Analysis of Variation

BIOL 426/626
Approaches to Molecular Biology

Reminder…

Next Tuesday you'll be doing ***peer review*** of each others term papers (first draft).

Remember that a completed peer review is worth 5% of your grade for the paper.

Bring two printed copies to class on Tuesday for peer review and send one to me by E-mail.

Class 17: Analysis of variation

- **Learning Goal**
  - To understand the methods for determining the extent of variation in sequence and expression among individuals in response to physiological or genetic differences
- **Learning Objectives**
  - List and define the types of genetic variation that can occur among individuals in a population
  - Define restriction fragment length polymorphisms (RFLPs) and explain how they are analyzed in genotyping and genetic fingerprinting
  - Define short tandem repeats (STRs) and variable number tandem repeats (VNTRs) and explain how they are analyzed
  - Explain the use of microarrays in analyzing single nucleotide polymorphisms (SNPs)
  - Describe how these technologies are used in analyzing human genomes through linkage analysis and genome wide association studies (GWAS)
- **Reading assignment:**
  - Dale From Genes to Genomes: Chapt 9

Phenotypic variation ⇒ genotypic variation

Phenotypic variation among humans

Where does phenotypic variation come from?

- With your group, talk about two topics:

  1. What kinds of changes to the structure of genomic DNA could cause changes in phenotype?

  2. Where in the genome might you expect to find these changes and what aspect of the expression of genes might they alter?

- Talk about these for about 10 minutes

Kinds of mutations causing phenotypic variation

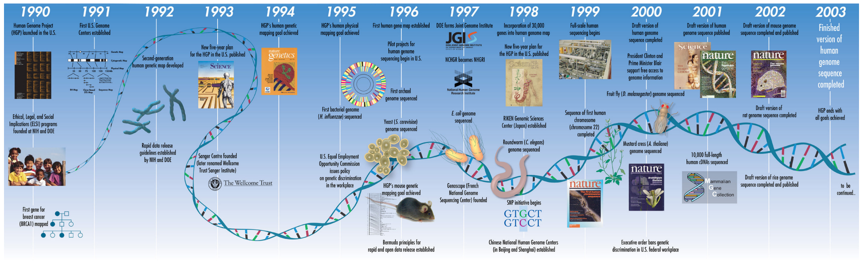
- SNPs (single nucleotide polymorphisms)
- Indels (insertion/deletions)
- Tandem duplications
- Translocations
- Inversions
- Loss of heterozygosity (LOH)
  - Resulting from large deletions or loss of chromosomes

Where might they happen? What aspect of gene expression?

- Where they might happen
  - Exon or introns
  - Promoters
  - 5’ or 3’ untranslated
  - Splice sites
  - Non-coding regulatory RNAs

- What aspect(s) of gene expression?

