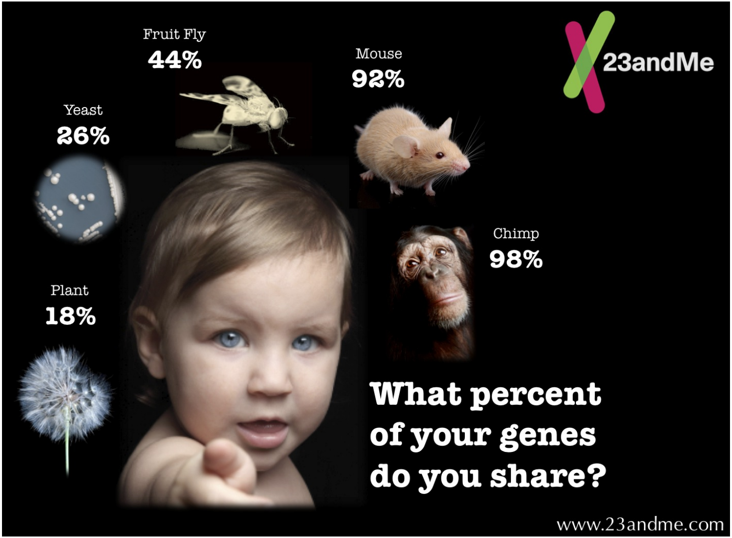
Genome sequencing intro

Mikhail Dozmorov

Spring 2018

What is genetic variation?



https://blog.23andme.com/23andme-and-you/genetics-101/genetic-similarities-of-mice-and-men/

What is genetic variation?

- Differences in DNA content or structure among individuals.
- Any two individuals have ~99.5% identical DNA.

Photo by Anthony Malloy

- But the human genome is big - each haploid set of 23 chromosomes has 3.1 billion nucleotides.
- There are >88,000,000 know genetic variants in the human genome.
- Effectively infinite combinations of alleles. The details matter.

Types of genetic variation

```
ctccgag
ctctgag
```

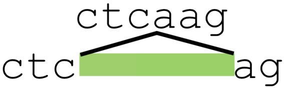
Single-nucleotide polymorphisms (**SNPs**)

“DNA spelling mistakes”

```
ctc--ag
ctctgag
```

Insertion-deletion polymorphisms (**INDELs**)

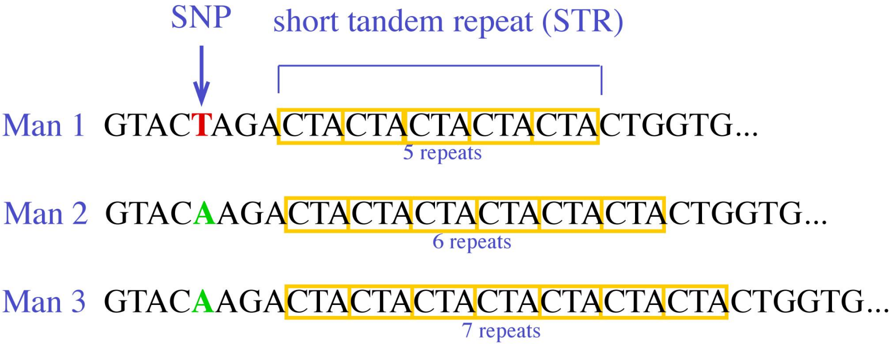
“extra or missing DNA”



Structural variants (**SVs**)

“Large blocks of extra, missing or rearranged DNA”

Types of genetic variation



SNP

short tandem repeat (STR)

Man 1 GTACTAGACTACTACTACTACTACTGGTG...

5 repeats

Man 2 GTACAAGACTACTACTACTACTACTACTGGTG...

6 repeats

Man 3 GTACAAGACTACTACTACTACTACTACTACTGGTG...

7 repeats

A typical human genome variation

- "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**.

https://www.nature.com/nature/journal/v526/n7571/full/nature15393.html

Why do we care?

