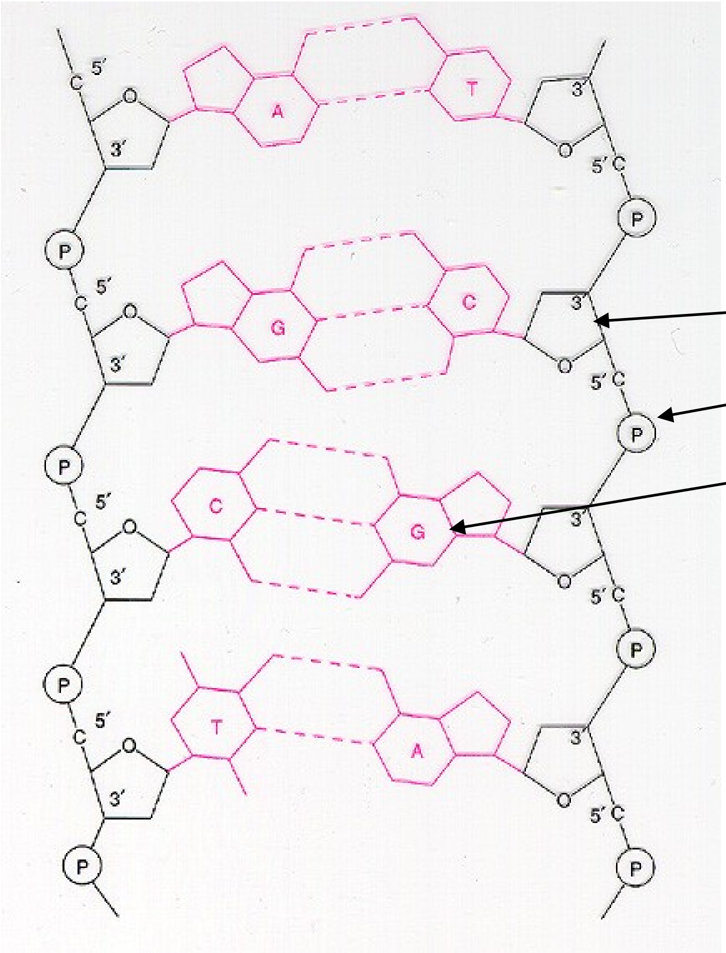
CS681: Advanced Topics in Computational Biology

Week 1, Lectures 2-3

Can Alkan
EA224
calkan@cs.bilkent.edu.tr

http://www.cs.bilkent.edu.tr/~calkan/teaching/cs681/

DNA structure refresher



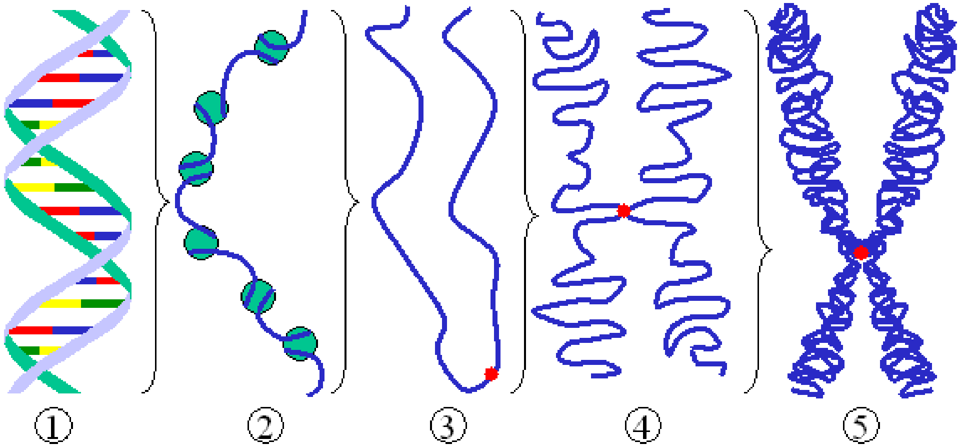
- DNA has a double helix structure which composed of
  - sugar molecule
  - phosphate group
  - and a base (A,C,G,T)

- DNA always reads from 5’ end to 3’ end for transcription replication

  5’ ATTTAGGCC 3’

  3’ TAAATCCGG 5’

Refresher: Chromosomes



- (1) Double helix DNA strand.
- (2) Chromatin strand (**DNA** with **histones**)
- (3) Condensed chromatin during interphase with **centromere**.
- (4) Condensed chromatin during prophase
- (5) Chromosome during metaphase

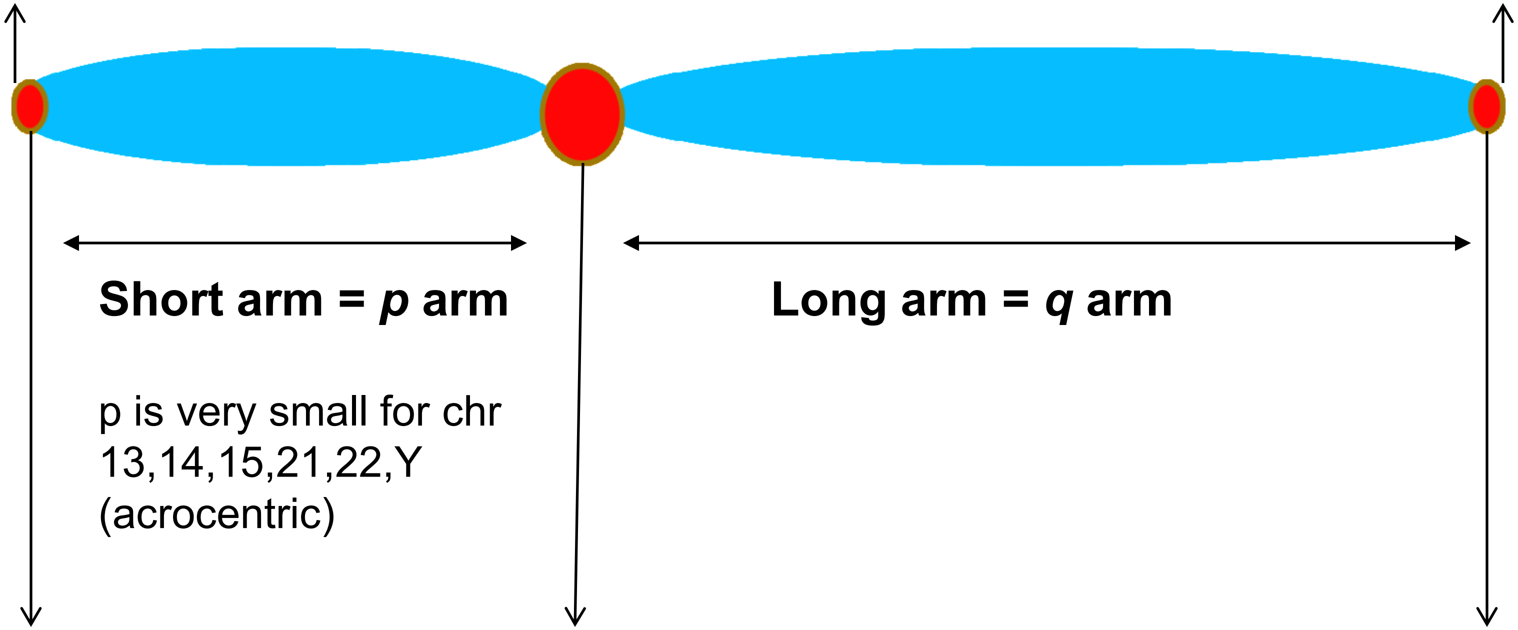
Chromosomes

Organism	Number of base pairs	number of chromosomes (n)
Prokayotic		
Escherichia coli (bacterium)	$4x10^6$	1
Eukaryotic		
Saccharomyces cerevisiae (yeast)	$1.35x10^7$	17
Drosophila melanogaster(insect)	$1.65x10^8$	4
Homo sapiens(human)	$2.9x10^9$	23
Zea mays(corn)	$5.0x10^9$	10

Chromosome structure

End of telomere
= T-loop (300 bp)

End of telomere
= T-loop (300 bp)



Short arm = *p* arm

p is very small for chr
13,14,15,21,22,Y
(acrocentric)

