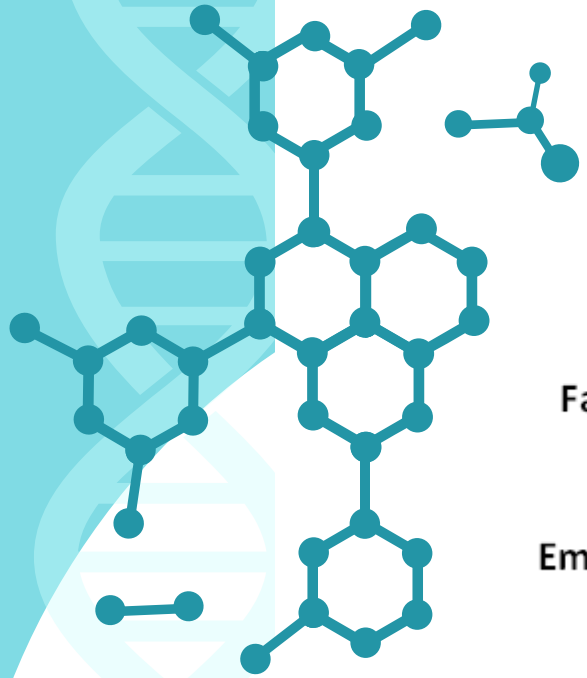


Program of Medical Informatics
Faculty of Computers and Information
Mansoura University

Genetic Variations

(Genomics)

Sara El-Metwally, Ph.D.
Faculty of Computers and Information,
Mansoura University, Egypt.

Email: sarah_almetwally4@mans.edu.eg
sara.elmetwally.2007@gmail.com

Office: Faculty of CIS, third floor

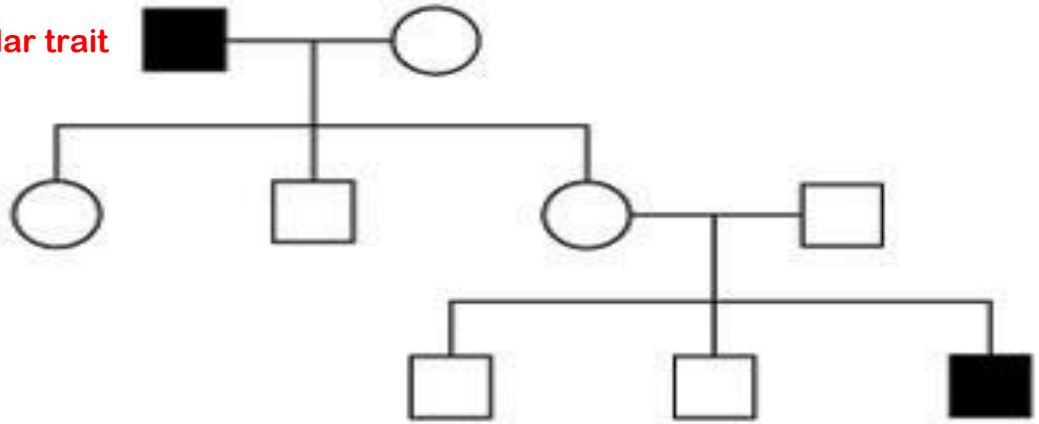
Basics

- Genes come in different varieties, called alleles.
- Somatic cells contain two alleles for every gene, with one allele provided by each parent of an organism.
- It is impossible to determine which two alleles of a gene are present within an organism's chromosomes based on its phenotype characteristics.
- An allele that is hidden, or not expressed by an organism, can still be passed on to that organism's offspring and expressed in a later generation.

https://www.nature.com/scitable/topicpage/inheritance-of-traits-by-offspring-follows-predictable-6524925

Basics

Male with a particular trait



The family tree in the above figure shows how an allele can disappear or "hide" in one generation and then reemerge in a later generation.

https://www.nature.com/scitable/topicpage/inheritance-of-traits-by-offspring-follows-predictable-6524925

Basics

- An individual gene may code for a specific physical trait, and that gene can exist in different forms, or alleles.

- One allele for every gene in an organism is inherited from each of that organism's parents.

- In some cases, both parents provide the same allele of a given gene, and the offspring is referred to as homozygous ("homo" meaning "same") for that allele.

- In other cases, each parent provides a different allele of a given gene, and the offspring is referred to as heterozygous ("hetero" meaning "different") for that allele.

https://www.nature.com/scitable/topicpage/inheritance-of-traits-by-offspring-follows-predictable-6524925

Basics

- Alleles produce phenotypes (or physical versions of a trait) that are either dominant or recessive

- The dominance or recessivity associated with a particular allele is the result of masking, by which a dominant phenotype hides a recessive phenotype.

