Linkage into a haplotype

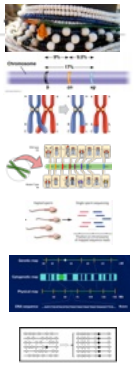
Stretches of DNA can stay together across generations

Pascal Gagneux, PhD

September 26, 2025

Learning objectives

- understand what a **haplotype** is.
- Understand the concept of **genetic linkage**
- Understand **genetic recombination** and its effect on linkage.
- Describe the role of allele co-segregation in **gene mapping**.
- Describe the basic principles/intuitions behind **family** versus **population mapping**.
- Describe the difference between **physical and genetic maps.**
- Understand the important concept of **Linkage Disequilibrium**.



Key Terms:
Physical map; genetic map; genetic recombination; cytogenetics; linkage; linkage disequilibrium; haplotype, haplotype blocks, centimorgan, identity by descent, genetic marker, genotype, beads on a string, two-fold cost of sex, HLA, MHC, SNP, MSAT (simple sequence Repeats)

Lecture in 3 Modules:

Module 1: Haplotypes and Linkage

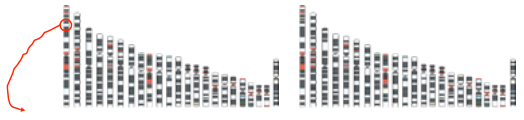
Module 2: Mapping Genes.

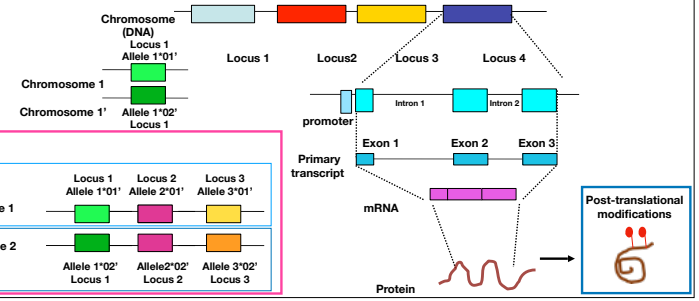
Module 3: Linkage and Linkage Disequilibrium.

Module 1: haplotypes and linkage

Genome
(diploid)

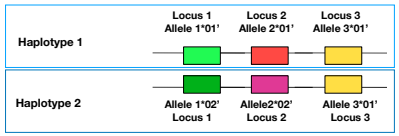
Gene, Locus
Allele





Haplotype

Haplotypes
in a diploid individual



Unique combination
of **linked alleles**

"Beads on a String", recombination can disrupt haplotypes.

The further apart on the string, the more likely to be recombined during meiosis.

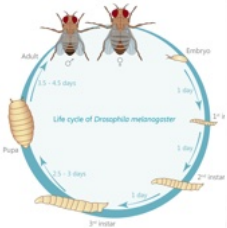
In genetics, the unit for measuring genetic linkage between two different "beads on the string" is the **centimorgan** (cM).

One **centimorgan** is the distance that has a **1% probability of recombination**, it corresponds roughly to 1 million basepairs in humans.

The human genome is ~ 3000 cM long (4460 cM in females / 2590 cM in males).

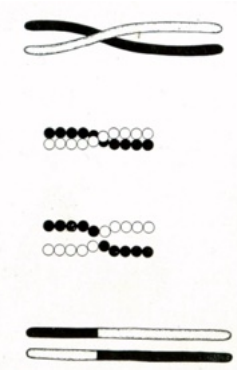
Linked traits in fruit flies

T.H. Morgan
1866-1945



Chose ***Drosophila melanogaster*** (fruti fly) as model animal because of its small size and short (ten day) generation time.

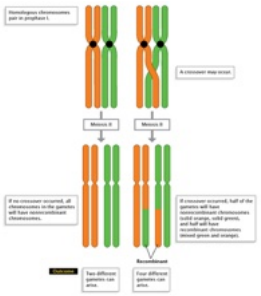
Some genes do not segregate randomly/independently.



Genes are like **beads on a string**, each string = a chromosome, are inherited together, except when chromosomes cross over to rearrange.

Morgan's Chromosomal Theory of Inheritance

T.H. Morgan:



Chromosomal Theory of Inheritance:

Genes are located on chromosomes like beads on a string.

Some genes are physically linked (same chromosome), and thus are (almost) always inherited together (early description of linkage disequilibrium).

"Exceptions" to Mendel's Second Law: Independent assortment

Noticed that "linked" traits would occasionally separate.

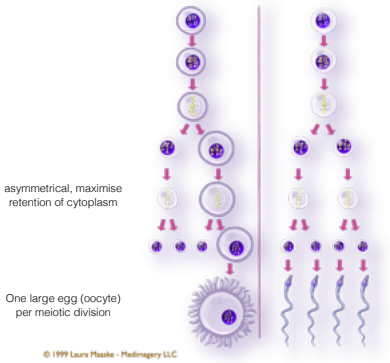
Other traits on the same chromosome showed little detectable linkage.

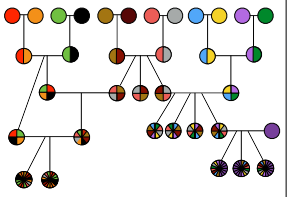
Proposed that recombination, might explain his results.

Proposed that the two paired chromosomes could "cross over" to exchange information during prophase of meiosis, and this produces different combinations of alleles in the gametes.