- Complex diseases (multiple genes contribute to risk)
- Understanding the relationship between genetic variation and traits or disease phenotypes

NATIONAL CANCER INSTITUTE
PRECISION MEDICINE
IN CANCER TREATMENT

Discovering unique therapies that treat an individual's cancer based on the specific genetic abnormalities of that person's tumor.

Mutation vs. polymorphism

- Mutation: *private* to this chromosome / individual

acctccgagta	acctccgagta
acctccgagta	acctccgagta
acctccgagta	acctccgagta
acctccgagta	acctccgagta
acctccgagta	acctc**T**gagta

Mutation vs. polymorphism

- From private mutation to a more common polymorphism

acctccgagta

acctccgagta

acctccgagta

acctc**T**gagta

acctccgagta

acctc**T**gagta

acctccgagta

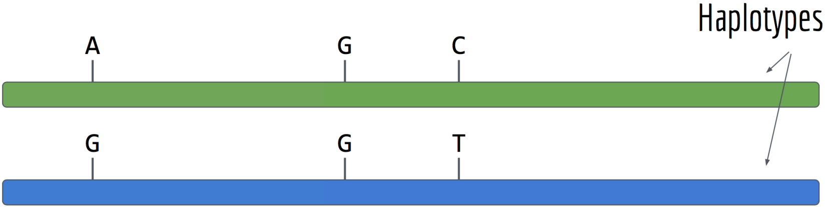
acctc**T**gagta

acctccgagta

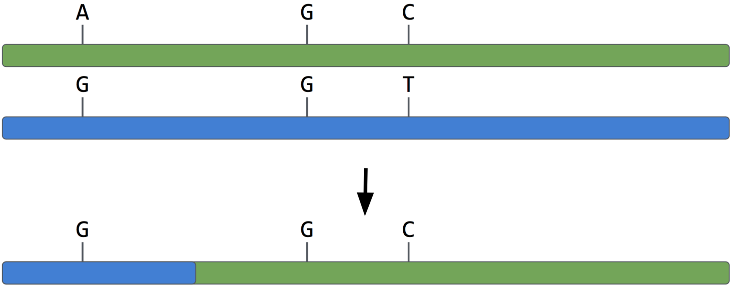
acctc**T**gagta

How SNPs arise

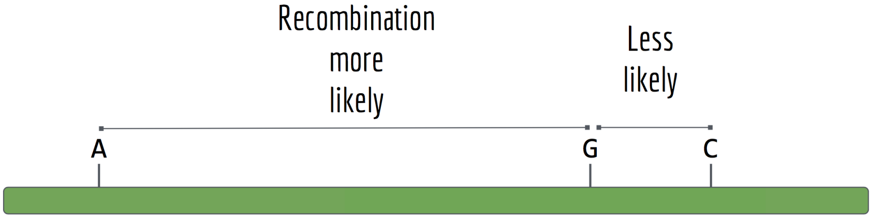
- Haplotype - a group of genes or DNPs *inherited together*
- A child inherits two haplotypes - one from dad and one from mom



Meiotic recombination shuffles alleles and generates new haplotypes



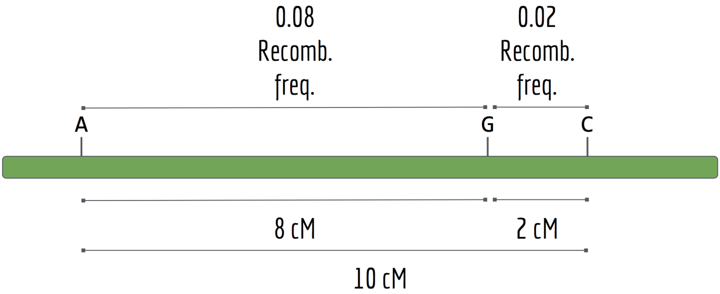
Genetic linkage



- The greater the frequency of recombination (segregation) between two genetic markers, the further apart they are assumed to be.