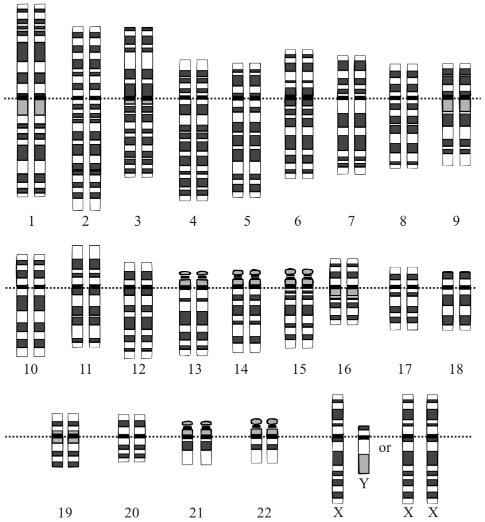
Human Genome Project timeline



1990
Human Genome Project (HGP) launched in the U.S.
Ethical, Legal, and Social Implications (ELSI) programs founded at NIH and DOE
First gene for breast cancer (BRCA1) mapped
1991
First U.S. Genome Centers established
Genetic Map
Cytogenetic Map
Physical Map
RH Map
Clone-Based STS Map
Sequence Map
1992
Second-generation human genetic map developed
Rapid data release guidelines established by NIH and DOE
1993
New five-year plan for the HGP in the U.S. published
Sanger Centre founded (later renamed Wellcome Trust Sanger Institute)
The Wellcome Trust
1994
HGP's human genetic mapping goal achieved
1995
HGP's human physical mapping goal achieved
First bacterial genome (H. influenzae) sequenced
U.S. Equal Employment Opportunity Commission issues policy on genetic discrimination in the workplace
1996
First human gene map established
Pilot projects for human genome sequencing begin in U.S.
First archael genome sequenced
Yeast (S. cerevisiae) genome sequenced
HGP's mouse genetic mapping goal achieved
Bermuda principles for rapid and open data release established
1997
DOE forms Joint Genome Institute
NCHGR becomes NHGRI
National Human Genome Research Institute
E. coli genome sequenced
Genoscope (French National Genome Sequencing Center) founded
1998
Incorporation of 30,000 genes into human genome map
New five-year plan for the HGP in the U.S. published
RIKEN Genomic Sciences Center (Japan) established
Roundworm (C. elegans) genome sequenced
SNP initiative begins
GTGCT
GTCCT
Chinese National Human Genome Centers (in Beijing and Shanghai) established
1999
Full-scale human sequencing begins
Sequence of first human chromosome (chromosome 22) completed
2000
Draft version of human genome sequence completed
President Clinton and Prime Minister Blair support free access to genome information
Fruit fly (D. melanogaster) genome sequenced
Mustard cress (A. thaliana) genome sequenced
Executive order bans genetic discrimination in U.S. federal workplace
2001
Draft version of human genome sequence published
10,000 full-length human cDNAs sequenced
Mammalian Gene Collection
2002
Draft version of mouse genome sequence completed and published
Draft version of rat genome sequence completed
Draft version of rice genome sequence completed and published
2003
Finished version of human genome sequence completed
HGP ends with all goals achieved
to be continued..