Long arm = *q* arm

Telomere
6bp tandem repeats
TTAGGC

Centromere
171bp tandem repeats
(alpha satellites)

Telomere
6bp tandem repeats
TTAGGC

Back to Genomes

- To understand the biology of species, we need to read their genomes:
  - Genome sequencing
- Basically
  - Collect DNA
  - Shear into pieces
  - Read pieces
  - Join them together
    - Sequence assembly ->very hard problem (week 7)

Sequenced Genomes

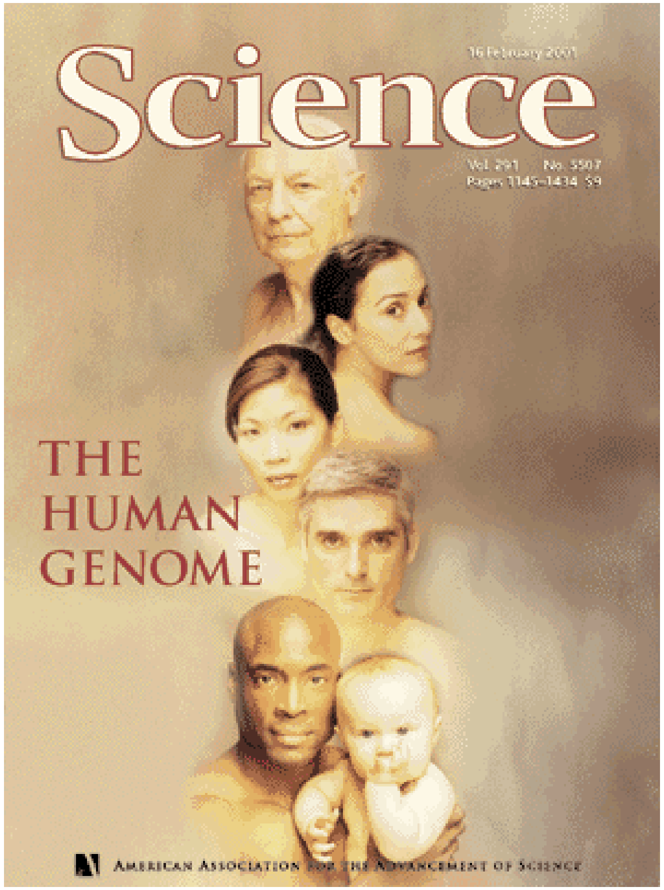
- Many many bacteria & single cell organisms (E. coli, etc.)
- Plants: rice, wheat, potato, tomato, grape, corn, etc.
- Insects: ant, mosquito, etc.
- Nematodes: C. elegans, etc.
- Many fish
- Mammals: human, chimp, bonobo, gorilla, orangutan, macaque, baboon, marmoset, horse, cat, dog, pig, panda, elephant, mouse, rat, opossum, armadillo, etc.

Non-human genomes

- BGI (China) has 1000 Plants and Animals Project
- Genome 10K (www.genome10k.org): Open-source like collaboration network that aims to sequence the genomes of 10.000 vertebrate species
  - Computational challenges / competition:
    - Alignathon
    - Assemblathon
- i5K: 5.000 insect species

Human genome project

- 1986: Announced (USA+UK)
- 1990: Started
- 1999: Chromosome 22 sequenced
- 2001: First draft
- 2004: Finished (kind of)

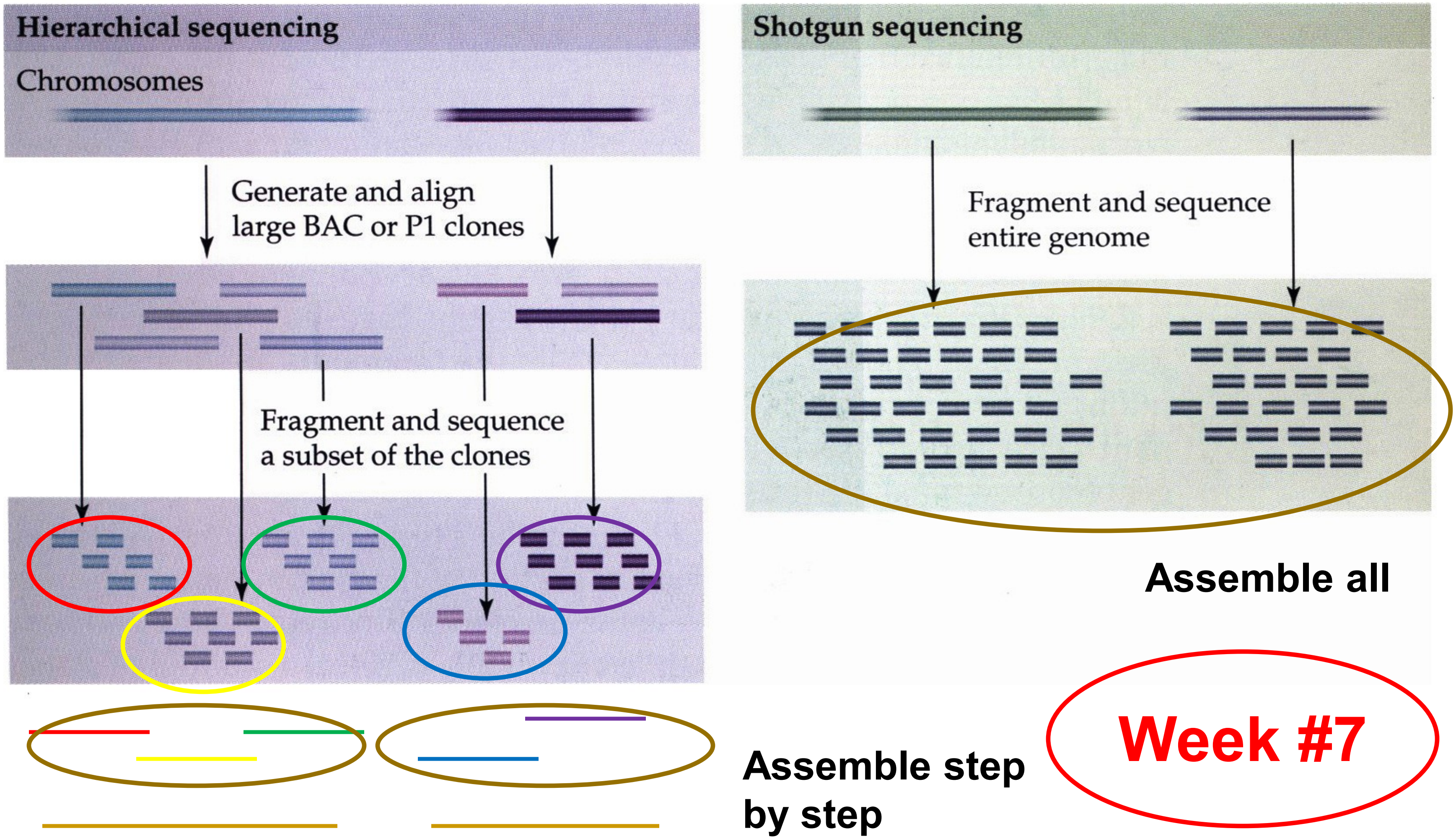


Many human samples, 14 years, 3-10 billion dollars

Sequencing basics

- No technology can read a chromosome from start to finish; all sequencers have limits for read lengths
- Two major approaches
  - Hierarchical sequencing (used by the human genome project)
    - High quality, very low error rate, little fragmentation
    - Slow and expensive!
  - Whole genome shotgun (WGS) sequencing
    - Lower quality, more errors, assembly is more fragmented
    - Fast and cheap(er)

