Genetic Variation I
Genomics: Bio5488, Spring 2021

Sheng Chih (Peter) Jin, Ph.D.
Assistant Professor of Genetics
Washington University School of Medicine
jin810@wustl.edu

(with thanks to Ira Hall, Don Conrad and slides from past years)
Outline

• Organizing principles: the forces that shape genetic variation

• The landscape of genome variation: definitions and numbers

• Genome-wide detection and interpretation of genome variation
The scope of the problem: basic genome facts

• The human genome is big
  - 3 billion nucleotide, 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 ~800bp.
  - 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.

https://bio.libretexts.org/
Genetic variations underlie phenotypic differences

~4 million germline variants

Human diversity

Rare ‘Mendelian’ disorder

Environment

Common ‘Complex’ disease

Cancer-related somatic mutation

https://doi.org/10.1016/j.cell.2013.09.001

https://doi.org/10.1016/j.ymeth.2019.11.002
Outline

• Organizing principles: the forces that shape genetic variation

• The landscape of genome variation: definitions and numbers

• Genome-wide detection and interpretation of genome 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</td>
<td>515</td>
</tr>
<tr>
<td></td>
<td>AP sites</td>
<td>10 000</td>
<td>513</td>
</tr>
<tr>
<td></td>
<td>Deamination</td>
<td>100–300</td>
<td>517</td>
</tr>
<tr>
<td>Oxidation</td>
<td>8-oxoG</td>
<td>27 000</td>
<td>503</td>
</tr>
<tr>
<td></td>
<td>Thymine glycol</td>
<td>270</td>
<td>531</td>
</tr>
<tr>
<td>Methylation</td>
<td>N^7^-methylguanine</td>
<td>4000</td>
<td>543</td>
</tr>
<tr>
<td></td>
<td>N^3^-methyladenine</td>
<td>600</td>
<td>543</td>
</tr>
<tr>
<td></td>
<td>O^6^-methylguanine</td>
<td>10–30</td>
<td>543</td>
</tr>
<tr>
<td>Glucose</td>
<td>Glucose adducts</td>
<td>3</td>
<td>541</td>
</tr>
<tr>
<td>Sun exposure</td>
<td>Pyrimidine dimer/6–4 photoprotein</td>
<td>60 000–80 000</td>
<td>552</td>
</tr>
<tr>
<td>Smoking</td>
<td>PAHs</td>
<td>100–2000</td>
<td>545–547</td>
</tr>
<tr>
<td>Coke ovens</td>
<td>BaP diol epoxide</td>
<td>7000–70 000</td>
<td>553</td>
</tr>
<tr>
<td>Radon</td>
<td>SSBs</td>
<td>2</td>
<td>556</td>
</tr>
</tbody>
</table>

Vijg, Aging of the genome 2007
The vast majority of lesions are repaired

Fig. 4.4 Schematic depiction of the main DNA-repair pathways in mammalian cells subdivided on the basis of the specific forms of DNA damage that prompt their action. TERT, telomerase reverse transcriptase.

Vijg, Aging of the genome 2007
**Methods for assaying germline mutation rate**

<table>
<thead>
<tr>
<th>Methods</th>
<th>Sequence Sampled</th>
<th>Rate Estimate</th>
</tr>
</thead>
<tbody>
<tr>
<td>Incidence of dominant disorders</td>
<td>Disease gene</td>
<td>$1 \times 10^{-9} - 2 \times 10^{-8}$ (e.g. Haldane, 1932: Haemophilia, $1 \times 10^{-5}$)</td>
</tr>
<tr>
<td>Species comparison</td>
<td>Pseudogenes</td>
<td>$1 - 4 \times 10^{-8}$</td>
</tr>
<tr>
<td>Direct observation by sequencing in pedigrees</td>
<td>mtDNA</td>
<td>$1 \times 10^{-8} - 1 \times 10^{-7}$</td>
</tr>
<tr>
<td></td>
<td>Y chromosome</td>
<td></td>
</tr>
</tbody>
</table>

**DNA sequences**

- **Chimp**: ...ATCGGCTGG...
- **Human major**: ...ATCGGCTGG...
- **Human minor**: ...ATCGCCTGG...