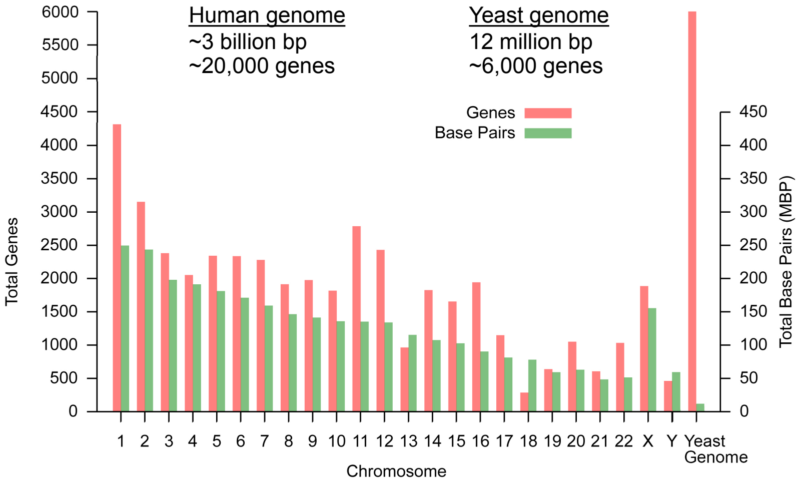
https://www.mun.ca/biology/scarr/

Human genome



Wikipedia

DNA and gene content of the genome

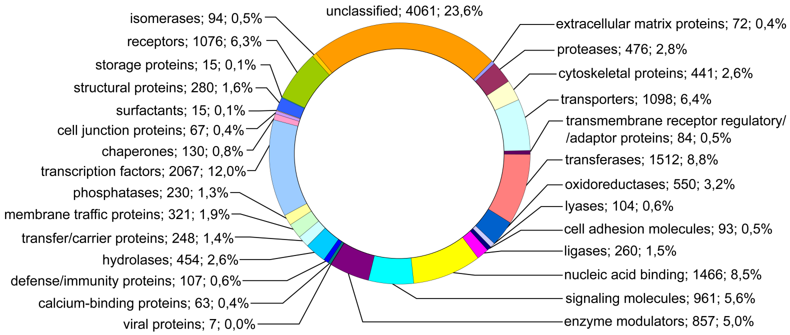


Human genome
~3 billion bp
~20,000 genes

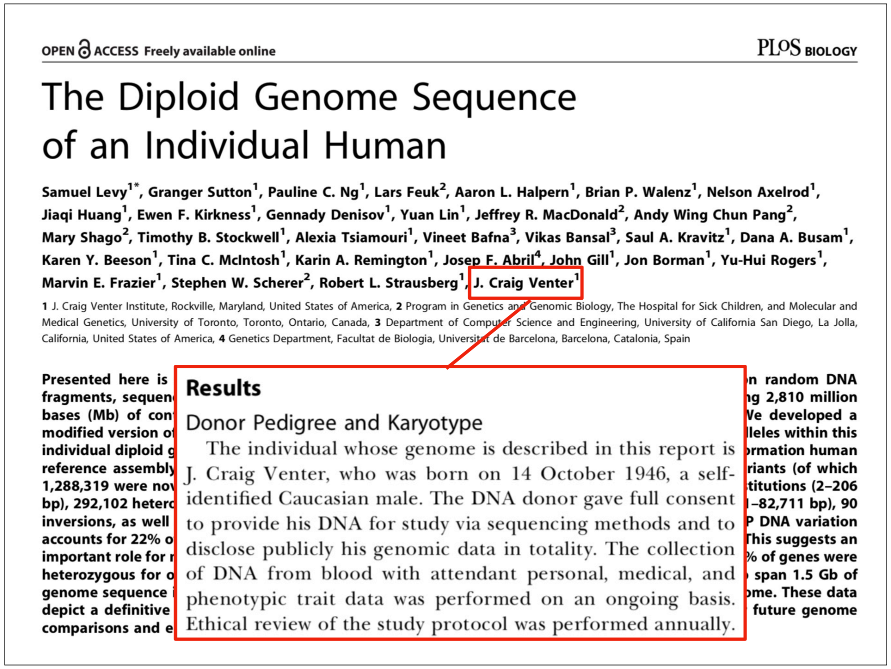
Yeast genome
12 million bp
~6,000 genes

modified from Wikipedia

Deduced functions of encoded proteins



Wikipedia



OPEN ACCESS Freely available online

PLOS BIOLOGY

The Diploid Genome Sequence of an Individual Human

Samuel Levy[1*], Granger Sutton[1], Pauline C. Ng[1], Lars Feuk[2], Aaron L. Halpern[1], Brian P. Walenz[1], Nelson Axelrod[1], Jiaqi Huang[1], Ewen F. Kirkness[1], Gennady Denisov[1], Yuan Lin[1], Jeffrey R. MacDonald[2], Andy Wing Chun Pang[2], Mary Shago[2], Timothy B. Stockwell[1], Alexia Tsiamouri[1], Vineet Bafna[3], Vikas Bansal[3], Saul A. Kravitz[1], Dana A. Busam[1], Karen Y. Beeson[1], Tina C. McIntosh[1], Karin A. Remington[1], Josep F. Abril[4], John Gill[1], Jon Borman[1], Yu-Hui Rogers[1], Marvin E. Frazier[1], Stephen W. Scherer[2], Robert L. Strausberg[1], J. Craig Venter[1]

1 J. Craig Venter Institute, Rockville, Maryland, United States of America, **2** Program in Genetics and Genomic Biology, The Hospital for Sick Children, and Molecular and Medical Genetics, University of Toronto, Toronto, Ontario, Canada, **3** Department of Computer Science and Engineering, University of California San Diego, La Jolla, California, United States of America, **4** Genetics Department, Facultat de Biologia, Universitat de Barcelona, Barcelona, Catalonia, Spain

Presented here is
fragments, sequen
bases (Mb) of con
modified version o
individual diploid g
reference assembly
1,288,319 were nov
bp), 292,102 hetero
inversions, as well
accounts for 22% o
important role for r
heterozygous for o
genome sequence i
depict a definitive
comparisons and e

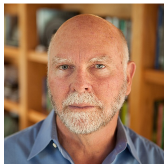
Results

Donor Pedigree and Karyotype

The individual whose genome is described in this report is J. Craig Venter, who was born on 14 October 1946, a self-identified Caucasian male. The DNA donor gave full consent to provide his DNA for study via sequencing methods and to disclose publicly his genomic data in totality. The collection of DNA from blood with attendant personal, medical, and phenotypic trait data was performed on an ongoing basis. Ethical review of the study protocol was performed annually.

n random DNA
ng 2,810 million
We developed a
lleles within this
rmation human
riants (of which
titutions (2–206
1–82,711 bp), 90
P DNA variation
This suggests an
% of genes were
span 1.5 Gb of
me. These data
future genome

