Key experiments:

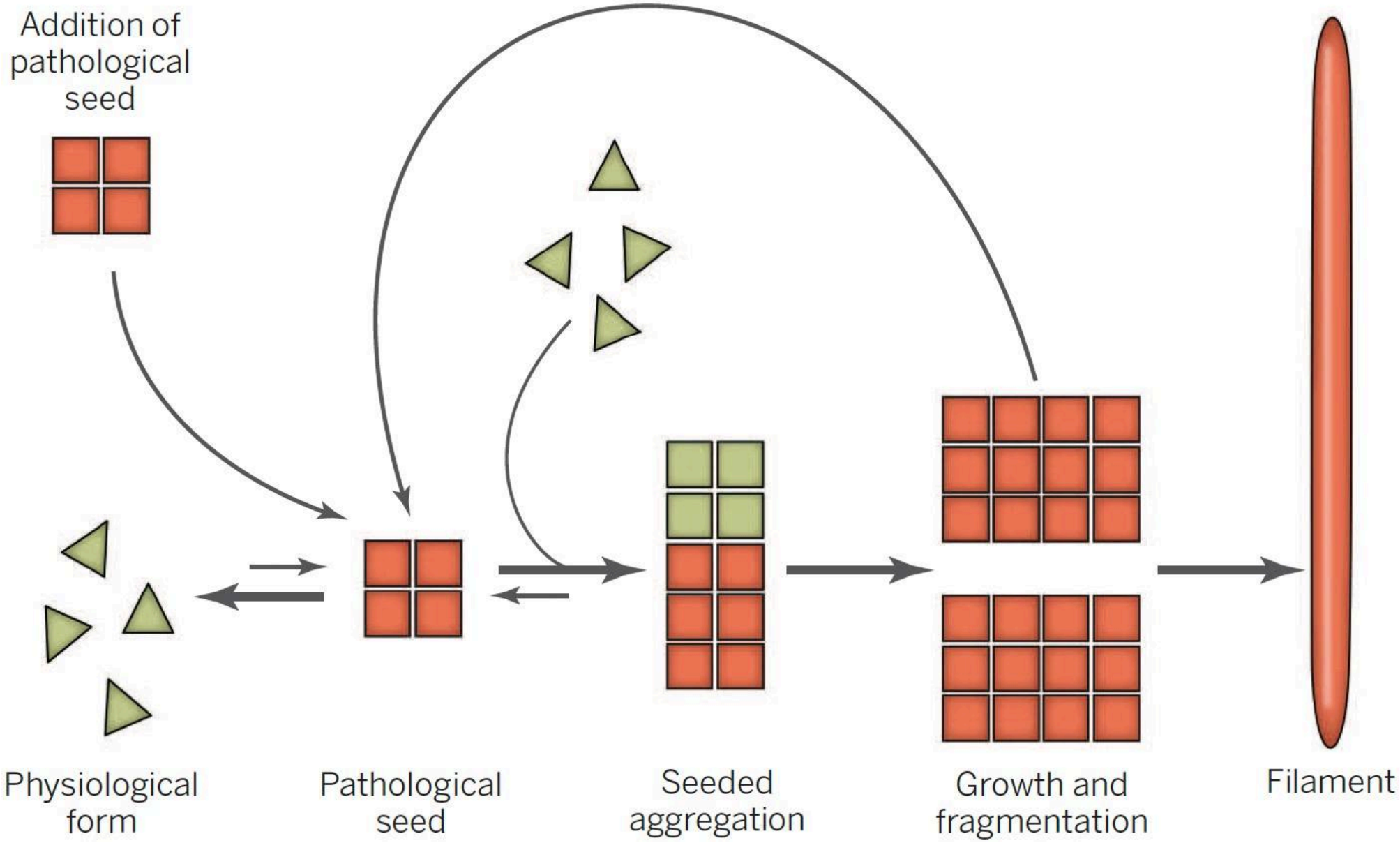
--Genetic: autosomal dominant

--Mice that express human PrP result in disease with the characteristics of the human disease.

--Mice that express PrPSc spontaneously develop scrapie

--Mice that lack PrPC are resistent to infection with PrPSc

The model


Addition of pathological seed
Physiological form
Pathological seed
Seeded aggregation
Growth and fragmentation
Filament

Prions in yeast behave as dominant traits, but with non-Mendelian inheritance

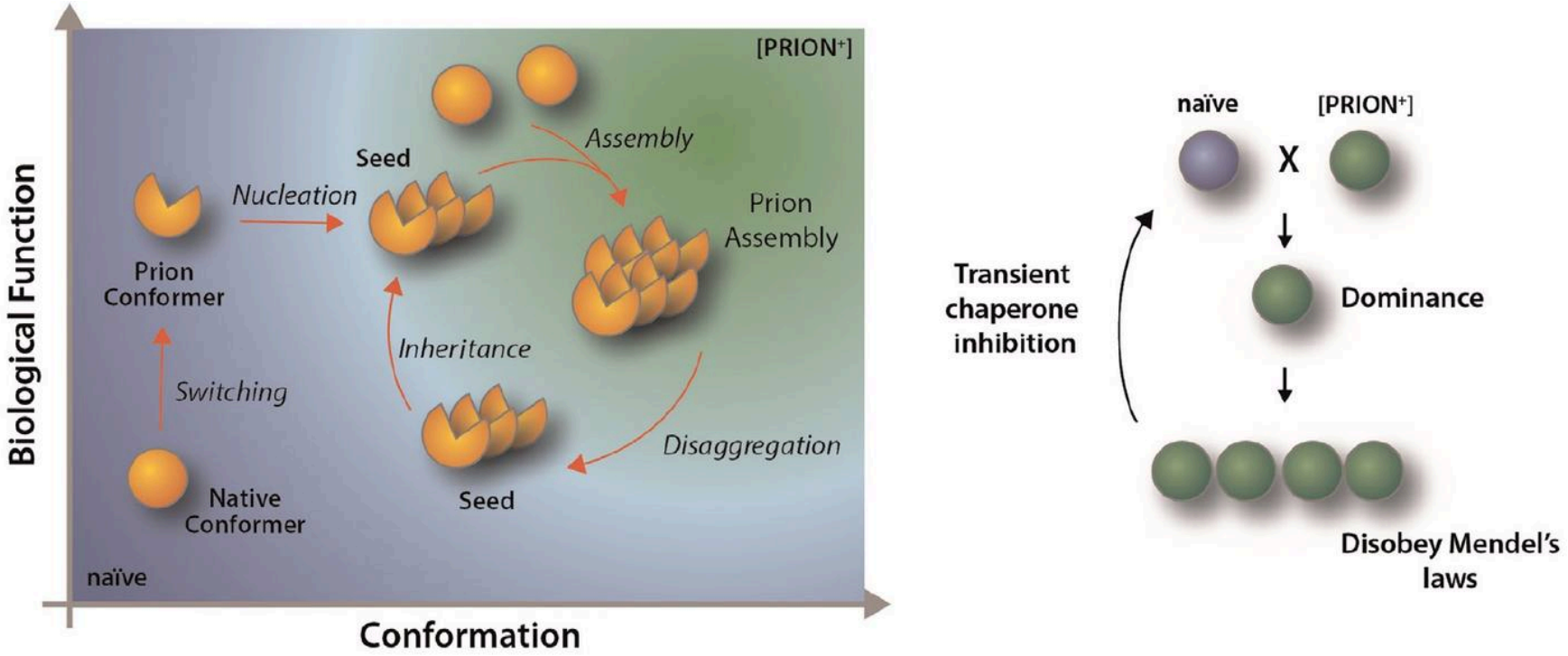


Fig. 2. Molecular and genetic hallmarks of prions. (Left panel) The *x*-axis represents the conformational space explored by a prion protein. On this axis, the multimeric prion assembly and native conformer depict two distinct and stable structural states of the same prion polypeptide. On the *y*-axis, the functional states of these two distinct structural states are mapped. The functional difference in these two distinct and stable structural states manifests as separate naïve and [*PRION*$^+$] traits, depicted in blue and green. The mechanism of prion formation and transmission to progeny fueling such epigenetic inheritance involves disaggregation by molecular chaperones to generate propagons that can be passed to the next generation. (Right panel) Accessing two distinct phenotypic states (naïve in blue and [*PRION*$^+$] in green) from the same polypeptide engenders distinctive behavior in genetic crosses. Prion-based traits are transmissible via cytoplasmic transfer, dominant in diploids, and passed to meiotic progeny in a non-Mendelian fashion. Most can also be permanently reverted to a naïve [*prion*$^-$] state from by transient chaperone inhibition. The unusual [*PRION*$^+$] nomenclature for these protein-based genetic elements is derived from these behaviors.

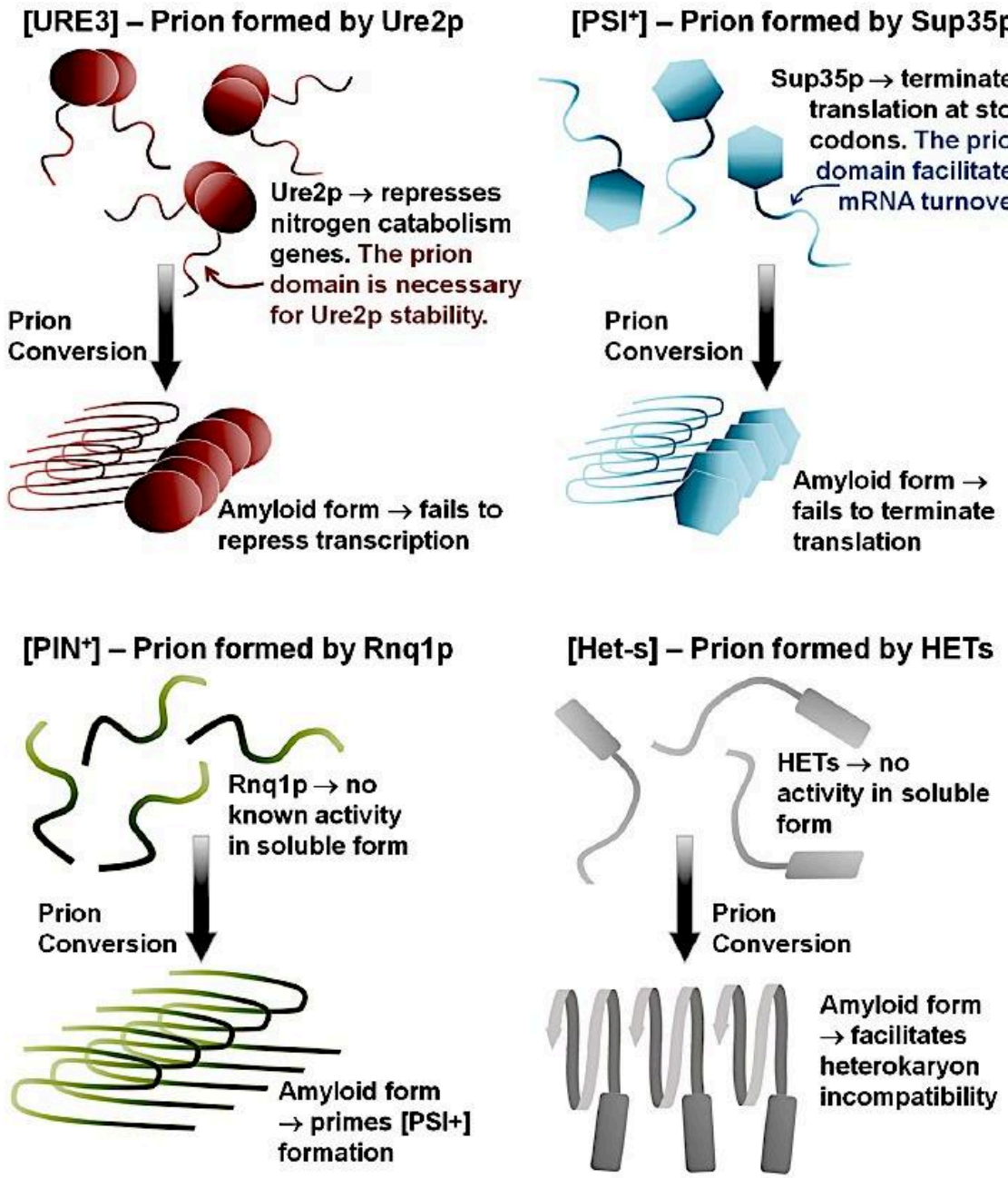
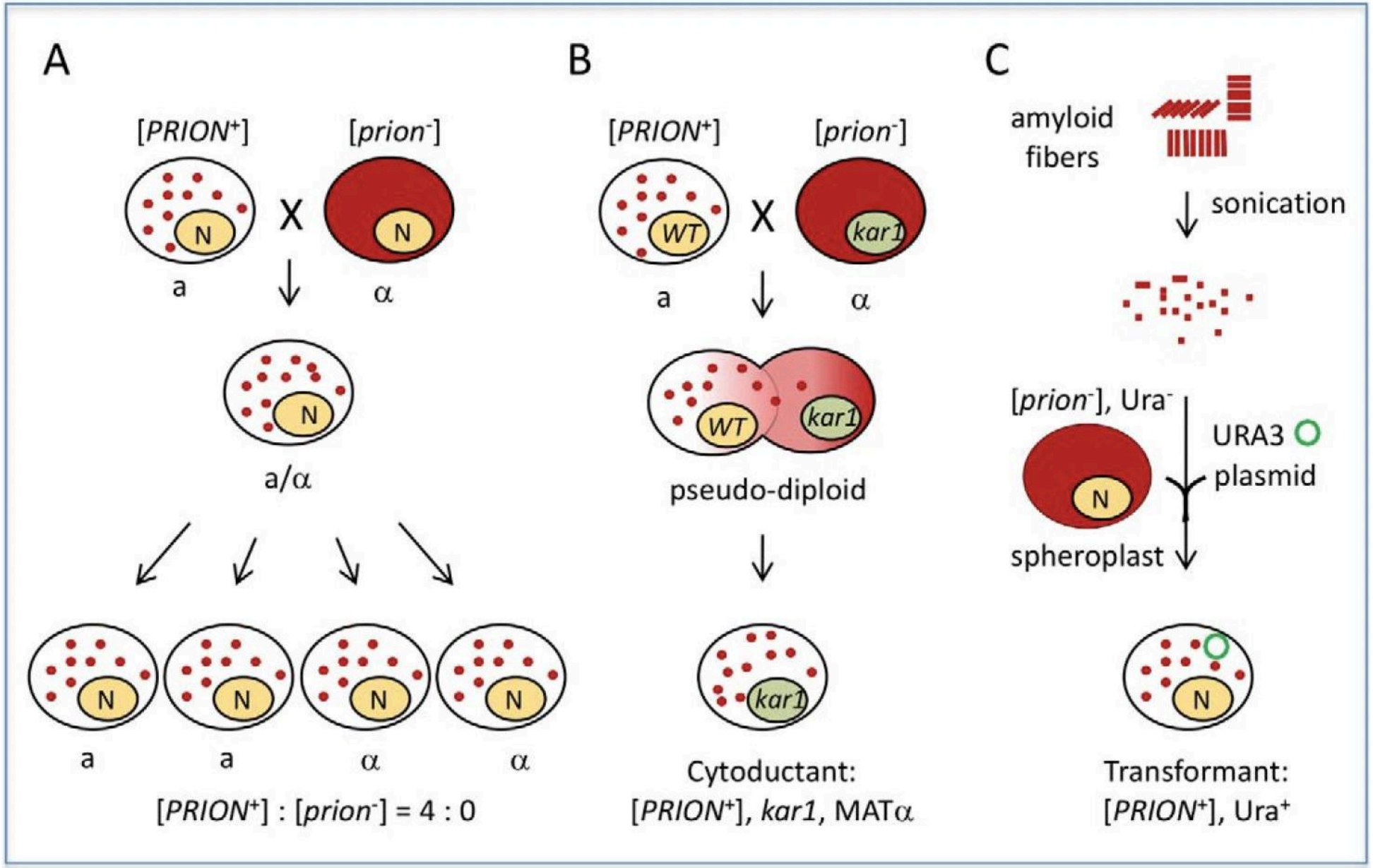


FIG 1 Prions [URE3], [PSI+], and [PIN+] of *S. cerevisiae* and [Het-s] of *Podospora anserina*. These prions are based on self-propagating amyloids of Ure2p, Sup35p, Rnq1p, and HET-s, respectively. The prion domains of Ure2p and Sup35p have nonprion functions, explaining their retention in evolution despite detrimental prion formation.


A
[PRION+]
[prion-]
N
X
N
a
α
N
a/α
N
N
N
N
a
a
α
α
[PRION+] : [prion-] = 4 : 0
B
[PRION+]
[prion-]
WT
X
kar1
a
α
WT
kar1
pseudo-diploid
kar1
Cytoductant:
[PRION+], kar1, MATα
C
amyloid
fibers
sonication
[prion-], Ura-
URA3
plasmid
N
spheroplast
N
Transformant:
[PRION+], Ura+

Prion formation is regulated by cellular signaling.
Heat shock chaperones in this case

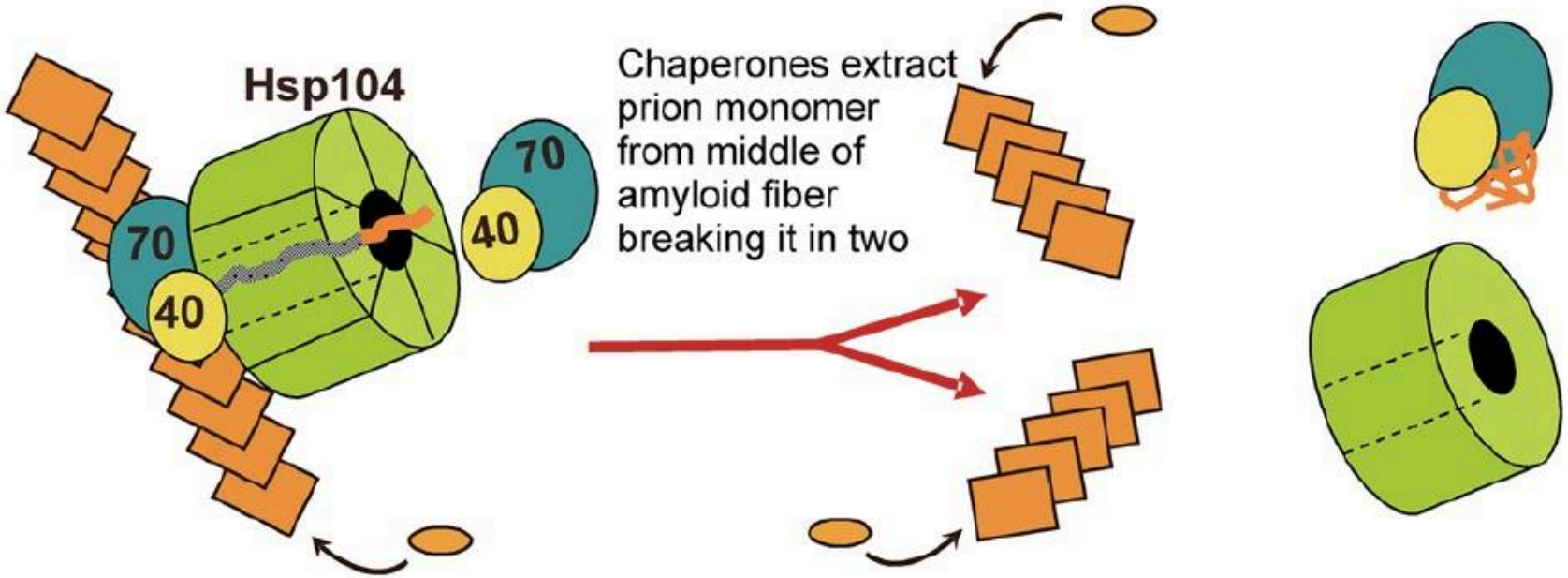


FIG 5 Mechanism of seed generation by Hsp104-Hsp70-Hsp40. Hsp104, working with Hsp40 (mainly Sis1p) and Hsp70 (Ssa proteins), extracts a monomer from an amyloid filament, resulting in the formation of two filaments, with two more growing ends. (Adapted from reference 217.)

Intrinsically disordered proteins/domains and prion formation.

The creation of unique epigenetic states that are reversible

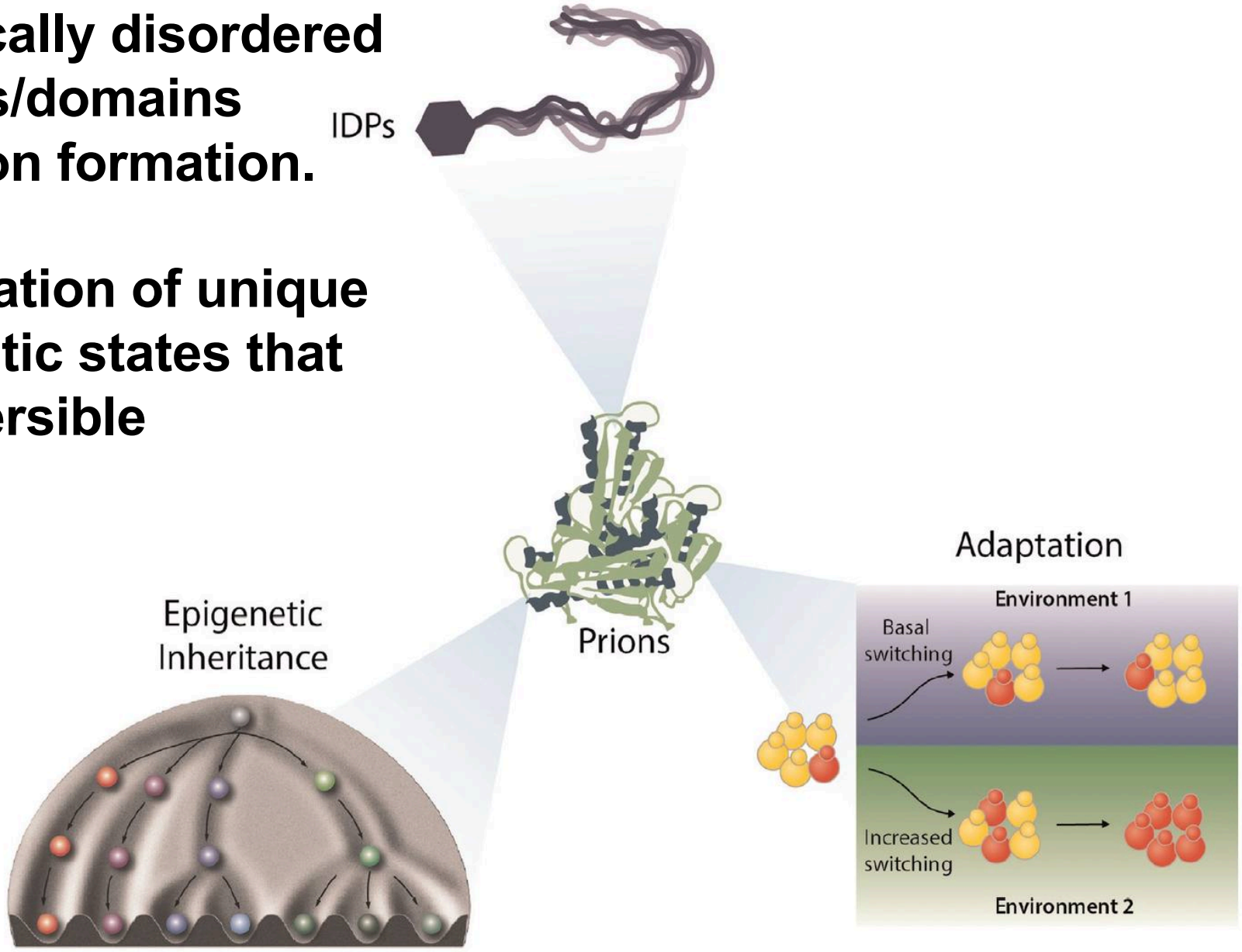


Fig. 1. Prions at the crossroads of three distinct ideas—IDPs, epigenetic inheritance, and evolutionary adaptation in fluctuating environments. IDRs in IDPs fuel the exploration of multiple conformational states. For prions, at least one such conformational state has the ability to self-template and is therefore heritable from one generation to the next. Epigenetic inheritance driven by prions produces different cell states in isogenic backgrounds. Such epigenetic states can be adaptive when faced with changing environments.

Asymptomatic deer excrete infectious prions in faeces

Gültekin Tamgüney[1,2], Michael W. Miller[3], Lisa L. Wolfe[3], Tracey M. Sirochman[3], David V. Glidden[4], Christina Palmer[1], Azucena Lemus[5], Stephen J. DeArmond[1,5] & Stanley B. Prusiner[1,2]

Infectious prion diseases[1]—scrapie of sheep[2] and chronic wasting disease (CWD) of several species in the deer family[3,4]—are transmitted naturally within affected host populations. Although several possible sources of contagion have been identified in excretions and secretions from symptomatic animals[5–8], the biological importance of these sources in sustaining epidemics remains unclear. Here we show that asymptomatic CWD-infected mule deer (*Odocoileus hemionus*) excrete CWD prions in their faeces long before they develop clinical signs of prion disease. Intracerebral inoculation of irradiated deer faeces into transgenic mice overexpressing cervid prion protein (PrP) revealed infectivity in 14 of 15 faecal samples collected from five deer at 7–11 months before the onset of neurological disease. Although prion concentrations in deer faeces were considerably lower than in brain tissue from the same deer collected at the end of the disease, the estimated total infectious dose excreted in faeces by an infected deer over the disease course may approximate the total contained in a brain. Prolonged faecal prion excretion by infected deer provides a plausible natural mechanism that might explain the high incidence and efficient horizontal transmission of CWD within deer herds[3,4,9], as well as prion transmission among other susceptible cervids.

The Spokesman-Review, Spokane, Wash.

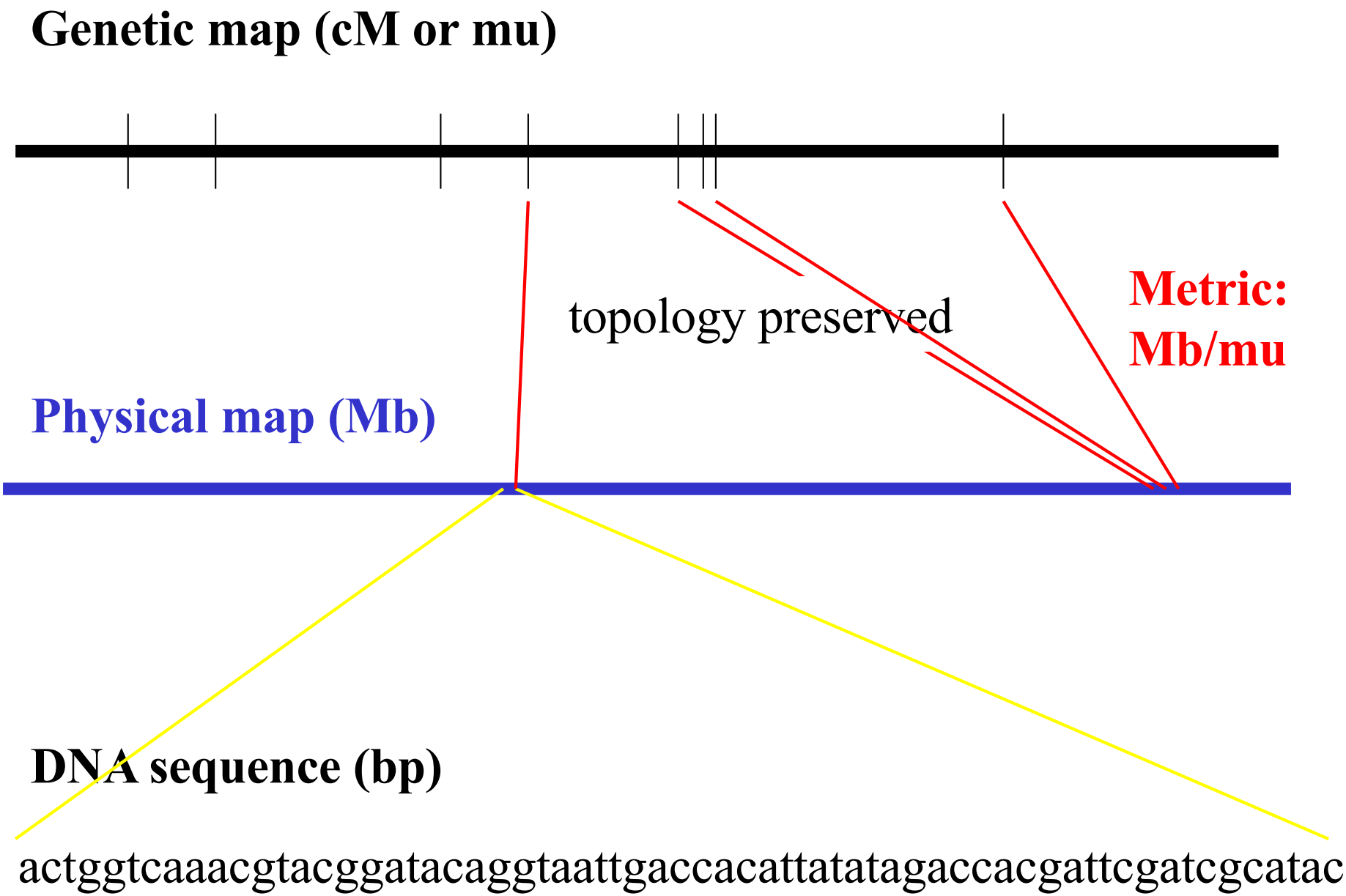
For the first time in a decade, Washington wildlife officials sample for chronic wasting disease on opening day of deer hunting season

Eli Francovich, The Spokesman-Review, Spokane, Wash.