https://en.wikipedia.org/wiki/Genetic_linkage

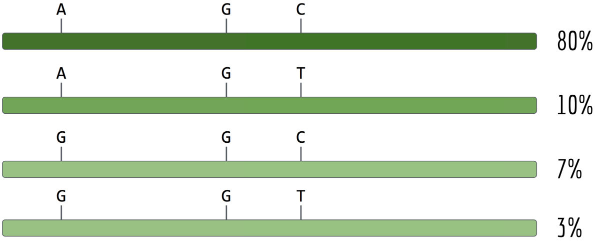
One centimorgan (cM) is the equivalent to a recombination frequency of 0.01 (1%)



In humans, 1 cM corresponds to approximately 1 million bp on average

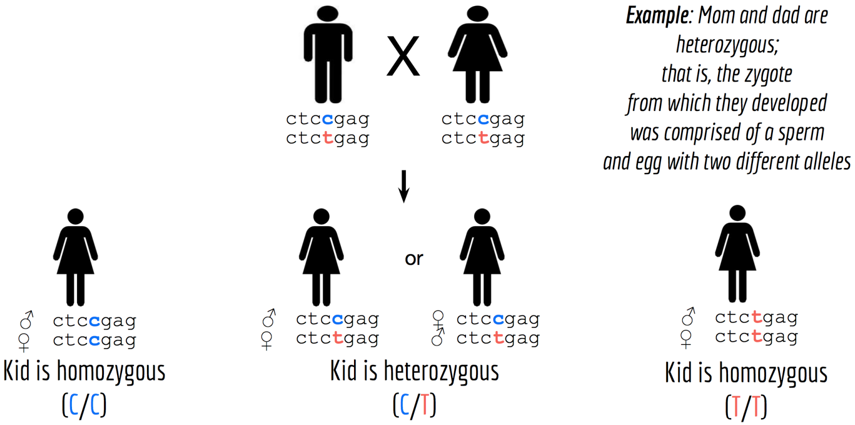
Linkage (dis)equilibrium

- Linkage equilibrium: random association of alleles at different loci
- Linkage disequilibrium: non-random association of alleles at different loci



- Therefore, knowing one allele (e.g., the first A) is a strong predictor of other alleles on a haplotype

Existing (germline) variants are inherited



ctccgag
ctctgag

X

ctccgag
ctctgag

***Example**: Mom and dad are heterozygous; that is, the zygote from which they developed was comprised of a sperm and egg with two different alleles*

♂ ctccgag
♀ ctccgag

Kid is homozygous
(C/C)

♂ ctccgag
♀ ctctgag

or

♀ ctccgag
♂ ctctgag

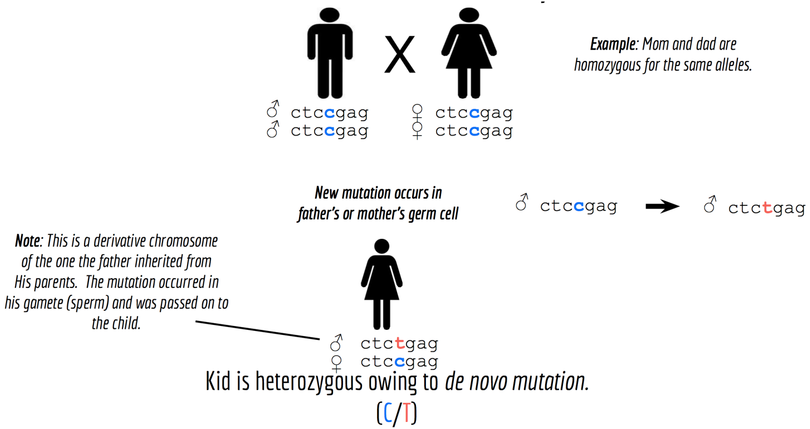
Kid is heterozygous
(C/T)

