SISG Module 3: Molecular Genetics and Genomics

19th Summer Institute in Statistical Genetics

UNIVERSITY of WASHINGTON

(This page left intentionally blank.)

Summer Institute in Statistical Genetics

2013

Molecular Genetics and Genomics 1
Genes and Inheritance

ggibson.gt@gmail.com
http://www.cig.gatech.edu

Course Outline

1. Genes and Inheritance (GG)
2. Molecular Biology of the Genome (JA)
3. Population and Evolutionary Genomics (JA)
4. Gene Expression (JA) / Association Mapping (GG)
5. Personalized Genomics (GG)

Genotype and Phenotype

The genotype of an organism is the sequence of it's genes.

The phenotype of an organism the way it appears.

In general, genes are not deterministic. Genotypic variation among organisms specifies the information that, in combination with the environment, influences the phenotype.

Pleiotropy refers to the ability of single genes to influence multiple phenotypes.

Penetrance is the proportion of individuals with a genotype who have the phenotype / disease.

Expressivity is the degree of the phenotype of in affected individuals.

Mendelian Genetics

$$\frac{A}{A} \times \frac{a}{a}$$ F0: Pure breeding parents

$$\Downarrow$$

$$\frac{A}{a} \times \frac{A}{a}$$ F1: Heterozygous offspring

$$\Downarrow$$

$$\frac{A}{A} ; \frac{A}{a} ; \frac{a}{A} ; \frac{a}{a}$$ F2: Mendelain proportions of Homozygotes + Heterozygotes

3 Models of Complex Disease

CDCV: Common Disease / Common Variant

The proposition that most disease susceptibility can be attributed to 10 to 20 loci, each of which explain around 5% of disease risk.

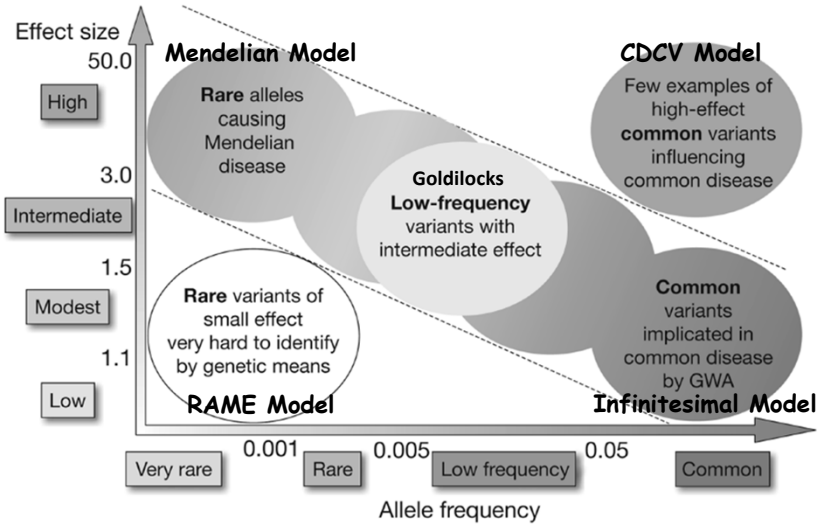
RAME: Rare alleles of Major Effect

The proposition that diseases are highly heterogeneous, with hundreds or thousands of rare mutations causing individual cases of disease.

Infinitesimal:

The proposition that we all carry thousands of very weak susceptibility alleles, and those unlucky enough to have too many are at highest risk, where rare variants or environmental triggers push us over the edge.

Models of the Genetics of Complex Traits



Manolio et al (2009) *Nature* **461**: 747-753

Heritability

Heritability is the proportion of variance in a population that can be attributed to genotypic differences

$$h^2 = V_G/V_P \text{ where } V_P = V_G + V_E$$

The phenotypes may be discrete, such as disease status; categorical, such as number of digits; or continuous, such as height or a biochemical measure.

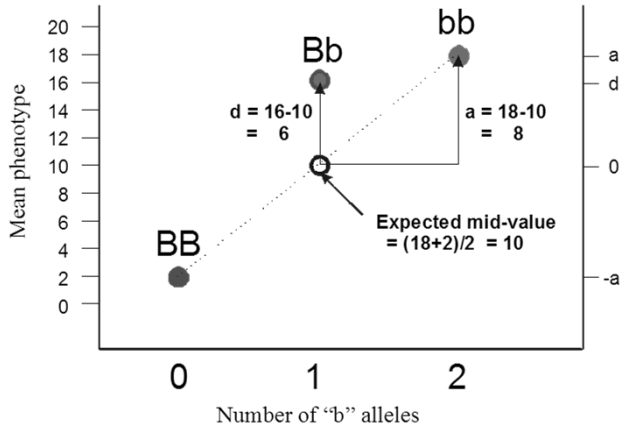
1. Heritability is not a statement about individuals.
A heritability of 50% for diabetes does not imply that half the reason why someone is diabetic is genetic, the other half environmental. Rather, it suggests that there would be half as much diabetes in the population if everyone was genetically identical.

2. Heritability is only a statement about a single population.
A heritability of 80% for height does not imply that most of the average difference in height between populations is due to genetic differences. Heritability estimates alone should not be used to draw inferences about genetic divergence between groups.

3. Heritability is not the same as inheritance.
Inheritance is the correspondence between children and their biological parents. It can be due to environmental, including cultural, factors that are shared by family members, or to effects. The only way to confidently interpret heritability is to actually measure the genotypic contribution.

4. Very low heritability does not imply very little genetic contribution.
It may either be due to relatively high environmental variance (hence, a large demominator V_P), or to an absence of variance in the genes that contribute. Many important genes, including drug targets, are not polymorphic and will only be discovered through other types of approach including model organism research.

Dominance ratio



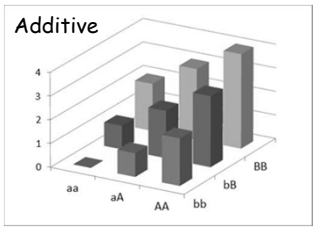
$$V_P = V_A + V_D + V_I + V_{GxE} + V_E$$

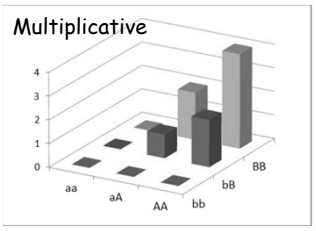
- Loci are said to have *Additive* effects if the contributions of each individual allele can simply be added algebraically to arrive at a prediction of a phenotype given a genotype.
- *Dominance* refers to the observation that heterozygotes resemble one class of homozygotes more than the other.
- *Epistasis* refers to a locus-by-locus Interaction, such as when alleles at two loci antagonize or synergize with one another.
- V_E is the *environmental* variance

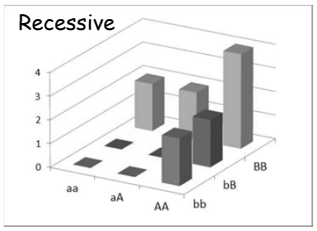
Broad Sense Heritability

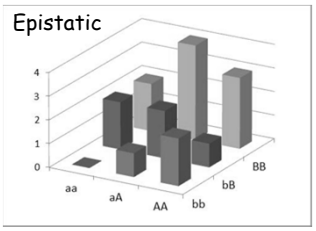
$$V_G = V_A + V_D + V_I + V_{G\times E}$$

Narrow sense heritability is only the additive component whereas Broad sense heritability includes dominance, interaction and genotype-by-environment effects.

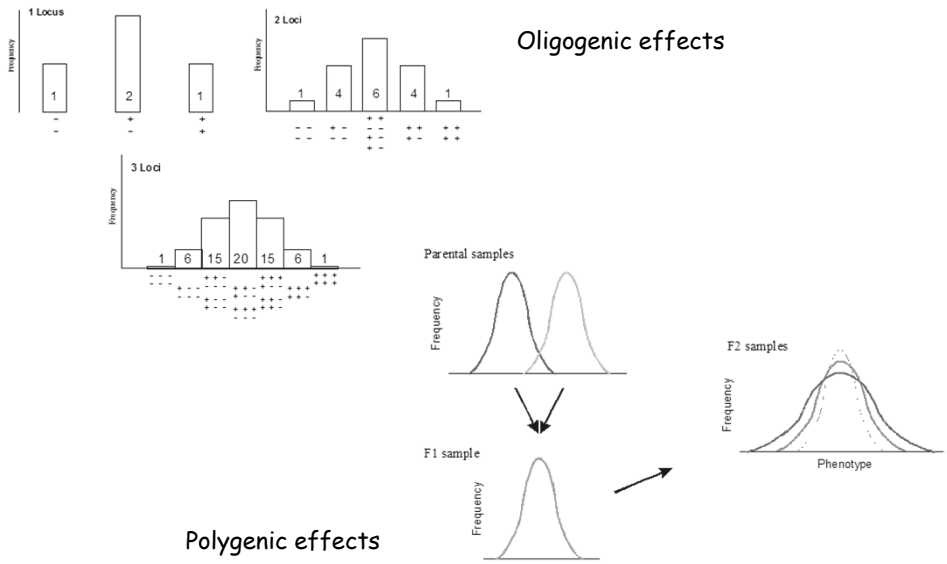






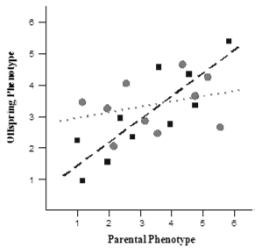


From Mendelian to Quantitative genetics

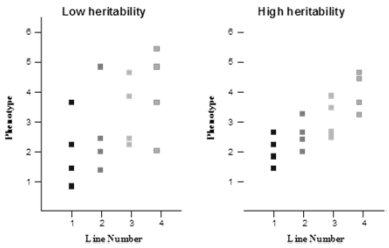


Estimating Heritability

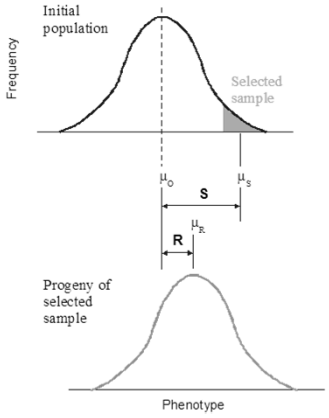
Parent-Offspring Regression



Variance among lines

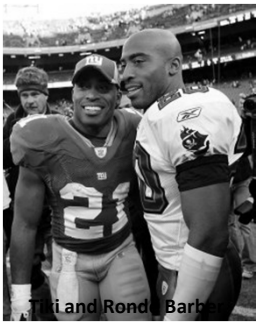


Realized Heritability



Twin Studies

Identical / Maternal



$r_{mz} = A + C$

Dizygotic / Fraternal

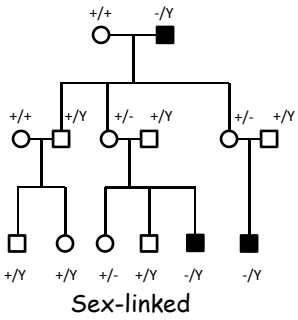
$r_{dz} = ½A + C$

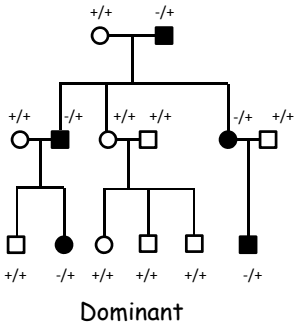
A = Additive Genetic component; C = Common Environment (smaller if reared apart)

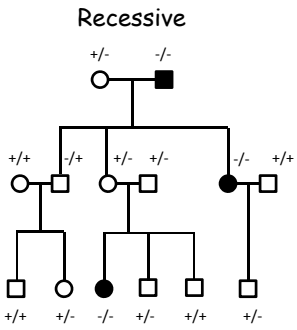
E = unique environment $= 1 - r_{mz}$

r_{dz} should be greater than r_{sib} since C is larger where the womb/upbringing is shared

Mendelian Pedigree Studies







Monogenic Disorders

Approximately 1 in 3,700 Americans have Cystic Fibrosis
Assuming $p^2 = 0.00027$, then $p = 0.016$, the mutant allele frequency

That is, 1 in 30 people are carriers (which is 120 times as many people as have CF),
that is, 3% of Caucasians are carriers, and less than 0.03% sufferers.
It is very likely that someone in this class is a carrier of a CF mutation

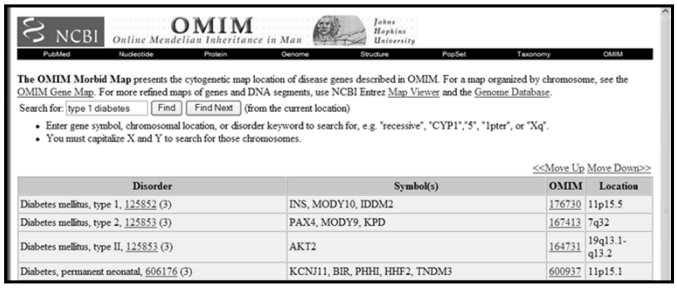
A CF carrier has a 1 in 30 chance of marrying another carrier by chance, and
1 in 4 of their children will be expected to have CF, and only half of all 2-child families will have an affected child

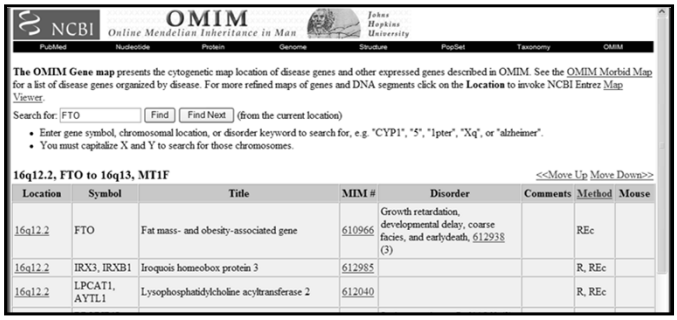
There are hundreds of similar conditions (rare recessives with p ~ 0.01), so we are all carriers for multiple Mendelian disease genes. Collectively, as many as 1 in 25 couples should expect to be dual carriers for a recessive Mendelian disorder, corresponding to an approximate 1% affected rate in all children

Around 1 in 400 children have an inherited Inborn Error of Metabolism, namely an enzyme deficiency affecting amino acid, lipid or other organic molecule biosynthesis. For example:

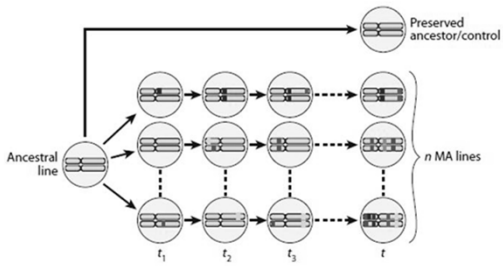
- Phenylketonuria	1/15,000	mental retardation syndrome
- Galactosemia	1/40,000	liver dysfunction and cataracts
- Gaucher's Disease	1/60,000	facial dysmorphology, liver disease
- Zellweger Syndrome	1/50,000	seizures, low muscle tone
- Lesch-Nyhan Syndrome	1/380,000	self-inflicted injury, gout / kidney disease

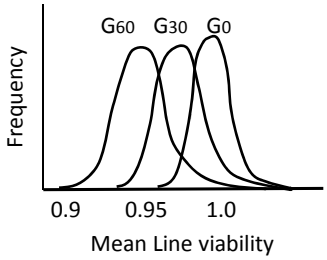
Online Mendelian Inheritance in Man (OMIM)





Mutation Accumulation

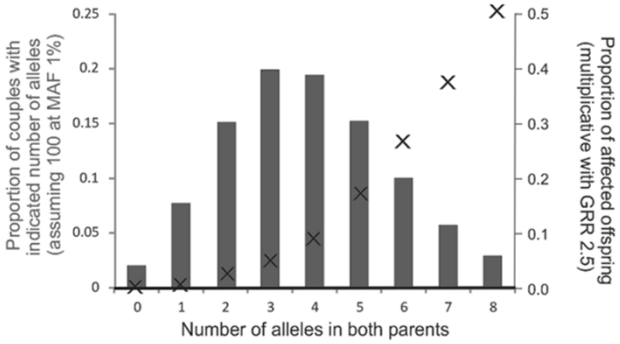




In each generation, slightly deleterious mutations add ~ 0.1% of the standing environmental variance to the heritability of traits, also reducing viability.

Halligan and Keightley (2009) *Annu Rev Ecol Evol Syst.* **40**: 151-172

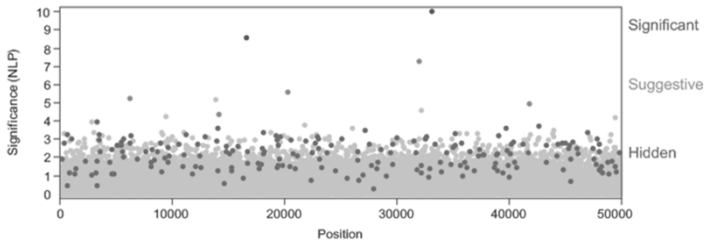
Multiplicative Rare Alleles of Major Effect



Assume there are 100 mutations at 1% frequency, each of which increases the risk of disease 2.5-fold over a baseline environmental risk of 1%.
Whence 0 alleles have a risk of 1%, 1 of 2.5%, 2 of 6%, 3 of 15%, 4 of 39%, 5 or more is highly penetrant.

Then compute the number of alleles per person, hence per couple under random mating, per child, and finally the concordance of siblings within families.

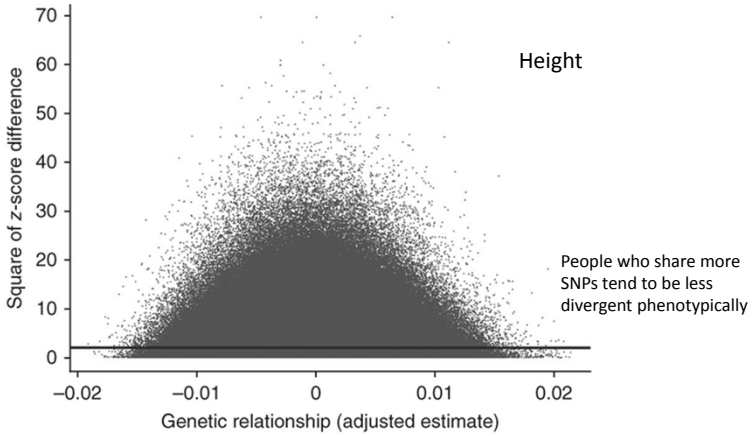
The Missing Heritability Problem



The missing heritability problem is that variants discovered by GWAS only explain a minor fraction of the expected heritability. This may be because:

- The effect sizes are much smaller than previously thought (GRR 1.1 rather than 2)
- Narrow sense heritability has been over-estimated in pedigree studies
- It is rare, not common, variants, that contribute most of the variation
- Epigenetic inheritance accounts for much of the resemblance among relatives
- Broad sense heritability is prevalent, but hard to detect
- Genotyping chips do not tag causal variants effectively enough

SNP-based heritability



Squared Phenotypic difference = 1.98 – 1.01*(Genetic distance)

=> Heritability explained by SNPs (genetic relatedness) is ~ 51%

Yang, Visscher et al (2010) *Nature Genetics* **42**: 565-569

Molecular Biology of the Genome

Joshua Akey
Department of Genome Sciences
akeyj@uw.edu

1

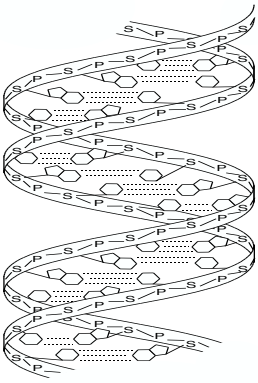
Outline

- Information Flow in Genomics
- Gene Structure
- Genetic Linkage
- Chromatin Structure
- Genome Sequencing

2

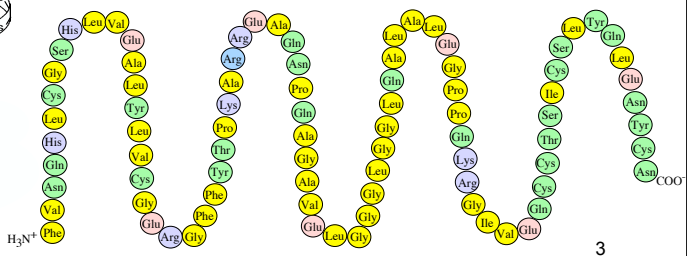
DNA and the Flow of Information

The genetic material: DNA
- Four kinds of subunits (bases A, C, G, T)



A coding problem

Activities within the cell performed by proteins
- Twenty kinds of subunits (amino acids)

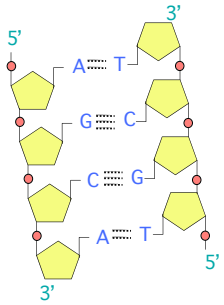


3

The "Central Dogma" of Molecular Biology

4

DNA Structure



- Information content is in the sequence of bases along a DNA molecule

  rules of base pairing → each strand of the double helix has all the info needed to recreate the other strand

- Redundancy in the code

  multiple ways that DNA can specify a single amino acid

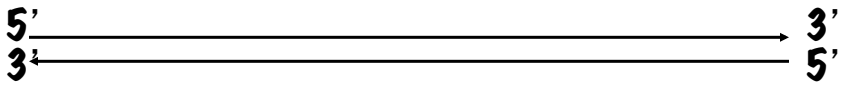
- Genetic variation — differences in the base sequence between different individuals

  why individuals vary in their phenotypes

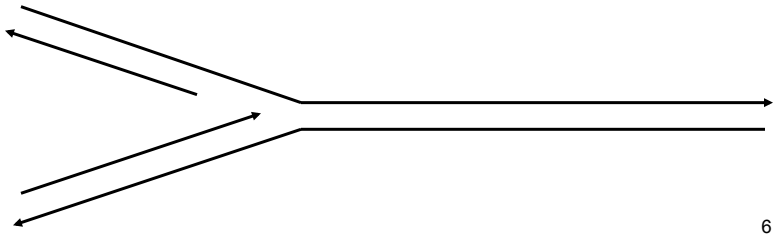
5

Central Dogma: DNA Replication

DNA structure: polarity and base pairing



DNA replication: what's the point?

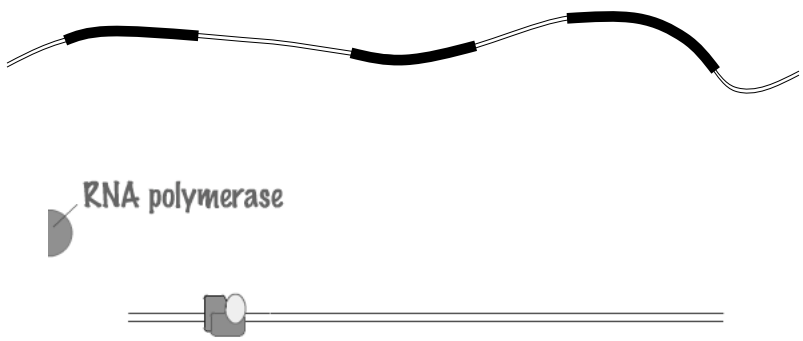


6

Central Dogma: Transcription

Genes — specific segments along the chromosomal DNA that code for some function

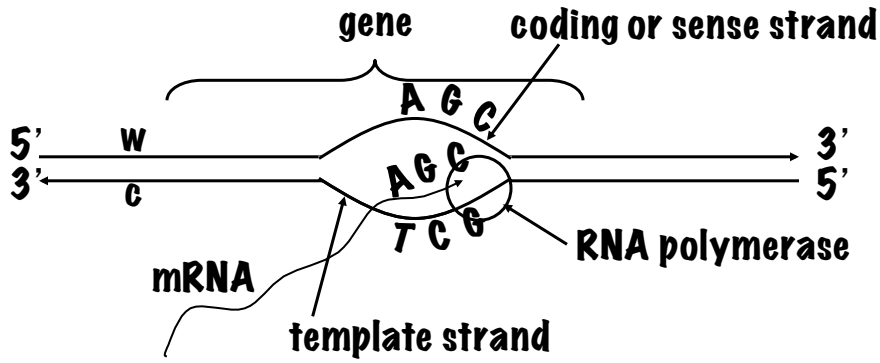
Transcription: "copy" gene into RNA (to make a specific protein)



7

Transcription

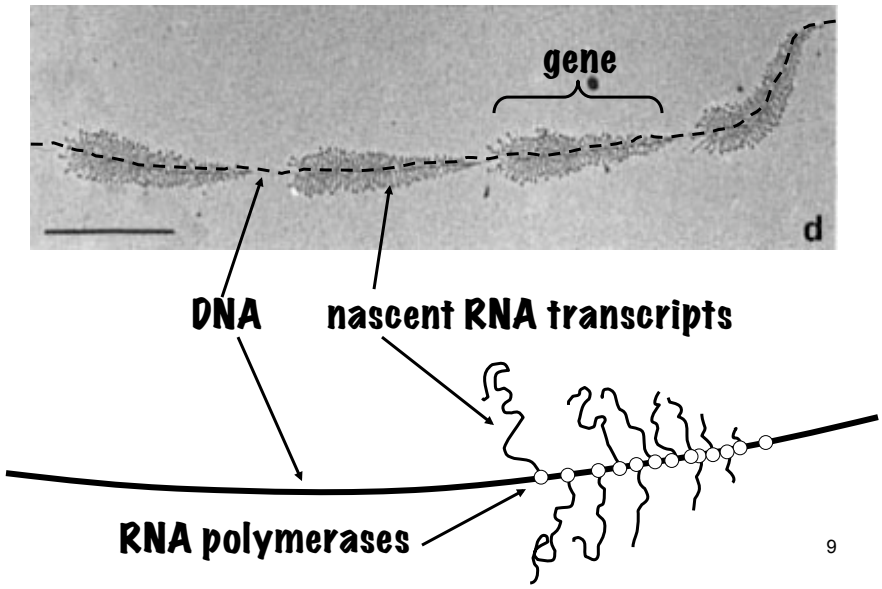
Transcription: "copy" gene into RNA to make a specific protein



Where's the 5' end of the gene? of the mRNA?
Which way is RNA polymerase moving?

