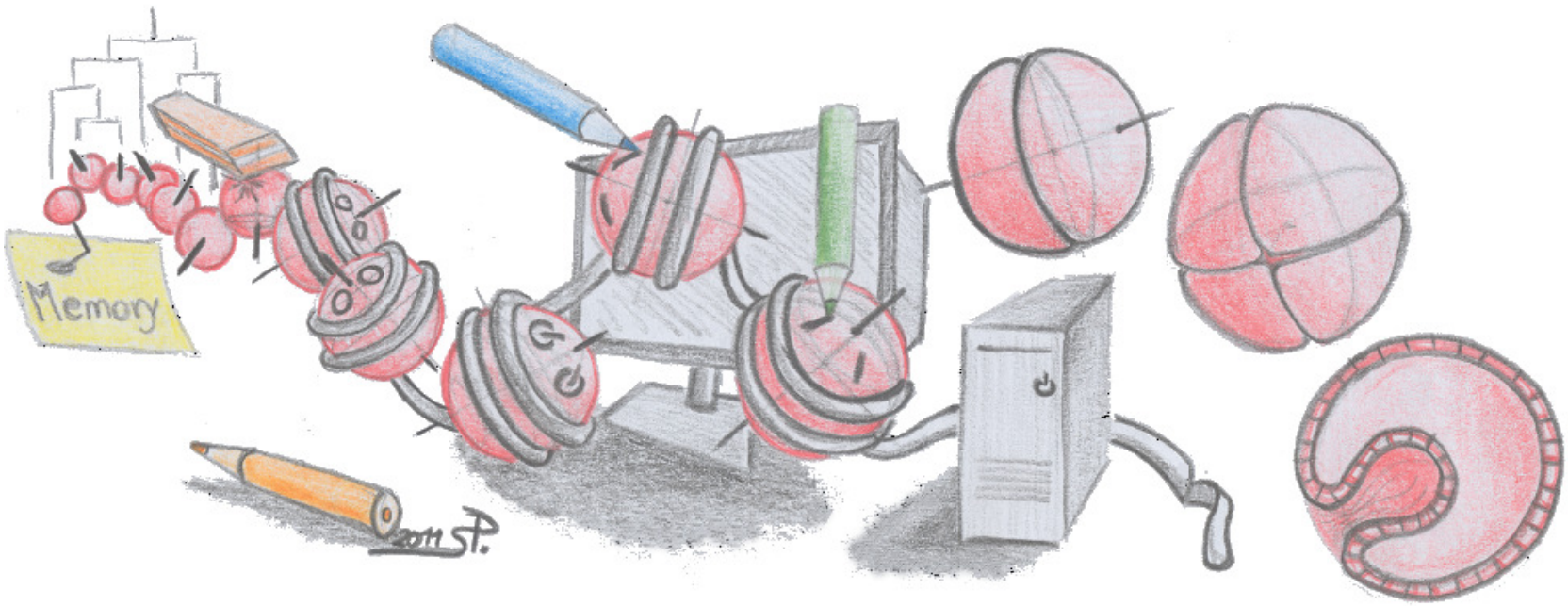
Theoretische Biologie

Prof. Sonja Prohaska

Computational EvoDevo, University of Leipzig

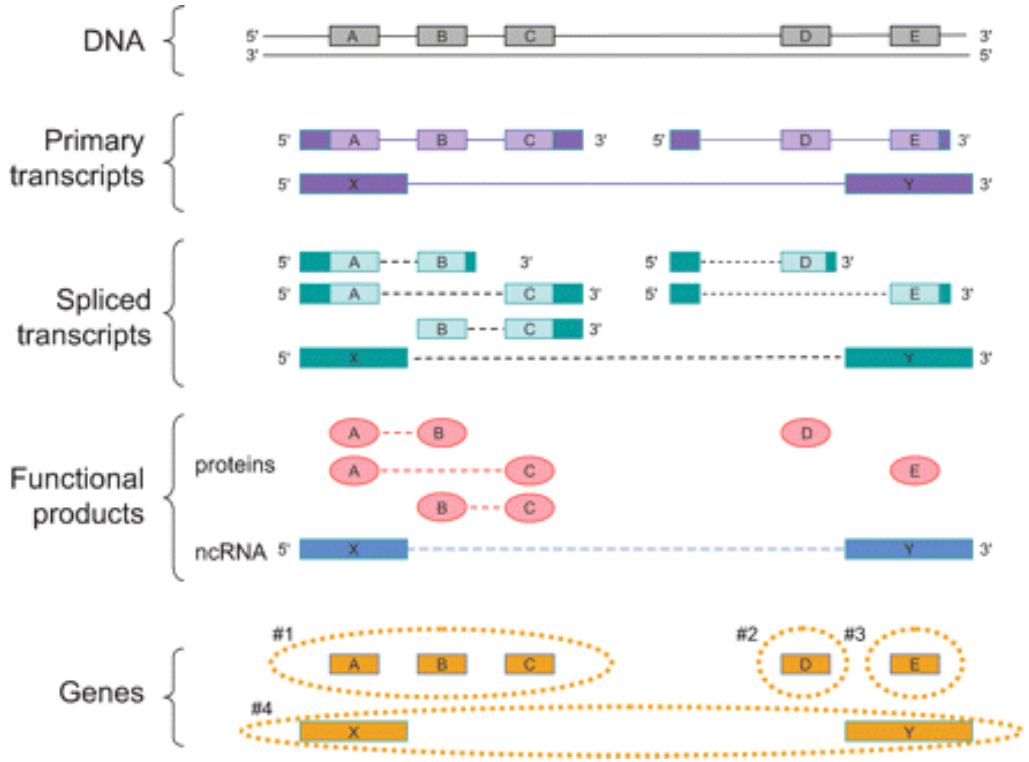


SS 2017

Two Gene Concepts in Comparison

Gerstein-Snyder gene definition

Gerstein MB, Bruce C, Rozowsky JS, Zheng D, Du J, Korbel JO, Emanuelsson O, Zhang ZD, Weissman S and Snyder M (2007) *What is a gene, post-ENCODE? History and updated definition.* Genome Res. 17:669-681.

