

Genome sequence 021822

All cellular life on earth evolved from a single common ancestor - LUCA (the Last Universal Common Ancestor).

Yet, our planet is now occupied by untold diversity of living forms.

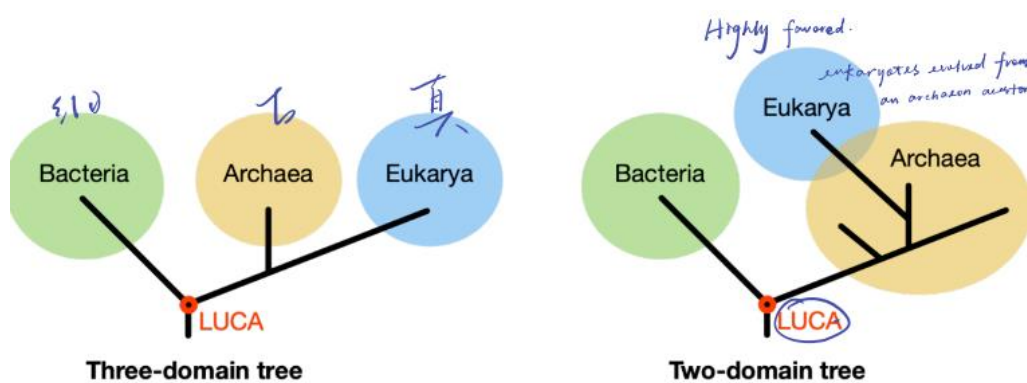
The feature of all cellular life

"Life is a self-sustaining chemical system capable of Darwinian evolution"

life-sustaining: replicable, making copies of themselves

capable of Darwinian: When making nearly exact copies of itself (with some mistakes), it allows variance of the species to potentially be more fit in the environment.

Modern views on the tree of life



Features of LUCA inferred from universal homologies

Feature	Trait
Core cellular features	
Genetic material	DNA
Bases used in DNA	A, C, T, G
Bases used in RNA	A, C, U, G
Genetic code	Three letter
Cell envelope	Lipoprotein membrane
Protein composition	20 core amino acids
Cellular complexes found in all organisms (examples)	
Translation	Small-subunit rRNA Large-subunit rRNA Multiple ribosomal proteins Aminoacyl-tRNA synthetases tRNA
Transcription	RNA polymerase
Membrane transport systems	ABC transporters

Inheritance (genetics), evolution, and a double-stranded DNA genome are all key features of cellular life on Earth.

How did we get from LUCA to this?

genetic/genomic variation time population dynamics selection:

- Variation in gene sequences → variation in gene function → phenotypic variation
- Variation in genome sequences is fodder[饲料] for evolution
- Linking genetic variants to phenotypic variants can reveal how phenotypes are regulated

What do we want to learn from the genome sequence?

- How does the genome sequence explain/ predict the biology of the organism?
- How does genetic variation (polymorphisms) between individuals in a population explain phenotypic differences (e.g. disease)?
- How do differences and similarities between genomes of different species contribute to our understanding of evolution?

What is genomics?

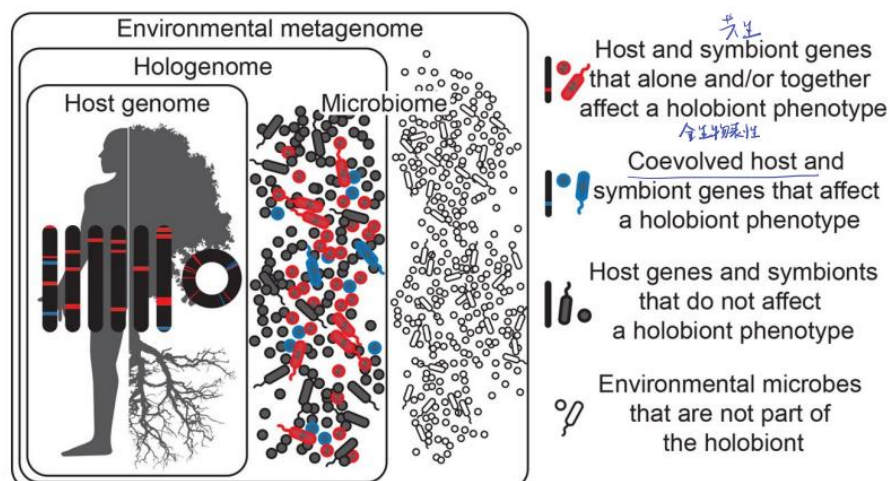
the branch of molecular biology concerned with the **structure, function, evolution, and mapping** of genomes

Nuts and bolts of genome sequencing

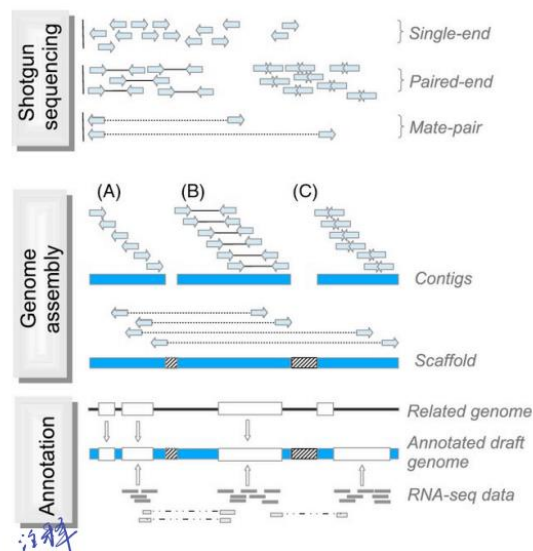
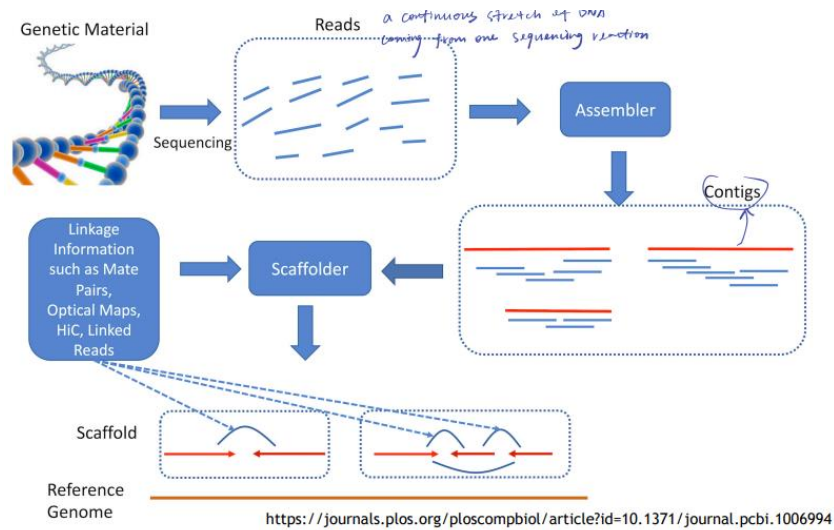
humans are **diploid**

23 pairs (so 46 total) of chromosomes, which are long, linear chains of DNA Range in size from 50,000,000 to 250,000,000basepairs

The hologenome[全基因组] concept



Major steps in a sequencing project



The idealized genome assembly

- Number of contigs should equal number of chromosomes (They should essentially represent the full chromosome sequence)
- Every nucleotide should be accounted for -- no gaps
- No errors in sequence

Break it down

Using DNA polymerase

Challenge - DNA polymerase is not processive enough to copy an entire genome

Solution - Sequence **small** stretches of DNA and then stitch the genome back together later.

Shotgun sequencing

Breaking large DNA molecule to small fragmentation, then sequence the fragmentation and assemble the fragmentations based on the overlapping part of DNA on the fragmentations

Restriction enzymes cut DNA

Longer restriction sites are rarer: Probability for the recognition site of n bp length to be found at a given position in a genome is $(1/4)^n$, so the longer restriction sites are rarer.

A modern approach: Shearing DNA by sonication

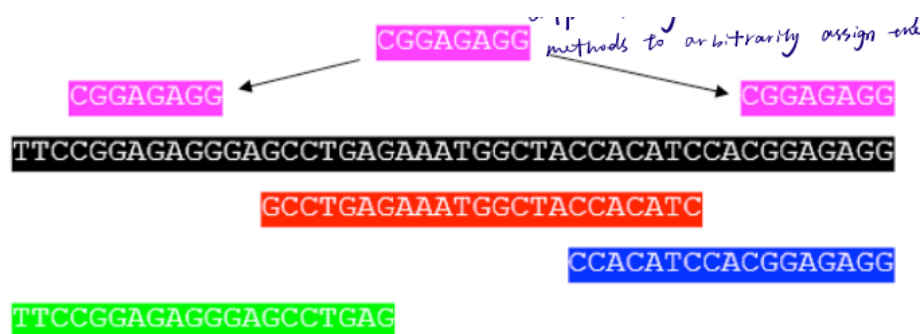
Shearing by sonication is “tunable”[可调节的], even and random

Sequences that overlap can be collapsed into contigs

- **end read (n)**: short stretch of sequence read from one end of a clone; length of read is constrained by sequencing technology, not by the length of the clone
- **contig (n)**: a stretch of contiguous sequences assembled from the sequences of multiple overlapping reads without gaps and hopefully no errors
- **scaffold (n)**: represents a large stretch of sequence built up from multiple contigs; may contain gaps

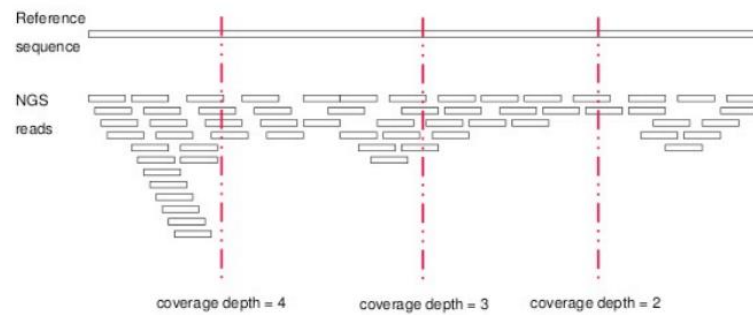
Genome assembly 022122

Read assembly



Different algorithms[算法] will use different methods to arbitrarily assign[任意匹配] the sequence. Shorter the sequence, more likely it is going to be found in multiple places in the genome. (Shown in the restriction enzyme frequency)

One partial solution - high “sequence coverage”

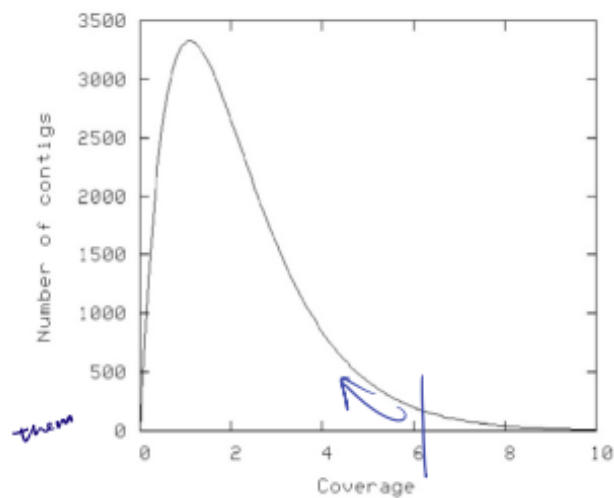


How to calculate average coverage

$$C = LN / G$$

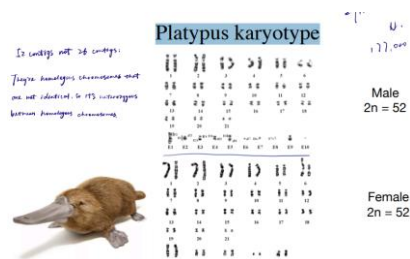
- C stands for coverage
- G is the haploid genome length
- L is the read length
- N is the number of reads

Number of contigs scales with genome coverage



- A higher number of contigs means, there more gaps between the contigs
- Low coverage still ends up with low contig number, but impossible to sequence one fragment covering each part of the chromosome in order.

