Genetic Variation I

Ira Hall, Ph.D.
Associate Director, McDonnell Genome Institute
Associate Professor, Department of Medicine
Washington University School of Medicine
ihall@wustl.edu
(1) Organizing principles: the forces that shape genetic variation

(2) The landscape of genome variation: definitions and numbers

(3) Genome-wide detection and interpretation of genome variation
The human genome is big
- 3 billion nucleotides, 2 copies.

Mutations are arising constantly
- Roughly 90 per human generation.
- Roughly 1 per somatic cell division.

Human genomes are diverse
- Germline: ~4 million genetic differences between 2 humans; 1 per ~800 bp.
- Cancer: $10^2$-$10^5$ somatic mutations.

Most variants are neutral, or benign
- 1-10% of genome is “functional”.
- But, much buffering and redundancy.
- My guess: 0.1% of variants are “functional” = 4,000 germline variants per person
~4 million germline variants

Environment

Cancer

Common "Complex" Disease

Rare "Mendelian" Disease

Human Diversity

~7,000 exomes

Extremes of phenotypic distribution for NHLBI-relevant traits

NHGRI Medical Sequencing Project

NHLBI Exome Sequencing Project ("ESP-GO")

20+ mendelian diseases

~200 exomes

(+ $10^2 - 10^5$ somatic mutations)
Part I. Organizing principles: forces that shape genetic variation
DNA lesions are arising constantly

Table 6.1 Estimated numbers of DNA lesions induced in human cells each day

<table>
<thead>
<tr>
<th>Source</th>
<th>Lesion</th>
<th>Estimated number of lesions induced per cell/day</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>Spontaneous hydrolysis</td>
<td>SSBs</td>
<td>20 000–40 000(^b)</td>
<td>515</td>
</tr>
<tr>
<td></td>
<td>AP sites</td>
<td>10 000(^b)</td>
<td>513</td>
</tr>
<tr>
<td></td>
<td>Deamination</td>
<td>100–300(^b)</td>
<td>517</td>
</tr>
<tr>
<td>Oxidation</td>
<td>8-oxoG</td>
<td>27 000(^c)</td>
<td>503</td>
</tr>
<tr>
<td></td>
<td>Thymine glycol</td>
<td>270(^d)</td>
<td>531</td>
</tr>
<tr>
<td>Methylation</td>
<td>(N^7)-methylguanine</td>
<td>4000(^b)</td>
<td>543</td>
</tr>
<tr>
<td></td>
<td>(N^3)-methyladenine</td>
<td>600(^b)</td>
<td>543</td>
</tr>
<tr>
<td></td>
<td>(O^6)-methylguanine</td>
<td>10–30(^b)</td>
<td>543</td>
</tr>
<tr>
<td>Glucose</td>
<td>Glucose adducts</td>
<td>3(^b)</td>
<td>541</td>
</tr>
<tr>
<td>Sun exposure</td>
<td>Pyrimidine dimer/6–4 photoprodut</td>
<td>60 000–80 000(^e)</td>
<td>552</td>
</tr>
<tr>
<td>Smoking</td>
<td>PAHs</td>
<td>100–2000(^f)</td>
<td>545–547</td>
</tr>
<tr>
<td>Coke ovens</td>
<td>BaP diol epoxide</td>
<td>7000–70 000(^g)</td>
<td>553</td>
</tr>
<tr>
<td>Radon</td>
<td>SSBs</td>
<td>2(^h)</td>
<td>556</td>
</tr>
</tbody>
</table>

Viijg, Aging of the genome 2007
The vast majority of lesions are repaired through DNA repair pathways. Bulky adducts, single-strand breaks, small base damage, and inter-strand crosslinks are repaired by NER (Nucleotide Excision Repair). Double-strand breaks and mismatches are repaired by BER (Base Excision Repair) and MMR (Mismatch Repair). Telomere attrition is repaired by TERT/RecQ telomerase transcriptase. Non-Homologous End Joining (NHEJ) and Homologous Recombination (HR) also play roles in the repair process.
Methods for assaying germline mutation rate