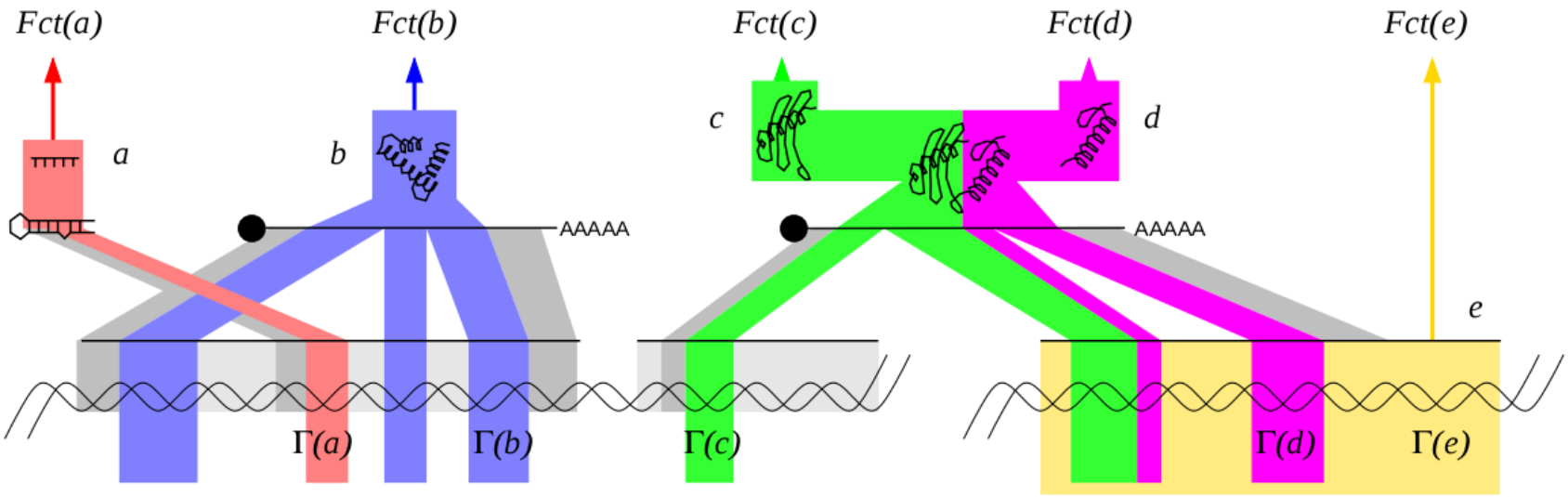


"The gene is a [set] of genomic sequence [fragments] encoding a coherent set of potentially overlapping functional products." Gerstein et al. 2007

"Coherent" meaning that any two functional products (on either RNA or protein level) are directly or indirectly sharing at least one genomic sequence fragment of the set.

The Stadler-Prohaska gene definition

Stadler PF, Prohaska SJ, Forst CV and Krakauer DC (2009). *Defining genes: a computational framework.* Theory in Biosciences 128:165-170.



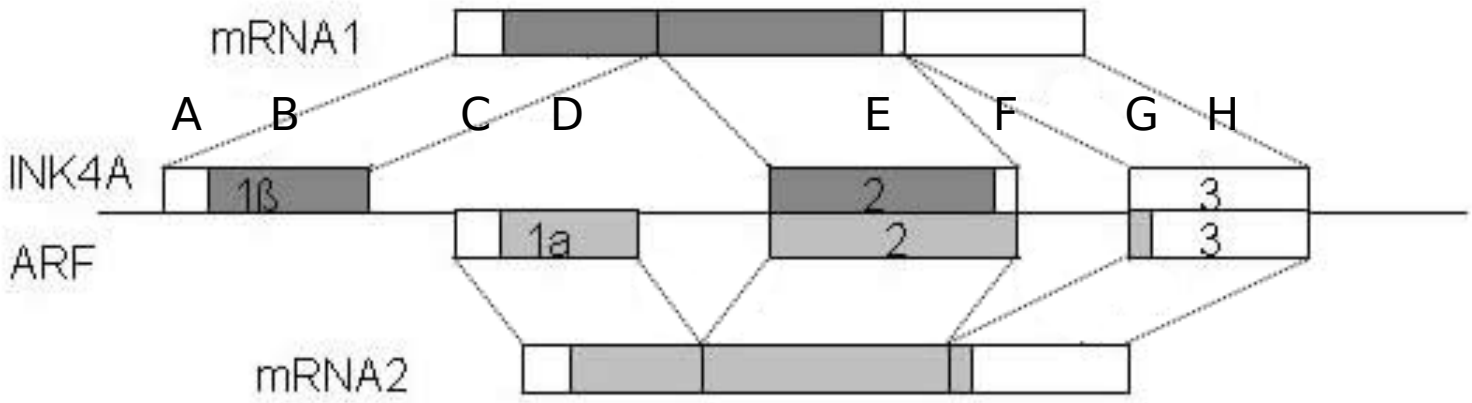
A functional RNA molecule (e.g. a miRNA) with function $Fct(a)$ is processed in two steps from an intronic sequence. Its image on the DNA is the genomic footprint $\Gamma(a)$. The genomic footprint $\Gamma(b)$ of the functional protein b is a discontinuous stretch of DNA corresponding to the coding sequence (CDS). The mRNA includes UTRs that also map back to the DNA. The functional proteins c and d are obtained by cleavage of the (non-functional) precursor cd. The later is encoded by a trans-spliced mRNA. The footprint $\Gamma(c)$ is distributed over two DNA molecules. The primary transcript e has an additional function $Fct(e)$ that is independent of its role as precursor of the mRNA of cd. As a consequence, $\Gamma(e)$ overlaps with both, $\Gamma(c)$ and $\Gamma(d)$.