Hierarchical vs. shotgun sequencing

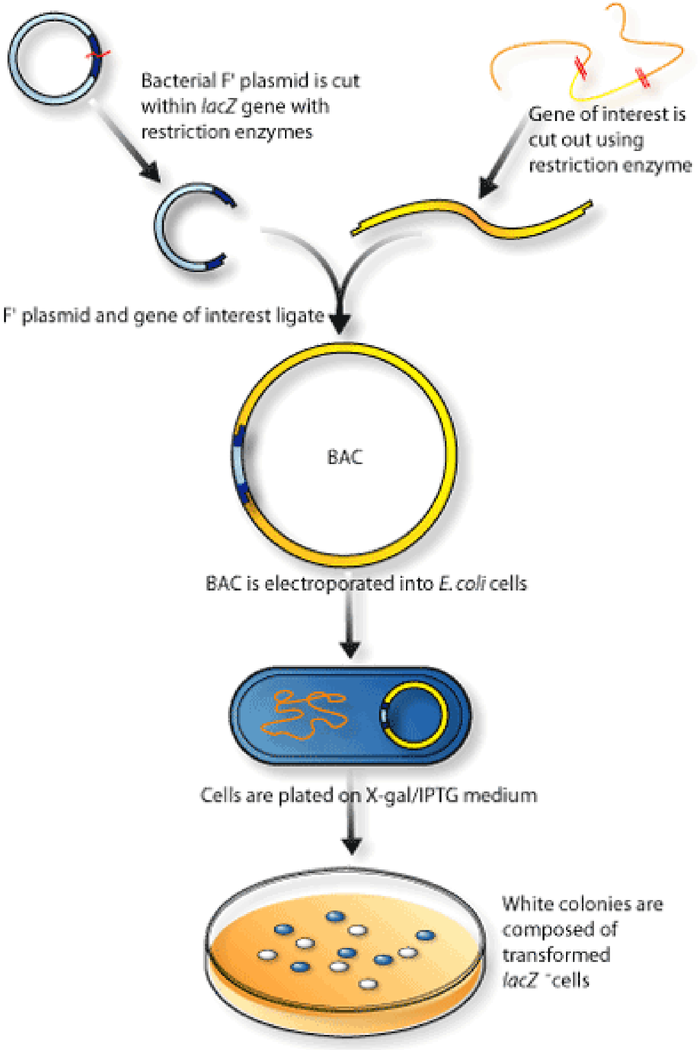

Hierarchical sequencing
Chromosomes
Generate and align large BAC or P1 clones
Fragment and sequence a subset of the clones
Shotgun sequencing
Fragment and sequence entire genome

Assemble all

Assemble step by step

Week #7

Cloning vectors



Bacterial F' plasmid is cut within *lacZ* gene with restriction enzymes

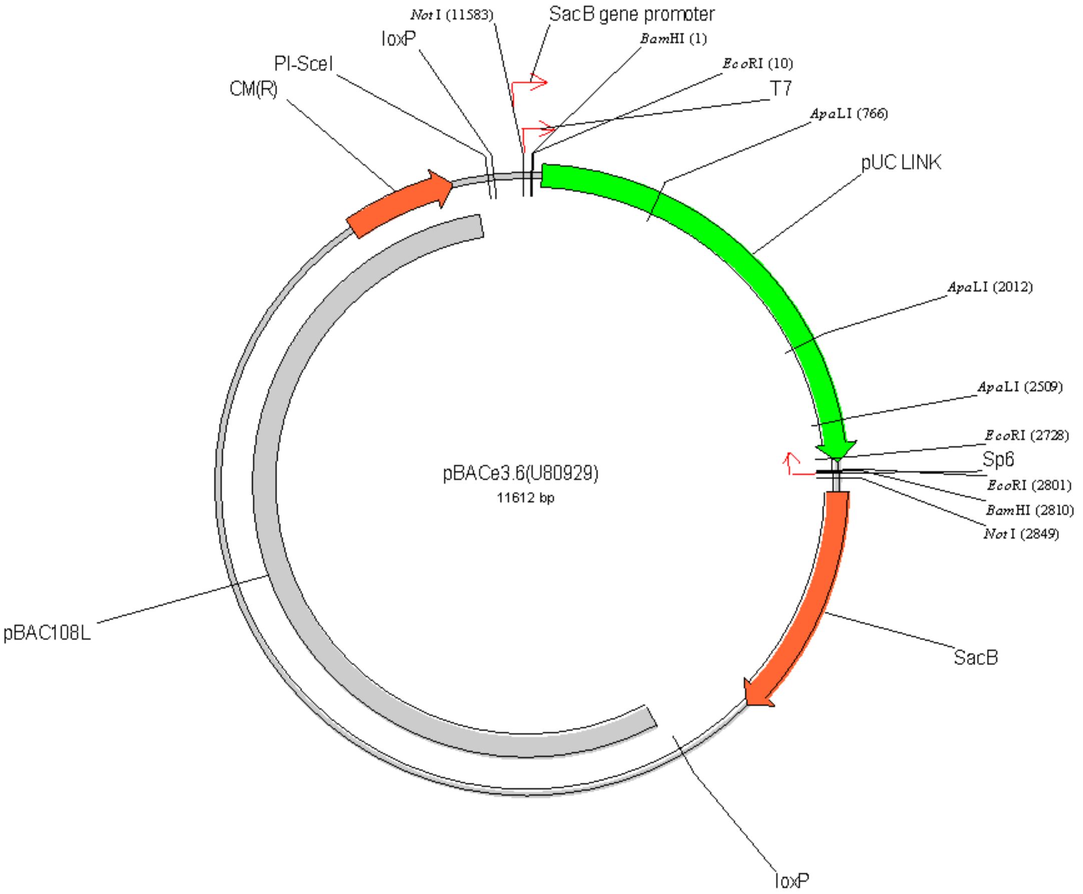
Gene of interest is cut out using restriction enzyme

F' plasmid and gene of interest ligate

BAC is electroporated into *E. coli* cells

Cells are plated on X-gal/IPTG medium

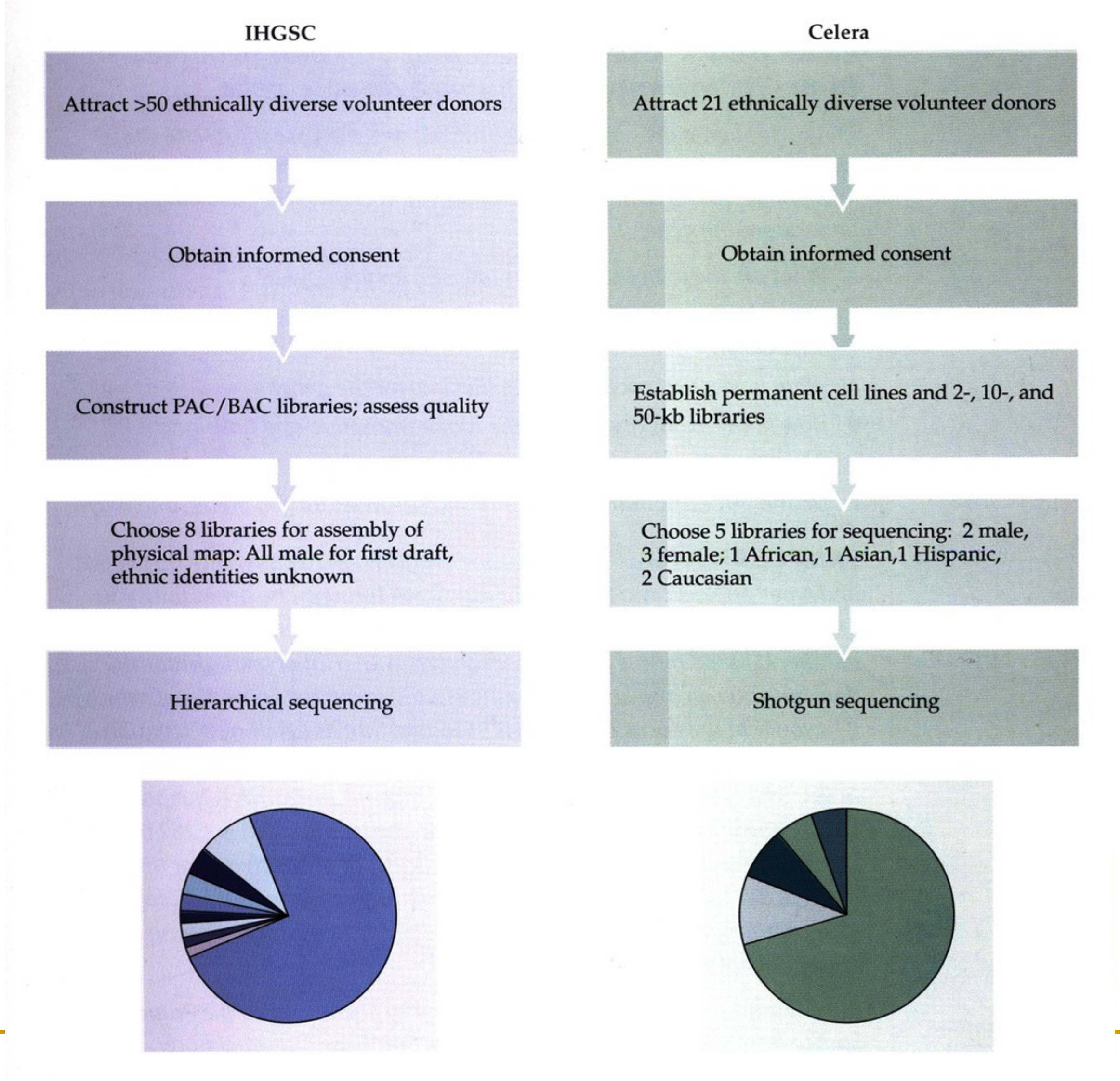
White colonies are composed of transformed *lacZ* ⁻cells