<table>
<thead>
<tr>
<th>Method</th>
<th>Sequence Sampled</th>
<th>Rate Estimate</th>
</tr>
</thead>
<tbody>
<tr>
<td>Incidence of dominant disorders</td>
<td>Disease genes</td>
<td>1 x 10^{-9} - 2 x 10^{-8} (e.g., Haldane, 1932: Haemophilia, 1 x 10^{-5})</td>
</tr>
<tr>
<td>Species Comparison</td>
<td>Pseudogenes</td>
<td>1 - 4 x 10^{-8}</td>
</tr>
<tr>
<td>Direct observation by sequencing in pedigrees</td>
<td>mtDNA Y chromosome</td>
<td>1 x 10^{-8} - 1 x 10^{-7}</td>
</tr>
</tbody>
</table>
Direct germline mutation rate estimates

Analysis of Genetic Inheritance in a Family Quartet by Whole-Genome Sequencing

Jared C. Roach,1* Gustavo Glusman,1* Arian F. A. Smit,1* Chad D. Huff,1,2* Robert Hubley,2
Paul T. Shannon,2 Lee Rowen,3 Krishna P. Pant,3 Nathan Goodman,3 Michael Bamshad,4
Jay Shendure,5 Radoje Drmanac,6 Lynn B. Jorde,7 Leroy Hood,1† David J. Galas1†
30 APRIL 2010 VOL 328 SCIENCE

These three studies produced rates of 1.1 x 10^{-8}, 1 x 10^{-8}, 1.2 x 10^{-8}, and 1.20 x 10^{-8}

The consensus: 1.2 x 10^{-8}

Variation in genome-wide mutation rates within and between human families

Donald F Conrad1,2, Jonathan E M Keebler3,4, Mark A DePristo5, Sarah J Lindsay1, Yujun Zhang1,
Ferran Casals3, Youssef Idaghdour3, Chris L Hartl5, Carlos Torroja1, Kiran V Garimella, Martine Zilversmit3,
Reed Cartwright6, Guy A Rouleau7, Mark Daly8, Eric A Stone6,6, Matthew E Hurles1 & Philip Awadalla3 for
the 1000 Genomes project8

VOLUME 43 | NUMBER 7 | JULY 2011 NATURE GENETICS

Rate of de novo mutations and the importance of father’s age to disease risk

Augustine Kong1, Michael L. Frigge1, Gisli Masson1, Soren Beesenbacher1,2, Patrick Sulem1, Gisli Magnusson1,
Sigurjon A. Gudjonsson1, Asgeir Sigurdsson1, Aslaug Jonasdottir1, Adalbjorg Jonasdottir1, Wendy S. W. Wong3,
Gunnar Sigurdsson1, G. Bragi Walters1, Stacy Steinberg1, Hannes Helgason1, Guðmör Thorleifsson1, Daniel F. Gudbjartsson1,
Ágmar Helgason1,4, Olafur Th. Magnusson1, Unnur Thorsteinsdottir1,5 & Kari Stefansson1,5

23 AUGUST 2012 | VOL 488 | NATURE | 471
• Mutation rate varies between germline and somatic cells. Somatic cells are 5-20 fold higher.

• Mutation rate varies among individuals & cells
  - Environment
  - Males vs. Females
  - Defective DNA repair genes can cause inherited diseases (e.g., radiation sensitivity; Hereditary nonpolyposis colorectal cancer (HNPCC))
  - Various DNA repair genes are tumor suppressors (e.g., BRCA)
Sex-based mutation rate variation in mammals

Alpha = ratio of male : female mutation rate

The biological basis for sex-biased mutation rates

Replications at puberty:

Replications at fertilization:

30 year old gamete:

60 year old gamete:

<table>
<thead>
<tr>
<th>Gender</th>
<th>Replications at puberty</th>
<th>Replications at fertilization</th>
<th>30 year old gamete</th>
<th>60 year old gamete</th>
</tr>
</thead>
<tbody>
<tr>
<td>Male</td>
<td>35</td>
<td>35 + 23 / year</td>
<td>380</td>
<td>1070</td>
</tr>
<tr>
<td>Female</td>
<td>22</td>
<td>22</td>
<td>22</td>
<td>22*</td>
</tr>
</tbody>
</table>

*But, older females have higher rates of aneuploidy |

From Crow, 2000, *Nature Reviews Genetics*
Direct observation of the mutation rate age effect

Rate of *de novo* mutations and the importance of father’s age to disease risk

Augustine Kong¹, Michael L. Frigge¹, Gisli Masson¹, Soren Besenbacher¹,², Patrick Sulem¹, Gisli Magnusson¹, Sigurjon A. Gudjonsson¹, Asgeir Sigurdsson¹, Aðalbjörg Jonasdottir¹, Adalbjorg Jonasdottir¹, Wendy S. W. Wong³, Gunnar Sigurdsson¹, G. Bragi Walters¹, Stacy Steinberg⁴, Hannes Helgason⁵, Gudmar Thorleifsson⁶, Daniel F. Gudbjartsson¹, Arnar Helgason⁷, Olafur Th. Magnusson¹, Únnum Thorsteinsdottir⁸ & Karl Stefansson¹,²

23 August 2012 | Vol 488 | Nature | 471

- 78 trios sequenced
- Variation in father’s age explains 94% of rate variation
- Dad contributes ~2 new mutations per year
- Mom contributes a fixed number of ~14 mutations
- On average, alpha is ~4

![Graph showing the relationship between number of de novo mutations and age of father at conception of child.](image)
Somatic mutation rates are tissue dependent.

Possible sources of differences: replicative age, mutagens.

Vijg, Aging of the genome 2007
Somatic mutations as markers for lineage tracing?

The *C. elegans* cell lineage

**The logic:** in human, there is ~1 mutation per somatic cell division. Thus, by comprehensively defining mutations among somatic cells, we should be able to learn how cells from different parts of the body are related to one another. This would inform models of development and aging.

*Note: John Sulston won the Nobel prize for this work (2002). It was based on meticulous cell biology, not DNA.*
Sequencing somatic cells to learn about development

Genome sequencing of normal cells reveals developmental lineages and mutational processes

Sam Behjat1,2, Meritzell Huch2*, Ruben van Boxtel3*, Wouter Karthaus4*, David C. Wedge5, Asif U. Tamuri6, Iñigo Martincorena1, Mia Petljak7, Ludmil B. Alexandrov8, Gunes Gundem9, Patrick S. Tarpey10, Sophie Roerink11, Joyce Blokker12, Mark Maddison13, Laura Mudle14, Ben Robinson15, Serena Nik-Zainal16, Peter Campbell17, Nick Goldman1, Marc van de Wetering18, Edwin Cappen1, Hans Clevers9 & Michael R. Stratton1

422 | NATURE | VOL 513 | 18 SEPTEMBER 2014

**Design:**
- “clone” single somatic cells via organoid tissue culture
- Twenty-five lines obtained from the stomach, small bowel and large bowel of two mice
- Whole genome sequencing

**Findings:**
- Different daughter cells from early divisions can contribute unequally.
- 6714 somatic SNVs discovered.
- Total 1.1 mutations per cell division.
- More in small bowel consistent with more divisions in this tissue.
- Early mutations C→T at CpG; bowel rich in C→A (reactive oxygen?)
Sequencing somatic cells to learn about development

Genome sequencing of normal cells reveals developmental lineages and mutational processes

Design:
- “clone” single somatic cells via organoid tissue culture
- Twenty-five lines obtained from the stomach, small bowel and large bowel of two mice
- Whole genome sequencing

