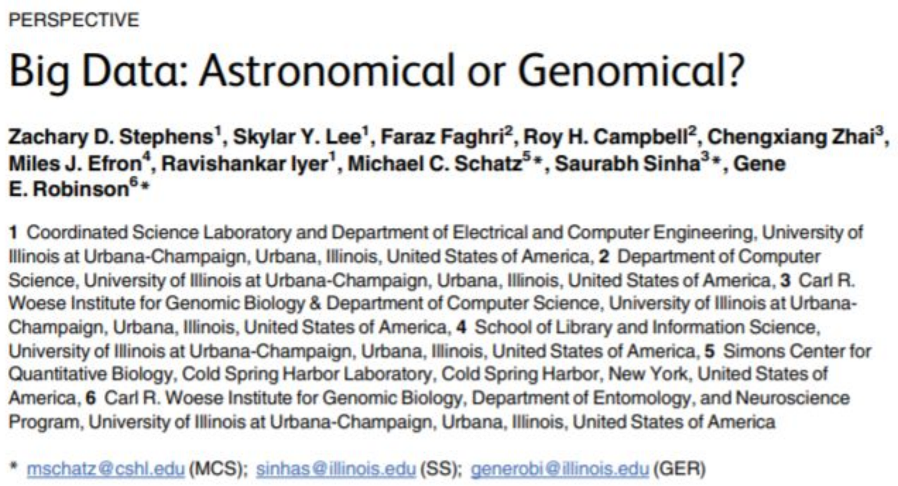
GE461: applications in genomics

Can Alkan
EA509
calkan@cs.bilkent.edu.tr

Genomics: the “new” Big Data



PERSPECTIVE

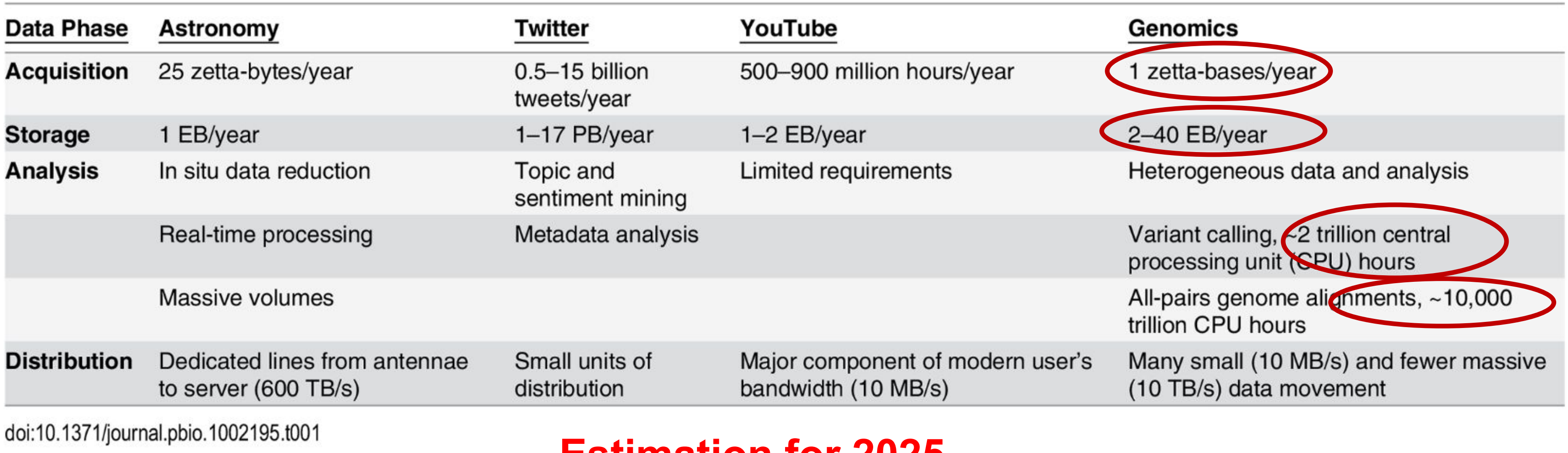
Big Data: Astronomical or Genomical?

Zachary D. Stephens[1], Skylar Y. Lee[1], Faraz Faghri[2], Roy H. Campbell[2], Chengxiang Zhai[3], Miles J. Efron[4], Ravishankar Iyer[1], Michael C. Schatz[5]*, Saurabh Sinha[3]*, Gene E. Robinson[6]*

1 Coordinated Science Laboratory and Department of Electrical and Computer Engineering, University of Illinois at Urbana-Champaign, Urbana, Illinois, United States of America, 2 Department of Computer Science, University of Illinois at Urbana-Champaign, Urbana, Illinois, United States of America, 3 Carl R. Woese Institute for Genomic Biology & Department of Computer Science, University of Illinois at Urbana-Champaign, Urbana, Illinois, United States of America, 4 School of Library and Information Science, University of Illinois at Urbana-Champaign, Urbana, Illinois, United States of America, 5 Simons Center for Quantitative Biology, Cold Spring Harbor Laboratory, Cold Spring Harbor, New York, United States of America, 6 Carl R. Woese Institute for Genomic Biology, Department of Entomology, and Neuroscience Program, University of Illinois at Urbana-Champaign, Urbana, Illinois, United States of America