In all cases, the gene is the pair $(\Gamma(x), x)$ composed of the genomic footprint $\Gamma(x)$ and the resulting functional molecule x.

Example

Gene Annotation: Case 2

The human INK4A/ARF (ARF = alternate reading frame) tumor suppressor region on the short arm of chromosome 9 produces two mRNA transcripts through alternative cis-splicing (intra-molecular splicing) that encode two divergent products, the proteins p16INK4a and p19ARF. Both play important, though very different, roles in cell growth, survival and senescence. The two transcripts have different first coding exons but share the coding sequences of exon 2 and 3. However, the presence of different first exons in the two transcripts causes exons 2 and 3 to be read in different reading frames. Hence there is no amino acid identity between the resulting proteins.



Stotz K, Griffiths PE, and Knight R (2004). *How biologists conceptualize genes: an empirical study* Studies in History and Philosophy of Science Part C: Studies in History and Philosophy of Biological and Biomedical Sciences 35(4):647-673.

Example

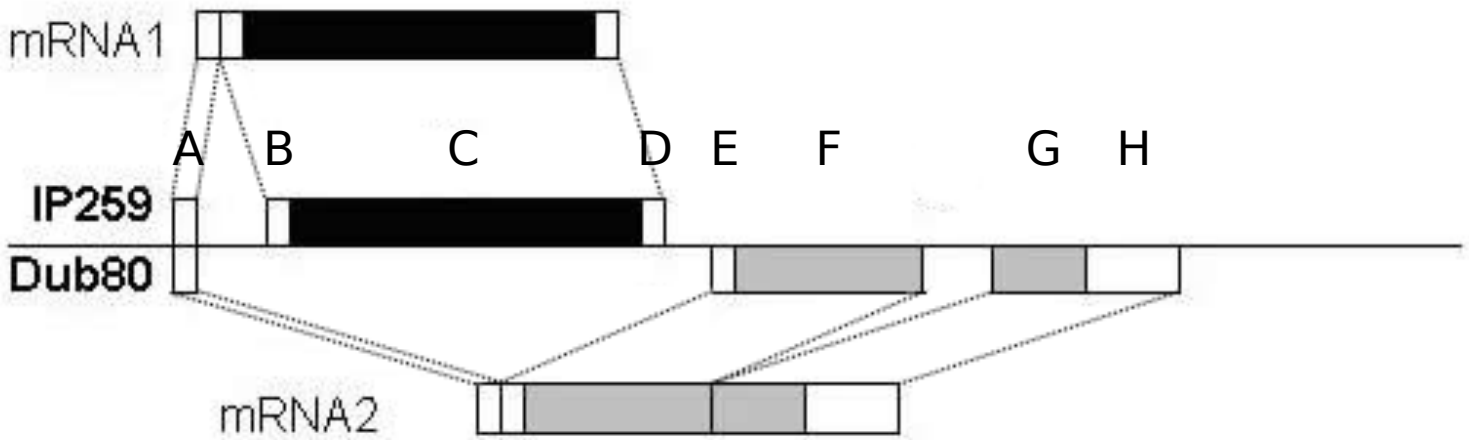
Gene Annotation: Case 2

- **Gerstein's gene definition**: (p677 "frameshift exons") "... there are two sequence sets with common elements ..." and despite that "...the protein products [are] completely different..." "these two proteins' sequences are simultaneously constraint". Therefore, INK4A and ARF are one gene: B,D,(E,F),G.
- **Stadler's gene definition**: two different functional products mapped back to the genome overlap in their footprints. The two genes INK4A (B,E) and ARF (D,(E,F),G) overlap in E only.