Comparison of the Venter genome to the reference genome…



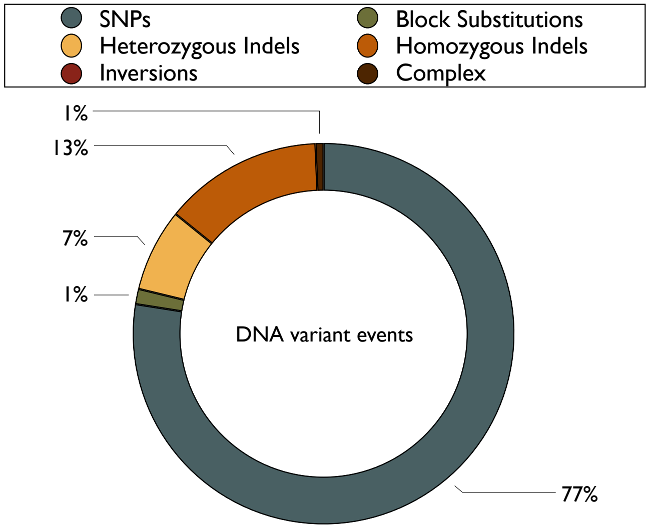
Craig Venter

- 4.1 million DNA variants (12.3 Mb of total bases)
- 3,213,401 single nucleotide polymorphisms (SNPs)
- 53,823 *block substitutions** (2–206 bp)
- 292,102 heterozygous indels (1–571 bp)
- 559,473 homozygous indels (1–82,711 bp)
- 90 inversions
- "Numerous segmental duplications and copy number variations"

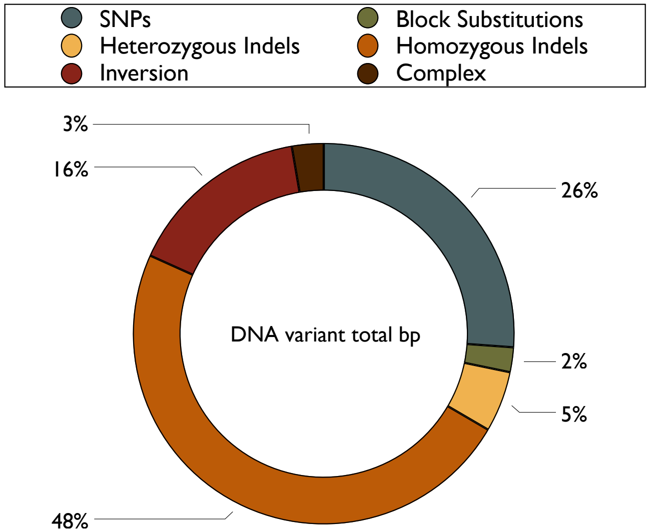
* small regions with many substitutions

Levy S et al. PLoS Biol. 2007 5:e254.

Fraction of events of each type



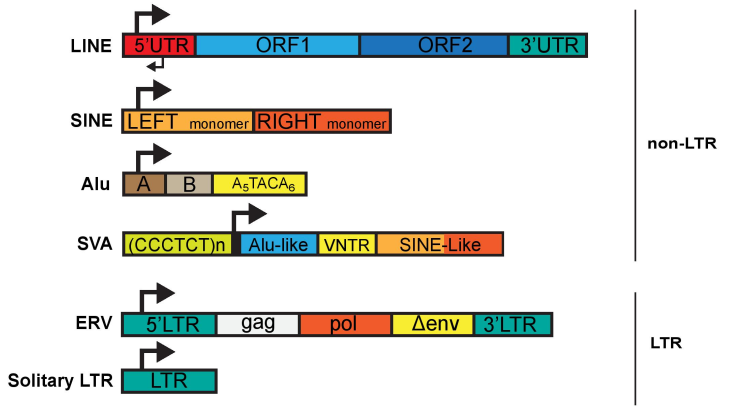
Number of total base pairs for each type



Range of sizes of each type of variation

Type	Minimum	Maximum	Mean
SNP	1	1	1
Block substitution	2	206	4.8
Heterozygous indels	1	321	2.4
Homozygous insertions	1	82,711	11.3
Homozygous deletions	1	18,484	9.9
Inversion	7	670,345	21,272
Complex	2	571	11.7

Large insertions are mainly transposons



http://www.operamedphys.org

Where do SNPs occur?

Classification	Percent
Intronic	40
Intergenic	32
Within non-coding sequence of a gene	10
Upstream	8
Downstream	4
Non-synonymous coding*	3
3′ untranslated region	~1
Synonymous coding	~1

* Only these change protein coding!

Nearly half the genes are heterozygous

- 44% of genes are heterozygous for one or more mutations
- For simplicity, the two alleles are termed A and B alleles
- Analyzing SNPs or other mutations involve distinguishing between the A and B alleles
- How is that done?

A allele	B allele
G - C	G - C
A - T	A - T
A - T	A - T
T - A	T - A
T - A	C - G
C - G	C - G
G - C	G - C
C - G	C - G
T - A	T - A

SNP example: sickle cell anemia

	Thr	-	Pro	-	Glu	-	Glu	beta^{A} protein
	ACT	-	CCT	-	GAG	-	GAG	beta^{A} gene
Codon #	4		5		6		7	
	ACT	-	CCT	-	GTG	-	GAG	beta^{S} gene
	Thr	-	Pro	-	Val	-	Glu	beta^{S} protein

⬆

Sickle Cell Anemia SNP

- Despite being deleterious, the sickle cell anemia SNP is common (~8% among African-Americans)

Frequency of SNPs in the genome?