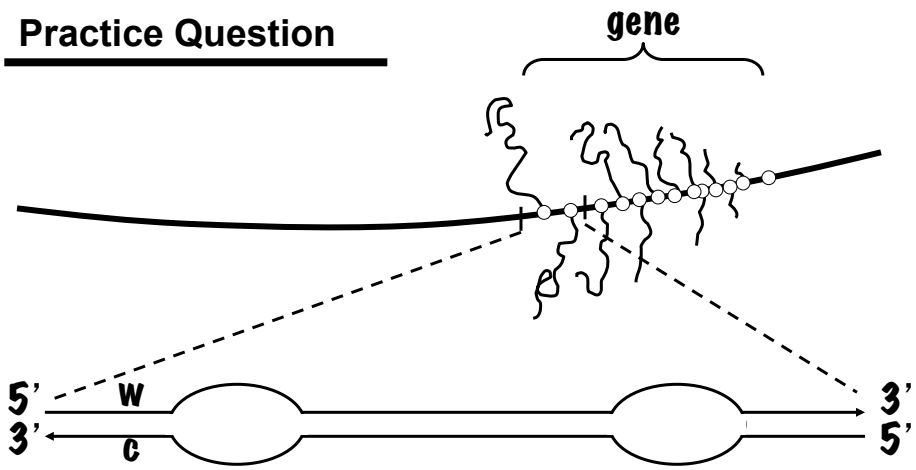
8

Transcription in vivo



9

Practice Question



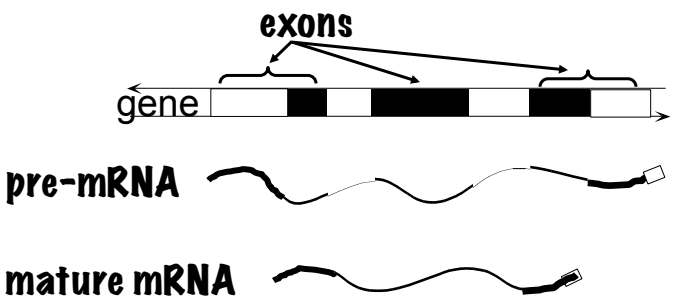
1. Which way (to the right or left) are RNA polymerases moving?

2. Which strand (W or C) is the template strand?

10

Processing of pre-mRNA

Eukaryotic genes are interrupted by introns (non-coding information). They must be removed from the RNA before translation in a process called "splicing."



11

Review of the Central Dogma: Translation

Translating the nucleic acid code to a peptide code…

Possible coding systems:

1 base per amino acid → Could only code for 4 amino acids!

2 bases per amino acid →

3 bases per amino acid →

12

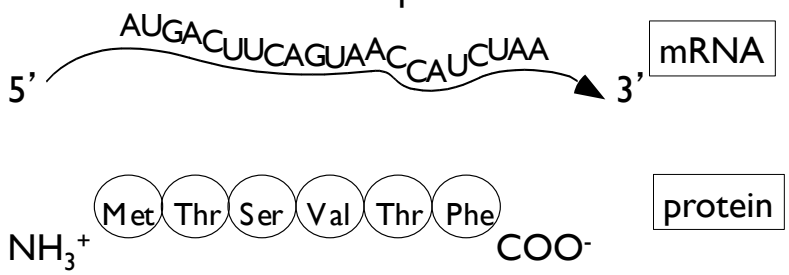
The triplet code

3 bases = 1 amino acid

More than 1 triplet can code for the same amino acid

codon

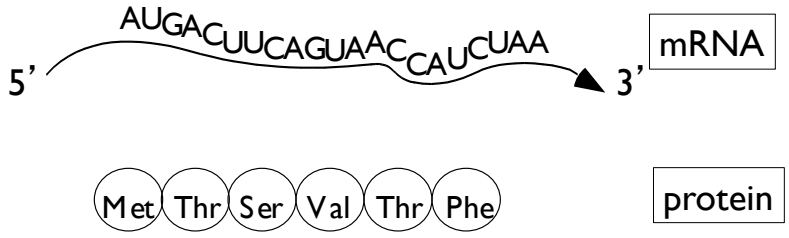
Translation: read the information in RNA to order the amino acids in a protein



13

Punctuation: start:

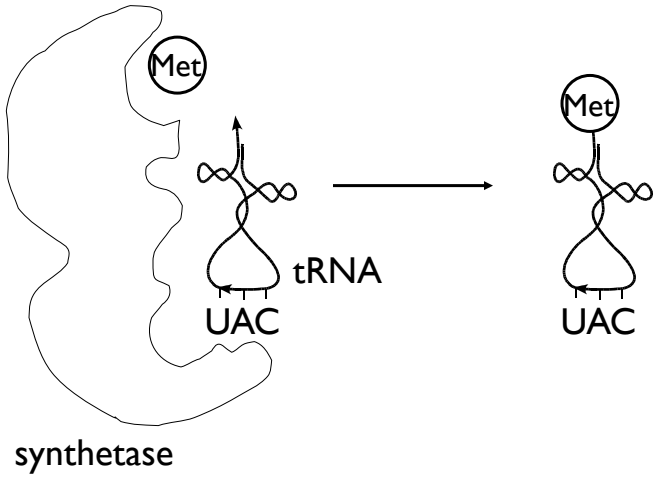
stop:



14

The Genetic Code: Who is the interpreter? Where's the dictionary? What are the rules of grammar?

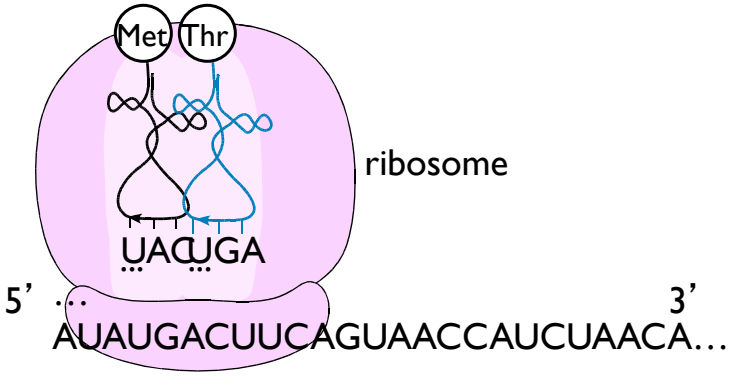
tRNA = transfer RNA



anticodon recognizes codon in mRNA

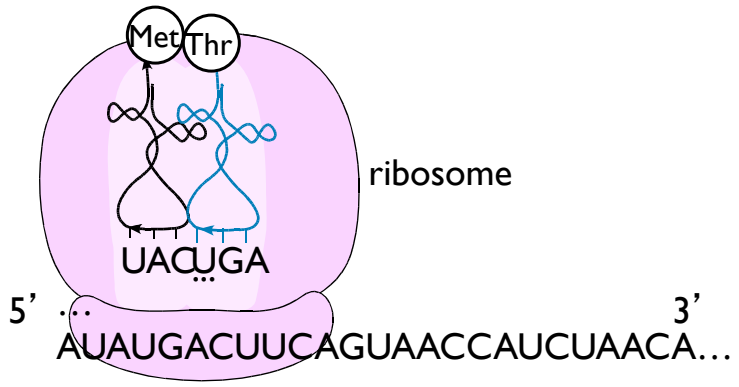
15

The **ribosome**: mediates translation

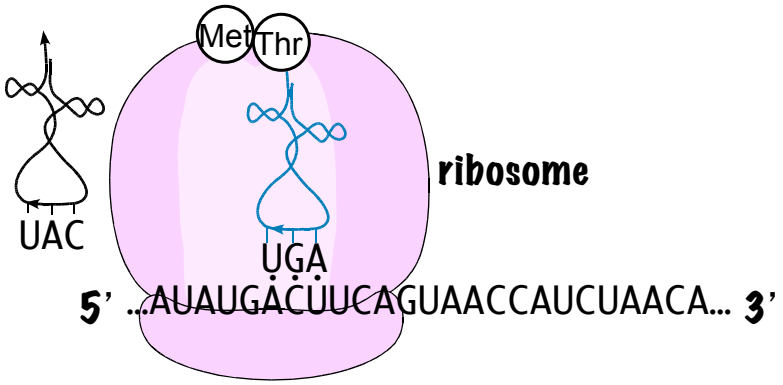


16

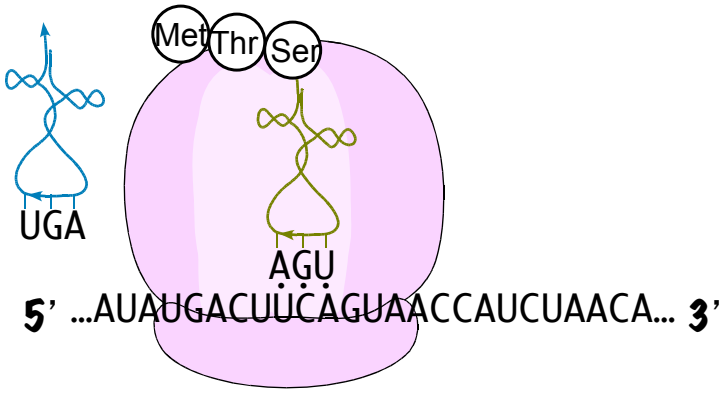
the ribosome breaks the Met-tRNA bond; Met is instead joined to the second amino acid



17

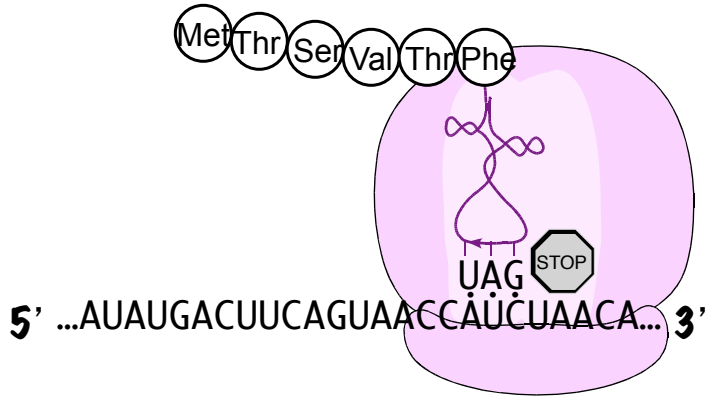


18



19

When the ribosome reaches the Stop codon…
termination

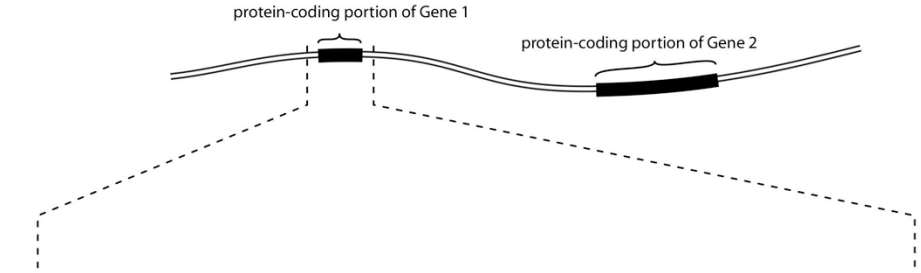


20

5' ...AUAUGACUUCAGUAACCAUCUAACA... 3'

21

Practice Question



5'-GATTAAAAGAATGAGATTTCCTTCAATTTTTACTGCAGTTTTATTCGCATAAGCCCGACT-3'
3'-CTAATTCTTTTAAGGAAGTTAAAAATGACGTCAAAATAGCGTATTCGGGCTGA-5'

transcription

5'-AAAAGAAUGAGAUUUCCUUCAAUUUUUACUGCAGUUUUAUUCGCAUAAGCCCGACU-3'

1. Which strand on the DNA sequence is the coding (sense) strand? How can you tell?

22

Finding Sense in Nonsense

cbdryloiaucahjdhtheflybitthedogbutnotthecatjhhajctipheq

GGGTATAGAAAATGAATATAAACTCATAGACAAGATCGGTGAGGGAACATTTTCGTCAGTGTATAAAGCCAAAGA
TATCACTGGGAAAATAACAAAAAAATTTGCATCACATTTTTGGAATTATGGTTCGAACTATGTTGCTTTGAAGAAA
ATATACGTTACCTCGTCACCGCAAAGAATTTATAATGAGCTCAACCTGCTGTACATAATGACGGGATCTTCGAGA
GTAGCCCCTCTATGTGATGCAAAAAGGGTGCGAGATCAAGTCATTGCTGTTTTACCGTACTATCCCCACGAGGA
GTTCCGAACTTTCTACAGGGATCTACCAATCAAGGGAATCAAGAAGTACATTTGGGAGCTACTAAGAGCATTGA
AGTTTGTTCATTCGAAGGGAATTATTCATAGAGACATCAAACCGACAAATTTTTTATTTAATTTGGAATTGGGGCG
TGGAGTGCTTGTTGATTTTGGTCTAGCCGAGGCTCAAATGGATTATAAAAGCATGATATCTAGTCAAAACGATTA
CGACAATTATGCAAATACAAACCATGATGGTGGATATTCAATGAGGAATCACGAACAATTTTGTCCATGCATTATG
CGTAATCAATATTCTCCTAACTCACATAACCAAACACCTCCTATGGTCACCATACAAAATGGCAAGGTCGTCCAC
TTAAACAATGTAAATGGGGTGGATCTGACAAAGGGTTATCCTAAAAATGAAACGCGTAGAATTAAAAGGGCTAAT
AGAGCAGGGACTCGTGGATTTCGGGCACCAGAAGTGTTAATGAAGTGTGGGGCTCAAAGCACAAAGATTGAT
ATATGGTCCGTAGGTGTTATTCTTTTAAGTCTTTTGGGCAGAAGATTTCCAATGTTCCAAAGTTTAGATGATGCG
GATTCTTTGCTAGAGTTATGTACTATTTTTGGTTGGAAAGAATTAAGAAAATGCGCAGCGTTGCATGGATTGGGT
TTCGAAGCTAGTGGGCTCATTTGGGATAAACCAAACGGATATTCTAATGGATTGAAGGAATTTGTTTATGATTTG
CTTAATAAAGAATGTACCATAGGTACGTTCCCTGAGTACAGTGTTGCTTTTGAAACATTCGGATTTCTACAACAA
GAATTACATGACAGGATGTCCATTGAACCTCAATTACCTGACCCCAAGACAAATATGGATGCTGTTGATGCCTAT
GAGTTGAAAAAGTATCAAGAAGAAATTTGGTCCGATCATTATTGGTGCTTCCAGGTTTTGGAACAATGCTTCGA
AATGGATCCTCAAAAGCGTAGTTCAGCAGAAGATTTACTGAAAACCCCGTTTTTCAATGAATTGAATGAAAACAC
ATATTTACTGGATGGCGAGAGTACTGACGAAGATGACGTTGTCAGCTCAAGCGAGGCAGATTTGCTCGATAAG
GATGTTCT

How do you find out if it contains a gene? How do you identify the gene?

23

Reading Frame: the ribosome establishes the grouping of nucleotides that correspond to codons by the first AUG encountered.

5' ...AUAUGACUUCAGUAACCAUCUAACA... **3'**

ORF: from the first AUG to the first in-frame stop. The ORF is the information for the protein.

More generally: a reading frame with a stretch of codons not interrupted by stop

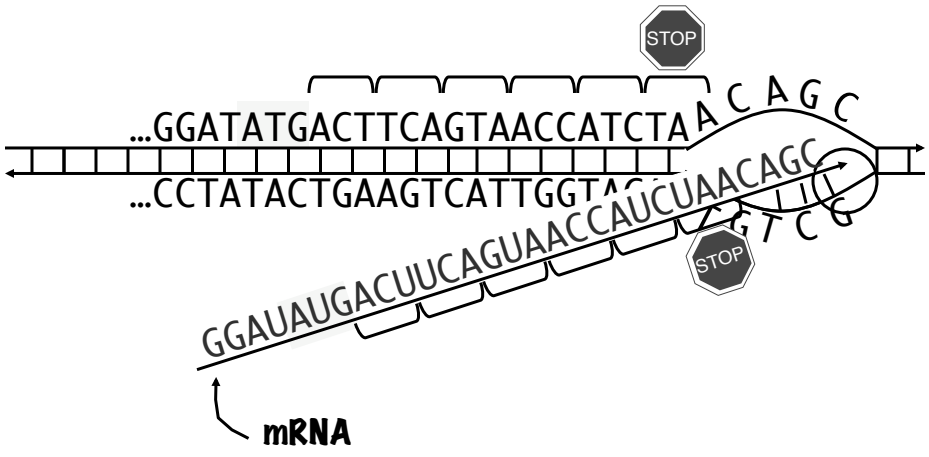
24

Looking for ORFs

- read the sequence 5' → 3', looking for stop
- try each reading frame
- since we know the genetic code—can do a virtual translation if necessary

25

Identifying ORFs in DNA Sequence



26

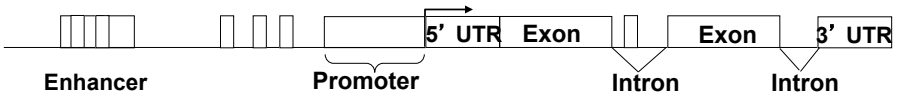
Outline

- Information Flow in Genomics
- Gene Structure
- Genetic Linkage
- Chromatin Structure
- Genome Sequencing

27

Gene Structure: The Parts List

Genomic DNA for a protein coding eukaryotic gene is comprised of regulatory and coding sequences

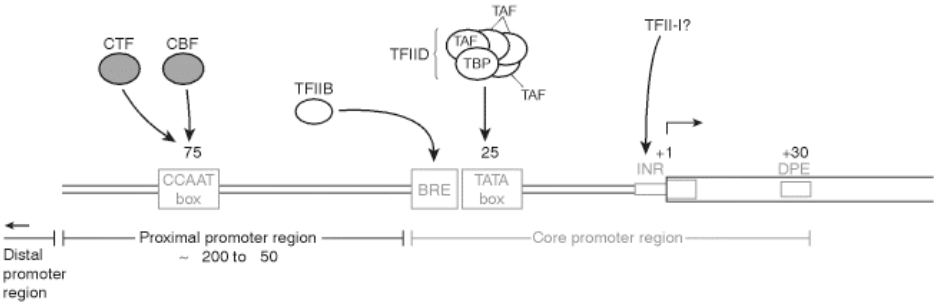


▯ = CRM (cis-regulatory motif)
- Can be upstream or downstream of promoter

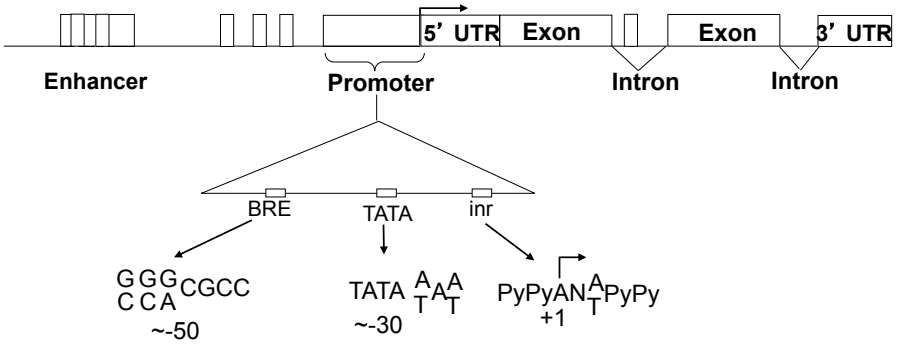
28

Promoters

- Promoters are specific sites on DNA that RNA polymerase first binds to initiate the transcription of a gene
- Composed of a variety of different *cis* sequence elements which recruit *trans*-acting factors through DNA-protein interactions



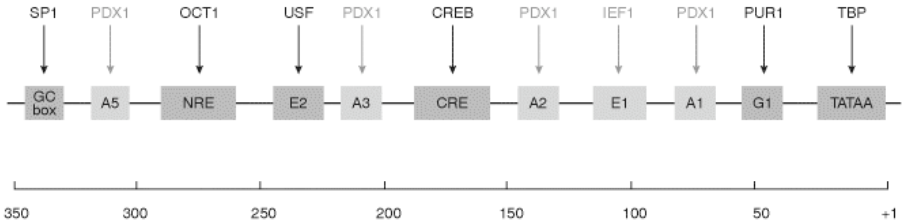
Core Promoter Elements



- not all elements required
- many promoters lack a TATA box, using instead the functionally analogous initiator (INR) element

Combinatorial Gene Regulation

- Most eukaryotic genes have multiple *cis* regulatory motifs located outside of the core promoter region
- Can be located in promoter proximal regions, 3' downstream regions, and many kb away from target gene
- Allows for *combinatorial control* of gene expression

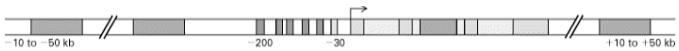


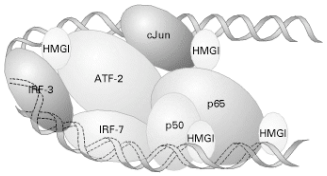
31

Distant Regulation: Enhancers

Enhancer elements:

- Can function in either orientation
- Can occur far (>50 kb) from the gene
- Can be up or downstream
- Range in size between ~50-200 bp
- Contain multiple TF binding sites





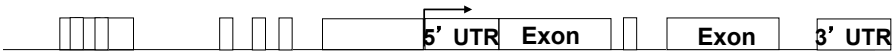
"Enhancesome"

32

http://www.ncbi.nlm.nih.gov/books/bv.fcgi?rid=mcb.figgrp.2601

Untranslated Regions (UTRs)

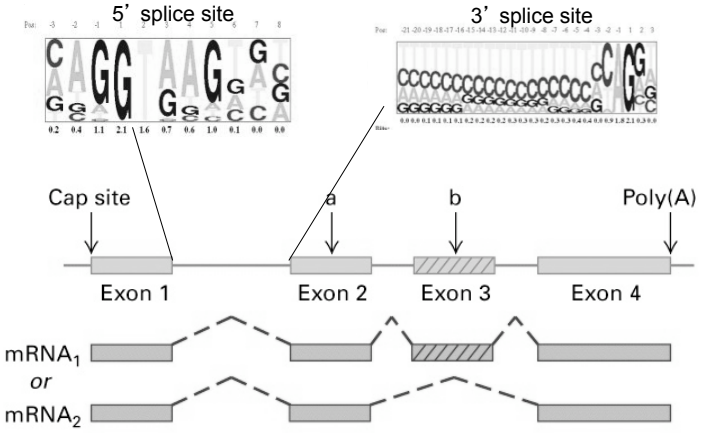
- Most eukaryotic mRNAs contain untranslated regions in their 5' and 3' ends
- The 5' UTR is the region between the start of transcription and the start of translation
- The 3' UTR is the region between the stop codon and poly-A tail
- Both the 5' and 3' UTRs can contain *cis* regulatory sequences that bind TFs, influence transport to the cytoplasm, mediate transcript stability, and translational control