* mschatz@cshl.edu (MCS); sinhas@illinois.edu (SS); generobi@illinois.edu (GER)

Stephens et al. PLoS Biology, 2015



Data Phase	Astronomy	Twitter	YouTube	Genomics
Acquisition	25 zetta-bytes/year	0.5–15 billion tweets/year	500–900 million hours/year	1 zetta-bases/year
Storage	1 EB/year	1–17 PB/year	1–2 EB/year	2–40 EB/year
Analysis	In situ data reduction	Topic and sentiment mining	Limited requirements	Heterogeneous data and analysis
	Real-time processing	Metadata analysis		Variant calling, ~2 trillion central processing unit (CPU) hours
	Massive volumes			All-pairs genome alignments, ~10,000 trillion CPU hours
Distribution	Dedicated lines from antennae to server (600 TB/s)	Small units of distribution	Major component of modern user’s bandwidth (10 MB/s)	Many small (10 MB/s) and fewer massive (10 TB/s) data movement

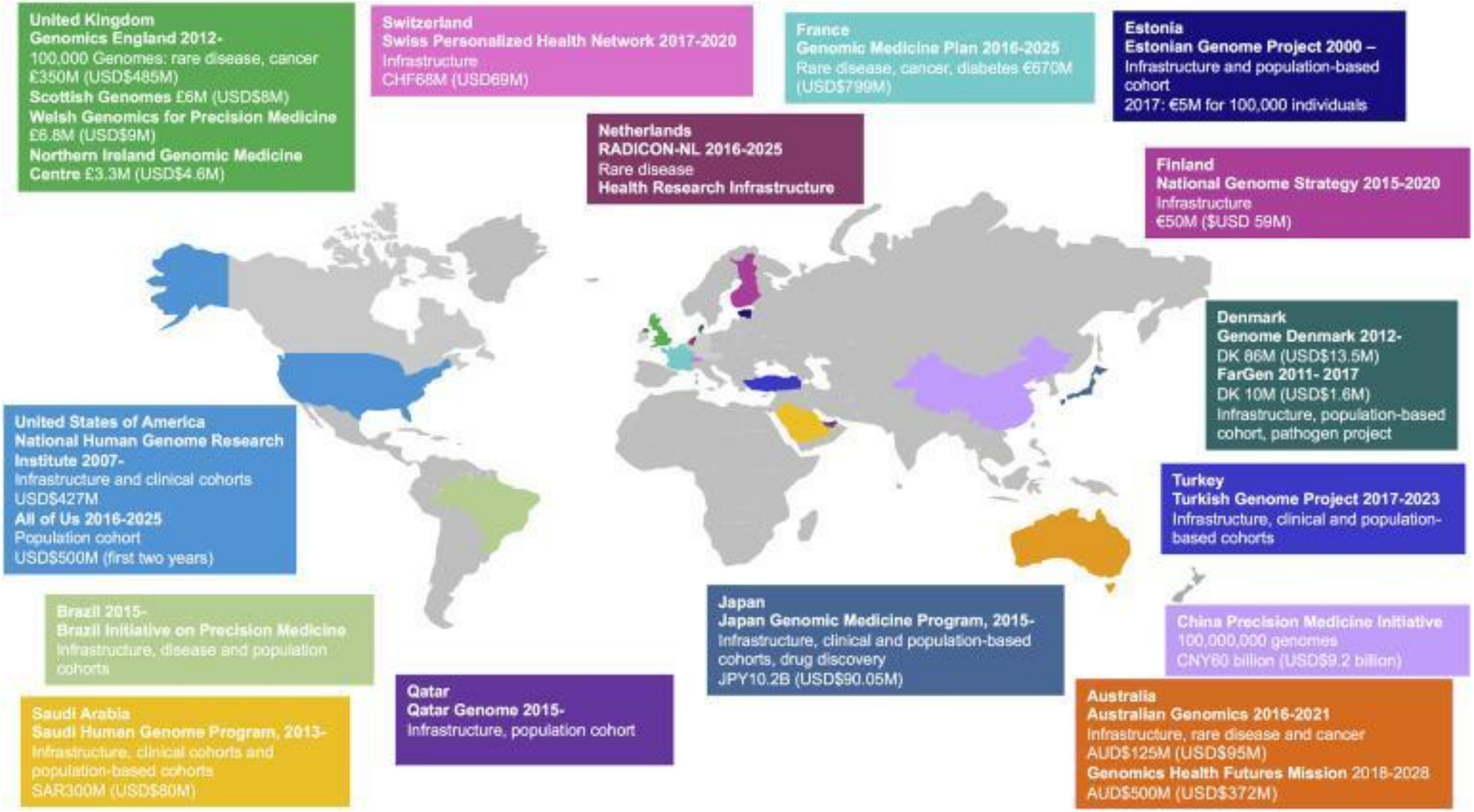
doi:10.1371/journal.pbio.1002195.t001

Estimation for 2025

Data size & processing

- Human reference genome: 3 GB
- One sequenced human sample (average):
  - 150 GB raw (compressed)
  - 150 GB “aligned” (analysis-ready)
  - ~20 CPU days
- One human (current) clinical sequencing data
  - 30-40 GB aligned
  - ~1 CPU week

