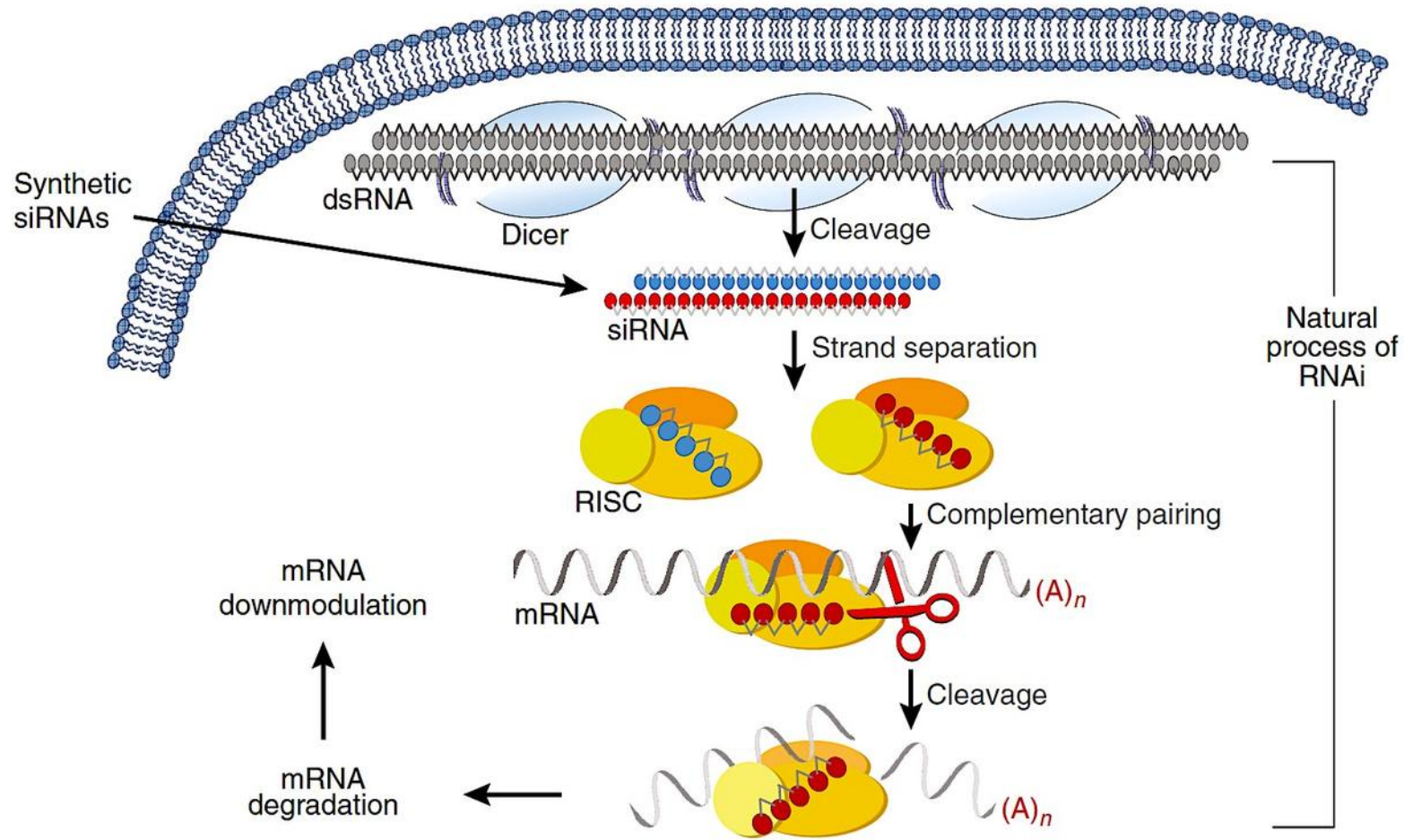


Figure 1 Cellular mechanism of RNA interference. Long double-stranded RNA (dsRNA) is cleaved, by the enzyme Dicer, into small interfering RNA (siRNA). These siRNAs are incorporated into the RNA-induced silencing complex (RISC), where the strands are separated. The RISC containing the guide or antisense strand seeks out and binds to complementary mRNA sequences. These mRNA sequences are then cleaved by Argonaute, the enzyme within the RISC responsible for mRNA degradation, which leads to mRNA down-modulation. A, adenosine.

WHY RNAi?

RNAi components also involved in:

Transposon silencing

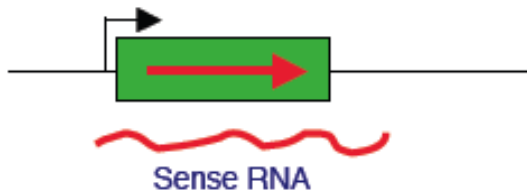
Viral defense

Gene regulation

How would a transposon trigger RNAi?

- Different copies of the same transposon can exist in forward and reverse orientations in the genome
- Expression of the transposon DNA (e.g. transposase gene) can come under the control of other regulatory sequences in the genome regardless of the orientation of the transposon
- Although rare, this scenario can lead to transcription of both sense and antisense RNA transcripts for the transposase gene
- This could lead to dsRNA formation, which in turn could trigger the RNAi cascade leading to the degradation of the normal sense transposase mRNA
- Ultimately, this would cause less transposase protein to be generated and as a result significantly limit the amount of future transposition
- Could this be a mechanism that have evolved to protect the genome against excessive transposition of these mobile elements?