Platypus karyotype



How many contigs needed for genome sequencing

52 contigs not 26 contigs, their homologous chromosomes are not identical, so it's heterozygous between homologous chromosomes.

Three major challenges to overcome

1) repetitive sequences 2) polymorphisms 3) sequencing errors

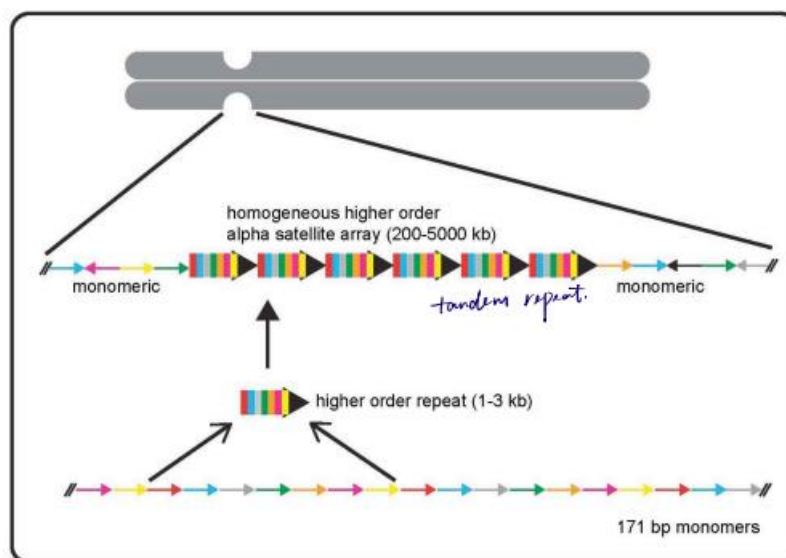
Repetitive sequences in platypus genome (Diverse type, 2 listed below)

Short interspersed[穿插] nuclear elements (SINEs): 100 to 700 base pairs in length

Long interspersed nuclear elements (LINEs): ~7,000 base pairs long

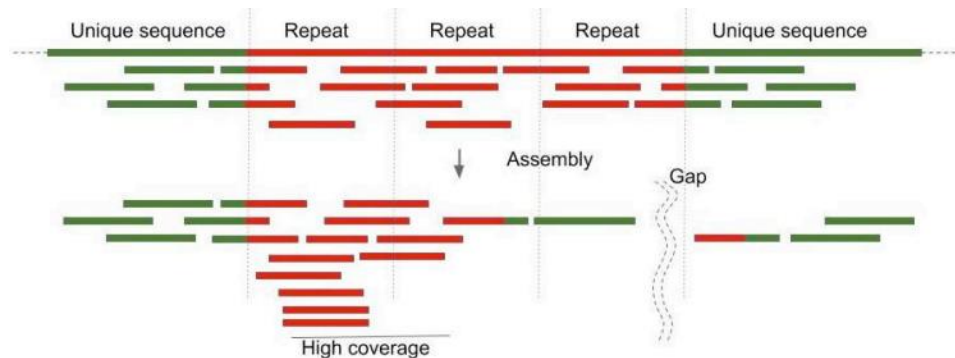
Repetitive sequences in human centromeres

Clustered around the human centromeres



Tandem [串联] arrays of rDNA repeats

Sequence repeats disrupt genome assembly



Sequence repeats disrupt genome assembly, because the repeat sequence can collapse on each other.

Additional common challenges in genome assembly

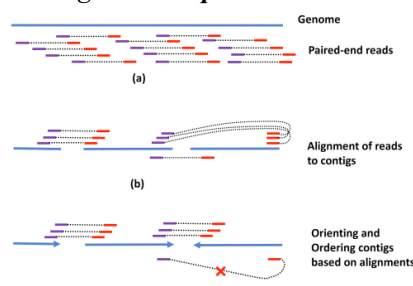
- genetic polymorphism (e.g. need to sequence from paired chromosomes of diploids or from preparing DNA from multiple individuals) → There's some difference between individuals, the most common one chosen as the reference.
 - sequencing errors
 - contamination
 - misassembled contigs
-
- Sequence coverage
 - Real-life example of a genome project
 - Repetitive elements and how to deal with them
 - Modern strategies for building a genome assembly
 - All genome sequences are incomplete

assemble contigs into larger stretches of sequence

Solution #1: Paired-end reads

*Sequences from both ends of a piece of DNA fragment can **connect two contigs into a scaffold***

*Paired end reads help with assembling **across repeats***




After shearing the Genome to the fragmentations and assembling them to contigs according to the shared sequence in different reads, we collect some paired-end reads, which is continuous in the sequence where the two contigs are not, so by alignment of reads to contigs, orienting and ordering contigs based on alignment, we can connect contigs into scaffold.

Solution #2: Use a mixture of sequencing approaches

Balance read length against error rate, because two things can't be combined.

Relative strengths of different sequencing technologies



	Illumina (HiSeq 4000)	PacBio (Sequel)	Oxford Nanopore (MinION)
Read length	Up to 150 bp	10-15kb	Up to 900kb
Number of reads	2.5-5 Million	500 K	Up to 1 M
Processing time	<1-3.5 days	Up to 10 hours	~ 6 hours
Error rate	<1%	10-15%	5-15%
Cost per run	~\$3000	~\$850	\$500-\$900
Instrument price	\$900 K	\$350K	\$1K
Advantages	Highly accurate	Sequence long reads	Sequence long reads Portable device

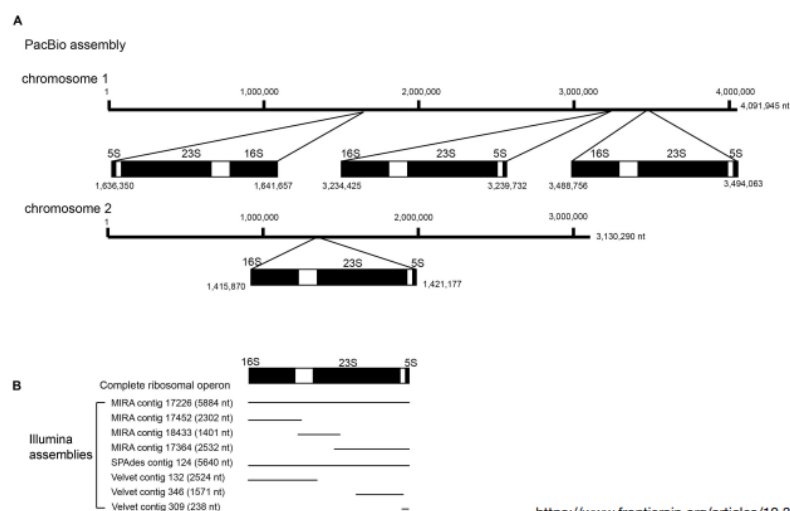
Errors are random in sequencing.

Get accuracy from Illumina, and couple it to the length that we get from PacBio.

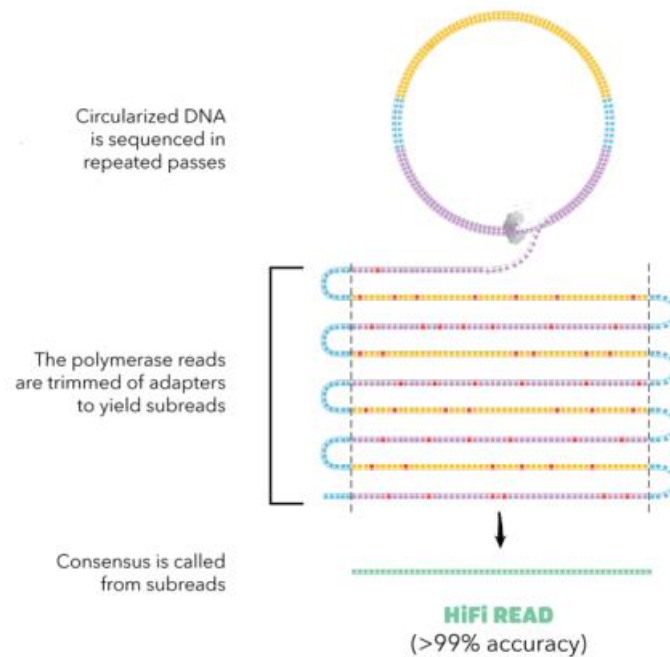
More deeply sequenced, more accurate.

The throughput of PacBio is not that high as Illumina

PacBio is superior for assembling contigs.



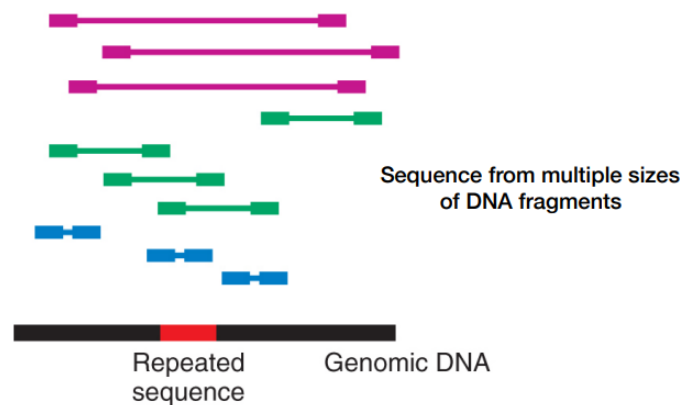
PacBio: solution to long reads with high error rates



Hairpin reads: they can be aligned to each other to give a high fidelity read with more than 99% accuracy, with giving up some of the link benefits of PacBio, but get back the high fidelity and accuracy.

Red dot: kinds of error.