- By this logic, in heterozygous offspring only the dominant phenotype will be apparent.

https://www.nature.com/scitable/topicpage/inheritance-of-traits-by-offspring-follows-predictable-6524925

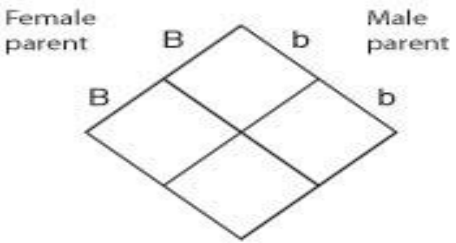
Basics

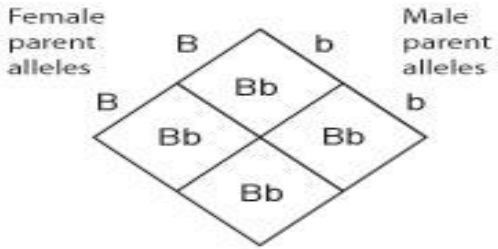
✓In fruit flies, the gene for body color has two different alleles: the black allele and the brown allele.

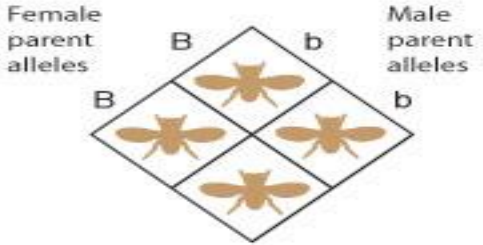
✓Moreover, brown body color is the dominant phenotype, and black body color is the recessive phenotype.

https://www.nature.com/scitable/topicpage/inheritance-of-traits-by-offspring-follows-predictable-6524925

Basics





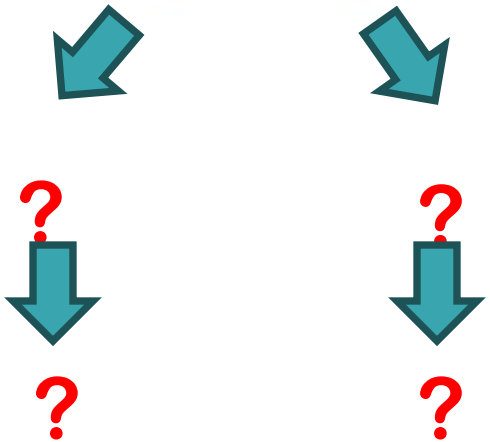


https://www.nature.com/scitable/topicpage/inheritance-of-traits-by-offspring-follows-predictable-6524925

Basics

Phenotype

Genotype



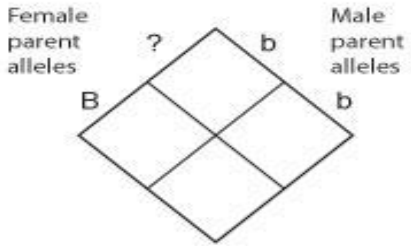
https://www.nature.com/scitable/topicpage/inheritance-of-traits-by-offspring-follows-predictable-6524925