Genomics and healthcare



United Kingdom
Genomics England 2012-
100,000 Genomes: rare disease, cancer
£350M (USD$485M)
Scottish Genomes £6M (USD$8M)
Welsh Genomics for Precision Medicine
£6.8M (USD$9M)
Northern Ireland Genomic Medicine Centre £3.3M (USD$4.6M)

Switzerland
Swiss Personalized Health Network 2017-2020
Infrastructure
CHF68M (USD69M)

France
Genomic Medicine Plan 2016-2025
Rare disease, cancer, diabetes €670M
(USD$799M)

Estonia
Estonian Genome Project 2000 –
Infrastructure and population-based cohort
2017: €5M for 100,000 individuals

Netherlands
RADICON-NL 2016-2025
Rare disease
Health Research Infrastructure

Finland
National Genome Strategy 2015-2020
Infrastructure
€50M ($USD 59M)

Denmark
Genome Denmark 2012-
DK 86M (USD$13.5M)
FarGen 2011- 2017
DK 10M (USD$1.6M)
Infrastructure, population-based cohort, pathogen project

United States of America
National Human Genome Research Institute 2007-
Infrastructure and clinical cohorts
USD$427M
All of Us 2016-2025
Population cohort
USD$500M (first two years)

Turkey
Turkish Genome Project 2017-2023
Infrastructure, clinical and population-based cohorts

Brazil 2015-
Brazil Initiative on Precision Medicine
Infrastructure, disease and population cohorts

Japan
Japan Genomic Medicine Program, 2015-
Infrastructure, clinical and population-based cohorts, drug discovery
JPY10.2B (USD$90.05M)

China Precision Medicine Initiative
100,000,000 genomes
CNY60 billion (USD$9.2 billion)

Saudi Arabia
Saudi Human Genome Program, 2013-
Infrastructure, clinical cohorts and population-based cohorts
SAR300M (USD$80M)

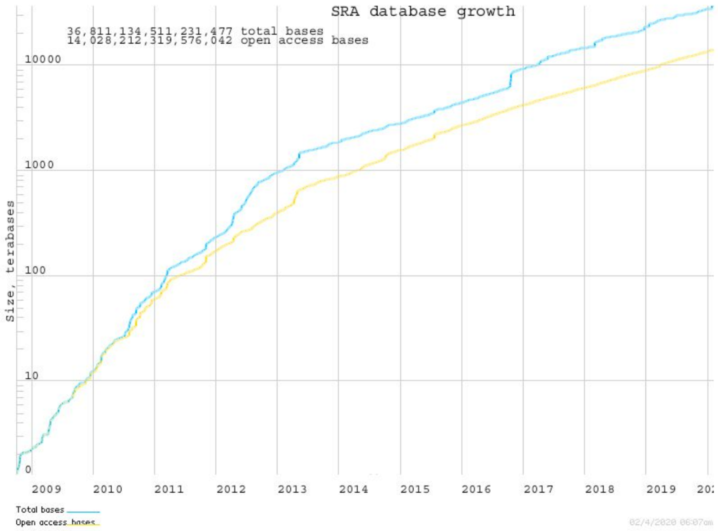
Qatar
Qatar Genome 2015-
Infrastructure, population cohort

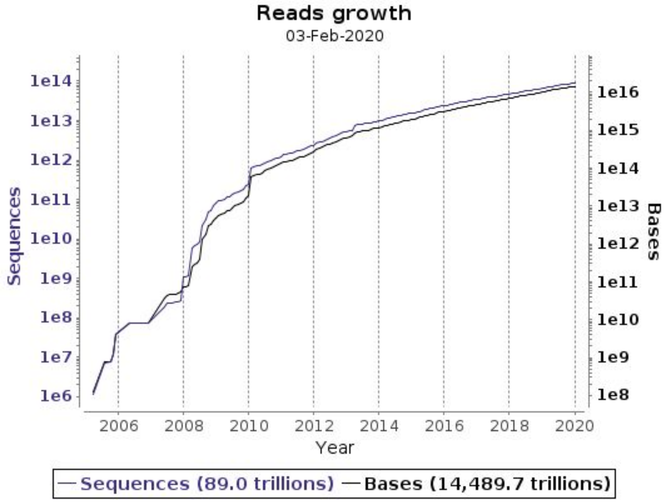
Australia
Australian Genomics 2016-2021
Infrastructure, rare disease and cancer
AUD$125M (USD$95M)
Genomics Health Futures Mission 2018-2028
AUD$500M (USD$372M)