Whole-genome shotgun sequencing strategy



Sequence from multiple sizes of DNA fragments

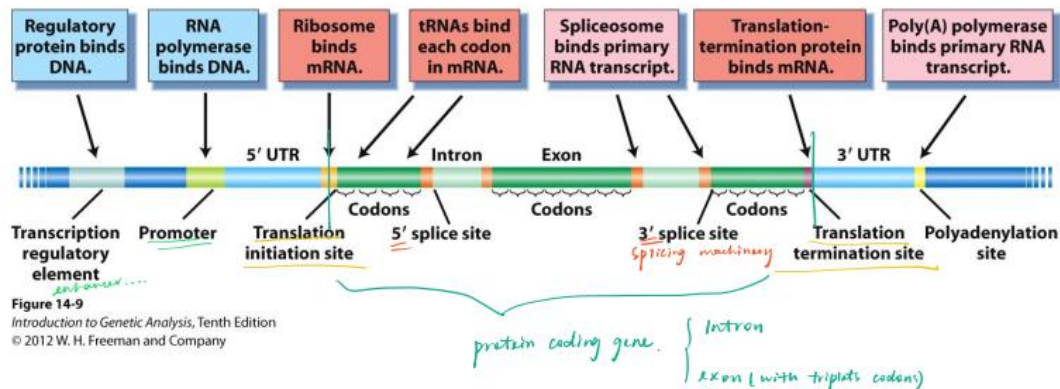
All genome sequences are incomplete and reflect a more-or-less accurate version of a species' genome sequence

Genome annotation 022322

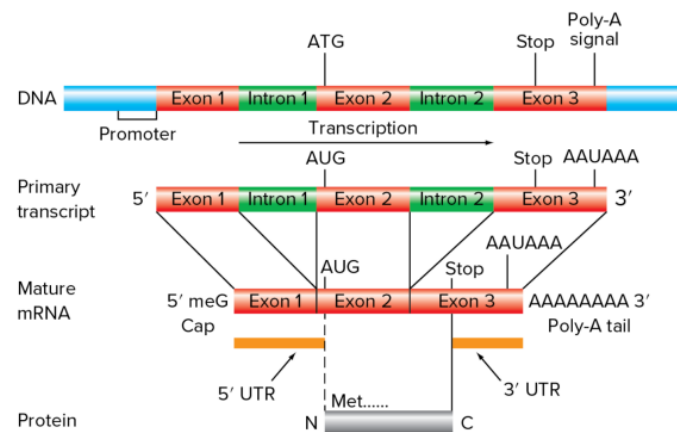
Transcriptional regulation leads to differentiation

Insight: to find binding sites for RNA polymerase/ transcription factor

The information content of the genome includes binding sites



Example: collagen gene structure



Genome annotation: Finding genes and other genetic elements in genomes

Strategy #1: Find long open reading frames

if there's a protein encoded by a stretch of DNA, there's going to be selection against stop codons, so we're looking for long stretches of DNA in a particular reading frame that do not have stop codons.

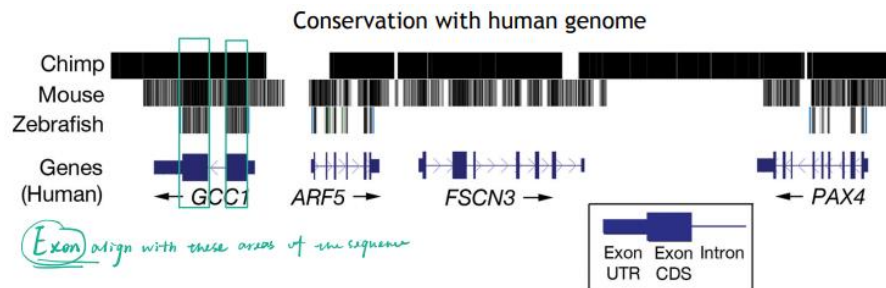
Frame 1 → 5' ...CCG ATG CTG AAT AGC GTA GAG GTT AGG TAA TCA TCA... 3'
 Frame 2 → 5' ... CGA TGC TGA ATA GCG TAG AGG TTA GGT AAT CAT CA... 3'
 Frame 3 → 5' ... GAT GCT GAA TAG CGT AGA GGT TAG GTA ATC ATC A... 3'

3' ...GGC TAC GAC TTA TCG CAT CTC CAA TCC ATT AGT AGT... 5' ← Frame 4
 3' ...GG CTA CGA CTT ATC GCA TCT CCA ATC CAT TAG TAG ... 5' ← Frame 5
 3' ...G GCT ACG ACT TAT CGC ATC TCC AAT CCA TTA GTA ... 5' ← Frame 6

no stop codon

Strategy #2: Look for conserved sequences

Do comparisons and see if you can find places and notice genomes that align well, reflecting protein-coding genes conserved across evolutionary time.



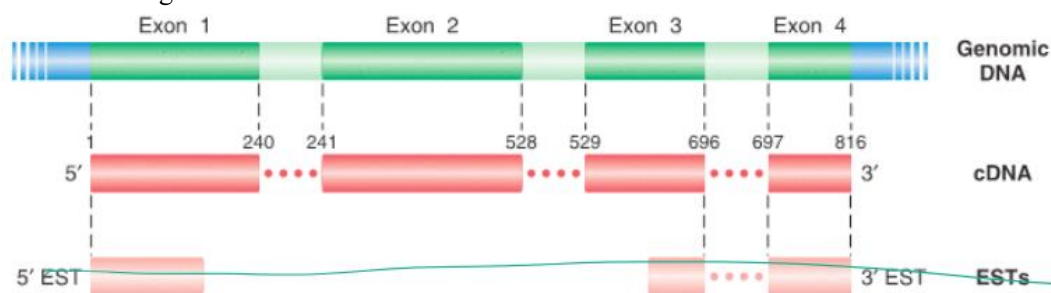
Exons align with these areas of the sequence, an area on open reading frame and high conservation suggesting that there might be a gene.

Strategy #3: Use functional information (RNA-Seq along with conservation of protein-coding sequences)

RNA-Seq: converting RNA into cDNA → Sequence (dsDNA) → align to DNA

Telling us:

- Location of genes and exons



- Validation of genome sequence
- Expression dynamics (different tissues or developmental stages)
- Length and location of introns
- Evidence for alternative splicing
- Can reveal coding sequence polymorphism

*Putting all this information together gives a draft roadmap for making predictions about the biology of an organism. Much is missed in the initial annotation of a genome and has to be filled in by **future analyses and experiments***

Major aspects of genome analysis

- Bioinformatics: analysis of the information content of entire genomes
- Comparative genomics: considers the genomes of closely and distantly related species for evolutionary insight, enables conserved sequences to be used as a guide to analyzing gene function

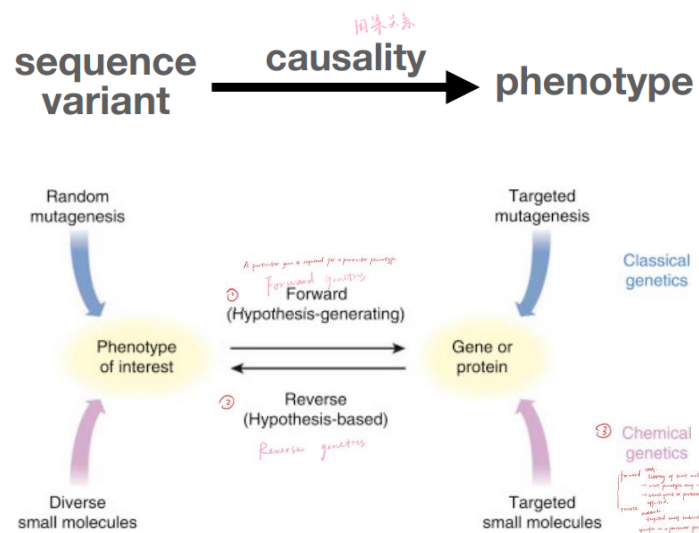
- Functional genomics: uses an expanding variety of methods, including reverse genetics, to understand gene function and to delineate[绘制] networks of interacting genes and proteins in biological processes.

Goal

- Use genome sequences to **build predictions** about basic biology and molecular mechanisms
- Leverage[杠杆; 利用] genomics where classical approaches (mutagenesis and screens, targeted disruptions) are not possible or informative

Genome Screens 022522

How do we identify the genes responsible for phenotypes of interest?



Classical genetics	Forward genetics (Hypothesis-generating)	Random mutagenesis → phenotype of interest → Gene or protein
	Reverse genetics (Hypothesis-based)	Target mutagenesis → Gene or protein → phenotype of interest
Chemical genetics	Forward genetics (Hypothesis-generating)	Diverse small molecules → phenotype of interest → Gene or protein
	Reverse genetics (Hypothesis-based)	Target small molecule → Gene or protein → phenotype of interest

How could you start to figure out how a genome controls a phenotype?

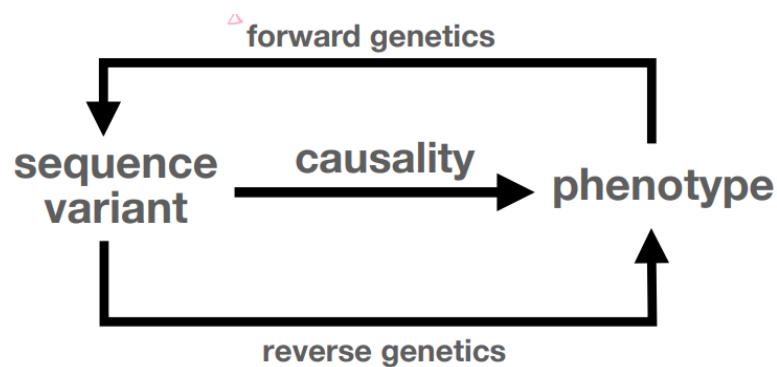
Systematically induce mutations and find alleles of genes that when mutated affect that phenotype!

What genes regulate proper development in vertebrates?

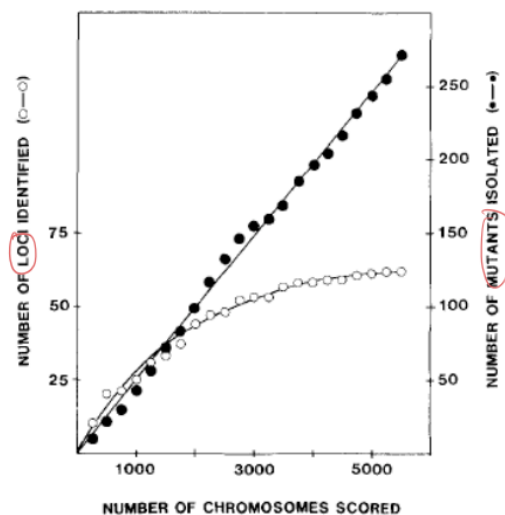
Prohibitively long and expensive scanning the genome by mutating to know, required gene for proper development

penetrance: the same mutation can have subtly different phenotypes in different individuals based on the genetic background.

How do we find all the genes that regulate our trait of interest?



Saturation



Because the genome is finite, a genetic screen is saturated when you stop finding new loci (genes), but rather just more alleles (mutants) of the same loci.

How do we know if a screen has saturated the genome?