Sun, October 17, 2021, 8:01 AM

Oct. 17—For the first time in a decade, Washington wildlife officials surveyed for a deadly neurological disease during Saturday's modern deer hunting season opener.

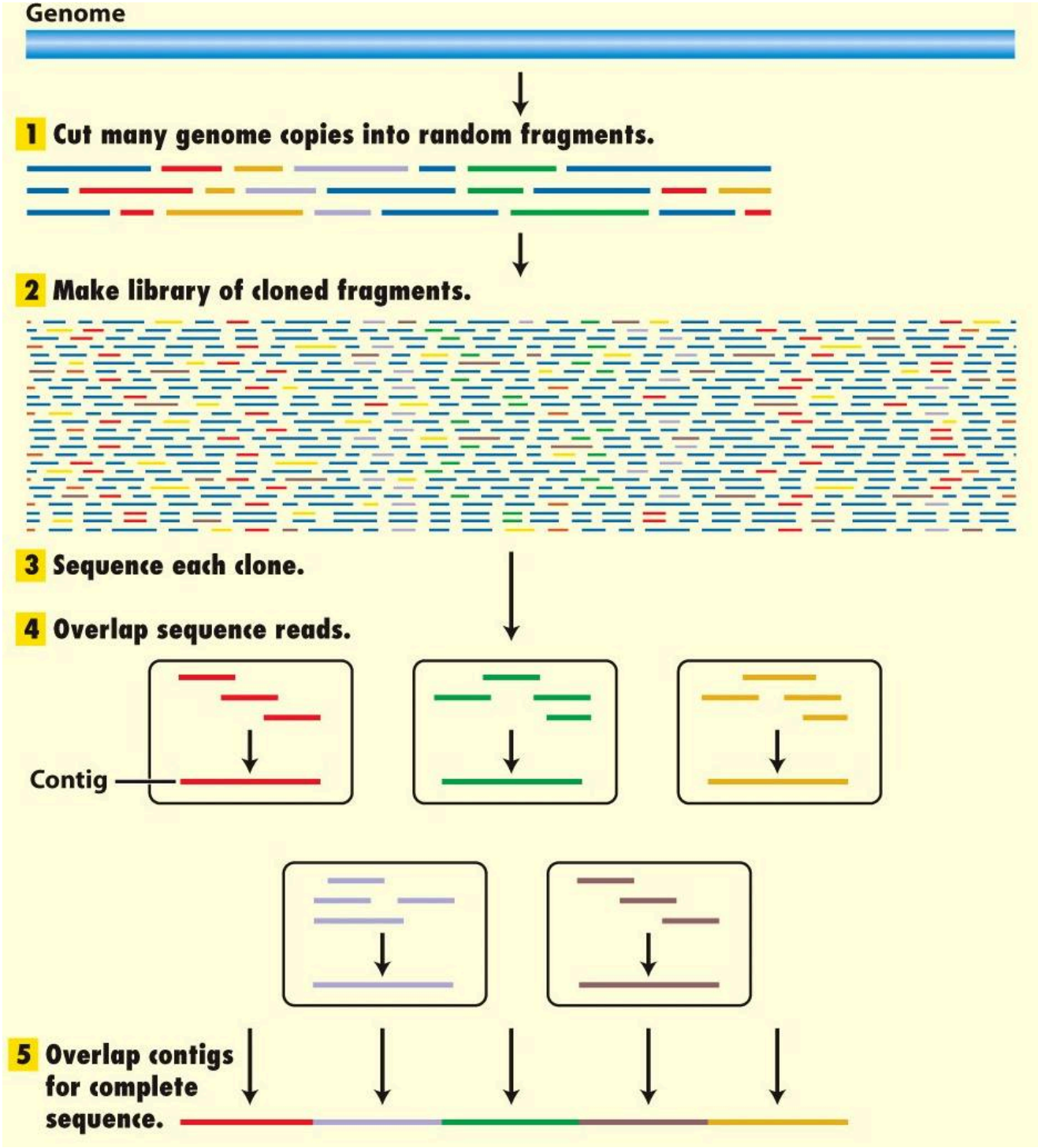
Three types of maps



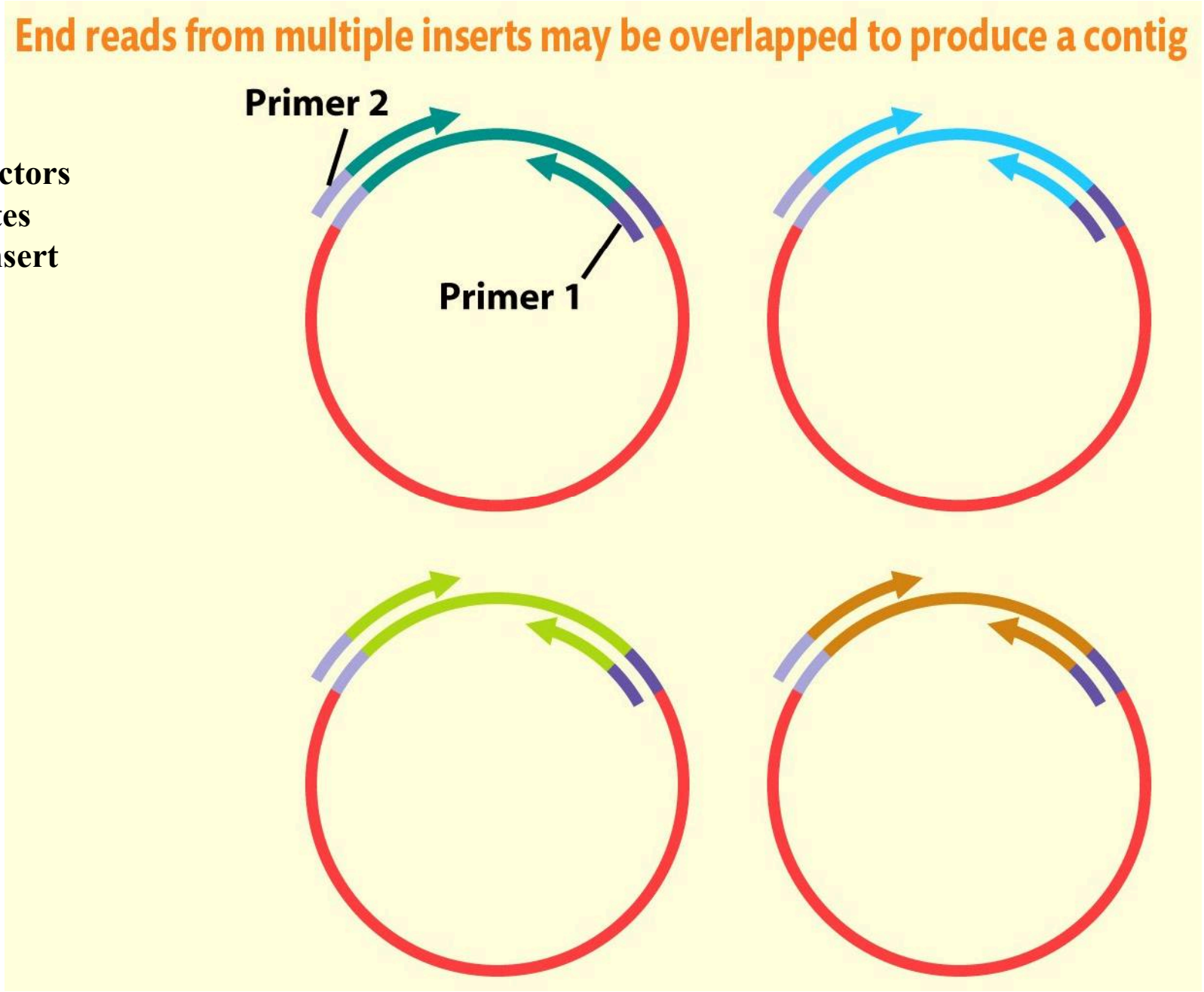
Sequencing Strategies

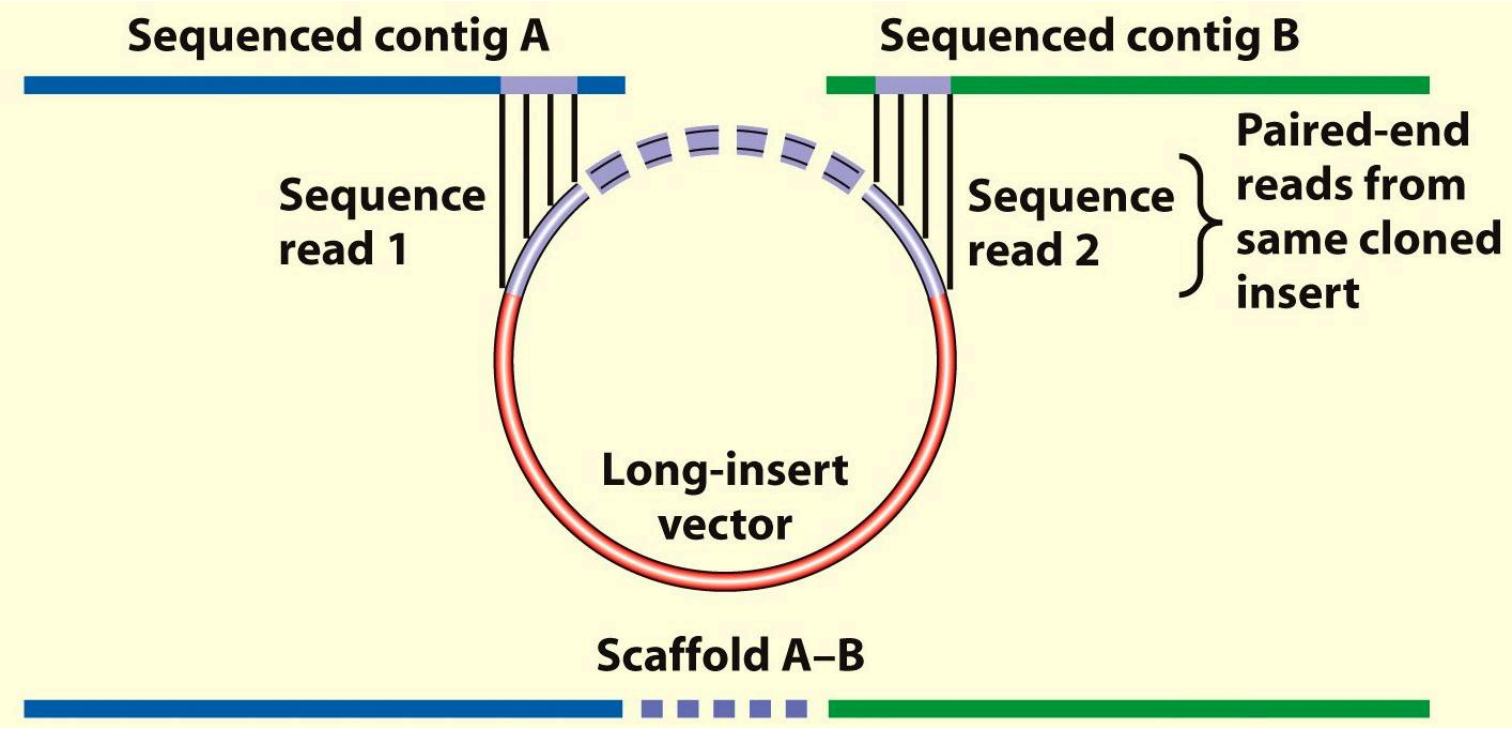
Whole genome “shotgun” and assembly

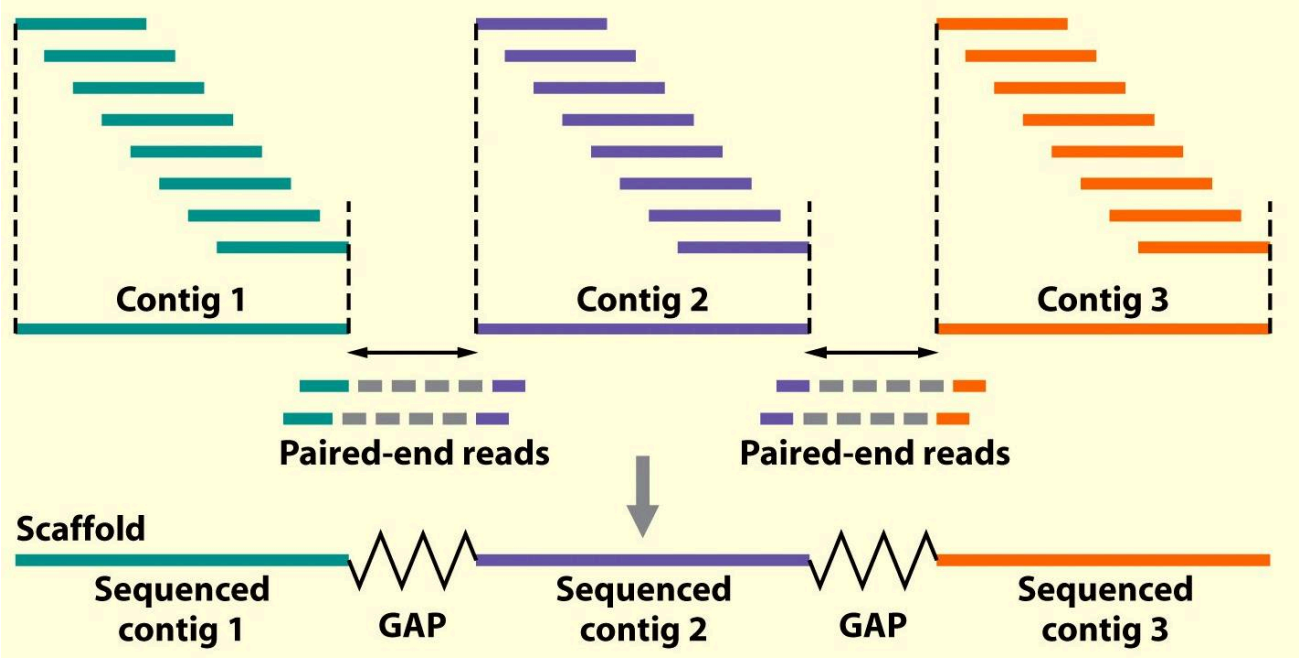
Shotgun sequencing
A genome



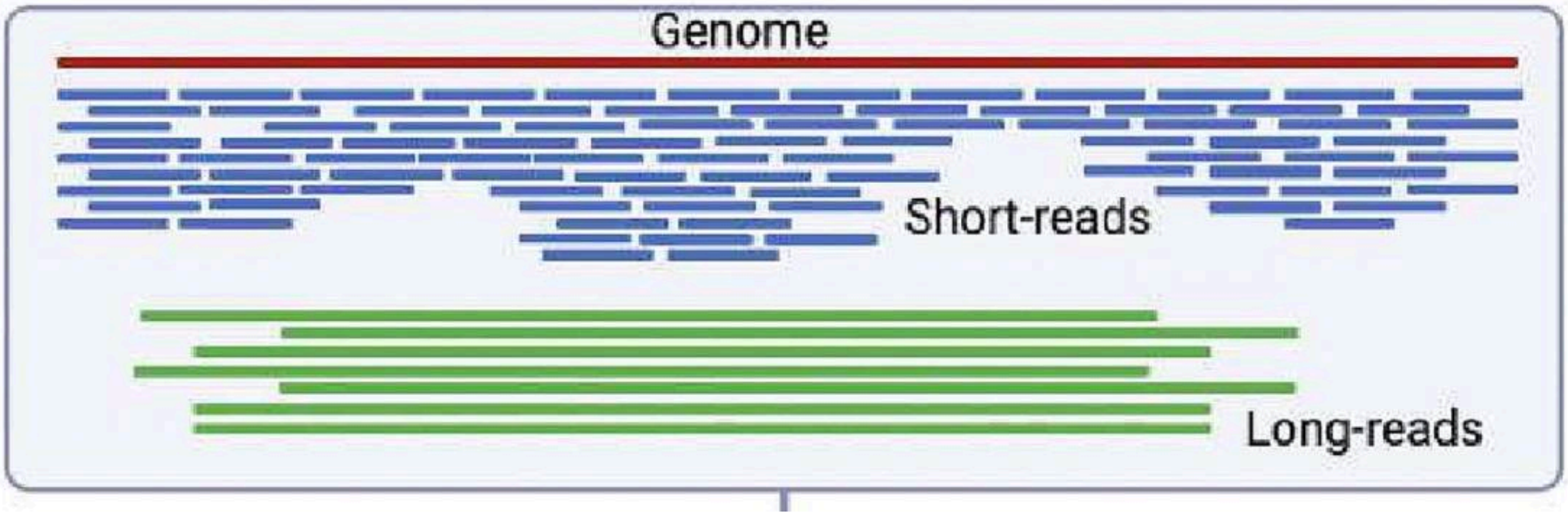
Shear the DNA and then clone it into vectors that have primer sites that read into the insert




Sequenced contig A
Sequenced contig B
Sequence read 1
Sequence read 2
Paired-end reads from same cloned insert
Long-insert vector
Scaffold A–B


Contig 1
Contig 2
Contig 3
Paired-end reads
Paired-end reads
Scaffold
Sequenced contig 1
GAP
Sequenced contig 2
GAP
Sequenced contig 3

De Novo Genome Assembly



Long and short read sequencing metrics

Company	Systems	Data type	Read Length (Maximum)	Accuracy (%)	Maximum throughput per flow cell	Sequencing cost per Gb (USD)
Pacific Biosciences (PacBio)	Sequel II/Sequel IIe	PacBio CLR	>100*kb	87–92	160Gb	13–26
		PacBio HiFi	>20 kb	>99	30Gb	43–86
	Revio	PacBio HiFi	15–20 kb	>99.9	90Gb	11
ONT	Onso	short read	up to 2 × 150 bp[b]	> 99.9	2400–3000 Gb	15
	MinION/GridION	Long	10–100 kb	>99%	48 Gb	$~14-24[d]
		Ultra-long	>100 kb			~72[d]
	PromethION	Long	10–100 kb		50–200 Gb	$ ~ 3–4.6[d]
		Ultra-long	>100 kb			
	Flongle	Long	10–100 kb		2.8 Gb	$ ~ 9.31[d]
		Ultra-long	>100 kb			
Illumina	NextSeq 1000 & 2000	Single-end	1 × 50 bp[c]	>99.9%	60Gb[c]	30/20
		Paired-end	up to 2 × 300 bp[c]		60–180 Gb[c]	
	NovaSeq 6000 series	Single-end	1 × 35 bp[d]		280–350 Gb[d]	10–35
		Paired-end	up to 2 × 250 bp[d]		325–400 Gb[d e]	
	NovaSeq X Series	Paired-end	up to 2 × 150 bp[f]		up to ~8 Tb[f]	2

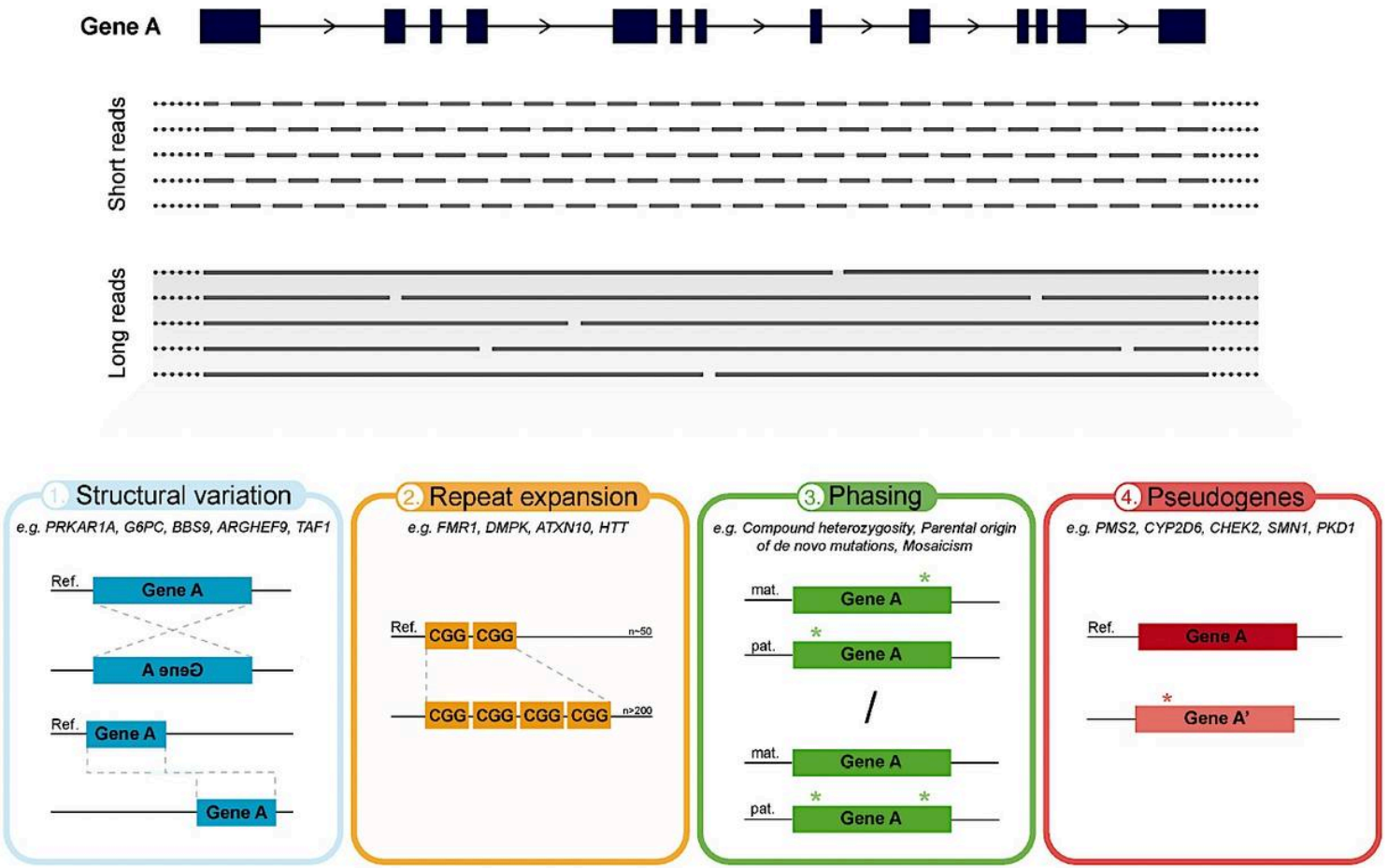


FIGURE 1 | Overview of the main advantages of current long-read sequencing (LRS) approaches in medical genetics. The predominant difference between LRS and the conventional SR-NGS approaches is the significant increase in read length. In contrast to short reads (150–300 bp), LRS has the capacity to sequence 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

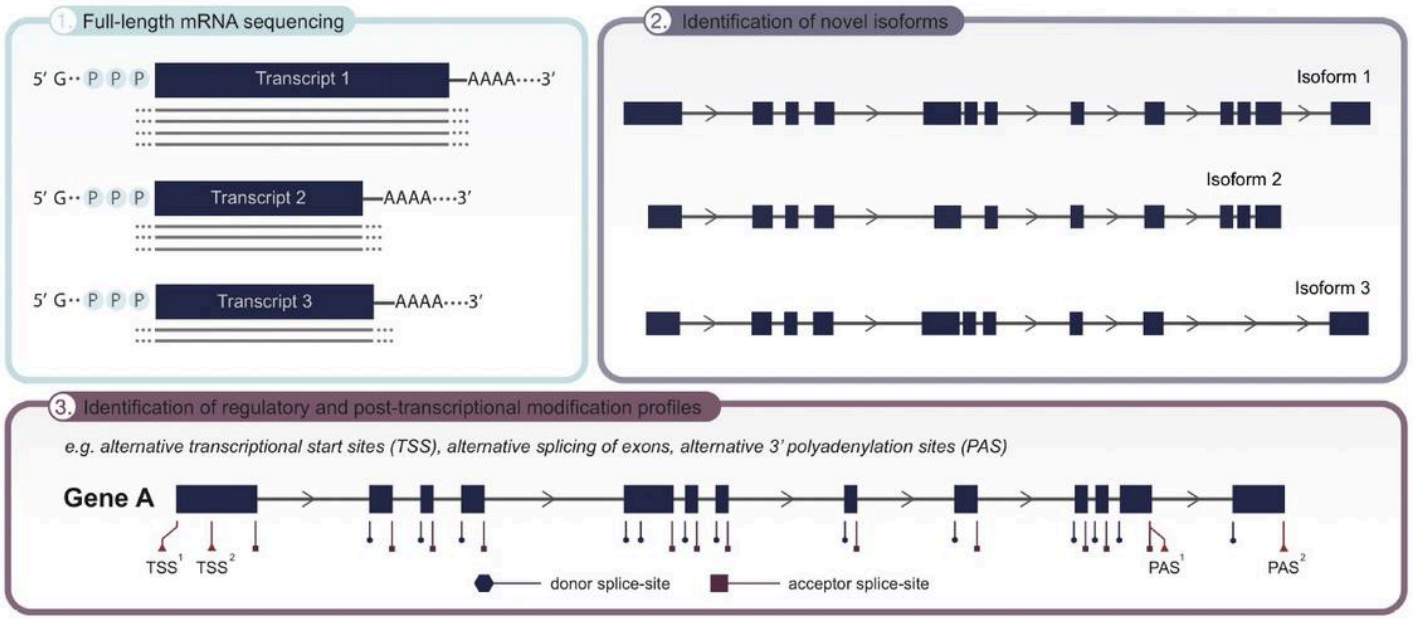


FIGURE 2 | Applicability of long-read sequencing (LRS) to unveil the transcriptome landscape of cells and tissues. Given the significant improvements in read length, employing LRS on RNA level now allows for full-length isoform sequencing, covering the complete mRNA transcript in one single read **(panel 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.

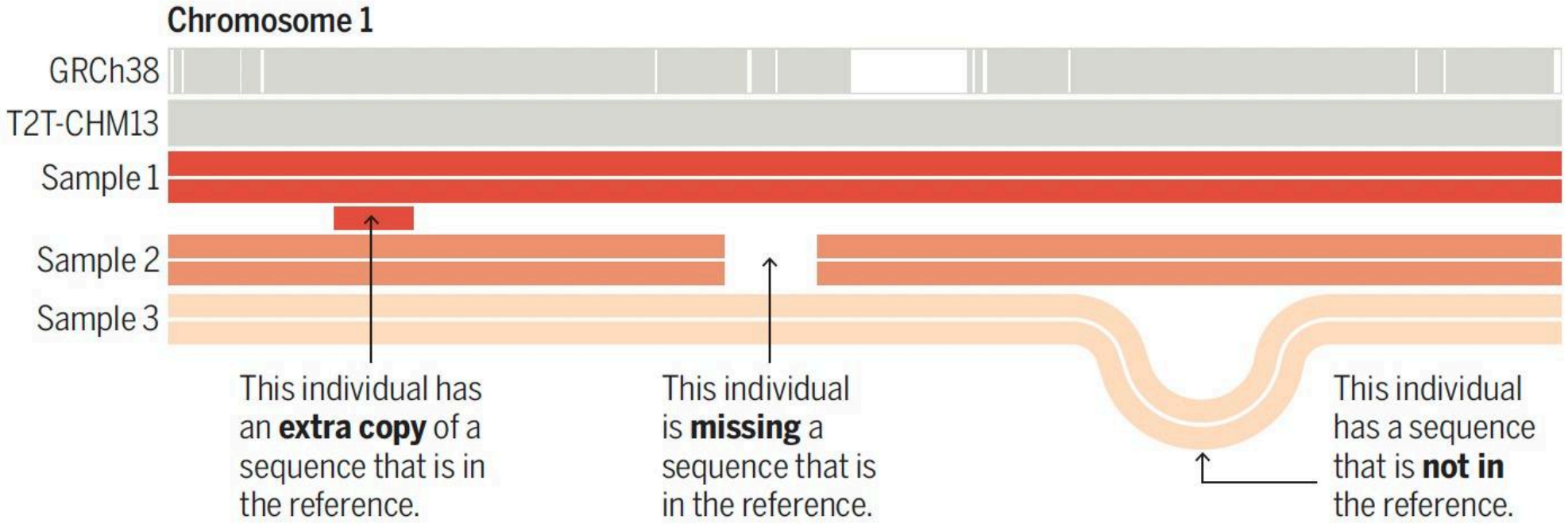


Table 1 Summary of variants in the euchromatic portion of a human genome

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.

