BS222

Genetic variation in human populations

- Discuss the types of genetic variation in human populations
- Explain the mechanisms that give rise to genetic variation, including small scale changes and chromosome structure

Strachan, Goodship and Chinnery (2014) Chapter 4

Every person is unique

- MZ twins start out genetically identical but develop into distinct people
- Aside from that everyone is genetically unique, and not in a small way
- Human population is 7×10^9 persons

How many nucleotides would you need to make that many unique sequences?

- Human genome is 3×10^9 nucleotides

POINT MUTATIONS

- Occasional errors in DNA replication

- Chemical damage due to cellular processes
  - Hydrolysis
  - Oxidative damage
  - Spontaneous deamination of 5-methyl cytosine
  - Abnormal methylation

Tomas Lindahl

- Chemical damage due to external mutagens
  - Ionising radiation
  - UV radiation
  - Polycyclic Aromatic Hydrocarbons (cigarette smoke, exhaust fumes)
  - Other chemicals that interact with DNA or generate ROS

There are repair mechanisms that correct much of this damage, removing damages bases and using the intact strand as a template to synthesis a new section.

We're all mutants

- Germline mutation rate in human is about 10^{-8} per nucleotide per generation
- That means you have about 60 new mutations relative to your parents
- These are almost all neutral
- A few are deleterious, contributing to our mutation load
- Each person probably carries about 10 deleterious mutations

Human genetic variation: terminology

- Point mutations
  - Single base substitution
- Indels
  - Term used for insertions or deletions from 1 to about 50 bp
- Copy number variation
  - Large indels (100 bp to megabases in length)

DNA variants named on basis of frequency in a population:

<0.01 rare variant

>0.01 polymorphism

DEFECTIVE DNA REPAIR

Many genetic disorders arise from mutations in DNA repair enzymes (more than 170), symptoms include:

- Cancer
  - Loss of DNA repair means rapidly dividing tissues are prone to cancer; skin cancer is observed in individuals with Xeroderma pigmentosum
- Progeria
  - Premature aging disorders lead to features of old age appearing in younger people followed by a premature death due to cancer or atherosclerosis
- Neurological damage
  - Neurons are rarely replaced during life and have high ROS generating activity so are very susceptible to DNA damage causing ataxia or learning difficulties
- Immunodeficiency
  - Some DNA repair enzymes function in generating immune diversity so loss of these genes also causes for example severe combined immunodeficiency disease (SCID)

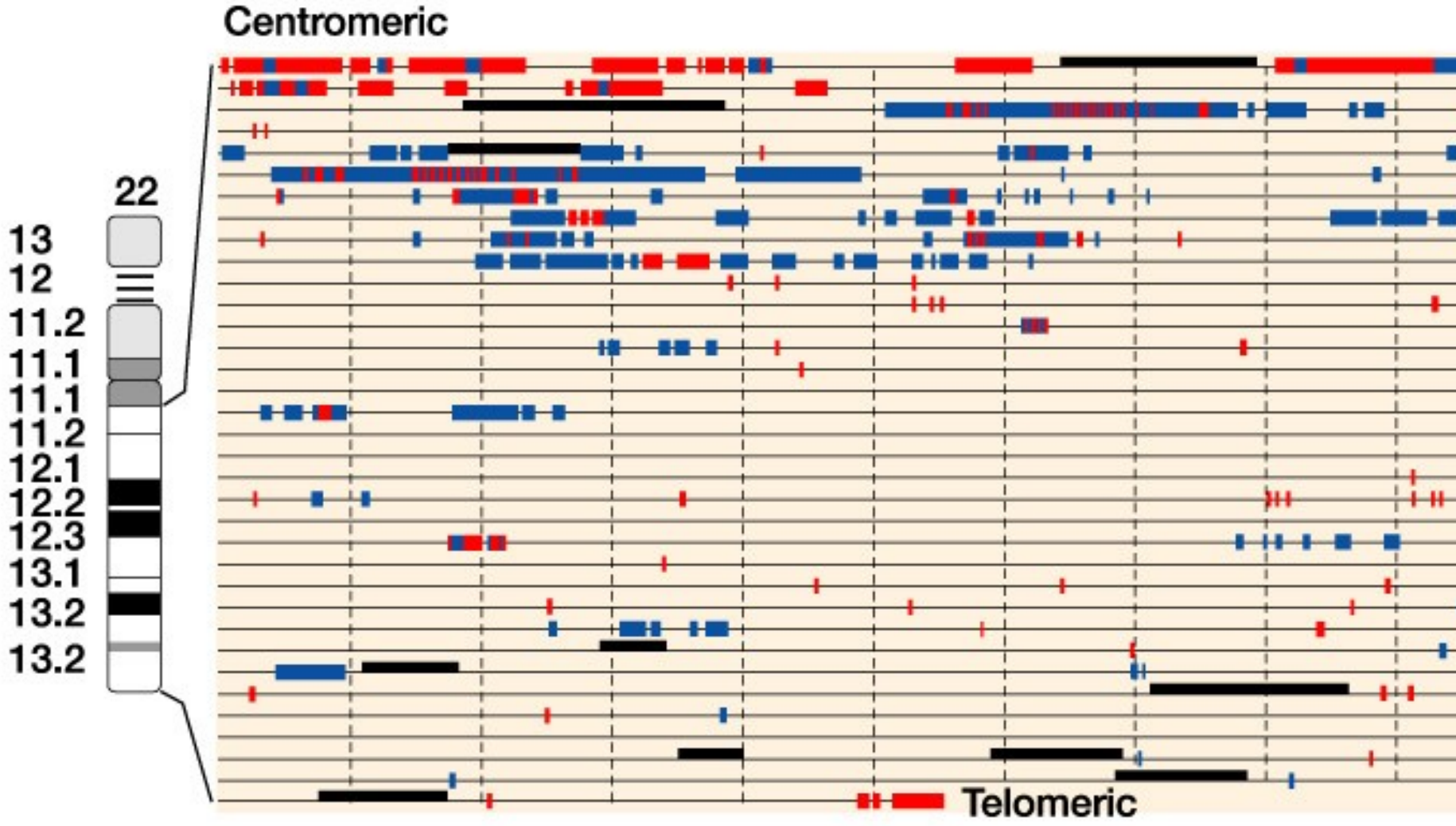
COPY NUMBER VARIATION (CNV)