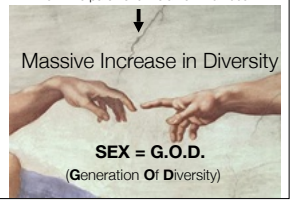
Sex: Meiosis (reduction division, from diploid to haploid)

Mixing of genomes via meiosis and fusion of gametes

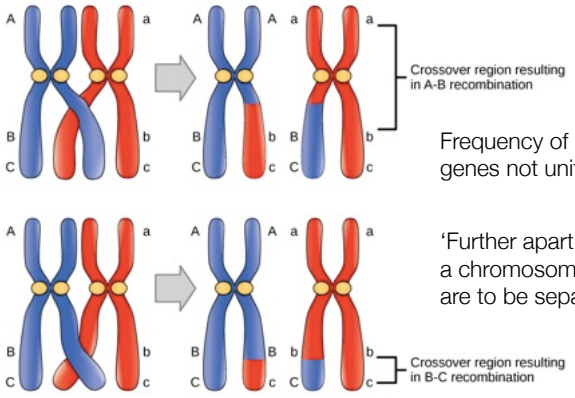




Each generation, half of the reshuffled DNA from two parents forms a new individual



Recombination



Frequency of recombination between genes not uniform.

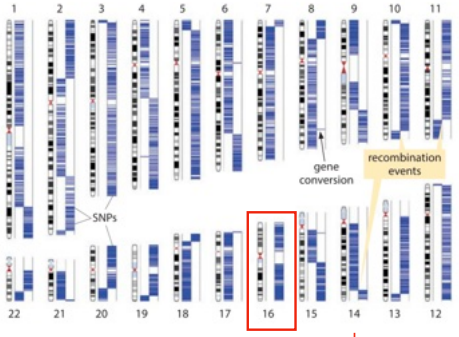
'Further apart two genes are on a chromosome, the more likely they are to be separated by recombination'

https://www.reddit.com/r/askscience/comments/3hq4zl/does_crossover_occur_in_all_4_nonsister/

Single sperm genome analyses:

During each meiosis:
1 to 3 recombinations per chromosome

Average human sperm shows evidence for ~26 recombinations

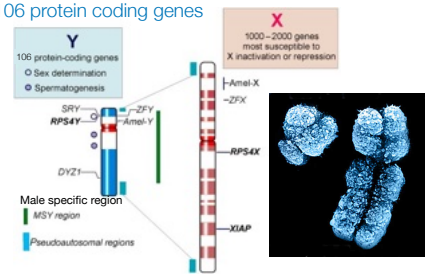


Jianbin Wang, H. Christina Fan, Barry Behr, Stephen R. Quake,
Genome-wide Single-Cell Analysis of Recombination Activity and De Novo Mutation Rates in Human Sperm,
Cell, Volume 150, Issue 2, 2012,Pages 402-412,

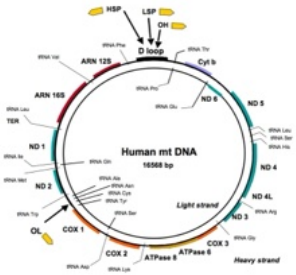
some chromosomes are non-recombined

The human Y-chromosome, mostly one haplotype

lates T2T Y data:
106 protein coding genes

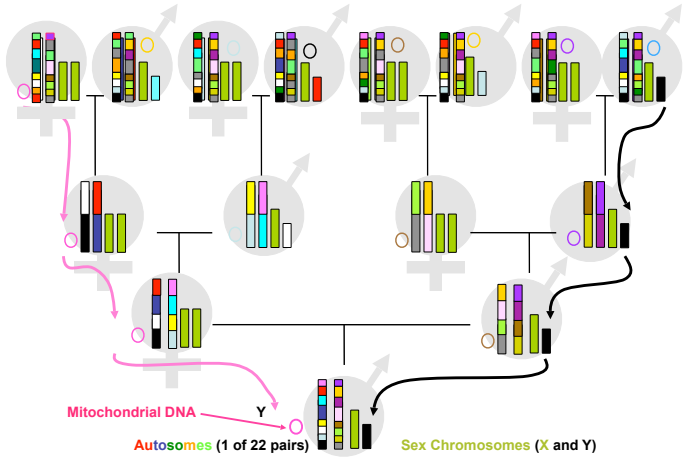


Human **mitochondrial DNA** is also one haplotype:

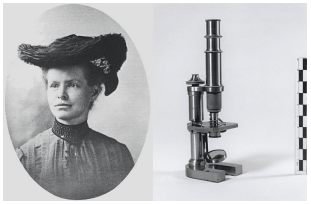


just small caps of the Y chromosome, the pseudoautosomal regions (PAR), recombine with homologous regions on the X!

Modes of inheritance: uniparental & biparental



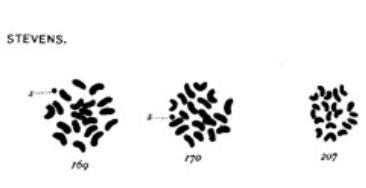
Sex Chromosomes: discovered by Nettie Stevens



Nettie Maria Stevens

She was a trained expert in the modern sense—in the sense in which biology has ceased to be a playground for the amateur and a plaything for the mystic. Her single-mindedness and devotion, combined with keen powers of observation; her thoughtfulness and patience, united to a well-balanced judgment, accounts, in part, for her remarkable accomplishment.