Example

Gene Annotation: 3

In the sequence complex IP259/Dub80 in D. melanogaster, alternative cis-splicing (intra-molecular splicing) creates two mature mRNAs that share a noncoding exon as a common translation start site but no overlapping coding sequences, and that encode structurally unrelated proteins (the 259aa protein and a ubiquitin fusion protein). The entire coding region of the IP259 transcript is found in the first intron of Dub80. Note: There is a very similar case in the nematode C. elegans, where the DNA complex ChAT/VAChT (also known as cha-1/unc-17) encodes the structurally unrelated proteins choline acetyl-transferase and vesicular acetylcholine transporter protein. In both the Drosophila case we focus on here and the similar nematode case, the genomic complexes and their products are highly conserved and homologues are found in nematodes, insects and mammals (incl. humans). In both cases each individual mRNA transcript can produce only one of the two products.



Stotz K, Griffiths PE, and Knight R (2004). *How biologists conceptualize genes: an empirical study* Studies in History and Philosophy of Science Part C: Studies in History and Philosophy of Biological and Biomedical Sciences 35(4):647-673.

Example

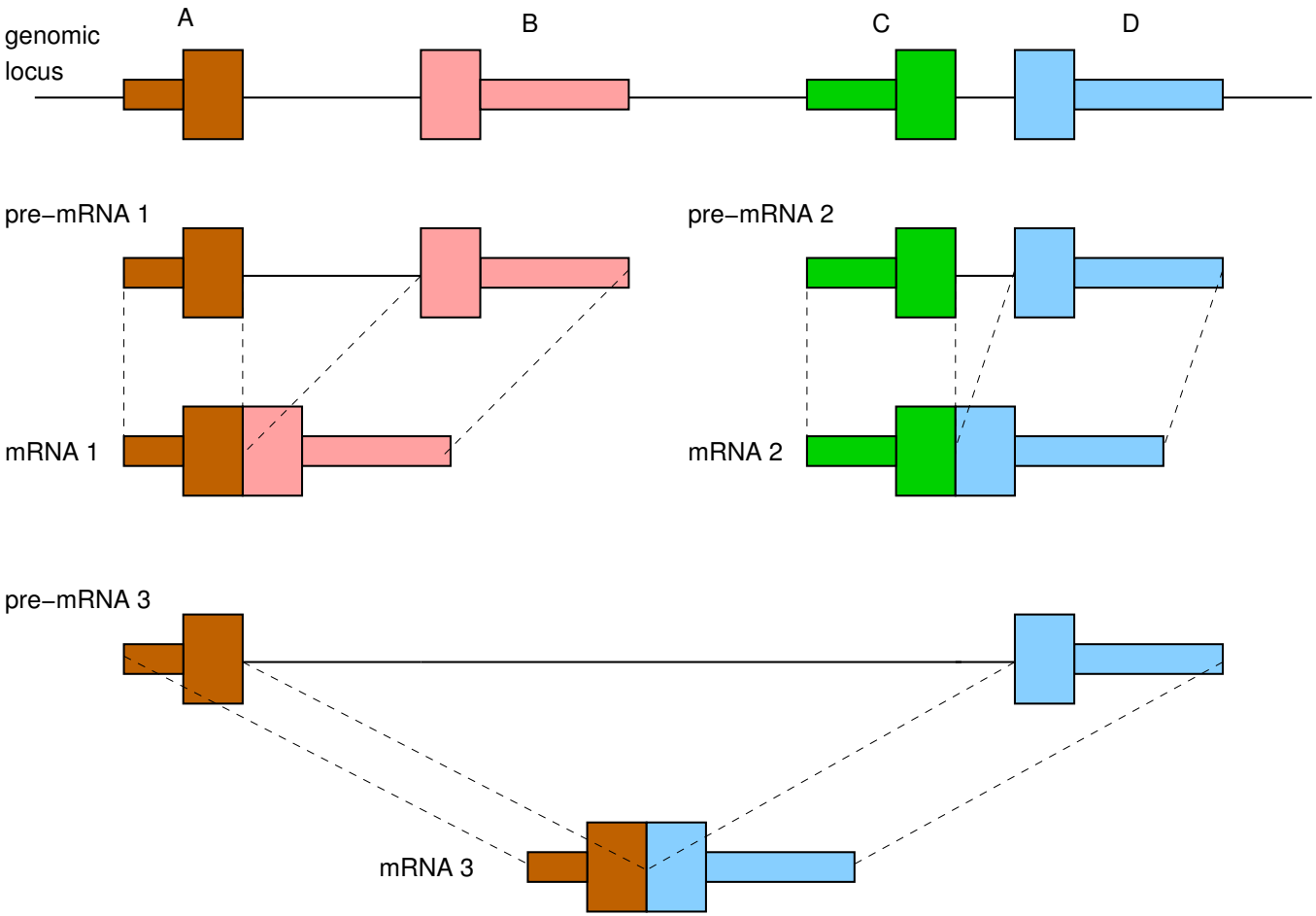
Gene Annotation: Case 3

- **Gerstein's gene definition**: (p677 "regulatory regions not included") "... regulatory regions [...] should not be considered in deciding whether multiple products belong to the same gene." " their is obviously a many-to-many relationship between regulatory regions and genes." Therefore, IP259 (C) and Dub80 (F,G) are two gene.
- **Stadler's gene definition**: two different functional products mapped back to the genome have their own footprints. There are two genes IP259 (C) and Dub80 (F,G) which is in agreement to the above.

Example

Gene Annotation: a Difficult Case

locally duplicated genes and a natural occuring fusion protein



What Any Good Gene Concept Needs

A **structural** and a **functional** component.