- occurs by recombination- and replication-based mechanisms
- mutation rates appear much higher for CNVs than for SNPs
- resulting gene dosage, gene disruption, gene fusion, position effects, etc. can cause Mendelian or sporadic traits, or be associated with complex diseases
- can also represent benign polymorphic variants
- mechanism driving gene and genome evolution

Segmental Duplication

- About 5% of the human genome is located in so-called segmental duplications
- relatively large (1- to >200-kbp) regions present in at least two copies per haploid genome and are 90% to 100% identical to each other
- Among the fully sequenced genomes, the human genome has the highest amount of segmental duplications (rodent 2-3%)
- 40 myr but many in last 12 myr

recent segmental duplications on 22q
interchromosomal duplications – red (note: near centromere and telomere)
intrachromosomal duplications - blue



significance of segmental duplications

- difficult to correctly identify and localize in a genome
- can lead to the alignment of *paralogous*, instead of *orthologous*, regions
- segmental duplications are hotspots for chromosomal rearrangements (often associated with genetic disease)
- potential to create “new” genes
- deletions may be important but harder to study

Example: analysis of duplications on chromosome 22

11 cases of transcribed genes that were created either by whole gene duplications, modified by segmental duplications, or contained exons derived from different duplication events

Comparing human genomes

Rapid cheap sequencing methods have made it possible to generate genome sequences from increasing numbers of humans