T. H. MORGAN

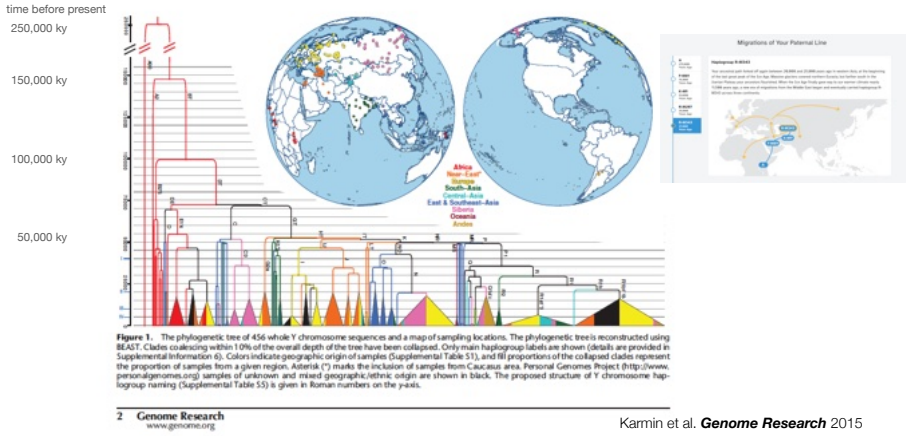


19 large chromosomes + 1 small chromosome in male

20 large chromosomes in female

Plate IV from Stevens (1905) showing hand-drawn micrographs from *Tenebrio molitor* (Meal worm beetle) samples

Human Y-chromosome haplotypes over time and space:



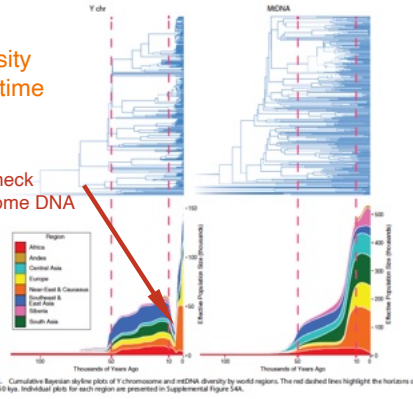
Karmin et al. ***Genome Research*** 2015

Cultural Effects on the Human Gene Pool:

Y-haplotype diversity projected back in time

second bottle neck for Y chromosome DNA

Massive restriction of Y-chromosome diversity: Male clans are causing the extinction of other male clans.



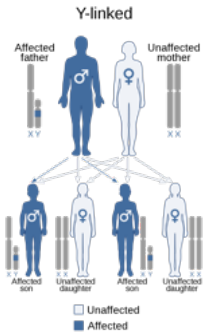
mtDNA haplotype diversity projected back in time

No second bottle neck for mt DNA!

Strong bottleneck in Y-chromosome ~6000 years ago, no such effect on mt DNA!
Male variance in reproductive success with adoption of agriculture and subsequent conflict/wars between paternal sibships?

Karmin et al. ***Genome Research*** 2015

Y-linked genes and disease



The majority of the 106 Y-chromosome genes are located in the "non-recombining region"

They are inherited together as one haplotype!

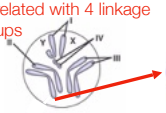
- ASMTY (acetylserotonin methyltransferase), melatonin biosythesis
- TSPY (testis-specific protein), 35 copies exist, spermatogenesis
- IL3RAY (interleukin-3 receptor),
- SRY (sex-determining region),
- ZFY (zinc finger protein),
- PRKY (protein kinase, Y-linked),
- ANT3Y (adenine nucleotide translocator-3 on the Y),
- AZF2 (azoospermia factor 2),
- BPY2 (basic protein on the Y chromosome) Male germ cell development
- AZF1 (azoospermia factor 1),
- DAZ (Spermatogenes is deleted in azoospermia),
- RBM1 (RNA binding motif protein, Y chromosome, family 1, member A1), spermatogenesis
- RBM2 (RNA binding motif protein 2), spermatogenesis
- UTY (ubiquitously transcribed TPR gene on Y chromosome).Minor Histocompatibility Antigen
- USP9Y (spermatogenesis, mutate s assoc with Sertoli cell only syndrome)
- AMELY (biomineralization e.g. tooth enamel)

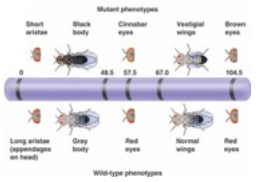
Module 2: Mapping genes

Mohammed Al Idrisi's map of the world, 12th century,
one of the first maps of the Old World
used cities, mountains ranges, lakes & rivers as marks

As a 19 year old undergrad, **Alfred Sturtevant** created the first genetic map of fruit fly genes:
used "factors" responsible for traits as marks…..
The further apart on the chromosome, the more likely to be inherited independently (because more likely to be recombined).

4 pairs of chromosomes correlated with 4 linkage groups





Ethnocentric Maps

Mohammed Al Idrisi's map of the world, 12th century,
one of the first maps of the Old World
used cities, mountains ranges, lakes & rivers as marks

Sturtevant: the First Genetic Map

His key idea:

Realized **frequency of crossing over was related to distance**

Further apart two genes were on a chromosome, the more likely it was that these genes would separate during recombination.

"proportion of crossovers could be used as an index of the distance between any two factors" (Sturtevant, 1913)

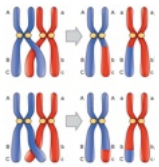
Created **first chromosomal linkage map** for the genes located on the X chromosome of fruit flies (Weiner, 1999)

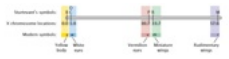
Sturtevant then worked out the order and the linear distances between these linked genes, thus forming a linkage map, with distance in an arbitrary unit he called the "map unit".