Actively transcribed region containing transposon



Silenced region containing same transposon



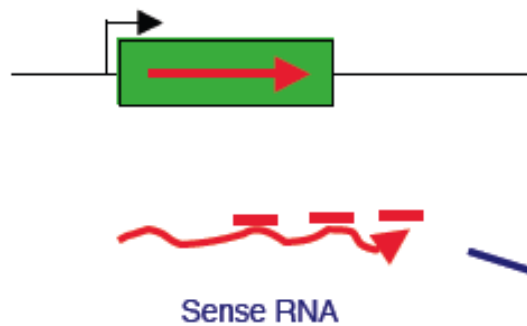
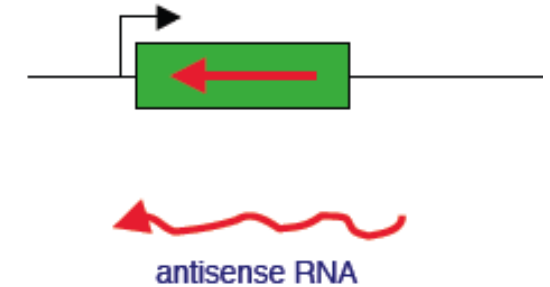
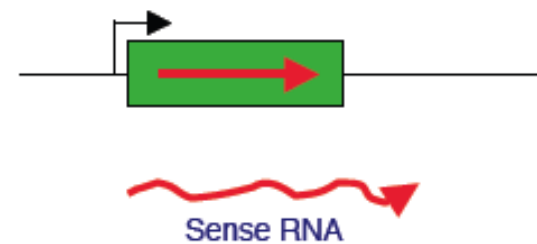
A second actively transcribed region



One copy of transposon
"jumps" to new location
in reverse order

RNA hybridization
& formation of dsRNA

Trigger RNAi to eliminate
transposon transcript



Translation of transposase
is blocked

RNA viruses make dsRNA as replication intermediates. These are targets for an RNAi-based immune system

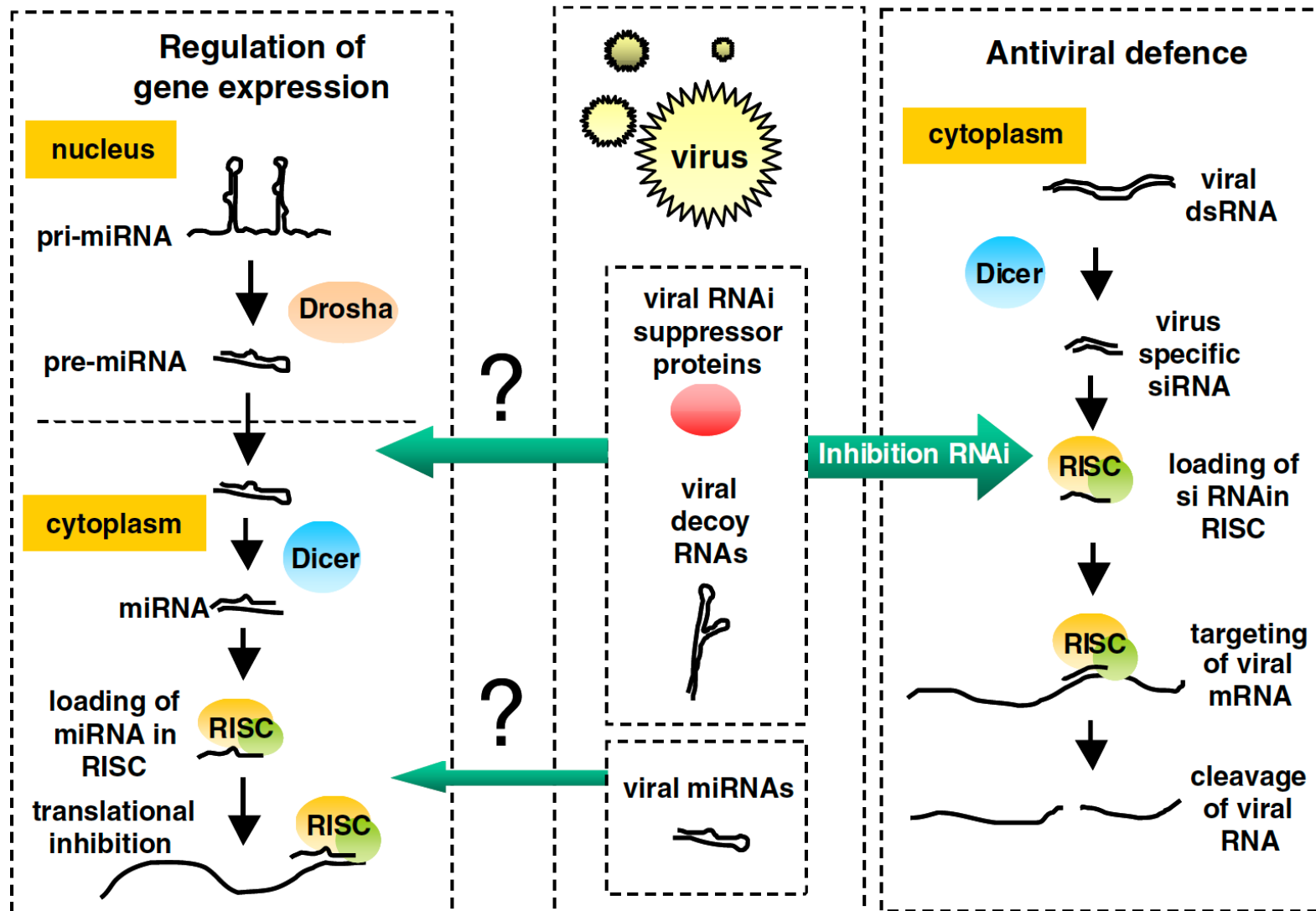
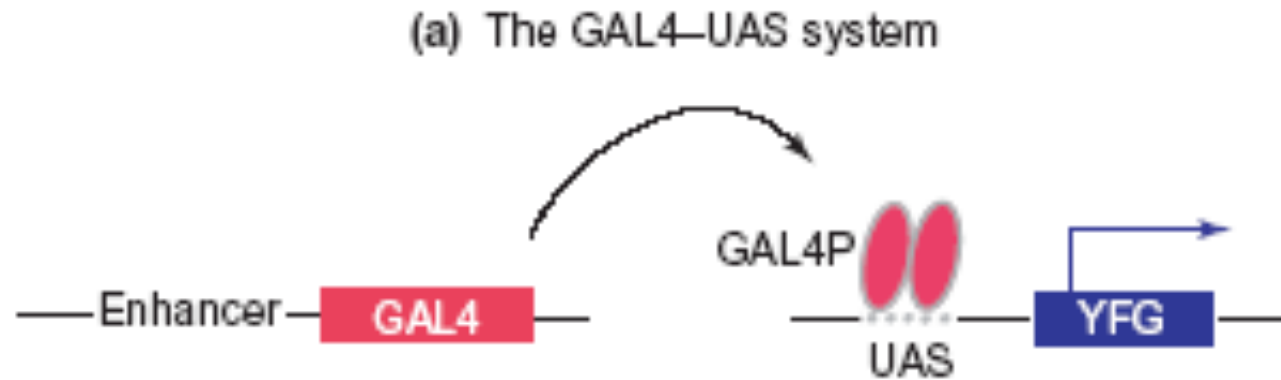