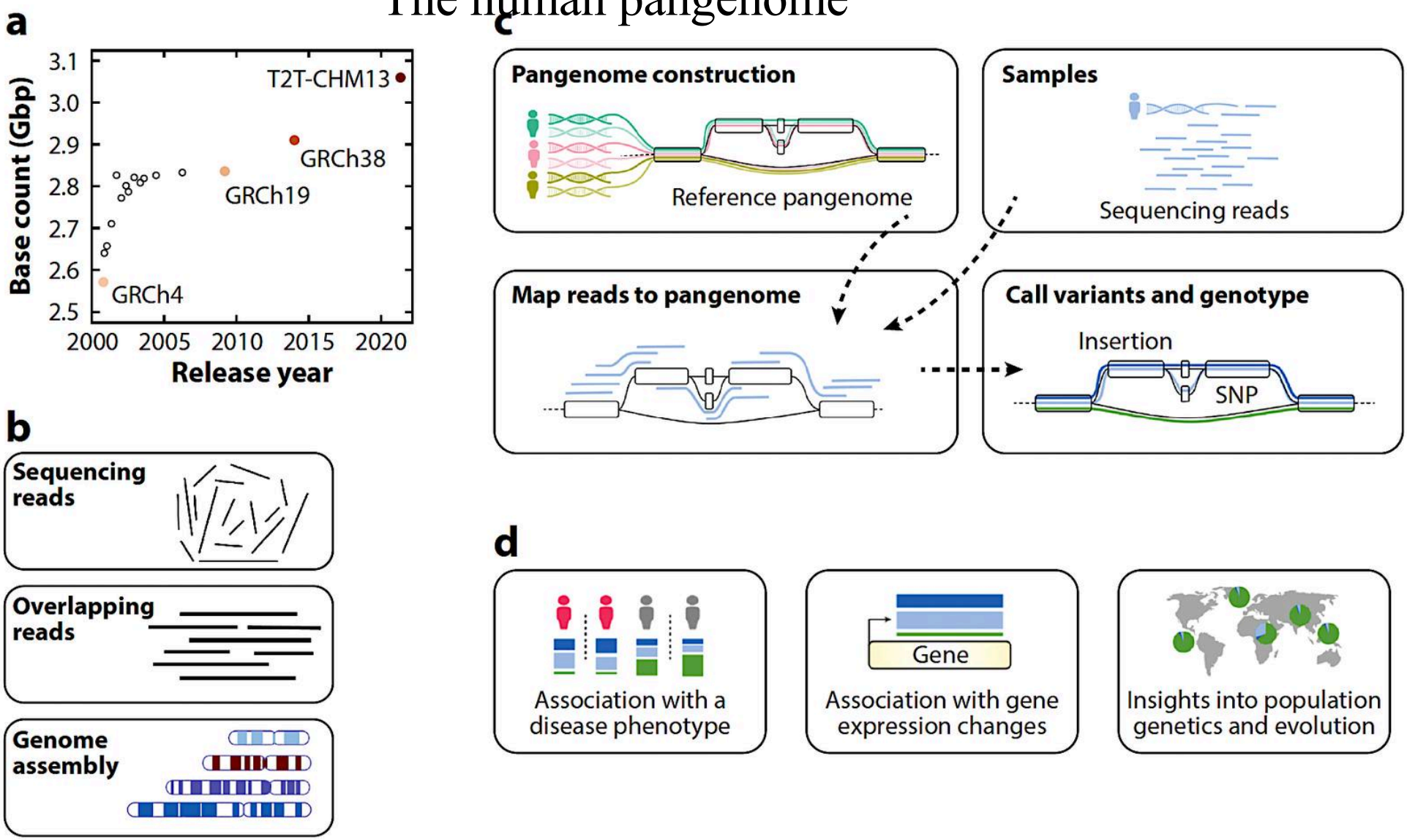
a The average number of sites observed of a given variant type within each genome.

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 Days 00 Hours 00 Minutes 00 Seconds

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

	Whole Genome Sequencing **$99** ~~$29~~ 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** ~~$99~~ 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** ~~$29~~ 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

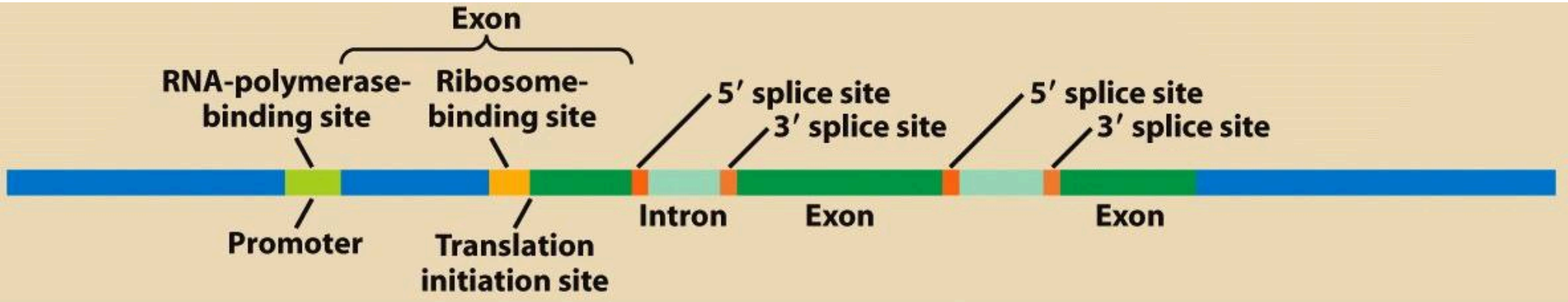
Assembling random DNA sequence reads into sequence contigs

```
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
GTTGGGAAAACTAGTTAGCCATATGCAGAAAACGGAAACTGGACCCCTTCCTTACACCTTATACAAAAAT
TAACTCAAGATGGATTAACGACTTAAACATAAAACATAAAAACATAAAAACCCTAGAAGAAAACCTAGAC
AATATGATTCAGGACATAGGCATGGGAGAAGACTTTATGACTAAAACACCAAAAGCAATAGCAACAAAAG
CCAAAATTGACAAATGGGATCTAATTAAGCTAAAGAGCTTCTGCACAGCAAAAGAAACTATCATCAGAGT
GAATAGGCAACCTACATAATGGGAGAAAATTTTTGCAATCTATCCATCTGACAAAGGGCTAATATCCAGA
ATCTACAAAGAACTTGAACAAATTTACAAGAAAAAAACAACTCCATCAAAAAGTGGGCAGTAAGAACAGA
CACTTCTCAAAAGAAGACATTTATGCAGCCAATACACATATGAAAAAAAGCTCATCATCACTGGTCATTA
GAGAAATGCAAATGAAAATCACAATGAGATACCATCTCATGCCAGTTAGAATGGTGATCCTTAGGAGGTC
AGGAAACAACAGATGTTGAAGAGGATGTGGAGAAATAGGAACACTTTTACACTGTTGGTGGGAGTATCAA
TTAGTTCAACCACTGTGGAAGACAGTGTGGCAATTCCTCAAGGATCTAGAACTAGAAATACCATTTGACC
CAGCAATTCCATTGCTAGGTATATACCCAAAGGATTATAAATCATTCTACTATAAAGACACACACACACG
TATGTTTACTGCGGCACTGTTCACAATAGCAAATACTTGGAACCAACCCAAATGCCCATCAGTGATAGAC
TGGATAAAGAAAACATGGCACATATACACCATGGAATACTATGCAGCCATAAAAAGGTTGAGTTCACGTC
CTTTGCAGGGACATGGCTGAAGCTGGAAACCATCATTCTCAGCCAACTAACGTAAGAACAGAAAACCAAA
CACTGCATATTTTCACTGATAAGCGTGAGTTGAACAATGAGAACACATGGACATGGGGAAGGGAACATCA
CACACTGGGGCCTGTCAGGGGTTGAGGGGTTAGGGGAGGGATAGTATTAGGAGAAATACCTAATGTAGAT
GACAAGTTGATGGGTGCAGCAAACCACCATGGCACATGGATATCTATGTAACAAACCTGCATGTTCTGCA

Finding genes in genome sequence can be hard

As your english teacher always told you,
grammer and punctuation are the key!!!!!!!!!!

ajisufjjjulibnbzvxcqplmjueyyyetyttb
winsoutherncaliforniawithgraceand
splendorboundwherethelofty mountai
npeakslookouttolandbeyondproudlys
tandsouralmamaterglorioustoseewer
aiseourvoicesproudlyhailinghailin
gtheeechoesringingwhilewearesingin
goverlandandseathehallsoffamereso
undthynamenoblecitjiuienenebstijz
inzyiplmnhtgbswbneemnimistiungjzx

ajisufjjulibnbzvxcqplmjueyyetyttb
winsoutherncaliforniawithgraceand
splendorboundwherethetheloftymountai
npeakslookouttolandbeyondproudlys
tandsouralmamaterglorioustoseewer
aiseourvoicesproudlyhailinghailin
gtheeechoesringingwhilewearesingin
goverlandandseathehallsoffamereso
undthynamenoblecitjiuienenebstijz
inzyiplmnhtgbswbneemnimistiungjzx

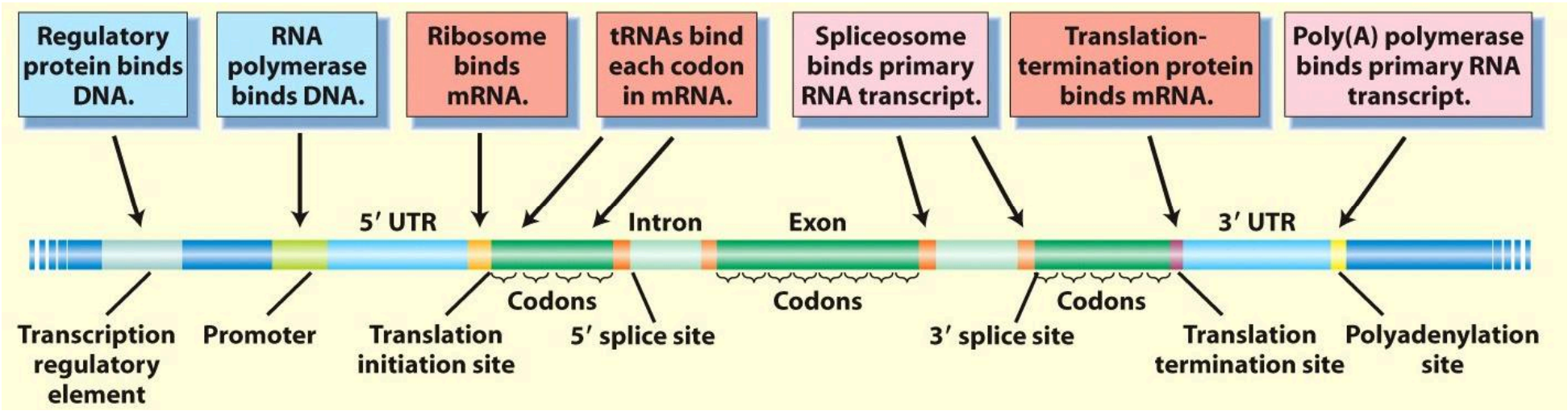
ajisufjjulibnbzvxcqplmjueyyetyttb
winsoutherncaliforniawithgraceand
splendorboundwherethelofty mountai
npeakslookouttolandbeyondproudlys
tandsouralmamaterglorioustoseewer
aiseourvoicesproudlyhailinghailin
gtheeechoesringingwhilewearesingin
goverlandandseathehallsoffamereso
undthynamenoblecitjiuienenebstijz
inzyiplmnhtgbswbneemnimistiungjzx

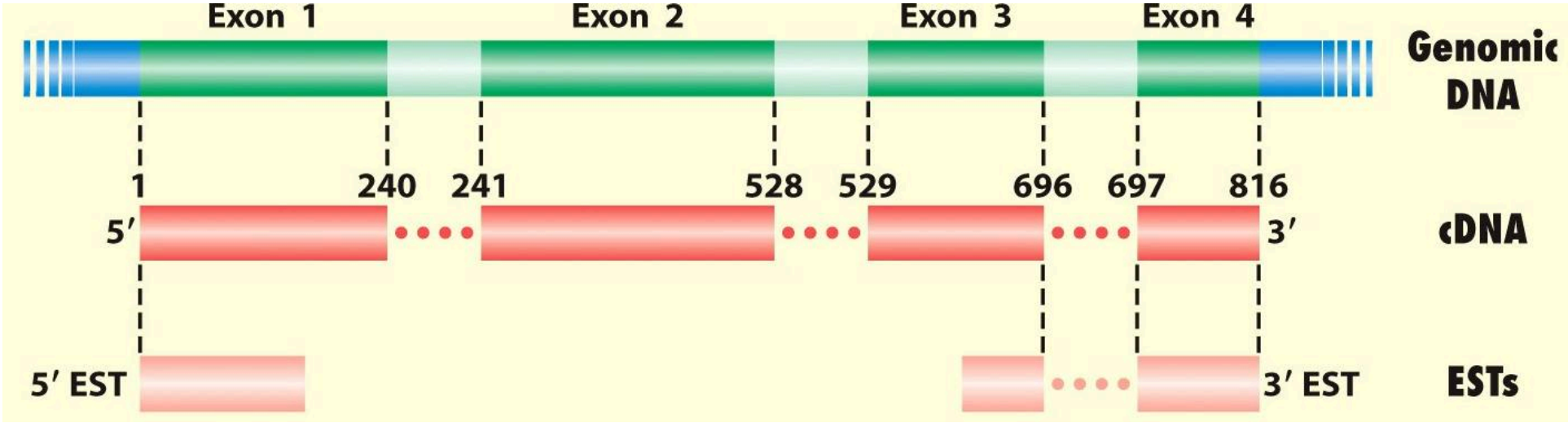
ajisufjjulibnbzvxcqplmjueyyetyttb
wInsouthernCaliforniawithgraceand
splendorboundWheretheloftymountai
npeakslookouttolandbeyondProudlys
tandsourAlmaMatergloriousstoseeWer
aiseourvoicesproudlyhailinghailin
gTheeEchoesringingwhilewearesingin
goverlandandseaThehallsoffamereso
undthynamenobleCITjiuienenebstijz
inzyiplmnhtgbswbneemnimistiungjzx

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

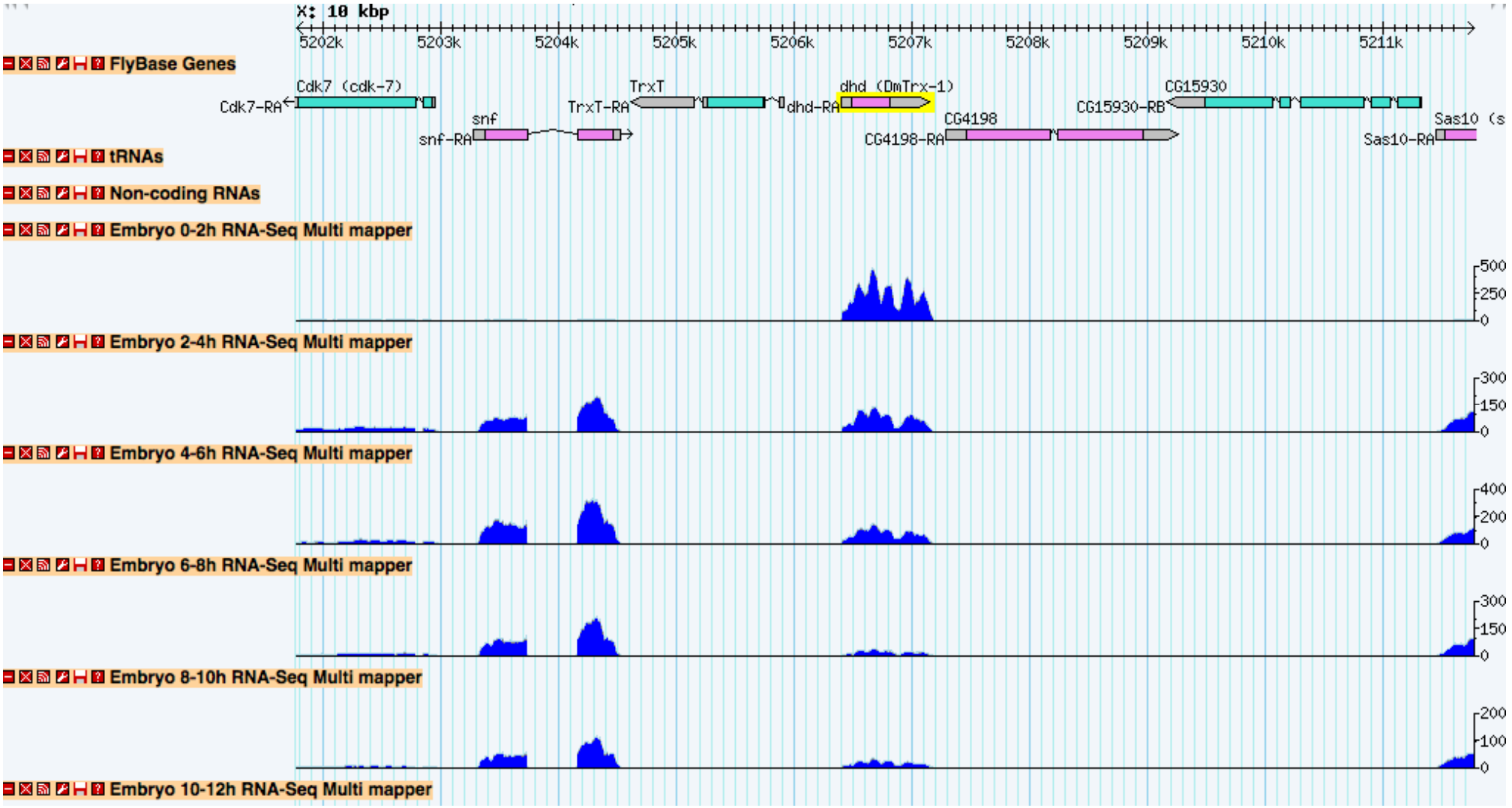
In southern California
with grace and splendor bound,
Where the lofty mountain peaks
lookout to land beyond,
Proudly stands our Alma Mater,
glorious to see;
We raise our voices proudly,
Hailing, hailing Thee!
Echoes ringing while we're singing
over land and sea;
The halls of fame resound thy name,
noble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

Manton Barnes, BS '21 EE


Regulatory protein binds DNA.
RNA polymerase binds DNA.
Ribosome binds mRNA.
tRNAs bind each codon in mRNA.
Spliceosome binds primary RNA transcript.
Translation-termination protein binds mRNA.
Poly(A) polymerase binds primary RNA transcript.
5′ UTR
Intron
Exon
3′ UTR
Codons
Codons
Codons
Transcription regulatory element
Promoter
Translation initiation site
5′ splice site
3′ splice site
Translation termination site
Polyadenylation site


Exon 1
Exon 2
Exon 3
Exon 4
Genomic DNA
1
240
241
528
529
696
697
816
5′
3′
cDNA
5′ EST
3′ EST
ESTs

http://www.modencode.org/



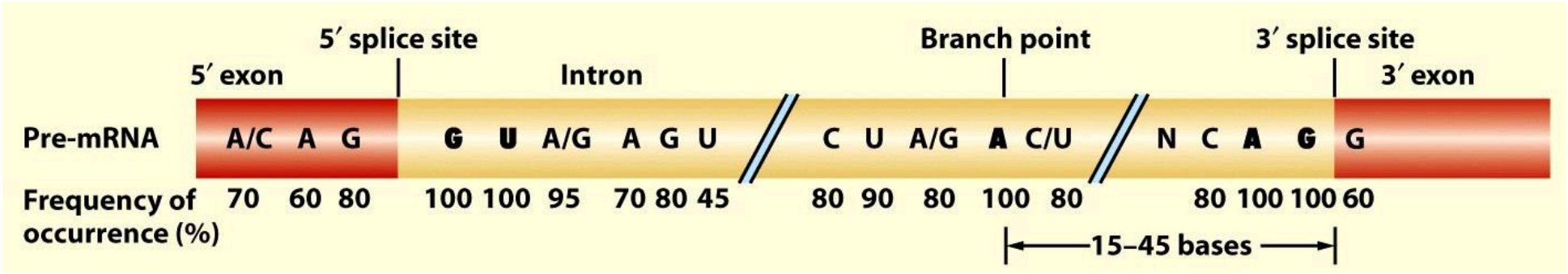
Gene finding

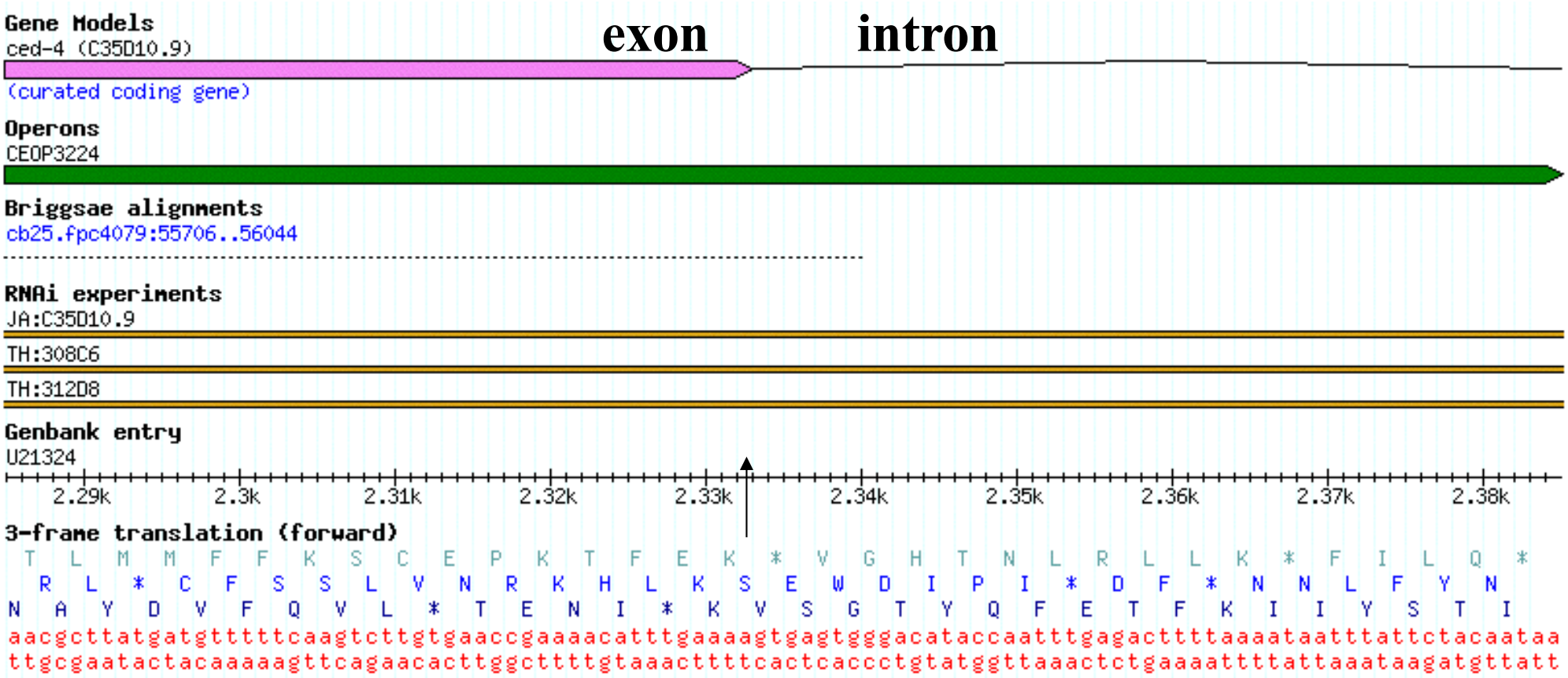
actggtcaaacgtacggatacaggtaattgaccacattatatagaccacgattcgatcgcatac

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?

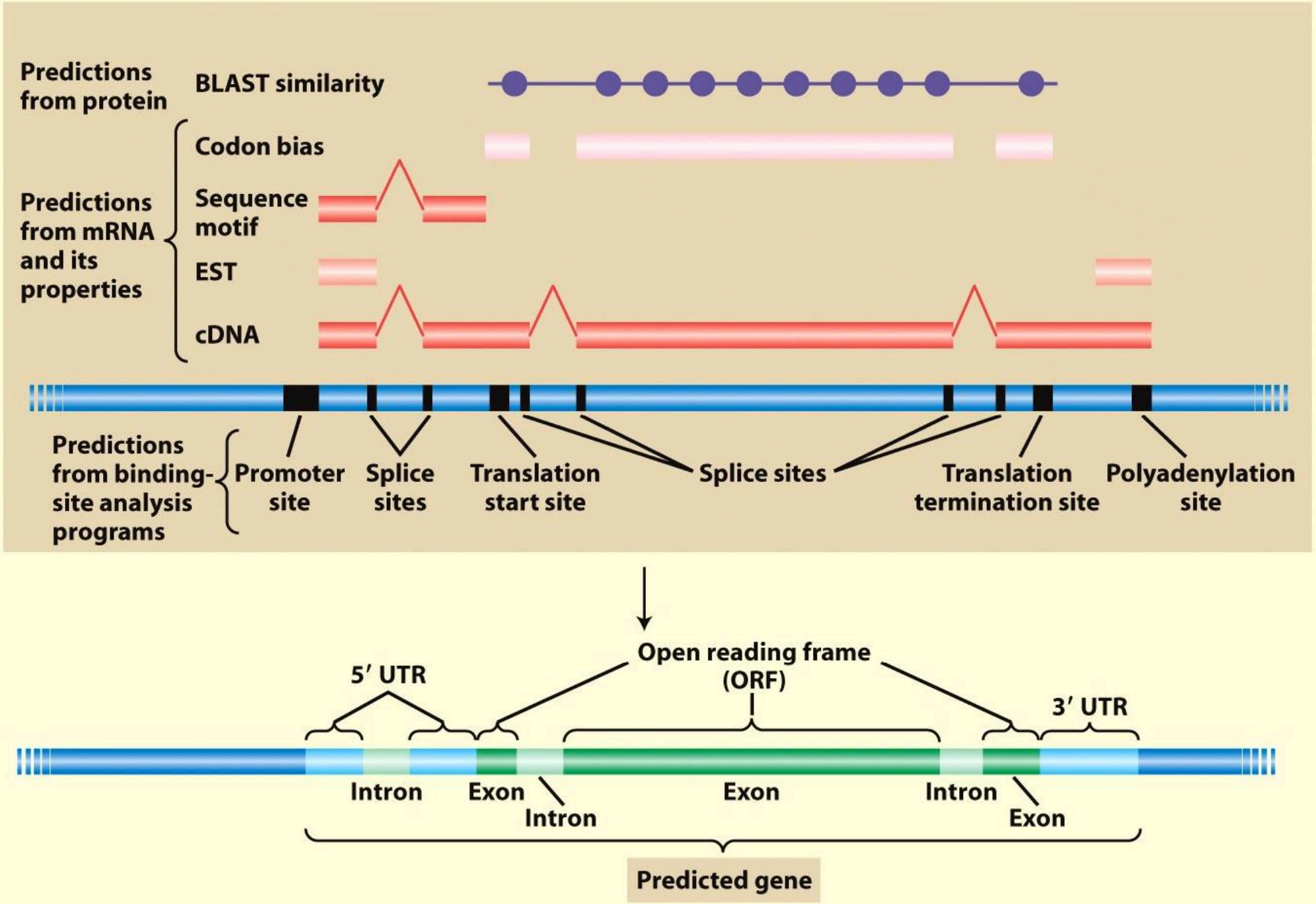




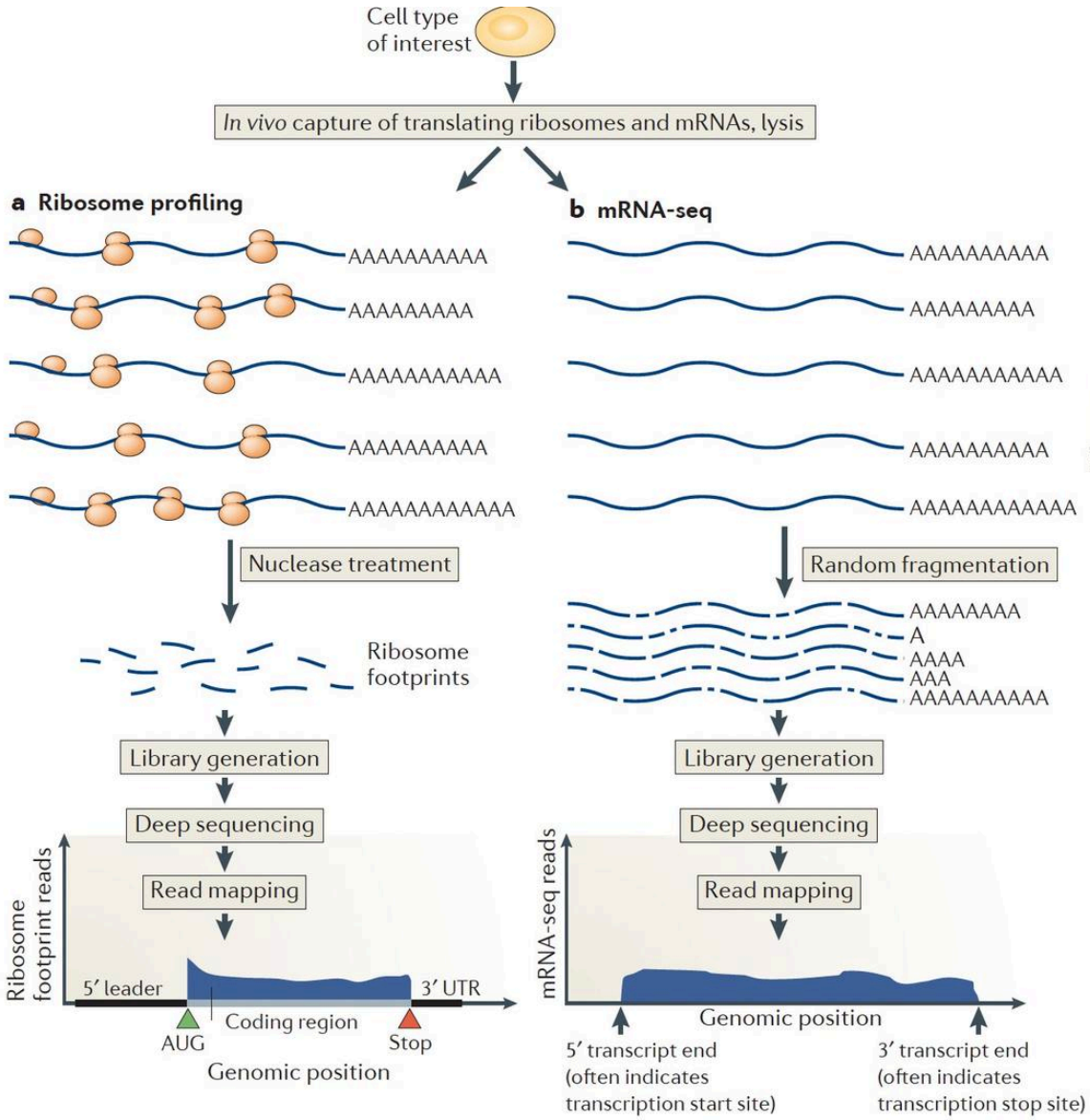
***= stop codons**

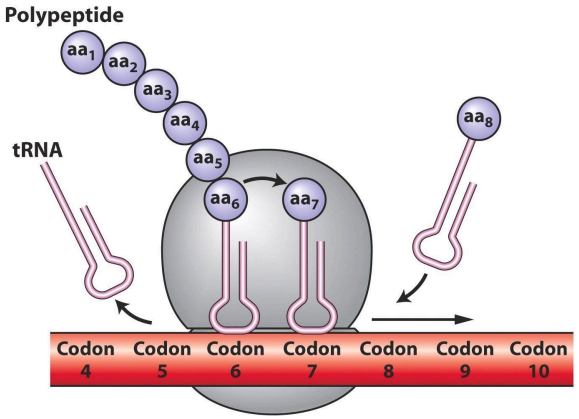
XXXgt = splice donor

Many forms of evidence are integrated to make gene predictions



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





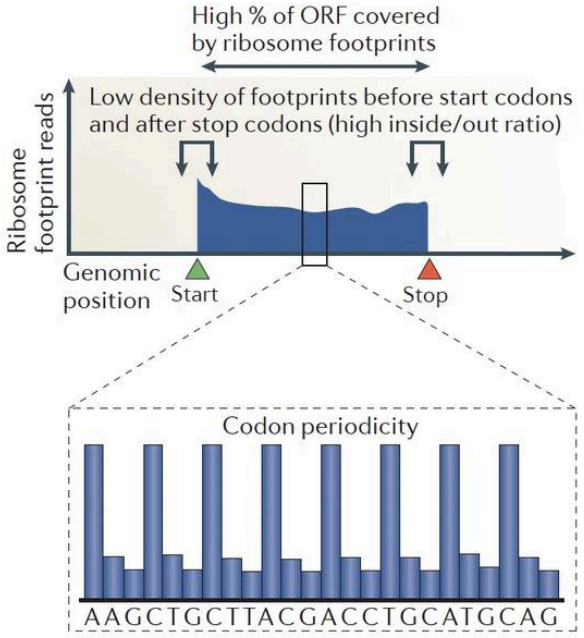
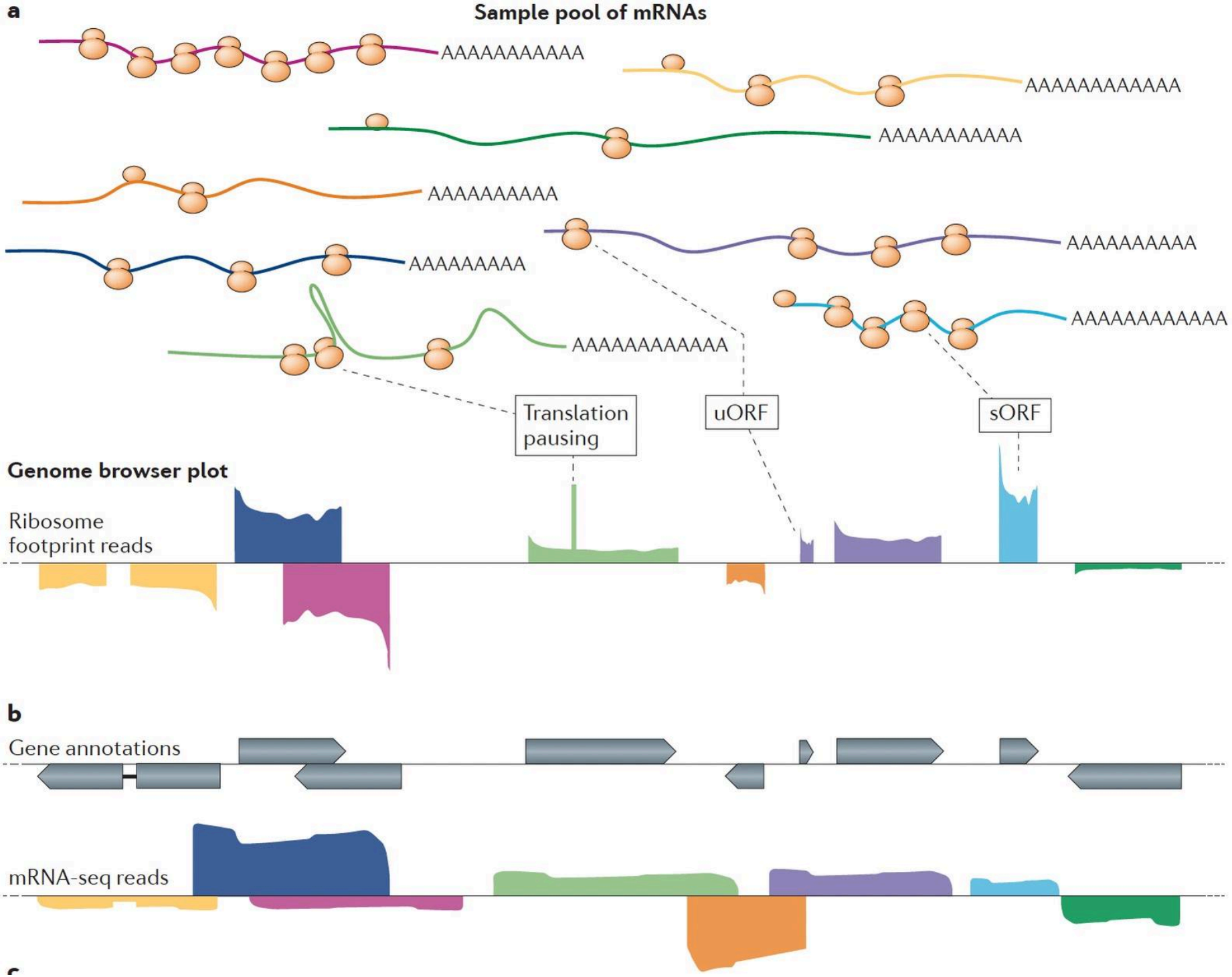


Figure 1 | **An overview of ribosome profiling.** **a** | Ribosome-bound mRNAs are isolated by size and treated with a nonspecific nuclease (typically RNase I or micrococcal nuclease), resulting in protected mRNA fragments termed 'footprints'. These ribosome footprints are isolated and converted to a 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.


a
Sample pool of mRNAs
AAAAAAAAAAA
AAAAAAAAAAAA
AAAAAAAAAAA
AAAAAAAAAA
AAAAAAAAA
AAAAAAAAAA
AAAAAAAAAAAA
Translation pausing
uORF
sORF
Genome browser plot
Ribosome footprint reads
b
Gene annotations
mRNA-seq reads

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

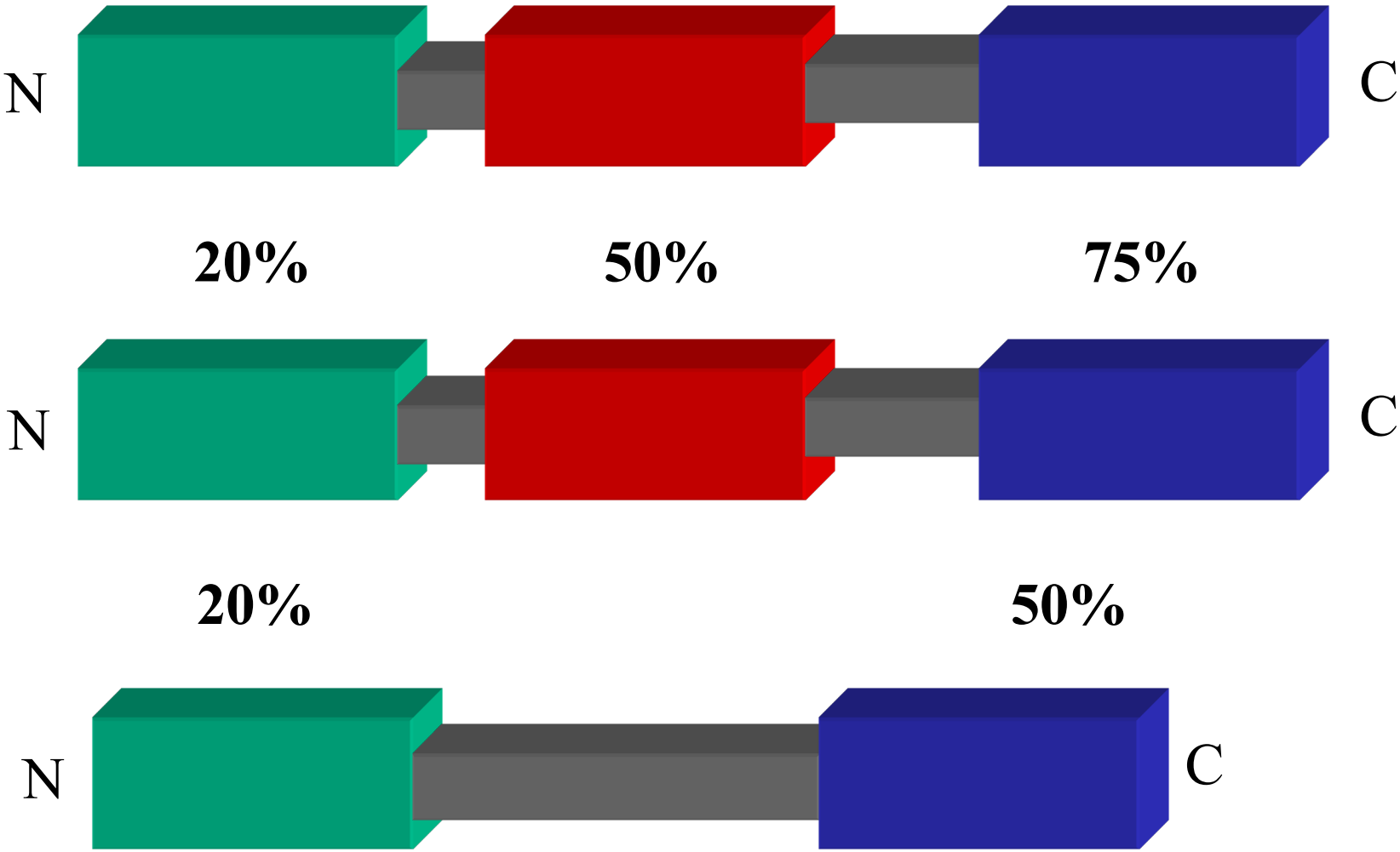
```
In southern California with grace  and splendor bound,
In Smoggy   California Where nerds and geeks   abound,