Defined the **"map unit": if two positions are 1cM away**,

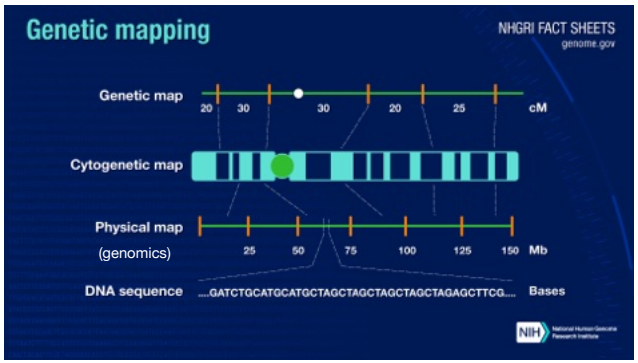
it means they have a **recombination frequency of 0.01 or 1%**.

The map unit was renamed the centimorgan (cM), in honor of Thomas Hunt Morgan.





Mapping Genes through genetic markers



Gene Mapping

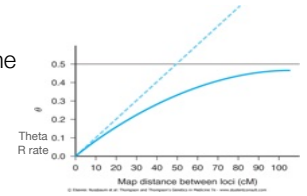
Concepts:

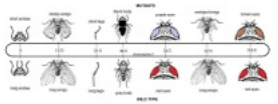
The process of establishing the locations of genes on the chromosomes.

Used to determine the genetic position of a disease-causing gene.

Goal: identifying the approximate location of a disease-causing gene can help researchers to identify the actual gene.

Even if the specific disease-causing gene is not yet identified, knowing its approximate genetic location can facilitate genetic counseling.





Thomas Hunt Morgan's *Drosophila melanogaster* genetic linkage map

Genetic "Markers" = Observable Polymorphic Loci

Useful polymorphic loci have a high likelihood of being different in two individuals, even heterozygous in one individual and are scattered throughout the genome at regular intervals.

Markers are used to differentiate each of the homologs inherited from the parents

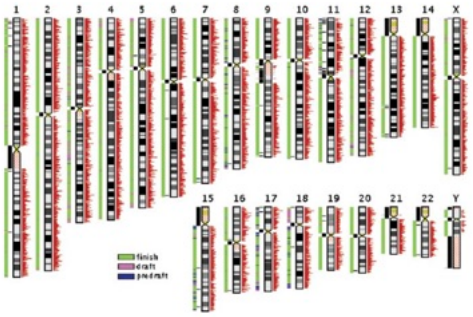
Examples of markers:

DNA sequence polymorphisms identified via sequencing studies. Microsatellites (Simple Sequence Repeats: SSR, Variable Tandem Repeats: VNTR), which are most frequently used for family studies, due to their high degree of polymorphism.

Single Nucleotide Polymorphisms (SNPs), which are most frequently used for population-based studies (a.k.a. SNV single nucleotide variants).

Restriction Fragment Length Polymorphisms (RFLPs), DNA cutting patterns by using bacterial enzymes (restriction enzymes) that cut at specific DNA sequences. The patterns of short DNA segments generated reflects the distribution of recognized sequences cut by each enzyme.

MSAT markers are scattered cross all chromosomes



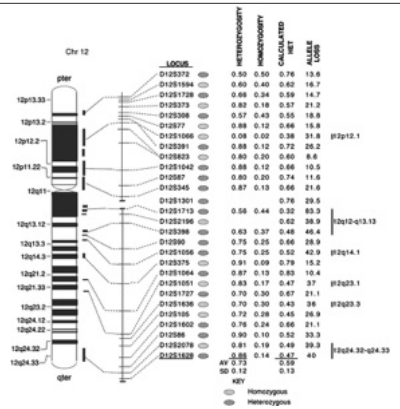
Using highly variable genetic markers to find associations with disease genes.

27,039 polymorphic microsatellite loci across the human genome =1 MSAT locus every 100kb

Genetic markers are highly variable sites found across our genomes, most are in non-coding regions, but many are located near genes of interest.

Tamiya *et al. Whole genome association study of rheumatoid arthritis using 27 039 microsatellites.* ***Human Molecular Genetics,*** *2005, Vol. 14, No. 16 2305–2321*

Mapping Genes through MSATs: example



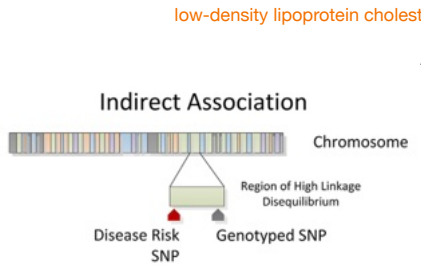
Using highly variable genetic markers to find associations with disease genes.

Mapping of candidate tumor suppressor genes on chromosome 12 in **adenoid cystic carcinoma (ACC)**

Genetic markers are highly variable sites found across our genomes, most are in non-coding regions, but many are located near genes of interest.

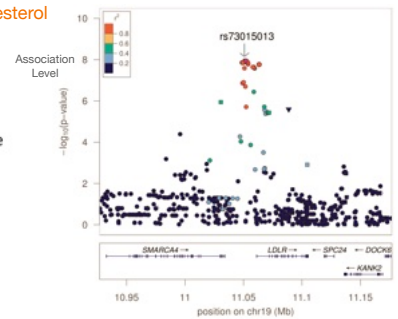
Rutherford, S., Hampton, G., Frierson, H. *et al.* Mapping of candidate tumor suppressor genes on chromosome 12 in adenoid cystic carcinoma. ***Lab Invest*** 85**,** 1076–1085 (2005).