Fig. 1. Viral factors affecting cellular RNAi pathways. miRNA regulated gene expression starts with pri-miRNAs in the nucleus, which are processed into mature miRNAs by Drosha and Dicer (left panel). The antiviral RNAi response is triggered by virus-derived dsRNAs during infection (right panel). Viral RNAi suppressor factors (proteins or decoy RNAs; middle panel) counter these effects. Viruses can also encode miRNA-like molecules targeting cellular mRNAs. Both RNAi suppressors and viral miRNAs can potentially affect cellular miRNA processing and function (arrows with question mark).

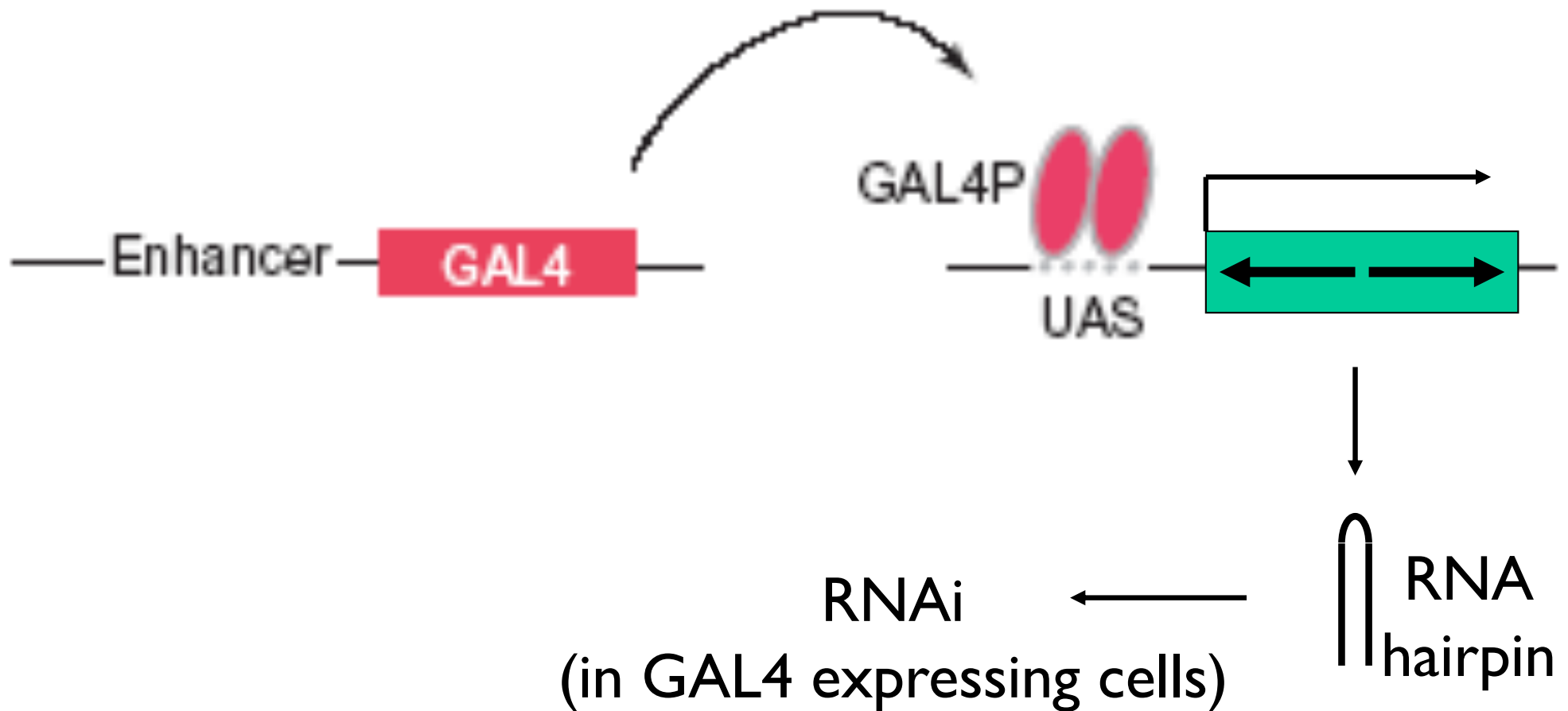
Remember the bipartite GAL4-UAS system



- GAL4 is a transcriptional activator from yeast that recognizes a DNA sequence called the UAS (upstream activating sequence)
- We can use this to control expression of YFG in a tissue specific manner by using enhancer elements specific for the tissue we are interested in

Can express an inverted repeat representing sequences of any gene of interest.

(a) The GAL4-UAS system



LASKER AWARDS 2008



Albert Lasker Basic Medical Research Award >



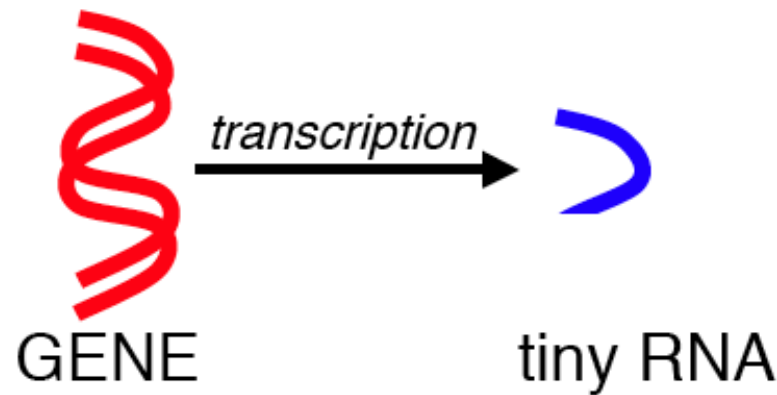
[Victor Ambros](#) [David Baulcombe](#) [Gary Ruvkun](#)

For discoveries that revealed an unanticipated world of tiny RNAs that regulate gene function in plants and animals.