33

Alternative Splicing

- mRNA from some genes can be spliced into two or more distinct transcripts
- Creates protein diversity (isoforms)



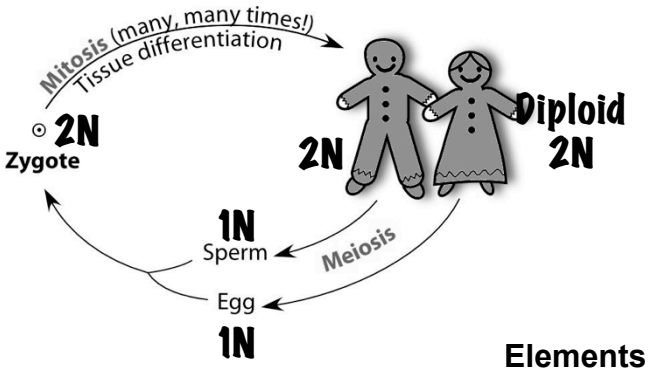
34

Outline

- Information Flow in Genomics
- Gene Structure
- Genetic Linkage
- Chromatin Structure
- Genome Sequencing

35

Transmission of Genetic Information



Elements of cell division

- Cell growth
- Chromosome duplication
- Chromosome segregation

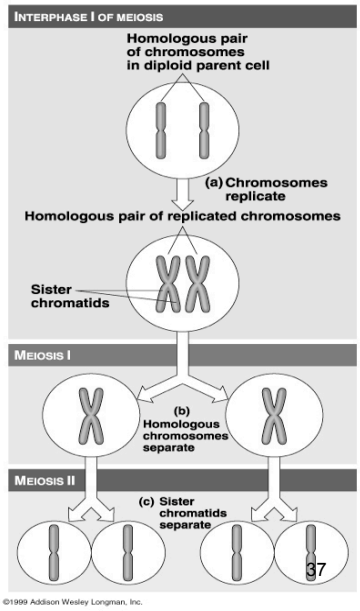
36

Meiosis

Interphase:
Chromosomes replicate

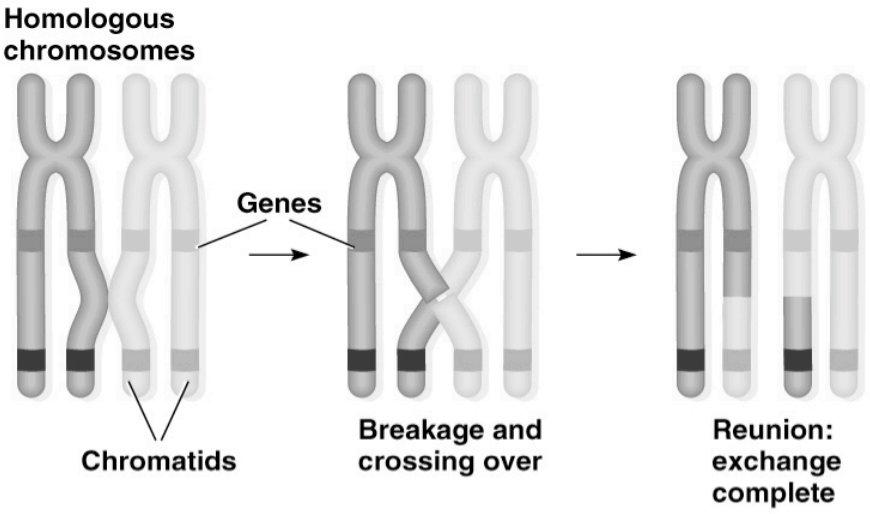
Meiosis I:
Reductive division, homologous chromosomes separate

Meiosis II:
Sister chromatids separate



37

Recombination



38

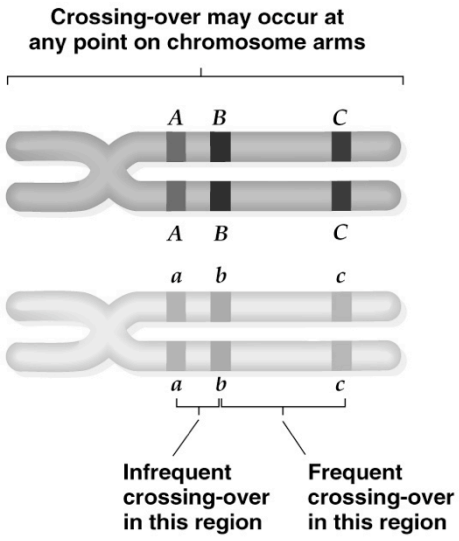
How Does Distance Between Loci Affect Transmission?

Independent Assortment: loci are far enough apart that they are transmitted independently from one another

Genetic linkage: loci are close enough together on a chromosome to be transmitted together

39

Genetic Mapping



The amount of recombination between loci is based on the distance between them

40

Recombination Is A Measure of Distance

- Recombination fraction, θ = the probability that a recombinant gamete is transmitted
- If two loci are on different chromosomes, they will segregate independently
  => recombination fraction $\theta = 0.5$
- If two loci are right next to each other, they will segregate together during meiosis
  => recombination fraction $\theta = 0$
- Jargon:
  $\theta < 0.5$ the loci are close (they are linked)
  $\theta = 0.5$ the loci are far apart (they are not linked)

41

Recombination Is A Measure of Distance

$$\text{Map Distance} = \frac{\text{Number Recombinant Gametes}}{\text{Total Number of Gametes}} \times 100$$

Centimorgan (cM): a unit of chromosome length, equals the length of chromosome over which crossing-over occurs with 1% frequency

42

Practice Question

- In maize, consider three recessive phenotypes: lazy growth (ll), glossy leaves (gg), and sugary endosperm (ss).
- The following cross was made: Ll Gg Ss x ll gg ss and the observed progeny distribution was

Phenotype	Number
wildtype	286
lazy	33
glossy	59
sugary	4
lazy, glossy	2
lazy, sugary	44
glossy, sugary	40
lazy, glossy, sugary	272
Total	740

- Determine order and distances among the three genes

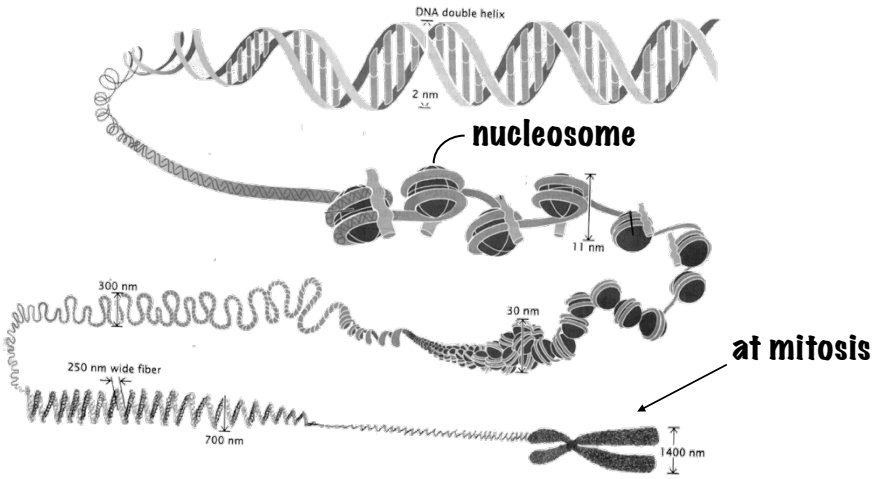
43

Outline

- Information Flow in Genomics
- Gene Structure
- Genetic Linkage
- Chromatin Structure
- Genome Sequencing

44

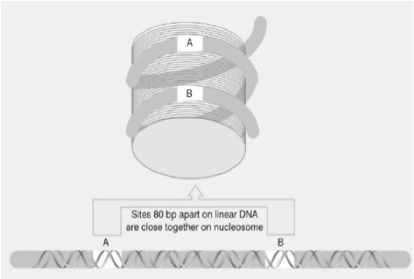
Chromosome Structure: Coils of Coils of Coils…

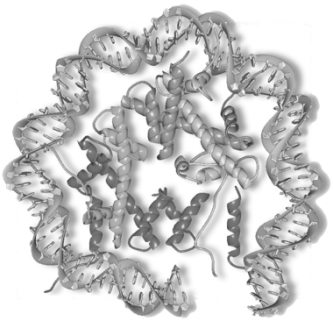


Local unpacking of chromatin…to allow gene expression and replication

45

Nucleosomes

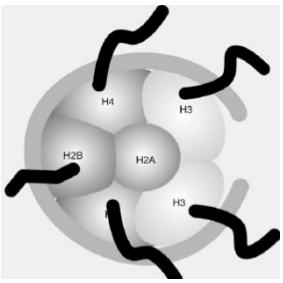


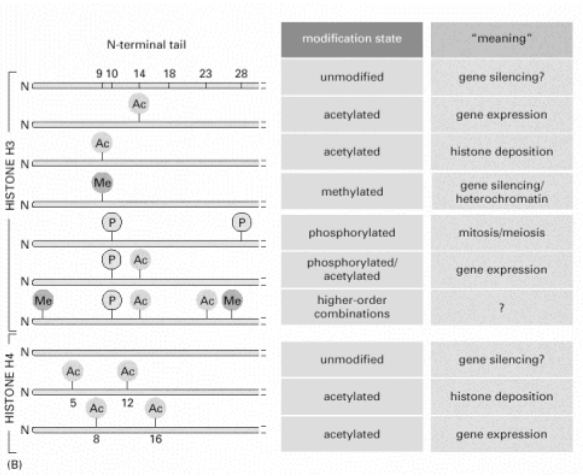


~146 bp of DNA wrapped around histone proteins

46

Histone Modification and Chromatin Structure



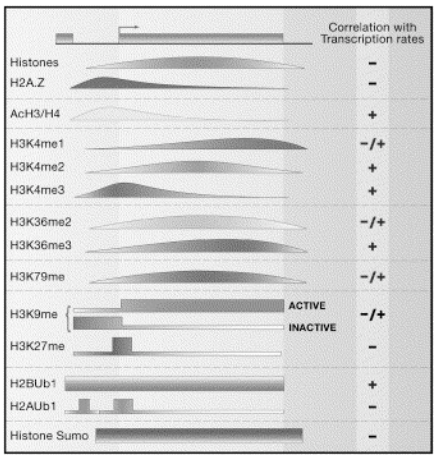


Mol. Biology of the Cell Fig. 4-35

47

What Do These Modifications Do? A Histone Code?

"Distinct histone modifications, on one or more tails, act sequentially or in combination that is read by other proteins to bring about distinct downstream events" (Strahl and Allis, 2000, Nature, 403:41)



48

Carey et al. Cell (2007) 128:707

Outline

- Information Flow in Genomics
- Gene Structure
- Genetic Linkage
- Chromatin Structure
- Genome Sequencing

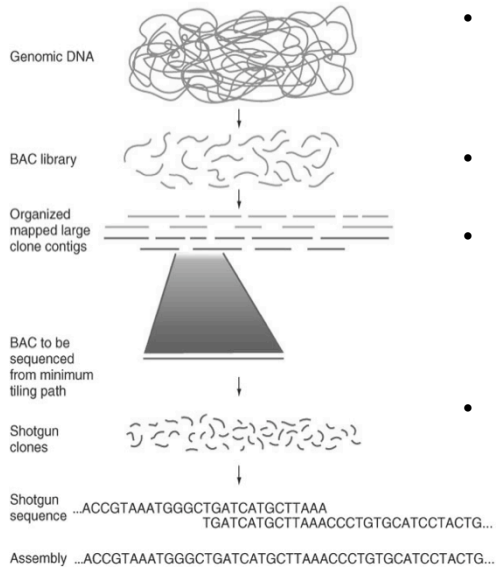
49

DNA Sequencing Technology

- **Sanger sequencing**
- **Next-Generation**
- **3rd Generation**

50

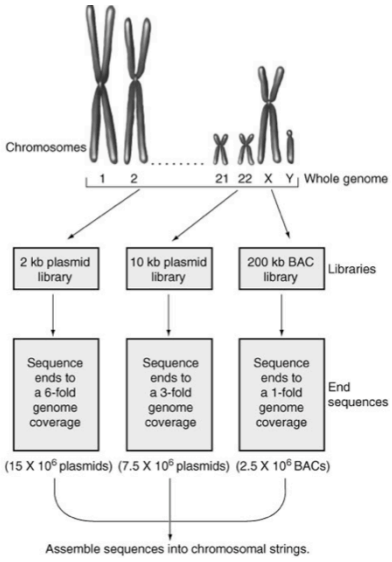
Genome Sequencing: Hierarchical Shotgun Sequencing



- Shear genomic DNA into smaller pieces and subclone into library (such as BACs, Cosmids, etc.)
- Create physical map
- Shotgun sequence each BAC from minimal tiling path (shearing of ~150kb BAC clone into ~ 2kb fragments)
- Data from linkage and physical maps used to assemble sequence maps of chromosomes

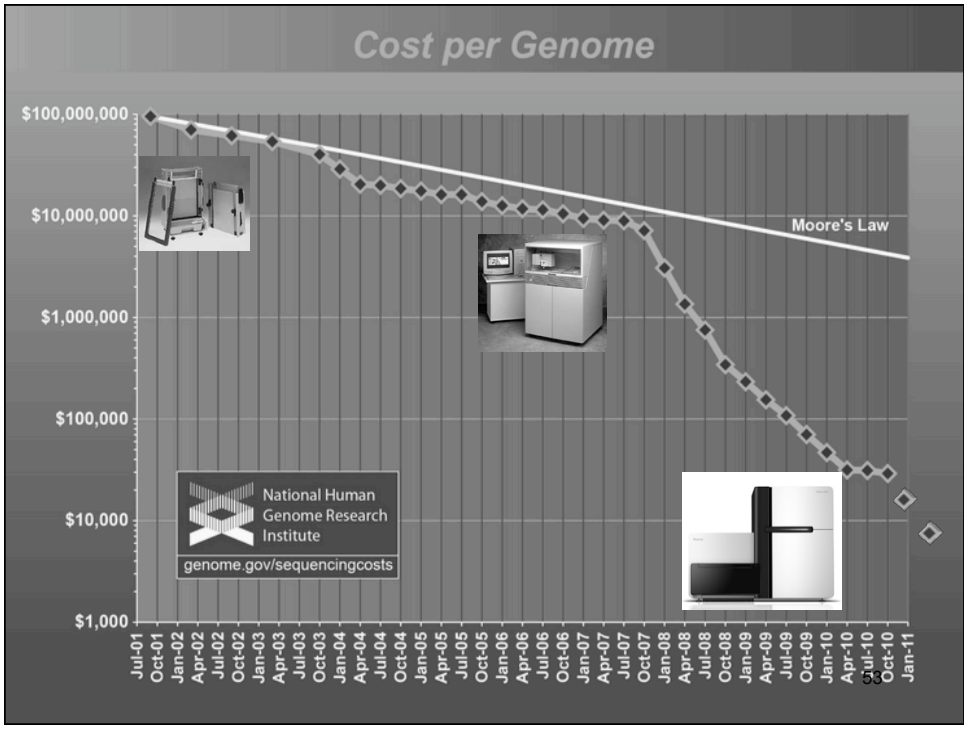
51

Genome Sequencing: Whole Genome Shotgun Sequencing



- Whole genome randomly sheared three times
  - Plasmid library constructed with ~ 2kb inserts
  - Plasmid library with ~10 kb inserts
  - BAC library with ~200 kb inserts
- Computer program assembles sequences into chromosomes
- No physical map construction
- Only one BAC library
- Overcomes problems of repeat sequences

52

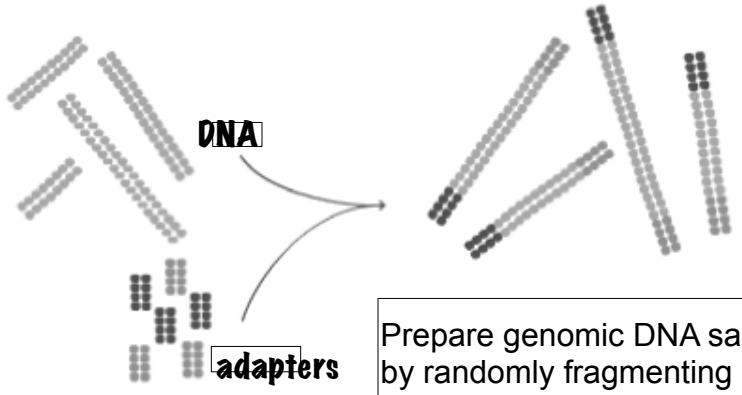


Next-Generation Sequencing Technology

- Illumina HiSeq:
  - 100 million reads per sample X 200 bases = 20 Gbp
  - 8 samples per run = 200 Gbp
    (one human genome = 3 Gbp)
  - Reagent cost ~$8K per run
- ABI SOLID: claims similar yield
- Roche 454: 1 million reads X 500 bases = 0.5 Gbp
  - Can be subdivided and barcoded almost without limit
  - Reagent cost ~$10K per run

54

Illumina Sequencing: DNA Library Prep



Prepare genomic DNA sample by randomly fragmenting DNA and ligating adapters to both ends of the fragments

55

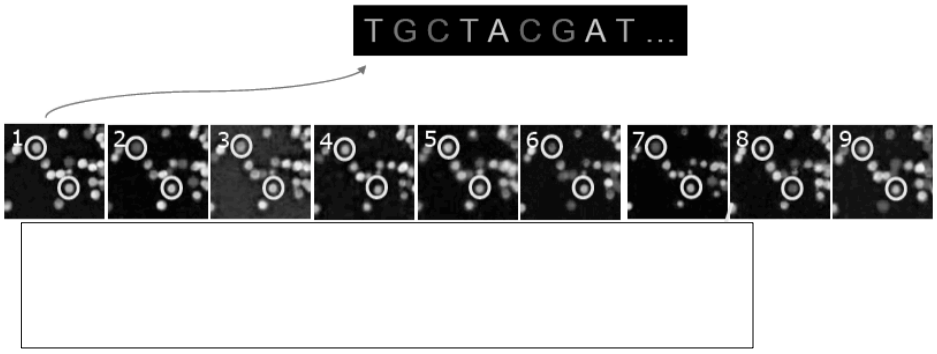
Illumina Sequencing: DNA clusters

Attach DNA to surface

Use primers on surface for PCR to generate DNA clusters

56

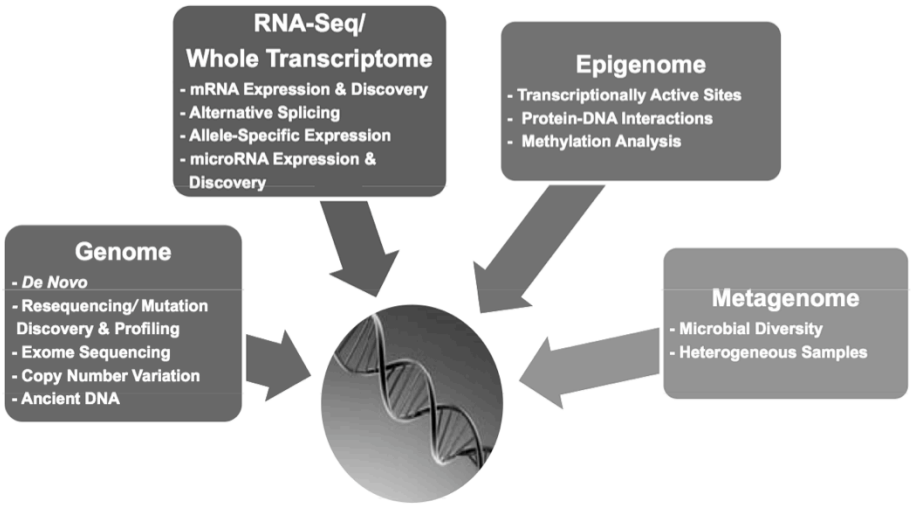
Illumina next generation sequencing: Sequence reaction



Sequence clusters using reversible terminators, by imaging after each cycle of synthesis

57

Next-Gen Sequencing - What's All the Fuss About?



58

First step of nearly all next-gen studies: mapping reads

- **Hash Table (Lookup table)**
  - FAST, but requires perfect matches
- **Array Scanning**
  - Can handle mismatches, but not gaps
- **Dynamic Programming (Smith Waterman, Forward, Viterbi)**
  - Indels
  - Mathematically optimal solution
  - Slow (most programs use Hash Mapping as a prefilter)
- **Burrows-Wheeler Transform (BW Transform)**
  - FAST (memory efficient)
  - But for gaps/mismatches, it lacks sensitivity

59

Some Aligners…

Aligner	Version	Developer	License
BFAST	0.3.1	UCLA	Academic
Bowtie	0.9.8	Salzberg/UMD	GNUPL
cross_match	1.080721	U. Wash.	Academic
CELL	2.0	CLC bio	Commercial
Maq	0.7.1	Sanger	GNUPL
Novoalign	2.00.12	Novocraft	Academic
RMAP	0.41	CSHL	Public
SeqMap	1.0.12	Stanford	Academic
SHRiMP	1.1.0	U. Toronto	Public
SOAP	2.01	BGI	Academic

Slide Courtesy of Dan Kobolt, Genome Sequencing Center
See www.massgenomics.org

60

NGS File Formats…

- Raw data from various vendors
- Different quality metrics (with different levels of empirical validity)
- GenBank formats (SRA)
- Alignments = SAM/BAM
- Genome Browser formats (wig, bed, gff, etc)
- Variants (SNPs, indels, etc)

61

FASTQ Format: Sequence plus Quality information

@SRR350953.5 MENDEL_0047_FC62MN8AAXX:1:1:1646:938 length=152

NTCTTTTTCTTTCCTCTTTTGCCAACTTCAGCTAAATAGGAGCTACACTGATTAGGCAGAAACTTGATTAACAGGGCTTAAG GTAACCTTGTTGTAGGCCGTTTTGTAGCACTCAAAGCAATTGGTACCTCAACTGCAAAAGTCCTTGGCCC

+SRR350953.5 MENDEL_0047_FC62MN8AAXX:1:1:1646:938 length=152

+50000222C@@@@@22::::8888898989::::::<<<:<<<<<<:<<<<::<<::::<<<<<:<:<<<IIIIIGFEEGGGGGGGII@IGDGB GGGGGGDDIIGIIEGIGG>GGGGGGDGGGGGIIHIIBIIIGIIIHIIIIGII

@SRR350953.6 MENDEL_0047_FC62MN8AAXX:1:1:1686:935 length=152

NATTTTTACTAGTTTATTCTAGAACAGAGCATAAACTACTATTCAATAAACGTATGAAGCACTACTCACCTCCATTAACATG ACGTTTTTCCCTAATCTGATGGGTCATTATGACCAGAGTATTGCCGCGGTGGAAATGGAGGTGAGTAGTG

+SRR350953.6 MENDEL_0047_FC62MN8AAXX:1:1:1686:935 length=152

#---+83355@@@CC@C22@@C@@CC@@C@@@CC@@@@@@@@@@@@C? C22@@C@::::@@@@@@C@@@@@@@@@CIGIHIIDGIGIIIIHHIIHGHHIIHHIFIIIIIHIIIIIIBIIIEIFGIIIFGFIBGDGGGG GGFIGDIFGADGAE

@SRR350953.7 MENDEL_0047_FC62MN8AAXX:1:1:1724:932 length=152

NTGTGATAGGCTTTGTCCATTCTGGAAACTCAATATTACTTGCGAGTCCTCAAAGGTAATTTTTGCTATTGCCAATATTCCTC AGAGGAAAAAAGATACAATACTATGTTTTATCTAAATTAGCATTAGAAAAAAAATCTTTCATTAGGTGT