Indirect Association of a SNP with a heritable trait



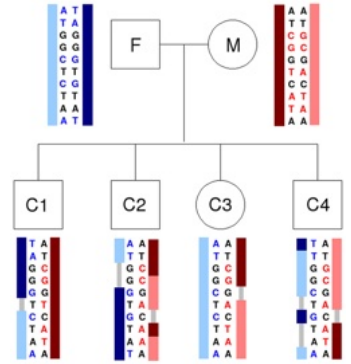
Bush, William & Moore, Jason. (2012). Chapter 11: Genome-Wide Association Studies. **PLoS computational biology.** 8. e1002822. 10.1371/journal.pcbi.1002822.

low-density lipoprotein cholesterol



Sanna S, Li B, Mulas A, *et al*. Fine mapping of five loci associated with low-density lipoprotein cholesterol detects variants that double the explained heritability. ***PLoS Genet.*** 2011;7(7):e1002198. doi:10.1371/journal.pgen.1002198

Identification of recombination events.



Chowdhury R, ***et al.***(2009) Genetic Analysis of Variation in Human Meiotic Recombination. ***PLOS Genetics*** 5(9): e1000648.

Unit of recombination probability = length of genome

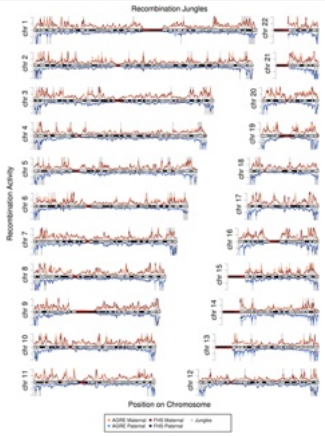
Genes on different chromosomes have a 50% chance of being inherited together (Theta, their recombination rate is 0.5).

Genes located at the very ends of the same chromosome also have a near 50% chance of being inherited together.

1 cM : 1% chance of recombination or Theta = 0.01

Size of the human genome in centimorgans:
The human genome contains about 3300 centimorgan (cM)
~350 evenly spaced markers are included in a typical genome-wide screen to obtain a coverage rate of 1 marker per 10 cM.

Recombination Jungles across the Human Genome

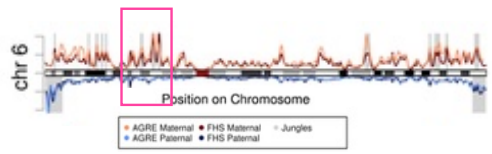


Recombination events are not distributed evenly across the human genome.

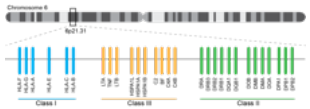
Genomic regions with higher recombination counts are referred to as “recombination jungles” or hots spots (hot spots used for very short segments).

Chowdhury R, ***et al.***(2009) Genetic Analysis of Variation in Human Meiotic Recombination. ***PLOS Genetics*** 5(9): e1000648.

Recombination Jungles: HLA region on Chromosome 6

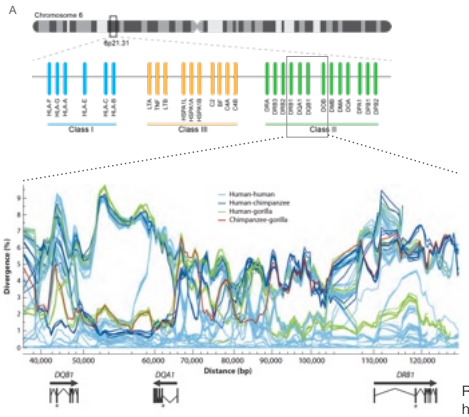


HLA is a “super locus”, >150 genes,
many of them with key roles in immunity

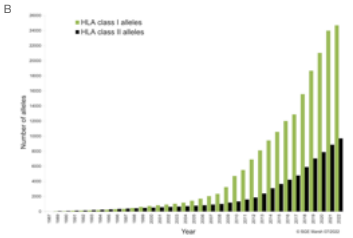


Chowdhury R, ***et al.***(2009) Genetic Analysis of Variation in Human Meiotic Recombination. ***PLOS Genetics*** 5(9): e1000648.

HLA haplotypes: Very diverse, Massively Shuffled & Retained

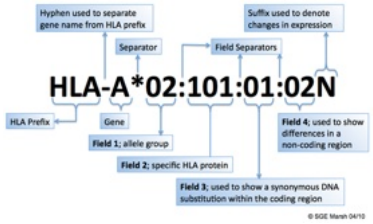


Great clinical Importance:
transplantation medicine
disease risk:
autoimmune e.g. celiac. RA, MS
infectious: HIV-AIDS, Covid 19, TB etc..



Raymond CK, Kas A, Paddock M, Qiu R, Zhou Y, et al. 2005. Ancient haplotypes of the HLA Class II region. **Genome Res**. 15:1250–57

HLA haplotypes: e.g.