[Watch a video about the 2008 Basic Award winners](#)

Nobel, 2024 to Ambros and Ruvkun

Gene expression regulated by a tiny RNA!



The RNA

**The *C. elegans* Heterochronic Gene *lin-4*
Encodes Small RNAs
with Antisense Complementarity to *lin-14***

and

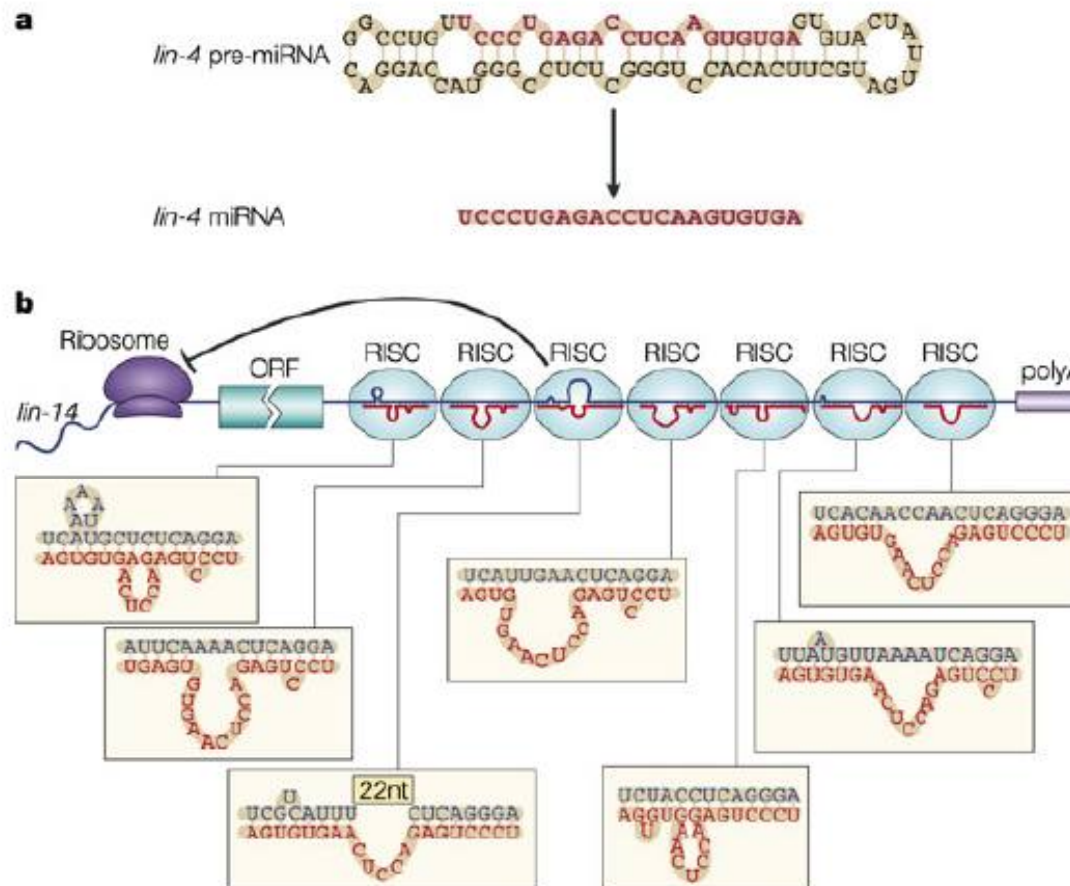
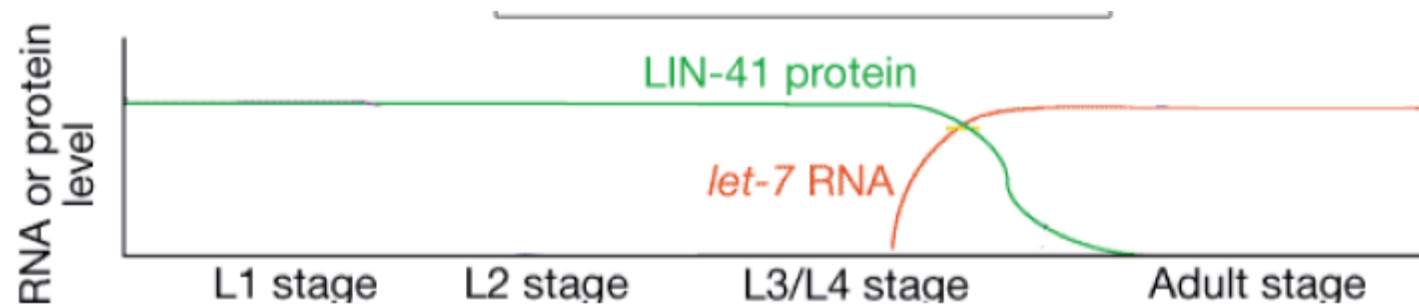
Rosalind C. Lee,^{*,†} Rhonda L. Feinbaum,^{*,†}
and Victor Ambros[†]
Harvard University
Department of Cellular and Developmental Biology
Cambridge, Massachusetts 02138

its Target

**Posttranscriptional Regulation
of the Heterochronic Gene *lin-14* by *lin-4*
Mediates Temporal Pattern Formation in *C. elegans***

Bruce Wightman,^{*,†} Ilho Ha,^{*} and Gary Ruvkun^{*}
Department of Molecular Biology
Massachusetts General Hospital
Boston, Massachusetts 02114