**Diagram**

- Pedigree for incidence of dominant disorders.
- Pedigree for direct observation by sequencing in pedigrees.
Direct germline mutation rate estimates

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

Jared C. Roach,1* Gustavo Giusman,1* Arian F. A. Smit,1* Chad D. Huff,1,2* Robert Hubley,2
Paul T. Shannon,3 Lee Rowen,4 Krishna P. Pant,5 Nathan Goodman,5 Michael Bamshad,6
Jay Shendure,7 Radoje Ormanac,3 Lynn B. Jorde,6 Leroy Hood,5† David J. Galas5†

30 April 2010 | Vol 328 | Science

These three studies produce rates of $1.1 \times 10^{-8}$, $1 \times 10^{-8}$, and $1.2 \times 10^{-8}$

The consensus: $1.2 \times 10^{-8}$
Mutation is a quantitative trait

• 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 radiation (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

\[ \alpha = \text{ratio of male : female mutation rate} \]

The biological basis for sex-biased mutation rates

<table>
<thead>
<tr>
<th></th>
<th>Male</th>
<th>Female</th>
</tr>
</thead>
<tbody>
<tr>
<td>Replications at puberty</td>
<td>35</td>
<td>22</td>
</tr>
<tr>
<td>Replications at fertilization</td>
<td>35 + 23 / year</td>
<td>22</td>
</tr>
<tr>
<td>30 year old gamete:</td>
<td>380</td>
<td>22</td>
</tr>
<tr>
<td>60 year old gamete:</td>
<td>1,070</td>
<td>22*</td>
</tr>
</tbody>
</table>

*But, older females have higher rate 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, Gislí Masson¹, Søren Besenbacher¹,², Patrick Salem¹, Gislí Magnússon¹,
Sigurjón A. Gudjónsson¹, Ásgeir Sigurðsson¹, Aslaug Jonasdóttir¹, Adalbjörg Jonasdóttir¹, Wendy S. W. Wong³,
Gunnar Sigurðsson¹, G. Bragi Walters¹, Stacy Steinberg¹, Hannes Helgason¹, Gudmar Thorleifsson¹, Daniel F. Gudbjartsson¹,
Agnar Helgason¹,², Olafur Th. Magnússon¹, Unnur Thorsteinsdóttir¹ &³ & Kari Stefánsson¹,²

23 August 2012 | Vol 488 | Nature | 473

- 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
Somatic mutation rates are tissue dependent

Possible sources of differences: replicative age, mutagens

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

The C. elegans cell lineage

*Note: John Sulston won the Nobel prize for this work (2002). It was based on meticulous cell biology, not DNA.

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.
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.
- 6,714 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 (cont.)
Mutational diversity in cancer

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

214 | Nature | Vol 499 | 11 July 2013
Clonal evolution

Sources of genetic mutation:
• Normal processes of DNA replication & cell division
• Genomic instability = increased mutation rate
• Environmental mutagens: UV, tobacco, etc.

Sources of natural selection:
• Growth rate
• Apoptosis/senescence
• Competition for limited resources
• 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, but a collection of related cell lineages. The number of lineages can vary dramatically depending on time, mutation rate, selective pressure, and stochastic process. 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

A homogeneous tumor

A heterogeneous tumor

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

Trees of genetic relatedness between single cells!!

Navin et al., Nature 2011
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 of reads identifying the variant base / total reads aligning to that base

This yield 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 that variant.
2. The genotype of the variant in those cells (e.g., heterozygous or homozygous)

[Welch et al., Cell 2012]
**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, Bo Yuan, Caroline Robberecht, Rolph Pfundt, Przemyslaw Szafranski, Meriel E. McEntagart, Sandesh C.S. Nagamani, Ayelet Erez, Magdalena Bartnik, Barbara Wiśniewska-Kowalnik, Katie S. Plunkett, Amber N. Pursley, Sung-Hae L. Kang, Weimin Bi, Seema R. Lalani, Carlos A. Bacino, Mala Vast, Karen Marks, Michael Patton, Peter Olofsson, Ankita Patel, Joris A. Veltman, Sau Wai Cheung, Chad A. Shaw, Lisinka E.L.M. Vissers, Joris R. Vermeesch, James R. Lupski, and Paweł Stankiewicz

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

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

Parental age effect disorders:
- Apert syndrome (caused by \textit{FGFR2} mutations)
- Achondroplasia, and thanatophortic dysplasia (\textit{FGFR3})
- Costello syndrome (\textit{HRAS})

The bottleneck effect is an extreme example of genetic drift that happens when the size of a population is severely reduced. Events like natural disasters (earthquakes, floods, fires) can decimate a population, killing most individuals and leaving behind a small, random assortment of survivors.