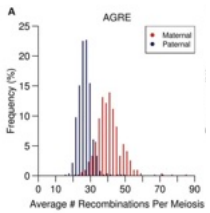


Nomenclature	Indicates
HLA	the HLA region and prefix for an HLA gene
HLA-DRB1	a particular HLA locus i.e. *DRB1*
*HLA-DRB1*13*	a group of alleles that encode the DR13 antigen or sequence homology to other *DRB1*13* alleles
*HLA-DRB1*13:01*	a specific HLA allele
*HLA-DRB1*13:01:02*	an allele that differs by a synonymous mutation from *DRB1*13:01:01*
*HLA-DRB1*13:01:01:02*	an allele which contains a mutation outside the coding region from *DRB1*13:01:01:01*
*HLA-A*24:09N*	a 'Null' allele - an allele that is not expressed
*HLA-A*30:14L*	an allele encoding a protein with significantly reduced or 'Low' cell surface expression
*HLA-A*24:02:01:02L*	an allele encoding a protein with significantly reduced or 'Low' cell surface expression, where the mutation is found outside the coding region
*HLA-B*44:02:01:02S*	an allele encoding a protein which is expressed as a 'Secreted' molecule only
*HLA-A*32:11Q*	an allele which has a mutation that has previously been shown to have a significant effect on cell surface expression, but where this has not been confirmed and its expression remains 'Questionable'

HLA haplotype provide much important information, from disease risk, to reproductive compatibility, to tissue transplantation.

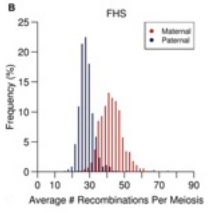
More recombinations in females than males!

Autism Genetic Research Exchange (AGRE) Database



For our analysis, we used the genotype data from members of two-generation families that have two or more children to infer recombination phenotypes of the parents in these families. The 511 AGRE families have an average of 2.26 children (median=2; range: 2 to 7) and provided data for 1,155 female and 1,155 male meioses. Using ~400,000 SNP genotypes of the parents and children in these families, we inferred the recombination phenotypes of 511 mothers and 511 fathers.

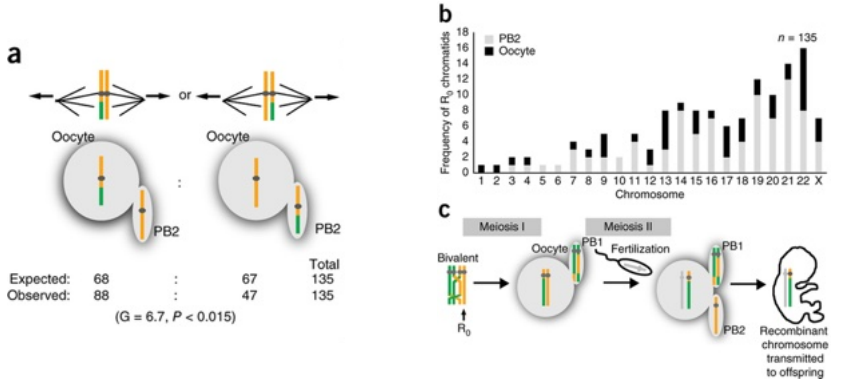
Framingham Heart Study (FHS) Database



For the second population, we analyzed genotypes for ~500,000 SNP markers from members of 784 two-generation families from the FHS. This dataset provided us with recombination phenotypes for 654 mothers and 639 fathers, with an average of 2.86 children per individual (median=3; range: 2 to 9).

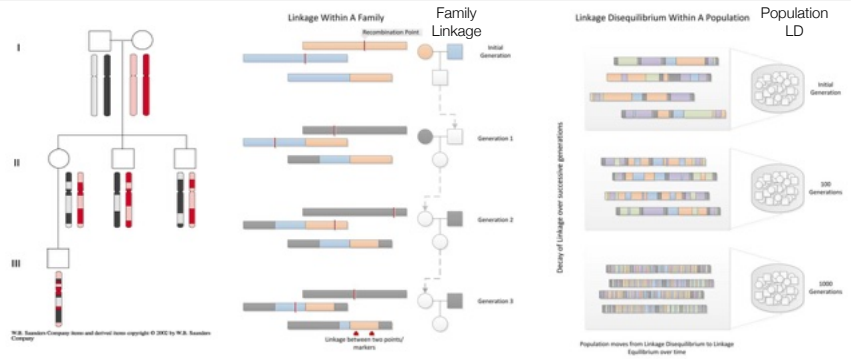
Chowdhury R, ***et al.***(2009) Genetic Analysis of Variation in Human Meiotic Recombination. ***PLOS Genetics*** 5(9): e1000648.

The Ovum favors recombined chromatids!



Ottolini *et al.* ***Nature Genetics 2015***

Module 3: Linkage and Linkage Disequilibrium



Bush, William & Moore, Jason. (2012). Chapter 11: Genome-Wide Association Studies. ***PLoS computational biology***. 8. e1002822. 10.1371/journal.pcbi.1002822.

If two loci are 10 cM away from each other, what is the probability of recombination at meiosis?

Answer: 10%

Completion of the Human Genome Project : What Next?

A focus on variation in human genome sequence and its role in phenotypic expression:

The HapMap & 1000 Genomes Project
Catalogue of human variation and genotype data
(https://www.internationalgenome.org/)

The Cancer Genome Atlas (TCGA)
Molecularly characterized over 20K primary cancer and matched normal samples spanning 33 cancer types (assess the role of DNA sequence, gene expression, and protein expression and structure on cancer)
(http://cancergenome.nih.gov/)

The ENCODE Project (ENCyclopedia Of DNA Elements)
Seeks to interpret DNA sequence, focus on non-coding DNA
(https://www.encodeproject.org)

Human Pangenome Reference Consortium
Genomes from individuals from diverse populations in order to better represent the genomic landscape of diverse human populations.
https://humanpangenome.org

These projects raise questions about how genetic variations behave in the population, as well what they do (physiologically) within an individual.