+SRR350953.7 MENDEL_0047_FC62MN8AAXX:1:1:1724:932 length=152

#.,')-2--/@@@@@@@@@@<:<<: 778789979888889:::::99999<<::<::::<<<<<@@@@@::::::IHIGIGGGGGGDGGDGGDDDIHIHIIIII8GGGGGIIHHIIIGII GIBIGIIIIEIHGGFIHHIIIIIIIGIIFIG

62

SAM Format: Aligned to the Reference Genome

```
MCL-SRR350952.1     99      chr13     28330526  29      76M     =
        28330636    183
        NAAAGACACAGTTACATGAAGAACATACTCCTCTCTCAGACTGCCCAGGTTCAGTGATTCATTCAACAAACTTTAT
        #,())*3--.@@@@@@@@@@<<<::87999<<<<<@@@@@@@:@@<:<<<<<8<<::7::::::::@@@22::::@
        XT:A:U      NM:i:1      SM:i:29     AM:i:29     X0:i:1      X1:i:0      XM:i:1      XO:i:0
        XG:i:0      MD:Z:0T75   XA:Z:chr13,+28330526,76M,1;
MCL-SRR350952.1     147     chr13     28330636  29      73M3S   =
        28330526    -183
        TATAGGATTCAACTGTGAGAAAGACATATTAATCTCTTCCATTGTGCAGACTACATTCTTTTTTTTTTTTTTTGAG
        ######################B7:?2,?8;;@+=+A@3DEBB?2?5B7=A?=?4<8;;AIGIDIHFAIIGHIIBII
        XT:A:M      NM:i:2      SM:i:29     AM:i:29     XM:i:2      XO:i:0      XG:i:0      MD:Z:
18A18G35
MCL-SRR350952.2     99      chr9      16437227  60      76M     =
        16437304    153
        NGAAATGCAAGGCTGTTTGGGATGTTTTCGAAGTGATGAATGCTGGAAGGATTGCTGTTCTCTAAGTGAGCAAGGA
        ##############################################################################
        XT:A:U      NM:i:1      SM:i:37     AM:i:37     X0:i:1      X1:i:0      XM:i:1      XO:i:0
        XG:i:0      MD:Z:0A75   XA:Z:chr9,+16437227,76M,1;
MCL-SRR350952.2     147     chr9      16437304  60      76M     =
        16437227    -153
        GCTGGGACTCCTGGTGCGATTATTGCTCTCAATGAAAGTCCTTATATCTGAGTCTGTCTTTGAAGATGGTACAGCC
        DBAEA<>G>GGDGG<DD3E<CDCC>E?D?E@DGDDDG8BGGGEGEG@E@@BCCFEE,IHIIGEGGEFCCF<FFFFD
        XT:A:U      NM:i:0      SM:i:37     AM:i:37     X0:i:1      X1:i:0      XM:i:0      XO:i:0
        XG:i:0      MD:Z:76     XA:Z:chr9,-16437304,76M,0;
```

63

VCF Format: Variants

```
chr1        901559      .       G       A       23      .
        DP=3;VDB=0.0298;AF1=1;AC1=2;DP4=0,0,3,0;MQ=60;FQ=-36        GT:PL:GQ    1/1:55,9,0:15
chr1        905974      .       C       T       187     .
        DP=22;VDB=0.0249;AF1=0.5;AC1=1;DP4=5,4,9,3;MQ=59;FQ=170;PV4=0.4,0.2,0.2,0.49   GT:PL
:GQ         0/1:217,0,197:99
chr1        907622      .       G       C       82.3    .
        DP=8;VDB=0.0474;AF1=1;AC1=2;DP4=0,0,2,3;MQ=54;FQ=-42        GT:PL:GQ    1/1:115,15,0:
27
chr1        909073      .       C       T       39      .
        DP=16;VDB=0.0504;AF1=0.5;AC1=1;DP4=0,9,0,7;MQ=59;FQ=42;PV4=1,7.5e-09,0.14,0.49          GT:PL
:GQ         0/1:69,0,156:72
chr1        909238      .       G       C       115     .
        DP=10;VDB=0.0503;AF1=1;AC1=2;DP4=0,0,4,4;MQ=60;FQ=-51        GT:PL:GQ    1/1:148,24,0:
45
chr1        909309      .       T       C       3.01    .
        DP=6;VDB=0.0018;AF1=0.4998;AC1=1;DP4=0,2,0,3;MQ=60;FQ=4.72;PV4=1,0.02,1,0.0071          GT:PL
:GQ         0/1:30,0,47:28
chr1        909768      .       A       G       137     .
        DP=21;VDB=0.0530;AF1=1;AC1=2;DP4=0,0,11,9;MQ=60;FQ=-87       GT:PL:GQ    1/1:170,60,0:
99
chr1        935222      .       C       A       4.61    .
        DP=4;VDB=0.0216;AF1=1;AC1=2;DP4=0,0,0,2;MQ=60;FQ=-33         GT:PL:GQ    1/1:34,6,0:5
```

64

The Era of Personal Genomics?

James D. Watson (5/31/2007)

http://www.ffrf.org/day/img/0406_watson.gif, http://www-news.uchicago.edu/releases/07/images/070601.watson.jpg

J. Craig Venter (8/4/2007)

http://www.usnews.com/usnews/images/news/photos/venter051022.jpg

THE DNA AGE

Gene Map Becomes a Luxury Item

Fred

BIG SPENDER Dan Stoicescu is paying a company to map his genome.

By AMY HARMON
Published: March 4, 2008

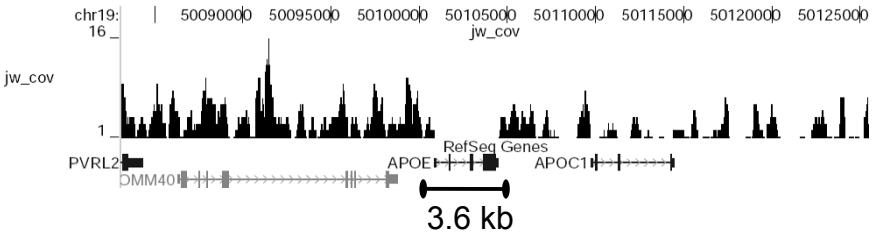
On a cold day in January, Dan Stoicescu, a millionaire living in Switzerland, became the second person in the world to buy the full sequence of his own genetic code.

The DNA Age
$350,000 Genome
Articles in this series explore the

New York Times 3/4/08

65

"Censoring" of Watson's ApoE gene



Important ethical issues confront personal genomics

66

3rd Generation Sequencing Technology

- Single Molecule Sequencing Technology (tSMS)
- No amplification
- Single Molecule Real Time (SMRT) sequencing technology (PacBio RS)

67

Interpreting Genome Sequences

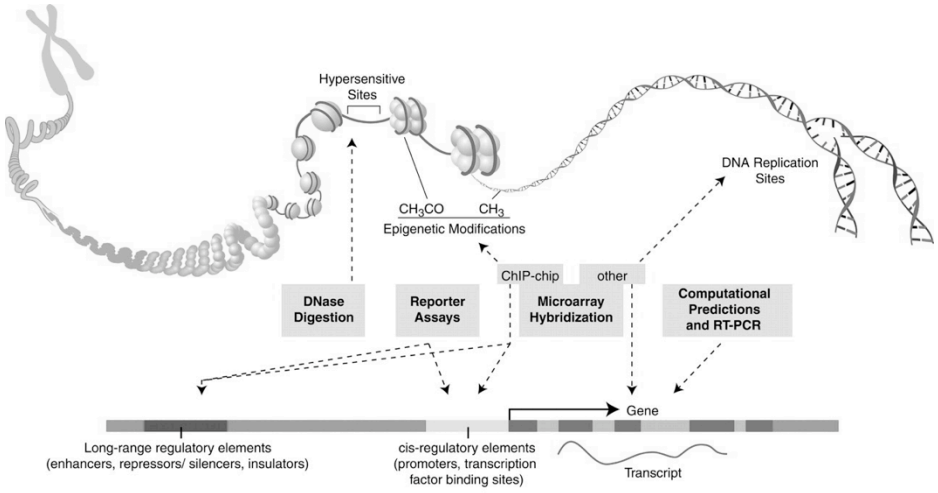
The ENCODE Project: comprehensive parts list of the functional elements in the human genome

Cover art: Darryl Leja, NHGRI

- Pilot Project Description
  - ENCODE Project Consortium et al. The ENCODE (ENCyclopedia Of DNA Elements) Project. Science (2004) vol. 306 (5696)
- Pilot Project Results
  - ENCODE Project Consortium et al. Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. Nature (2007) vol. 447 (7146)

68

ENCODE Project



69

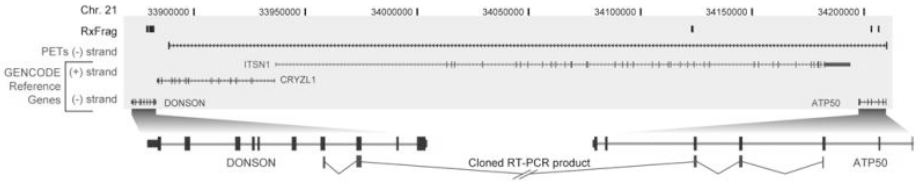
ENCODE Highlights

- Transcription is more complex than expected
- Transcription Start Sites (TSS) more numerous than protein-coding genes
- Regulatory information is distributed:
  - Clustered across the genome, distribution near TSSs is symmetrical.
- There are many distal DNaseI Hypersensitive Sites (DHSs)
- Replication is correlated with histone structure in a more detailed manner than previously known
- Evolutionary constraint be can detected at far higher resolution than was previously feasible; surprisingly, many experimentally identified features did not show constraint

70

Transcriptional Complexity

- 93% of bases covered in a primary transcript!
- 10x the number of TSS compared to number of protein coding genes
- Protein complexity:



71

Let's Play "Gene" or "No Gene"

A gene is <u>often</u> a segment of DNA that encodes a protein.

How about DNA that encodes:

a micro RNA that binds to an mRNA to inhibit translation?

an RNA spliced out of an intron and used for another function?

an antisense transcript?

a long non-coding RNA of unknown function?

a pseudogene?

72

Population Genomics

Joshua Akey
Department of Genome Sciences
akeyj@uw.edu

1

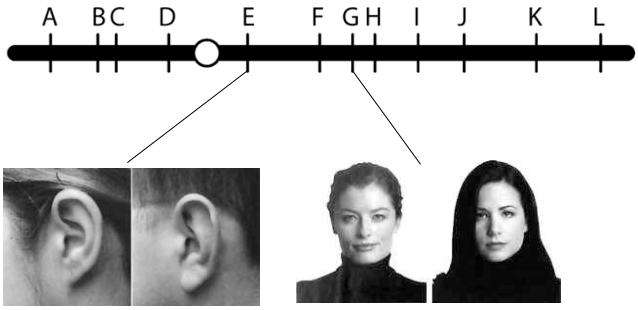
Outline

- Genetic markers
- Measures of genetic variation
- Inferences from genetic variation
  - Linkage Disequilibrium
  - Population structure
  - Recombination rate
  - Natural Selection

2

Markers

Genetic markers - inherited variations that are used to test genetic hypotheses



3

What Is A Polymorphic Molecular Marker?

A polymorphic site or locus…

A location in the genome where at least two versions of the sequence exist in the population,

each at a frequency of at least 1%

UW student population - N ~ 40,000

⇒ 2N = 80,000 copies of (e.g.) chromosome 2

70,000 copies have A-T base pair
10,000 copies have C-G base pair

each is at > 1% of total population, so this <u>is</u> a polymorphic site

4

Types of Polymorphic Molecular Markers

1. Single Nucleotide Polymorphisms (SNPs)

```
..TCTTGATC..
..TCTCGATC..
```

2. Insertion/Deletions (Indels)

```
..TCTTGATC..
..TCTTTGATC..
```

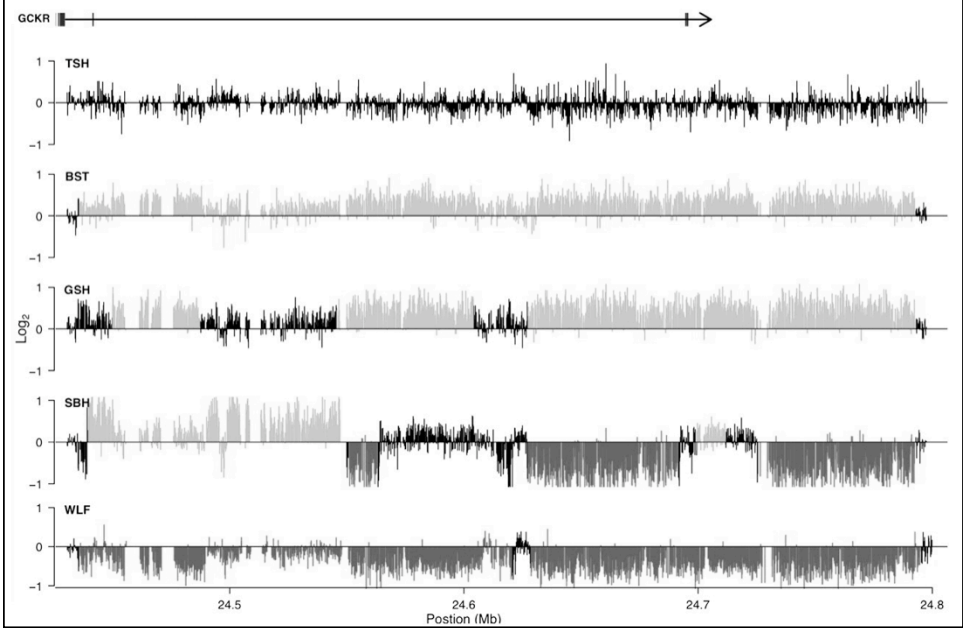
3. Variable Number of Tandem Repeats (VNTRs)

```
..CCGCAGCAGCAGCAGCAGATTC..
..CCGCAGCAGCAGCAGCAGCAGATTC..
..CCGCAGCAGCAGCAGCAGCAGCAGATTC..
```

4. Restriction Fragment Length Polymorphisms (RFLPs)

5

Polymorphisms of DNA Content



Outline

- Genetic markers
- Measures of genetic variation
- Inferences from genetic variation
  - Linkage Disequilibrium
  - Population structure
  - Recombination rate
  - Natural Selection

7

Measuring Nucleotide Variation

```
..TCTTGAATCGGACGTATGCTCAATTACGATC..
..TCTCGATTCGGACGTATACTCAATTACGATC..
```

- If it was possible to sequence your genome, how many SNPs would we expect to find?

  ~ 1 SNP per 1000 bp => 3 million

- Stable genetic markers: mutation rate ~ 2×10^{-8}/site/gen
  How many new SNPs do you carry?

  ~ $3 \times 10^{9} \times 2 \times 10^{-8} \times 2$ = 120 new SNPs

8

Quantifying Nucleotide Variation

1. *Nucleotide Polymorphism*: the proportion of nucleotide sites that are expected to be polymorphic in a sample of size n chromosomes:

$$\theta_w = \frac{S}{\sum_{i=1}^{n-1} \frac{1}{i}}$$

2. *Nucleotide Diversity* (aka nucleotide heterozygosity): the average number of nucleotide differences between two randomly selected DNA sequences:

$$\pi = \left(\frac{n}{n-1}\right) \sum_{i=1}^{S} 2 p_i q_i$$

9

Example

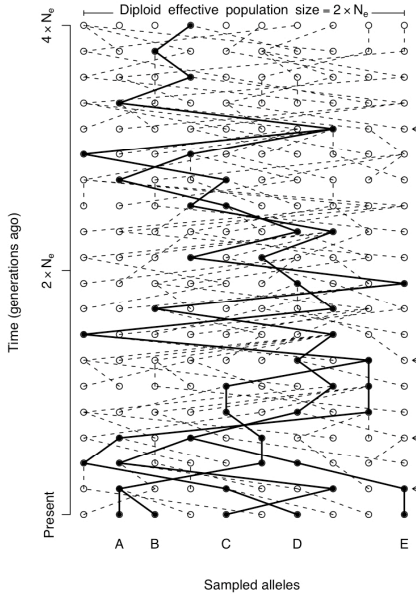
Sample 1 `TACCGCATT`**`G`**`CGTAGGCTCATAAGCTACTAAGATAC`
Sample 2 `TACCGCATTCCGTAGGCTCATAAGCTACTAAGATAC`
Sample 3 `TACCGCATTCCGTAGGCTCATAAGCTACTA`**`T`**`GATAC`

Divide by number of bp sequenced to get per nucleotide estimates

Under certain assumptions $\theta_W = \pi = 4N_e\mu$

10

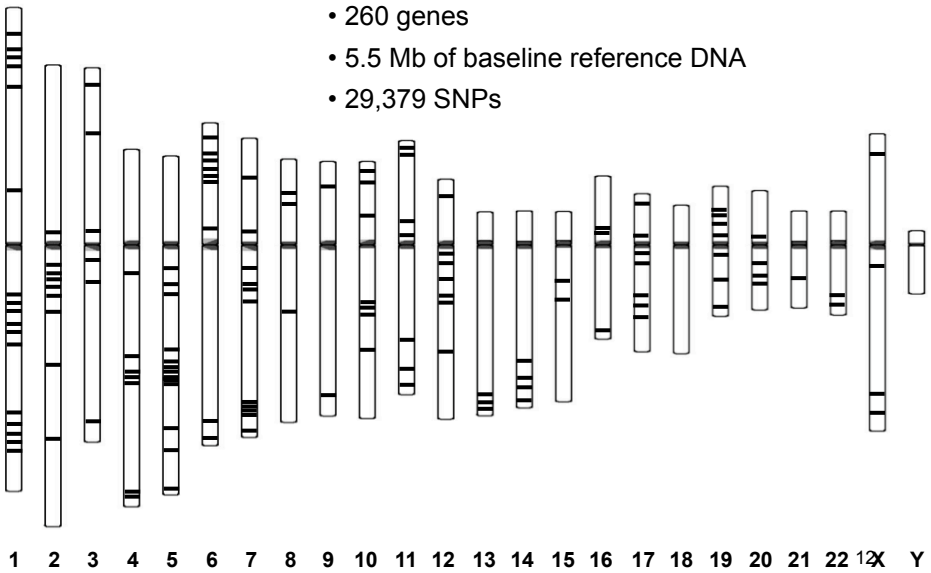
The Coalescent



11

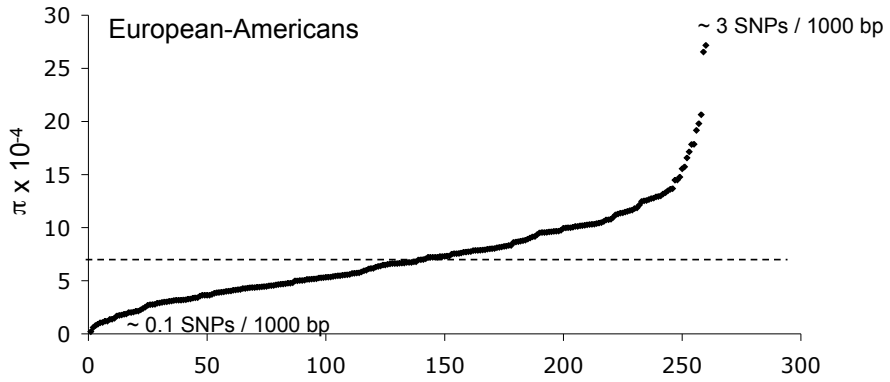
Empirical Observations: SeattleSNP

- 260 genes
- 5.5 Mb of baseline reference DNA
- 29,379 SNPs



12

Empirical Observations: Diversity Varies Across Genome



Average π across all 260 genes is 7.35×10^{-4}, but considerable variation!

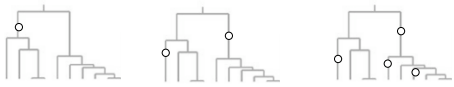
13

Why So Much Variation?

Patterns of polymorphism are determined by both *genealogical* and *mutational* history

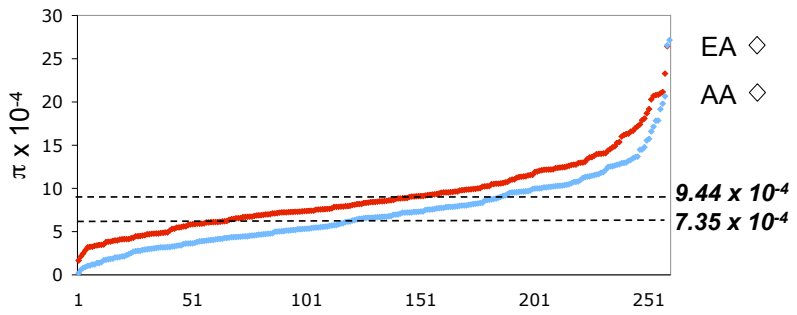
Gene genealogies vary across the genome due to the stochastic nature of coalescent events:

Mutation rates vary by several orders of magnitude across the genome



14

Empirical Observations: Diversity Varies Across Populations

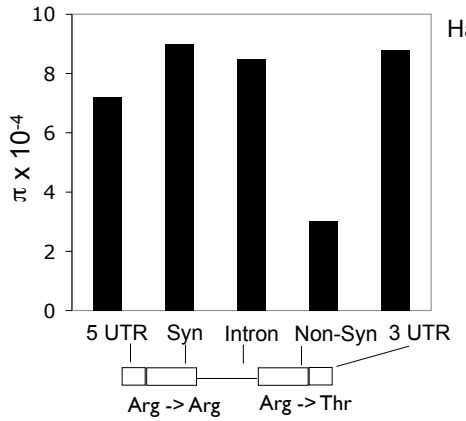


Why?

- Differences in population demographic history?
- Adaptation to distinct selective pressures?

15

Empirical Observations: Diversity Varies Across Types of Genomic Regions



Halushka et al (2003)

Why?