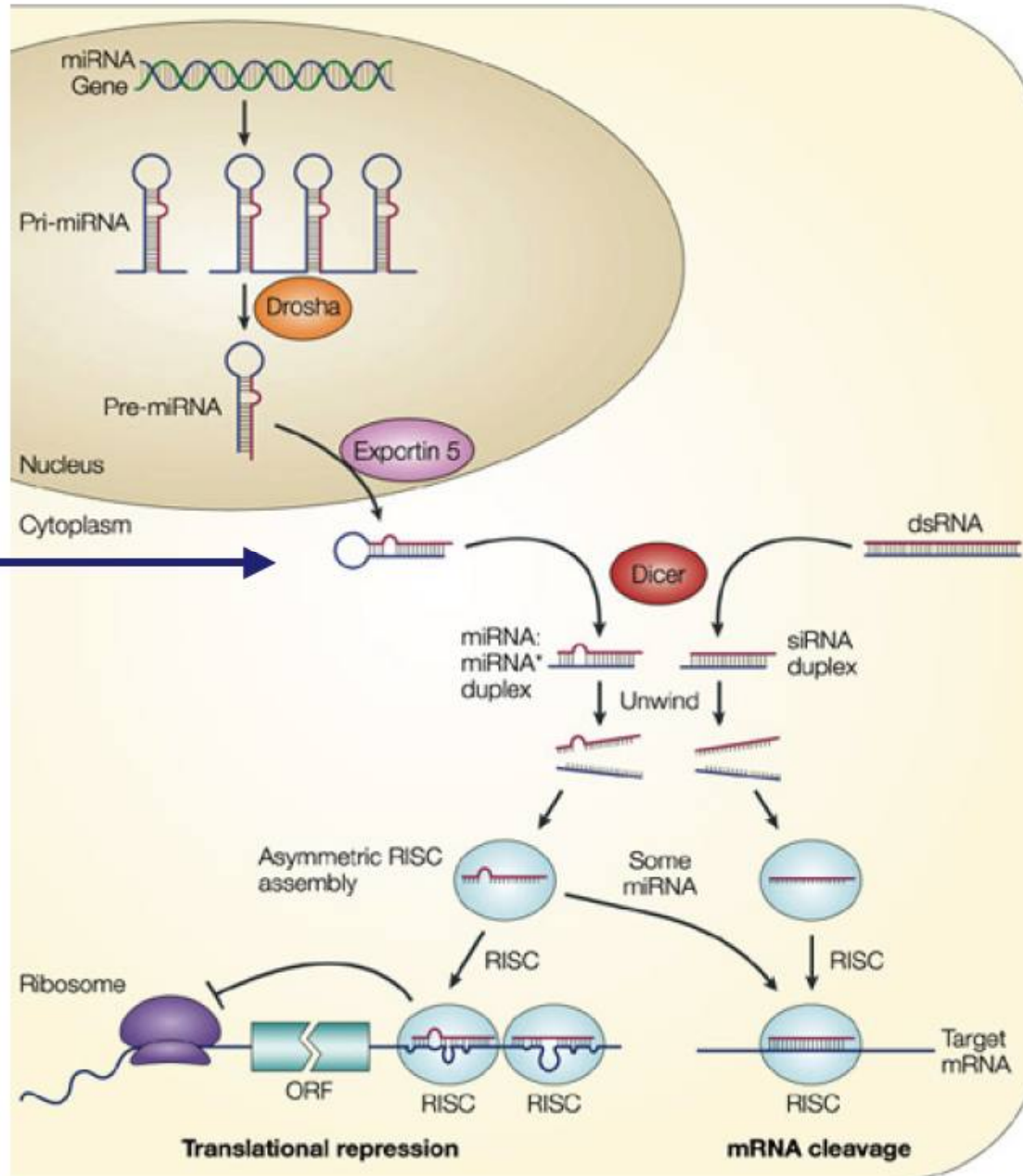
Cell, 1993



The same processed miRNA sequence (e.g. *lin-4*) can hybridize (not perfectly, but sufficiently) to multiple complementary sites in the 3'UTR of target mRNAs (e.g. *lin-14*)

Processing pri-miRNA transcripts or exogenous dsRNA

Naturally occurring
in the genome



Introduced by the
researcher
experimentally
as shRNA

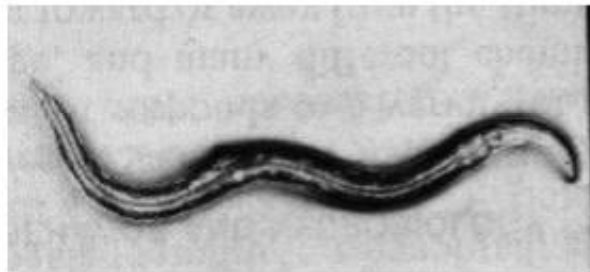
Introduced by the
researcher
experimentally
as dsRNA

MicroRNAs (miRNAs)--Not just in worms!