Findings:
- Different daughter cells from early divisions can contribute unequally.
- 6714 somatic SNVs discovered.
- Total 1.1 mutations per cell division.
- More in small bowel consistent with more divisions in this tissue.
- Early mutations C→T at CpG; bowel rich in C→A (reactive oxygen?)
Sequencing somatic cells to learn about development

Genome sequencing of normal cells reveals developmental lineages and mutational processes

Design:
- “clone” single somatic cells via organoid tissue culture
- Twenty-five lines obtained from the stomach, small bowel and large bowel of two mice
- Whole genome sequencing

Findings:
- Different daughter cells from early divisions can contribute unequally.
- 6714 somatic SNVs discovered.
- Total 1.1 mutations per cell division.
- More in small bowel consistent with more divisions in this tissue.
- Early mutations C→T at CpG; bowel rich in C→A (reactive oxygen?)
Mutational diversity in cancer

Mutational heterogeneity in cancer and the search for new cancer-associated genes

214 | Nature | Vol 499 | 11 July 2013

![Graph showing mutational diversity in cancer](image_url)
**Clonal evolution**

**Sources of genetic mutation**

1) normal processes of DNA replication & cell division
2) Genomic instability = increased mutation rate
3) Environmental mutagens: UV, tobacco, etc.

**Sources of natural selection**

1) Growth rate
2) Apoptosis/senescence
3) Competition for limited resources
4) Resistance to drug treatment

..........many other potential sources

Each cancer is a unique evolutionary experiment!
Clonal evolution generates intra-tumor heterogeneity

**KEY POINT:** A tumor is not a single entity, it is a collection of related cell lineages. The number of lineages can vary dramatically depending on time, mutation rate, selective pressure, and stochastic processes. Related lineages may have different properties and may compete and/or cooperate. This process is not well understood.

Adapted from Campbell et al., Nature (2010)
Two extreme examples: imagine everything in between

“monoclonal” evolution

“polyclonal” evolution

A homogenous tumor

A heterogenous tumor

Navin et al., Mol. Oncology (2010)
Single cell sequencing reveals fine-scale heterogeneity

Trees of genetic relatedness between single cells!!!

Digital DNA sequencing data reveals intra-tumor allele frequencies. Clever algorithms can infer clonality.

Intra-tumor variant allele frequency (VAF) can be estimated by:
number reads identifying the variant base / total reads aligning to that base

This yields an estimate of the fraction of chromosomes in a tumor that carry the variant, as determined by:

1) The fraction of tumor cells that carry the variant.
2) The genotype of the variant in those cells (e.g., heterozygous or homozygous).

**Germline versus somatic mutation**

**Germline mutation:** genetic mutations that arise in the egg or sperm, or in the cell lineages that give rise to germ cells, and are transmitted to progeny. These underlie numerous sporadic human disorders.

**Somatic mutation:** genetic mutations that arise after fertilization and are present in a subset of somatic cells.
There is not always a clear distinction between germline and somatic mutations

Parental Somatic Mosaicism Is Underrecognized and Influences Recurrence Risk of Genomic Disorders

Ian M. Campbell,1,11 Bo Yuan,1,13 Caroline Robberecht,2 Rolph Pfundt,3 Przemyslaw Szafrański,1 Meriel E. McEntagart,4 Sandesh C.S. Nagamani,1,5 Ayelet Erez,1,5 Magdalena Bartnik,6 Barbara Wiśniewiecka-Kowalnik,6 Katie S. Plunkett,1 Amber N. Pursley,1 Sung-Hae L. Kang,1 Weimin Bi,1 Seema R. Lalan,1,5 Carlos A. Bacino,1,5 Mala Vast,4 Karen Marks,4 Michael Patton,4 Peter Olofsson,7 Ankita Patel,1 Joris A. Veltman,3 Sau Wai Cheung,1 Chad A. Shaw,1 Lisenka E.L.M. Vissers,3 Joris R. Vermeesch,2 James R. Lupski,1,5,8,9,* and Paweł Stankiewicz1,10,*