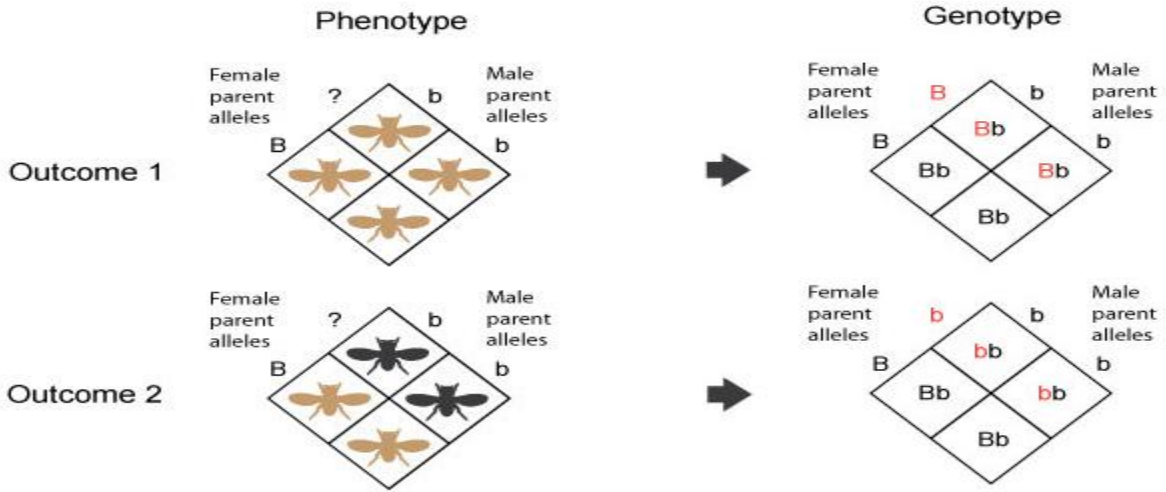
Basics

- Brown flies can be either homozygous (BB) or heterozygous (Bb).
- Is it possible to determine whether a female fly with a brown body has the genotype BB or Bb?
- To answer this question, an experiment called a test cross can be performed.
- Test crosses help researchers determine the genotype of an organism when only its phenotype (i.e., its appearance) is known.

https://www.nature.com/scitable/topicpage/inheritance-of-traits-by-offspring-follows-predictable-6524925

Basics





https://www.nature.com/scitable/topicpage/inheritance-of-traits-by-offspring-follows-predictable-6524925

What is genetic variation?

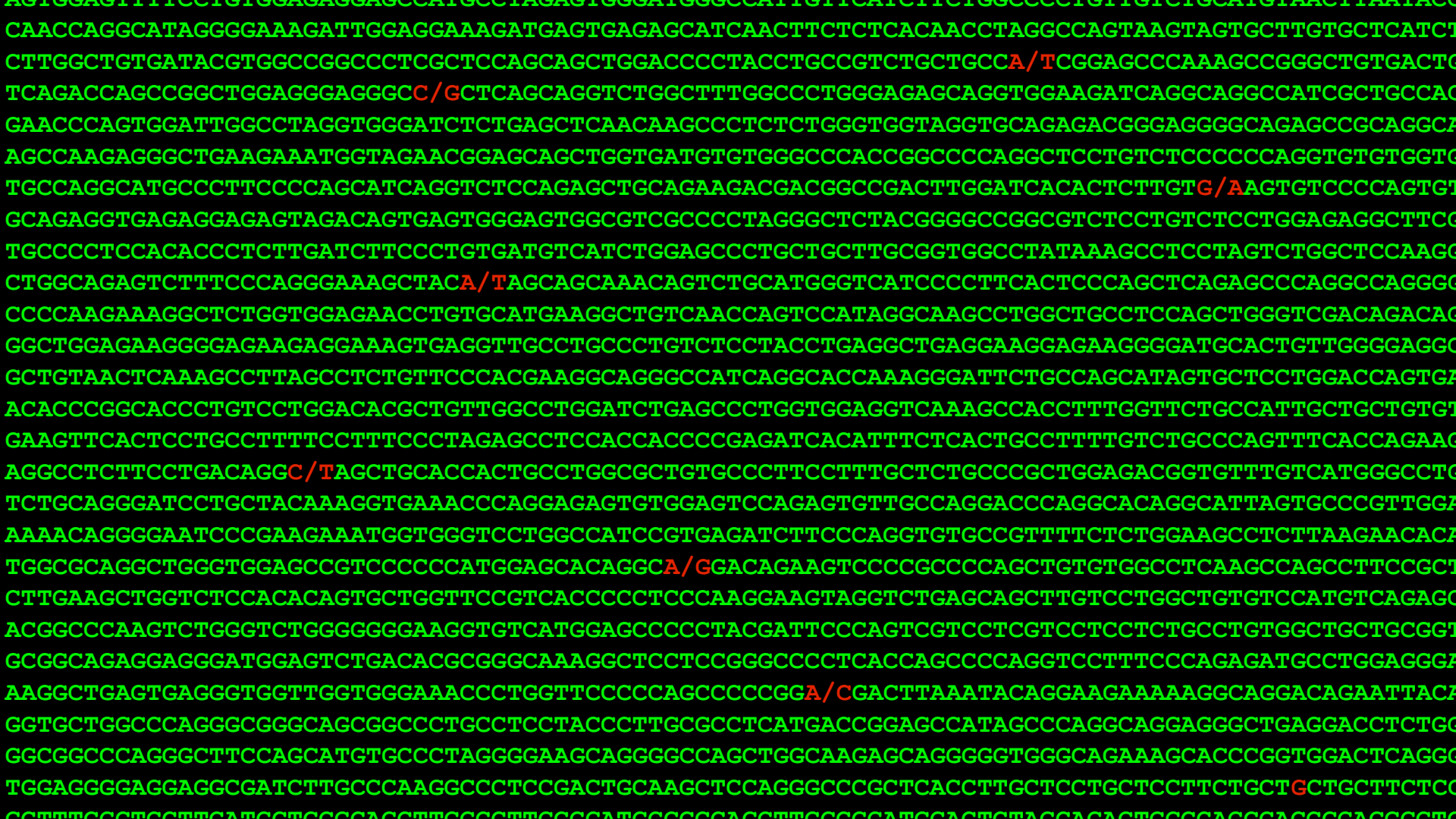
- Differences in DNA content or structure among individuals
  - Any two individuals have ~99.5% identical DNA.
- But the human genome is big - each haploid set of 23 chromosomes has 3.1 billion nucleotides.
  - There are >100,000,000 know genetic variants in the human genome
- Effectively infinite combinations of alleles. The details matter.