Stark et al., AJHG 2019

Publicly available data

- Two “main” sources for genomics/transcriptomics:
  - NCBI Sequence Read Archive (SRA)
    - > 14 PB public / free
    - > 36 PB total (~22 PB controlled access)
  - EBI Nucleotide Archive (ENA)
    - > 10 PB

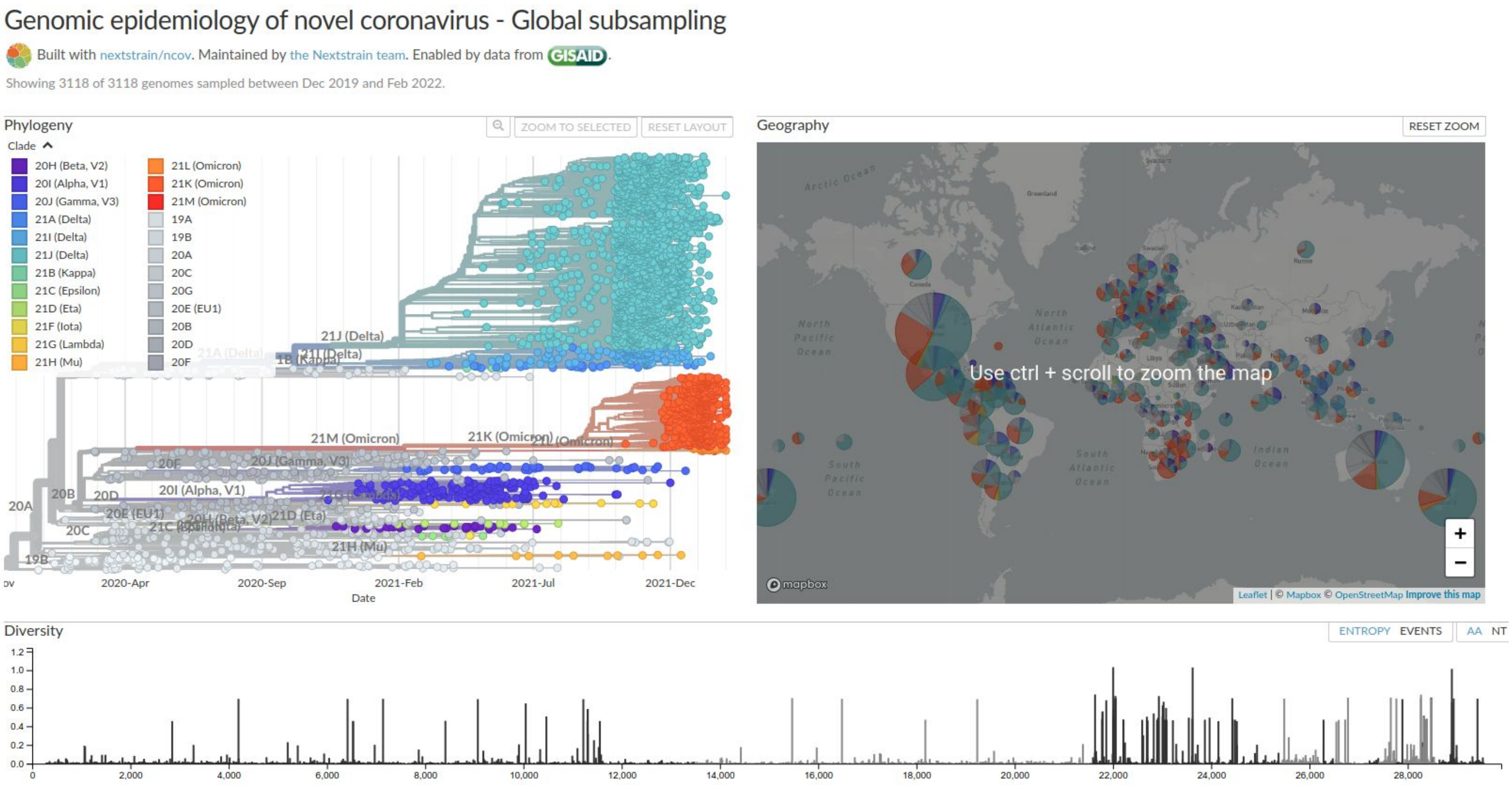




Genomics: many use cases

- Catalog “normal” human genome variation
  - Population genetics / analysis of migration
  - Filtering data set for disease studies
- Genetic diseases
  - Find genetic causes of diseases
  - Guide diagnostics
  - Guide treatment
  - Identify cancer type / subtype
- Infection / outbreaks
  - Bacterial infections:
    - Guide antibiotics treatment
    - Sepsis: find cause & treat
  - Viral disease outbreaks:
    - Guide vaccine development
    - SARS-CoV-2!
      - Tracking new mutations
- Pharmacogenomics
  - Drug efficacy
    - Warfarin (blood thinner)
      - *VKORC1* and *CYP2C9* gene mutations -> increased sensitivity