*From Graham Coop’s Website: http://gcbias.org/*
A real world example

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)
Finnish heritage disease types

There are 36 identified Finnish heritage diseases.[8][10]

- Amyloidosis, Finnish type
- Lethal arthrogryposis with anterior horn cell disease
- Aspartylglucosaminuria
- Autoimmune polyendocrinopathy syndrome, type I, with or without reversible metaphysisal dysplasia
- Cartilage–hair hypoplasia
- Ceroid lipofuscinosis, neuronal, 1
- Ceroid lipofuscinosis, neuronal, 3
- Ceroid lipofuscinosis, neuronal, 5
- Ceroid lipofuscinosis, neuronal, 8, Northern epilepsy variant (Synonyms: Northern epilepsy; Epilepsy, progressive, with mental retardation)
- Choroideremia
- Cohen syndrome
- Cornea plana 2
- Diarrhea 1, secretory chloride, congenital
- Diastrophic dysplasia
- Epilepsy, progressive myoclonic 1A (Unverricht–Lundborg)
- Glycine encephalopathy (Nonsyndromic hyperglycinemia)
- GRACILE syndrome
- Gyrate atrophy of choroid and retina
- Hydrolethalus syndrome 1
- Infantile-onset spinocerebellar ataxia (Mitochondrial DNA depletion syndrome 7)
- Lactase deficiency, congenital
- Lethal congenital contracture syndrome 1
- Lysinuric protein intolerance
- Meckel syndrome
- Megaloblastic anemia-1, Finnish and Norwegian type
- Mullbrey nanism
- Muscular dystrophy-dystroglycanopathy (congenital with brain and eye anomalies), type A, 3
- Nephrotic syndrome, type 1 (Finnish congenital nephrosis)
- Ovarian dysgenesis 1
- Polycystic lipomembranous osteodysplasia with sclerosing leukoencephalopathy (Nasu–Hakola disease)
- Progressive encephalopathy with Edema, Hypsarrhythmia and Optic atrophy
- RAPADILINO syndrome
- Retinoschisis 1, X-linked, juvenile
- Sialuria, Finnish type (Salla disease)
- Tibial muscular dystrophy, tardive
- Usher syndrome, type 3A
Due to population history, there are proportionally more loss-of-function variants in Finnish individuals compared to non-Finnish Europeans.

Lim et al, 2015, DOI: 10.1371/journal.pgen.1004494
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

nature genetics • volume 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

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

• Organizing principles: the forces that shape genetic variation

• The landscape of genome variation: definitions and numbers

• Genome-wide detection and interpretation of genome variation
Key resource of genome variation

(1) “Point” mutations:
- Single nucleotide variant/polymorphism (SNV/SNP)
- Indels (< 50 bp)

(2) Structural variation (> 50 bp)
- 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 -> C)
- **Single nucleotide polymorphism (SNP):** an SNV that is relatively common in the human population, defined as variant allele frequency \( \geq 1\% \)
- **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:

- \( \sim 4 \) million in each individual human (relative to the reference genome). By far the most common class of genome variation.
- \( 3.2 \) Gbp genome = \( 1/800\)bp = \( 0.13\% \) divergence (for context, human vs. chimp is \( 1\% \))
- 13 million “common” SNPs in the human population \((\geq 1\% \) allele frequency\()
- Impact: in a typical personal genome, \( \sim 22k \) SNVs affect exons, \( \sim 10k \) change coding sequence (non-synonymous); \( \sim 90 \) stop codons
- Very useful for genetic mapping (abundant, stable inheritance, easy and cheap to genotype with microarrays)
**Short insertions and deletions (indels)**

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

**Reference genome:** TCTAATGAATCTAG——-CCCAGAGCATCGGCTCTGCAATGCC

**Your genome:** TCTAATGAATCTAGGATCCAGAGCATCGGCTC——CAATGCC

**Key facts:**
- ~400k are identified in each personal genome; ~400 are in exons.
- Indels are a key source of gene loss of function mutations due to their ability to cause “frameshifts” in the coding sequence.
- In a typical human 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, and complex rearrangements.