~99.5% identical DNA (differ at 1/ 620 - 1/750 bp)

Drosophila - 1/180

99% identical DNA

Slides adapted from Aaron Quinlan: https://github.com/quinlan-lab/applied-computational-genomics


CAACCAGGCATAGGGGAAAGATTGGAGGAAAGATGAGTGAGAGCATCAACTTCTCTCACAACCTAGGCCAGTAAGTAGTGCTTGTGCTCATCT
CTTGGCTGTGATACGTGGCCGGCCCTCGCTCCAGCAGCTGGACCCCTACCTGCCGTCTGCTGCCA/TCGGAGCCCAAAGCCGGGCTGTGACTG
TCAGACCAGCCGGCTGGAGGGAGGGCC/GCTCAGCAGGTCTGGCTTTGGCCCTGGGAGAGCAGGTGGAAGATCAGGCAGGCCATCGCTGCCAG
GAACCCAGTGGATTGGCCTAGGTGGGATCTCTGAGCTCAACAAGCCCTCTCTGGGTGGTAGGTGCAGAGACGGGAGGGGCAGAGCCGCAGGCA
AGCCAAGAGGGCTGAAGAAATGGTAGAACGGAGCAGCTGGTGATGTGTGGGCCCACCGGCCCCAGGCTCCTGTCTCCCCCCAGGTGTGTGGTG
TGCCAGGCATGCCCTTCCCCAGCATCAGGTCTCCAGAGCTGCAGAAGACGACGGCCGACTTGGATCACACTCTTGTG/AAGTGTCCCCAGTGT
GCAGAGGTGAGAGGAGAGTAGACAGTGAGTGGGAGTGGCGTCGCCCCTAGGGCTCTACGGGGCCGGCGTCTCCTGTCTCCTGGAGAGGCTTCG
TGCCCCTCCACACCCTCTTGATCTTCCCTGTGATGTCATCTGGAGCCCTGCTGCTTGCGGTGGCCTATAAAGCCTCCTAGTCTGGCTCCAAGG
CTGGCAGAGTCTTTCCCAGGGAAAGCTACA/TAGCAGCAAACAGTCTGCATGGGTCATCCCCTTCACTCCCAGCTCAGAGCCCAGGCCAGGGG
CCCCAAGAAAGGCTCTGGTGGAGAACCTGTGCATGAAGGCTGTCAACCAGTCCATAGGCAAGCCTGGCTGCCTCCAGCTGGGTCGACAGACAG
GGCTGGAGAAGGGGAGAAGAGGAAAGTGAGGTTGCCTGCCCTGTCTCCTACCTGAGGCTGAGGAAGGAGAAGGGGATGCACTGTTGGGGAGGG
GCTGTAACTCAAAGCCTTAGCCTCTGTTCCCACGAAGGCAGGGCCATCAGGCACCAAAGGGATTCTGCCAGCATAGTGCTCCTGGACCAGTGA
ACACCCGGCACCCTGTCCTGGACACGCTGTTGGCCTGGATCTGAGCCCTGGTGGAGGTCAAAGCCACCTTTGGTTCTGCCATTGCTGCTGTGT
GAAGTTCACTCCTGCCTTTTCCTTTCCCTAGAGCCTCCACCACCCCGAGATCACATTTCTCACTGCCTTTTGTCTGCCCAGTTTCACCAGAAG
AGGCCTCTTCCTGACAGGC/TAGCTGCACCACTGCCTGGCGCTGTGCCCTTCCTTTGCTCTGCCCGCTGGAGACGGTGTTTGTCATGGGCCTG
TCTGCAGGGATCCTGCTACAAAGGTGAAACCCAGGAGAGTGTGGAGTCCAGAGTGTTGCCAGGACCCAGGCACAGGCATTAGTGCCCGTTGGA
AAAACAGGGGAATCCCGAAGAAATGGTGGGTCCTGGCCATCCGTGAGATCTTCCCAGGTGTGCCGTTTTCTCTGGAAGCCTCTTAAGAACACA
TGGCGCAGGCTGGGTGGAGCCGTCCCCCCATGGAGCACAGGCA/GGACAGAAGTCCCCGCCCCAGCTGTGTGGCCTCAAGCCAGCCTTCCGCT
CTTGAAGCTGGTCTCCACACAGTGCTGGTTCCGTCACCCCCTCCCAAGGAAGTAGGTCTGAGCAGCTTGTCCTGGCTGTGTCCATGTCAGAGC
ACGGCCCAAGTCTGGGTCTGGGGGGGAAGGTGTCATGGAGCCCCCTACGATTCCCAGTCGTCCTCGTCCTCCTCTGCCTGTGGCTGCTGCGGT
GCGGCAGAGGAGGGATGGAGTCTGACACGCGGGCAAAGGCTCCTCCGGGCCCCTCACCAGCCCCAGGTCCTTTCCCAGAGATGCCTGGAGGGA
AAGGCTGAGTGAGGGTGGTTGGTGGGAAACCCTGGTTCCCCCAGCCCCCGGA/CGACTTAAATACAGGAAGAAAAAGGCAGGACAGAATTACA
GGTGCTGGCCCAGGGCGGGCAGCGGCCCTGCCTCCTACCCTTGCGCCTCATGACCGGAGCCATAGCCCAGGCAGGAGGGCTGAGGACCTCTGG
GGCGGCCCAGGGCTTCCAGCATGTGCCCTAGGGGAAGCAGGGGCCAGCTGGCAAGAGCAGGGGGTGGGCAGAAAGCACCCGGTGGACTCAGGG
TGGAGGGGAGGAGGCGATCTTGCCCAAGGCCCTCCGACTGCAAGCTCCAGGGCCCGCTCACCTTGCTCCTGCTCCTTCTGCTGCTGCTTCTCC