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 f ame resound thy name, noble CIT!
The halls of flame resound thy name, Noble CIT.
```

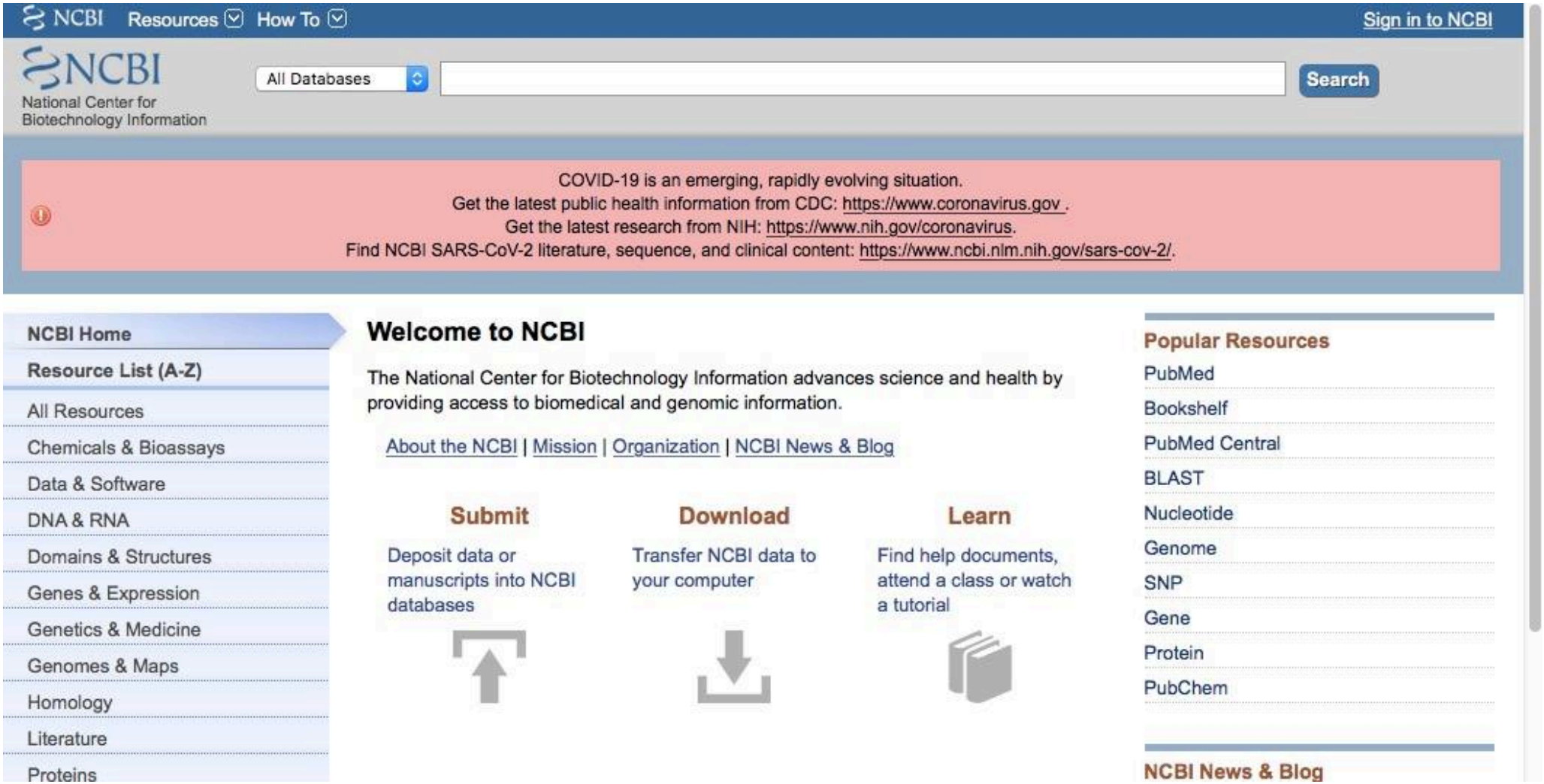
Comparative Sequence Alignment: two versions of Caltech Alma Mater song.

```
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,Chainsfor you and me.

The halls of f ame resound thy name, noble CIT!
The halls of flame resound thy name, Noble CIT.
```

Proteins are modular: compare domains and architecture

% amino acid identity




NCBI Resources How To
Sign in to NCBI
NCBI
National Center for Biotechnology Information
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

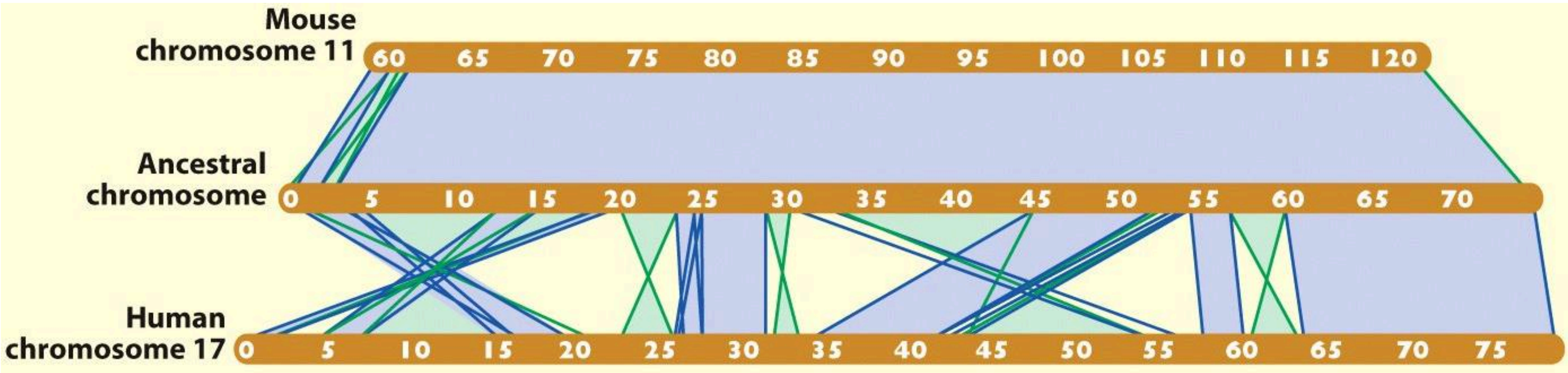
```
Score =  227 bits (573), Expect = 1e-57
 Identities = 122/264 (46%), Positives = 165/264 (62%), Gaps = 1/264 (0%)
 Frame = +1

Query: 206   SIVPATPHTLYKKQCA-NCGEGRAYKNISKDPALDGFTPKYYDELEKNEEHFIEIEDLLQ 264
              SIVP +   + K+       E  AY+ I K+P      P Y+   E   +HFIE++DLL
Sbjct: 78475 SIVPTSTGIVRKRISGLEDSEVHAYRLICKEPQTAQIVPAYFGIQEMQSQHFIELQDLLA 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   YMQFRERESSTAQLGFRIEAAKRLEGALEKNFKKVRTVEDVTTTFMDFFGTQRSRVRQQL 384
              YM FRE  SS+   GFRIEA +       K+ K  R+ E +  T   F   +RS V+++L
Sbjct: 78826 YMTFRESLSSSHSKGFRIEALRLRGRPPVKDLKTCRSSEQIAQTIEQFLAARRS-VQKEL 79002

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

Query: 445   WIPGNNEDGYLIGIDNLVKILEEL 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

Query  479  GATCAGGTGGATAACAACACGAACGCGACCCAGCTATTCAAAAATAATATAAACAAAACC  538
            || |||||||| ||||| ||||| || || ||||| ||||| || || ||  ||||| | |
Sbjct  73   GACCAGGTGGACAACAATACGAATGCTACGCAGCTGTTCAAGAACAACATCTACAACATC  132

Query  539  AGAATGAACGATTTAAACCGCGAGGAGACGCGATTAAAGACCTTCACCGACTGGCCGCTA  598
            |   |||||||  |  || || |||||  ||||  | || |||||| || | |||||  ||
Sbjct  133  A---TGAACGAGTACAATCGGGAGGAAGCGCGCCTGAAAACCTTCGCCAATTGGCCTCTG  189

Query  599  GATTGGCTGGATAAACGCCAATTGGCCCAAACCGGCATGTACTTCACACACGCCGGCGAC  658
            |  ||||| || || |  || ||||||| ||| ||||||| ||  ||  |||     |||
Sbjct  190  GCGTGGCTCGACAAGCATCAGTTGGCCCGAACGGGCATGTTCTATACGAACGATAACGAT  249

Query  659  AAAGTTAAATGCTTTTTCTGCGGCGTGGAAATCGGTTGCTGGGAGCAGGAGGATCAGCCC  718
            || ||  || |||| ||||||||  ||||| |||||  |||||||  |  || |||||  ||
Sbjct  250  AAGGTCAAGTGCTATTTCTGCGAGGTGGAGATCGGGCGCTGGGATCTCGATGATCAACCT  309

Query  719  GTGCCGGAACATCAGCGATGGTCGCCCAACTGTCCACTGTTGCGCCGGCGCACTACCAAC  778
            || ||  ||  || |  ||| ||||||||||||||||  ||   |||||| |||||  ||||||
Sbjct  310  GTACCTGAGCACCTGCGCTGGTCGCCCAACTGTCCGCTCCTGCGCCGTCGCACCACCAAC  369

Query  779  AATGTGCCGATCAATGCCGAAGCATTAGATCGCATCCTGCCGCCAATAAGCTACGATATC  838
            |||||||||  |  ||    |||  ||   |  ||  ||||||||||| ||  || |||||||  ||
Sbjct  370  AATGTGCCGCTGAACAGCGAGGCCCTCGACCGCATCCTGCCCCCCATCAGCTACGACATT  429

Query  839  TGCGGCGCCAACGACTCGACGCTAGAGATGAGGGAGCACGCCTACGCAGAAGGCGTCATA  898
            |||||||||||||||||||||  | |||  ||  | |||  ||||||  ||  ||  |||   ||||
Sbjct  430  TGCGGCGCCAACGACTCGACGGTCGAGCTGCGCGAGAGCGCCTATGCCGAGGGCCGCATA  489
```

Probability of sequence alignment to this extent by chance

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



>gi|19773481|dbj|AB072322.1|AB072322S1
Pan troglodytes COX8 gene for cytochrome oxidase subunit VIII, exon 1

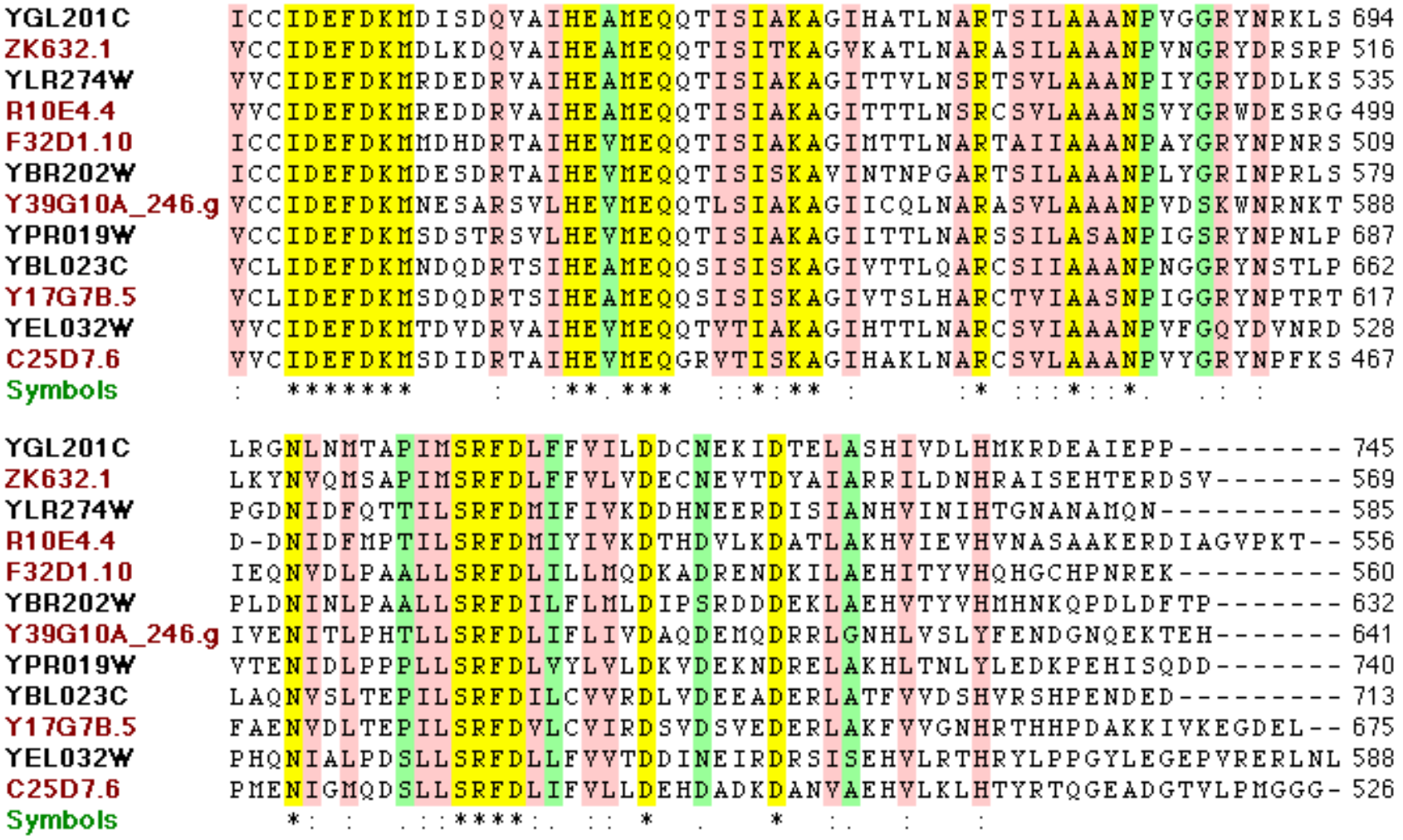
```
ATGTCCGTCCTGACGCCGCTGCTGCTGCGGGGCTTGACAGGCTCGGCCCGGCGGCTCCCAGTGCCGCGCG
CCAAGATCCATTCGTTGCCGCCGGAGGAGAAGCTTGGGATCATG
```

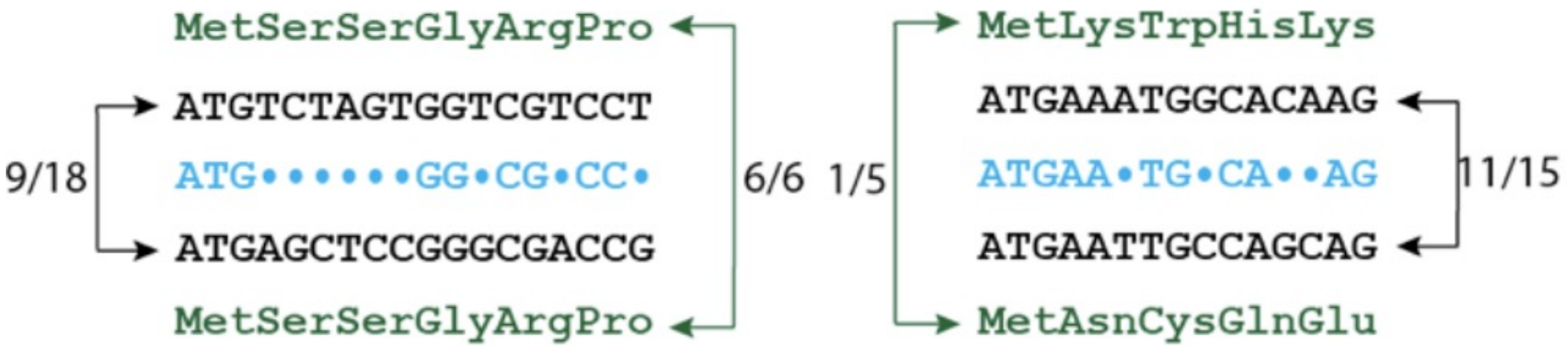
>gi|11436791|ref|XM_006132.1|
Homo sapiens cytochrome c oxidase subunit VIII (COX8), mRNA
Length = 473

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

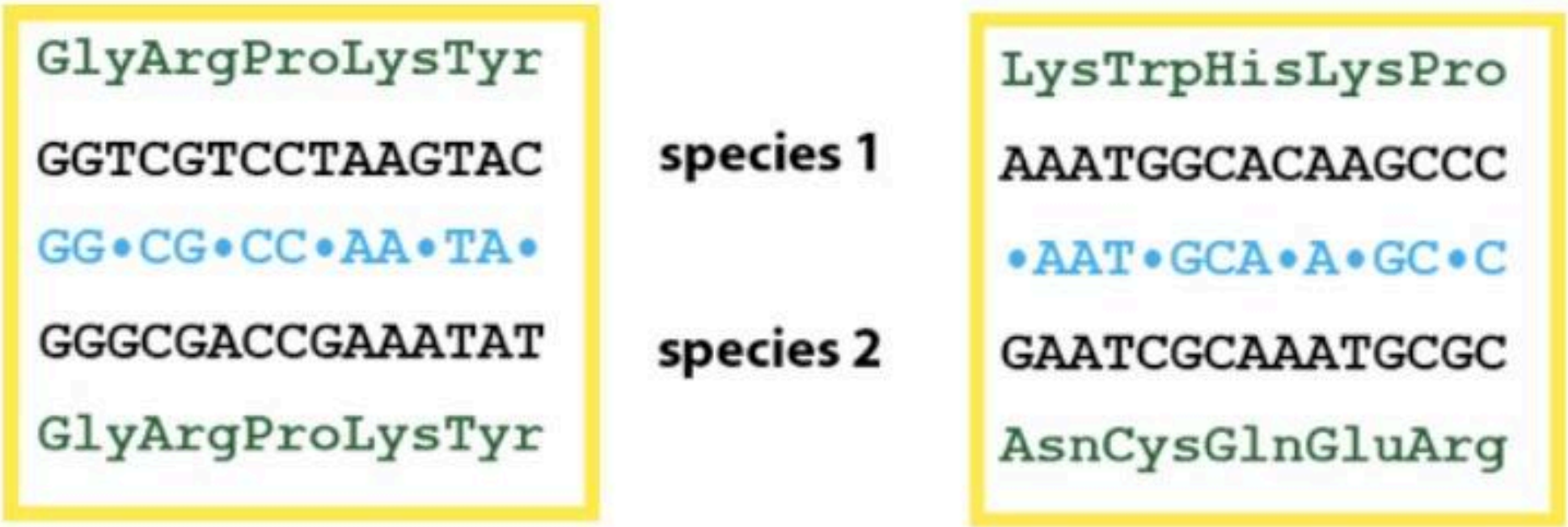
```
Chimp   Query: 1  atgtccgtcctgacgccgctgctgctgcggggcttgacaggctcggcccggcggctccca 60
                  ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Human   Sbjct:42  atgtccgtcctgacgccgctgctgctgcggggcttgacaggctcggcccggcggctccca 101