♂ ctctgag
♀ ctctgag

Kid is homozygous
(T/T)

New (*de novo*) mutations

- May be the cause of many developmental disorders



***Example**: Mom and dad are homozygous for the same alleles.*

♂ ctc**c**gag
♂ ctc**c**gag

X

♀ ctc**c**gag
♀ ctc**c**gag

New mutation occurs in father's or mother's germ cell

♂ ctc**c**gag → ♂ ctc**t**gag

***Note**: This is a derivative chromosome of the one the father inherited from His parents. The mutation occurred in his gamete (sperm) and was passed on to the child.*

♂ ctc**t**gag
♀ ctc**c**gag

Kid is heterozygous owing to *de novo mutation.*

(C/T)

http://massgenomics.org/2015/07/insights-human-de-novo-mutations.html

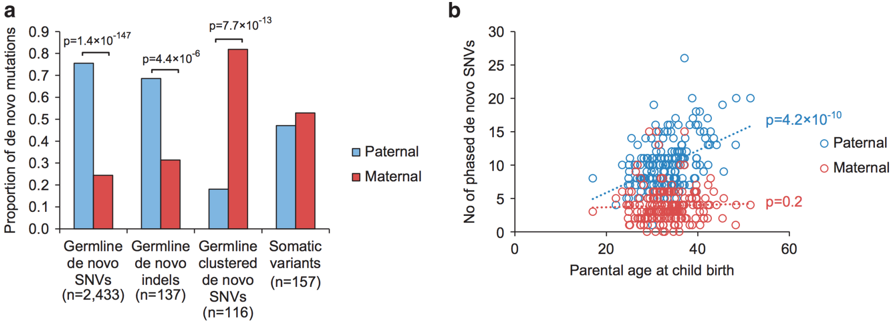
Frequency of *de novo* mutations

- Human mutation rate: $\sim 1.1x10^{-8}$ / bp / generation
- Other estimations: $\sim 2.5x10^{-8}$
- Size of the haploid genome: $\sim 3.1x10^{9}$ nucleotides
- So, $\sim 30 - 40$ *de novo* mutations per haploid genome or twice as many per diploid genome

Roach et al. (2010) Science, http://science.sciencemag.org/content/328/5978/636

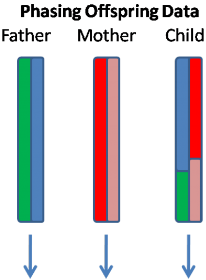
Nachman et al. (2000) Genetics, http://www.genetics.org/content/156/1/297

DNMs are more likely to occur in the paternal germline, and correlate with age



https://www.nature.com/articles/npjgenmed201627, however: Janecka, M, F Rijsdijk, D Rai, A Modabbernia, and A Reichenberg. "Advantageous Developmental Outcomes of Advancing Paternal Age." Translational Psychiatry 7, no. 6 (June 20, 2017): e1156. https://doi.org/10.1038/tp.2017.125. - Older dads have 'geekier' sons

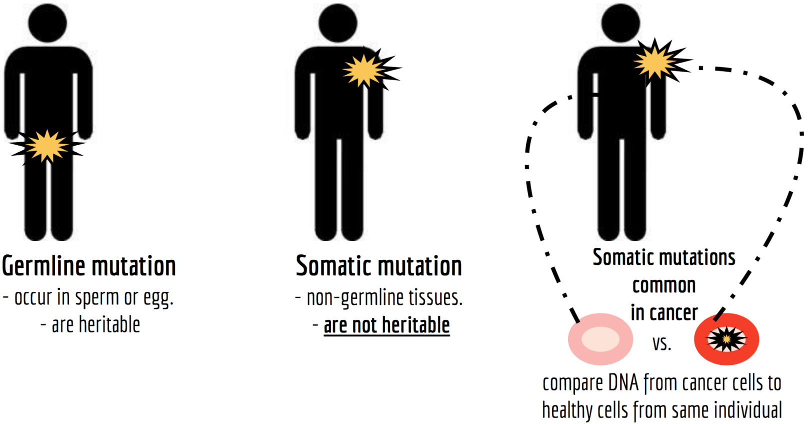
Identifying parental origion of DNMs - phasing