Cloning vectors

- Plasmids: carry 3-10 kbp of DNA
- Fosmids: carry ~40 kbp of DNA
- Cosmids: carry ~35-50 kbp of DNA
- BACs (bacterial artificial chromosomes): ~150-200 kbp of DNA
- YACs (yeast artificial chromosomes): 100 kbp – 3 Mbp of DNA

Human genomes: public vs private



IHGSC

Attract >50 ethnically diverse volunteer donors

Obtain informed consent

Construct PAC/BAC libraries; assess quality

Choose 8 libraries for assembly of physical map: All male for first draft, ethnic identities unknown

Hierarchical sequencing

Celera

Attract 21 ethnically diverse volunteer donors

Obtain informed consent

Establish permanent cell lines and 2-, 10-, and 50-kb libraries

Choose 5 libraries for sequencing: 2 male, 3 female; 1 African, 1 Asian,1 Hispanic, 2 Caucasian

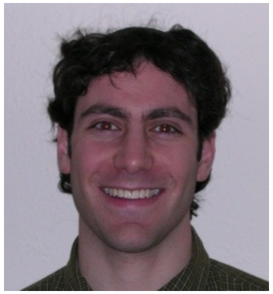
Shotgun sequencing

GENOMIC VARIATION:
CHANGES IN DNA SEQUENCE

The Diversity of Life

- Not only do different species have different genomes, but also different individuals of the same species have different genomes.
- No two individuals of a species are quite the same – this is clear in humans but is also true in every other sexually reproducing species.
  - Any two humans genomes are still 99.9% identical!

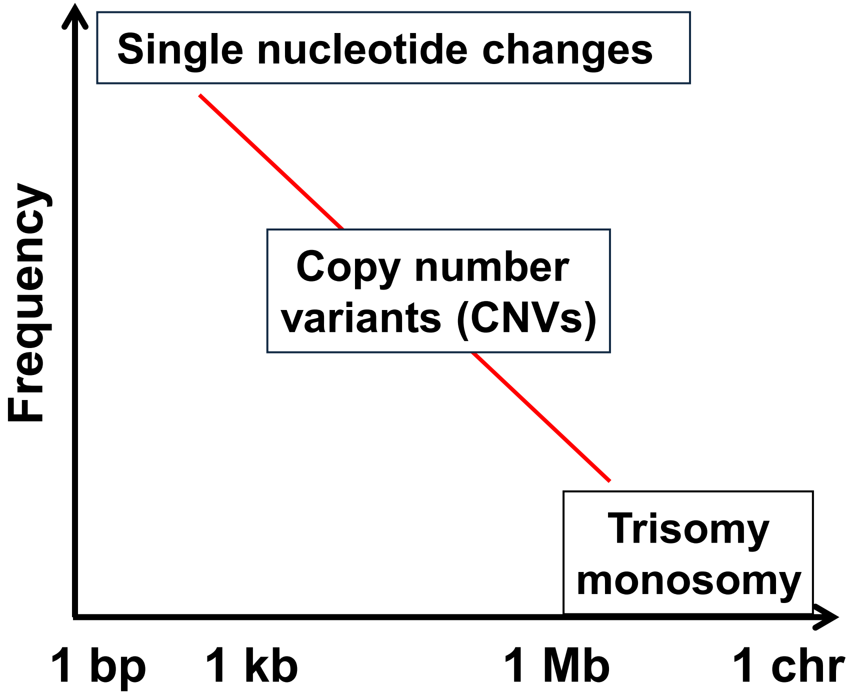
Human genome variation



- Genomic variation
  - Changes in DNA sequence
- Epigenetic variation
  - Methylation, histone modification, etc.

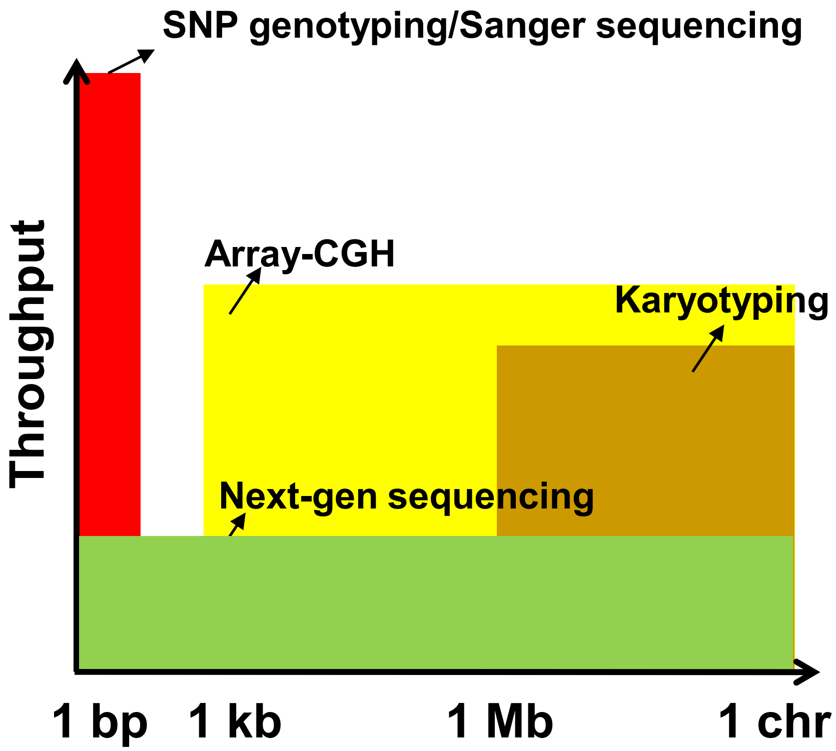
Human genetic variation

Types of genetic variants



Size of variant

How do we assay them?



Size of variant

Size range of genetic variation

- Single nucleotide (SNPs)
- Few to ~50bp (small indels, microsatellites)
- >50bp to several megabases (**structural variants)**:
  - ❑ Deletions
  - ❑ Insertions
    - Novel sequence
    - Mobile elements (*Alu*, L1, SVA, etc.)
  - ❑ Segmental Duplications
    - Duplications of size ≥ 1 kbp and sequence similarity ≥ 90%

CNVs

  - ❑ Inversions
  - ❑ Translocations
- Chromosomal changes

Genetic variation

If a mutation occurs in a codon:

- Synonymous mutations: Coded amino acid doesn't change
- Nonsynonymous mutations: Coded amino acid changes

GTT → Valine

GTA → Valine

SYNONYMOUS

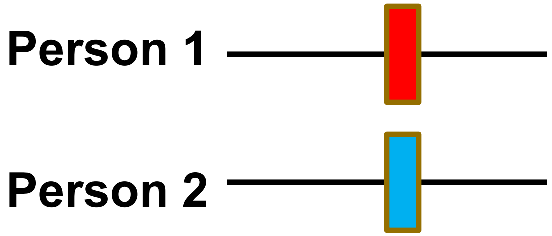
GTT → Valine

GCA → Alanine

NONSYNONYMOUS

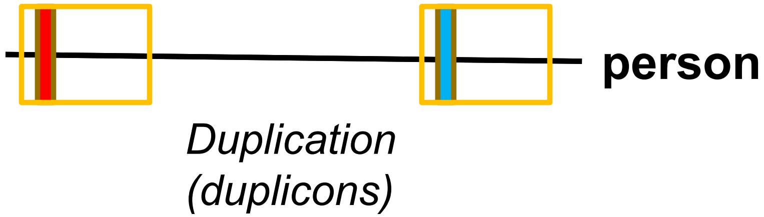
Genetic variation

Where in the genome?