        Query: 61   gtgccgcgcgccaagatccattcgttgccgccggaggagaagcttgggatcatg 114
                    ||||||||||||||||||||||||||||||||||||| ||||||||||||||||
        Sbjct: 102  gtgccgcgcgccaagatccattcgttgccgccggaggggaagcttgggatcatg 155
```

Multiple Sequence Alignments




MetSerSerGlyArgPro
ATGTCTAGTGGTCGTCCT
ATG••••••GG•CG•CC•
ATGAGCTCCGGGCGACCG
MetSerSerGlyArgPro
9/18
6/6
MetLysTrpHisLys
ATGAAATGGCACAAG
ATGAA•TG•CA••AG
ATGAATTGCCAGCAG
MetAsnCysGlnGlu
1/5
11/15

Which is protein-coding?

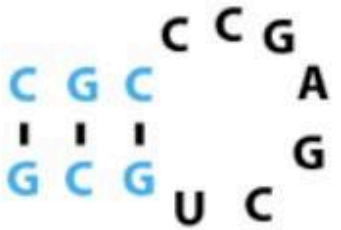


5/5 substitutions in wobble position

1/5 substitutions in wobble position

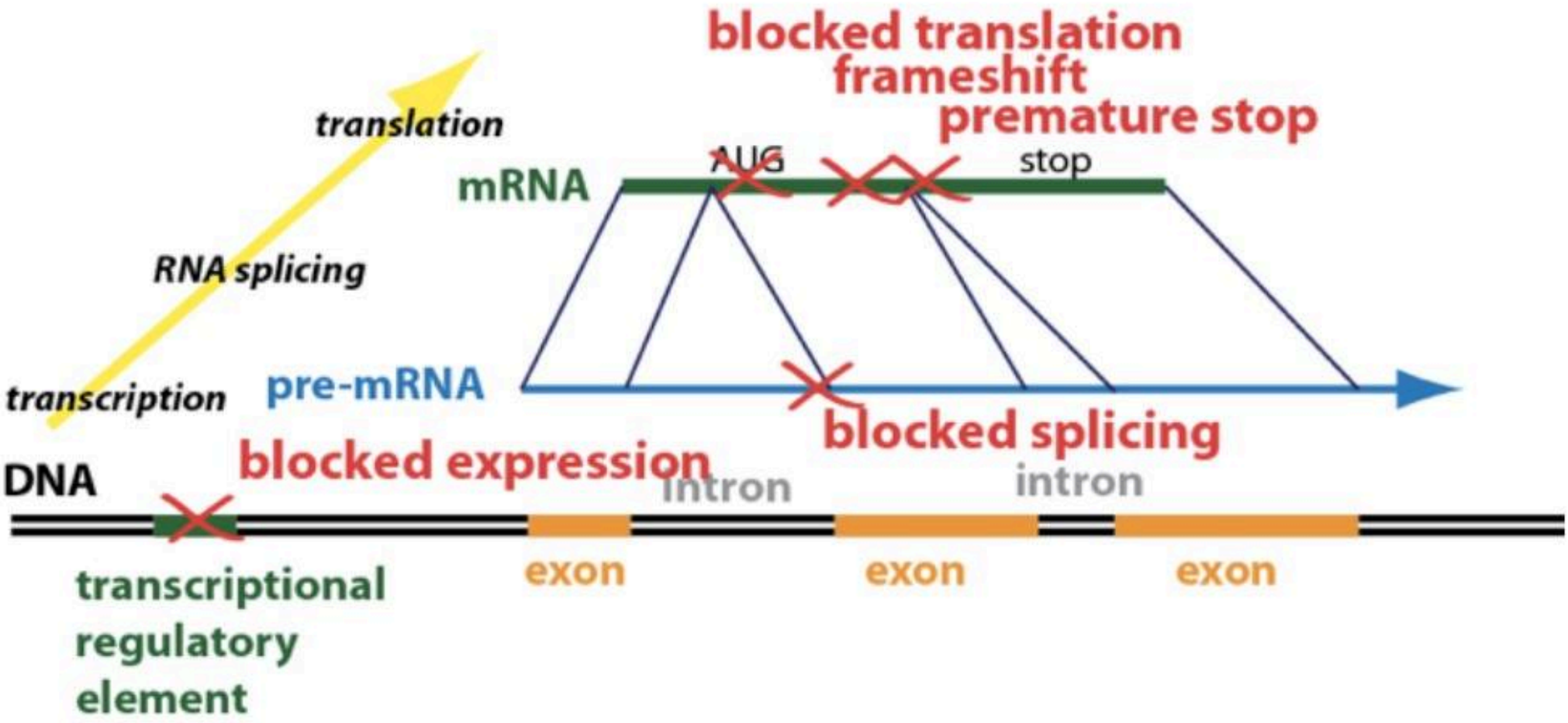
Sequence constraints are different for non-coding RNA

species 1



species 2

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
A(lf)	-	mutant
A(lf)	A	wild type
+	A	wild type
A(lf)	species2-A	wild type
+	species2-A	wild type

Genome sizes and gene numbers

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

https://doi.org/10.1038/s41586-020-2876-6

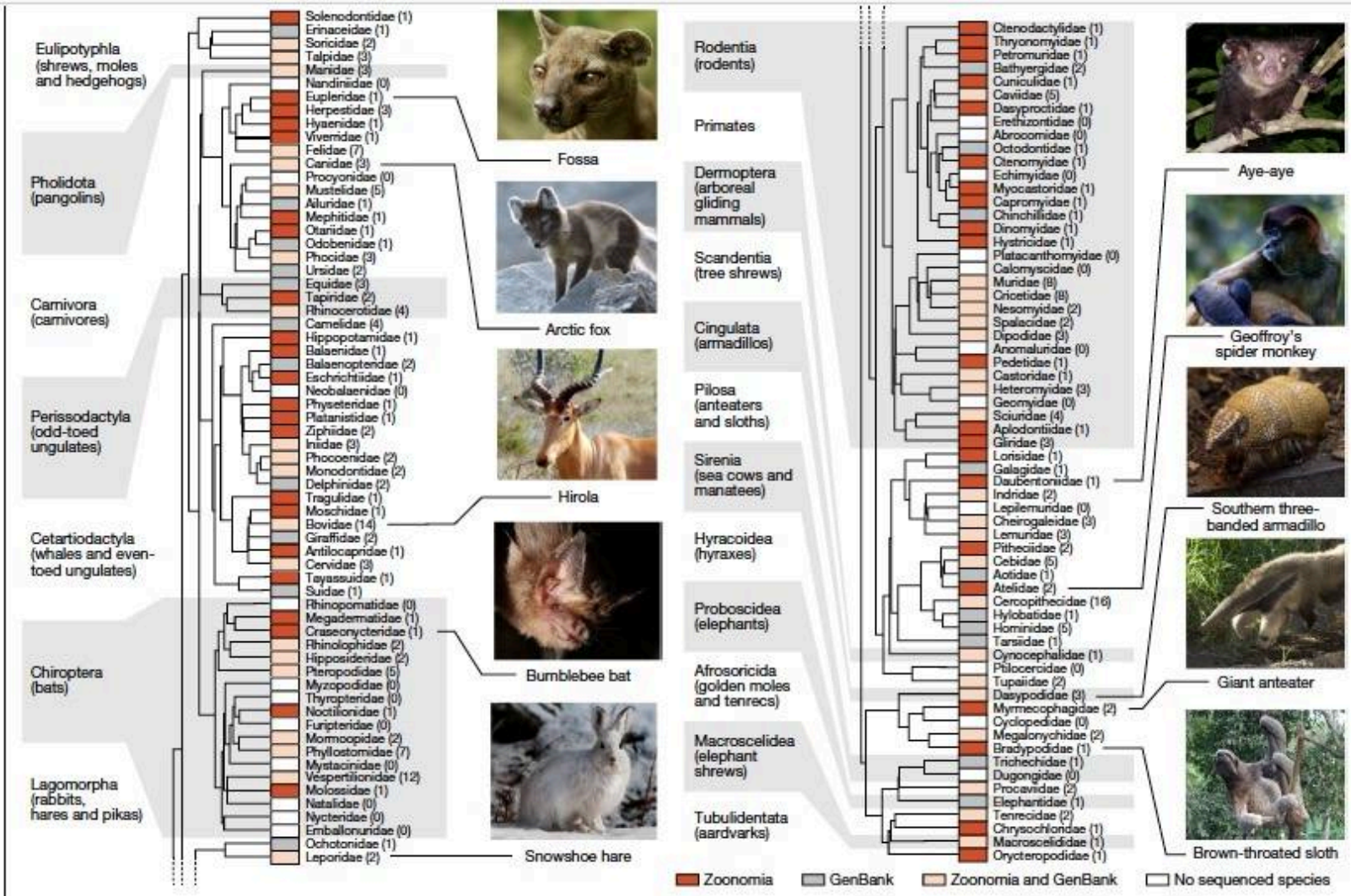


Fig. 1 | The Zoonomia Project brings the fraction of eutherian families that are represented by at least one assembly to 83%. Phylogenetic tree of the mammalian families in the Zoonomia Project alignment, including both our new assemblies and all other high-quality mammalian genomes publicly available 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 mammal[48], 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); hirola, JRProbert/Wikimedia (CC BY-SA); bumblebee bat, Sébastien J. Puechmaille (CC BY-SA); snowshoe hare, Denali National Park and Preserve/Wikimedia (public domain); aye-aye, Tom Junek/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).

https://doi.org/10.1038/s41586-020-2873-9

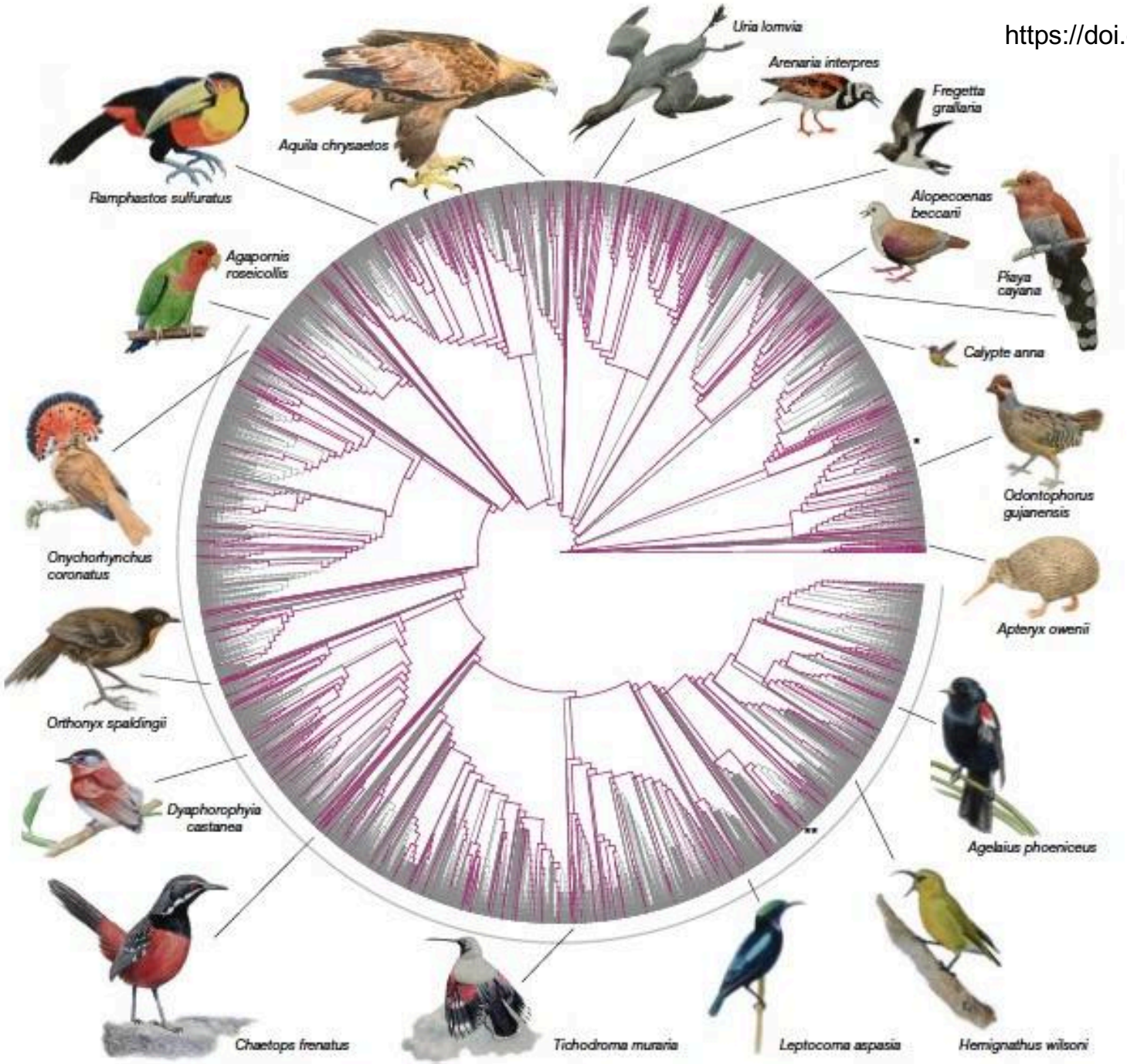


Fig. 1 | Newly sequenced genomes densely cover the bird tree of life. The 10,135 bird species[11,12] are shown on a draft phylogeny that synthesizes taxonomic and phylogenetic information[36] (Supplementary Data). In total, 363 species, covering 92.4% of all families, now have at least 1 genome assembly 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?

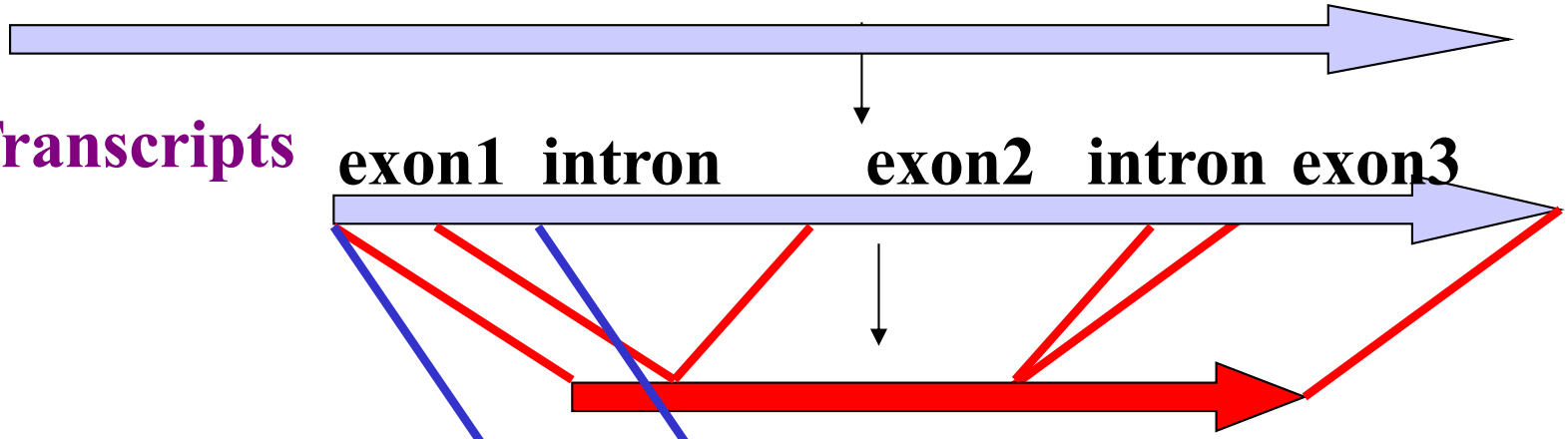
Genome sizes and gene numbers

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

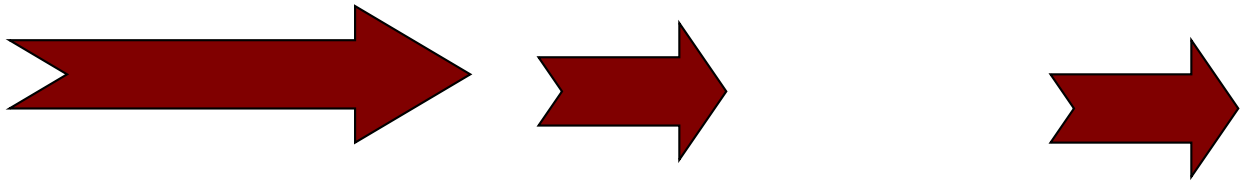
Complex genes

A. More Regulatory Sites

B. More Transcripts

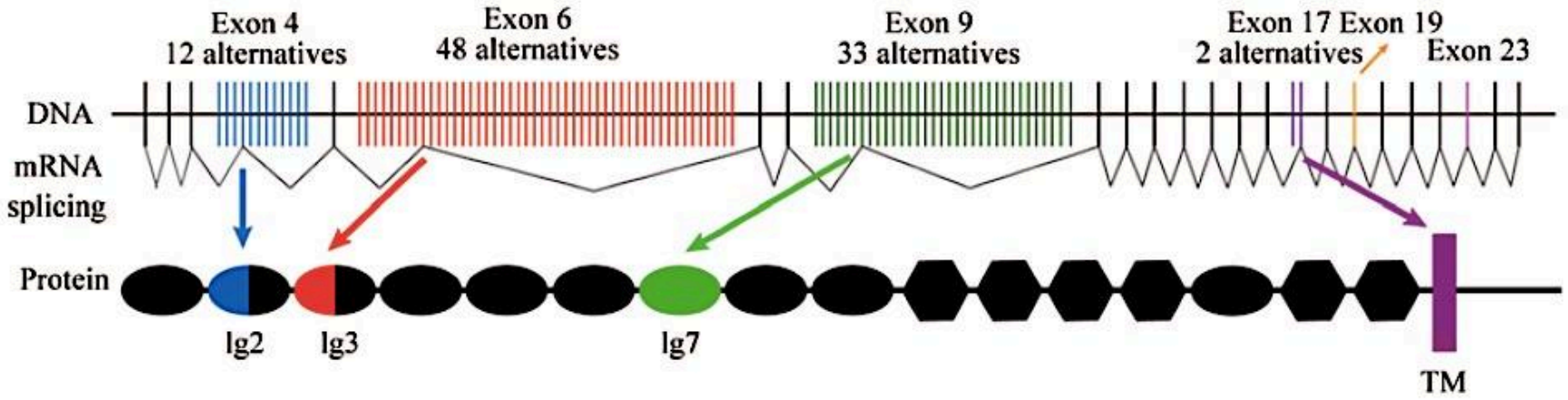


C. Alternative Splicing



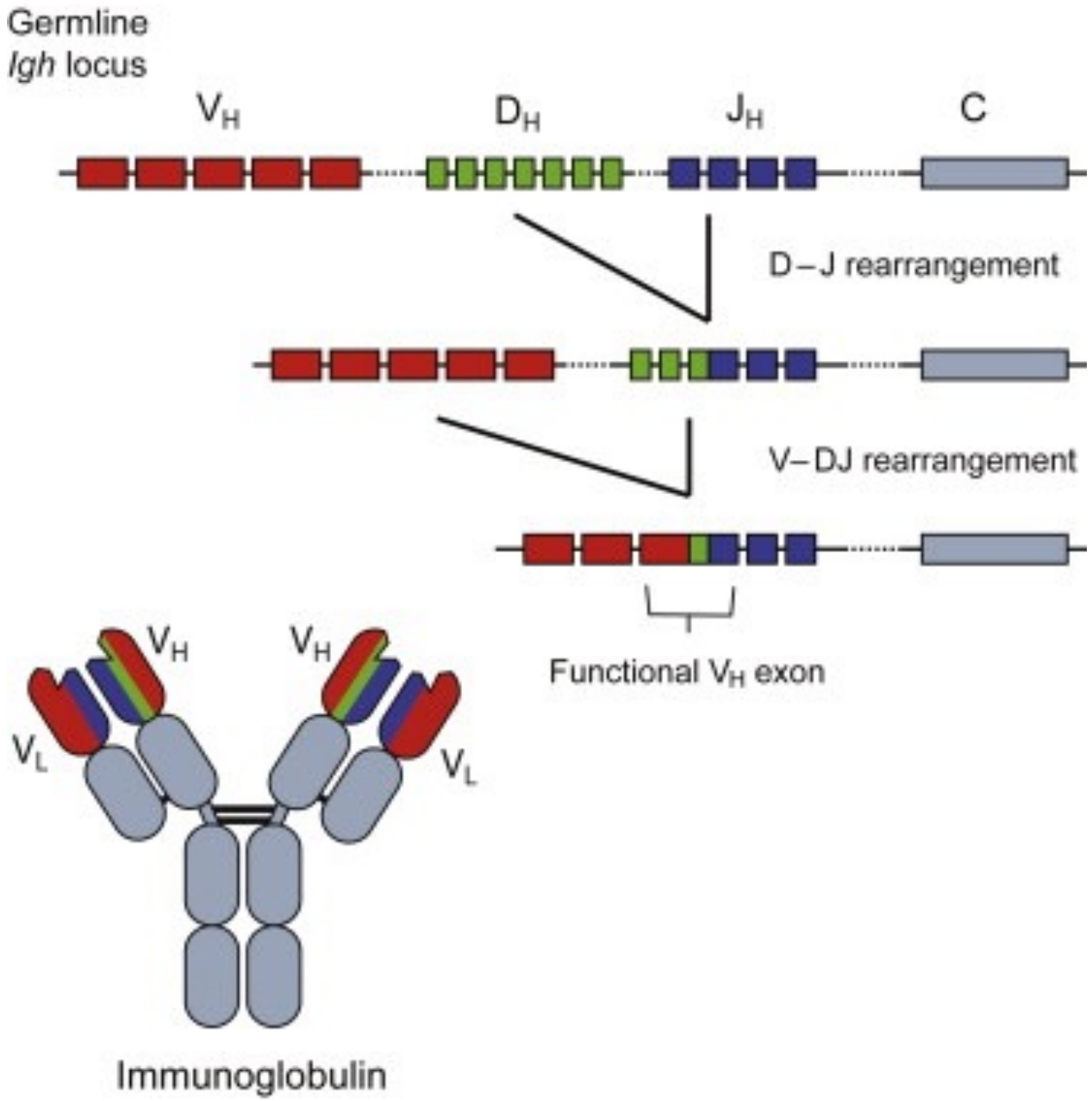
MULTIPLE PROTEINS MADE IN DIFFERENT AMOUNTS AND PLACES

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 largly 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 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.

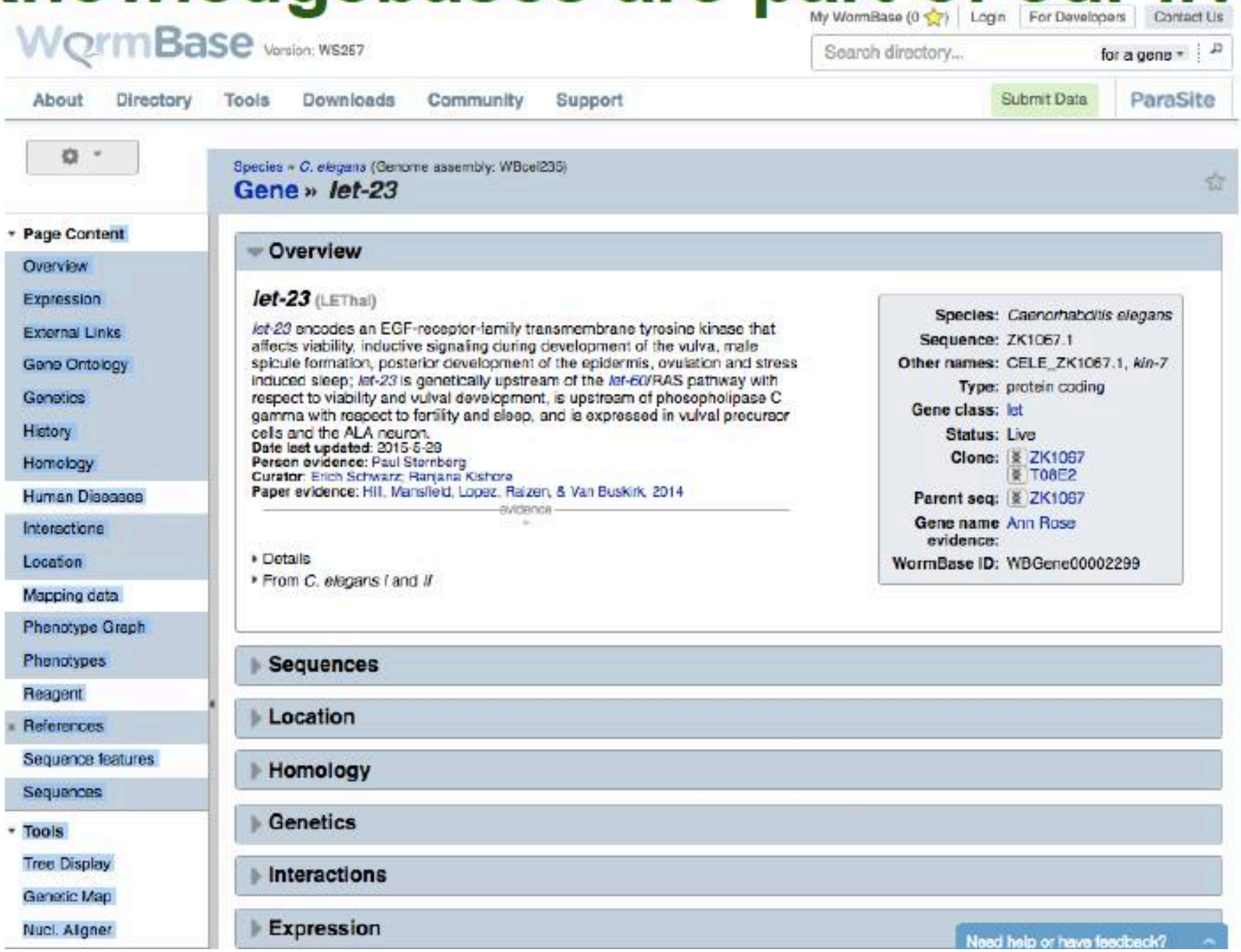
There are Model Organism Genome Databases such as

WormBase (www.wormbase.org)

FlyBase (www.flybase.org)

Yeast (yeastgenome.org)