The American Journal of Human Genetics 95, 173–182, August 7, 2014

- Screened 100 cases of genomic disorders caused by de novo mutation in a child (with normal parents)
- 4% were detectable in blood of a parent.
- These families will have high risk of disease recurrence
Clonal evolution in the germline: spermatogonial selection

Paternal age effect disorders:
- Apert syndrome (caused by FGFR2 mutations)
- Achondroplasia, and thanatophoric dysplasia (FGFR3)
- Costello syndrome (HRAS)

Population bottlenecks decrease genetic diversity

*From Graham Coop's Website: http://gcbias.org/*
Finland Population History

Early Settlement
- 2,000-10,000 years ago
- South and Coast

Late Settlement
- 16th century
- Multiple bottlenecks

Expansion
- 18th century (pop 250K)
- Today (pop 5.3M)
Real world implications

Finnish heritage disease

Finnish heritage diseases include:

- APECED (autoimmune polyendocrinopathy–candidiasis–ectodermal dystrophy)
- Aspartylglucosaminuria, a lysosomal storage disease
- Congenital adrenal hyperplasia[^6]
- Congenital nephrotic syndrome,[^7] a kidney disease of newborn babies
- Congenital chloride diarrhea
- Congenital stromal corneal dystrophy
- GRACILE syndrome
- Lethal arthrogryposis with anterior horn cell disease
- Lethal congenital contracture syndrome 1
- Meretoja syndrome
- Meckel syndrome
- Myotonia congenita
- Nonketotic hyperglycinemia
- Salla disease,[^8] a lysosomal storage disease
- PEHO syndrome
- Rapadilino syndrome
- Retinoschisis
- Usher syndrome
Due to population history, there are proportionally more LoF variants in Finnish individuals compared to non-Finnish Europeans.
Some other forces that shape genetic diversity

Many other population genetic and evolutionary forces affect patterns of genetic diversity, either in a locus-specific manner, or across the entire genome.

- Population dynamics: rapid growth, effective size
- Natural selection can produce regions of low or high diversity, depending on forces involved (e.g., negative, positive or balancing selection)
- Strong artificial selection can dramatically change patterns of genetic diversity (e.g., domestic plants and animals)
Genomic signatures of positive selection

a  Classic selective sweep

Neutral variation

An advantageous mutation arises

Over time, the advantageous mutation approaches fixation

b  Selection from standing variation

Neutral variation

A variant becomes adaptive in a new environment

Over time, the advantageous mutation approaches fixation

c  Selection on a complex trait

Neutral variation

A set of variants becomes adaptive in a new environment

Over time, the set of variants becomes more common

Scheinfeldt and Tishkoff, Nature Reviews Genetics (2013)
Genetic variation is shuffled by recombination. But, recombination occurs predominantly at hotspots.

Intensely punctate meiotic recombination in the class II region of the major histocompatibility complex

Alec J. Jeffreys¹, Liisa Kauppi¹ & Rita Neumann¹

tab 29 • October 2001
Recombination hotspots shape haplotype blocks

High-resolution haplotype structure in the human genome

Mark J. Daly¹, John D. Rioux¹, Stephen F. Schaffner¹, Thomas J. Hudson¹,² & Eric S. Lander¹,³

nature genetics • volume 29 • october 2001

haplotype: a set of DNA variations, or polymorphisms, that tend to be inherited together
Domestic Dog: bottlenecks + strong artificial selection

Karlsson and Lindblad-Toh, 2008, Nature Reviews Genetics
Part II. The landscape of genome variation: definitions and numbers
Key sources of genome variation

(1) “Point” mutations

- Single nucleotide variant/polymorphism (SNV/SNP)
  - Substitution
  - Your genome: GCATCGGCTCCGTCTAATGAACCTAG
  - My genome: GCCCTCGGCTCCGTCTAATGAACCTAG

- Indels (< 50 bp)
  - Insertion
  - Deletion

(2) Structural variation (> 50bp):
- Copy number variants (CNVs): deletion, duplication or amplification of large chromosomal segments
- Genomic rearrangements: translocations, inversions, complex

(3) Aneuploidy

(4) Transposons

(5) Simple repeats
Single nucleotide variants/polymorphisms

Key Definitions:

Single nucleotide variant (SNV): a single base substitution variant (e.g., A → G)

Single nucleotide polymorphism (SNP): an SNV that is relatively common in the human population, defined as variant allele frequency >= 1%.