Complementation test: a test for determining whether two mutations are in different genes (they complement) or the same gene (they do not complement)

- Reveals whether two mutations are in a single gene or in different genes
- "Complementation group" is synonymous[相似的] with a gene

Complementation: the production of a wild-type phenotype when two different recessive mutations are combined in a diploid

Caveats for complementation tests

- Complementation test should be done only when both loss-of-function mutations are fully **recessive**
- Use caution when doing complementation tests with mutants that have different phenotypes. Different mutations in the same gene can produce different phenotypes.
- The complementation test is only a test of gene function and provides no info regarding the position of mutations.

Genome Screens II 022822

Is a scanning saturated?

a fully saturated screen must hit all the genes multiple times. it is not saturated if some locus are only hit one time.

Challenges and goals

- The genome should be saturated with mutations. Ideally, this means **every position will be mutated** at some point.
- This works best if your mutagenesis strategy results in an **even distribution**[均匀分布] of **mutations** across the genome.
- If your mutagenesis strategy is **non-random**, **many more individuals** will need to be screened and some genes or genetic elements might not be mutated.

Mutagenesis is usually non-random: there are hotspots for mutation, heterogeneity of transposon insertion[转座子插入的异质性] across yeast chromosomes.

- Finally, you need a way to identify (1) where the mutations are, and (2) which mutations cause — the phenotype

Why can't we just sequence to find the mutations?

- Chemical mutagens can induce multiple mutations in a single organism/genome, and not all of the mutations are matters.
- The mutations can be hard to find, particularly if they induce SNVs (single nucleotide variants) or small indels[插入缺失].

To connect genes to phenotypes

We will take advantage of: phenotypic variants; recombination; phenotypic and genetic markers

We can use linkage and recombination to find genes on chromosomes

Construct genetic maps: the order of different genes on chromosomes

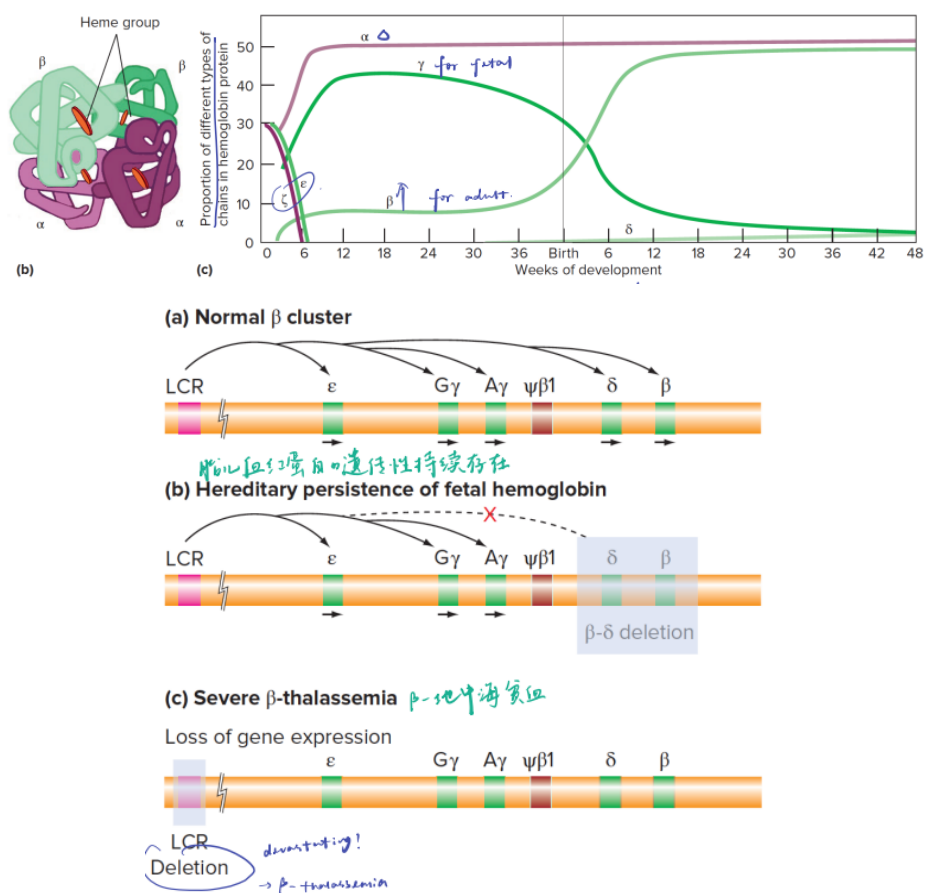
Approaches for mapping genes

- Two-point cross (pros/cons)
- Three-point cross (pros/cons)
- Modern “next-gen” approaches

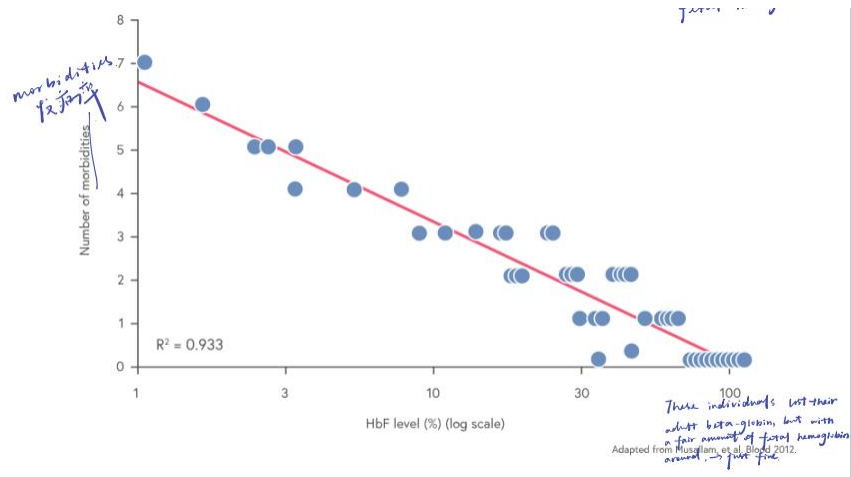
Reverse Genetics 020322

Genome editing through programmable nucleases

Hemoglobin switching during development

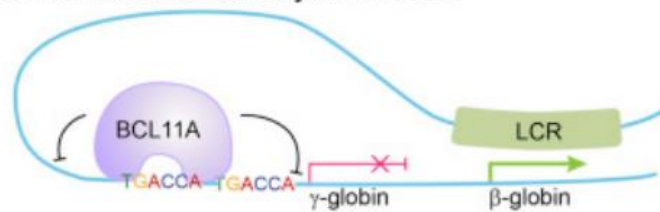


Reduced Symptoms in β -Thalassemia with Higher levels of HbF[fetal hemoglobin]

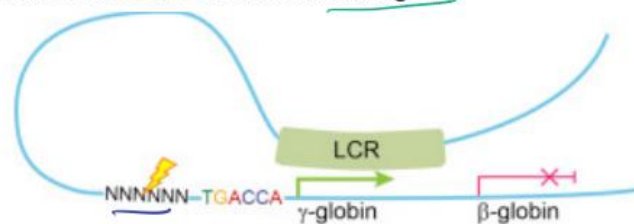


To rescue Sick Cell cells, turn fetal hemoglobin(γ Hb) back on after birth.

Normal adult human erythroid cells

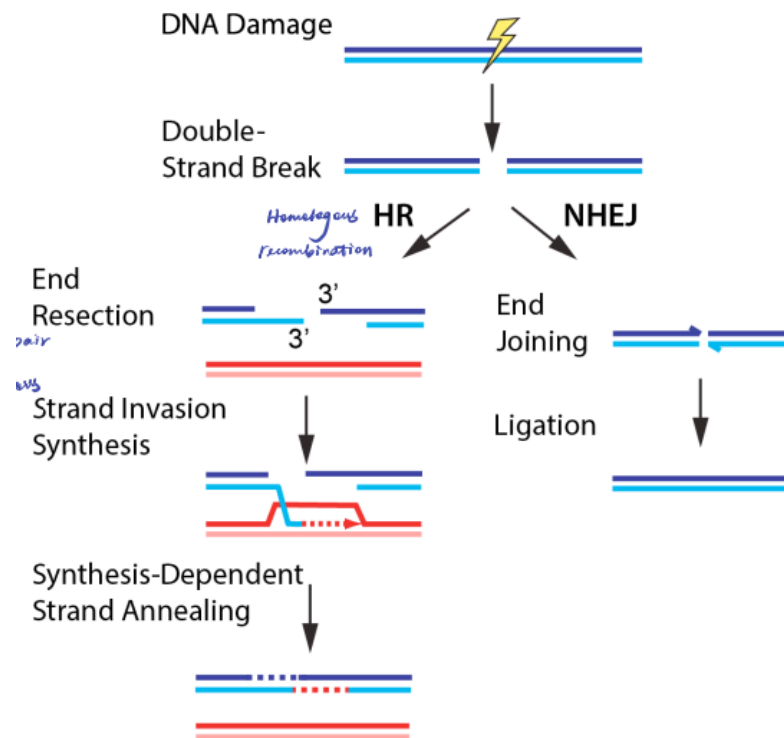


Mutated or edited BCL11A binding site



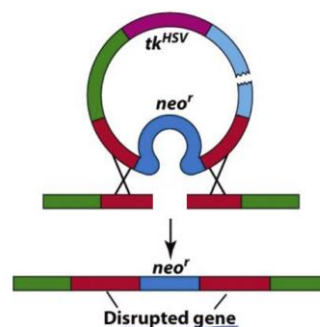
BCL11: a critical regulator (transcription factor) of whether a cell is producing fetal globin (γ) or adult globin (β), its binding inhibits expression of γ -globin, leading to LCR folding over and promoting the expression of β -globin.

Two paths to double-strand break repair



Double-strand breaks facilitate gene targeting by homologous recombination

Engineer a plasmid with a specific gene flanked by sequences that are homologous to the sequences to be disrupted.