A. thaliana/
O. sativa

>80 miRNAs



C. elegans

>100 miRNAs



D. melanogaster

>100 miRNAs



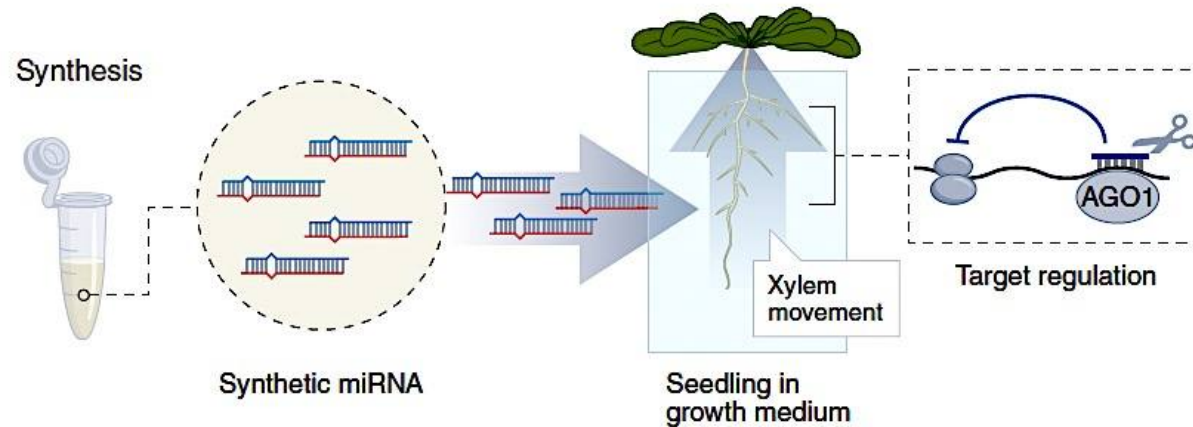
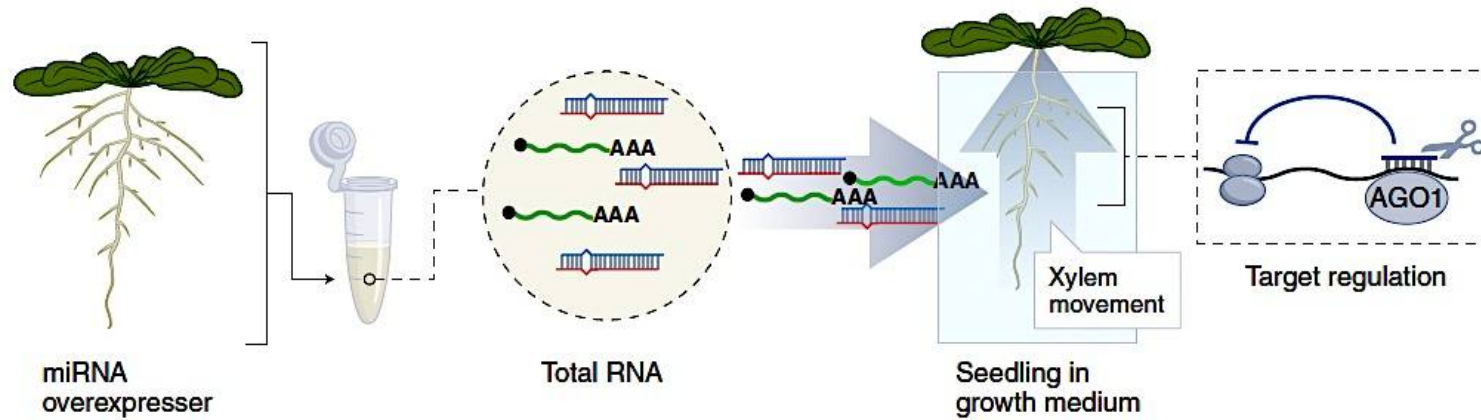
H. sapiens/
M. musculus

>200 miRNAs

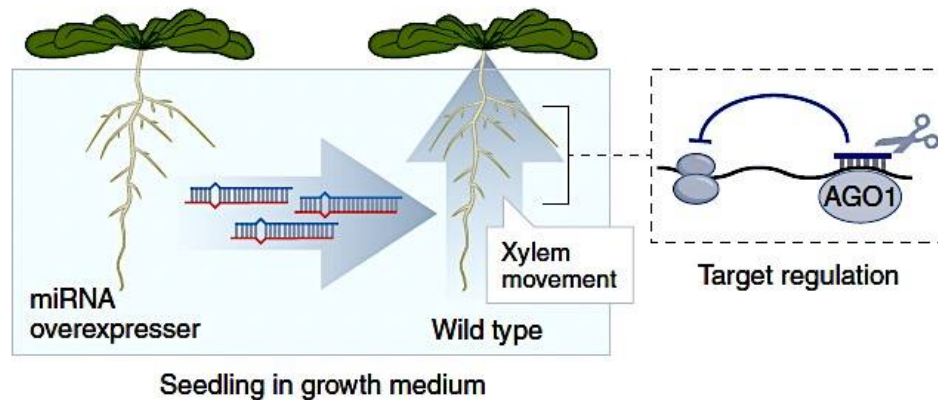
See the RNA Family database - <http://www.sanger.ac.uk/Software/Rfam/>