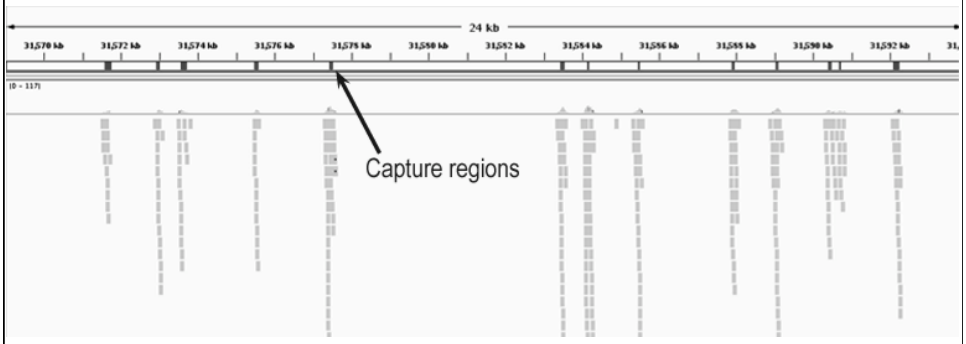
- Purifying selection eliminates deleterious mutations and therefore reduces genetic diversity in regions of functional constraint

16

Deep sequencing reveals new insights into population history

NHLBI Exome Sequencing Project (ESP)

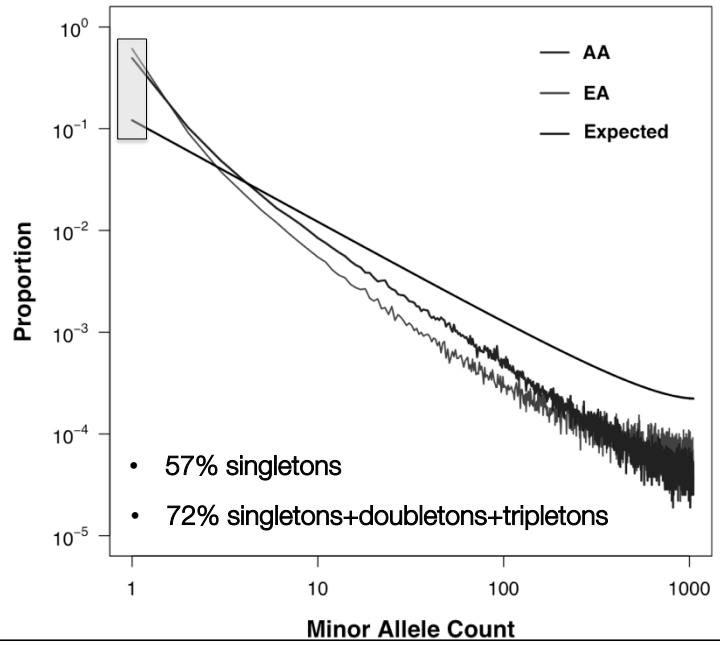
- Exome is the set of all protein-coding sequences (~1.5% of genome; 30 Mb)



- Deep exome sequencing of 6,515 individuals of European and African ancestry (**>100X** coverage) for ~16,000 genes

17

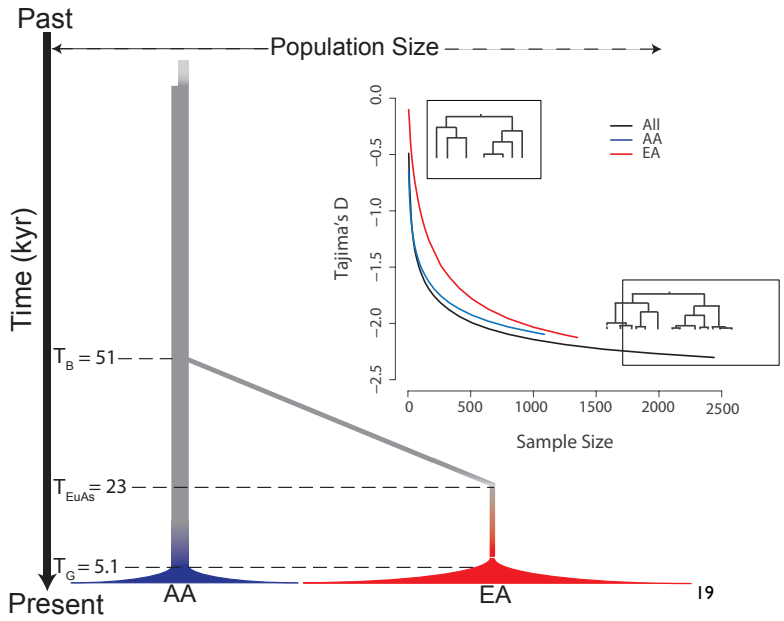
Most human variation is very rare



- 57% singletons
- 72% singletons+doubletons+tripletons

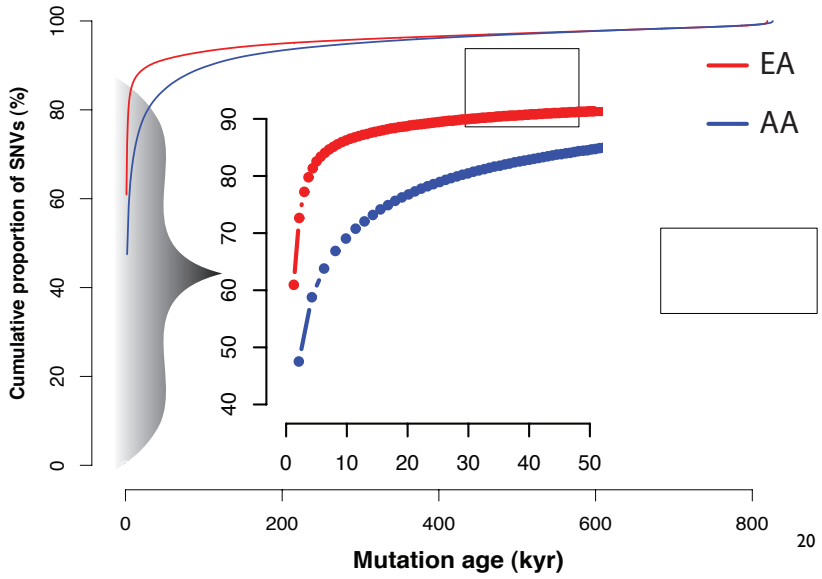
18

Primarily due to recent explosive growth



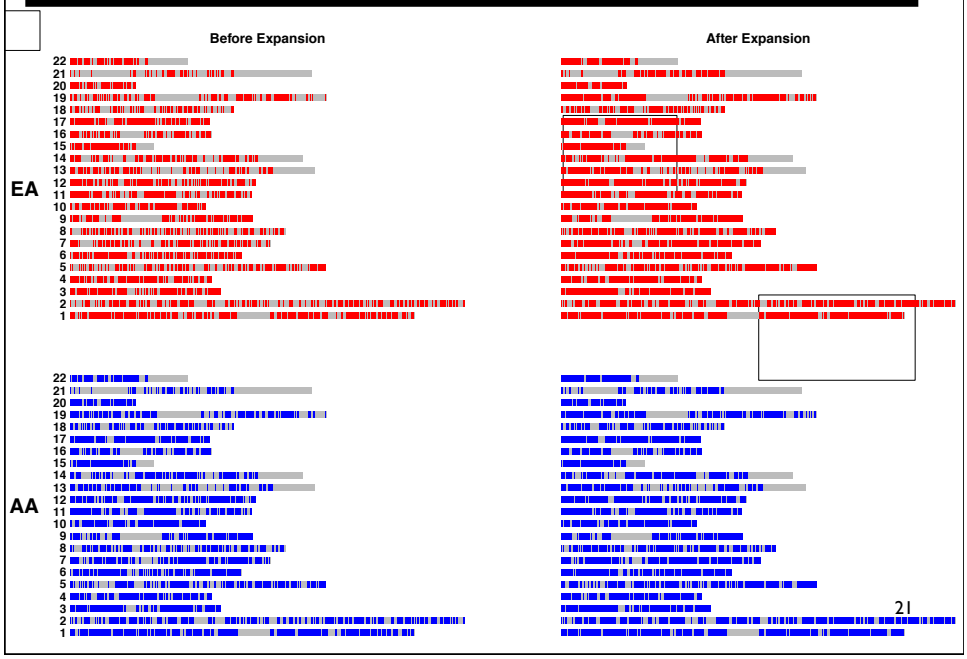
19

Most protein-coding variation arose recently



20

~90% predicted deleterious alleles arose < 5 kyr



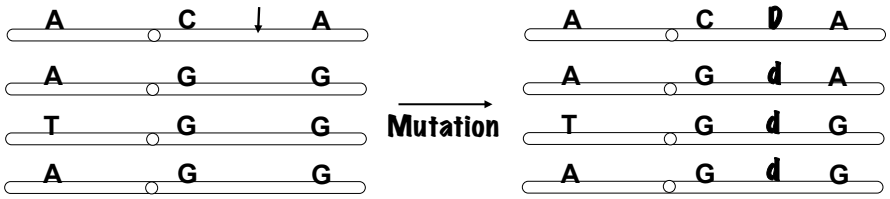
21

Outline

- Genetic markers
- Measures of genetic variation
- Inferences from genetic variation
  - Linkage Disequilibrium
  - Population structure
  - Recombination rate
  - Natural Selection

22

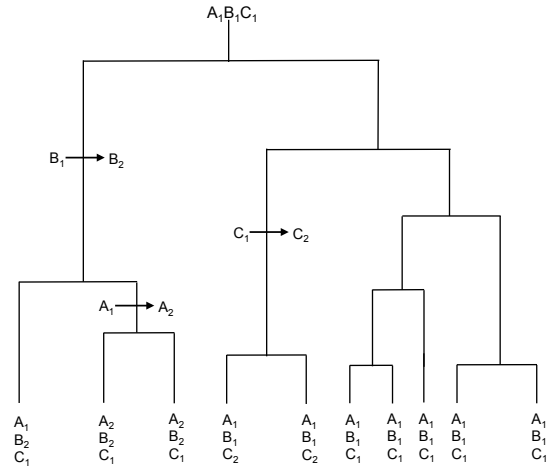
Linkage Disequilibrium



- Mutations arise on a particular haplotype background
- Creates association among alleles
- This is why association studies "work"

23

Coalescent Interpretation of Linkage Disequilibrium



Locus Pair	r^2
A-B	0.58
A-C	0.06
B-C	0.11

Modified from Nordborg and Tavare (2002)

Even in the ***absence of recombination*** pair-wise estimates of LD are extremely variable due to the stochastic history of mutation and coalescent events

24

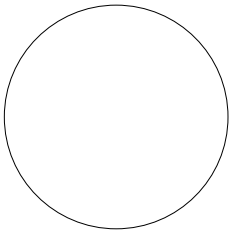
Outline

- Genetic markers
- Measures of genetic variation
- Inferences from genetic variation
  - Linkage Disequilibrium
  - Population structure
  - Recombination rate
  - Natural Selection

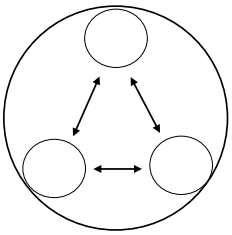
25

Population Structure

- Population Structure: rather than 1 *panmictic* population, the population is subdivided into several mating units



Unstructured population

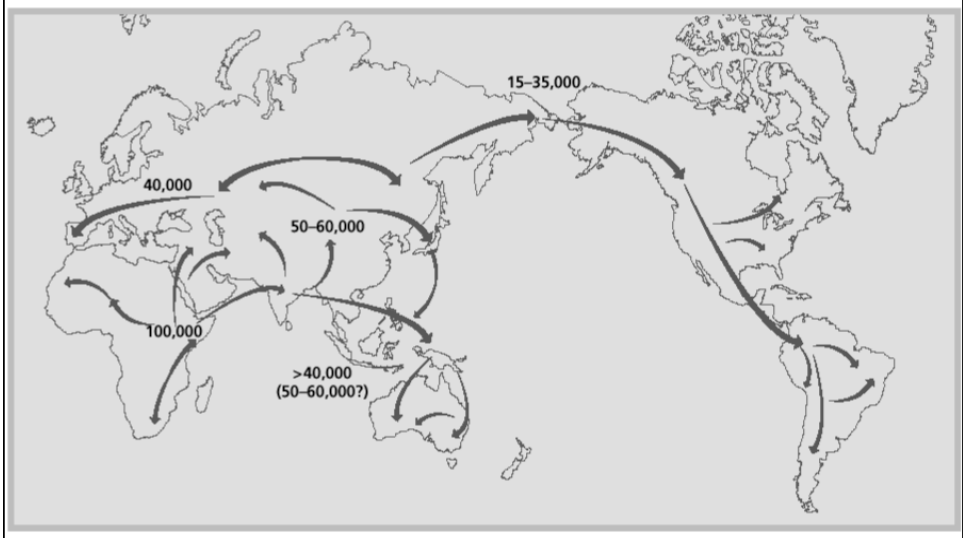


Metapopulation with 3 subpopulations

- Essential consequence is that allele frequencies slowly diverge between subpopulations (i.e., *genetic divergence*)

26

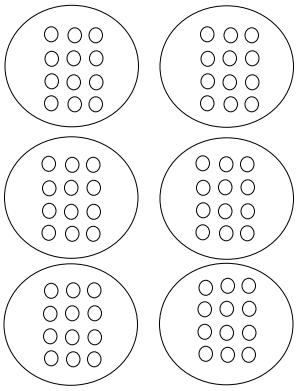
Human Population Structure



BOB CRIMI

Quantifying Population Structure: F_{ST}

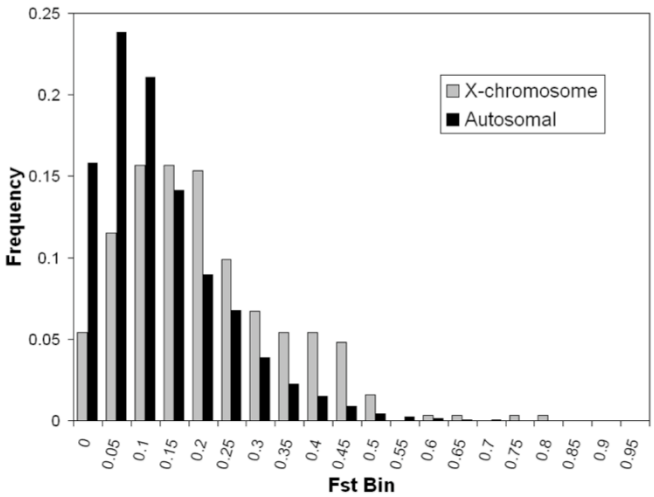
- Statistic that quantifies the allele frequency difference between populations
- Varies between 0 and 1
- Partitions genetic variation into within and between population components



	Variation between populations	Variation within populations
Case 1	100 %	0 %
Case 2	0 %	100 %
Case 3	11 %	89 %

28

Human Population Structure



F_{ST} distributions for 313 X-linked and 11,078 autosomal SNPs computed for 12 populations.

Average X-linked F_{ST} is 0.224

Average autosomal F_{ST} is 0.148

Although humans are "structured", most variation is shared between populations!

29

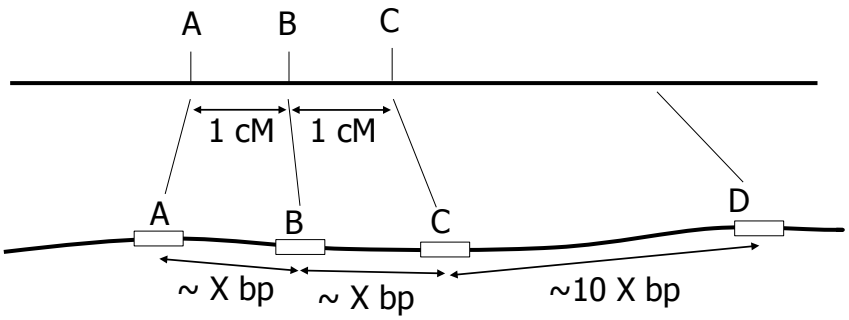
Outline

- Genetic markers
- Measures of genetic variation
- Inferences from genetic variation
  - Linkage Disequilibrium
  - Population structure
  - Recombination rate
  - Natural Selection

30

Integrated Genetic and Physical Maps

genetic map (recombination) units = cM

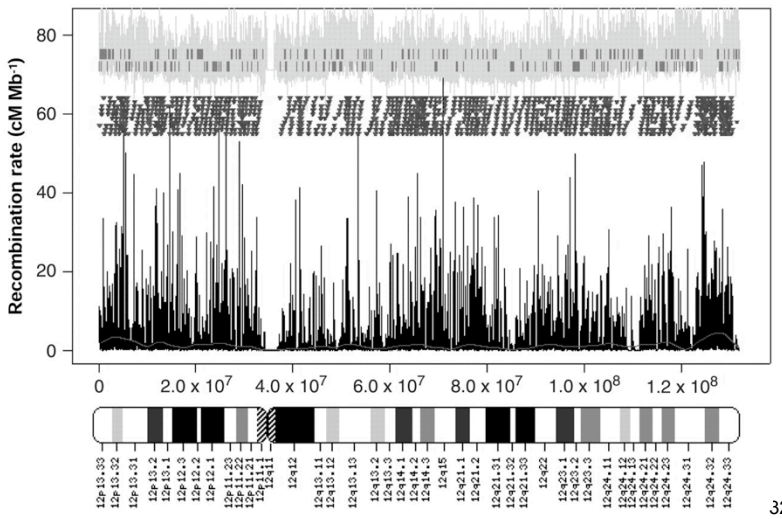


physical map (DNA) units = bp

- Order conserved between genetic and physical maps
- Distance separating loci in genetic and physical maps is proportional

31

Population Genomics of Human Recombination Rate Variation



32

Myers et al (2005) 310:321

Outline

- Genetic markers
- Measures of genetic variation
- Inferences from genetic variation
  - Linkage Disequilibrium
  - Population structure
  - Recombination rate
  - Natural Selection

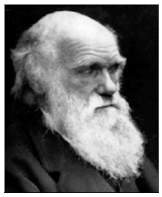
33

Natural Selection

Differential contribution of genetic variants to future generations

"Owing to this struggle for life, variations, however slight and from whatever cause proceeding, if they be in any degree profitable to the individuals of a species, …, will be [preferentially] inherited by the offspring. I have called this principle, by which each slight variation, if useful, is preserved, by the term of Natural Selection."

- Darwin in The Origin of the Species (1859)



34

Types of Natural Selection

Genetic Variation

Positive (Directional)

allele **INCREASES** the fitness and is therefore selected **FOR**

Purifying

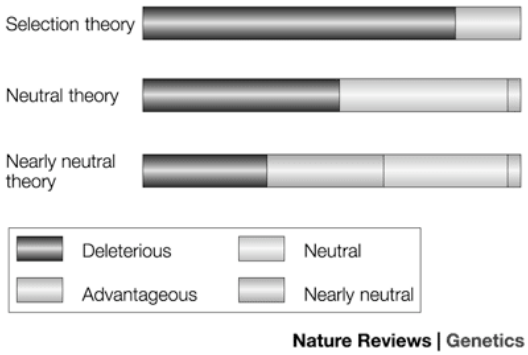
allele **DECREASES** the fitness and is therefore selected **AGAINST**

Balancing

2 or more alleles at a locus are maintained in a population

35

Neutral and Nearly Theories of Molecular Evolution



- Neutral Theory: Kimura (1968) and King & Jukes (1969) proposed that the bulk of genetic variation is *selectively neutral*
- Nearly Neutral Theory: Ohta (1973) proposed that the bulk of genetic variation is either neutral or *nearly neutral*

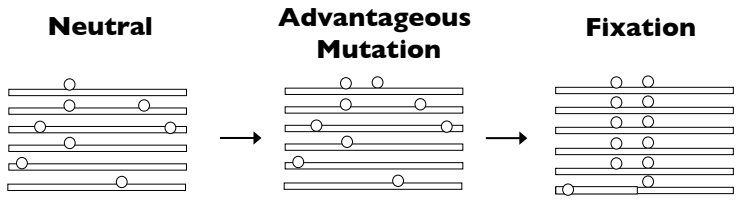
36

Adaptive Evolution: Who Cares?

- Differential contribution of genetic variants
- Identifying targets of selection will:
  - » Mechanisms of evolutionary change
  - » Windows into evolutionary history
  - » Insights into disease

37

Signatures of Positive Selection

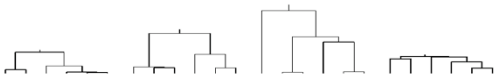


- Imparts "signatures" on patterns of genetic variation:
  - Reduced variation
  - Skew in the site frequency spectrum
  - Increased LD
  - Increased population structure

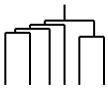
38

Selection is Difficult To Detect

1. Evolution is a noisy process:



2. These signatures can result from additional perturbations to the standard neutral model:

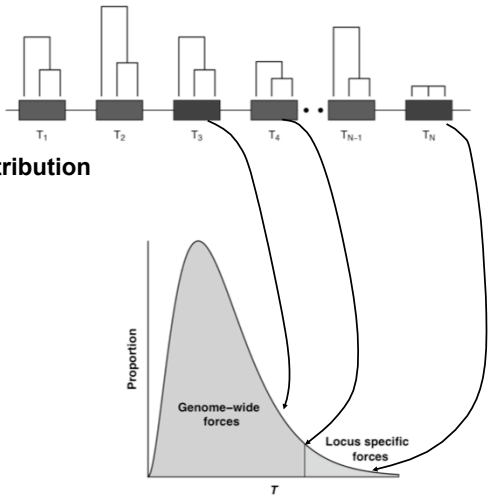


Selection	Positive
Neutral Deviation	Expansion, Bottleneck
Neutrality Tests	D, D*, F* < 0

39

The "Population Genomics" Approach

1. Sample loci and calculate statistic (T_i)

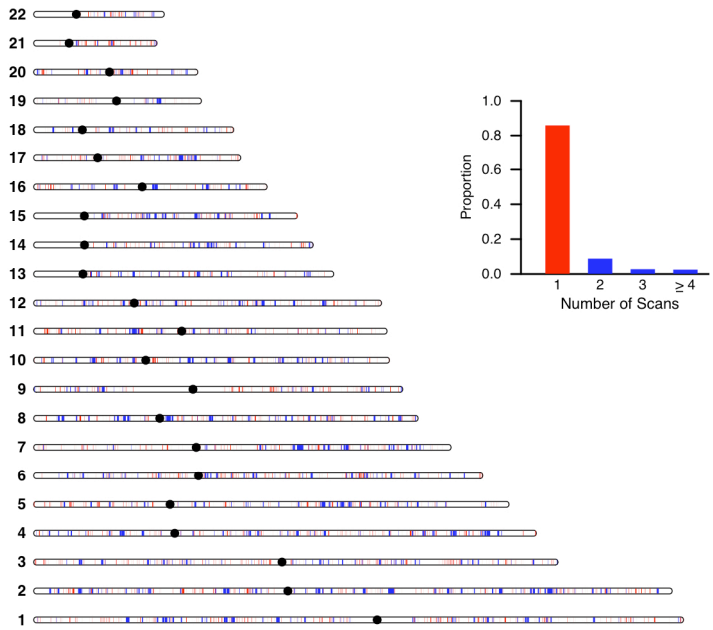


2. Construct empirical distribution

3. Identify "outlier" loci

40

Genome-wide Scans: "A Cottage Industry"



41

Methods for Detecting Selection

1. Divergence Between Species

- d_N / d_S

2. Polymorphism Within Species

- Site frequency based tests (Tajima's D, Fay and Wu's H)
- LD based tests (EHH, iHS, etc)

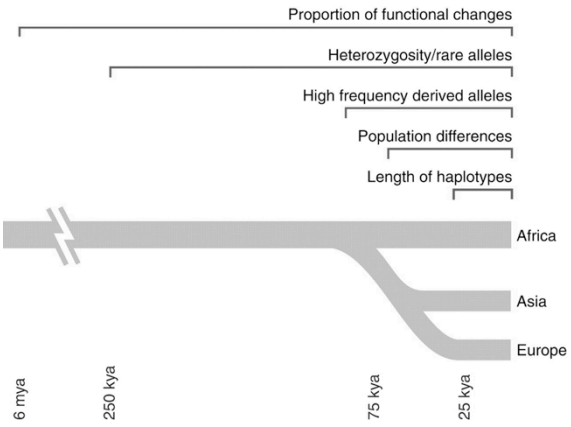
3. Hybrid approaches

- MK test
- HKA test

42

Why So Many Types of Tests?

Statistical tests capture selective events operating at different time scales:



43

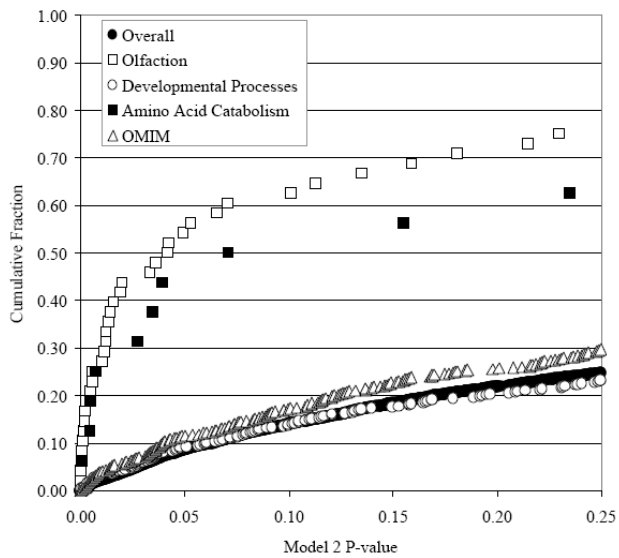
Sabeti et al., Science 312, 1614 -1620 (2006)

dN/dS

- dN – Non-synonymous substitution rate
  - Nucleotide differences that CHANGE the amino acid sequence in orthologous proteins
    CGC (Arg) -> GGC (Gly)
- dS – Synonymous substitution rate
  - Nucleotide changes that do not change the amino acid
    CGC (Arg) -> CGG (Arg)
- dN/dS Ratio
  - dN = dS indicates neutral change
  - dN/dS < 1 indicates constraint/negative selection
  - dN/dS > 1 indicates possible positive selection

44

What Types of Genes Does dN/dS Find?