Human Linkage Analysis

RFLP Markers for Linkage (1980)

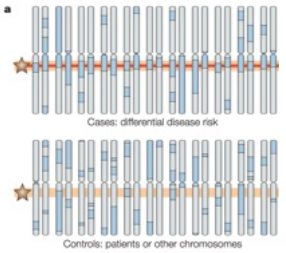
Huntington's Disease Linkage (1983)
Cystic Fibrosis Linkage (1985)

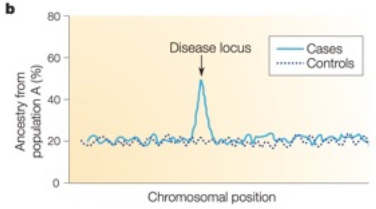
Cystic Fibrosis Gene (1989)
Huntington's Disease Gene (1993)

BRCA1 (1994)

Detecting disease-associated genomic regions using mapping by admixture linkage disequilibrium (MALD)

Using differences between recently admixed populations to detect disease associated loci (still embedded in long distinct haplotype of only one parent population.

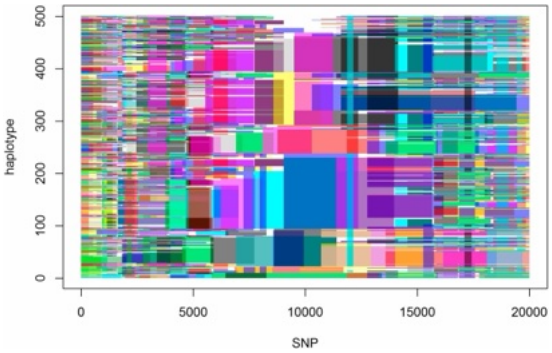


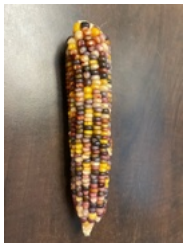


population analysis

Smith, M., O'Brien, S. Mapping by admixture linkage disequilibrium: advances, limitations and guidelines. ***Nat Rev Genet*** 6, 623–632 (2005)

Identifying Haplotype Blocks on a chromosome: e.g. #Chr1



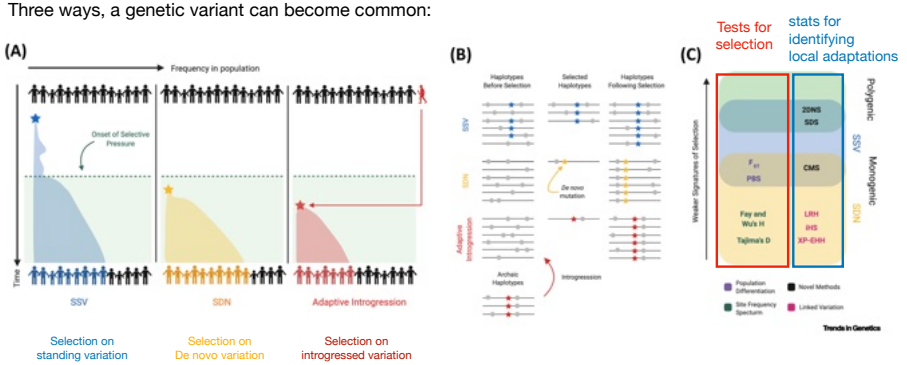


of maize!

HaploBlocker: Creation of Subgroup-Specific Haplotype Blocks and Libraries. Torsten Pook, T. ***et al*** ***GENETICS*** August 1, 2019 vol. 212 no. 4 1045-1061

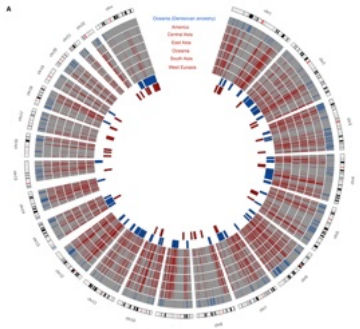
Characterisation of Selection on Standing Variation, Selection on *de Novo* Mutation and Adaptive Introgression.

Three ways, a genetic variant can become common:



Rees *et al.* ***Trends in Genetics 2020***

Mapping genes on the genome



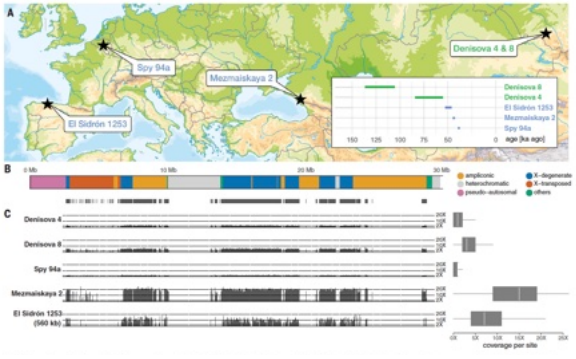
Neanderthal girl paleo-reconstruction,
(Kennis and Kennis)

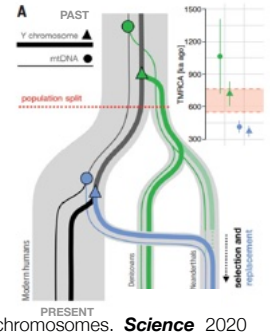
Non-overlapping 100 kb windows with inferred archaic ancestry in each of six populations (blue, Denisovans; red, Neanderthal).