- a well-defined structural representation on the genetic material
- a well-defined notion of function (by measurement)
- **heritability** (conveyed by the genetic material)
- compartibility with the concept of **homology** on structural and/or functional level
- processes of **gene expression** connecting the structural and functional components
- relation between mutations and changes in a gene product

We suggest: A gene is a heritable elementary functional unit.

Factors Impairing the Development of a Gene Concept

- limited knowledge
- false information or biases
- too wide or too narrow concept

  e.g. protein-coding mRNAs are polyadenylated

  e.g. genes are be protein coding
- lack of a predictive theory

  can the theory tell us how a gene should look like?

  can it tell us if something is a gene or not?
- limits due to measurment techniques

  mutation and rescue

  crosses and breeding

  transcriptome sequencing

  evolutionary conservation
- context

  e.g. genes expressed in very particular context

Finding (Measuring) Genes Experimentally

The most general strategy: Mutation and Rescue

- given a (haploid, unicellular) organism (wildtype (wt))
- introduce a **mutation** (point mutation)
- that results in a change in phenotype (e.g. loss of function X)
- conclusion: the sequence changed by the mutation must be part of a (gene) sequence responsible for function X

How to find the gene sequence?

- cut the genome of an indentical organism in gene-size pieces
- create a library
- transfect mutants (X^-) with a single piece each
- find the individual that no longer shows the mutant phenotype
- conclusion: the sequence in this individual contains the gene, it **rescues** the wildtype phenotype
- refinement: stepwise shortening of the inserted sequence and re-testing for rescue $\rightarrow$ minimal sequence required is "the gene"

Lazebnik Y (2004). *Can a Biologist Fix a Radio? – or, What I Learned while Studying Apoptosis*. Biochemistry (Moscow), 69(12):1403-1406.

How does a method like this influence the definition of a gene?