SARS-CoV-2 analysis in action



Genomic epidemiology of novel coronavirus - Global subsampling

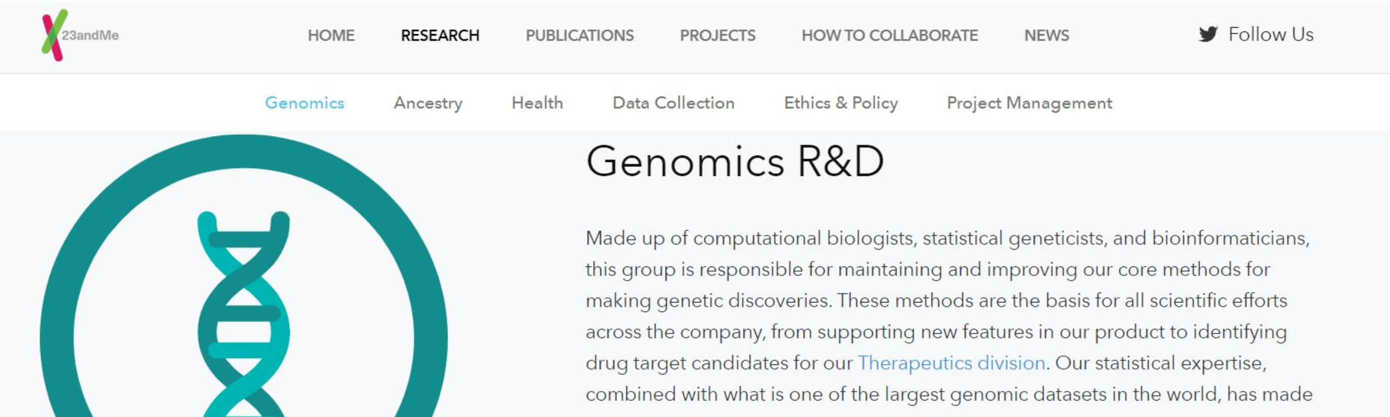
Built with nextstrain/ncov. Maintained by the Nextstrain team. Enabled by data from GISAID.

Showing 3118 of 3118 genomes sampled between Dec 2019 and Feb 2022.

Use ctrl + scroll to zoom the map

https://nextstrain.org/ncov/global

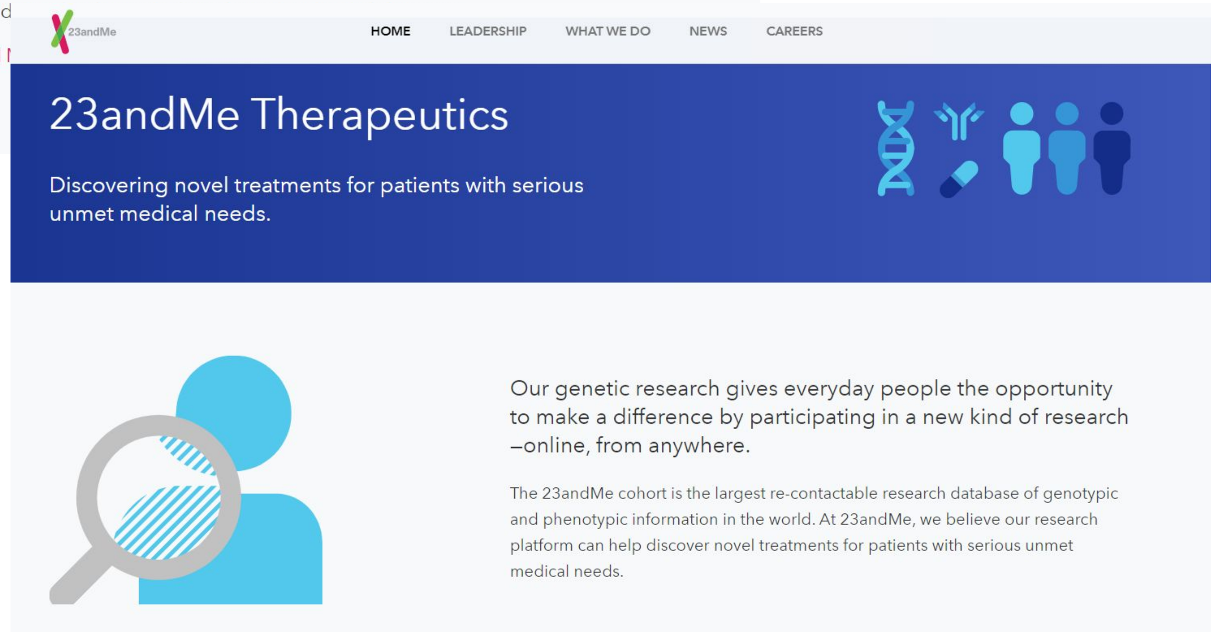
Data science in genomics industry



Genomics R&D

Made up of computational biologists, statistical geneticists, and bioinformaticians, this group is responsible for maintaining and improving our core methods for making genetic discoveries. These methods are the basis for all scientific efforts across the company, from supporting new features in our product to identifying drug target candidates for our Therapeutics division. Our statistical expertise, combined with what is one of the largest genomic datasets in the world, has made 23and

Read I



23andMe Therapeutics

Discovering novel treatments for patients with serious unmet medical needs.