Expanding horizons: Small RNA movement between plants

a Extraction



b Communication



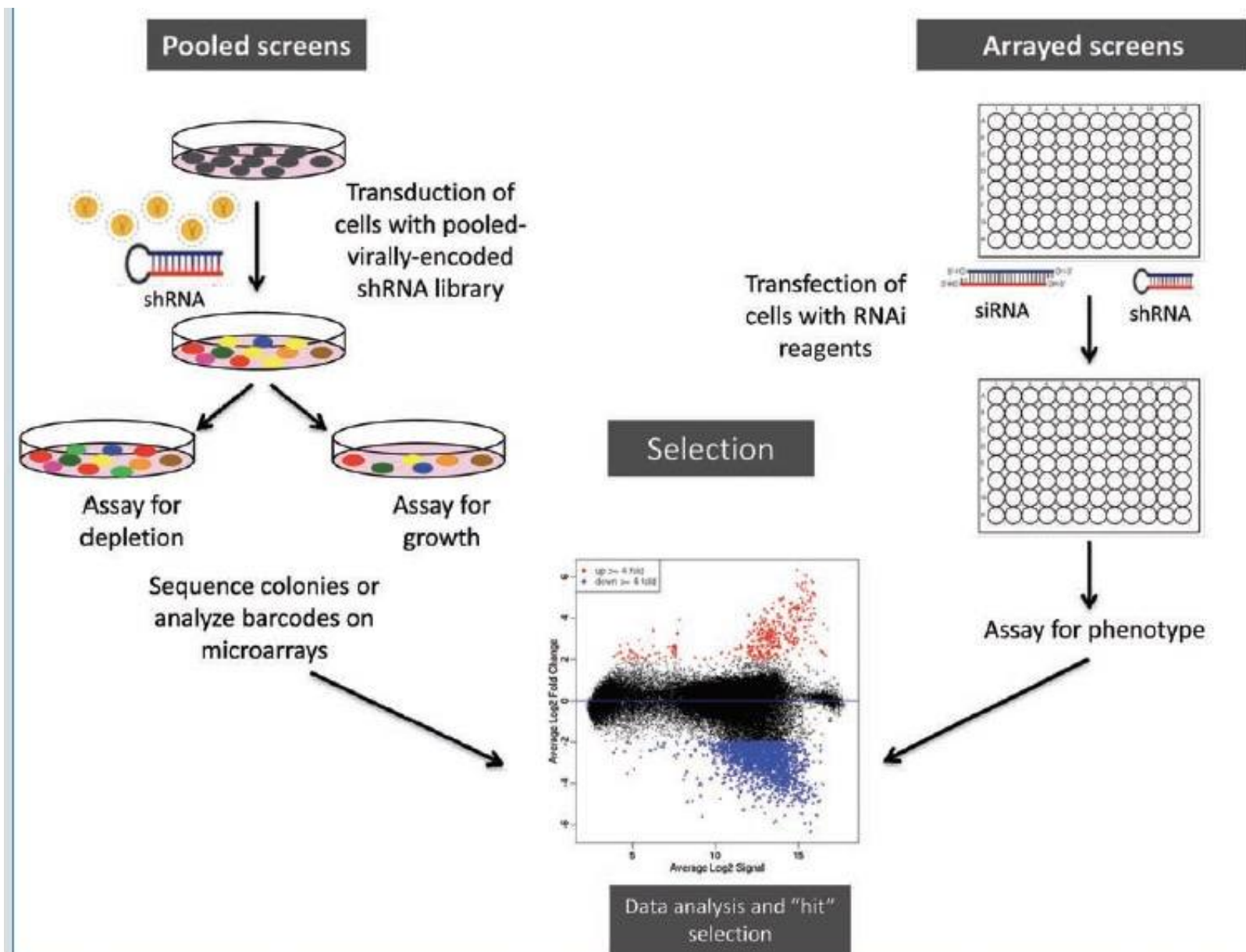


Figure 1 Schematic illustration of arrayed or pooled RNAi screens in cells. **Left panel.** Pooled-viral vectors encoding libraries of shRNAs targeting multiple genes can be used to transduce a target cell population in a single tissue culture dish. After selection for the desired phenotype, cells are analysed for the identification of genes whose inhibition by RNAi knockdown cause the specific phenotype as described in Table 1. The relative abundance of each shRNA or a random 60-mer barcode expressed from the same vector as the specific shRNA can be identified and quantified by labelling the PCR product with fluorescent dyes (e.g., Cy5 or Cy3). The PCR products are then hybridised to custom designed cDNA microarrays containing barcode or shRNA complementary oligonucleotides. The relative abundance of barcodes obtained from the cells that were exposed to selective pressure are compared to that detected in control cells that have been exposed to the same shRNA library, but not to the selective pressure (for example, drug treatment or genetic mutations). **Right panel.** Arrayed RNAi screen libraries consist of individual siRNA or shRNA reagents that target different genes and that are placed in each well of a multi-well plate. RNAi reagent libraries can comprise synthetic siRNAs, plasmid- or virally-encoded shRNAs. Various assay readouts are used to determine the effect of RNAi on the phenotype as described in Table 1. Adapted¹⁹.