Types of genetic variation

```
ctccgag
ctctgag
```

Single-nucleotide polymorphisms (**SNPs**)

"DNA spelling mistakes"

```
ctc--ag
ctctgag
```

Insertion-deletion polymorphisms (**INDELs**)

"extra or missing DNA"

Structural variants (**SVs**)

"Large blocks of extra, missing or rearranged DNA"

Slides adapted from Aaron Quinlan: https://github.com/quinlan-lab/applied-computational-genomics

SNPs vs. SNVs

- SNV stands for single nucleotide variation, which means at one base there is a difference. This definition does not imply any kind of how often this variation occurs.

- SNV example (somatic mutations in cancer)

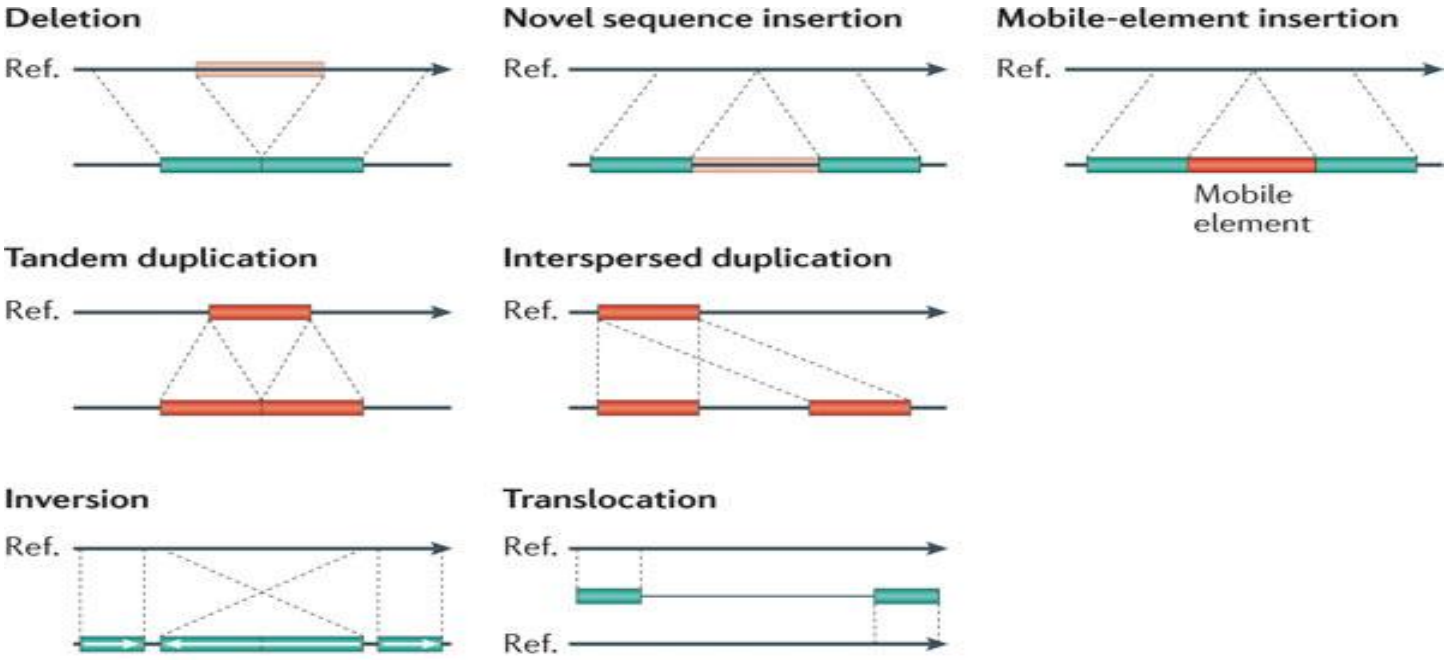
- SNP stands single nucleotide polymorphism does imply that there is an implication how often this variation occurs. when someone refers to a SNP that this variation is common (at least 1% of population.