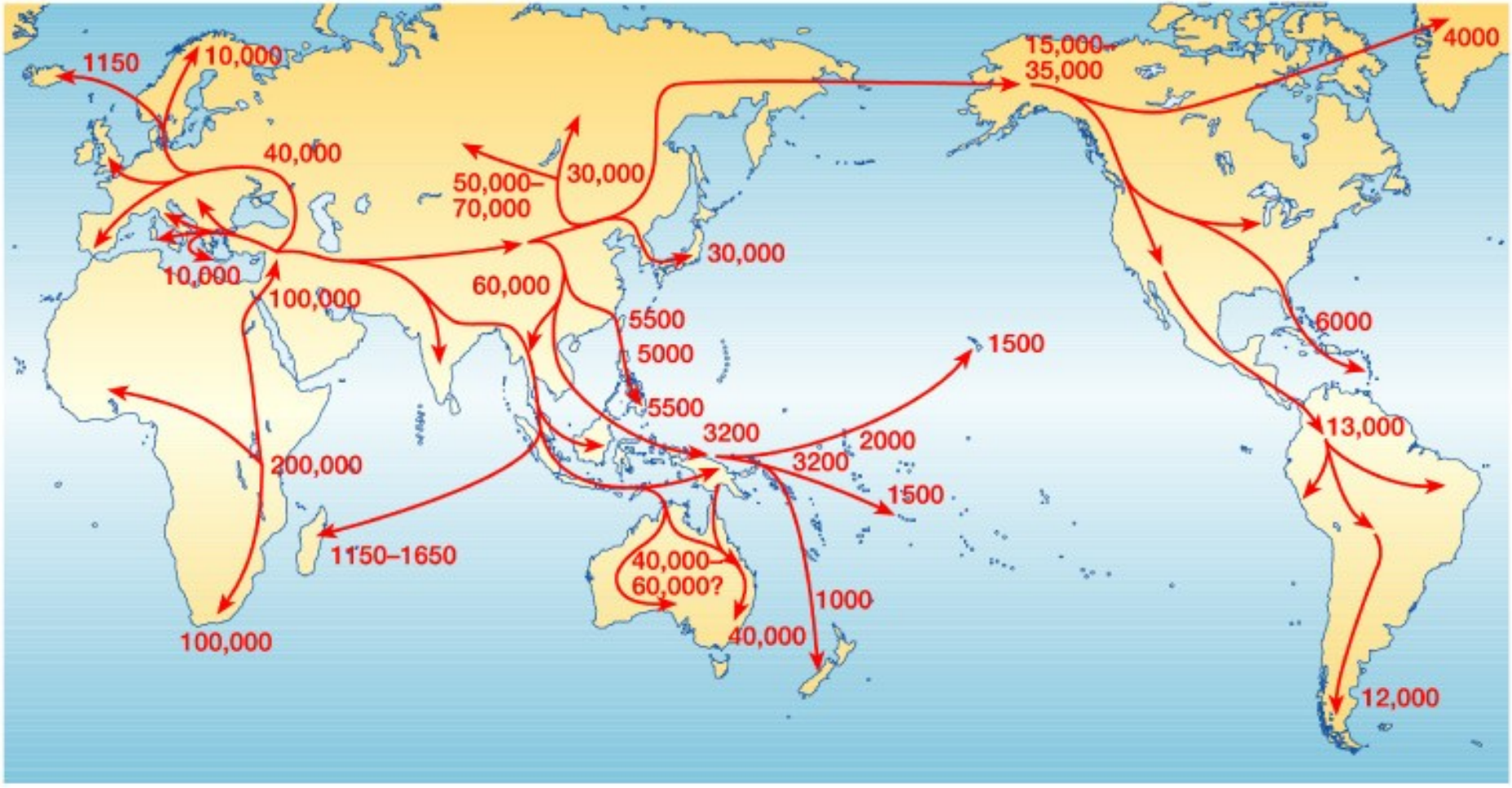
1000 Genomes Project
- expanded to 2500 individuals
- http://www.1000genomes.org/

UK10K Wellcome Trust funded UK project set up in 2010
- 4000 people from the TwinsUK study plus others in long-term monitoring projects; others with serious complex disease
- http://www.uk10k.org/

100,000 genomes project
- http://www.genomicsengland.co.uk/

The aim of human genome sequencing now is to catalogue genetic variability in human populations and to use this data to find correlations between genetic markers and disease.

1000 genome populations chosen from around the globe



Recent African origin model for human migration is supported by paleontological and archaeological evidence

Figure 12-30 Human Molecular Genetics, 3/e. (© Garland Science 2004)