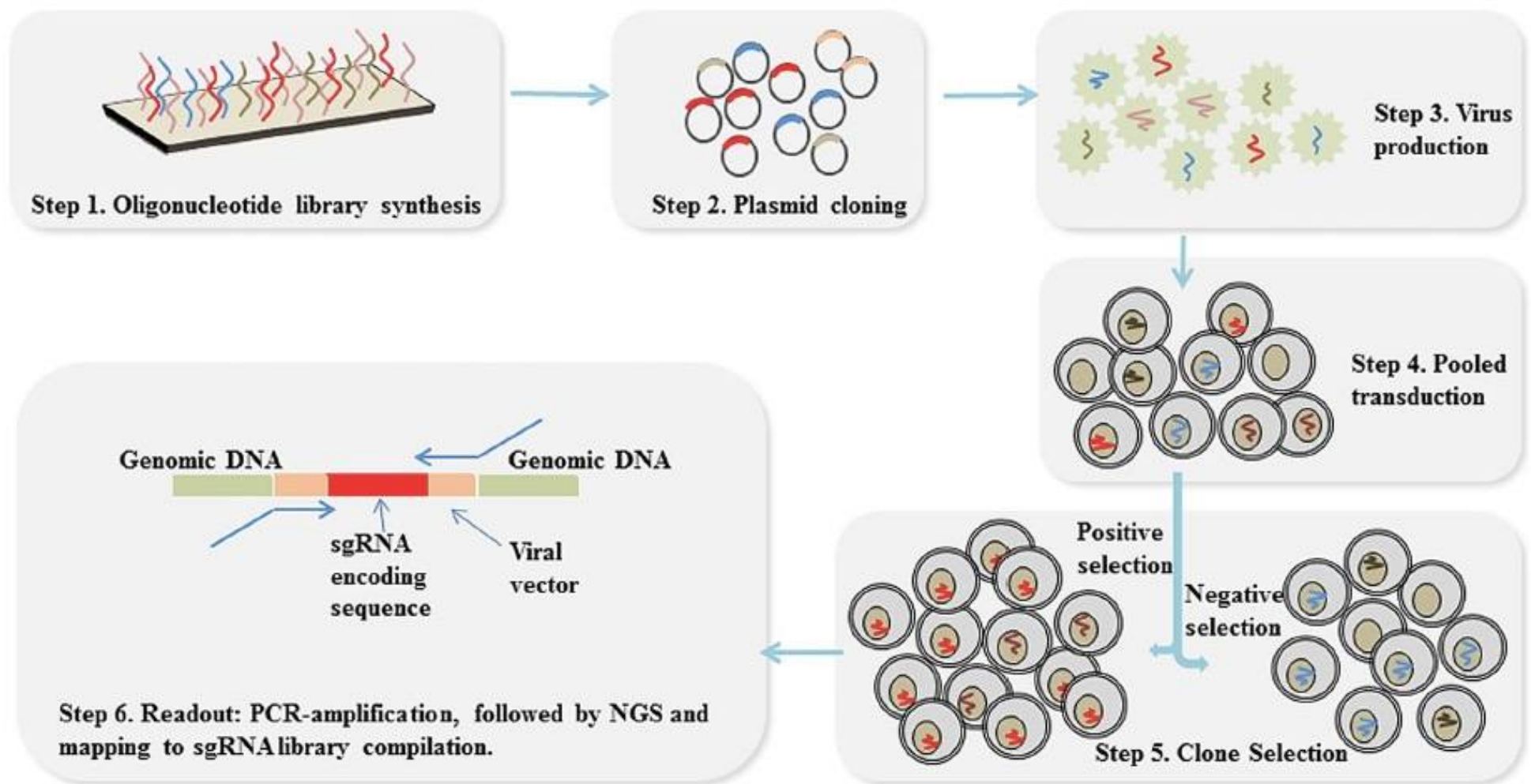
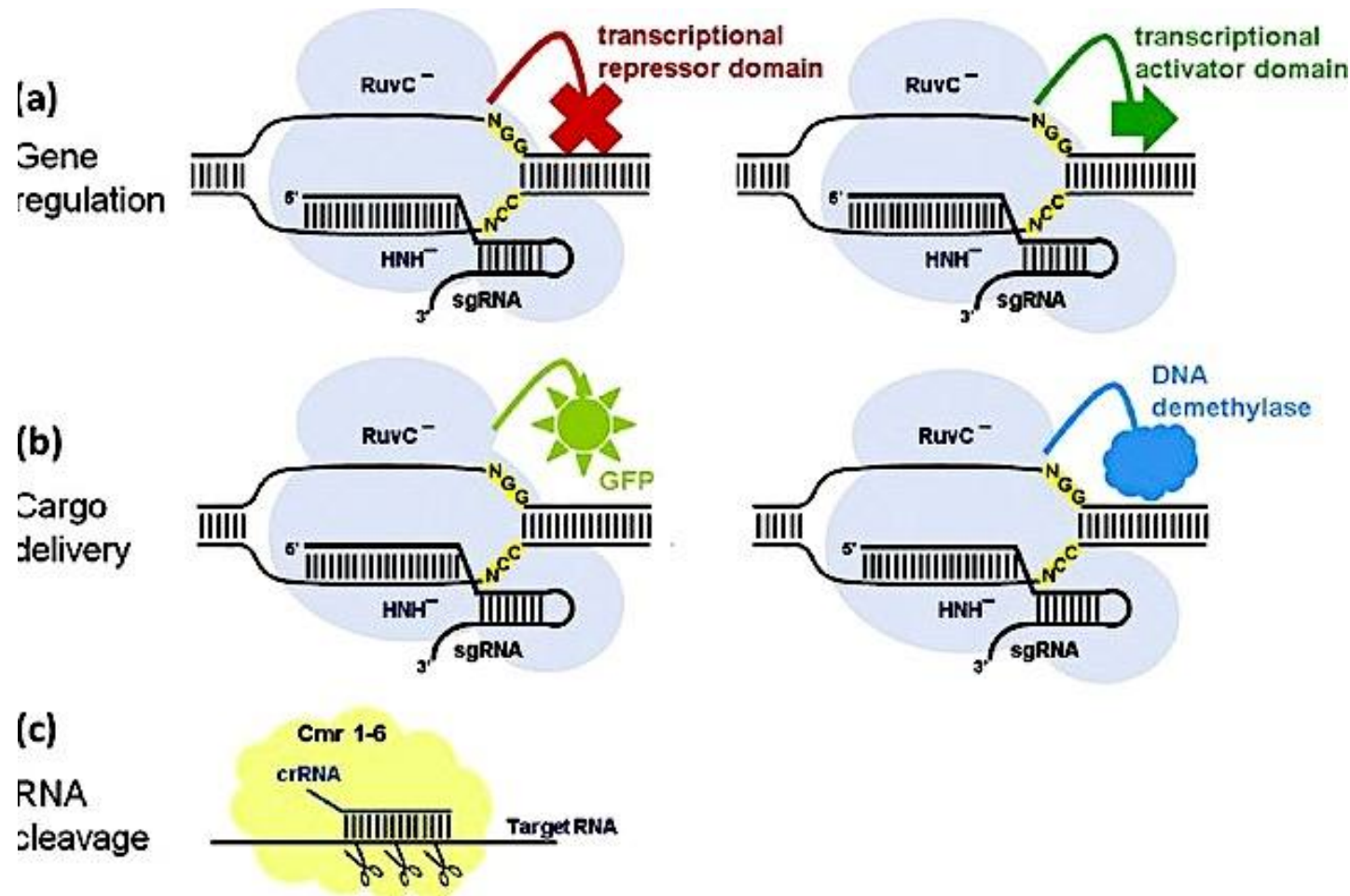


Figure 2 Flow chart of CRISPR (clustered *regularly interspaced short palindromic repeats*) screen in pooled format. sgRNA library is massively synthesised (step 1), cloned into plasmids (step 2) and packaged into viral vectors (step 3). Cas9-expressing cells are infected with these vectors at low multiplicity of infection in order that most cells receive only one virus integration (step 4). After positive or negative selection (step 5), sgRNA-encoding regions are PCR-amplification and sequenced, followed by NGS and mapping to the sgRNA library design to read out sgRNA representation in the selected cell population (step 6). sgRNA: single-guide RNA; NGS, next generation sequencing.



Applications of the CRISPR/Cas system beyond genome editing. In addition to targeted genome editing, the CRISPR/Cas9 technology is suitable for other interesting applications. (a) Gene regulation. A catalytically inactive dead Cas9 (dCas9, in light blue) can be fused to either a transcriptional repressor (left) or activator (right). When the dCas9-repressor fusion is recruited by a gRNA that matches the promoter, 5' untranslated region or coding sequence of an endogenous gene, it can block transcription initiation, elongation or the binding of transcription factors. When the dCas9-activator fusion is targeted to a promoter, the specific expression of the endogenous gene is induced. (b) Cargo delivery. The catalytically inactive dCas9 can also be exploited as a programmable DNA-binding protein to deliver diverse cargos to specific genomic locations. For example, fusion with a green fluorescent protein (left) provides a tool for visualizing chromosome structure or dynamics, and fusion with a demethylase (right) can be used for targeted epigenome editing. (c) RNA cleavage. The Type III-B CRISPR-Cas system (e.g., from *Pyrococcus furiosus*) is composed of nucleases that form the so-called Cmr complex (yellow) and represents a unique RNA silencing system. The Cmr-crRNA complex targets invading RNA in a PAM-independent process and is able to degrade complementary RNA sequences cleaving them at multiple sites.