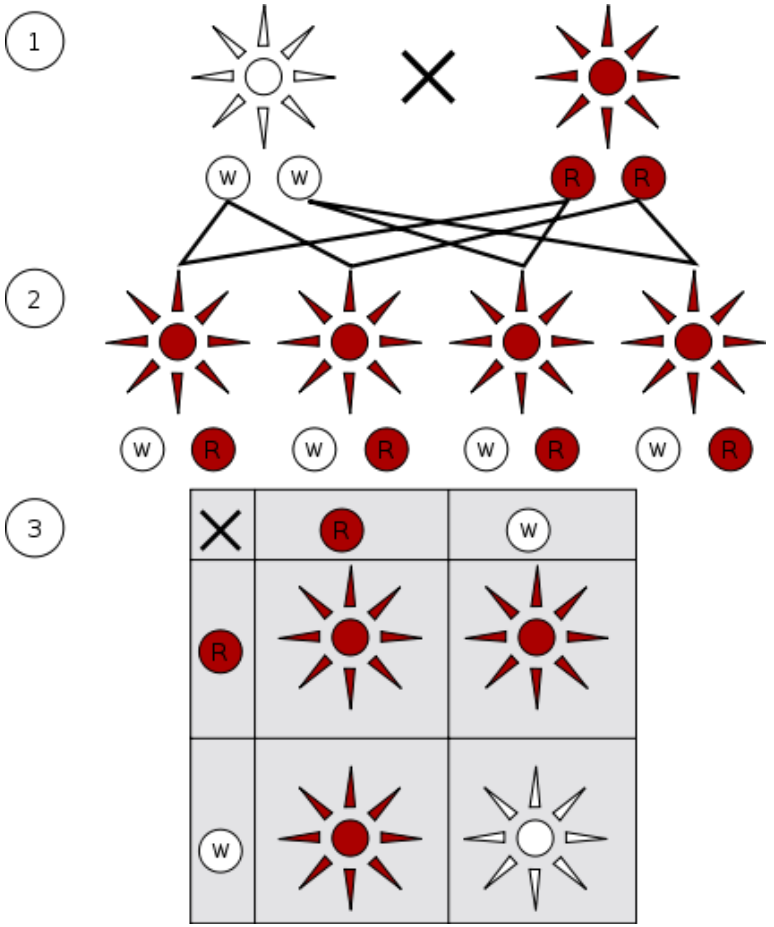
discuss!

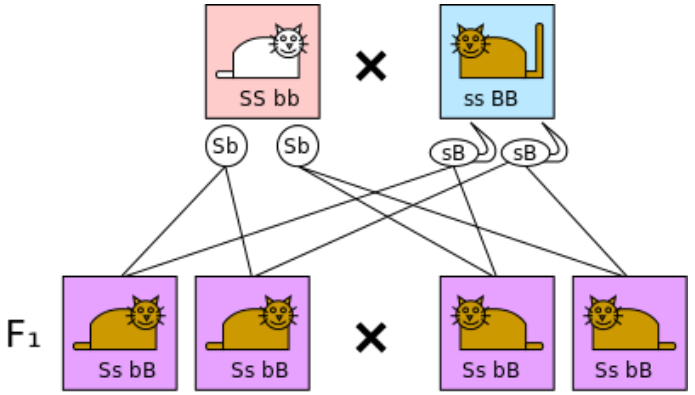
Reminder: Genetics in Eukaryots

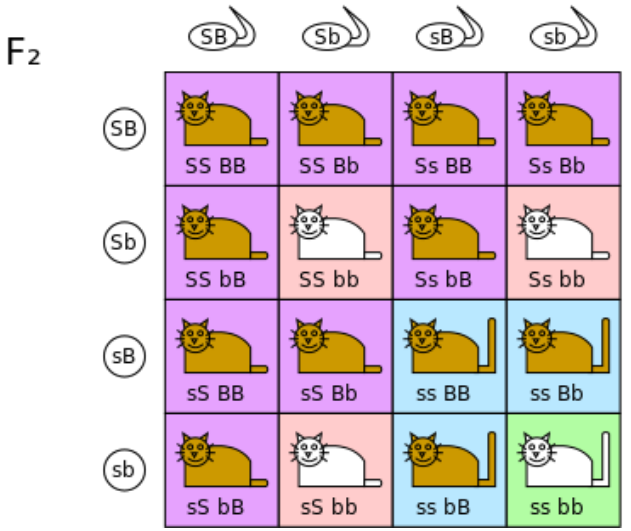
you should be familiar with the folowing terms