Our genetic research gives everyday people the opportunity to make a difference by participating in a new kind of research —online, from anywhere.

The 23andMe cohort is the largest re-contactable research database of genotypic and phenotypic information in the world. At 23andMe, we believe our research platform can help discover novel treatments for patients with serious unmet medical needs.

Data science in genomics industry

- 23andMe: ancestry, therapeutics
- Insitro: drug discovery, pharmacogenomics
- Regeneron: drug discovery, precision medicine
- Pretty much all major drug companies
  - Merck, Novartis, Bristol Myers-Squibb, Pfizer,…
- Many more

Bioinformatics: methods for -omics

- Bioinformatics: Development of methods based on computer science for problems in biology &medicine
  - Sequence analysis (combinatorial and statistical/probabilistic methods)
  - Graph theory
  - Data mining
  - Database
  - Statistics
  - Image processing
  - Visualization
  - …..

CS 481 and CS 681

All life depends on 3 critical molecules

- DNAs **Genomics**
  - Hold information on how cell works
    - □ RNA for retroviruses
- RNAs **Transcriptomics**
  - Act to transfer short pieces of information to different parts of cell
  - Provide templates to synthesize into protein
- Proteins **Proteomics**
  - Form enzymes that send signals to other cells and regulate gene activity
  - Form body’s major components (e.g. hair, skin, etc.)
- For a computer scientist, these are all strings derived from three alphabets.

Alphabets

DNA:

$\Sigma = \{A, C, G, T\}$

A pairs with T; G pairs with C

RNA:

$\Sigma = \{A, C, G, U\}$

A pairs with U; G pairs with C

Protein:

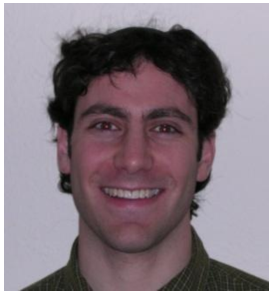
$\Sigma = \{A,C,D,E,F,G,H,I,K,L,M,N,P,Q,R,S,T,V,W,Y\}$ *and*

B = N | D
Z = Q | E
X = *any*

SAMPLE USE CASE

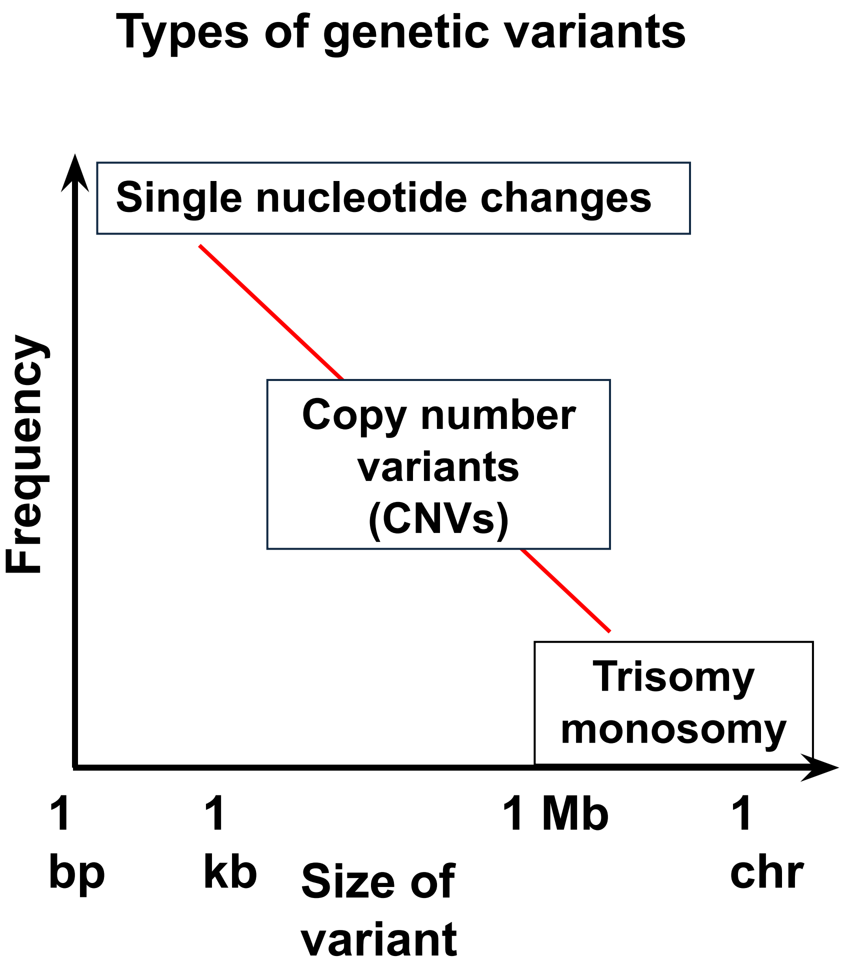
GENOMIC VARIATION: CHANGES IN DNA SEQUENCE