https://www.quora.com/Genomics-What-is-the-difference-between-an-SNP-and-an-SNV/answer/Nathan-Johnson-129

SNPs vs. SNVs

SNPs	SNVs
Occurrence expected at the position for any member in the species (well-characterized)	Occurrence seen in only one individual (not well characterized)
Occur in population at some frequency so expected at a given locus	Occur at low frequency so not common
Validated in population	Not validated in population
Catalogued in dbSNP (http://www.ncbi.nlm.nih.gov/snp)	

Classes of SVs



Alkan, C. et al. Genome structural variation discovery and genotyping. Nature Reviews Genetics 12, 363-376 (2011).

Slides adapted from NGS Workshop Variant Calling by Ramesh Nair

What is CNVs

- CNVs stands for copy number variations and defined as a DNA segment of one kilobase (kb) or larger that is present at a variable copy number in comparison with a reference genome.

- Some CNVs have no apparent influence on phenotype, while as many as 40 others have been definitively linked with disease.

- Evidence also indicates that interaction with additional genetic or environmental factors may influence whether CNVs have a detectable phenotypic effect.

https://www.nature.com/scitable/topicpage/copy-number-variation-445

What is CNVs

- The gene copy number (also "copy number variants" or CNVs) is the number of copies of a particular gene in the genotype of an individual.

- Recent evidence shows that the gene copy number can be elevated in cancer cells.

- It was generally thought that genes were almost always present in two copies in a genome. However, recent discoveries have revealed that large segments of DNA, ranging in size from thousands to millions of DNA bases, can vary in copy-number.