- Total of common human SNPs: ~10 million
  - Present in 10-50% of individuals
- Size of genome: ~ 3 billion
- About 20,000 genes (1.5% of total genome)
- Average distance between common SNPs = 300 bp
- There are many more uncommon SNPs (<10%)—40 million in the newest database

- 96% of coding regions include at least one common SNP
- Per transcription unit, there are 72 common SNPs on average (median = 43)

- Venter had only ~3 million of the 10 million common SNPs consistent with an average of 30% per SNP

What genetic analysis could exploit SNPs?

- In groups, try to think of every possible way SNPs could be used for genetic analysis

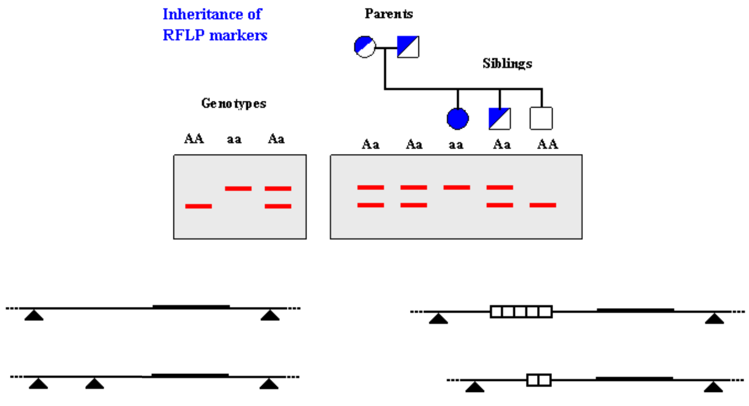
Polymorphisms provide a powerful genetic tool

- SNPs and other mutations can be used to…
  - …map genes associated with Mendelian genetic diseases
  - …map genes associated with incidence of polygenic diseases
  - …do forensic analysis
  - …test for paternity
  - …analyze variation in response to drugs
  - …more?

Genotyping SNPs?

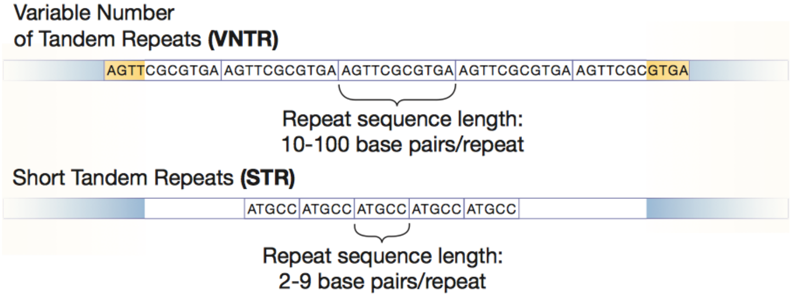
- Restriction fragment length polymorphisms (RFLPs)
  - Some SNPs by chance alter a restriction enzyme site
  - Most SNPs don't do that
  - Indels or repeated regions between restriction sites can create an RFLP
- SNP microarrays
  - Useful for any known SNP
  - Easily automated
  - Can not find novel SNPs
- Deep (next generation) sequencing
  - Can find novel SNPs
  - Too expensive for routine screening (at present)
  - Exome sequencing reduces cost but most SNPs are outside transcribed regions

RFLP analysis?