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

**Prevalence:** 5k–10k 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 rarer 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**

**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)

Key facts:
- Caused by retrotransposition. DNA -> RNA -> DNA.
- 500-1,000 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. 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., 2014 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, ~16k in human genome
- **Minisatellite**: 10-100 bp repeat unit, > 1k in human genome
- **Satellite**: centromeres, telomeres, and heterochromatin (e.g., 171 bp)
- **VNTR**: variable micro- & minisatellites (Sir Alec Jeffreys)
- In general, high mutation rates
Why “satellite”? 


http://www.umanitoba.ca/afs/plant_science/courses/PLNT3140/14/114.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

Microsatellite 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

5' - 3'  

- Replication slippage
- Realignment
- Misalignment
- Extension

5' - 3'  

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

-1 REPEAT MUTATION
8-repeat allele after subsequent DNA replication
Minisatellites mutate by recombination

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

- 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
DNA Fingerprinting

- Heterozygosity = probability that two alleles are different
- Minisatellites are highly heterozygous, typically > 90%
- Basis of “DNA fingerprinting” (Sir Alec Jeffreys)
- Now routinely done using PCR and microsatellites

https://www.genome.gov/genetics-glossary/DNA-Fingerprinting
• Alpha satellite is the most common, 170bp unit, present in most 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 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.
Long-read sequencing allows identification of large, complex variants in the structure of DNA

Haplotype-resolved diverse human genomes and integrated analysis of structural variation

Peter Ebert1, Peter A. Audano2, Qihui Zhu3, Bernardo Rodriguez-Martin4, David Porubsky5, Marc Jan Bonder6, Arvis Sulovari7, Jana Eblier1, Weichen Zhou1, Rebecca Serra Martí1, Feyza Yilmaz7, Xuefang Zhao5,8, Ping Hsieh9, Joyce Lee10, Sushant Kumar10, Jiadong Lin11, Tobias Rausch1, Yu Chen12, Jingwen Ren13, Martin Santamarina14,15, Wolfram Höps1, Hufsa Ashraf16, Nelson T. Chuang16, Xiaofei Yang17, Katherine M. Munson18, Alexandra P. Lewis18, Susan Fairley19, Luke J. Tallon20, Wayne E. Clarke19, Anna O. Basile19, Marta Byrskaa-Bishop19, André Corvelo19, Uday S. Evani20, Tsung-Yu Lu20, Mark J.P. Chaisson19, Junjie Chen20, Chong Li20, Harrison Brand20,21, Aaron M. Wenger21, Maryam Gherehgi21,22,19, William T. Harvey1, Benjamin Kaeder1, Patrick Hasenfeld2, Allison A. Regler23, Haley J. Abel21, Ira M. Hall15, Paul Flicek24, Oliver Stegle25, Mark B. Gerstein10, Jose M.C. Tubio14,15, Zepeng Mu23, Yang I. Li27, Xinghua Shi20, Alex R. Hastie28, Kai Ye21,28, Zechen Chong2, Ashley D. Sanders2, Michael C. Zody19, Michael E. Talkowski18,22, Ryan E. Mills26,27, Scott E. Devine29, Charles Lee30,31,*, Jan O. Korbel30,31,*, Tobias Marschall14,2, Evan E. Eichler2,31,†

Cite as: P. Ebert et al., *Science* 10.1126/science.abf7117 (2021).

- 64 assembled human genomes that represent 26 different human populations.
- Found 107,590 SVs, of which 68% are not discovered by short-read sequencing, and 278 SV hotspots
- Characterized 130 of the most active mobile element source element and found 63% of all SVs arise by homology-mediated mechanisms
- Allows reliable graph-based genotyping from short reads of up to 50,340 SVs, resulting in the identification of 1,526 expression quantitative trait loci