Clark et al. (2003)

45

Tajima's D

Test Statistic: $D = \frac{\pi - \theta_W}{\sqrt{\mathrm{var}(\pi - \theta_W)}}$ $\theta_w = \frac{S}{\sum_{i=1}^{n-1} \frac{1}{i}}$ $\pi = \left(\frac{n}{n-1}\right)\sum_{i=1}^{S} 2p_i q_i$

Motivation: Under standard neutral model $\theta_W = \pi$ and D = 0

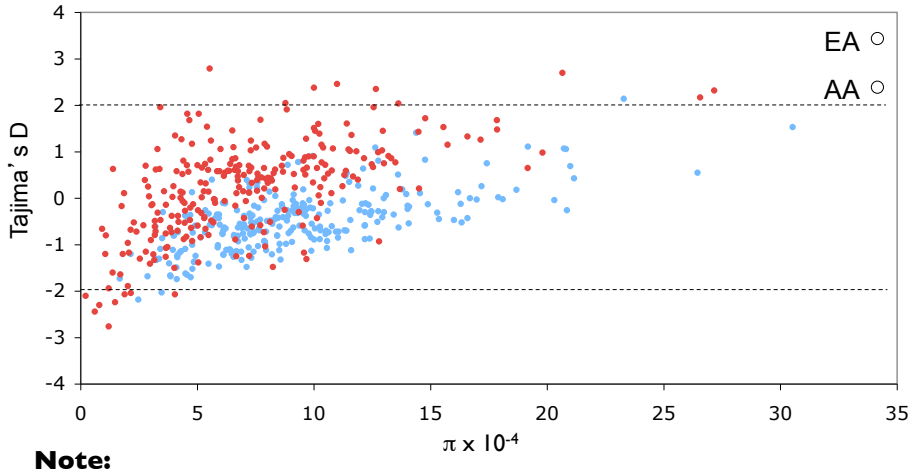
Selective Expectations:

- Positive Selection skews the allele frequency distribution toward lower frequency alleles: $\pi < \theta_w$ and D is < 0
- Balancing Selection causes an excess of intermediate frequency alleles: $\pi > \theta_w$ and D is > 0

Caveat: Significant results are not necessarily due to selection and can also be caused by demographic perturbations

46

Tajima's D For 260 Genes (SeattleSNPs)



Note:

Considerable variation within each population -> Why?

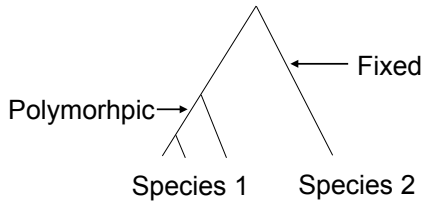
Average Tajima's D in EA is 0.34 and in AA is -0.43 -> Why?

Most genes consistent with neutral evolution

47

MK Test

- Pattern of evolution *within* a species is compared to the pattern of evolution *between* species for two different kinds of sites (SYN and NS aa substitutions)
- Test of within species polymorphism –vs- between species divergence



	Fixed	Polymorphic
ns	X_1	X_2
syn	Y_1	Y_2

Under neutrality the ratio:

$X_1/Y_1 = X_2/Y_2$

48

Software

DNASP: HKA, MK, Tajima D, Fay and Wu H, simulation, more

SITES: Comparative DNA sequence data

ARLEQUIN: Many population genetic analyses

HKA: You guessed it…

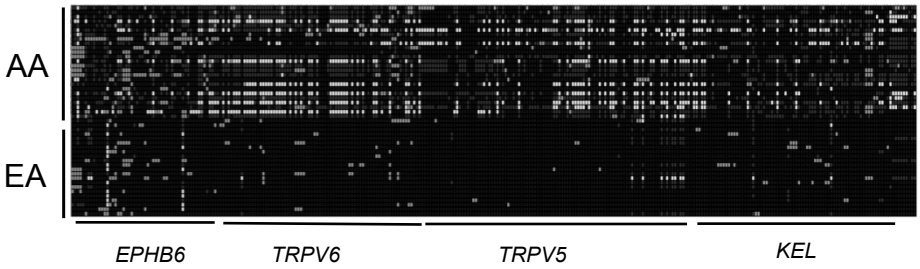
PAML: estimates d_n/d_s ratio using ML for DNA and proteins

MEGA: General evolutionary analyses, tree drawing

49

Chromosome 7q: A Recent Selective Sweep

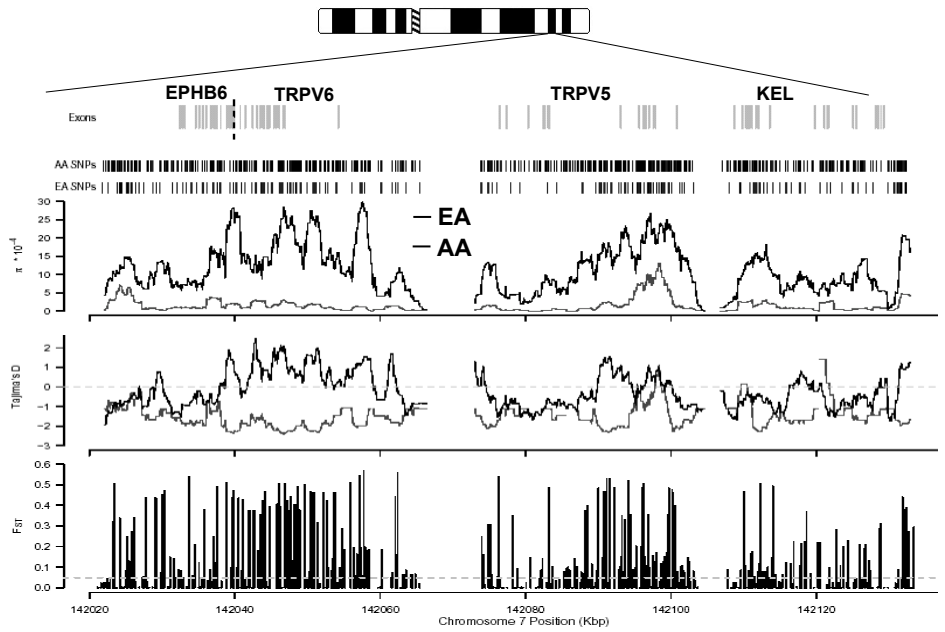
Genetic variation "swept away" in EA sample



- ☐ Homozygous Common Allele
- ☐ Heterozygous
- ☐ Homozygous Rare Allele

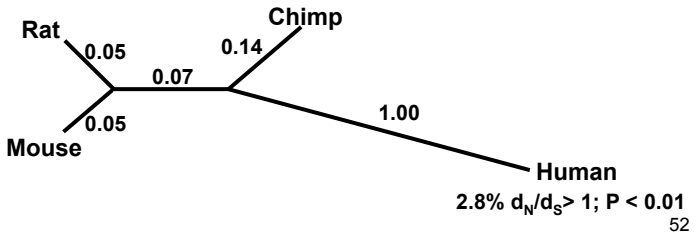
50

Chromosome 7q: A Recent Selective Sweep



TRPV6 Protein Evolution

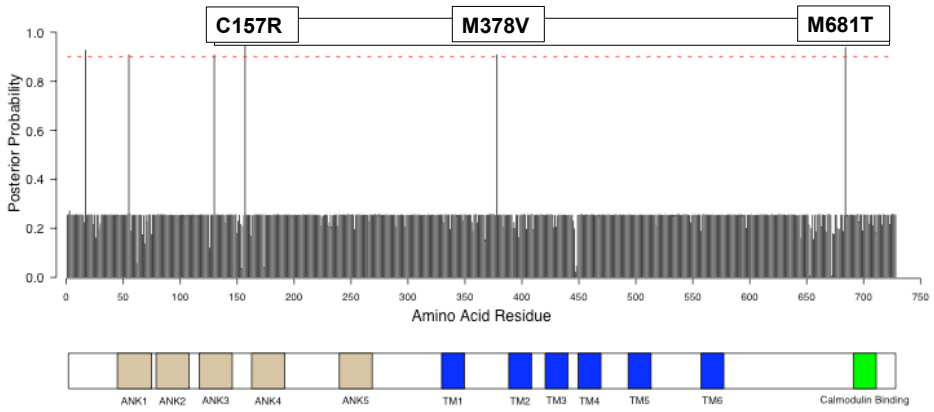
1. Construct a phylogenetic tree of orthologous sequences
2. Codon based maximum likelihood tests of d_N/d_S
   - Allow variation in d_N/d_S between lineages
3. Codon based maximum likelihood tests of d_N/d_S
   - Allow variation in d_N/d_S between lineages AND sites



2.8% $d_N/d_S > 1$; $P < 0.01$

52

Identifying Putatively Selected Residues

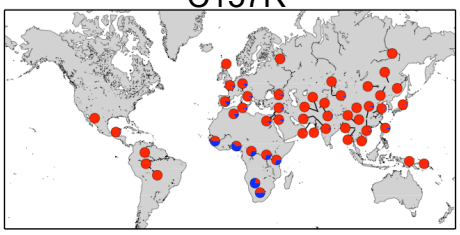


Analysis performed with PAML

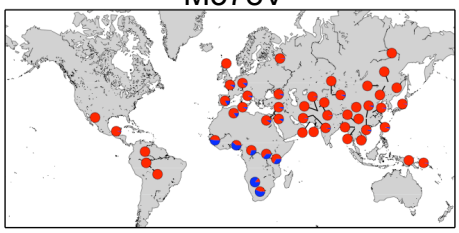
53

Worldwide Distribution of Selected Alleles

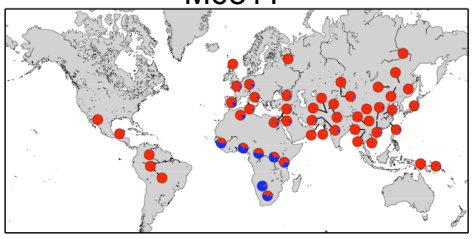
C157R



M378V



M681T

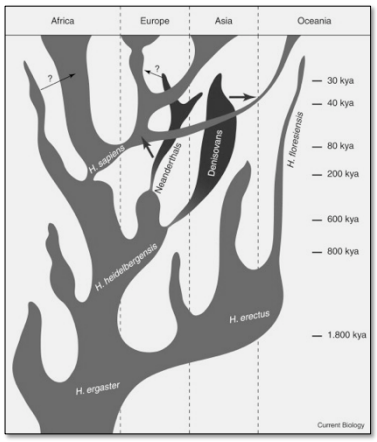


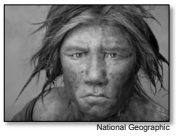
- Derived haplotype (CMM) nearly fixed in all non-African populations
- Time of selection coincident with out of Africa dispersal

Ancestral Amino Acid
Derived Amino Acid

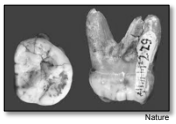
54

Archaic Introgression




National Geographic

RESEARCH ARTICLE
A Draft Sequence of the Neanderthal Genome


Nature

ARTICLE
Genetic history of an archaic hominin group from Denisova Cave in Siberia

But we don't even need bones to "sequence" archaic genomes!

55

A little introgression goes a long way!

Genome

56

Gene Expression

Joshua Akey
Department of Genome Sciences
akeyj@u.washington.edu

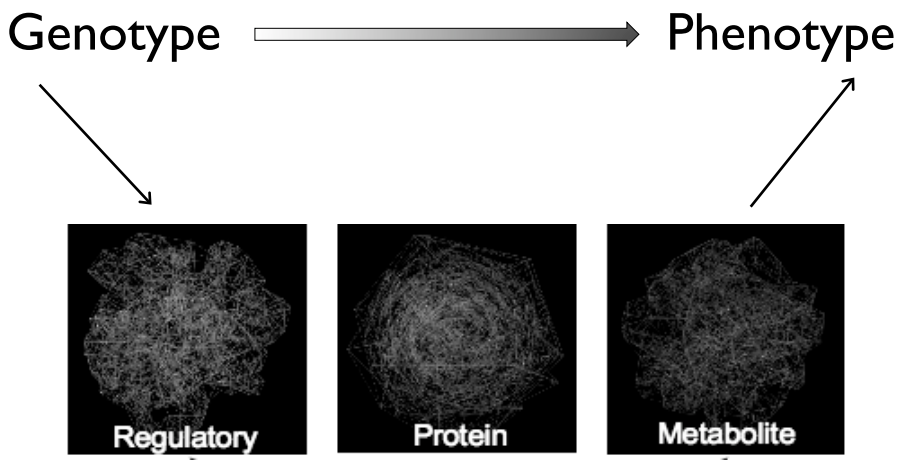
1

Outline

- Overview of gene expression analyses
- cDNA microarrays
- Genetics of gene expression
- RNA-Seq

2

The road from genotype to phenotype



3

Gene expression workflow

1. Ask a good question
2. Design the experiment with adequate statistical power
3. Collect the samples
4. Measure expression levels (qPCR, arrays, RNA-Seq)
5. Extract the raw data
6. Assess the significance of differential expression
7. Perform high level bioinformatics
8. Return to the biology

4

Types of questions

- Tissue/Cell type-specific
  - -e.g. skin cell vs. brain cell
  - -e.g. keratinocyte vs. melanocyte
- Developmental stage
  - -e.g. embryonic skin cell vs. adult skin cell
- Disease state
  - -e.g. normal skin cell vs. skin tumor cell
- Environment-specific (drugs, toxins)
  - -e.g. skin cell untreated vs. treated

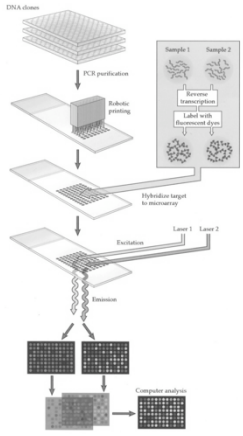
5

Outline

- Overview of gene expression analyses
- cDNA microarrays
- Genetics of gene expression
- RNA-Seq

6

cDNA microarrays



Prepare PCR products

⇩

Print Microarray

⇩

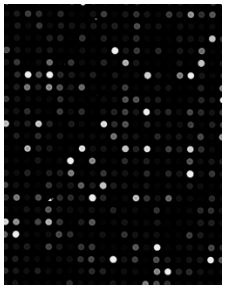
Synthesize labeled target

⇩

Hybridize

⇩

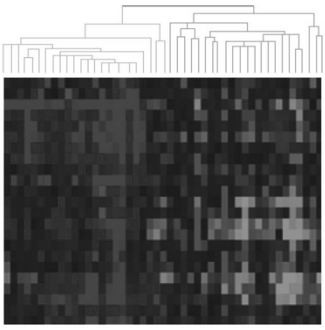
Scan image and analyze.



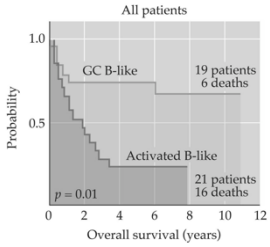
7

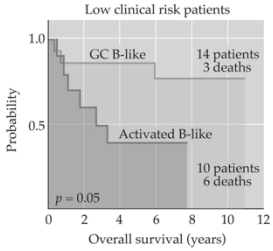
Cancer prediction

Lymphoma class discrimination



Survival prediction



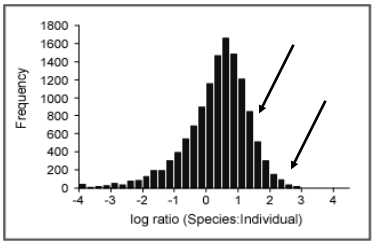


Alizadeh et al. *Nature* 403: 503-511 (2000)

The evolving human brain

Ne = 10,000
T = 500,000 gen

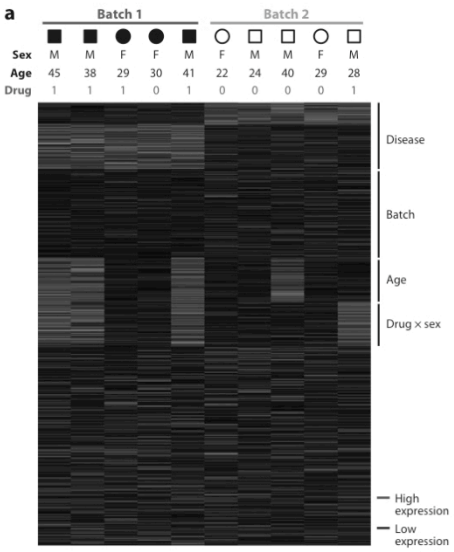
Ne = 50,000
T = 100,000 gen



Enard et al. *Science* 296: 340-343 (2002)

9

A word of caution: study design is critical



10

Outline

- Overview of gene expression analyses
- cDNA microarrays
- Genetics of gene expression
- RNA-Seq

11

Expression Genetics (aka Genetical Genomics)

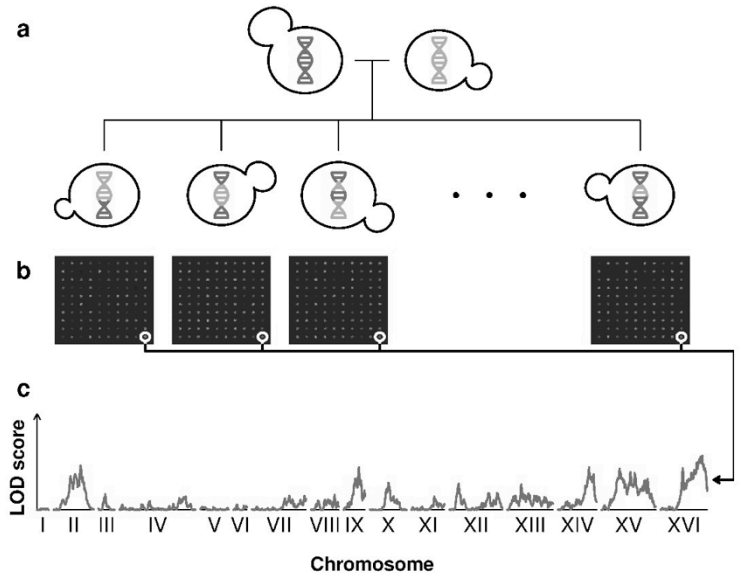
- Systematic study of the genetic architecture of gene expression variation
- Connects genotypes to phenotypes

DNA ⟶ mRNA ⟶ Protein ⟶ Phenotypes

- Source of evolutionary change between species

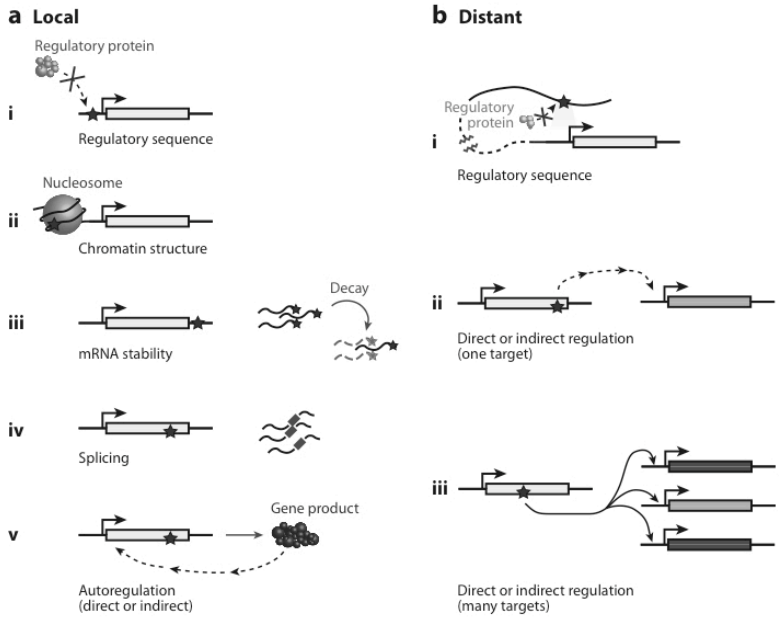
12

A Typical eQTL Study Design

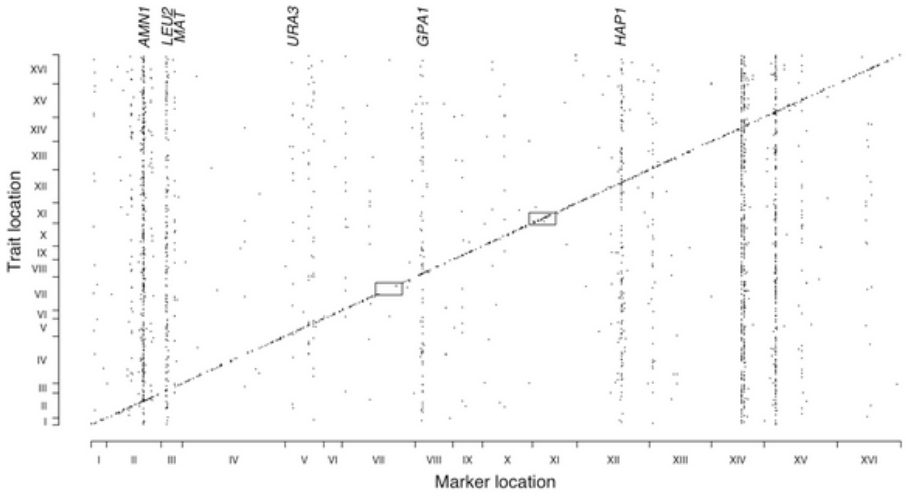


13

Mechanisms of Transcriptional Variation

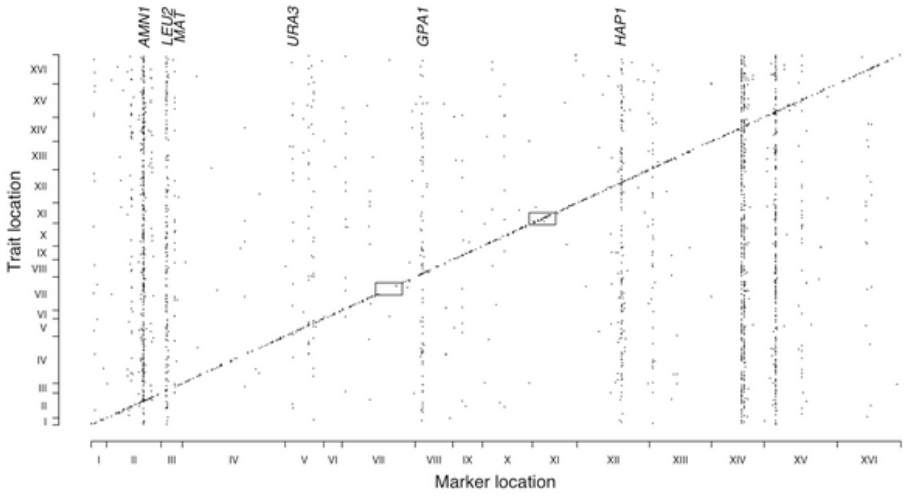


eQTL results in yeast



15

eQTL results in yeast



16

Outline

- Overview of gene expression analyses
- cDNA microarrays
- Genetics of gene expression
- RNA-Seq

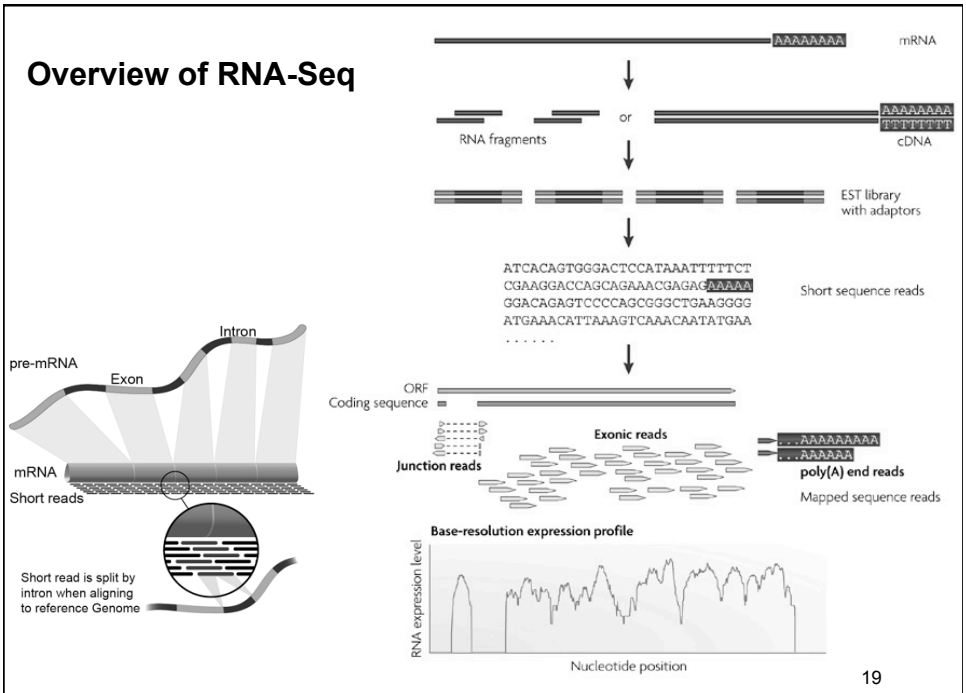
17

Applications of RNA-Seq

- Define transcript structure (start, stop, isoforms, etc.)
- Identifying new transcribed regions
- Expression profiling
- Resequencing to find genetic polymorphisms:
  - SNPs, micro-indels
  - CNVs