Goal: target endonucleases to specific DNA sequences

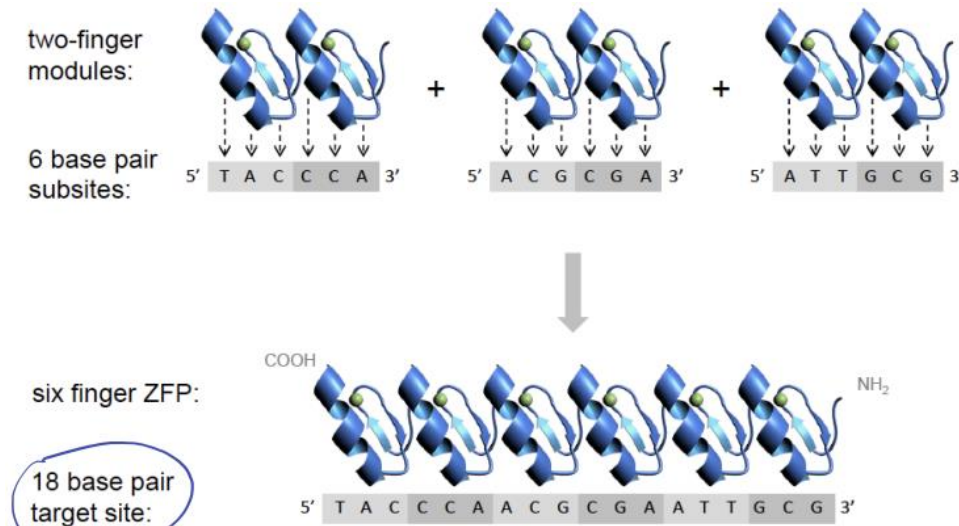
Strategy #1: Restriction Enzyme

- Not existing in many different places
- Most transcription factors have degenerate DNA sequence preferences.
- Most transcription factors also require contextual information, it needs complex DNA structure & modifications

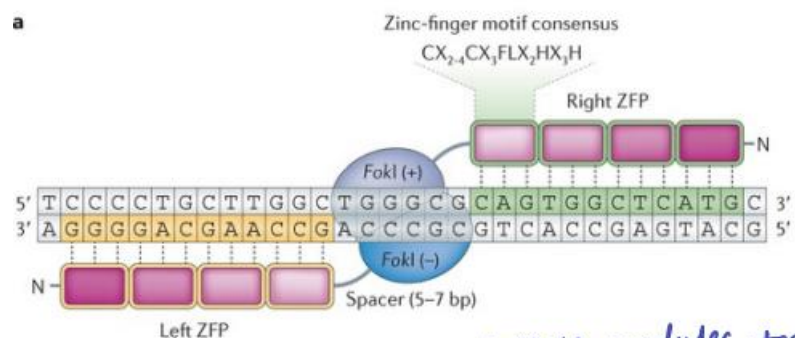
Programmable nucleases allow genome engineering: Zinc Finger Nucleases (ZFNs), TALENs, CRISPR/Cas9. All have in common that **a programmable, DNA sequence specific binding domain is coupled to an endonuclease activity**.

Strategy #2: Zinc Finger Nucleases (ZFNs)

Zinc finger motif is used in a lot of transcription factors, two-finger modules used for design.



Zinc fingers fused to endonuclease (ZFN) provide precise DNA cutting and DNA insertion:



not these modules together

Put continuous Zinc-finger modules together flanking a sequence to target and fuse it to the enzyme FockI,

Each motif has its preference, best there are more modules.

TALEN is like the same

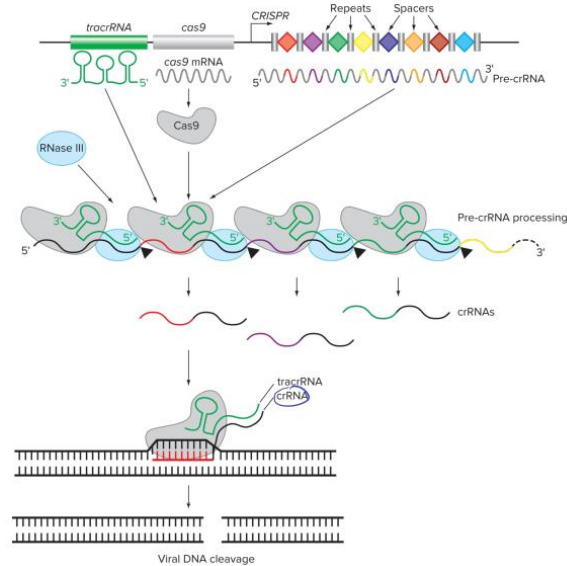
Different approaches offer different strengths

Precision: target any desired nucleotide in the genome	Specificity: edit the targeted nucleotide without editing elsewhere in genome	Efficiency: frequency of modification at the desired target nucleotide	Throughput: ability to rapidly design and implement gene editing across genome
---	--	---	---

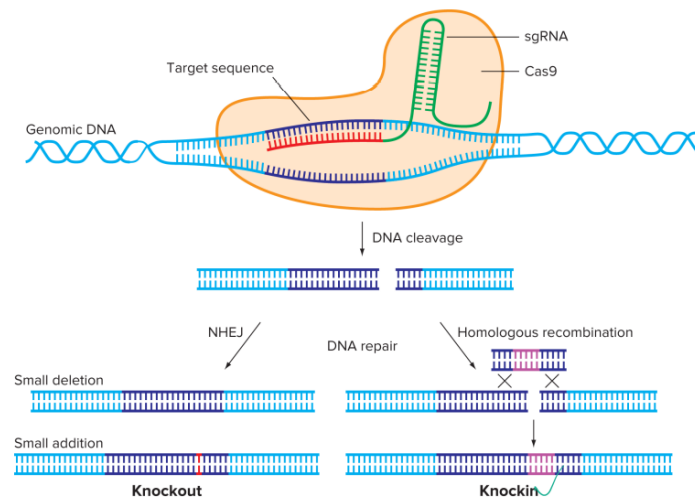
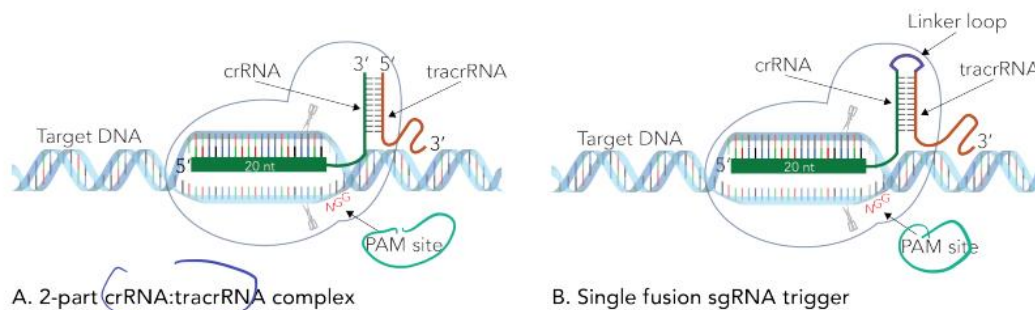
Clustered Regularly Interspaced Short Palindromic Repeats (CRISPR)

Spacers were found to be homologous to sequences from bacteriophage, the CRISPR elements part of a phage defense system.

→ A new method based on bacterial adaptive immunity



A key breakthrough: the single fusion guide RNA



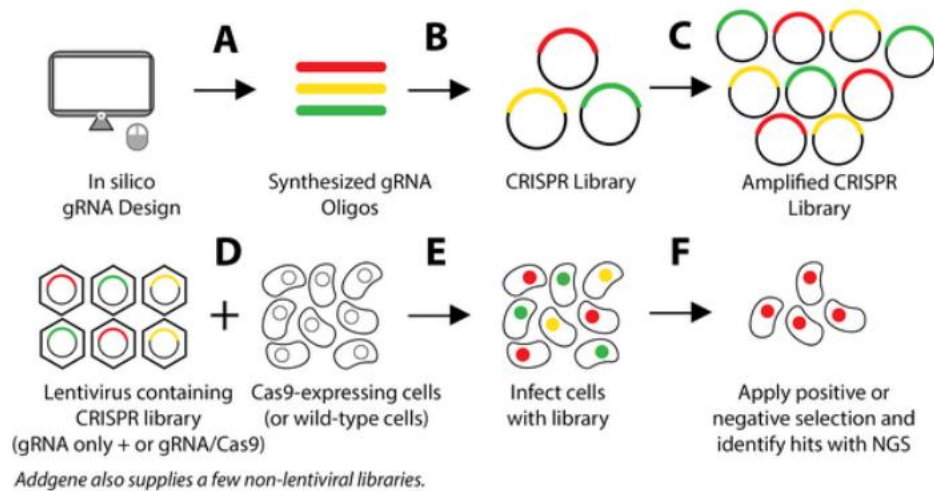
To rescue Sickle cell patient:

Option A: Fix, CRISPR and DNA template fix the mutation in the adult hemoglobin gene.

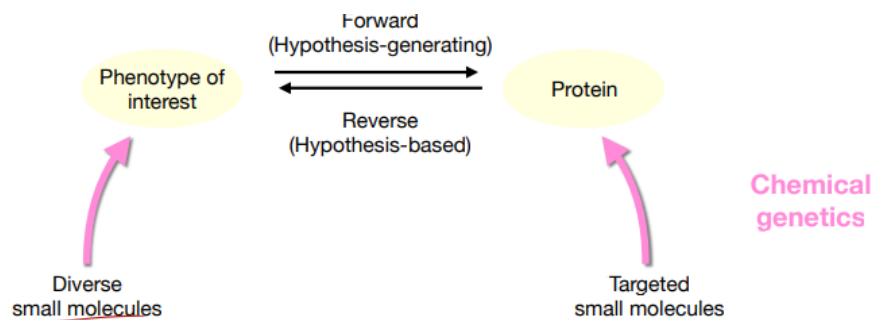
Option B: Swap, CRISPR reactivates the fetal hemoglobin gene by turning off the BCL71A gene.

Option C: Interrupt a sequence in the BCL11A

CRISPR-based genetic screen



Chemical Genetics 040322



*Chemical biology is using chemistry to understand biology at a **molecular** level (atoms & bonds)*

Test diverse small molecules in a system to see if we get a phenotype, use that phenotype to lead us towards a protein of interest.

small molecules

< 1000 Daltons, Organic (Sulfur, Phosphorus, Oxygen, Nitrogen, Carbon, Hydrogen)

natural or synthetic

Forward

Test diverse small molecules in a system to see if we get a phenotype, use that phenotype to lead us towards a protein of interest.

Artemisinin

Forward chemical genetics is a way to use small molecules to find genes or proteins that are important for *P. falciparum* growth

309474 diverse small molecules tested to get those inhibiting *P. falciparum* growth

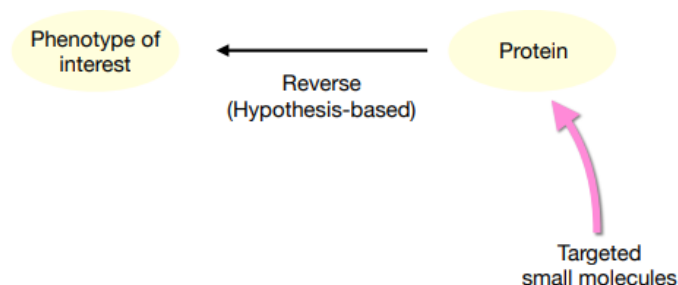
	<u>SJ000030570</u>	SJ000101247	SJ000025081
Known target?	Unknown	Unknown	Unknown
50% <i>P. falciparum</i> death dose	2 nM	40 nM	25 nM
Effective against chloroquine-resistant <i>P. falciparum</i> ?	Yes	Yes	Yes
Toxic to human cells?	No	No	No
Dissolves easily? Easy to make?	Yes / Yes	No / Yes	No / Yes

(+)-SJ733 was discovered through a forward chemical genetic study and is currently in a phase

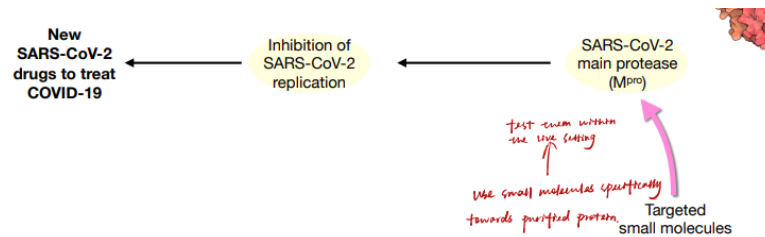
II clinical trial

- (+)-SJ733, an improved version of one of the molecules found, is more effective than artesunate
- To find the relevant protein, researchers grew *P. falciparum* in the presence of (+)-SJ733 to obtain resistant parasites
- *P. falciparum* with mutations in **pfatp4**, a predicted sodium efflux channel protein, were resistant to (+)-SJ733. Treatment of *P. falciparum* with (+)-SJ733 increases intracellular sodium in the parasite. Pfatp4 mutants are resistant to (+)-SJ733 and accumulate less sodium.