Person 1

Person 2



person

Duplication (duplicons)

ALLELIC VARIATION

NONALLELIC (PARALOGOUS) VARIATION

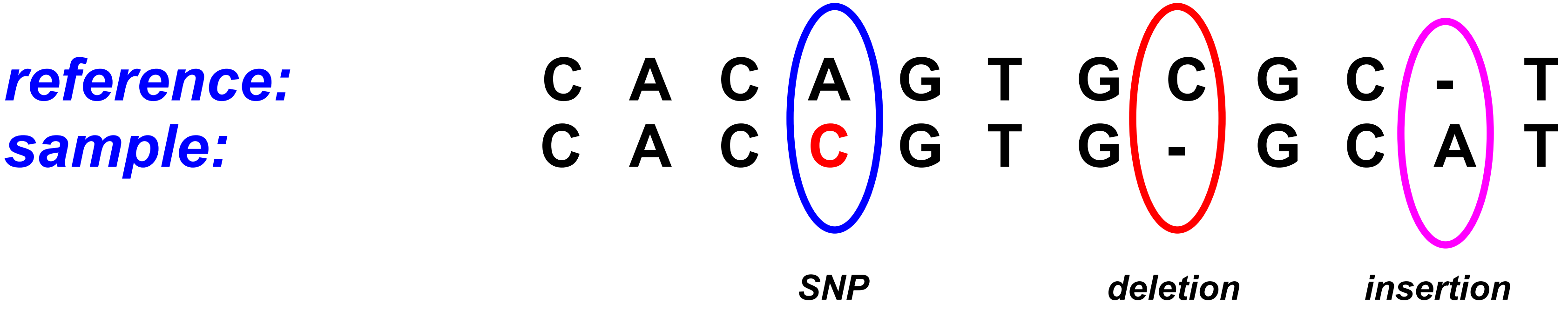
Where in the body?

Germ cells or gametes (sperm egg) -> Transmittable -> Germline Variation

Other (somatic cells) -> Not transmittable -> Somatic Variation

SNPs & indels

SNP: Single nucleotide polymorphism (substitutions)
Short indel: Insertions and deletions of sequence of length 1 to 50 basepairs



reference:	C	A	C	A	G	T	G	C	G	C	-	T
sample:	C	A	C	C	G	T	G	-	G	C	A	T

- Neutral: no effect
- Positive: increases fitness (resistance to disease)
- Negative: causes disease
- Nonsense mutation: creates early stop codon
- Missense mutation: changes encoded protein
- Frameshift: shifts basepairs that changes codon order

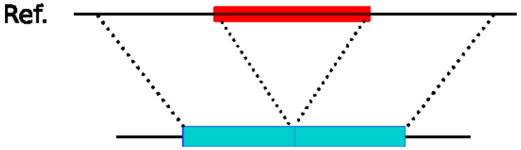
Short tandem repeats

reference: CAG CAG CAG CAG
sample: CAG CAG CAG *CAG* CAG

- Microsatellites (STR=short tandem repeats) 1-10 bp
  - Used in population genetics, paternity tests and forensics
- Minisatellites (VNTR=variable number of tandem repeats): 10-60 bp
- Other satellites
  - Alpha satellites: centromeric/pericentromeric, 171bp in humans
  - Beta satellites: centromeric (some), 68 bp in humans
  - Satellite I (25-68 bp), II (5bp), III (5 bp)
- Disease relevance:
  - Fragile X Syndrome
  - Huntington’s disease