18

Overview of RNA-Seq

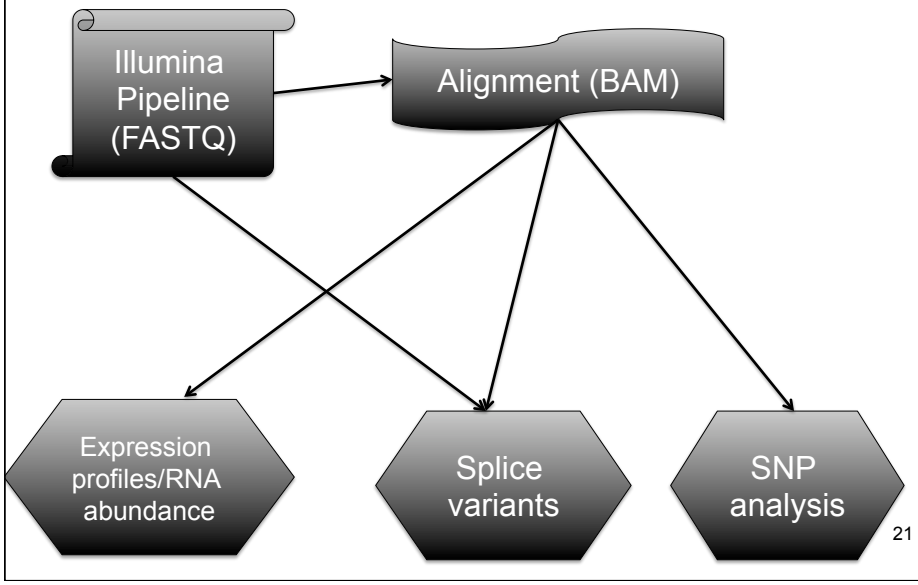


RNA-Seq Technology

- Single-end, paired-end
- Typically 30-400bp reads
- Popular platforms: Illumina, 454, SOLID
- >10 million reads in a single "lane"
- Alignment tools: Bowtie, BWA, Eland, etc.
- Automated pipeline for RNA-Seq:
  – Tophat : for alignment
  – Cufflinks : for calculating expression levels

20

Overview of Analysis



21

Units of Measurement

- **RPKM : Reads per kilobase per million mapped reads**

1kb transcript with 1000 alignments in a sample of 10 million reads (out of which 8 million reads can be mapped) will have
RPKM = 1000/(1 * 8) = 125

- **FPKM : for paired-end sequencing**
  – A pair of reads constitute one fragment

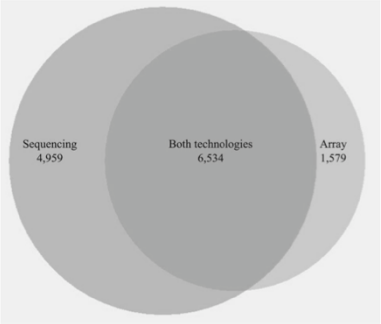
22

RNA-Seq -vs- Arrays

RNA-seq: An assessment of technical reproducibility and comparison with gene expression arrays

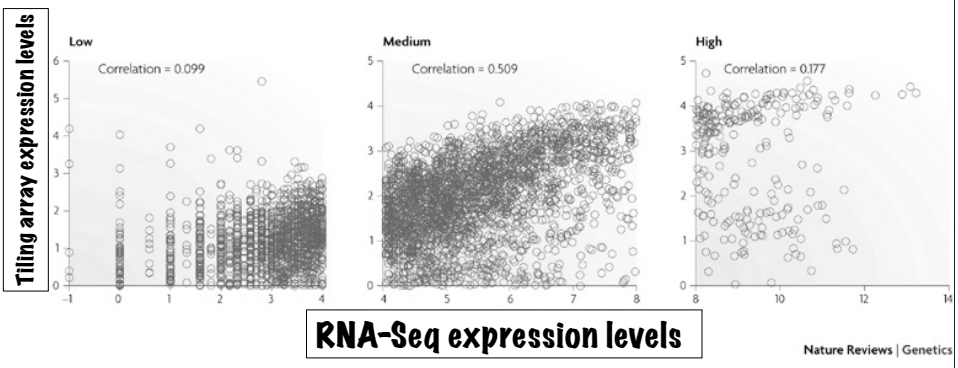
John C. Marioni,[1,6] Christopher E. Mason,[2,3,6] Shrikant M. Mane,[4] Matthew Stephens,[1,5,7] and Yoav Gilad[1,7]

- Correlation of fold change between arrays and RNAseq is similar to correlation between array platforms (0.73)
- Technical replicates are almost identical, no need to run
- Extra analysis: prediction of alternative splicing, SNPs
- Low- and high-expressed genes do not match



23

RNA-Seq -vs- Arrays



24

Advantages of RNA-Seq

- Does not require existing genomic sequence
  - Unlike hybridization approaches
- Very low background noise
  - Reads can be unabmiguously mapped
- Resolution
  - Up to 1 bp
- High-throughput
  - Better than Sanger sequencing of cDNA or EST libraries
- Cost
  - Lower than traditional sequencing
- Can reveal sequence variations (SNPs)

25

Analysis Issues

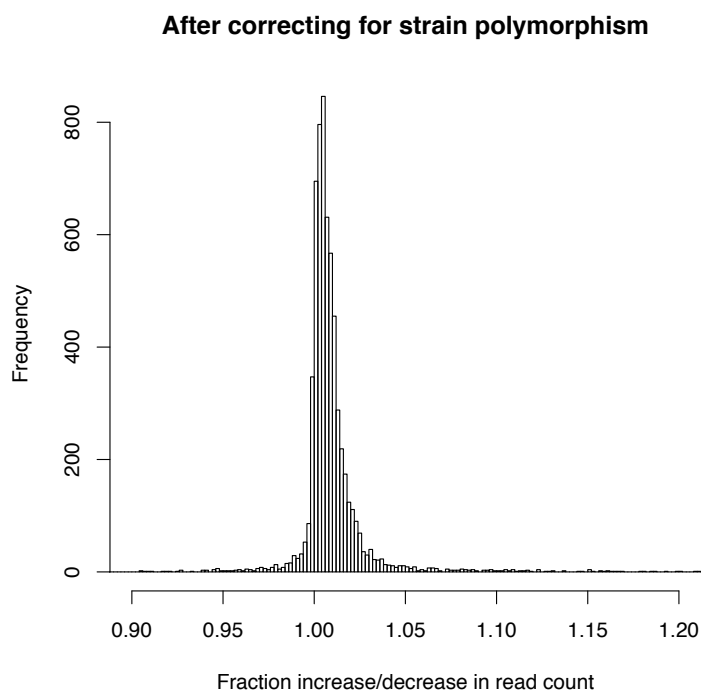
- Practical issues of data storage and manipulation
- How much read-depth?
- Mapping short reads to reference genome
- Best way to test for differential expression?

26

Reference mapping bias

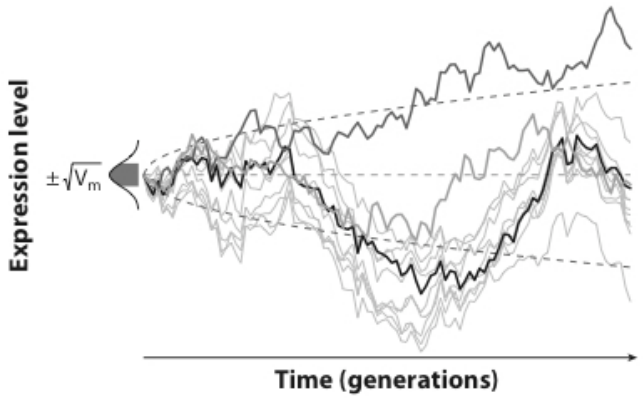
- RNA-Seq in 23 yeast strains
- Performed read mapping in two ways: 1) to a single "reference" genome and 2) to each strain's genome sequence
- Pervasive overall effect
  - 115 genes unchanged for any sample
  - 5742 genes have more reads mapping
  - 828 genes have less reads mapping
  - 8,000,000 more reads map (1% of total)

27



28

Models of Expression Evolution



— Neutral model
— Purifying selection

29

Outstanding Issues

- How biologically important is gene expression variation and how much transcriptional variation results in differences in protein abundance?
- From eQTL to eQTA
- How stable are patterns and characteristics of heritable expression variation in different environments and/or tissues?
- More integrative genomics approaches to interpret, understand, and predict topology, function, and evolution of regulatory networks

30

Summer Institute in Statistical Genetics

2013

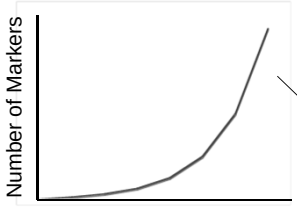
Molecular Genetics and Genomics 4B
Association Mapping

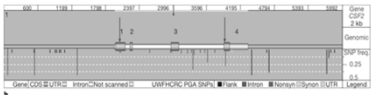
ggibson.gt@gmail.com
http://www.cig.gatech.edu

Outline

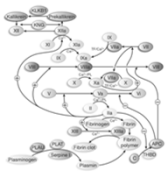
- General overview of association studies
- Sample results
- Three steps to GWAS:
  primary scan, replication, fine mapping

Scale of Association Studies

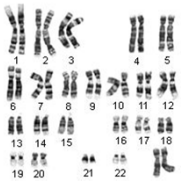




Candidate Gene



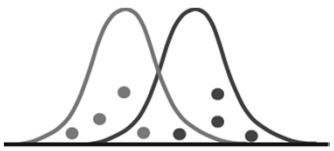
Candidate Pathway



Whole Genome (GWAS)

Priniciple of Association Studies

Individual	Site	Score
1	A T C C G A	9
2	A C T C G A	8
3	A C C A - G	3
4	T T C A G A	5
5	A T C A G A	2
6	A C C C - G	7
7	T C T A - G	4
8	A T C C G A	8

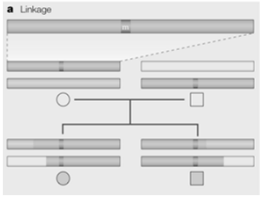


Are the phenotype scores associated with each class of SNP drawn from the same or different distributions ?

Linkage versus Association

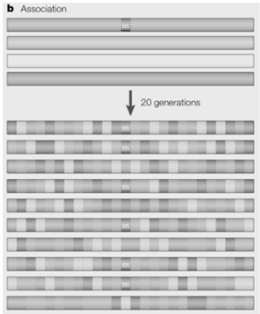
Linkage examines recent recombination events in a pedigree:

- over just several generations
- large chromosomal regions detected
- no information on allele frequency



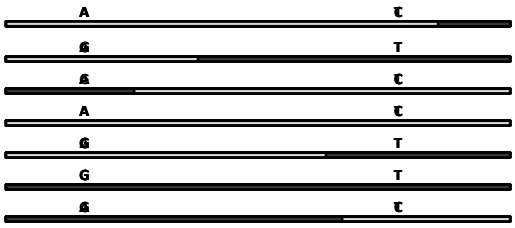
Association examines historical recombination events in a population:

- basically a 10,000 generation pedigree
- resolution to single genes
- estimates effect size and frequency



Cardon & Bell. 2003. Nat. Rev. Genet. 2:91

Why LD happens

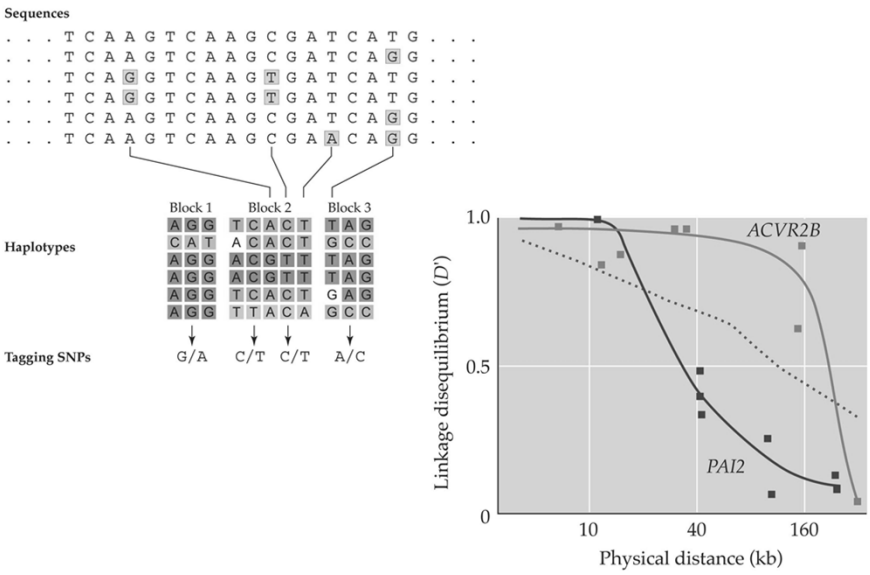


When a mutation occurs, by definition it is only on one chromosome and hence "associated" with the genotypes elsewhere on that chromosome.

Over time, the mutation increases in frequency and becomes a polymorphism. It remains in LD with the genotypes on the chromosome it appeared on.

Eventually recombination breaks up the LD, in proportion to genetic distance.

Haplotypes and Tagging SNPs



Measurement of LD

LD is the non-random association of genotypes.

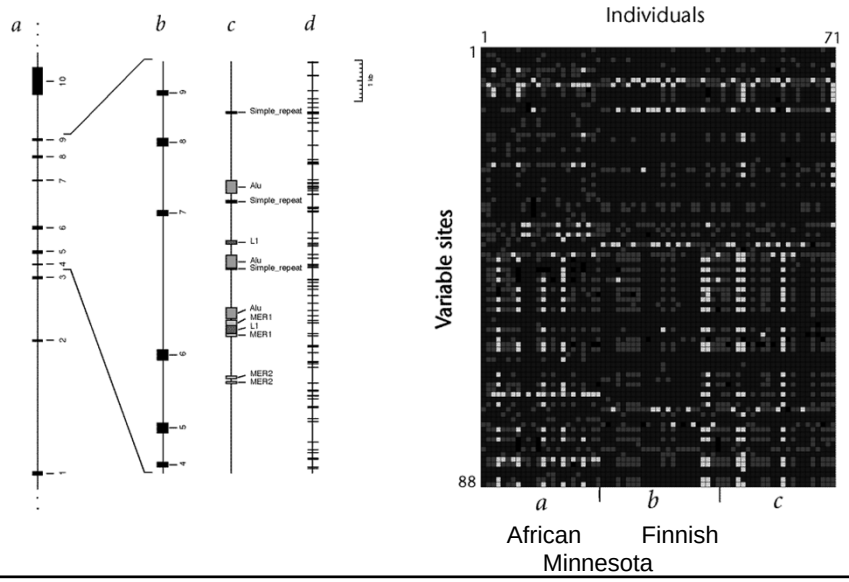
Expected

		AA	AG	GG
		24	48	24
TT	24	6	12	6
TC	48	12	24	12
CC	24	6	12	6

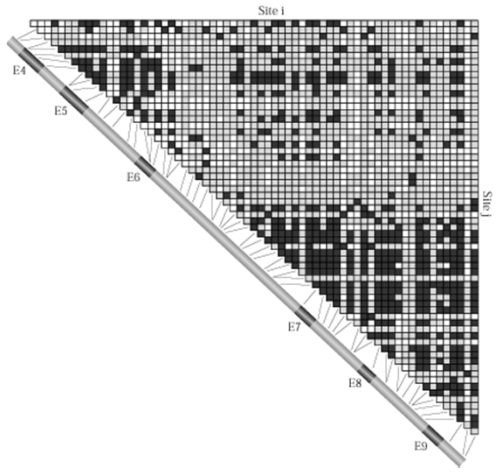
Observed

		AA	AG	GG
		24	48	24
TT	24	24	0	0
TC	48	0	48	0
CC	24	0	0	24

LPL example



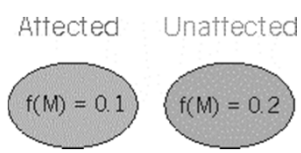
LD plots



Key Parameters for LD Mapping

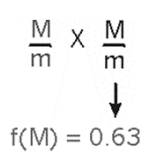
- Polymorphism: Flies 1/30 bp, 10 × > Human 1/kb
- Haplotype structure/LD:
  Fly LD decays over 200 bp, Human LD decays over 100 kb
- Population structure:
  Panmixia and clinality v. Structure and admixture
- Allele frequencies:
  Much more power for common alleles (infinitesimal model)

Case-Control and TDT designs



	Observed		Expected	
	Allele M	Allele m	Allele M	Allele m
Affected	34	278	61	265
Unaffected	69	256	62	269

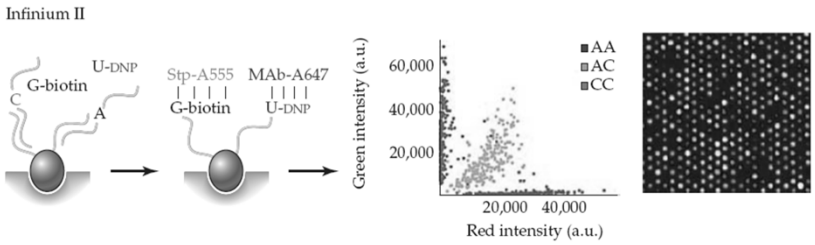
$\chi^2 = 14.0$ P <0.001



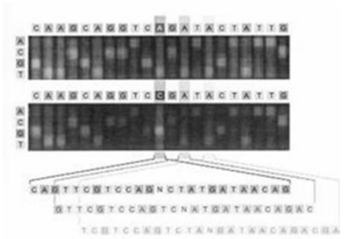
	Transmitted Allele	
	M	m
Observed	78	46
Expected	62	62

$\chi^2 = 8.2$ P < 0.001

Illumina and Affymetrix Genotyping

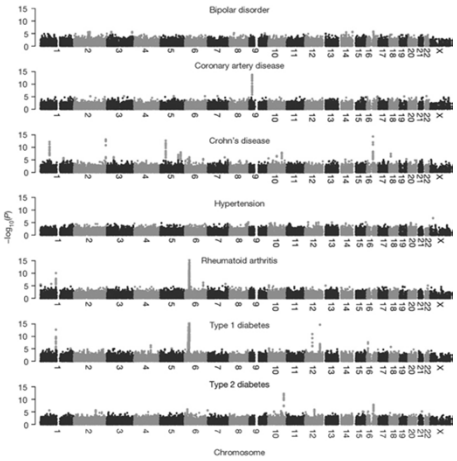


ILLUMINA: Single base extension assay, where the type of base that is incorporated is detected indirectly by hybridization to a fluorescent signal. Performed in parallel on > 800,000 SNPs, each present on between 20 and 80 beads so that the genotype is determined from the average signal.



AFFYMETRIX: Hybridization-based genotyping, based on preferential signal from the probe with a perfect match at the central base of a 25-mer. Performed for both strands provides cross-check of accuracy.

The Genetics of 7 Diseases

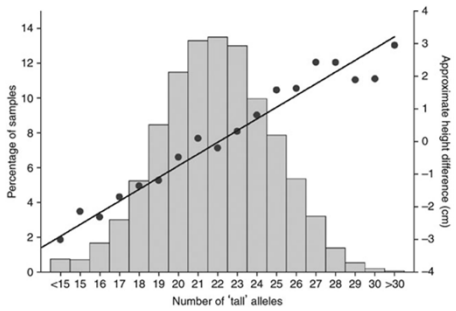


GWAS first appeared 24 months ago, now several new diseases each month

Inflammatory diseases show multiple associations, with some common variants (notably the MHC)

Depression and Hypertension show nothing: likely no variants with a relative risk greater than 1.5

The Genetics of Height



But … they only explain 10% of the variation for height, which is the most heritable human trait!

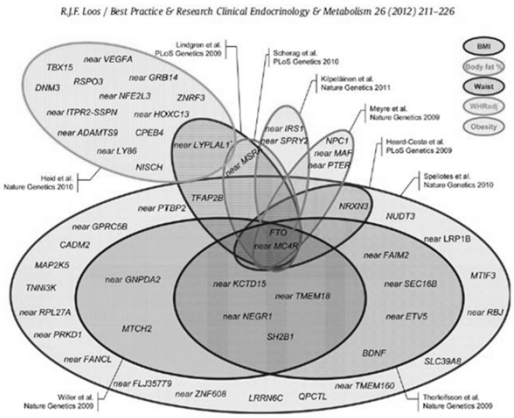
-> Height is infinitesimal

At least 180 loci clearly influence height in combined analysis of 180,000 people

Diverse roles:
Hedgehog signalling
Chromatin structure
Cell cycle regulation
Extracellular matrix deposition

Possible contribution of some of these loci to osteoarthritis, cancer, athleticism

The Genetics of Obesity



Heritability of obesity ~ 60%

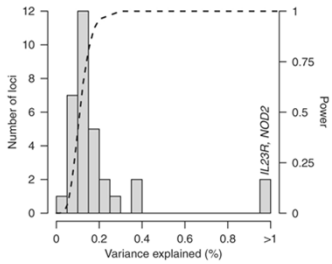
2/3 Americans BMI > 25

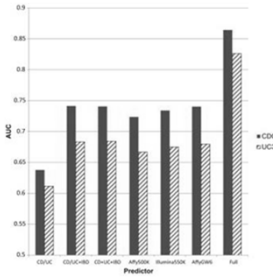
One gene, FTO, is repeatedly associated with BMI, hip circumference and weight, in most human populations

Homozygote classes differ in weight by up to 2 kg

Obesity and Waist/Hip Ratio GWAS show little overlap

The Genetics of IBD





Inflammatory Bowel Disease affects ~1% of adults either as Crohn's or Ulcerative Colitis

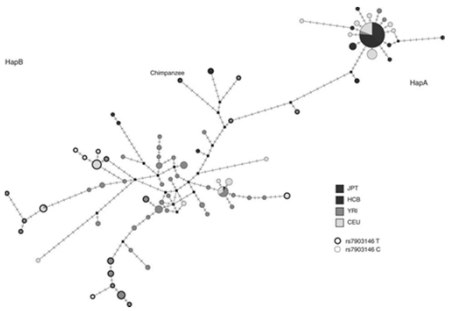
Two genes, IL23R and NOD2, each explain 1% of the variance

Another 30 genes each explain less than 0.25% of it

Together, they explain 20% of the genetic risk, or 10% of the population risk of disease

But, Wei et al (AJHG June 2013) achieve AUC ~ 85% for both, using thousands of Immunochip variants and a machine learning approach

The Genetics of Type 2 Diabetes

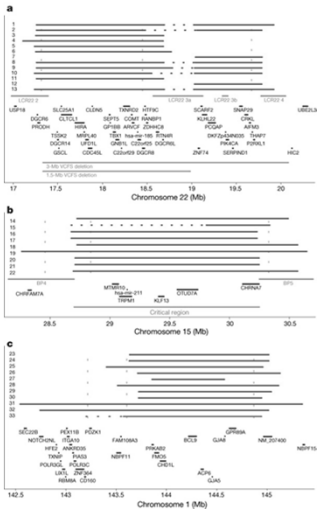


TCF7L2 is the strongest risk locus for T2D in Caucasians

The ancestral allele is the risk allele: it is found in 90% of Africans, 40% of Europeans, and just 5% of Asians

18 other replicated loci have been associated with T2D risk:
1% of Poms have >24 alleles
2% of Poms have <12 alleles
They differ 4-fold in odds of T2D

The Genetics of Schizophrenia



No replicated SNP associations until very recently (now several dozen known)

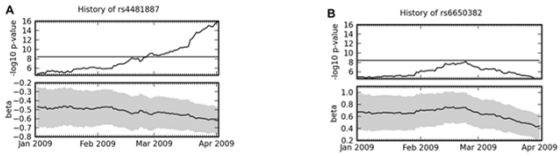
But at least 5% of cases attributable to CNV: copy number variation

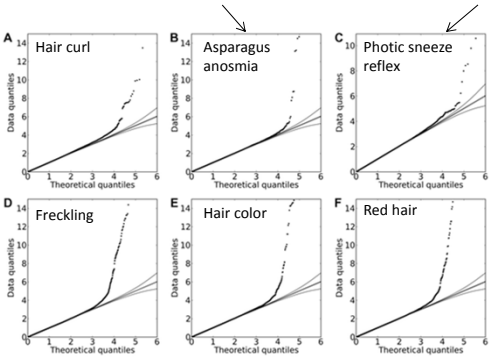
3 major chromosomal deletions of >100kb at frequency <1% almost exclusively found in schizophrenics

Deletions greater than 500kb highly enriched in general, as are some duplications

Similar early results for autism and mental retardation

23andme studies





Other interesting traits:

Endurance Runner vs Sprinter (30% of people change their answer if they know their ACTN3)

Left vs Right Handedness (nothing striking)

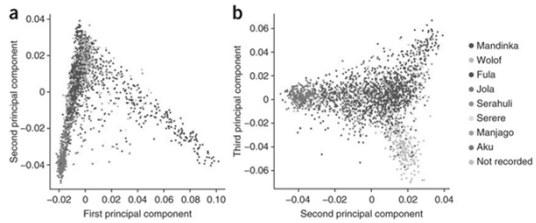
Have you ever needed braces or wisdom teeth surgery?