Reverse



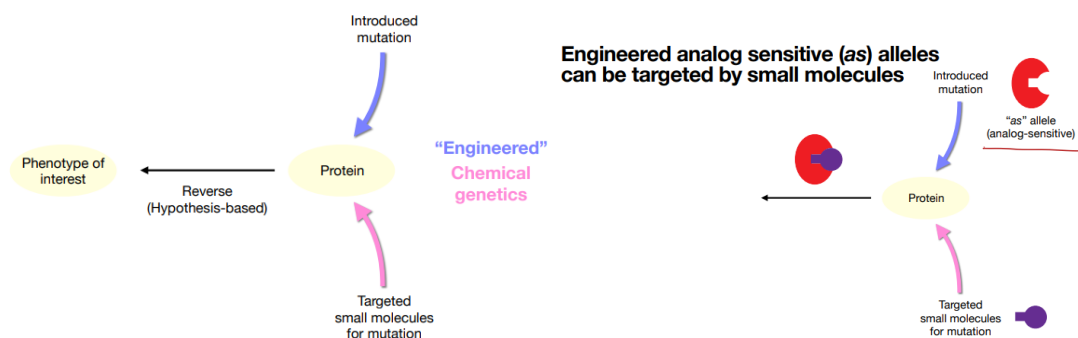
Coronaviruses like SARS-CoV-2 require proteases to cut the viral polyprotein into individual proteins necessary for viral replication



Use small molecules specifically towards purified protein, then test them within the live setting, see the function and phenotype.

- PF-00835231 was identified as an inhibitor of M^{pro} and optimized to make PF-07321332
- PF-07321332 rescues cells from virus-induced death
- PF-07321332 (Nirmatrelvir) combined with ritonavir (a molecule that prevents drug metabolism) effectively treats patients with early COVID-19
- By using reverse chemical genetics, Pfizer was able to develop PAXLOVIDTM to treat COVID-19

Chemical biologists can use “Orthogonality” [正交]



- *Arabidopsis thaliana* *cdka* mutants are hard to study due to a severe developmental phenotype
- The *as* *cdka* mutant rescues the developmental phenotype and allows for chemical inhibition with 1NM-PP1
- Treatment of the *as* *cdka* mutant with 1NM-PP1 results in the same growth defect but not with wild-type (less severe)
- Specific inhibition of CDKA;1 shows that the protein kinase activity is necessary for *A. thaliana* growth
- Chemical genetics is complementary to classical genetics
- Chemical genetics uses small molecules to assess the relevance of genes or proteins for a phenotype.
- Chemical genetics can be used to separate protein enzymatic activity from other protein functions.
- Chemical genetics is used in drug discovery and molecular and cell biology research

Canine Genetics 090322

Epistasis

Epistasis is a condition in which a mutation in a single gene phenotypically “wins” over a phenotype caused by a mutation in a different gene. [上位性: 单个基因中的突变在表型上“胜过”由不同基因中的突变引起的表型。]

A → B → C → 1

			<div> <div>A → B → C → 1</div> </div>		
Genotype			Phenotype		Inferred Pathway
			transformation	duplication/deletion	
i	$\frac{+}{+}$	$\frac{+}{+}$	none (both 1 and 2)	none	
ii	$\frac{a}{a}$	$\frac{+}{+}$	1 → 2 ✓	missing 1	A → 1 <i>2.2 x 10⁸ 1.6.</i>
iii	$\frac{+}{+}$	$\frac{b}{b}$	2 → 1 ✓	extra 1	B → 1 <i>160,000,000</i>
iv	$\frac{+}{+}$	$\frac{c}{c}$	1 → 2 ✓	missing 1	C → 1
v	$\frac{+}{+}$	$\frac{b}{b}$	1 → 2 ✓	missing 1	B → C → 1
vi	$\frac{a}{a}$	$\frac{b}{b}$	2 → 1 ✓	extra 1	A → B → 1

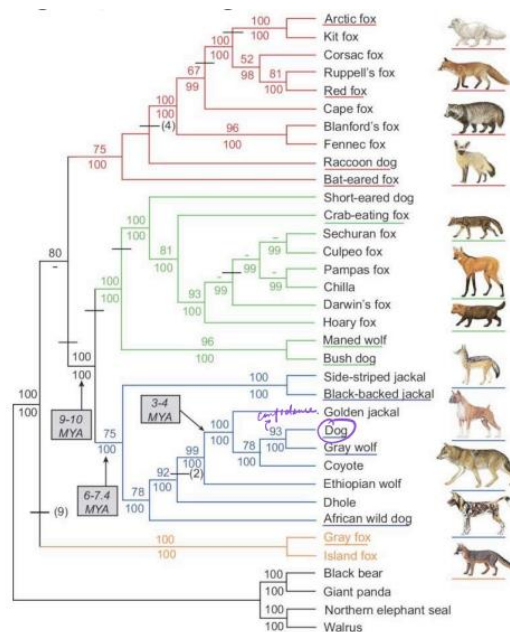
Intro to population genetics and quantitative genetics 110322

Canine genetics: dogs are incredibly diverse and that we can take advantage of dog genetics to try to understand how genomes encode morphology and different aspects of biology. So dogs turn out to be a really wonderful model for **linking variation in genomes to variation in phenotype**.

Domestication of dogs

- When were dogs first domesticated? How long have humans been co-evolving with dogs?
- From what (kind of wolf) were dogs domesticated?
- How many domestication events? One? More?

Phylogeny of dogs, wolves and other relatives



Phylogenies depict relationships between living organisms over evolutionary timescales.

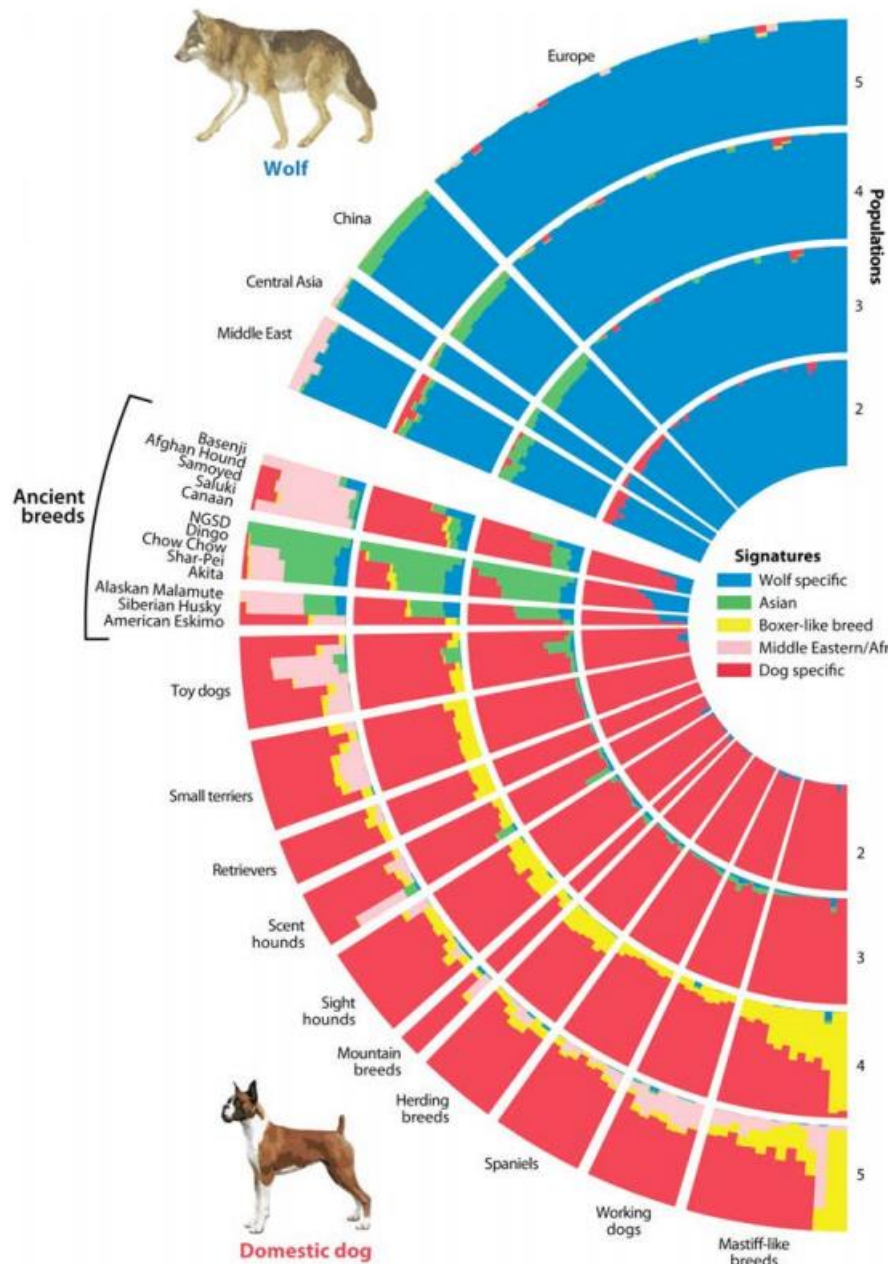
This phylogeny is actually based on conserved sequences that are shared and all these lineages.

Dog and the gray wolf: if we trace these lines back, their closest relatives are each other. You'll see here that the dog as close to the golden jackal along the axis. But if you trace the line back, you see that actually they are quite separated in terms of divergence is an evolutionary timescales

*These numbers at the branched actually give us a measure of our confidence in each of these branches points on the tree. So this relationship between dogs and gray wolves is actually highly supported. These values are often called **bootstrap values**. There are other kinds of statistical measures. And in this case on this tree, the, the highest support is a 100. So you can see that that this is a very highly supported branch.*

If we work our way back to the common ancestor of the Goldman tackle, the dog, the gray wolf, and the coyote. We think that these lineages[谱系] diverged around three to 4 million years ago.