			Transmitted Haplotypes	
Father	Mother	Child	Given by Father	Given by Mother
AB	BB	BB	B	B
AA	AA	AA	A	A
AB	AB	AA	A	A
AB	AA	AB	B	A
BB	BB	BB	B	B
BB	AB	AB	B	A
AB	AB	AB	?	?
AB	BB	BB	B	B
AA	AA	AA	A	A
BB	BB	BB	B	B

http://www.chromosomechronicles.com/2009/09/30/use-family-snp-data-to-phase-your-own-genome/

Somatic mutations are acquired over the lifetime



Germline mutation

- occur in sperm or egg.
- are heritable

Somatic mutation

- non-germline tissues.
- **<u>are not heritable</u>**

Somatic mutations common in cancer

vs.

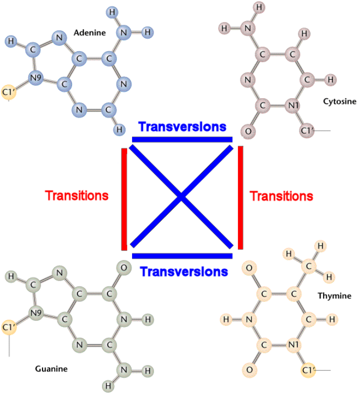
compare DNA from cancer cells to healthy cells from same individual

SNPs are not created equal

- Cytosine is the least stable DNA base. Its half-life is ∼19 days compared to a year or longer for other bases
- The spontaneous deamination of cytosine to uracil can cause polymerases to read the former C as T, making C-G to T-A an unusually common muration in genomes

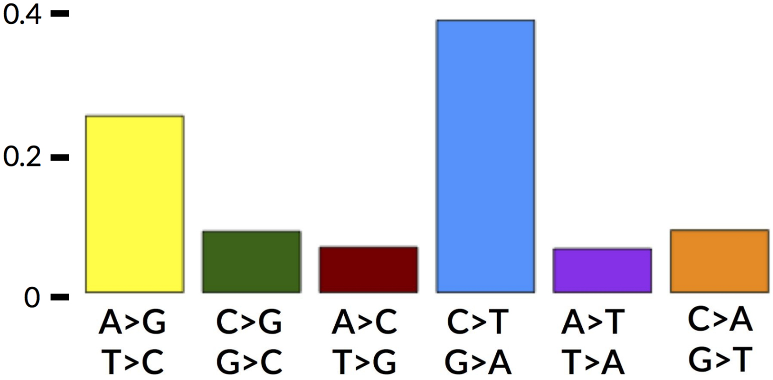
SNPs are not created equal

- Transitions are interchanges of two-ring purines (A <> G) or of one-ring pyrimidines (C <> T): they therefore involve bases of similar shape.
- Transversions are interchanges of purine for pyrimidine bases, which therefore involve exchange of one-ring and two-ring structures.



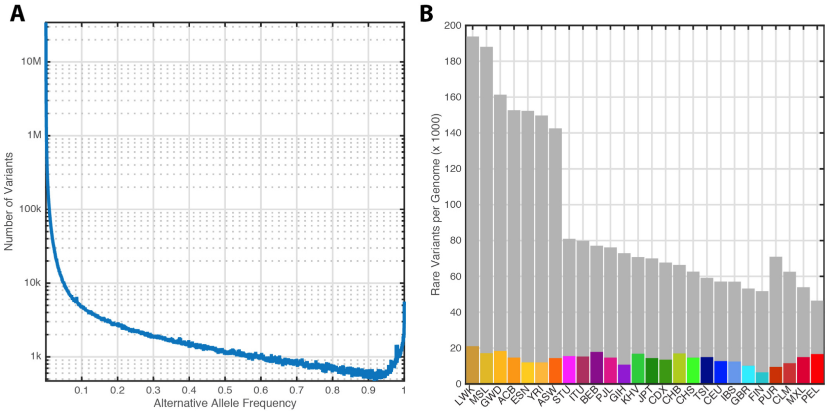
https://www.mun.ca/biology/scarr/Transitions_vs_Transversions.html