Breast size (finds breast cancer risk loci)

Hand-clasp dominance …

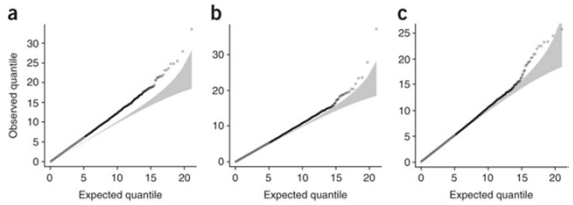
Eriksson et al, (2010) *PLoS Genetics* **6**: e1000993

Genetics of malaria susceptibility



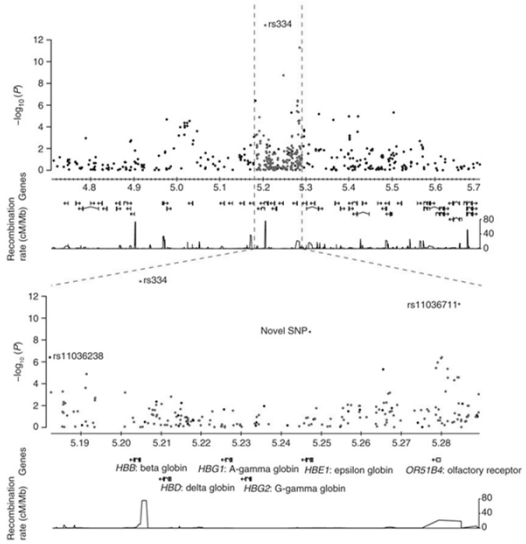
The Gambia

Population structure



Raw Self-report PC adjusted

Imputation and resequencing



Black SNPs on 550K genotyping panel; red ones imputed after sequencing a subset of participants

rs334 is the *HbS* sickle cell variant that explains 2% of malaria susceptibility

Direct typing rs334 yields NLP > 27! The original hit rs11036238 is not strictly GWAS

Some references

- The Wellcome Trust Case Control Consortium (2007) Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature* **447**: 661-678
- Scuteri , A. *et al* (2007) Genome-wide association scan shows genetic variants in the *FTO* gene are associated with obesity-related traits. *PLoS Genetics* **3**: e115
- Weedon, M.N., *et al* (2008) Genome-wide association analysis identifies 20 loci that influence adult height. *Nature Genet*ics **40**: 575-583
- Barrett, J.C. *et al* (2008) Genome-wide association defines more than 30 distinct susceptibility loci for Crohn's disease. *Nature Genetics* **40**: 955-962
- Zeggini, E. *et al* (2008) Meta-analysis of genome-wide association data and large-scale replication identifies additional susceptibility loci for T2D. *Nature Genet*ics **40**: 638-645
- The International Schizophrenia Consortium (2008) Rare chromosomal deletions and duplications increase risk of schizophrenia. *Nature* **455**: 237-241
- He, C. *et al* (2009) Genome-wide association studies identify loci associated with age at menarche and age at natural menopause. *Nature Genetics* **41**: 724-727
- Jallow, M. *et al* (2009) Genome-wide and fine-resolution association analysis of malaria in West Africa. *Nature Genetics* **41**: 657-665
- Ioannidis, J.P., G. Thomas and M.J. Daly (2009) Validating, augmenting and refining genome-wide association signals. *Nature Reviews Genetics* **10**: 318-329

Summer Institute in Statistical Genetics

2013

Molecular Genetics and Genomics 5
Personalized Genomics

ggibson.gt@gmail.com
http://www.cig.gatech.edu

Genome sequencing success story

Diagnosed at age 5 with dopa-responsive dystonia

Worsening respiratory and neuromuscular disease not responsive to dopamine precursor therapy

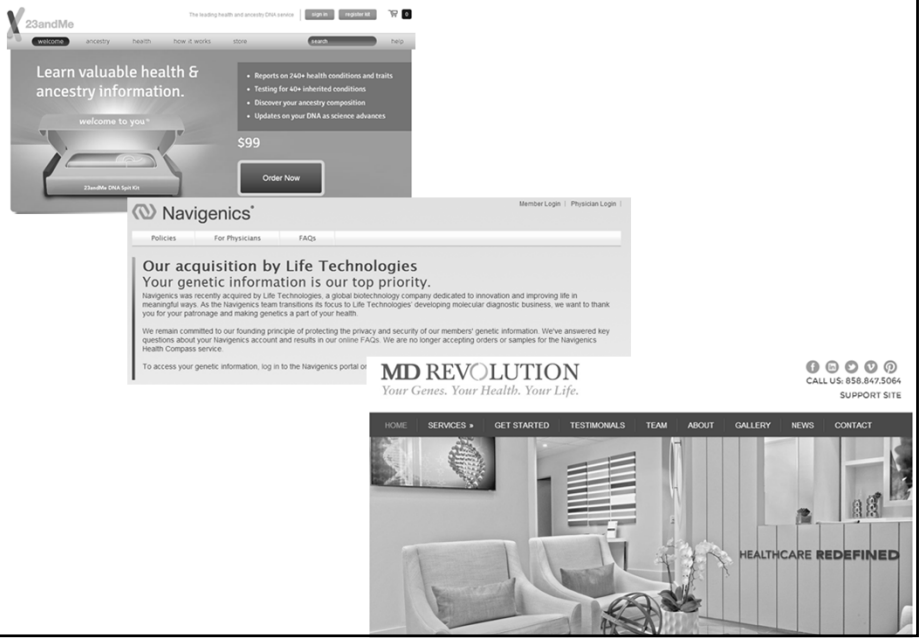
WGS shows mutation in SPR "sepiapterin reductase" gene

5-HT serotonin precusor supplementation had immediate impact

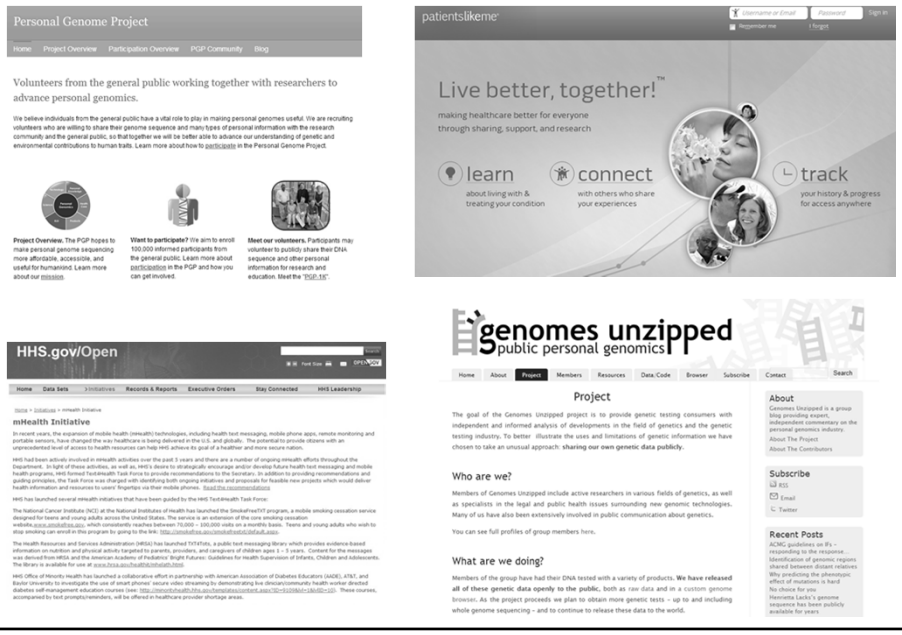
The Beery twins

Bainbridge, M. N. *et al.* (2011) *Sci. Transl. Med.* **3,** 87re3

Some Personal Genomics Companies



Some Public Initiatives



Types of Predictive Health Information

1. Clinical, Survey, and Anthropomorphic Data
2. Family History
3. Genotypes and Genome Sequence
   - Common variants
   - Rare mutations
4. Gene Expression Profile
5. Metabolite Profile
6. Microbiome Profile

Types of things you might like to know about

1. Your Ancestry
2. Things that are just plain interesting
   - eye and hair color, handedness, unusual features
3. Things that you might be able to act on
   - BRCA, warfarin resistance, metabolic deficiencies
4. (Scary) things that you cannot do anything about
   - Predispositions to late onset disease
5. Things you might want to know about before having kids
   - Mendelian disease carrier status
6. Things you might consider selecting for (or against) in the future
   - Intelligence, athletic ability, good looks
   - OR against depression, complex disease risk

Variation in 12 genomes

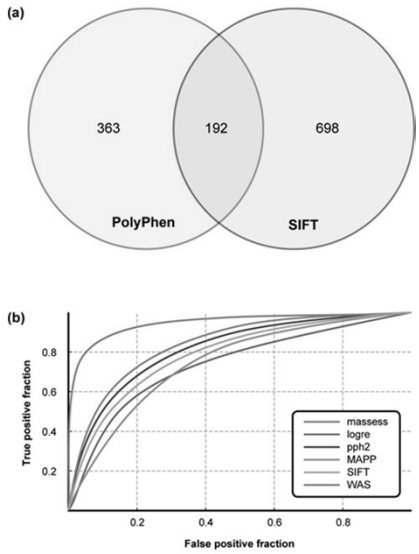
Table 1. Summary of genetic variations in genome sequences of twelve individuals.

Subject ID	Eth, sex	Total Variants (>q20)			Coding variants (based on Gencode v7)								Homozygous nsSNPs (based on dbSNP build 137)			
		SNPs	Indels	SVs	SNPs				Indels			SVs	#Known nsSNPs	#Unique genes	#*de novo* nsSNPs	#Unique genes
					S	MS	NS	Splice Overlap	Indels FS	Indels NFS	Overlap	SVs Overlap				
1	Afr, F	4513763	733596	4251	14793	14039	72	98	381	342	137	37	88	77	2	2
2	Afr, F	4472988	754399	4545	14500	13712	66	106	393	335	147	55	71	63	3	3
3	Afr, F	4287739	722922	4447	13755	13166	84	79	374	301	120	43	77	71	3	2
4	Afr, F	4443799	746111	4368	14488	13874	73	104	366	338	142	40	58	56	1	1
5	Cau, F	3734820	645032	3977	11929	11745	62	90	343	307	123	43	57	40	2	2
6	Cau, F	3691337	633475	4114	11757	11457	56	90	317	280	106	49	52	45	none	none
7	Cau, F	3691270	632544	4033	11912	11488	65	71	279	304	116	37	50	44	4	4
8	Cau, F	3722234	641792	4197	11887	11434	64	76	303	299	125	41	55	42	none	none
9	Cau, M	3647944	590064	3828	11619	11255	54	76	311	281	95	38	60	50	2	2
10	Cau, M	3643046	597363	4011	11814	11480	61	85	289	287	109	31	82	65	2	2
11	Cau, M	3650690	602744	3916	11560	11285	60	80	342	280	112	32	72	65	5	5
12	Cau, M	3701558	639005	4739	11842	11708	60	81	334	290	118	37	75	64	5	5
													#Total = 797	#Unique = 575	#Total = 29	#Unique = 25

Example Rare Variants

SNP type	Gene	Position, Base change (AA change)	rsID (%MAF EA/AA/All)	Disease [prediction counts]	Grantham score	Protein stability change	Site annotations
Disease-causal	*ATP6V0A4*	7:138417791 *A-->G (M580T)*	rs3807153* (4.8/18.5/9.4)	Distal renal tubular acidosis with preserved hearing (RTADR) [Del count: 2; Con count: 3]	81	Neutral	TRANSMEM
	MTMR2	11:95569448 *T-->C (N545S)*	rs558018 (0.02/3.9/1.3)	Charcot-Marie-Tooth disease type 4B1 (CMT4B1) [Del count: 2; Con count: 3]	46	Decrease	DOMAIN
	APOE	19:45412079 *C-->T (R176C)*	rs7412 (5.6/8.7/6.6)	lipoprotein glomerulopathy (LPG) [Del count: 5; Con count: 3]	180	Decrease	REPEAT
	BMP15	X:50658966 *G-->A (A180T)*	rs104894767 (1.4/0.3/1.0)	premature ovarian failure type 4 (POF4) [Del count: 0; Con count: 1]	58	Neutral	PROPEP
Probable pathogenic	*FRZB*	2:183699584 *G-->C (R324G)*	rs7775 (8.8/28.4/15.4)	Osteoarthritis type 1(OS1) [Del count: 1; Con count: 3]	125	Neutral	-
	HABP2	10:115348046 *G-->A (G534E)*	rs7080536 (3.9/0.7/2.8)	[Del count: 5; Con count: 3]	98	Decrease	DOMAIN
	HNF1A	12:121416650 *A-->C (I27L)*	rs1169288 (33.5/12.1/26.2)	Diabetes mellitus insulin-dependent type 20 (IDDM20) [Del count: 1; Con count: 3]	5	Neutral	REGION (Dimerization)
	XYLT1	16:17564311 *C-->A (A115S)*	rs61758388 (-/-/1.7)	[Del count: 0; Con count: 3]	99	Neutral	TOPO_DOM
	CYP2A6	19:41354533 *A-->T (L160H)*	rs1801272 (2.5/0.5/1.8)	[Del count: 1; Con count: 3]	99	Decrease	-
	ADA	20:43255220 *T-->C (K80R)*	rs11555566 (6.3/6.8/6.5)	Severe combined immunodeficiency autosomal recessive T-cell-negative/B-cell-negative/NK-cell-negative due to adenosine deaminase deficiency (ADASCID) [Del count: 1; Con count: 3]	26	Decrease	-

Predicting Pathogenicity



Predictions on the Venter Genome for 7,534 non-synonymous SNPS

WAS is a consensus deleterious score (Condel) averaging the others

Katsanis and Katsanis (2013) *Nat Rev Genet*. 14: 415-426

Association-Adjusted Consensus Deleterious Classification

		Features of SNP				Subject ID, Ethnicity, Sex												Total SNPs
		Disease-causing	Deleterious count ≥ 3	In disease genes	In trait genes	1	2	3	4	5	6	7	8	9	10	11	12	
						Afr, F	Afr, F	Afr, F	Afr, F	Cau, F	Cau, F	Cau, F	Cau, F	Cau, M	Cau, M	Cau, M	Cau, M	
Category	1	X				1	none	none	none	1	none	none	1	1	none	1	none	5
	2A			X		8	7	6	5	6	5	11	7	12	7	10	9	93
	2B		X	X		1	none	none	none	2	2	4	1	2	none	2	none	14
	3A				X	9	24	17	7	6	10	9	10	5	13	16	10	136
	3B		X		X	1	5	none	none	1	1	3	4	2	2	none	2	21
	4		X			5	5	8	6	none	5	3	7	4	6	5	4	58
	5			X	X	14	21	20	10	8	12	9	11	13	18	20	14	170
	6					68	40	49	42	46	32	33	33	40	56	45	55	539
Total #unique nsSNPs						88	71	77	58	57	52	50	55	60	82	72	75	797

ACMG Categories:

1: Known disease-causing mutation
2: Likely disease causing
3: Possibly disease causing
4: Likely not disease causing
4: Not disease causing
6: Variants of unknown clinical significance

Preeprem and Gibson (2013) Submitted

Levels of Evidence

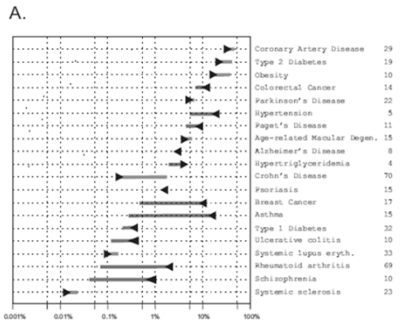
1. Previously ascribed clinical function
2. Bioinformatic prediction from protein structure or attributes
3. Evolutionary conservation
4. Experimental validation (animal models, cellular manipulation, in vitro studies)

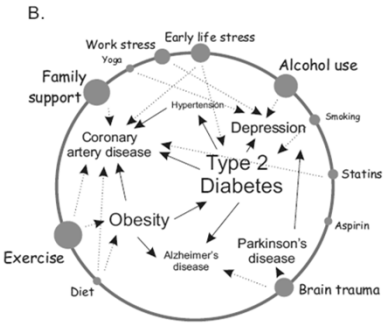
5. BUT, it is easy to get trapped in a genetically deterministic worldview:
   - even Mendelian variants have incomplete penetrance
   - expressivity is modified by genetics and environment
   - deleterious to the protein is not necessarily deleterious to the organism

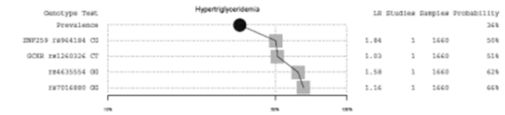
6. We do not have parallel methods for evaluating function of regulatory variants

Bottom line: N = 1 genetics is in its infancy

Common Variant Risk-o-Grams

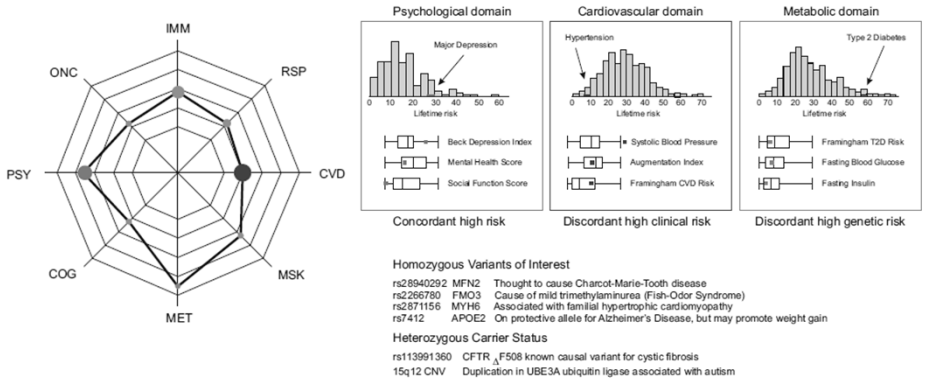






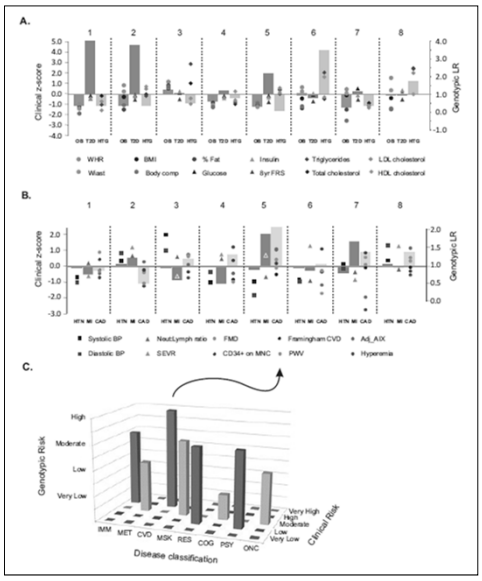
Ashley et al (2010) *Lancet* 375: 1525-1535

A Sample Predictive Health Profile



Patel, Sivadas et al (2013) *Genome Medicine* 5: 58

Focusing on Individual Attributes



Metabolic Risk

Cardiovascular Risk

Gridiron Plot

Types of Genetic Test

Table 1 | **Factors considered in selecting a genetic test**

Test	Description	Example	Embryo or blastocyst (pre-implantation genetic diagnosis)	Fetus (prenatal testing)	Child	Adult
Newborn screening	Targeted tests for recessive genetic disorders	Phenylketonuria, cystic fibrosis, sickle-cell anaemia	Not applicable	Not applicable	Tests provided at birth vary by country and state or region	Not applicable
Diagnostic testing	Confirmatory test or differential diagnosis testing for a symptomatic individual	Skeletal dysplasias, thalassaemias, craniosynostoses	Specimen type and limited available amount for sampling may restrict platform selection (for example, WES or WGS versus SNP or STR typing)		Where treatment is desired, turnaround time may restrict platform selection	
			Turnaround time necessary may restrict platform selection			
Carrier testing	Targeted testing for asymptomatic individuals potentially carrying one or more recessive mutation	Cystic fibrosis, thalassaemias, Tay–Sachs disease	Applied typically for rare disease but applicable for other familial mutations		Carrier testing of minors is considered in the context of individual paediatric cases[164,165]	According to standard of care
Predictive testing	Tests for variants causing or associated with diseases or disorders with a hereditary component, usually with adult-onset symptoms	Most cancers, cardiovascular disease, diabetes	Some have discouraged genetic testing of asymptomatic minors for adult-onset conditions			According to standard of care
Pre-symptomatic testing	Tests for variants causing or associated with diseases or disorders known to be inherited in the family, often with adult-onset symptoms	Huntington's disease, haemo-chromatosis, Alzheimer's disease	Some have discouraged genetic testing of asymptomatic minors for adult-onset conditions[152,153]			According to standard of care
			Interpretation of VUSs will depend on presenting phenotypes in the family			
Pharmaco-genetics	Targeted tests for variants associated with pharmaceutical dosage choice or adverse reactions	DNA tests for abacavir, warfarin, carbamazepine	Application not currently conducted but theoretically feasible	Application not currently conducted, but conceivably applicable for screening treatment approaches *in utero*	Pharmacogenetic testing is considered in context of individual paediatric cases[166]	According to standard of care

Katsanis and Katsanis (2013) *Nat Rev Genet*. 14: 415-426

Prediction and Heritability

For complex traits, the upper limit of the correlation between an as yet unobserved phenotype and a genetic predictor is h, the square root of the heritability (h2). For most human complex traits, heritability h^2 falls in the range of 40-80%, so an upper limit of precision of prediction of about 60-90%.

Imagine if we had a perfect genetic predictor for height (e.g. all causal variants known without error) then the prediction error for any individual would be $7*\sqrt{(1-0.8)} = 3.1$cm, assuming that $h^2=0.8$ and that the standard deviation for height is 7cm. So we would predict someone's height with a 95% confidence interval of about +/- 6cm. For IQ, supposing “optimistically” that we could capture 50% of the phenotypic variance with genotypes, the corresponding confidence interval of the prediction will be around 20 IQ points.

That’s about the same as we get from knowing the parental IQ. So what!

Prediction and Classification

However, it is not necessarily about prediction. What if it is about changing health behaviors? Then classifying risk might help people focus.

The top and bottom 10% of any sample are generally 1.6 standard deviation units above or below the average. For height that translates to $1.6*7*\sqrt{(0.8)} = 10$cm, so individuals at the genetic tails will tend to be 20cm apart. In fact, even with just 20 variants, empirical data shows they average 6cm, in line with theory.

In a family of 10, genetic predictors that capture half the heritability would separate the tails from the average by $1.6*s.d.*\sqrt{(0.5*(0.5*h^2))}$ since half the heritability is within the family. Siblings are thus expected to differ quite a bit in their predispositions to diseases and traits that might be of interest.

Effective reduction of chronic disease deaths by 2%, for example by targeting treatment to those most at risk, would potentially save economies a much greater percentage of lost national income.