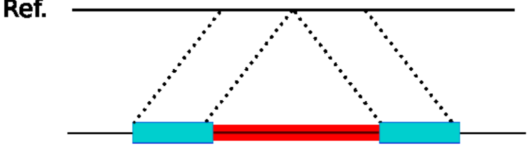
Structural Variation

DELETION

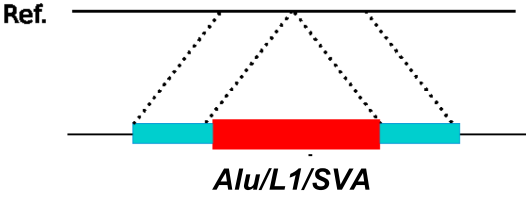


Autism, mental retardation, Crohn’s

NOVEL SEQUENCE INSERTION

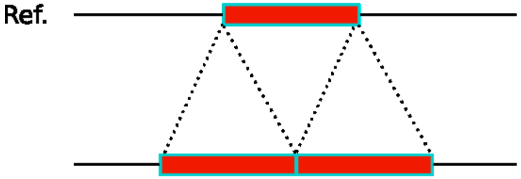


MOBILE ELEMENT INSERTION

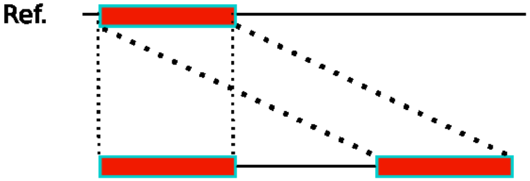


Haemophilia

TANDEM DUPLICATION

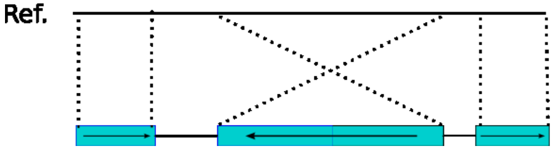


INTERSPERSED DUPLICATION

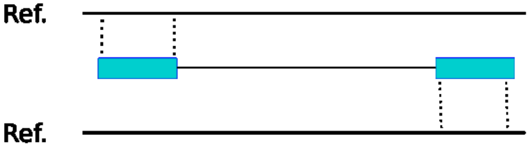


Schizophrenia, psoriasis

INVERSION



TRANSLOCATION

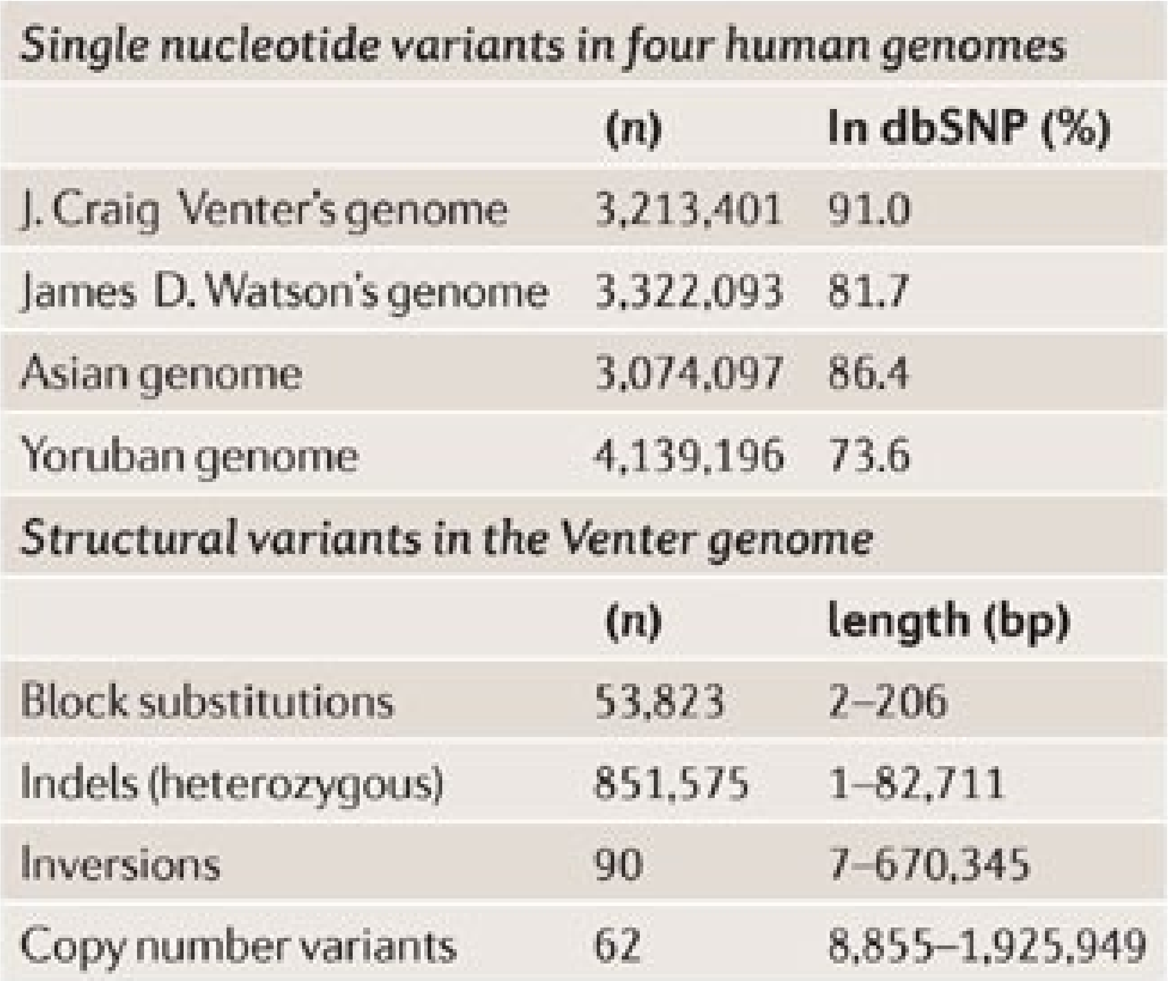


Chronic myelogenous leukemia

Chromosomal changes

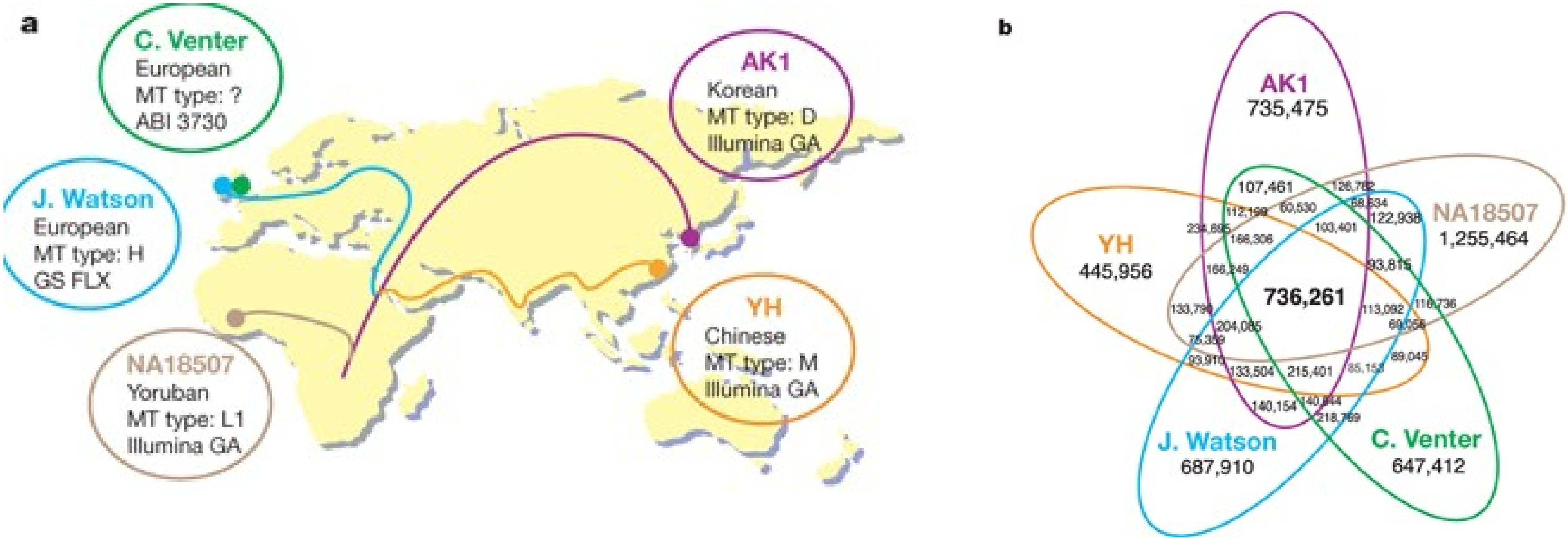
- “Microscope-detectable”
- Disease causing or prevents birth
- Monosomy: 1 copy of a chromosome pair
- Uniparental disomy (UPD): Both copies of *a* pair comes from the same parent
- Trisomy: Extra copy of a chromosome
    - chr21 trisomy = Down syndrome

Genetic variation among humans



Single nucleotide variants in four human genomes		
	(n)	**In dbSNP (%)**
J. Craig Venter's genome	3,213,401	91.0
James D. Watson's genome	3,322,093	81.7
Asian genome	3,074,097	86.4
Yoruban genome	4,139,196	73.6
Structural variants in the Venter genome		
	(n)	**length (bp)**
Block substitutions	53,823	2–206
Indels (heterozygous)	851,575	1–82,711
Inversions	90	7–670,345
Copy number variants	62	8,855–1,925,949

Genetic variation are “shared”



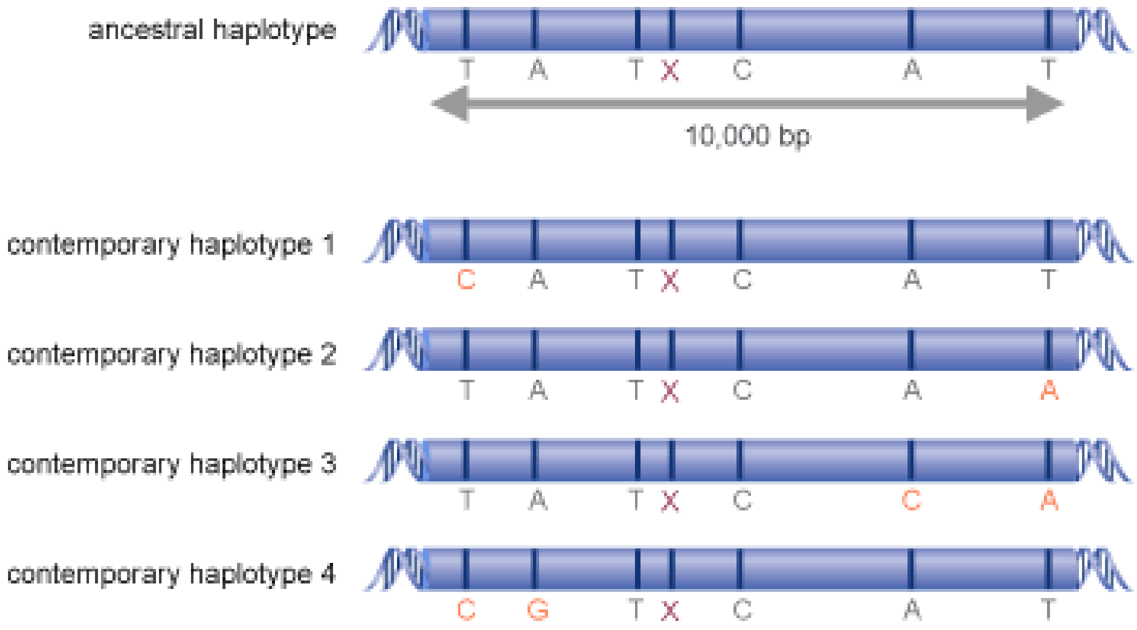
Kim *et al.* Nature, 2009

Zygosity

- Animals are diploid; i.e. 2 of each chromosome, this 2 of each location in the genome
- Any variation is one of:
  - Homozygous: both copies have the same genotype
  - Heterozygous: each copy has the same genotype
  - Hemizygous (for deletions): one copy has a segment missing, the other has it intact