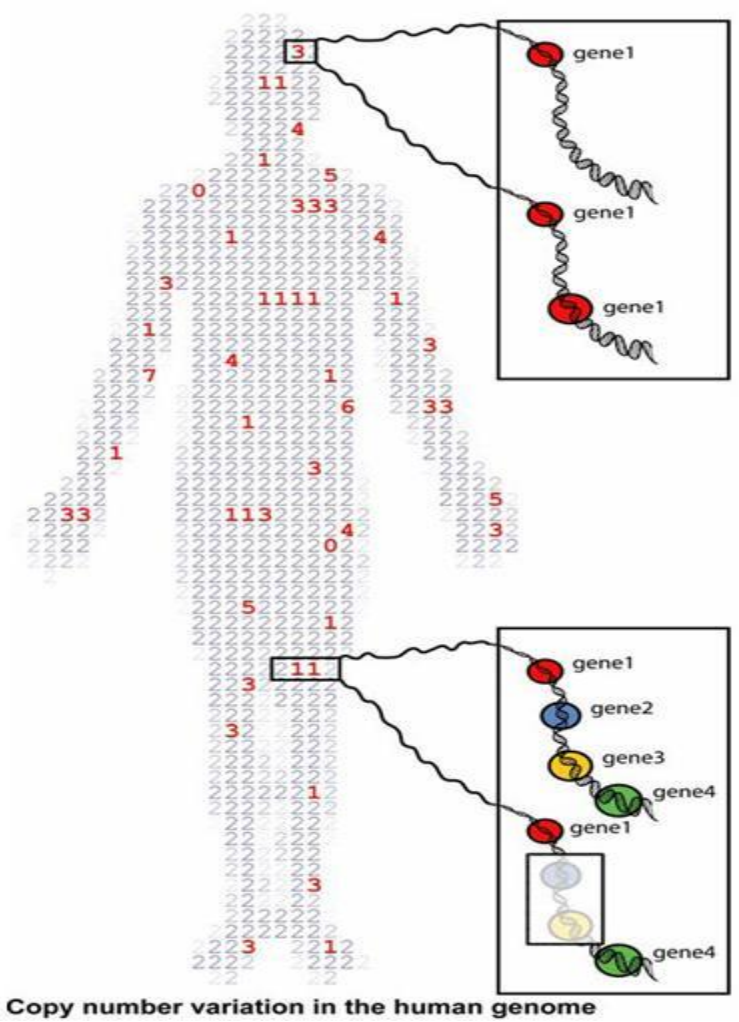
http://cnv.gene-quantification.info/

What is CNVs

- For example, genes that were thought to always occur in two copies per genome have now been found to sometimes be present in one, three, or more than three copies. In a few rare instances the genes are missing altogether

- **Important:** read this

  https://www.ncbi.nlm.nih.gov/dbvar/content/overview/



Copy number variation in the human genome

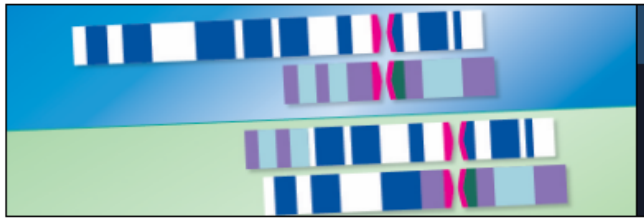
Genetic variation

https://www.ncbi.nlm.nih.gov/dbvar/

NCBI Resources How To Sign in to NCB

dbVar dbVar Search

Advanced He



dbVar

dbVar is NCBI's database of human genomic structural variation — insertions, deletions, duplications, inversions, mobile elements, translocations, and others

Getting Started

- Overview of Structural Variation
- Organism List
- FAQ
- Help

Accessing Data

- Structural Variation Data Hub
- Tools for analyzing dbVar data
- Study Browser
- Genome Browser

Other NCBI Resources

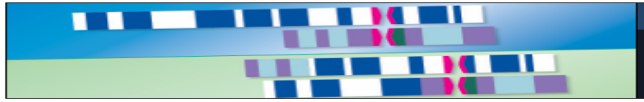
- dbSNP
- ClinVar
- Variation Portal
- Variation Tools

https://www.ncbi.nlm.nih.gov/dbvar/

NCBI Resources How To Sign in to NCBI

dbVar dbVar Search

Advanced



dbVar

dbVar is NCBI's database of human genomic struc... mobile elements, translocations, and others

Genetic variation

Getting Started

Overview of Structural Variation
Organism List
FAQ
Help

Accessing Data

Structural Variation Data Hub
Tools for analyzing dbVar data
Study Browser
Genome Browser

Other NCBI Resources

dbSNP
ClinVar
Variation Portal
Variation Tools

Variant Call Type	Sequence Ontology ID
copy number gain	**SO:0001742** A sequence alteration whereby the copy number of a given region is greater than the reference sequence.
copy number loss	**SO:0001743** A sequence alteration whereby the copy number of a given region is less than the reference sequence.
duplication	**SO:0001742** (copy number gain) A sequence alteration whereby the copy number of a given region is greater than the reference sequence.
deletion	**SO:0000159** The point at which one or more contiguous nucleotides were excised.
insertion	**SO:0000667** The sequence of one or more nucleotides added between two adjacent nucleotides in the sequence.
mobile element insertion	**SO:0001837** A kind of insertion where the inserted sequence is a mobile element.
novel sequence insertion	**SO:0001838** An insertion the sequence of which cannot be mapped to the reference genome.

Catalogs of Human Genetic Variation

- The 1000 Genomes Project
  - http://www.1000genomes.org/
  - SNPs and structural variants
  - genomes of about 2500 unidentified people from about 25 populations around the world will be sequenced using NGS technologies
- HapMap
  - http://hapmap.ncbi.nlm.nih.gov/
  - identify and catalog genetic similarities and differences
- dbSNP
  - http://www.ncbi.nlm.nih.gov/snp/
  - Database of SNPs and multiple small-scale variations that include indels, microsatellites, and non-polymorphic variants
- COSMIC
  - http://www.sanger.ac.uk/genetics/CGP/cosmic/
  - Catalog of Somatic Mutations in Cancer