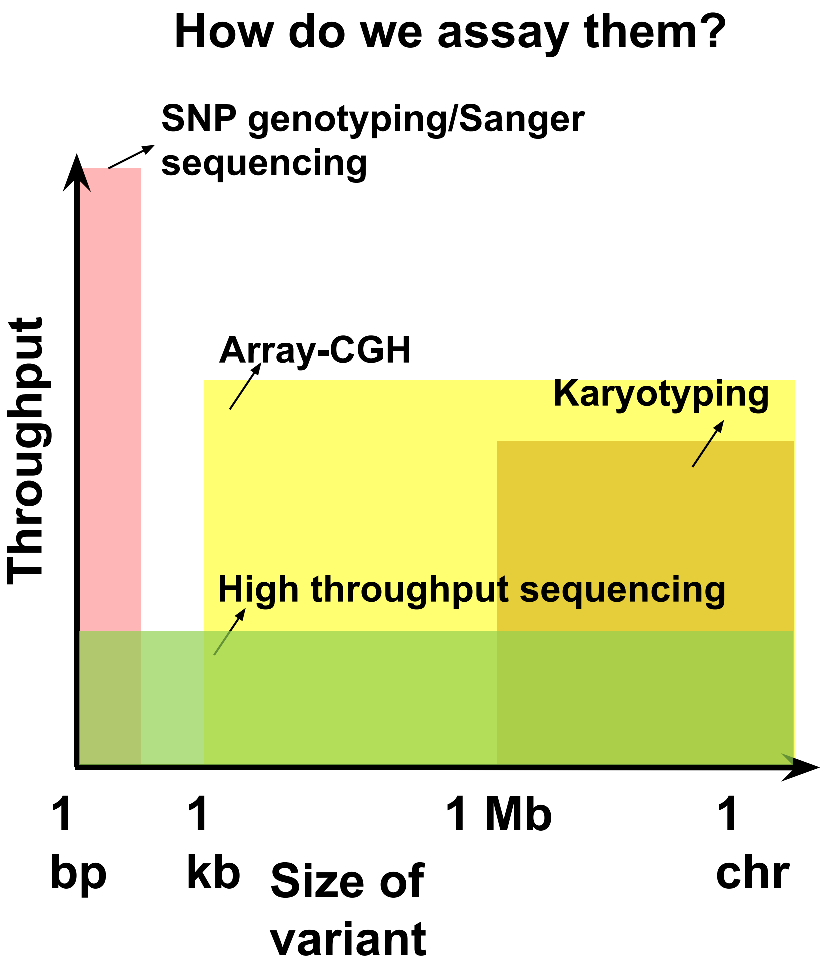
Human genome variation



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

Human genetic variation


Types of genetic variants
Frequency
Single nucleotide changes
Copy number variants (CNVs)
Trisomy monosomy
1 bp
1 kb
1 Mb
1 chr
Size of variant



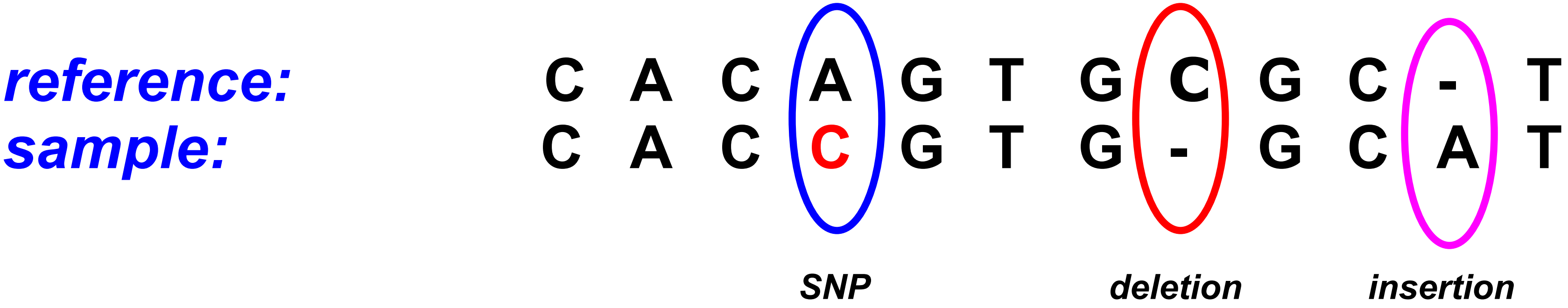
Size range of genetic variation

- 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%
  - Inversions
  - Translocations
- Chromosomal changes