SNPs

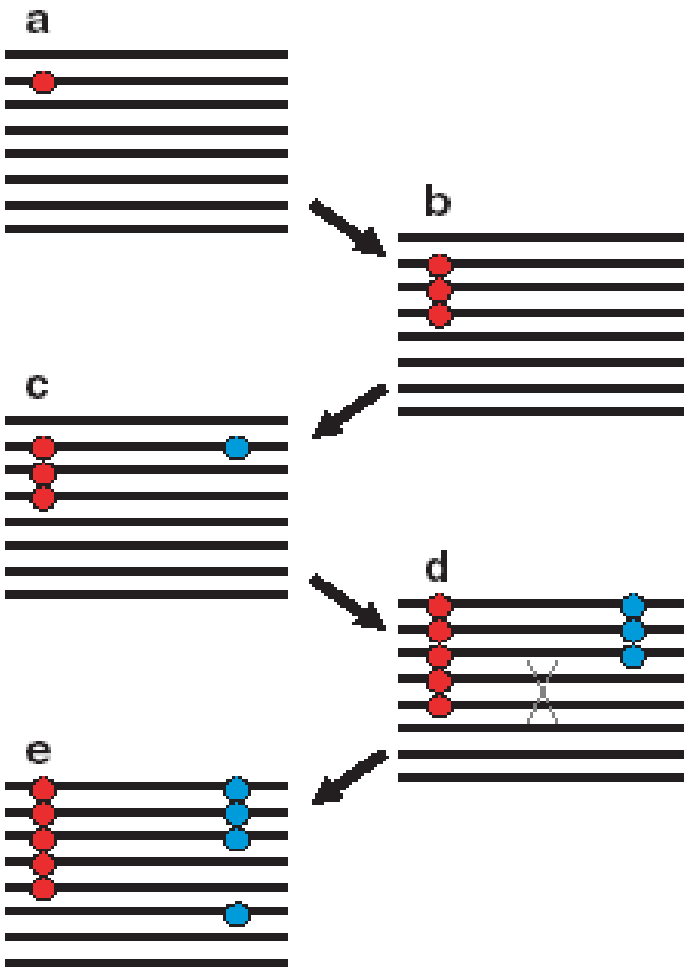
- single nucleotide polymorphisms

these have become the method of choice for studying variation in human populations: there are very large numbers and they are cheap to analyse

- SNPs are biallelic
- frequency 1 in 1000 bases (average)
- result of unrepaired DNA replication errors
- allow very detailed genetic maps to be developed

- combinations of SNPs are called “haplotypes”
- enable population studies of common diseases which have a genetic component

effect of genetic drift and recombination on haplotypes



mutation

genetic drift

2nd mutation

genetic drift
(linkage disequilibrium)

reduced linkage disequilibrium

haplotype: non-random combination of alleles
linkage disequilibrium: non-random association of alleles

International HapMap Project

Scientific rationale

Any two humans are approximately 99.9% identical in their DNA sequences, the 0.1% variation contributes to differences in :

1. risk of diseases
2. responses to drugs, infectious agents, toxins and other environmental factors

- so it must be important and useful to study SNP haplotypes
- October 2007 second generation HapMap published

HapMap Project Populations

270 DNA samples from 4 populations:

- 30 trios (two parents and an adult child) from the Yoruba people (Nigeria)
- 45 unrelated Japanese in the Tokyo area
- 45 unrelated Han Chinese in Beijing
- 30 trios from the Utah population

RELATED PROJECTS

dbSNP is a database of SNPs

- contains 27 million SNPs
- 4 million are in genes
- 6 million are common
- 10 x 500 Kb regions resequenced in 48 of the HapMap individuals by ENCODE (also a number of other gene resequencing projects)

DATABASE	DESCRIPTION	WEBSITE
dbSNP	SNPs and other short genetic variations	http://www.ncbi.nlm.nih.gov/SNP/index.html
dbVar	genomic structural variation	http://www.ncbi.nlm.nih.gov/dbvar/
DGV		http://dgv.tcag.ca/
ALFRED	allele frequencies in human populations	http://alfred.med.yale.edu/alfred/index.asp

Table 4.3 Genetics and Genomics in Medicine (© Garland Science 2015)

WHAT ARE THESE STUDIES LOOKING FOR AND WHY?

- regions that are different between populations
- regions that are very conserved or high linkage disequilibrium

both are candidate regions for selection so the genes in these regions could be important in

- genetic disorders
- mechanisms of disease

SNPs very good for finding *common* disease alleles

SNP analysis detects selection in a human population

- African sample (Nigeria) evidence of selection for two genes linked to the Lassa fever virus endemic in this region (21% of the population show signs of exposure)

- SNP study found strongest signal from LARGE gene; encodes a glycosylase that modifies the receptor for Lassa fever virus
- also a positive signal from DMD gene; this encodes a cytosolic adaptor protein that binds to receptor

- hypothesis: that Lassa fever created selective pressure on LARGE and DMD
- tested by correlating the geographical distribution of the selected haplotype with virus

changes in the approaches to studying genetics of human disease

before the HGP:

- single gene disorders: linkage and positional cloning

e.g. cystic fibrosis, Huntington's disease

now: association studies like HapMap

- identifying loci involved in complex, common human diseases

e.g. heart disease, stroke, diabetes and cancer

UNRESOLVED ISSUES ABOUT SNPs

1. What is the proportion of common variants versus rare variants responsible for common genetic diseases?

2. What proportion of functional polymorphisms involved in complex diseases affects protein structure, gene expression, or splicing?

3. Can we develop robust statistical methods to reliably identify cis-elements using association between SNPs and gene expression data?

4. What is the importance of insertions/deletions or rearrangements relative to SNPs in common ‘genetic’ diseases?