How dogs evolved from wolves



In this study, many different living individuals from different geographic locations, Europe, China, Central Asia, and the Middle East were sampled and end there and SNPS in their genome (Single nucleotide polymorphisms) were sampled across their genomes. Determine how related they are to their own wolf populations versus populations of dogs.

Chinese and Middle Eastern wolves have have dog-specific signatures

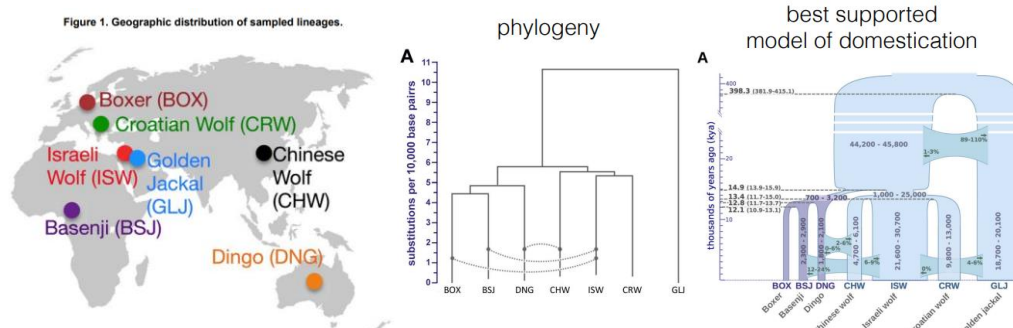
“Ancient Breeds” have wolf signatures

Implication:

Dogs might have evolved not from European groups, but instead from those that are found in China, Central Asia, and the Middle East. And these are the shared markers between these

breeds and wolves provide more support for the notion that **ancient breeds are the ones that are most closely related or that share more genetic similarity to their last common ancestor.**

When and how did dogs first originate?



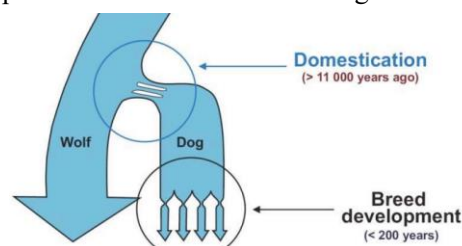
If we follow this line, then what we see is that we're starting at the top, we move left and then go down. And then we have what's called a bifurcation or split in this lineage. And on one side of the split, we have the wolves. And on the other side of the split, we have the dogs. But there are these little dotted lines. And I know you learned about admixture in the case of human evolution. And it looks like that also has happened with dogs and wolves. So it seems that after this diversification event, there have been some examples, dogs meeting with wolves.

On the right then is an attempt to build a scenario for how and when dogs evolved based on information that we have from the paleontological record and other ways of measuring the rate at which these nucleotide substitutions accumulation.

Bottlenecks

Bottlenecks: leading to the diversification between two lineages.

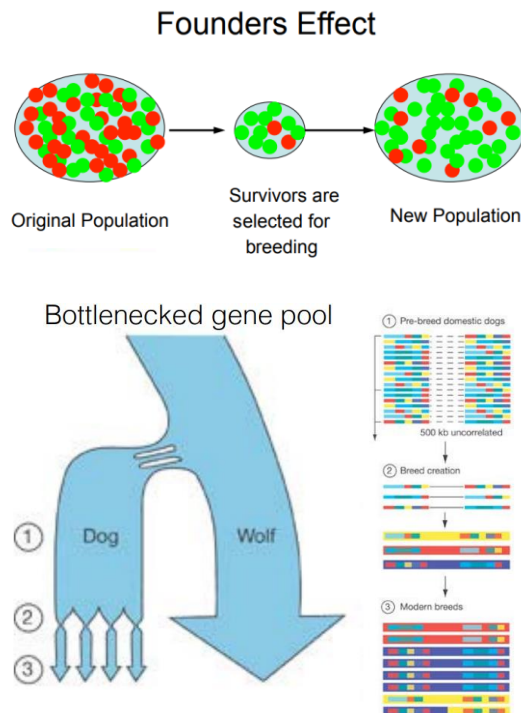
At least two bottlenecks shaped the evolution of modern dog breeds



founder effects and bottlenecks:

Initially: 50-50 ratio of A and a
 ↓ Bottleneck: large migration
 still 50-50

Initially: 50-50 ratio of A and a
 ↓ Bottleneck: strictive
 shifted



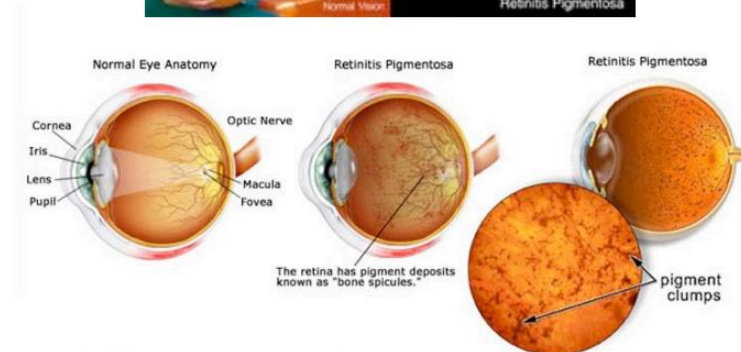
During this early phase of dog evolution, there might have been a lot of genetic diversity within the population. And, and now, you don't see clear haplotypes. But then at the onset of selecting for specific breeds, then this is breed creation. What you see is that in fact, the founding of these different breeds involves the establishment or essentially the transmission of very clear haplotype blocks. Okay, because the diversification of it is very new. And so as we look at modern breeds, because very little time has passed since the establishment of these breeds. And today, our modern breeds largely reflect the haplotypes of the founders from which they came. Okay, so there's the strong bottleneck up here.

Dogs as models for research into genetics of disease

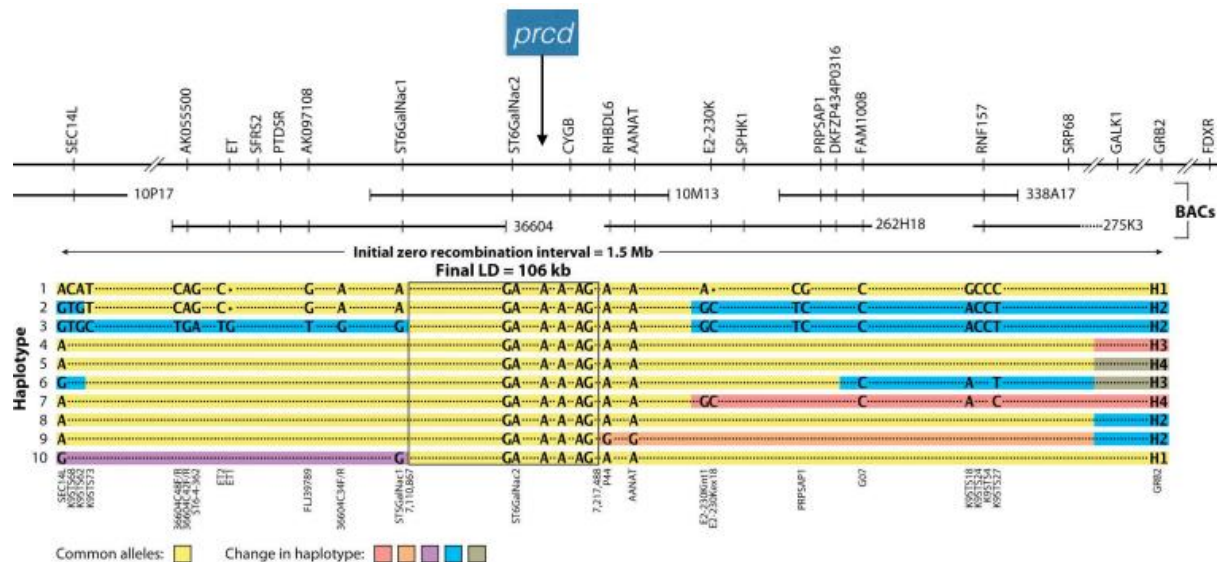
- Dog breeds arose through multiple genetic bottlenecks that have decreased recombinant variability
The incredible diversity of dogs that we see in different dog breeds today is the result of artificial selection over a very short period, 200 years. Okay, So there's still a strong connection between haplotype blocks, phenotypes. The dog breeds arose through multiple genetic bottlenecks and these decreased recombinant variability.
- Genetic diseases are problematic within inbred dogs. (At least 350 that are found in humans have been identified in different breeds.)
- Many breeds have many generations of breeding documentation

Progressive Retinal Atrophy [进行性视网膜萎缩]——Retinitis pigmentosa [色素性视网膜炎]

Over time and individuals who have what's called RP retinitis pigmentosa. You see that the field of view narrows greatly, leading to blindness.



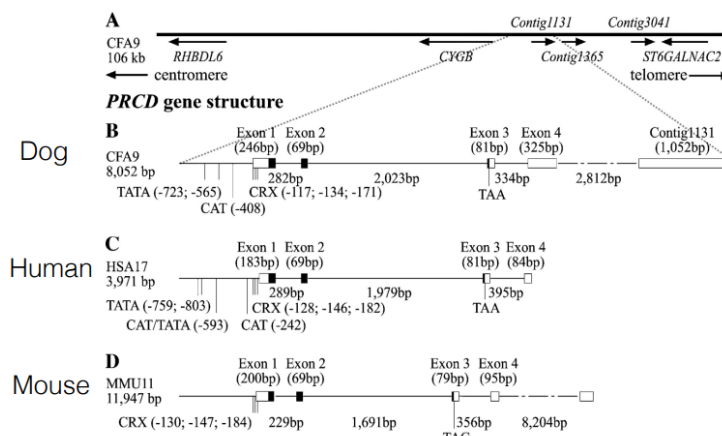
Identifying a small gene associated with progressive retinal atrophy



Yellow: common alleles → *prcd*

Not yellow: changes in the haplotype

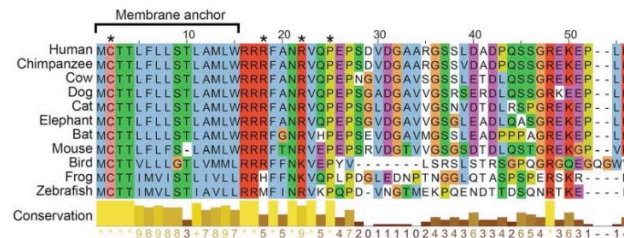
Pedigrees of six families showing segregation of the *prcd* mutant haplotype



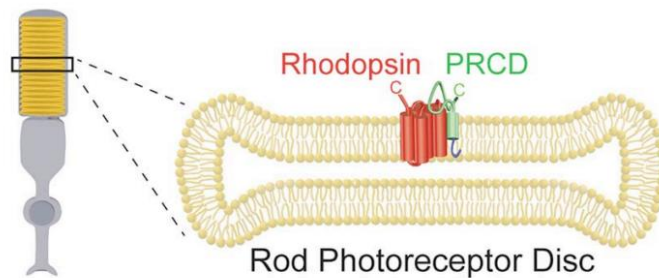
Gene sequence

Short exons & tiny protein, highly conserved in human and mouse.