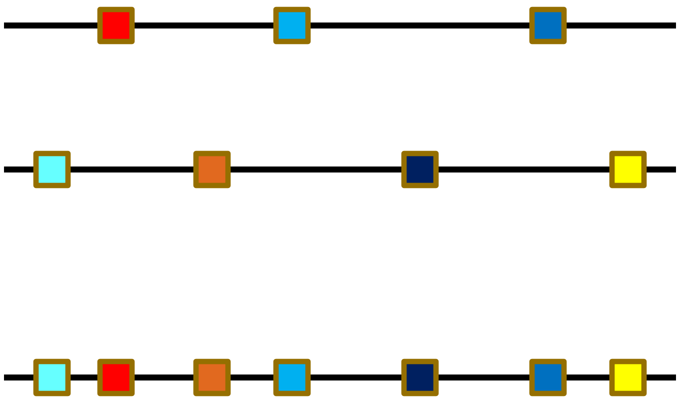
Haplotype

- “Haploid Genotype”: *a combination of alleles at multiple loci that are transmitted together on the same chromosome*



Haplotype resolution

- Variation discovery methods do not directly tell which copy of a chromosome a variant is located
- For heterozygous variants, it gets messy:



Chromosome 1, #1

Chromosome 1, #2

Discovered variants in Chromosome 1

Haplotype resolution or haplotype phasing: finding which groups of variants "go together"

Discovery vs. genotyping

- Discovery: no *a priori* information on the variant
- Genotyping: test whether or not a "suspected" variant occurs

Variation discovery & genotyping

- Targeted, low-cost methods:
  - SNP:
    - PCR
    - SNP microarray (genotyping)
  - Indel
    - PCR
    - “Indel microarray” (genotyping)
  - Structural variation
    - Quantitative PCR
    - Array Comparative Genomic Hybridization (array CGH)
    - Fluorescent *in situ* Hybridization (FISH) if variant > 500 kb
  - Chromosomal:
    - Microscope!

Next week

Variation discovery & genotyping

- Targeted methods are:
  - Cheap(er), but limited:
    - Variants that are not in reference genome cannot be found
    - One experiment yields one type of variant
    - Not always genome-wide
- Alternative:
  - Whole genome resequencing
    - More expensive
    - (Theoretically) comprehensive
    - Computational challenges

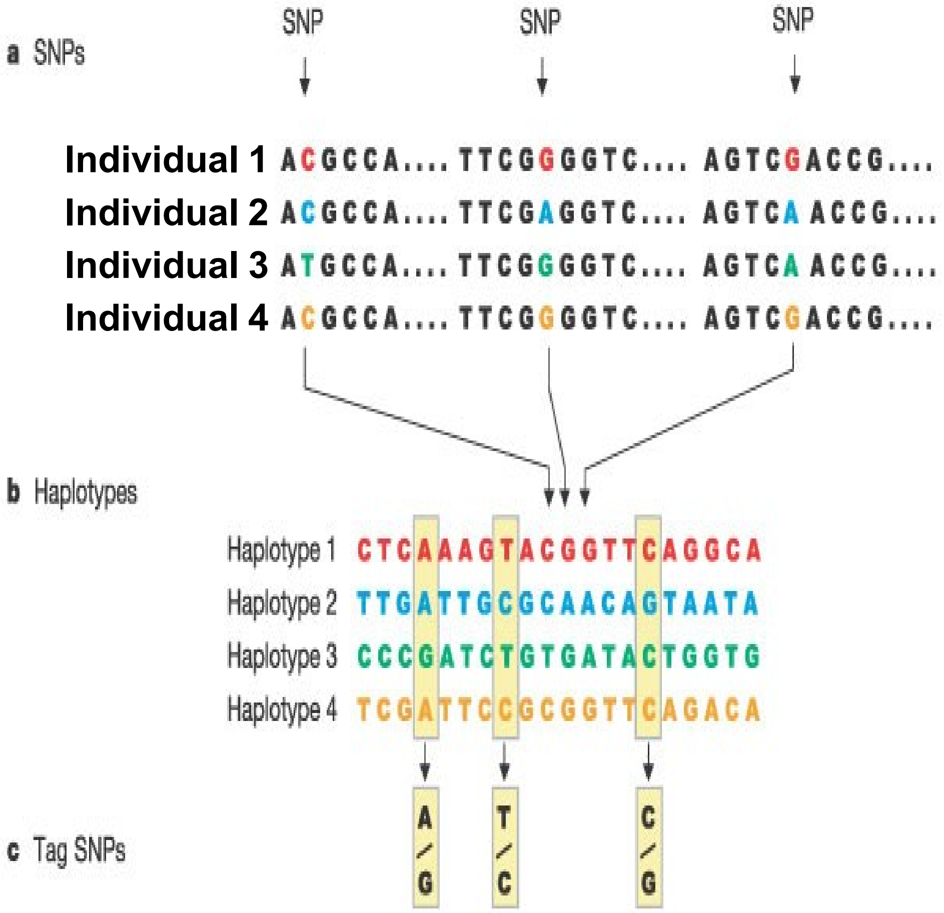
PROJECTS FOR GENOMIC VARIATION DISCOVERY

International HapMap Project

- Determine genotypes & haplotypes of 270 human individuals from 3 diverse populations:
  - Northern Americans (Utah / Mormons)
  - Africans (Yoruba from Nigeria)
  - Asians (Han Chinese and Japanese)
- 90 individuals from each population group, organized into parent-child **trios**.
- Each individual genotyped at ~5 million roughly evenly spaced markers (SNPs and small indels)

http://www.hapmap.org

HapMap Project


a SNPs
SNP
SNP
SNP
Individual 1 ACGCCA.... TTCGGGGTC.... AGTCGACCG....
Individual 2 ACGCCA.... TTCGAGGTC.... AGTCA ACCG....
Individual 3 ATGCCA.... TTCGGGGTC.... AGTCA ACCG....
Individual 4 ACGCCA.... TTCGGGGTC.... AGTCGACCG....
b Haplotypes
Haplotype 1 CTCAAAGTACGGTTCAGGCA
Haplotype 2 TTGATTGCGCAACAGTAATA
Haplotype 3 CCCGATCTGTGATACTGGTG
Haplotype 4 TCGATTCCGCGGTTCAGACA
c Tag SNPs
A/G
T/C
C/G

Step 1: SNPs are identified in DNA samples from multiple indivduals

Step 2: Adjacent SNPs that are inherited together are compiled into "haplotypes."

Step 3: "Tag" SNPs within haplotypes are identified that uniquely identify those haplotypes

By genotyping just the three tag SNPs shown above, one can identify which of the four haplotypes shown here are present in each individual.

Human Genome Diversity Panel

- More extensive set of genomic variation
- One aim is to build DNA resource libraries for large scale discovery & genotyping projects
- 1.050 human individuals from 52 populations

Initial HapMap and HGDP did not sequence the genomes of any samples.

Why?

- To understand “normal” human genomic variation
- To understand genetic transmission properties
- To understand *de novo* mutations
- To understand population structure, migration patterns
- To understand human disease:
  - Two views
    - Common variant common disease
    - Rare variant common disease

Human disease

- Rare variant common disease:
  - Most “complex” diseases, including neuropsychiatric diseases
- Common variant common disease
  - More “common”; diseases that follow Mendelian inheritance
    - If a common disease is caused by a recessive mutation, it can be found at high frequency in a population
      - MAF (minor allele frequency) > 5%

Why sequence whole genomes?