In the innermost rings plots “gene deserts” (windows >10 Mb).

Sankararaman *et al.*, 2016, ***Current Biology*** 26, 1241–1247
The Combined Landscape of Denisovan and Neanderthal Ancestry in Present-Day Humans

How Neanderthals lost their Y



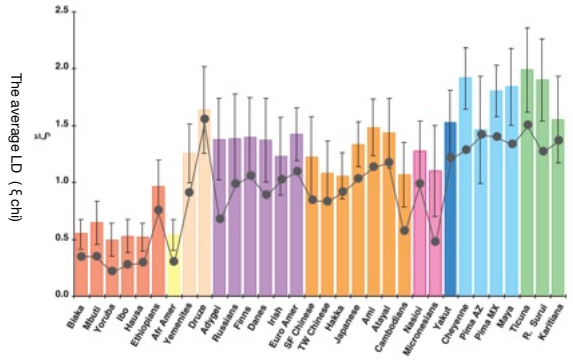


Petr, M *et al.* The evolutionary history of Neanderthal and Denisovan Y chromosomes. ***Science*** 2020

The average LD for 83 SNPs across 21 haplotypes for 32 populations

based on published data on CD4, DM1, DRD2, DRD4, PAH and COMT plus unpublished data.

Increase in LD with distance from Africa provides one of the strongest evidence of our common African ancestry.



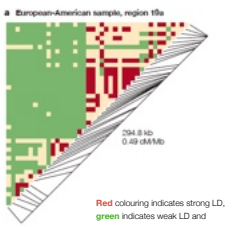
Tishkoff & Kidd VOLUME 36 | NUMBER 11 | NOVEMBER 2004 **NATURE GENETICS SUPPLEMENT**

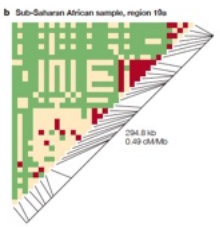
Haplotype Blocks in Human Populations

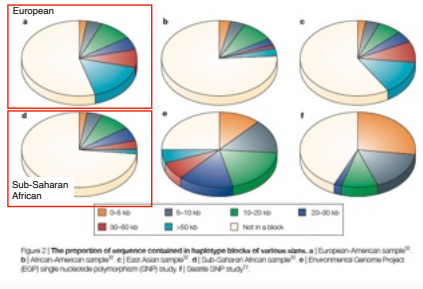
Europe

Sub-Saharan Africa

Many more long blocks outside Africa







Wall and Pritchard ***Nature Genetics*** 2003

Planetary Take over: timing reflected in LD patterns

stone tools in Mexican cave
Chiquihuite ~30ky

human foot prints in NM
~23ky

Human migration: Climate and the peopling of the world , Peter B. deMenocal & Chris Stringer
Nature 538, 49–50 (06 October 2016) doi:10.1038/nature19471

Two Contrasting Truths

Human populations are remarkably genetically similar

There also exists real genetic difference between them including gradient in Linkage disequilibrium.

Timbuktu to Tokyo

Timbuktu, Mali

Fes, Morocco

Algeciras, Spain

Toulouse, France

Istanbul, Turkey

Almati, Kazakhstan

Xining 西宁, China

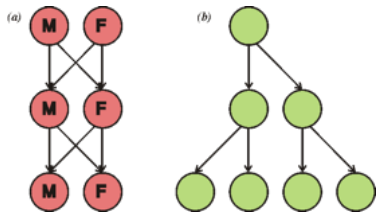
Tokyo 東京, Japan

Where does one “race” stop and another start??

G.O.D. is costly!

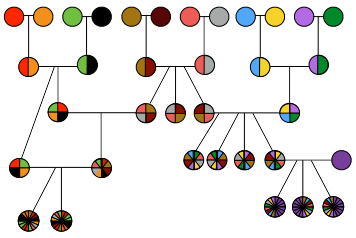
1. two-fold cost of sex

sex no sex



2. “shredding” winning combos

constant loss of potentially
excellent combinations!



Summary

Genes and regulatory regions are **like beads on a string**, along different chromosomes.

Fruit fly genetics (just 4 chromosomes) revealed **linkage and recombination via crossing over.**

Recombination frequency can be measured and **correlates with position along chromosome.**

Recombination frequency can be **directly observed via single sperm sequencing.**

Recombination rates are **higher in females** than males.

Ellie Stevens discovered the basis for sex determination by **sex chromosomes**.

Y-chromosomes (mostly) and mt DNA are haplotypes

Y-halpotypes studies can reconstruct global spread of humans and the effect of empires and conquest (male violence).

Genetic maps can be built by **breeding experiments** of lines with **distinct traits** (eye color, wing shape, bristle number)**.**

Genetic maps can be built by the use of **variable genetic markers**: MSATs, SNPs, SV, RFLP.

Physical (genomic) maps can be established by **whole genome sequencing** (genomics)

Linkage disequilibrium (LD) can indicate recent establishment, admixture and/or selection.

Haplotype blocks are shorter in Africa than outside, there is **more LD further from Africa.**

Sex is costly due to the **two-fold cost of having males,** and the **loss of perfectly adapted combinations** and:

Recombination is mutagenic, costly but apparently worth it!

The Human **HLA region isthe most variable part of our genomes and** contains famous **and infamous haplotypes.**

Thank You!