- **gene**: region on the DNA that determines a characteristic (abstract)
- **allele**: instance of a gene (concrete)
- **haploid**: n (single chromosome set – in germ cells)
- **diploid**: 2n (two homologous chromosome set – in somatic cells)
- **heterozygote**: individual posessing two different allels of a gene
- **homozygote**: individual posessing two of the same allele of a gene
- **dominant**: an allele X is dominant if the characteristic of X is expressed in homo- and heterozygotes (X/X and X/x)
- **recessive**: an allele x is recessive if the characteristic of x is only expressed in homozygotes (x/x)
- **Mendel's laws**: (Uniformitätsregel, Spaltungsregel, Unabhängigkeitsregel und Rückkreuzung)

Mendel's laws



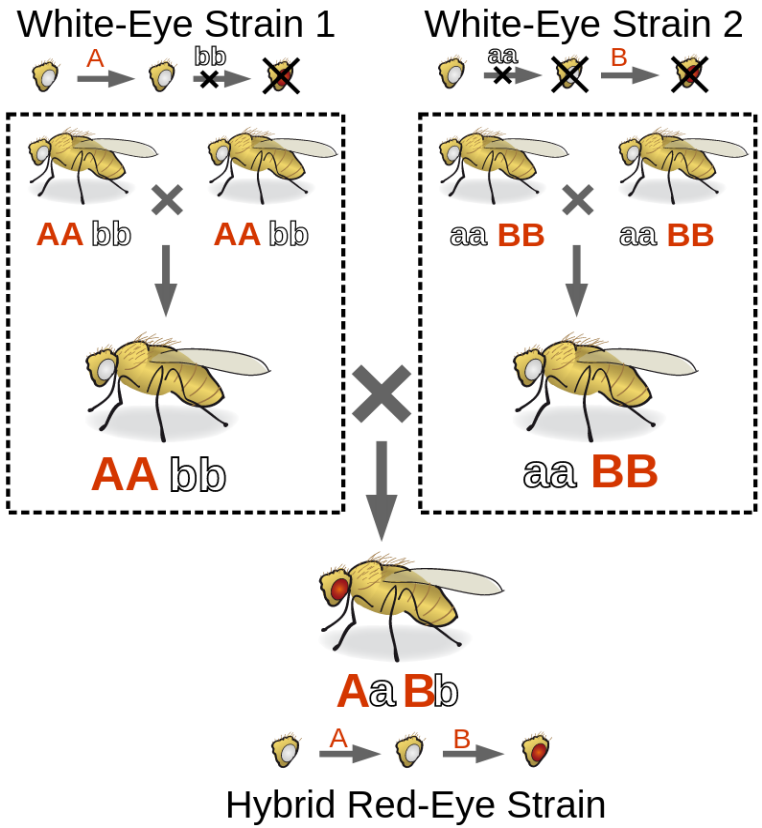




Complementation Test

Many genes are involved in pathways that leed to an observable phenotype.
A mutated allel a^- is most often recessive to the dominant wildetype allel A^{wt}.
Assume homozygote a^- and homozygote b^- individuals have the same phenotype.

Do the mutations a^- and b^- occur in the same gene?



Measurement and Definition are Tied Together

what you find and what you miss

- mutation and rescue (bacteria)

  miss: essential genes when any mutation is lethal

  find: regulatory sequences (e.g. promoters) behave like genes
- crosses and breeding (eukarya)

  similar to above

  generally more difficult due to complex phenotypic characters
- transcriptome sequencing

  miss: low expessed genes, outside the length range analysed

  find: hypotetical genes, “junk” transcripts
- evolutionary conservation

  find: conservative “genes” (any changes negatively selected)

  miss: diverging and novel genes under positive selection