SNPs are not created equal



- Due to spontaneous deamination of methylated cytosines, C>T transitions predominate in DNMs

The majority of variants in the data set are rare



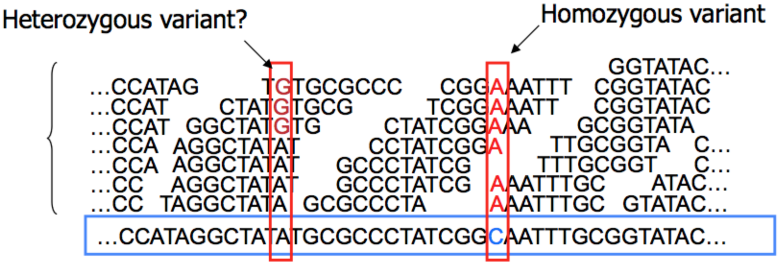
Extended Data Figure 3 | Variant counts. a, The number of variants within the phase 3 sample as a function of alternative allele frequency. **b,** The average number of detected variants per genome with whole-sample allele frequencies <0.5% (grey bars), with the average number of singletons indicated by colours.

- ∼64 million autosomal variants have a frequency <0.5%, ∼ 12 million have a frequency between 0.5% and 5%, and only ∼8 million have a frequency >5%

Distinguishing genomic variants from sequencing errors

Distinguishing SNPs from sequencing error typically a likelihood test of the coverage

- Hardest to distinguish between errors and heterozygous SNP.
- Coverage is the most important factor!
    - Target at least 10x, 30x more reliable


Heterozygous variant?
Homozygous variant
GGTATAC...
...CCATAG TGTGCGCCC CGGAAATTT CGGTATAC
...CCAT CTATGTGCG TCGGAAATT CGGTATAC
...CCAT GGCTATGTG CTATCGGAAA GCGGTATA
...CCA AGGCTATAT CCTATCGGA TTGCGGTA C...
...CCA AGGCTATAT GCCCTATCG TTTGCGGT C...
...CC AGGCTATAT GCCCTATCG AAATTTGC ATAC...
...CC TAGGCTATA GCGCCCTA AAATTTGC GTATAC...
...CCATAGGCTATATGCGCCCTATCGGCAATTTGCGGTATAC...

Exome-Capture Sequencing

Exome-capture reduces the costs of sequencing

- Currently targets around 50Mbp of sequence: all exons plus flanking regions
- WGS currently costs ~\$1500 per sample, while WES currently costs ~\$300 per sample
- Coverage is highly localized around genes, although will get sparse coverage throughout rest of genome

Bamshad et al. Exome sequencing as a tool for Mendelian disease gene discovery (2011) Nature Reviews Genetics. 12, 745-755
https://www.nature.com/nrg/journal/v12/n11/full/nrg3031.html

Defining the exome

- **Exome** - The subset of a genome that is protein coding. In addition to the exome, commercially available capture probes target non-coding exons, sequences flanking exons and microRNAs.
- Initial efforts at exome sequencing erred on the conservative side (for example, by targeting the high-confidence subset of genes identified by the Consensus Coding Sequence (CCDS) Project).
- Commercial kits now target, at a minimum, all of the RefSeq collection and an increasingly large number of hypothetical proteins.

Exome limitations

Limitations

- Knowledge of all truly protein-coding exons is incomplete.
- Efficiency of capture probes varies
- Not all regions sequenced efficiently
- Should other transcripts (e.g., miRNAs) be targeted?
- On average, 82% of genes have at least 90% bases called.