Variant allele frequency: The fraction of chromosomes in a population that carry a given genetic variant (not the number of people, or cells)

Key Facts:

• ~4 million in each individual human (relative to the reference genome). By far the most common class of genome variation.

• 3.2 Gbp genome = 1/800bp = 0.13% divergence
  - for context, human vs. chimp is 1%

• 13 million “common” SNPs in the human population (>1% allele frequency)

• Impact: in a typical personal genome, ~22,000 SNVs affect exons, ~10,000 change coding sequence (non-synonymous); ~90 stop codons

• Very useful for genetic mapping
  - abundant, stable inheritance, easy and cheap to genotype (microarrays)
Short insertions and deletions (indels)

**Indel**: a short insertion or deletion relative to the reference genome, <50 bp in size.

**Key Facts:**

- ~400,000 are identified in each personal genome; ~400 are in exons
- Indels are a key source of gene loss of function (LOF) mutations due to their ability to cause “frameshifts” in the coding sequence.
- In a typical personal genome, we identify ~90 frameshift mutations, 24 of which are homozygous
**Structural Variation (SV):** Differences in the copy number, orientation or location of "large" genomic segments (>50 bp). Includes deletions, duplications, inversions, insertions, translocations & complex rearrangements.

**Copy number variation:** SVs that involve a change in DNA copy number. CNV is often used specifically to refer to large multi-allelic CNVs present as tandem arrays.

**Prevalence:** 5,000-10,000 germline SVs in a "normal" genome, ~80% of which are small deletions. Tens to hundreds of SVs in a tumor genome, of diverse types.

**Impact:** Although much more rare than SNPs and indels, thought to be more impactful. Due to large size they affect more total base-pairs in any individual genome.

**Cancer:** SVs play a central role in many cancers: amplification of oncogenes, deletion of tumor suppressors, gene fusions, etc.
Aneuploidy: Changes in the copy number of entire chromosomes.
monosomy = 1 copy; trisomy = 3 copies; uniparental disomy = 2 copies from 1 parent

Key Facts:
• Extremely common in human cancers. Majority of tumors have aneuploidy due to chromosome mis-segregation
• In humans, most aneuploidies are embryonic lethal. Exceptions are trisomy 21 (Down Syndrome), 18 (Edwards Syndrome; rarely survive) and 13 (Pateau syndrome; rarely survive). Others can survive if mosaic or sub-chromosomal.
• Aneuploidy is a major cause of miscarriage and the main reason for prenatal screening. Increased risk with female age.
Mobile element insertions (MEIs)

- Caused by retrotransposition. DNA → RNA → DNA.
- 500-1000 variable MEIs exist in the typical human genome. 90% are Alu, 8% are L1, rest are SVA and HERV-K.
- Activity and element types vary widely across mammalian species. For example, transposons are >10X more active in mouse than in human. For example, bats have an active DNA transposon not found in any other lineage.
Retrogene insertions = “retroduplications”

- Caused by retrotransposon machinery acting on endogenous gene transcripts
- Results in insertion of gene without introns
- Long recognized to be a source of pseudogenes during evolution
- 48 “gene retrocopy insertion polymorphisms” (GRIPs) found in 1000 Genomes data (Ewing et al., 2013, Genome Biology; Abyzov et al., 2014, Genome Research).
- 174 “retroduplications” found in same data by Abyzov et al., 2013, Genome Research
- Unknown how many are variable between two humans, likely ~10 (my guess)
Retroelements can carry flanking sequence when they jump.