www.wikimedia.org

Sequence repeats


Variable Number of Tandem Repeats **(VNTR)**

AGTTCGCGTGA AGTTCGCGTGA AGTTCGCGTGA AGTTCGCGTGA AGTTCGCGTGA

Repeat sequence length: 10-100 base pairs/repeat

Short Tandem Repeats **(STR)**

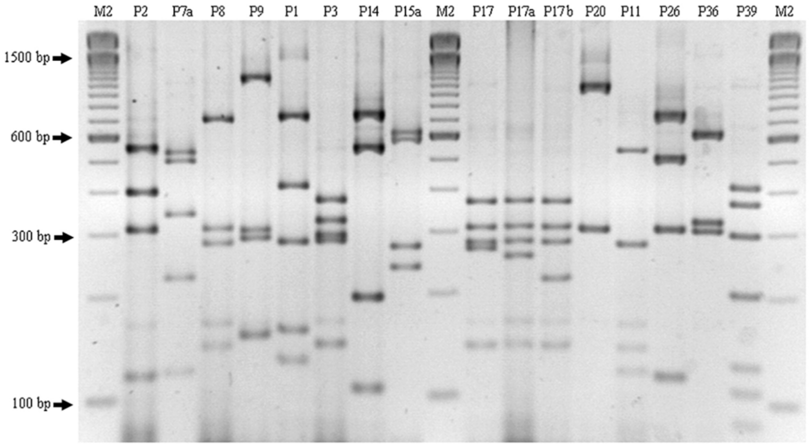
ATGCC ATGCC ATGCC ATGCC ATGCC

Repeat sequence length: 2-9 base pairs/repeat

AKA…

VNTRs = minisatellite sequences
STRs = microsatellite sequences

Real RFLP data can be more complicated



Moreno J. Clin. Microbiol. 2006 44:1453

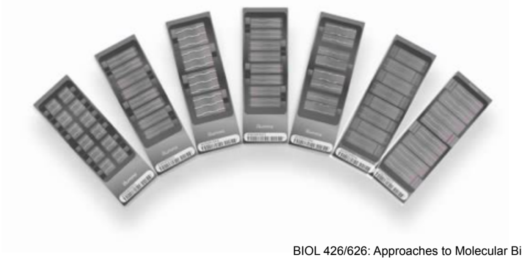
Genotyping SNPs?

- Restriction fragment length polymorphisms (RFLPs)
  - Some SNPs by chance alter a restriction enzyme site
  - Most SNPs don't do that
  - Indels between restriction sites can create an RFLP
- SNP microarrays
  - Useful for any known SNP
  - Easily automated
  - Can not find novel SNPs
- Deep (next generation) sequencing
  - Can find novel SNPs
  - Too expensive for routine screening (at present)
  - Exome sequencing reduces cost but most SNPs are outside transcribed regions

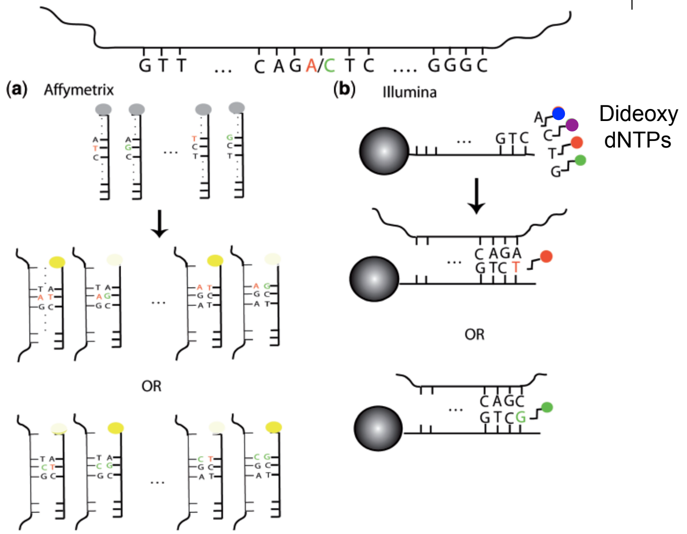
Genotyping with SNP microarrays

Affymetrix

Illumina



Two SNP microarray methods



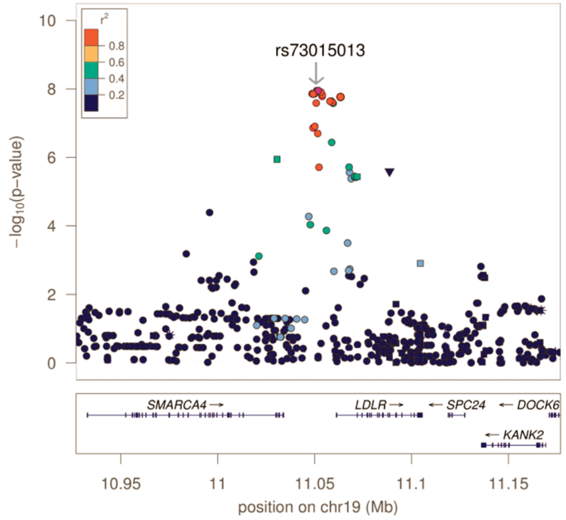
Genotyping SNPs?