Knowledgebases are part of our lives



ZFIN

Search

Research | General Information | ZIRC

Downloads Login

Home Genes / Markers / Clones BLAST GBrowse Expression Antibodies Mutants / Knockdowns / Tg Constructs Anatomy / GO / Human Disease Publications Maps

ZFIN ID: ZDB-GENE-030916-1

Your Input Welcom

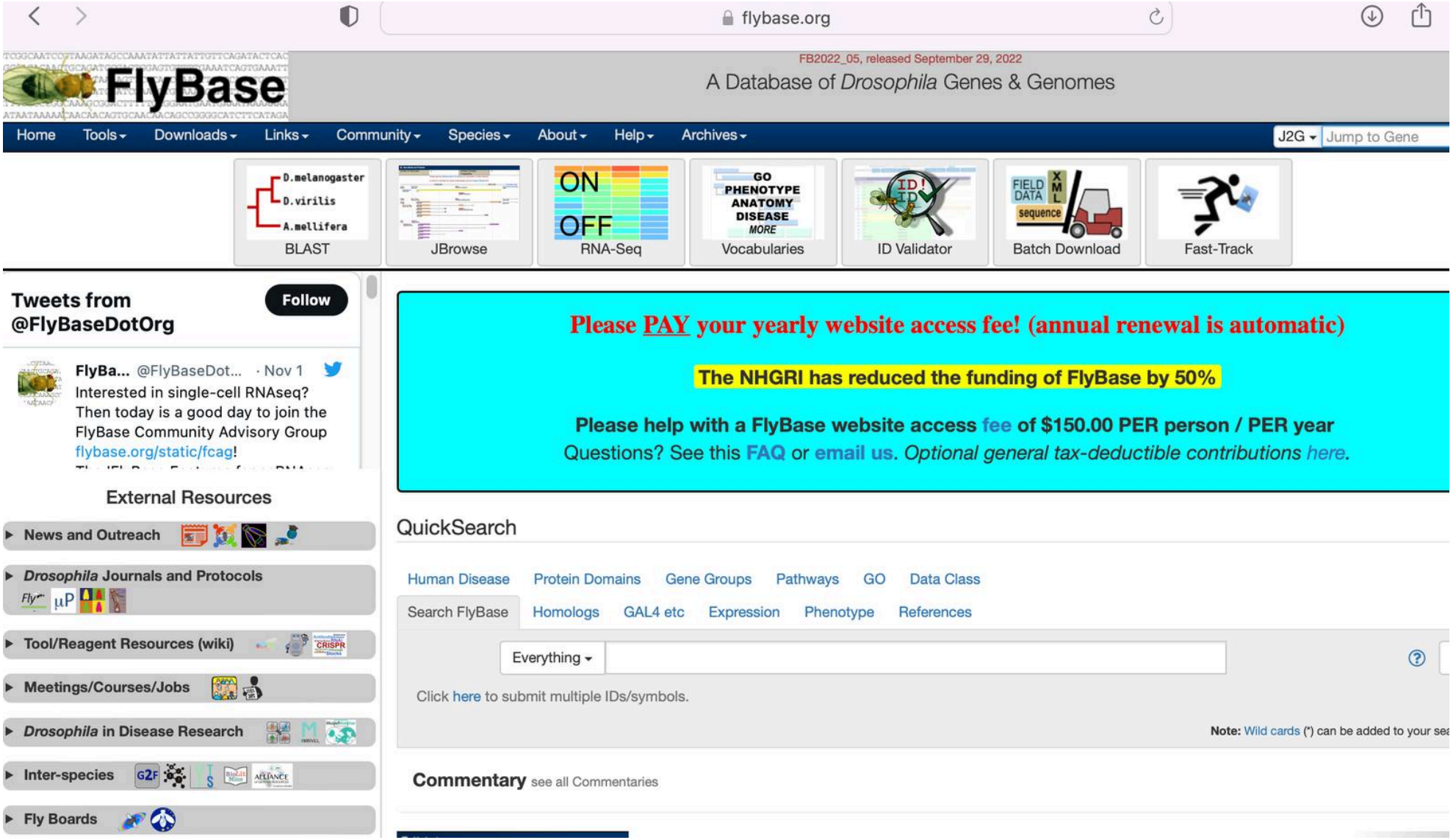
Gene Name: ***epidermal growth factor receptor a (erythroblastic leukemia viral (v-erb-b) oncogene homolog, avian)***
Gene Symbol: ***egfra***
Sequence Ontology ID : SO:0000704
Previous Names: *egfr*, *erbb1a* (1)
Location: Chr: 2 Mapping Details/Browsers
Nomenclature History

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) (orde this)
sa1683	Point Mutation	Unknown	Premature Stop	ENU	Zebrafish International Resource Center (ZIRC) (orde this)
sa11905	Point Mutation	Unknown	Premature Stop	ENU	European Zebrafish Resource Center (EZRC) (order this) Zebrafish International Resource Center (ZIRC) (orde this)
sa14931	Point Mutation	Unknown	Premature Stop	ENU	European Zebrafish Resource Center (EZRC) (order this) Zebrafish International Resource Center (ZIRC) (orde this)
sa15365	Point Mutation	Unknown	Premature Stop	ENU	European Zebrafish Resource Center (EZRC) (order

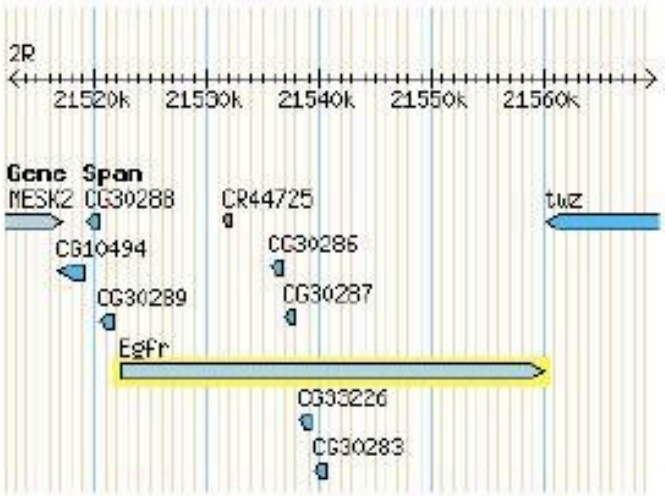

flybase.org
FlyBase
FB2022_05, released September 29, 2022
A Database of Drosophila Genes & Genomes
Home
Tools
Downloads
Links
Community
Species
About
Help
Archives
J2G
Jump to Gene
BLAST
JBrowse
RNA-Seq
Vocabularies
ID Validator
Batch Download
Fast-Track
Tweets from @FlyBaseDotOrg
Follow
FlyBa... @FlyBaseDot... · Nov 1
Interested in single-cell RNAseq? Then today is a good day to join the FlyBase Community Advisory Group flybase.org/static/fcag!
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.
External Resources
News and Outreach
Drosophila Journals and Protocols
Tool/Reagent Resources (wiki)
Meetings/Courses/Jobs
Drosophila in Disease Research
Inter-species
Fly Boards
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 sea
Commentary see all Commentaries

General Information			
Symbol	Dmel\Egfr	**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, flb, Elp, dEGFR, Egf-r, c-erbB		
Gene Snapshot	Epidermal growth factor receptor (Egfr) is the transmembrane tyrosine kinase receptor for signaling ligands in the TGFα family (grk, spi, vn, and Krn), 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

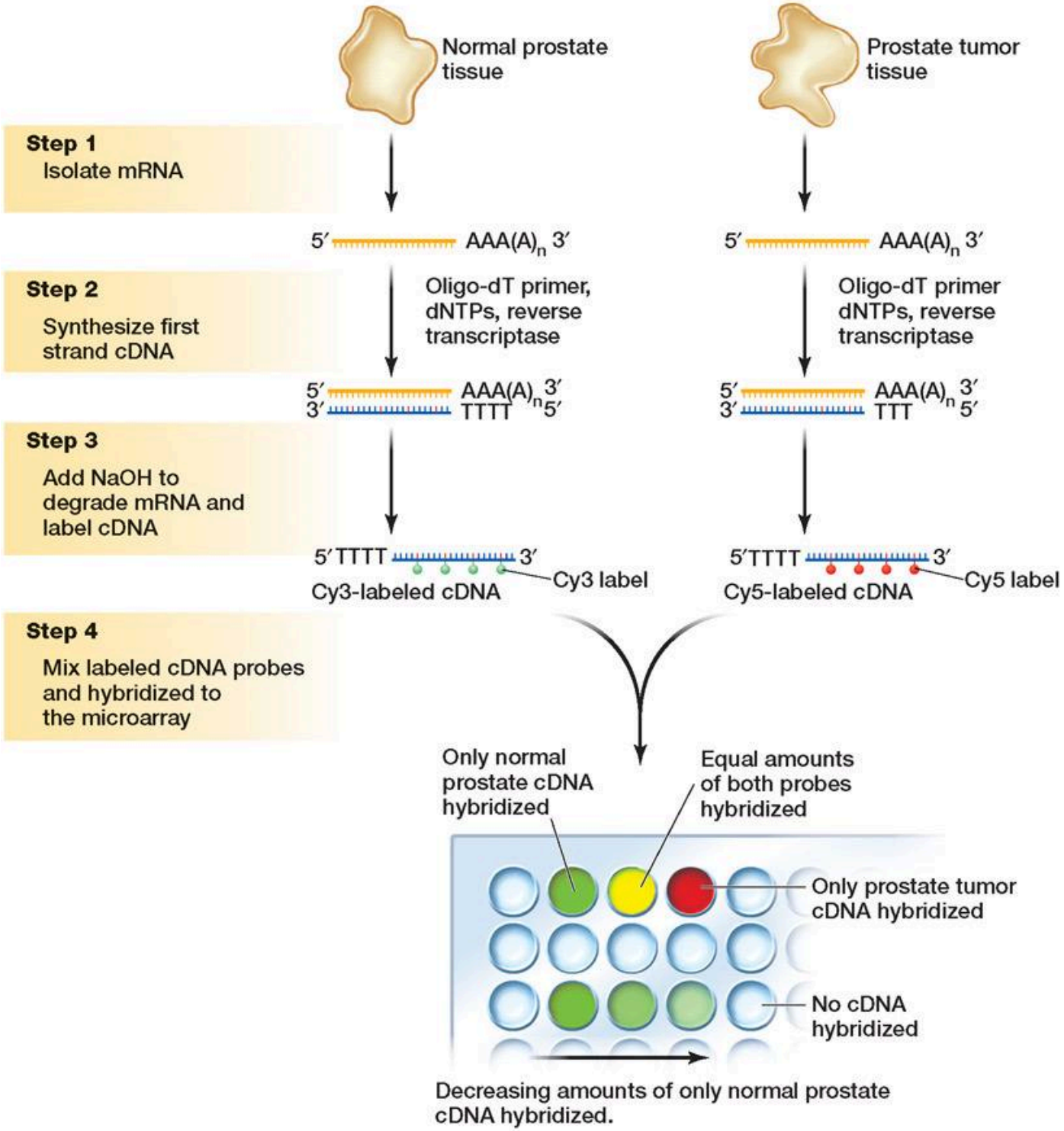
The following external sites may use different assemblies or annotations than FlyBase.

www.alliancegenome.org

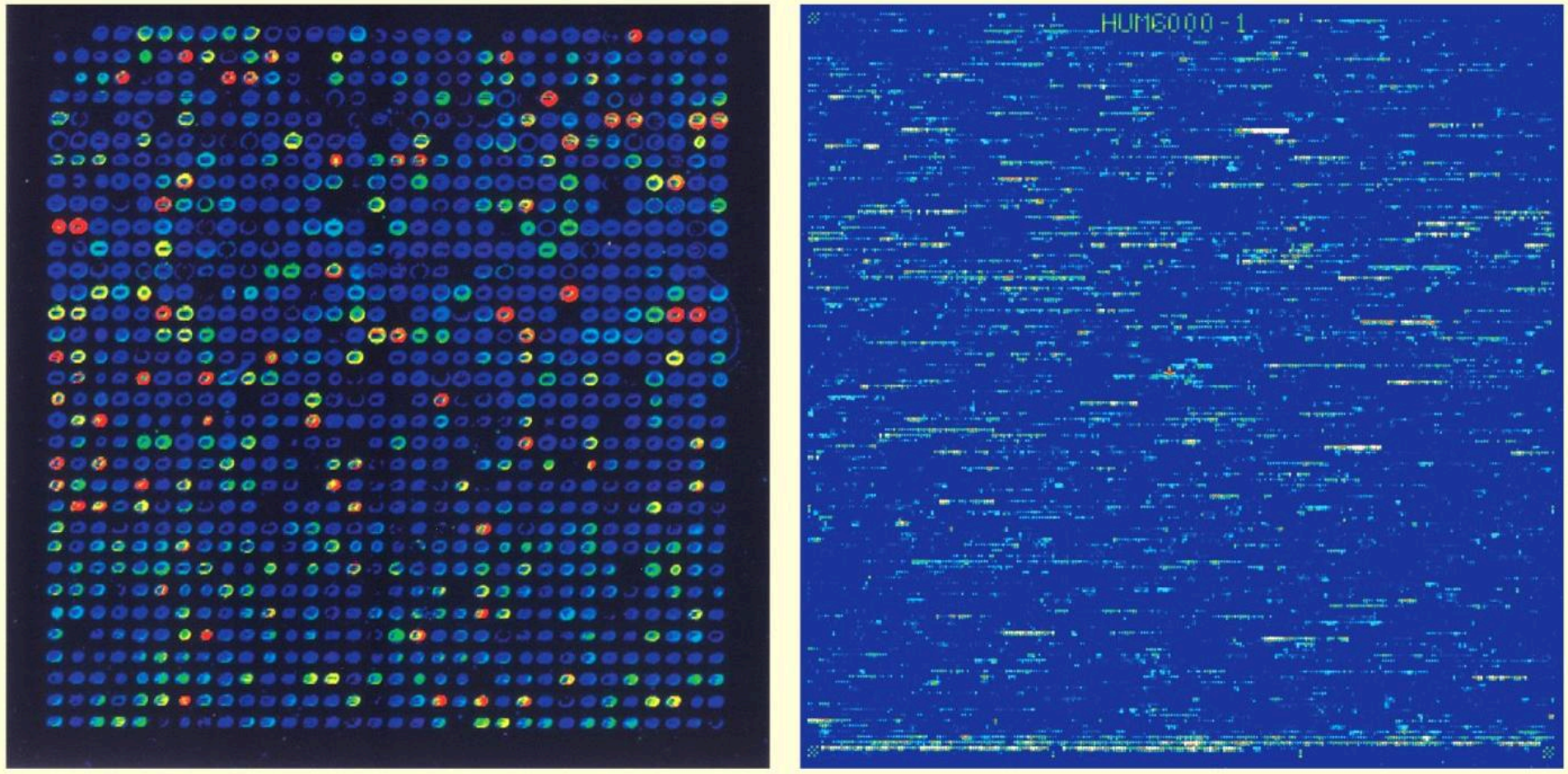
one-stop shopping for genes

NIH 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

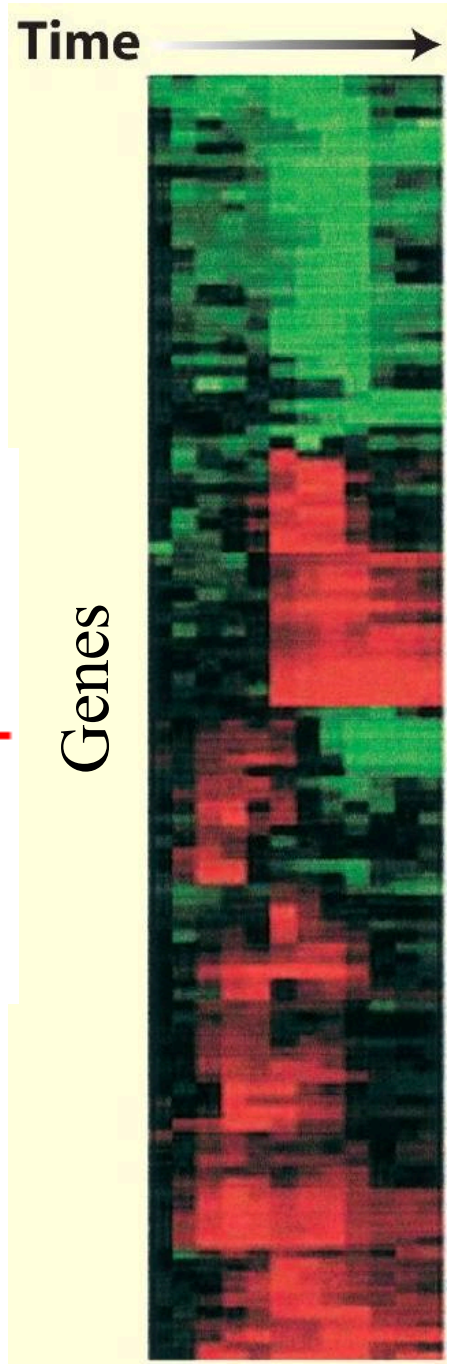


Transcriptional profiling

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

Conditions

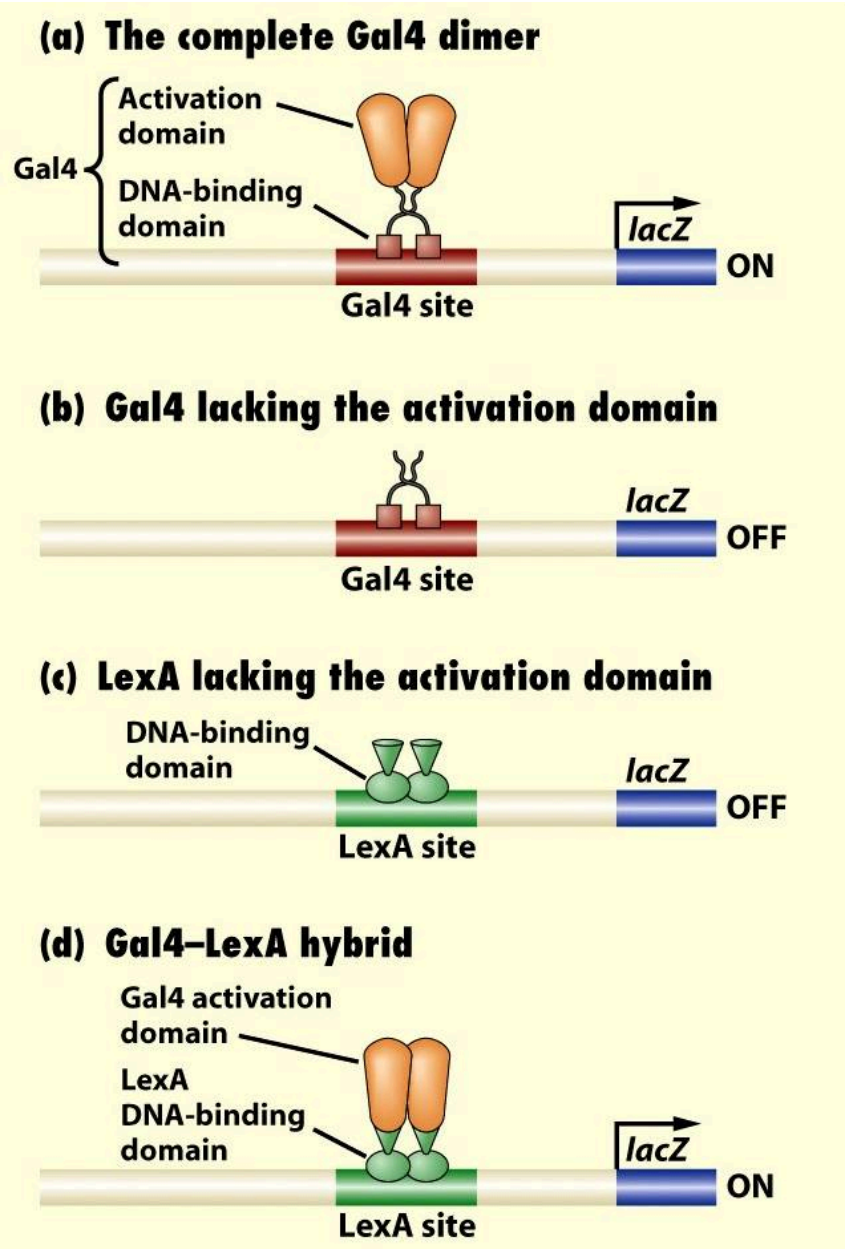
Genes	a	b	c	d	e	f
1	-	-	0	0	0	0