Conservation of *prcd*

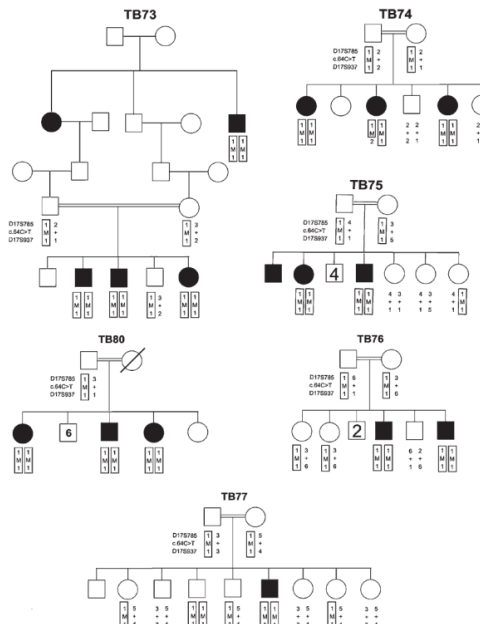


PRCD forms a complex with Rhodopsin in the rod photoreceptor disc



The PRCD binding to Rhodopsin stabilizes and maintains these healthy Rod Photoreceptor Disc.

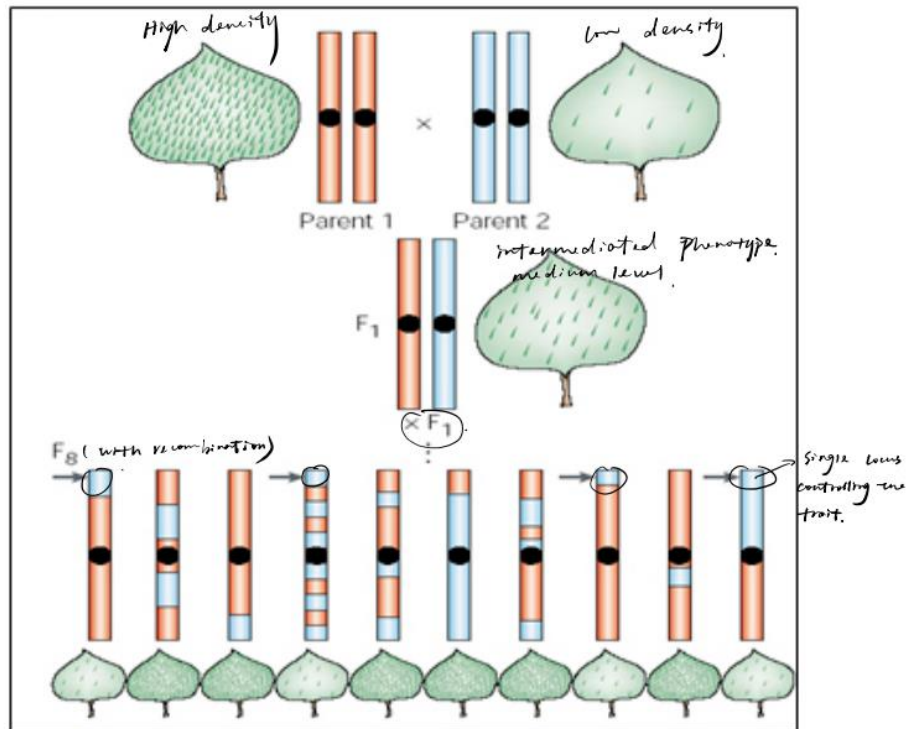
Pedigrees of six families showing segregation of the prcd mutant haplotype



By using, taking advantage of the really interesting and unusual population genetics of dogs. And there are two strong bottlenecks that they experienced in their domestication. We can actually

identify genes that are important for not only dog diseases, but we can take that information and apply it to our understanding of human diseases.

Quantitative genetics 140322



How do we measure the heritability of a complex trait?

(Focusing on the contribution of genes as opposed to the environment?)

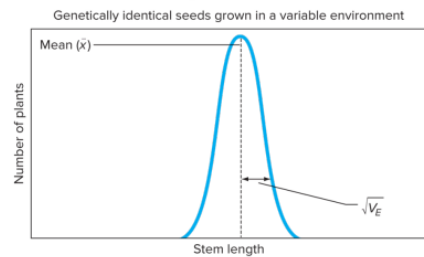
Measure the mean and variance

The variance provides a mathematical description of a distribution; the narrower the curve relative to the peak, the lower the value of the variance.

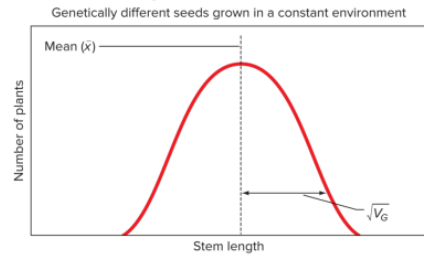
How do we design an experiment to distinguish between genetic and environmental variance?

Environment + genetics compounds variance

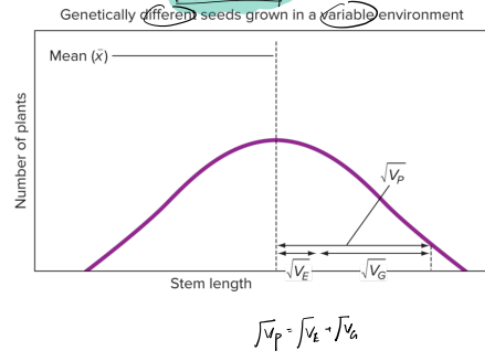
(a) Environmental variance (V_E)



(b) Genetic variance (V_G)



(c) Phenotypic variance ($V_P = V_G + V_E$)



Measuring heritability

We define heritability of a phenotypic trait as the proportion of total phenotypic variance (V_P) ascribable to genetic variation along (V_G).

- V_A : variance due to additive genetic effects
- V_D : variance due to dominance effects
- V_I : variance due to interactions between alleles at different loci (e.g. epistasis)
- V_E : variance due to the environment

$$\left. \begin{array}{l} V_A: \text{variance due to additive genetic effects} \\ V_D: \text{variance due to dominance effects} \\ V_I: \text{variance due to interactions between alleles at different loci (e.g. epistasis)} \\ V_E: \text{variance due to the environment} \end{array} \right\} V_P$$

Broad sense heritability (twin studies)

Two genetically identical individuals

$$\text{Broad-sense } H^2 = \frac{V_G}{V_P}$$

$$V_G = V_A + V_D + V_I, \text{ thus}$$

$$V_P = V_A + V_D + V_I + V_E$$

$$\text{Broad-sense } H^2 = \frac{V_A + V_D + V_I}{V_A + V_D + V_I + V_E} = \frac{V_G}{V_P}$$

We measure broad-sense heritability only in studies that compare identical twins to each other because they share the same alleles at all loci.

Narrow-sense heritability (compare parents and offspring)

$$\text{Narrow-sense } h^2 = \frac{V_A}{V_A + V_D + V_I + V_E} = \frac{V_A}{V_P}$$

Used by plant and animal breeders to predict how strongly a particular trait will respond to selection.

Quantitative genetics: genetic analysis of complex (polygenic) traits

More loci, more phenotypic classes

Quantitative traits[数量性状] are described by a frequency distribution

Types of Quantitative Traits

- Continuous traits: Vary continuously, human height
- Meristic traits: Measured in whole numbers, animal litter size
- Threshold trait: Measured by presence or absence, susceptibility to disease

How could you start to figure out how a genome controls a phenotype?

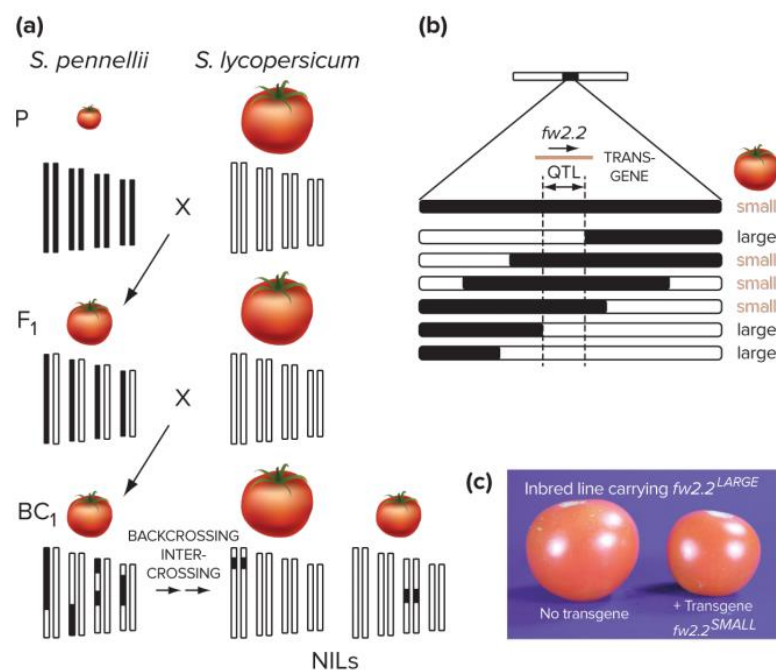
Set up genetic crosses between genomes that control different phenotypes and ask whether genotypes at any genomic regions control phenotype!

Quantitative trait loci (QTL)

Chromosome regions containing a gene or genes that **influence a quantitative trait**

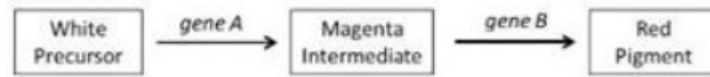
For a quantitative characteristic, each genotype may produce a range of possible phenotypes.

Fine QTL mapping



The power of comparative biology

In the flower petals of a particular plant species, the synthesis of red pigment requires two steps in a pathway as follows:



a) A pure breeding magenta flower plant is crossed to a pure-breeding white flowered plant to produce F1 progeny that consist entirely of red flowered plants. What is the genotype of the white parent?

P1: Magenta AAbb × aaBB White

b) If the red F1 progeny are self-crossed, what is the ratio of phenotypes observed in the F2 generation?

4 White: 3 Magenta: 9 Red

c) What is the term that describes the interaction between genes A and B? Please briefly explain the term in one sentence.

Epistasis, which can be defined as a gene interaction whereby one gene interferes with the phenotypic expression of another non-allelic gene or genes. The gene or locus which suppresses or masks the action of a gene at another locus is called epistatic gene.

d) Scientists have also found an enzyme encoded by the haplo-sufficient dominant allele (D) of gene D that blocks the function of gene B. A plant with genotype AAbbDD is crossed to a plant with genotype AABBdd producing an all magenta F1 generation. If the F1 progeny are self-crossed, what is the ratio of phenotypes observed in the F2 generation?

F1: generation: magenta (AABbDd)

F2: 13 Magenta: 3 Red