- Restriction fragment length polymorphisms (RFLPs)
  - Some SNPs by chance alter a restriction enzyme site
  - Most SNPs don't do that
  - Indels between restriction sites can create an RFLP
- SNP microarrays
  - Useful for any known SNP
  - Easily automated
  - Can not find novel SNPs
- Deep (next generation) sequencing
  - Can find novel SNPs
  - Too expensive for routine screening (at present)
  - Exome sequencing reduces cost but most SNPs are outside transcribed regions

SNPs can be used to map linked genes

- SNPs near a gene carrying a mutation causing or contributing to a disease will be inherited by affected individuals
  - GWAS (genome wide association study)
  - Compare SNPs across the genome
  - Requires large numbers of affected and unaffected people
  - A SNP near any gene contributing to the disease will appear more often in affected than unaffected people

SNP associated with abnormal low-density lipoprotein (LDL)

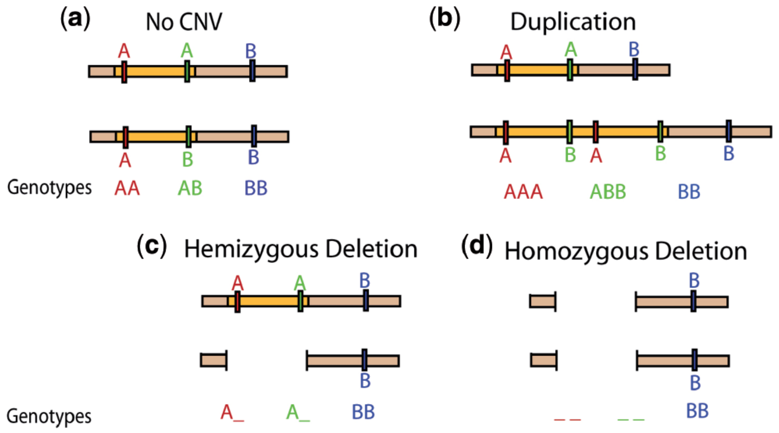


LDLR = gene for LDL receptor

Wikipedia

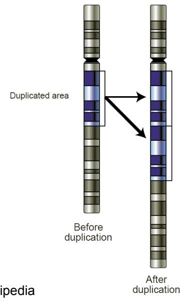
Changes in genetic locus copy number

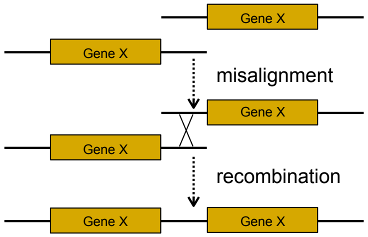
CNV = copy number variant



SNPs can identify "copy number variations"

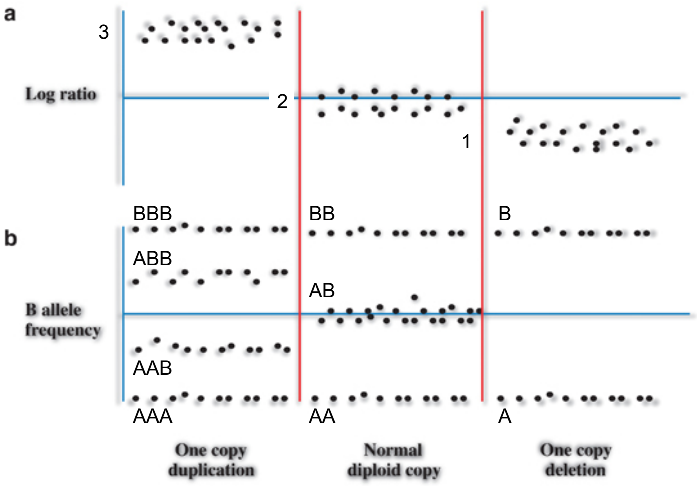
- Because pairs of chromosomes differ in the SNPs they carry they essentially are a "tag" that marks each of them
- From the genome sequence we know that large deletions and insertions are common; these change to "copy number" of the sequences in the two chromosomes
- SNP array analysis can identify regions of copy number variation (CNV)