- SNP/indel/arrayCGH platforms are mainly designed for individuals of West European descent
- For a disease common in somewhere else, like India:
  - Variants at high frequency in India may not be represented in the available platforms
  - Genome is a big entity; SNP/indel/arrayCGH can not cover the entire genome:
    - Largest has 2.1 million markers (compare to 3 billion)

High Throughput Sequencing

- More about HTS platforms, data properties, cost/benefit analyses: Week #3
- Take-home message for today:
  - Cheaper to sequence but harder and expensive to analyze

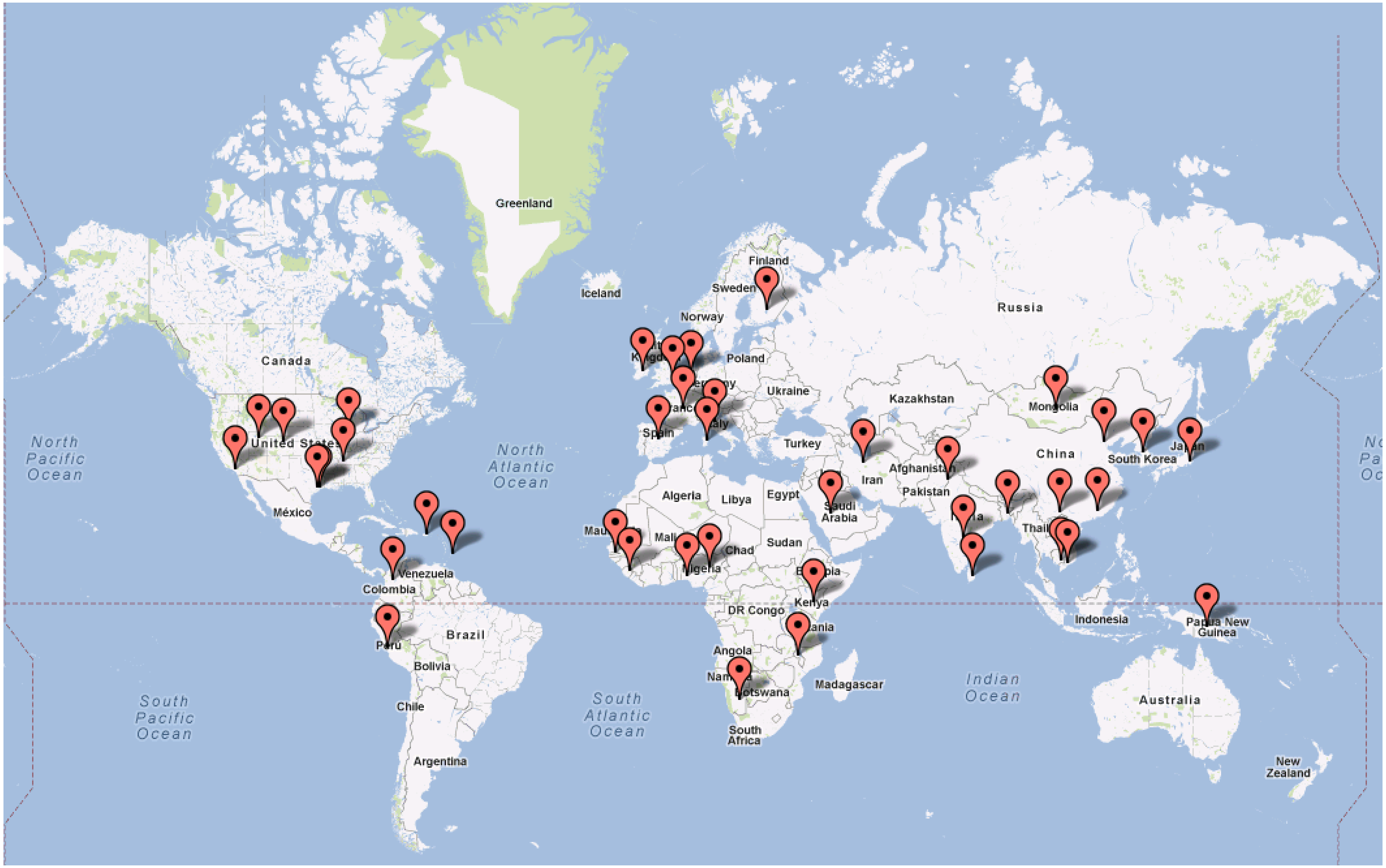
Sequencing-based projects

- The 1000 Genomes Project Consortium (www.1000genomes.org)
  - Large consortium: groups from USA, UK, China, Germany, Canada
  - 2.500 humans from 29 populations
    - 1.197 from 14 populations finished (September 2011)
- Independent
  - South African (Schuster et al., 2010), Korean, Japanese, UK (UK10K project), Ireland, Netherlands (GoNL project)
  - *Starting, early phase:* Saudi Arabia, Iran (led by American Iranians)
- Ancient DNA: Neandertal (Green et al., 2010); Denisova (Reich et al., 2010)

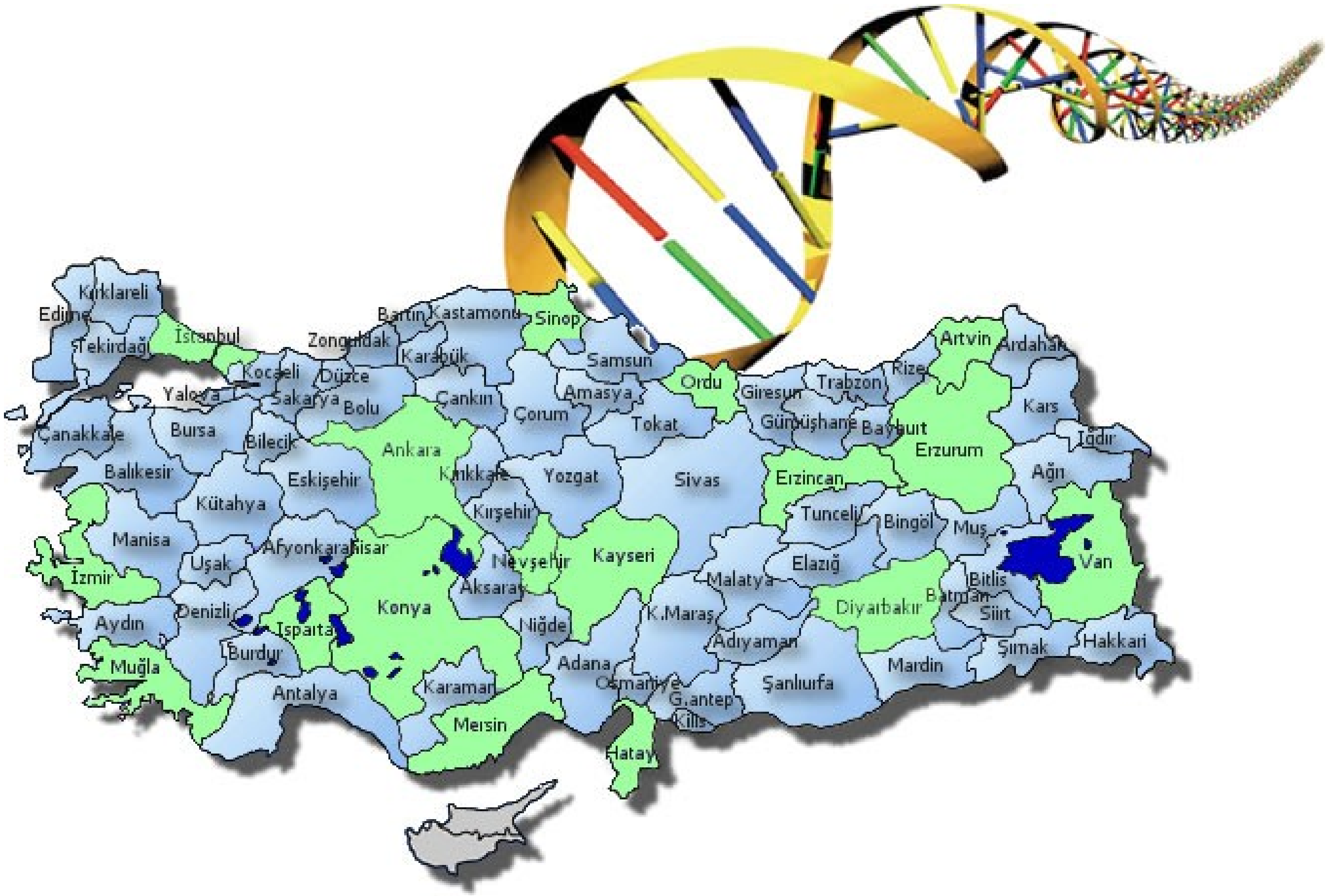
High Throughput Sequencing

- 2007: “Sanger”-based capillary sequencing; one human genome (WGS): ~ $10 million (Levy et al., 2007)
- 2008: First “next-generation” sequencer 454 Life Sciences; genome of James Watson: ~$2 million (Wheeler et al., 2008)
- 2008: The Illumina platform; genome of an African (Bentley et al, 2008) and an Asian (Wang et al., 2008): ~$200K each
- 2009: The SOLiD platform: ~$200K
- Today with the Illumina platform: ~$5K/ genome

Genome Sequence Map of the World



How about Turkey?



17 human genomes from 17 different provinces are sequenced

http://turkiyegenomprojesi.boun.edu.tr