Emergence of primate genes by retrotransposon-mediated sequence transduction

Jinchuan Xing†, Hui Wang†, Victoria P. Belancio†, Richard Cordaux†, Prescott L. Deininger‡, and Mark A. Batzer†§

†Department of Biological Sciences, Biological Computation and Visualization Center, Center for BioModular Multi-Scale Systems, Louisiana State University, 202 Life Sciences Building, Baton Rouge, LA 70803; and ‡Tulane Cancer Center SL-66, Department of Environmental Health Sciences, Tulane University Health Sciences Center, New Orleans, LA 70112

Edited by Susan R. Wessler, University of Georgia, Athens, GA, and approved June 28, 2006 (received for review April 20, 2006)

SVA has duplicated 53kb of sequence in past 25M years, including 3 duplications of entire AMAC gene!
Variable number tandem repeats (VNTRs)

Many classes, terms and unclear definitions

- **microsatellite**: 1-6 bp repeat unit, ~16,000 in human genome
- **minisatellite**: 10-100 bp repeat unit, >1000 in human genome
- **satellite**: centromeres, telomeres and heterochromatin (e.g., 171 bp)
- **VNTRs** = variable micro- & minisatellites (Sir Alec Jeffreys)
- In general, high mutation rates

Classes of simple sequence repeats (SSRs)
Why “satellite”?

http://www.umanitoba.ca/afs/plant_science/courses/PLNT3140/II4/II4.html]
Microsatellites

VNTR families are annotated as $([ACGT])_n$
- Core repeat unit, then number of copies “n”
- For example, $(AC)_n$ represents all dinucleotide AC repeats

Microsatellites are VNTRs of units 1-6bp in length

<table>
<thead>
<tr>
<th>Repeat unit</th>
<th>Sequence properties</th>
<th>Utility</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Mostly poly(A)/poly(T)</td>
<td>Problematic for PCR, next-gen seq</td>
</tr>
<tr>
<td>2</td>
<td>$(AC)_n/(GT)_n$ most common, $(GC)_n$ very rare</td>
<td>Problematic for PCR</td>
</tr>
<tr>
<td>3</td>
<td>Wide range of units $(AAT)_n$ and $(AAC)_n$</td>
<td>Little ‘stutter’, widely used in earlier human genetics studies</td>
</tr>
<tr>
<td>4</td>
<td>Wide range $(AAAC)$ and $(AAAT)$ most common</td>
<td>Little ‘stutter’, widely used</td>
</tr>
<tr>
<td>5</td>
<td>Range</td>
<td></td>
</tr>
</tbody>
</table>
Microsatellites mutate by replication slippage

START: 9 repeat allele

Slippage

Realignment

Misalignment

Extension

+1 REPEAT MUTATION
10-repeat allele after subsequent DNA replication

-1 REPEAT MUTATION
8-repeat allele after subsequent DNA replication

Repeat unit
DNA polymerase
Minisatellites mutate by recombination

- 10-100bp core sequences
- Not just big microsatellites
- “Scars” of sustained, localized recombination
- Hypervariable minisatellites may have highest mutation rate of any element, 14% per generation

Polarized mutation is sometimes observed, where one side of VNTR is more mutable than the other, perhaps due to flanking recombination hotspots.
DNA Fingerprinting

Heterozygosity = probability that two alleles are different

Minisatellites are highly heterozygous, typically over 90%

Basis of “DNA fingerprinting” (Sir Alec Jeffreys)

Now routinely done using PCR and microsatellites
Satellites: the final frontier of our genome

- Alpha satellite is most common, 170bp repeat unit, present in the majority of eukaryotic centromeres.
- Hundreds to thousands of copies at each human centromere, with complex organization.
- There is reported variability in Alpha satellite array length between humans, but it is not known whether or not this affects centromere function.
Rates of mutation for different variant classes

**Note**: this general landscape is thought to be relatively consistent across mammals, except for transposons, whose activity and element composition vary widely.