Slides adapted from NGS Workshop Variant Calling by Ramesh Nair

A typical human genome

"We find that a typical [human] genome differs from the reference human genome at **4.1 million to 5.0 million sites**. Although **›99.9% of variants consist of SNPs and short indels**, structural variants affect more bases: the typical genome contains an estimated **2,100 to 2,500 structural variants** (~1,000 large deletions, ~160 copy-number variants, ~915 Alu insertions, ~128 L1 insertions, ~51 SVA insertions, ~4 NUMTs, and ~10 inversions), **affecting ~20 million bases of sequence.**

A global reference for human genetic variation

The 1000 Genomes Project Consortium*

The 1000 Genomes Project set out to provide a comprehensive description of common human genetic variation by applying whole-genome sequencing to a diverse set of individuals from multiple populations. Here we report completion of the project, having reconstructed the genomes of 2,504 individuals from 26 populations using a combination of low-coverage whole-genome sequencing, deep exome sequencing, and dense microarray genotyping. We characterized a broad spectrum of genetic variation, in total over 88 million variants (84.7 million single nucleotide polymorphisms (SNPs), 3.6 million short insertions/deletions (indels), and 60,000 structural variants), all phased onto high-quality haplotypes. This resource includes >99% of SNP variants with a frequency of >1% for a variety of ancestries. We describe the distribution of genetic variation across the global sample, and discuss the implications for common disease studies.

Nucleotide diversity (π):

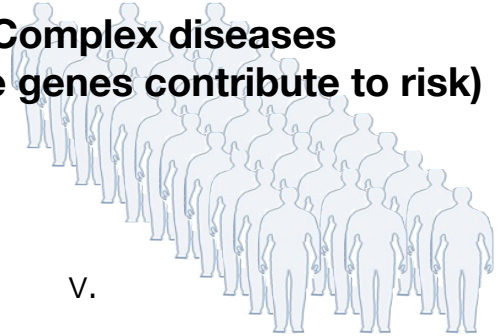
1/756 bp to 1/620 bp

Slides adapted from Aaron Quinlan: https://github.com/quinlan-lab/applied-computational-genomics

Why do we care?

Cases
(have disease)

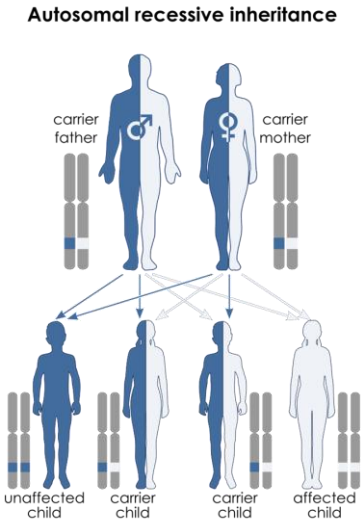
v.



Complex diseases
(multiple genes contribute to risk)

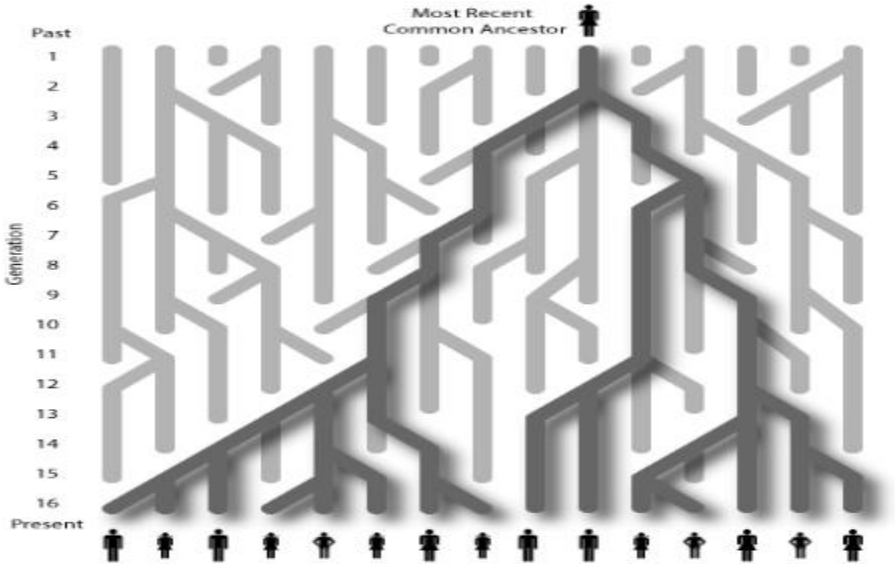
Controls
(no disease)

Understanding the relationship between genetic variation and traits or disease phenotypes



Slides adapted from Aaron Quinlan: https://github.com/quinlan-lab/applied-computational-genomics

Genetic variation



Slides adapted from Aaron Quinlan: https://github.com/quinlan-lab/applied-computational-genomics