4	-	-	0	0	0	0
7	-	-	0	0	0	0
5	+	-	+	0	0	+
2	+	-	+	0	+	+
3	+	-	+	0	+	+
6	+	-	+	-	+	+



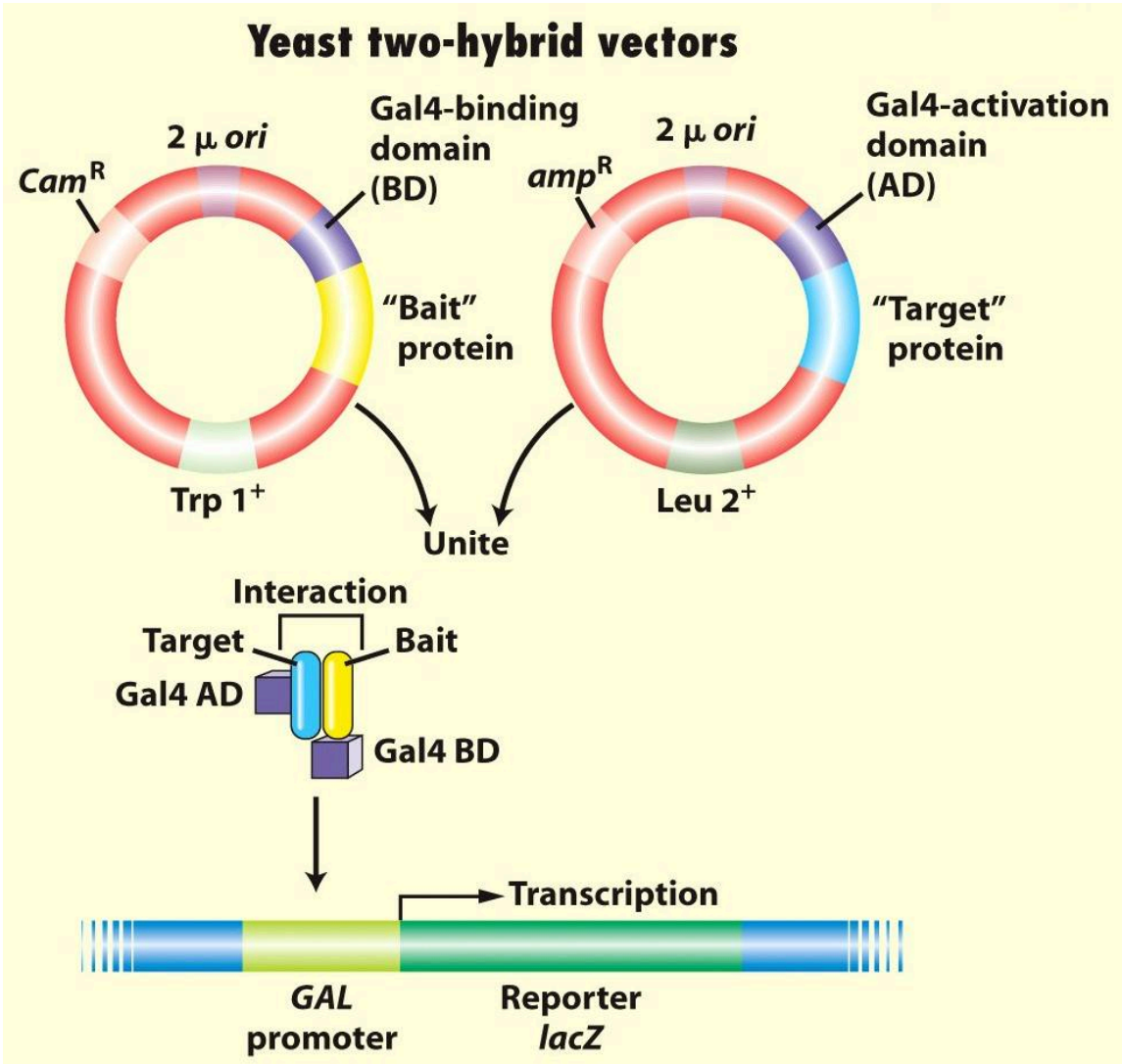
Gene expression pattern clustering algorithm

Compare normal and disease states

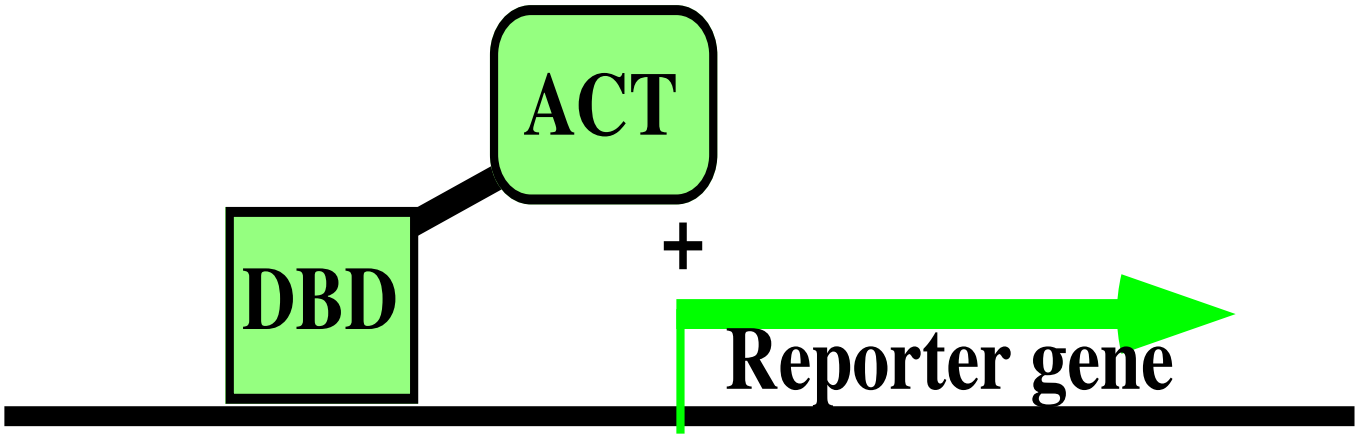
Identifying protein-protein interactions using the yeast two-hybrid system



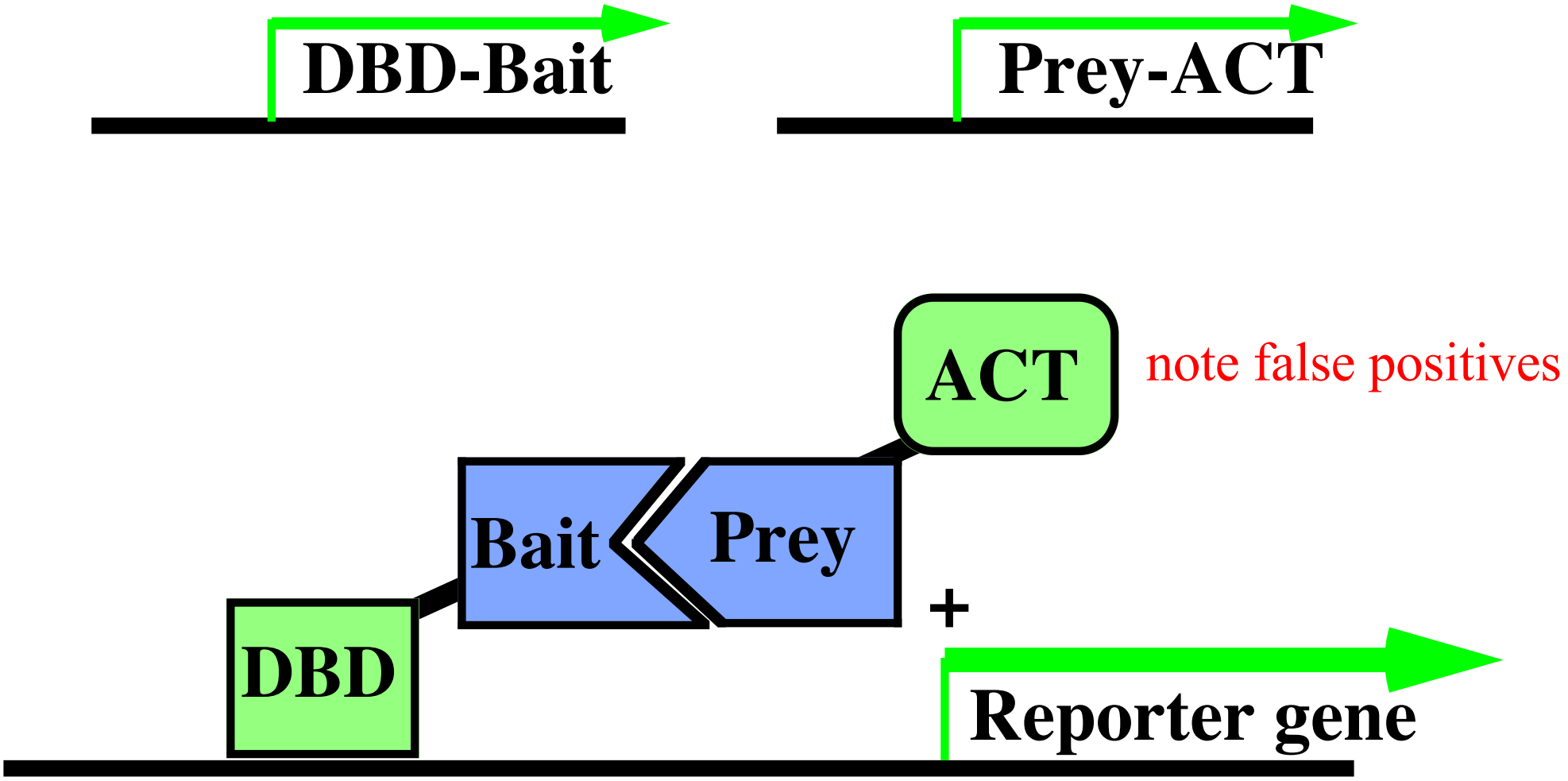
Transform plasmids into cells that are trp- and leu -, On minimal media. Only plasmid bearing cells survive

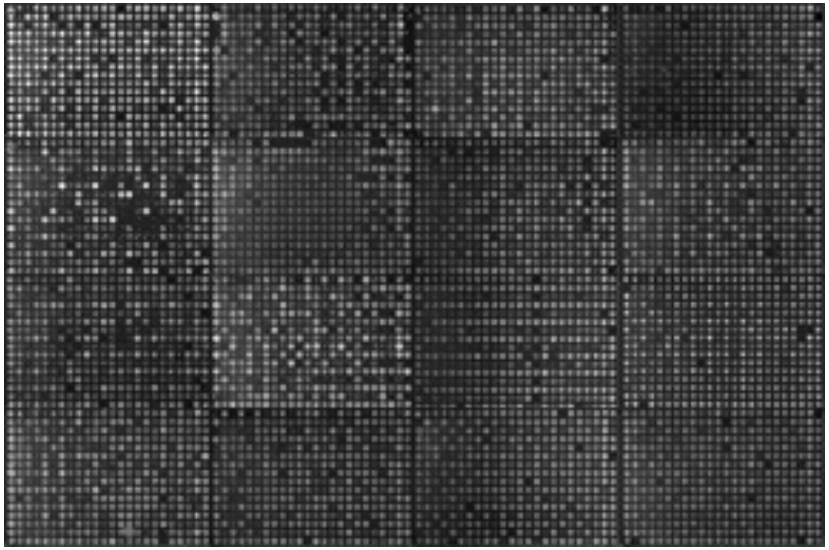


Bi-partite transcriptional activators

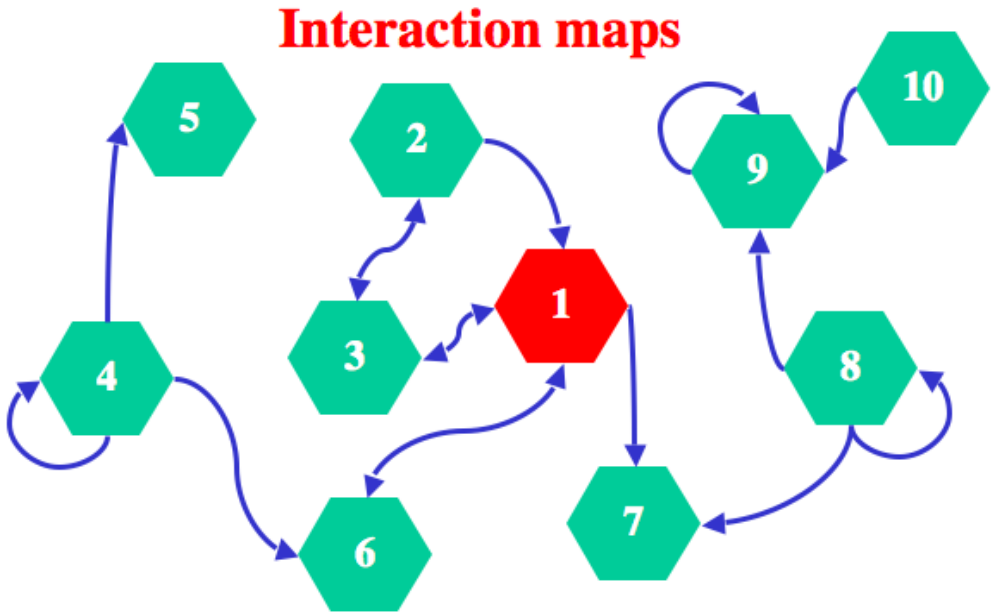


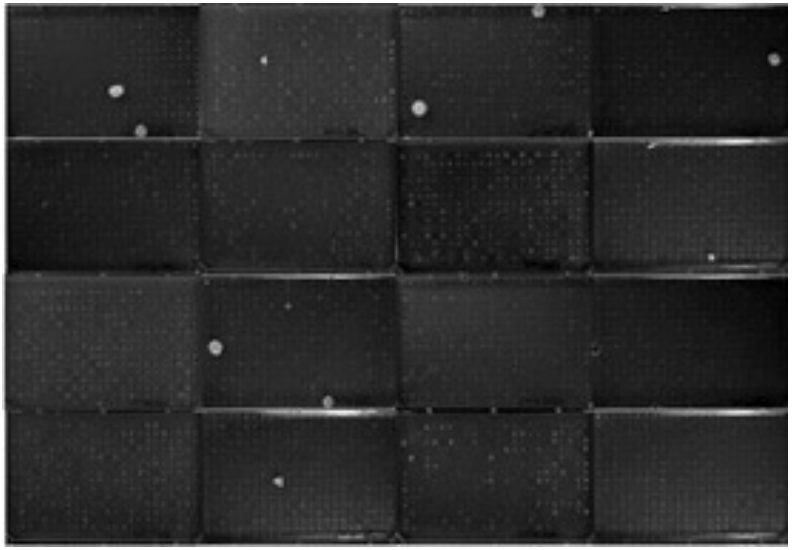
2-Hybrid Assay in yeast





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:

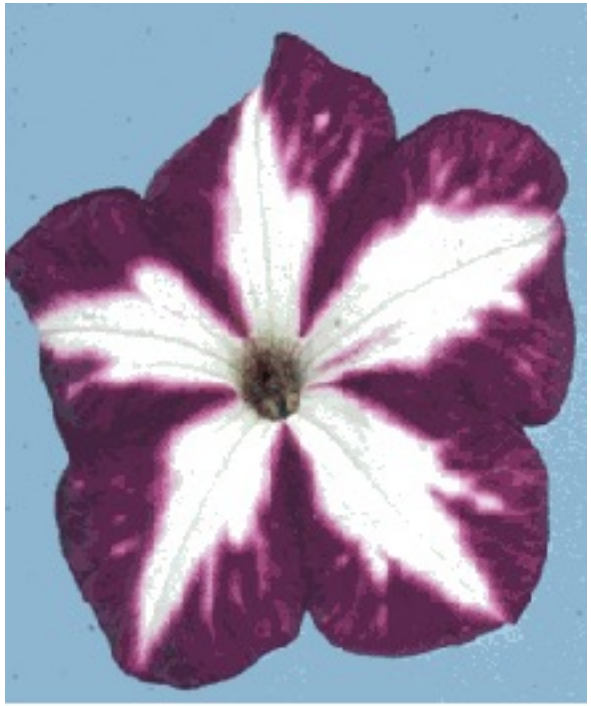
2006

RNA interference

Phenomena first observed in petunia

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

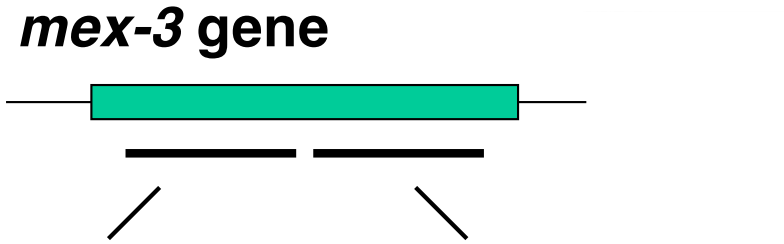
Caused the loss of pigment.



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

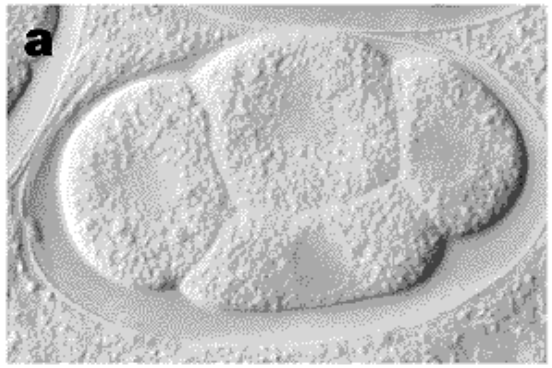
Embryogenesis in the worm C. elegans



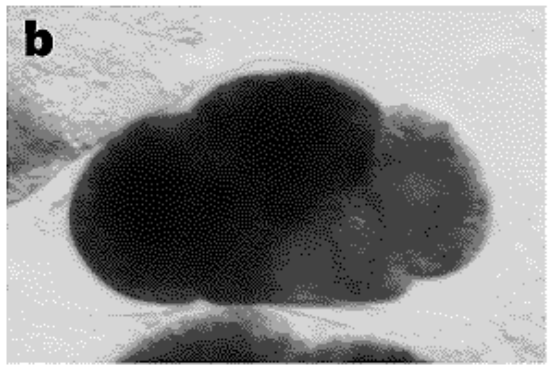
Injected dsRNA into hermaphrodite

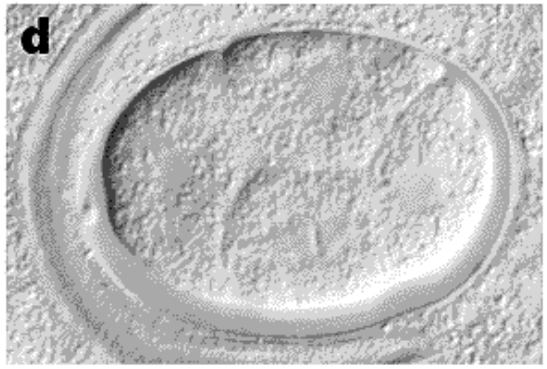
Antisense probe for her embryos

No probe



***mex-3* probe: wild type**





***mex-3* probe: dsRNA treatment**

General mechanism of RNAi

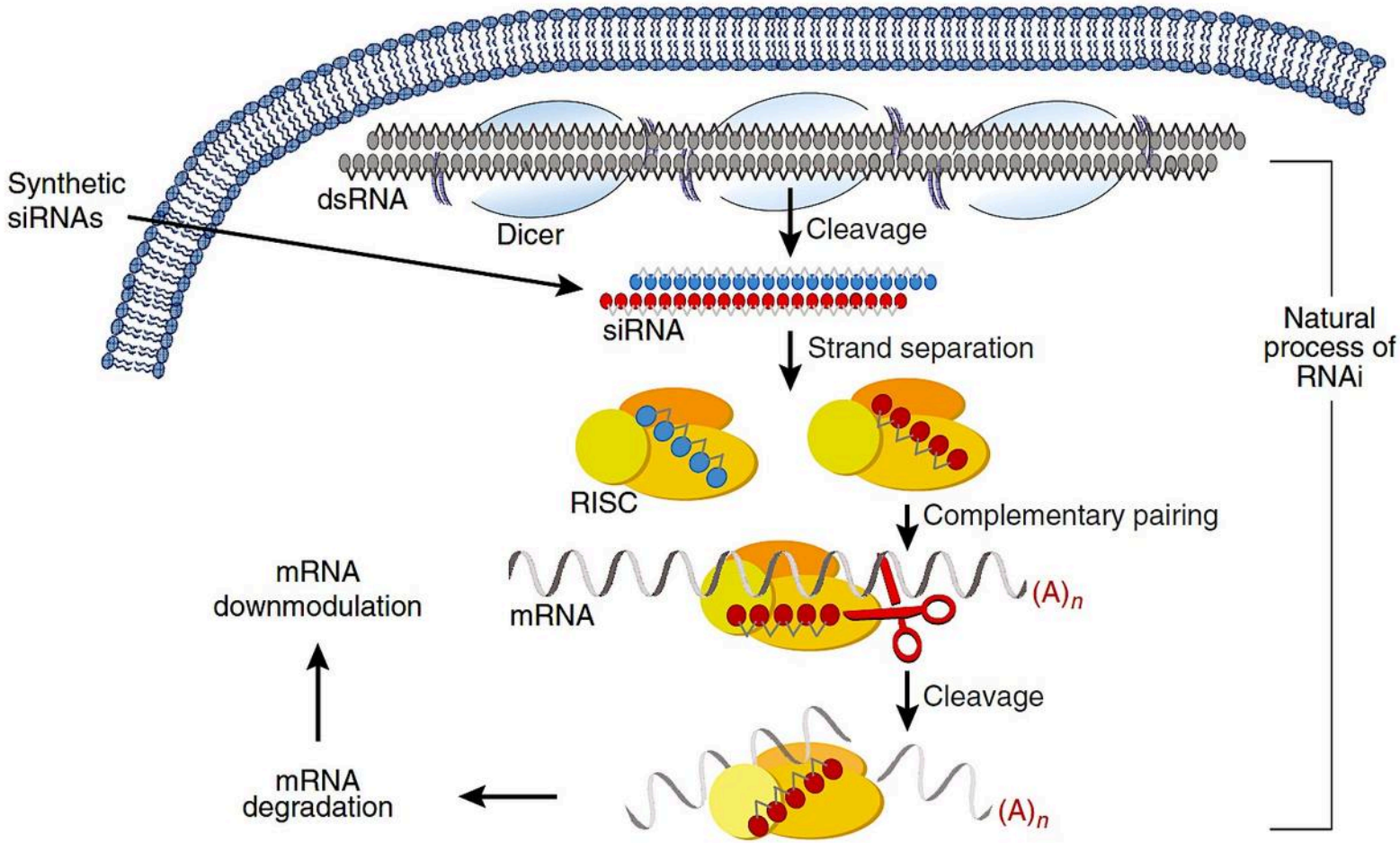


Figure 1 Cellular mechanism of RNA interference. Long double-stranded RNA (dsRNA) is cleaved, by the enzyme Dicer, into small interfering RNA (siRNA). These siRNAs are incorporated into the RNA-induced silencing complex (RISC), where the strands are separated. The RISC containing the guide or antisense 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:

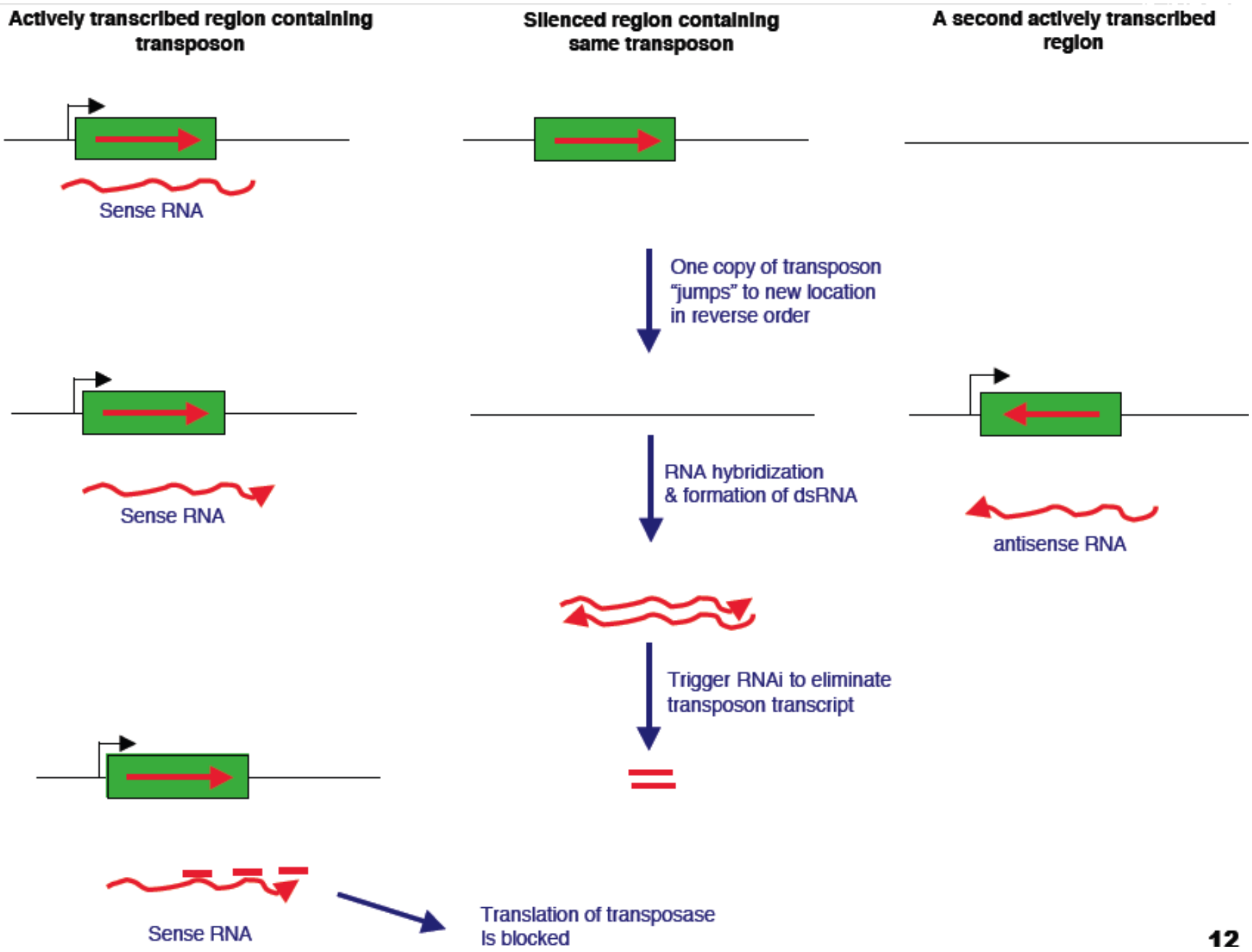
Transposon silencing

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
Sense RNA
One copy of transposon "jumps" to new location in reverse order
RNA hybridization & formation of dsRNA
antisense RNA
Trigger RNAi to eliminate transposon transcript
Sense RNA
Translation of transposase is blocked

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

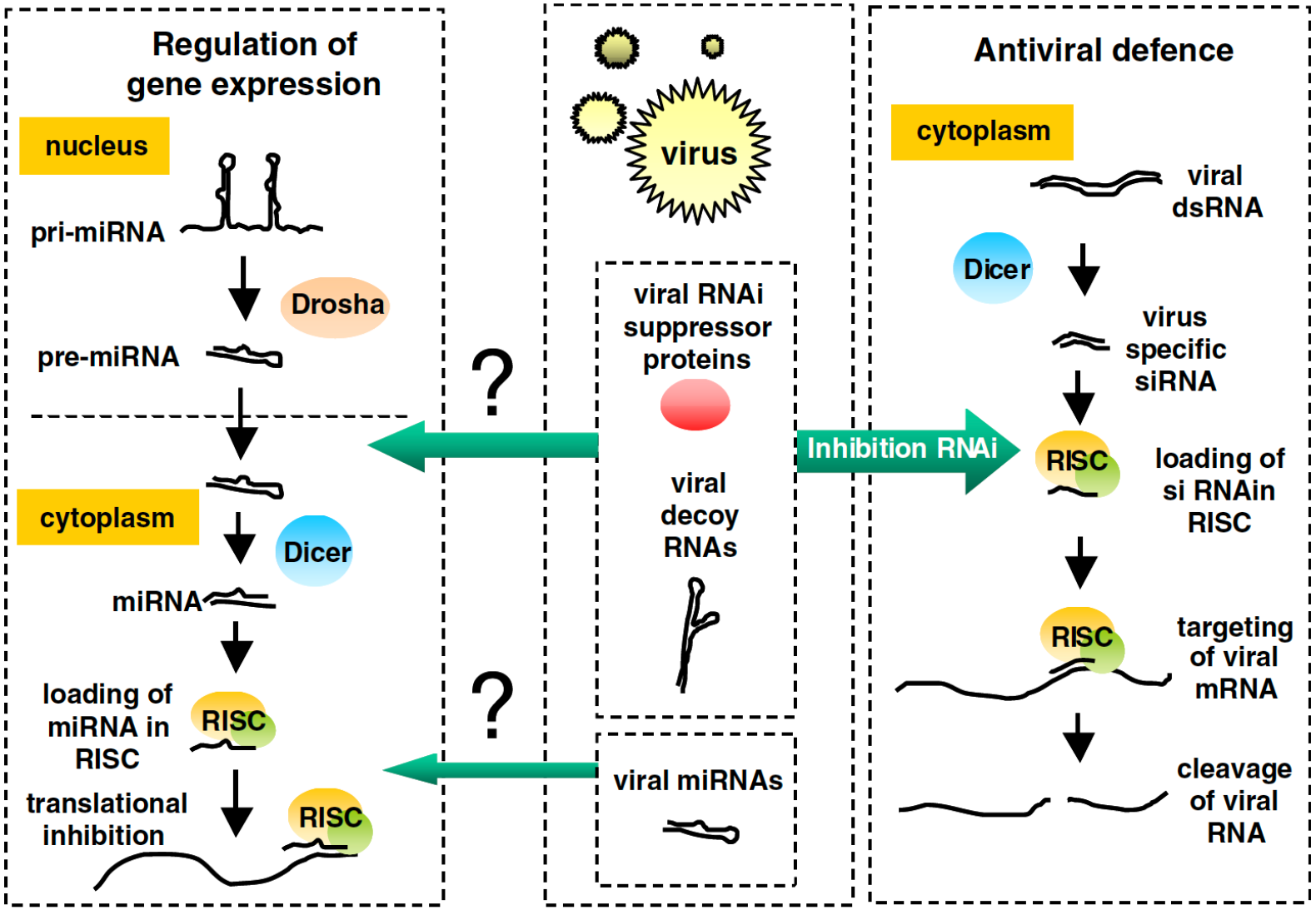
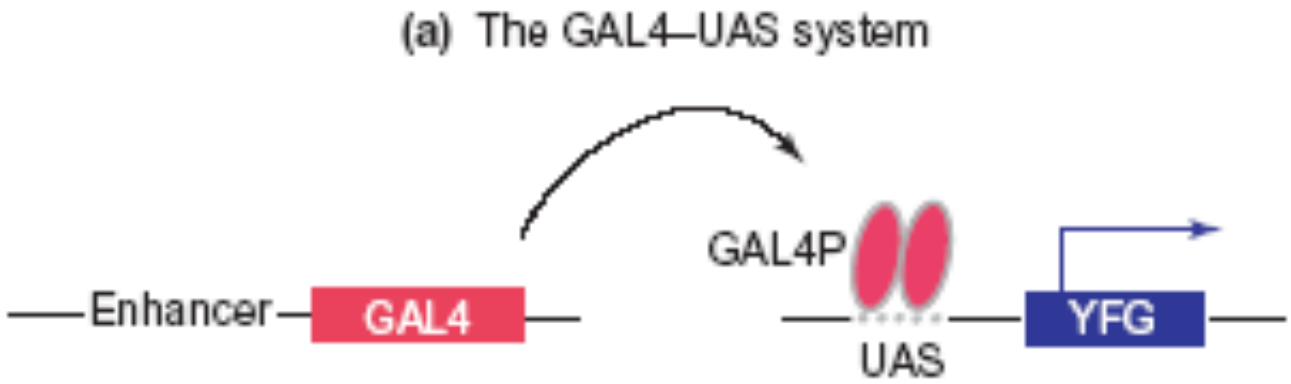


Fig. 1. Viral factors affecting cellular RNAi pathways. miRNA regulated gene expression starts with primiRNAs in the nucleus, which are processed into mature miRNAs by Drosha and Dicer (left panel). The antiviral RNAi response is triggered by virus-derived dsRNAs during infection (right panel). Viral 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).

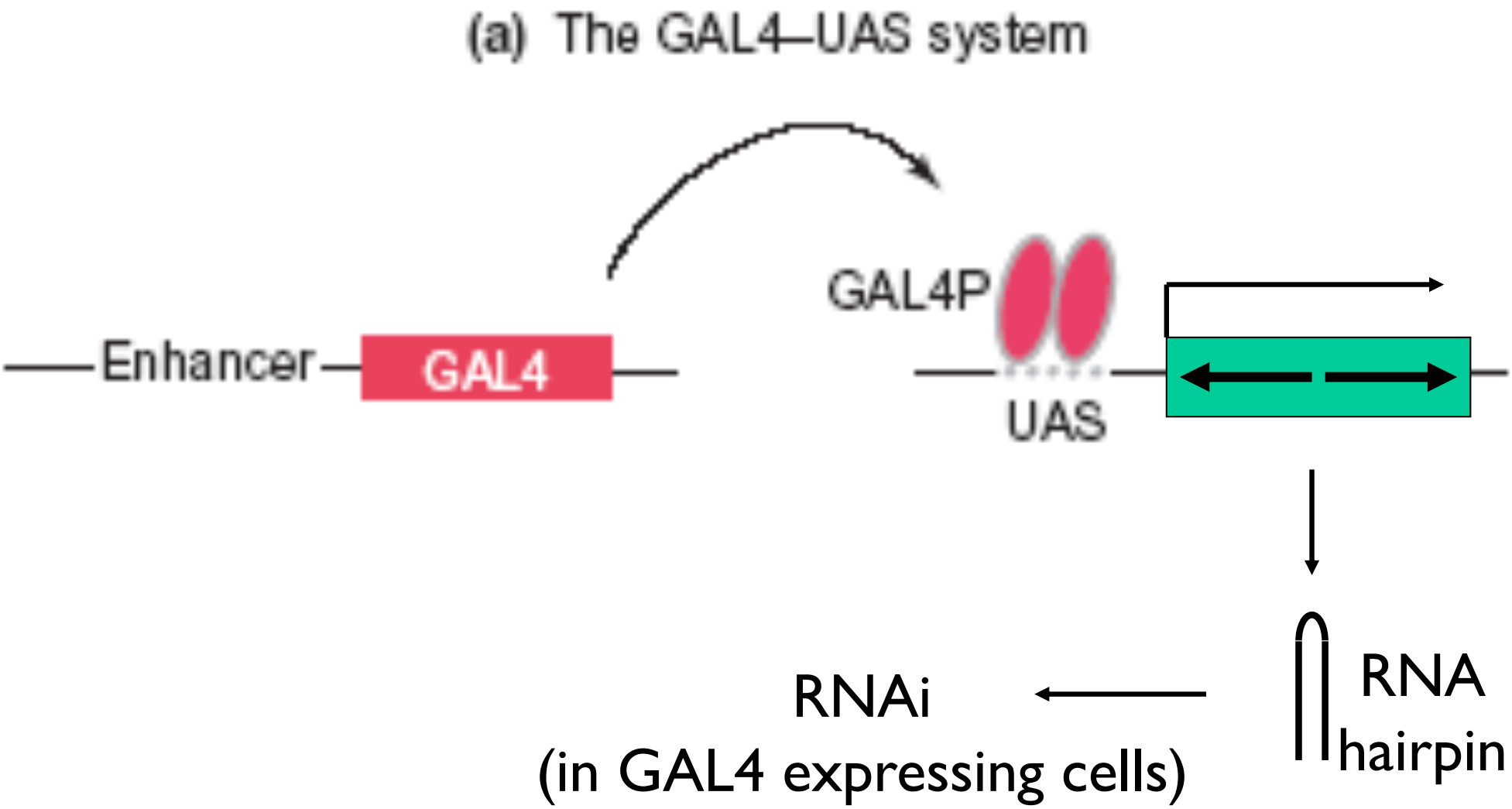
Remember the bipartite GAL4-UAS system



- 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.



Albert Lasker Basic Medical Research Award >



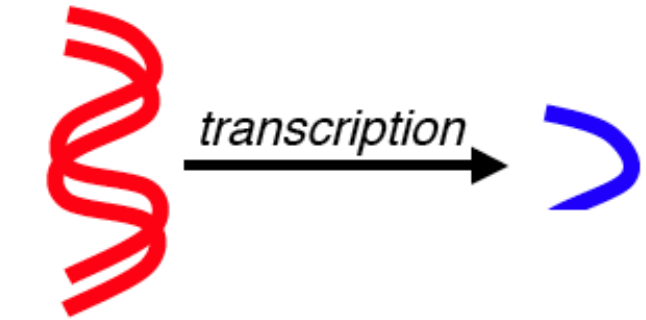
Victor Ambros David Baulcombe Gary Ruvkun

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