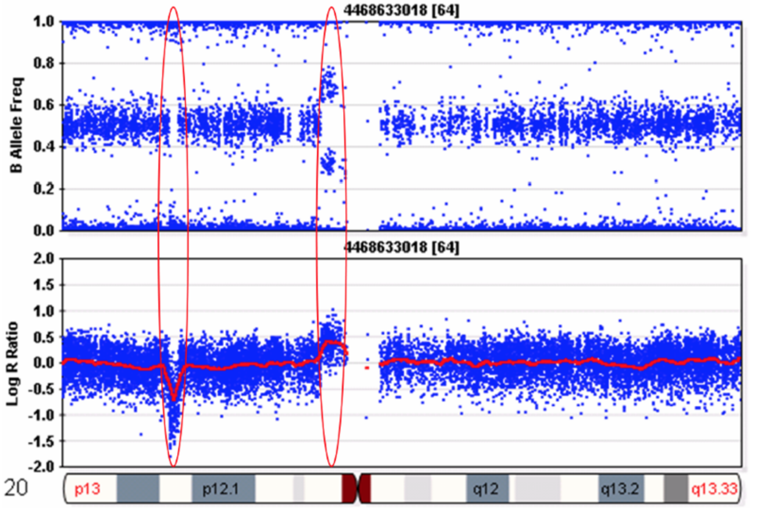
Wikipedia

SNP arrays report copy number



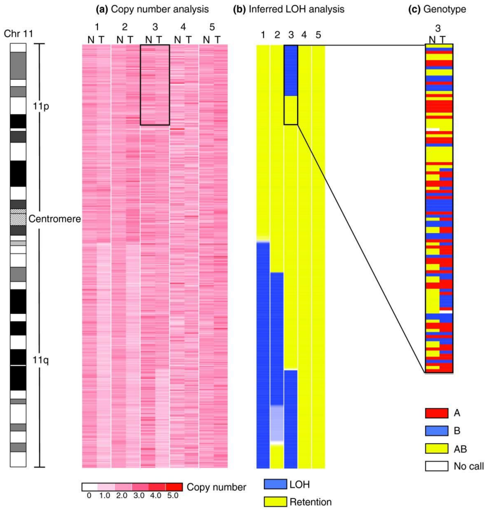
Ku et al. Mol. Psych. 2013 18:141

A one copy deletion and a duplication on human Chr 20



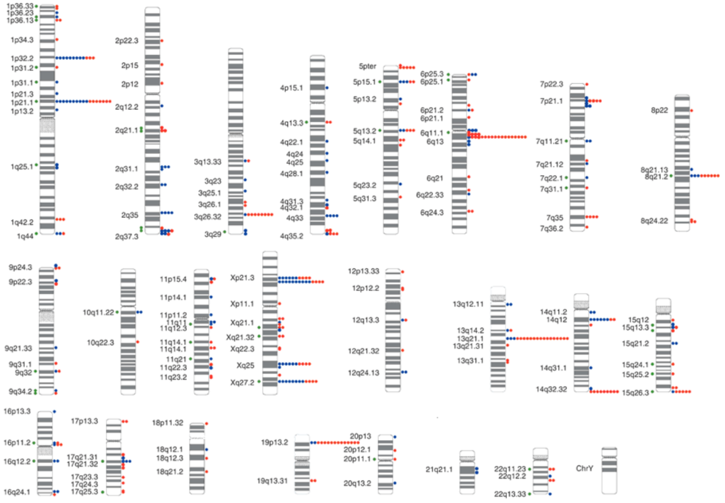
Wang & Bucan (2008) Cold Spring Harbor Protocols 3:1

Loss of heterozygosity (LOH) in tumor cells



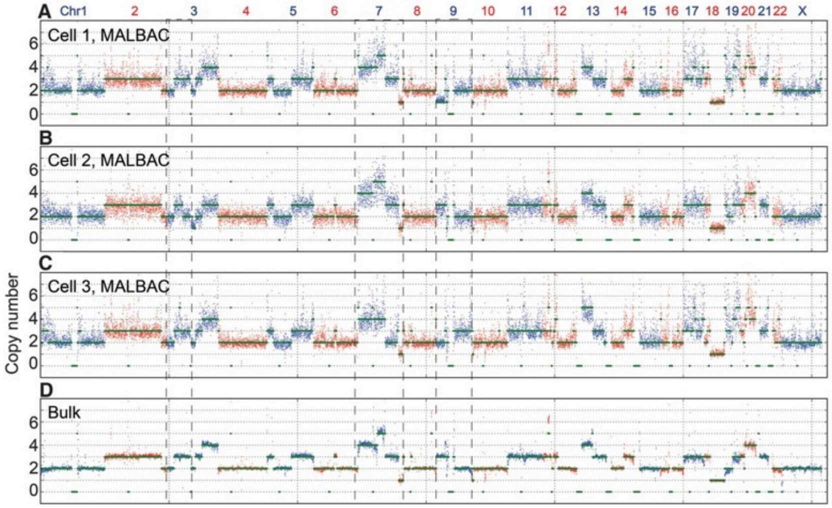
Heinrichs Genome Biol. 2007 8:219

Sites of copy number variation across the human genome



Iafrate et al. Nat Genet. 2004 36:949

Wide spread CNV in a cancer cell



Zong Science. 2012 338:1622

Reading for next time:

- Next class we will be doing peer review. Please bring two printed copies of your first draft to class.
- For the following class, read:
  - From Genes to Genomes: Concepts and Applications of DNA Technology by Dale et al., Chapter 10 “Transcriptomics & proteomics”