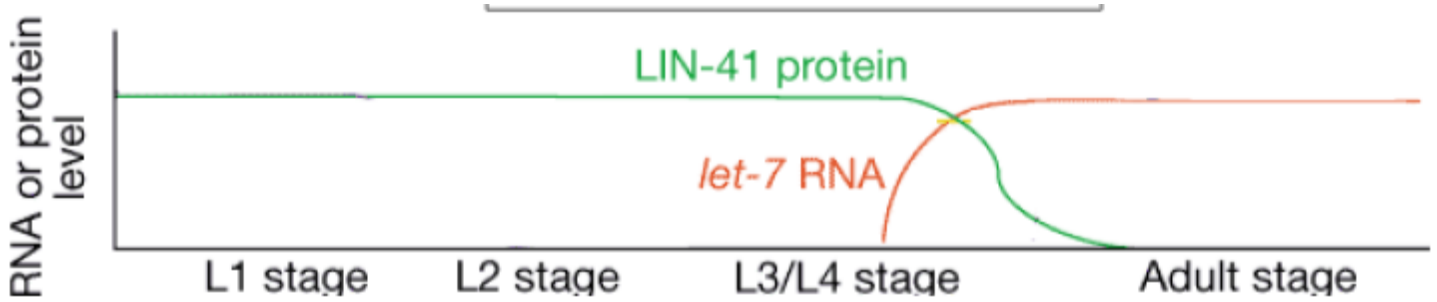
Rosalind C. Lee,*† Rhonda L. Feinbaum,*‡
and Victor Ambros†
Harvard University
Department of Cellular and Developmental Biology
Cambridge, Massachusetts 02138

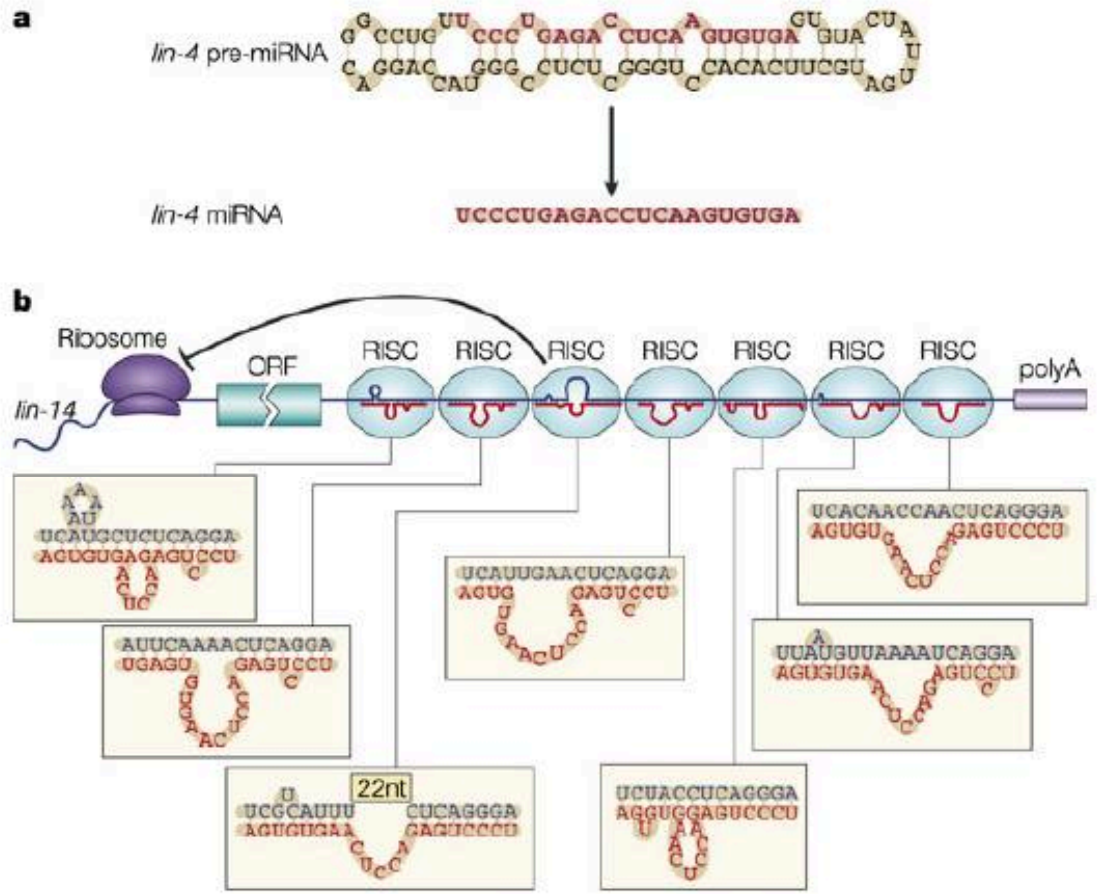
its Target

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

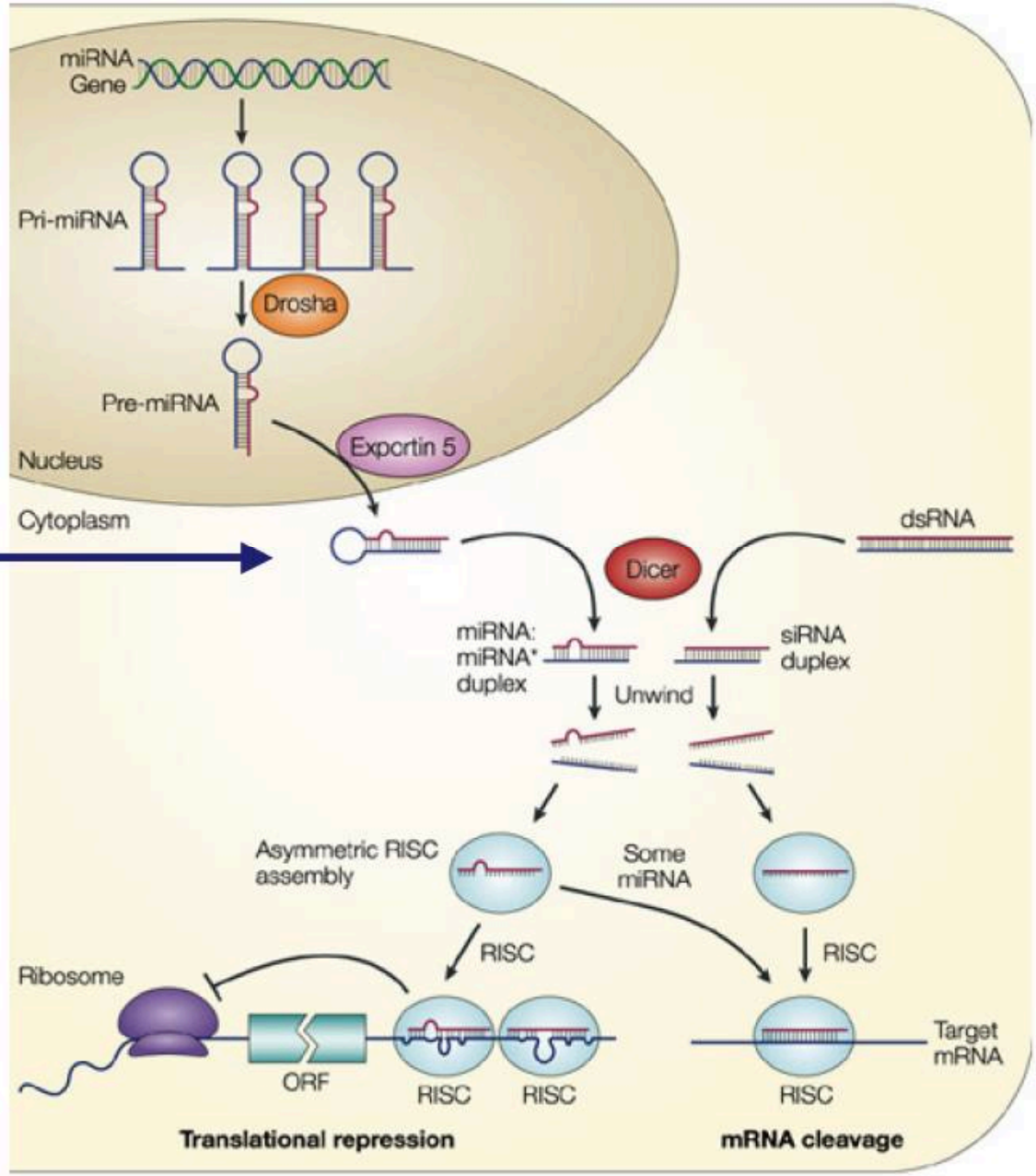




The same processed miRNA sequence (e.g. *lin-4*) can hybridize (not perfectly, but sufficiently) to multiple complimentary sites in the 3'UTR of target mRNAs (e.g. *lin-14*)

Processing pri-miRNA transcripts or exogenous dsRNA

Naturally occurring in the genome

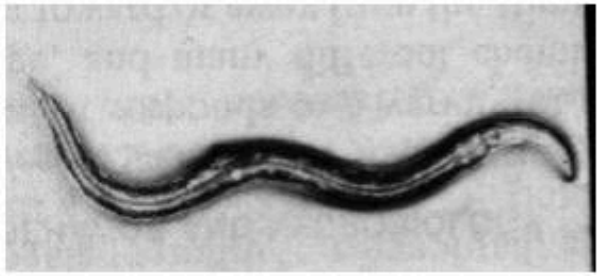


Introduced by the researcher experimentally as shRNA

Introduced by the researcher experimentally as dsRNA

MicroRNAs (miRNAs)--Not just in worms!

A. thaliana/
O. sativa
>80 miRNAs



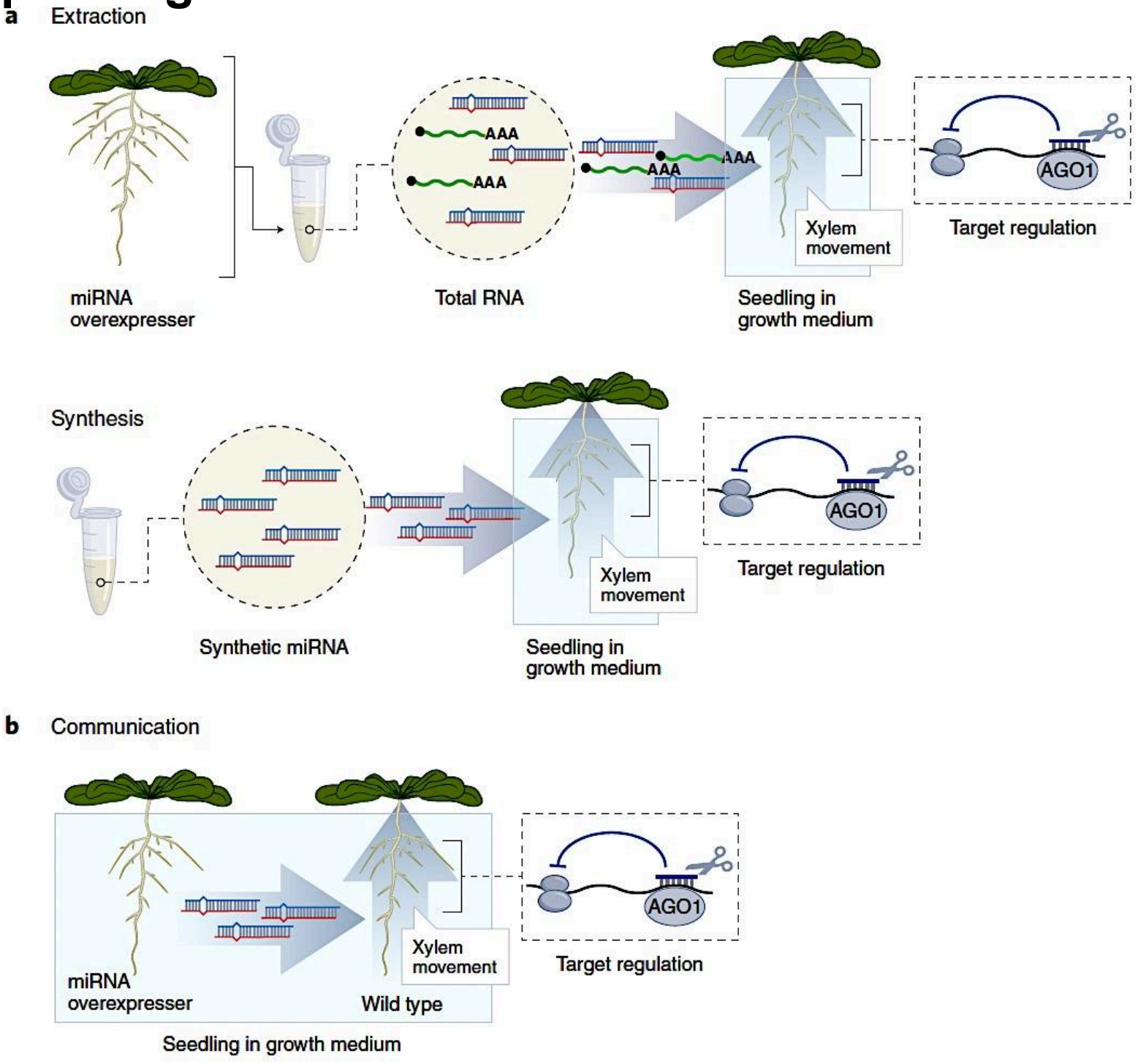
C. elegans
>100 miRNAs

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



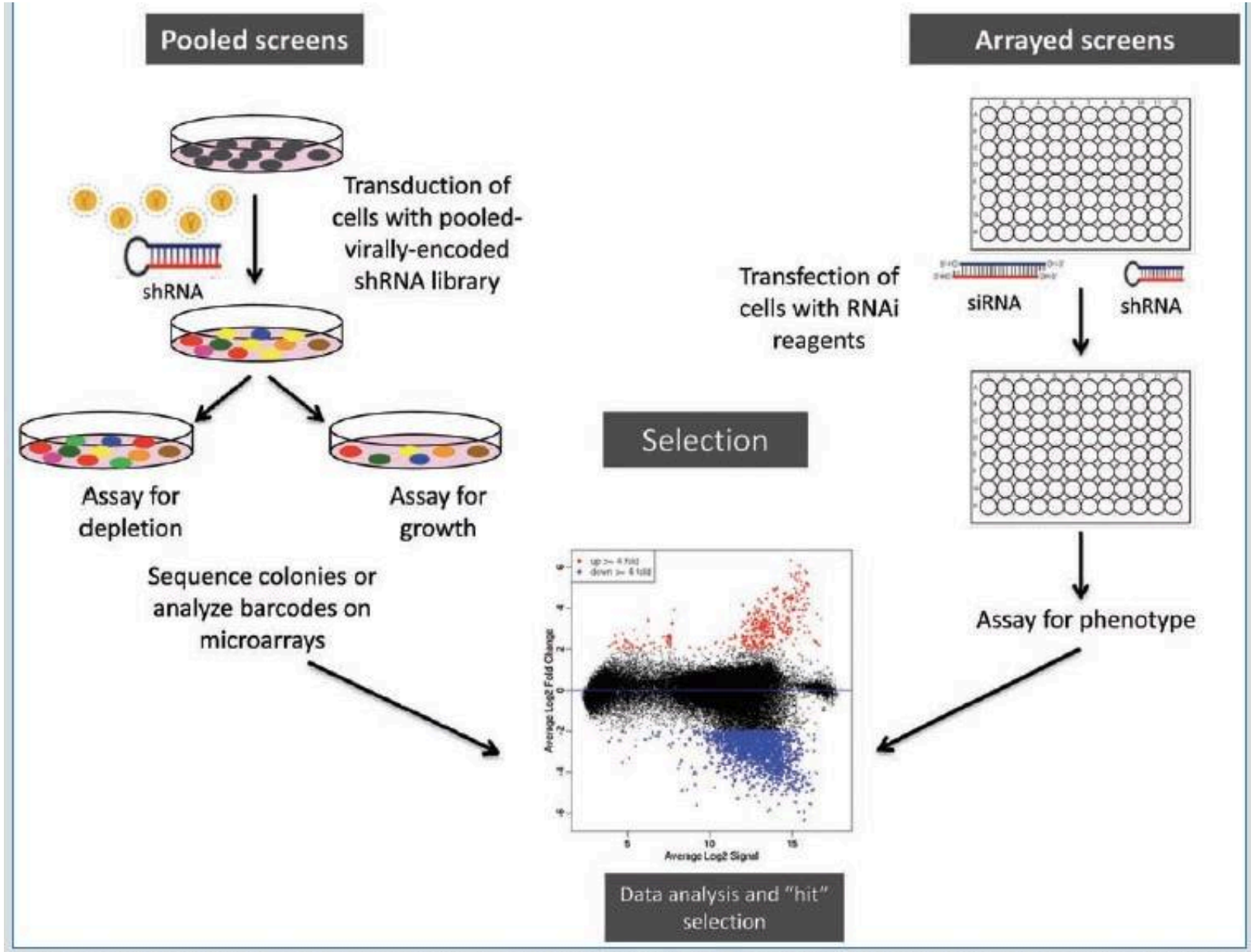


Figure 1 Schematic illustration of arrayed or pooled RNAi screens in cells. **Left panel.** Pooled-viral vectors encoding libraries of shRNAs targeting multiple genes can be used to transduce a target cell population in a single tissue culture dish. After selection for the desired phenotype, cells are 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. Adapted[18].

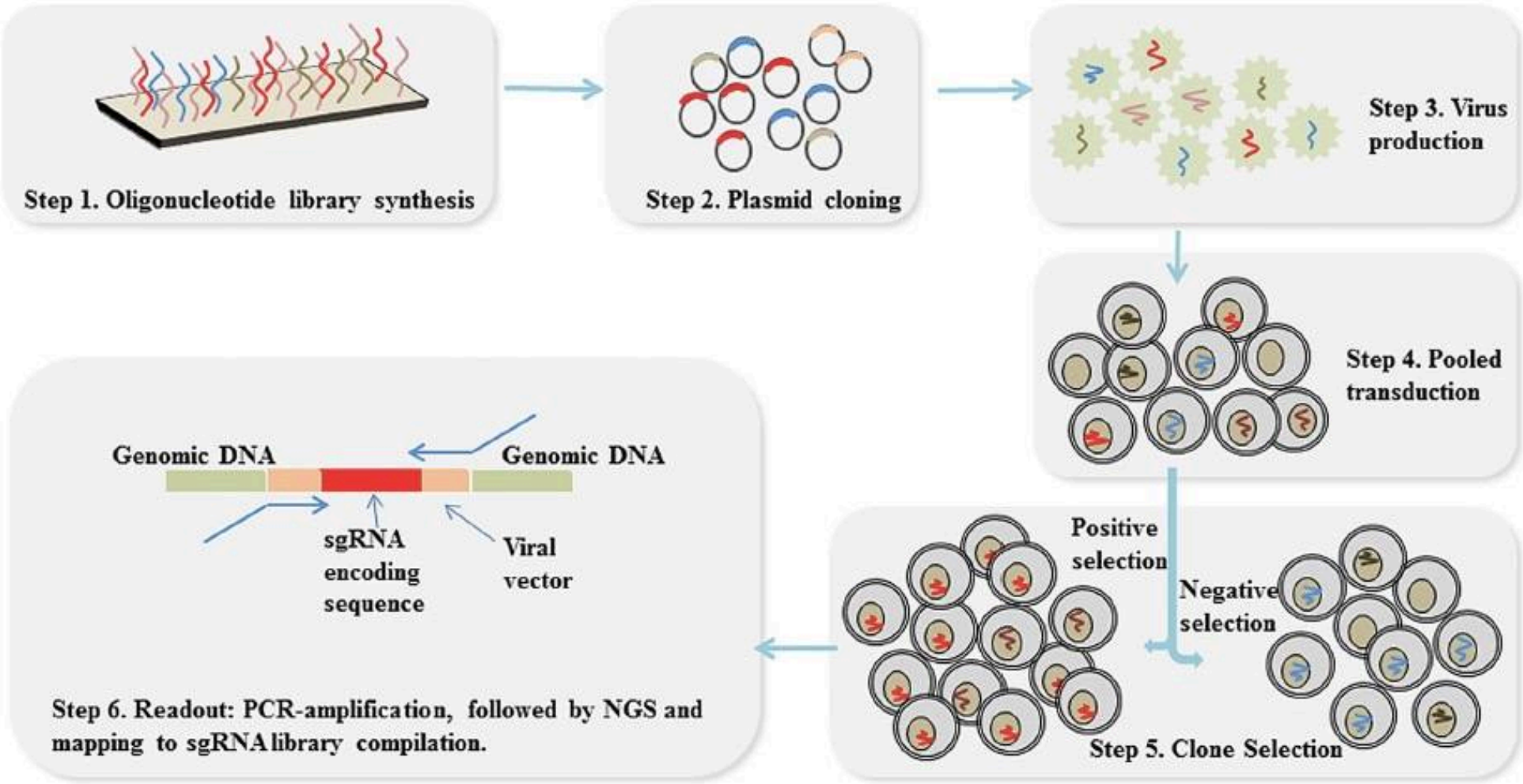
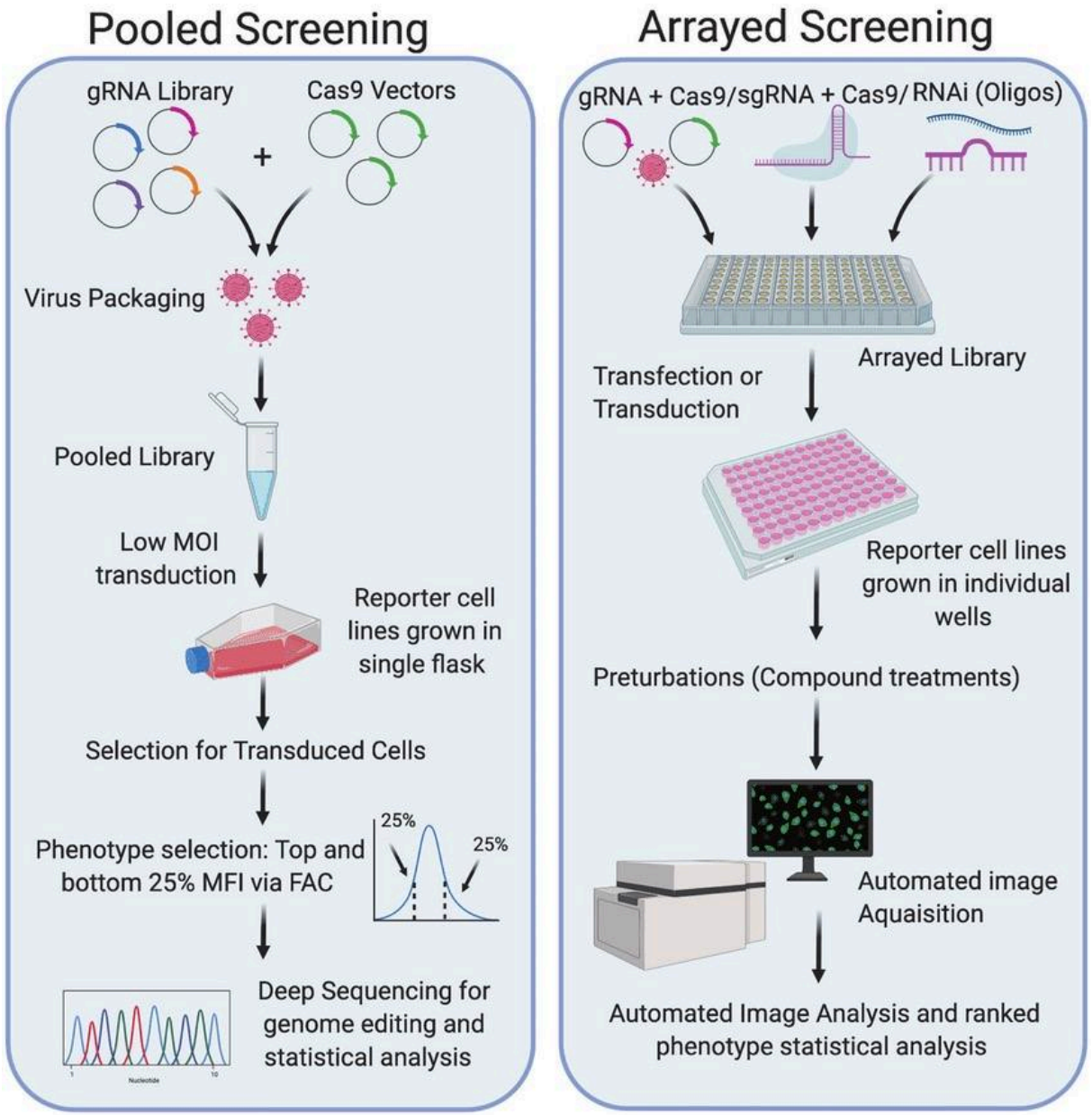
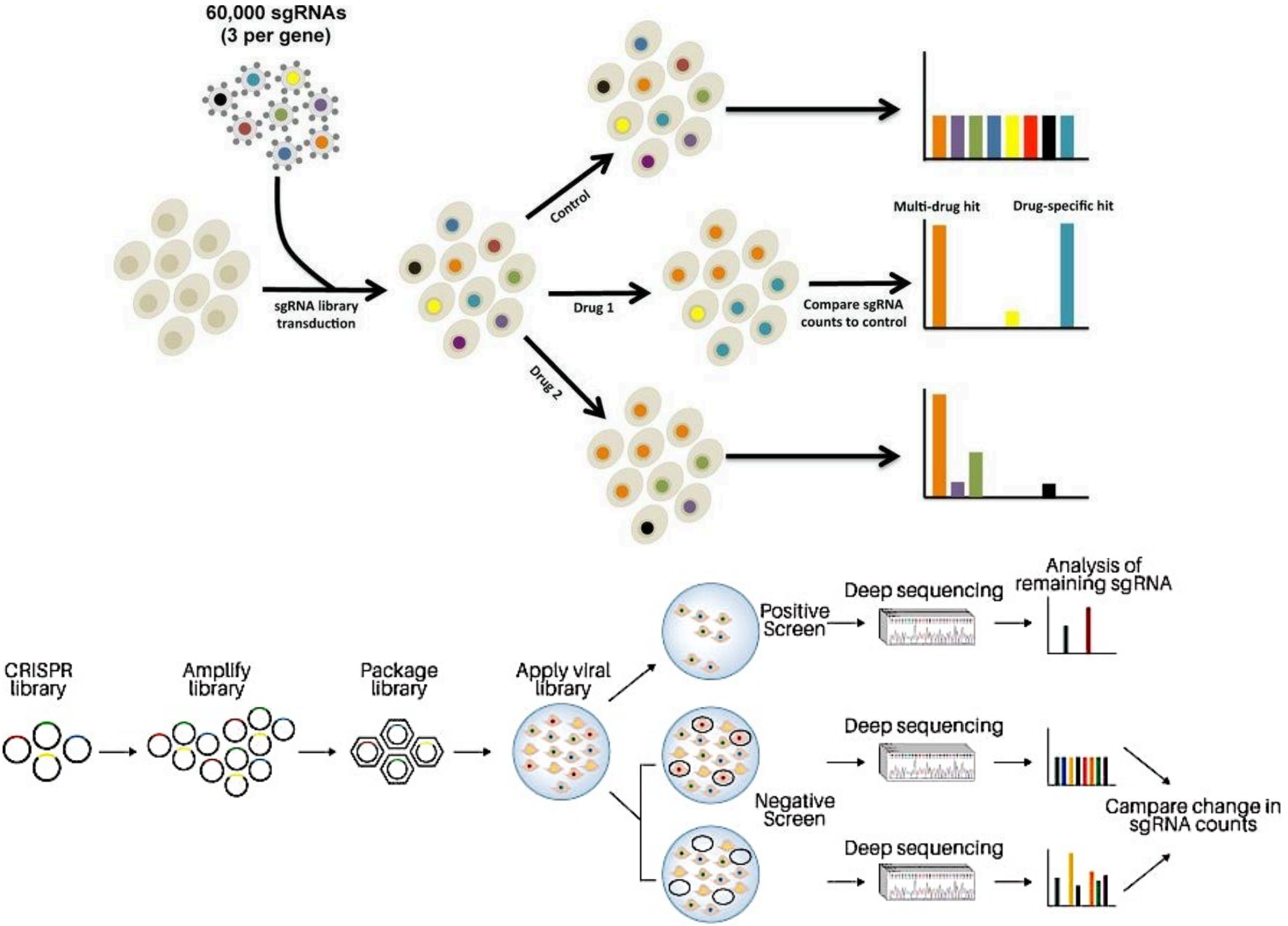
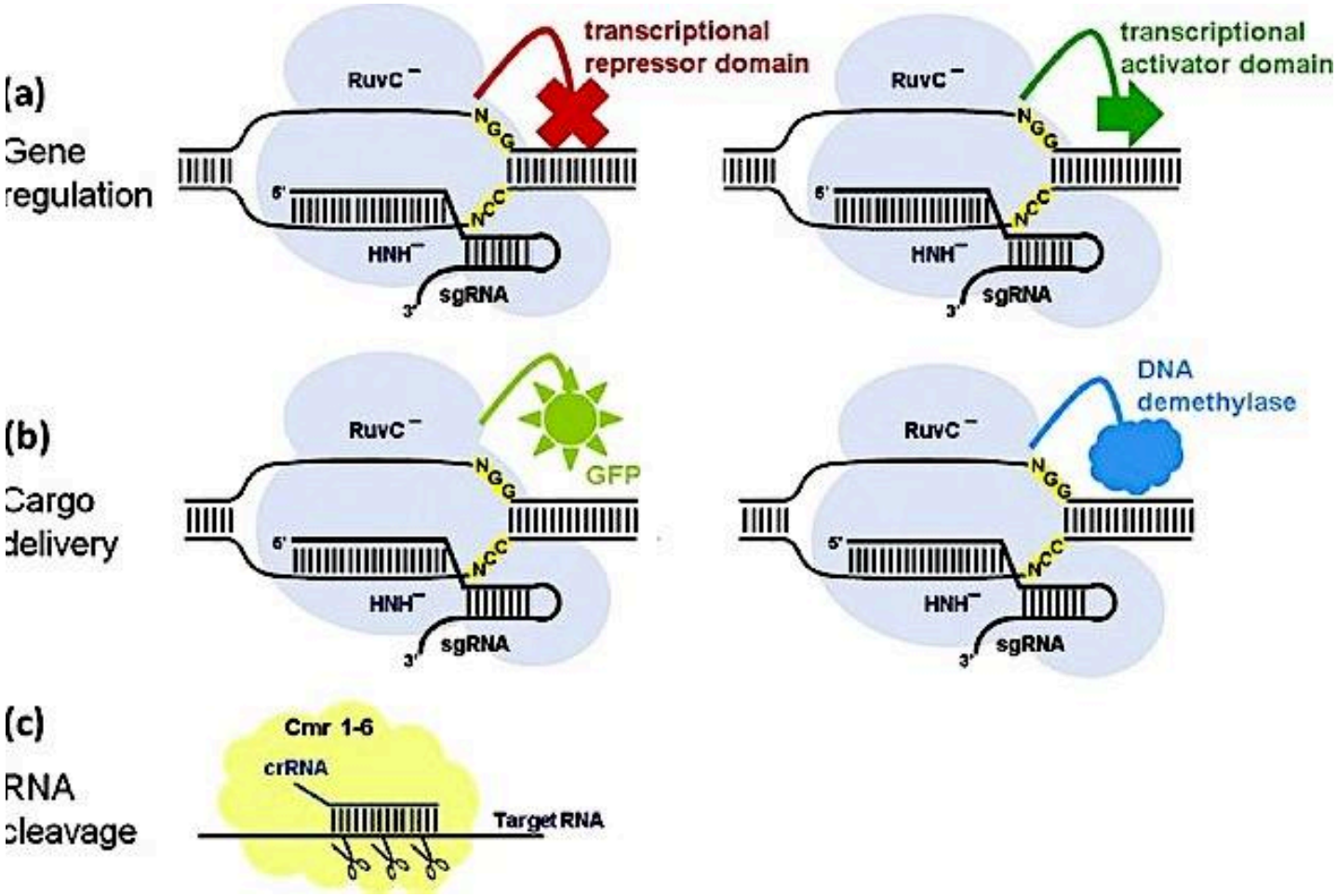


Figure 2 Flow chart of CRISPR (clustered *r*egularly *i*nterspaced *s*hort *p*alindromic *r*epeats) screen in pooled format. sgRNA library is massively synthesised (step 1), cloned into plasmids (step 2) and packaged into viral vectors (step 3). Cas9-expressing cells are infected with these vectors at 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.


Pooled Screening
gRNA Library
Cas9 Vectors
+
Virus Packaging
Pooled Library
Low MOI transduction
Reporter cell lines grown in single flask
Selection for Transduced Cells
Phenotype selection: Top and bottom 25% MFI via FAC
25%
25%
Deep Sequencing for genome editing and statistical analysis
Arrayed Screening
gRNA + Cas9/sgRNA + Cas9/ RNAi (Oligos)
Arrayed Library
Transfection or Transduction
Reporter cell lines grown in individual wells
Preturbations (Compound treatments)
Automated image Aquaisition
Automated Image Analysis and ranked phenotype statistical analysis


60,000 sgRNAs
(3 per gene)
sgRNA library
transduction
Control
Drug 1
Drug 2
Compare sgRNA
counts to control
Multi-drug hit
Drug-specific hit
CRISPR
library
Amplify
library
Package
library
Apply viral
library
Positive
Screen
Deep sequencing
Analysis of
remaining sgRNA
Negative
Screen
Deep sequencing
Deep sequencing
Campare change in
sgRNA counts



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.