CNVs

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

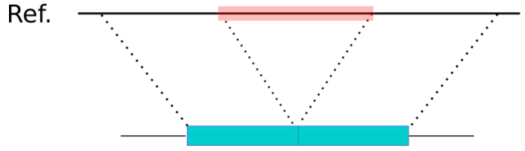
reference: CAG CAG CAG CAG

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

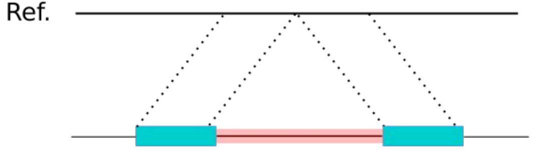
Structural Variation

DELETION

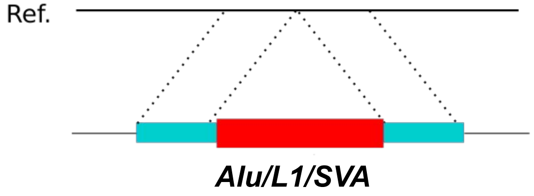


Autism, developmental delay, Crohn's

NOVEL SEQUENCE INSERTION

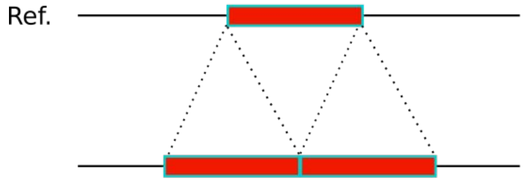


MOBILE ELEMENT INSERTION

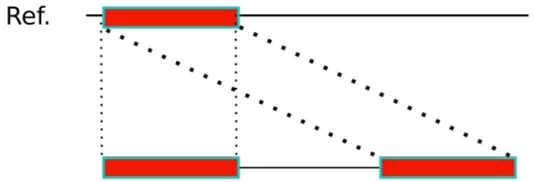


Haemophilia

TANDEM DUPLICATION

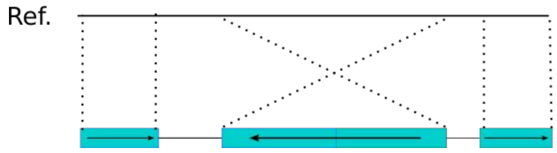


INTERSPERSED DUPLICATION

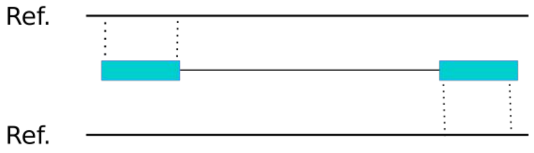


Schizophrenia, psoriasis

INVERSION



TRANSLOCATION

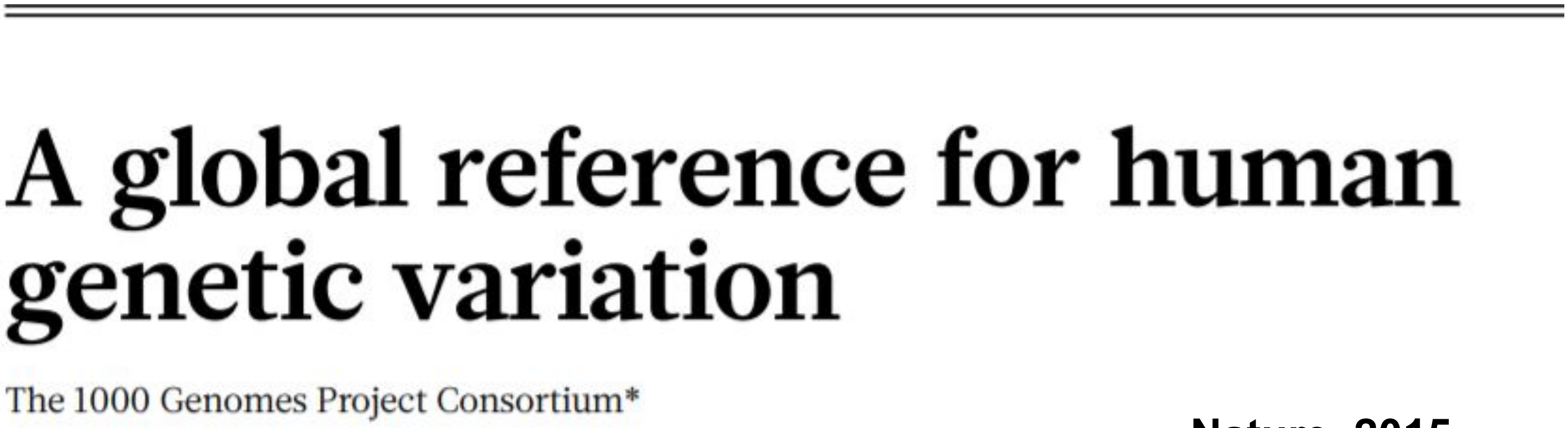


Chronic myelogenous leukemia

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



A global reference for human genetic variation

The 1000 Genomes Project Consortium*

Nature, 2015

Genetic variation among humans

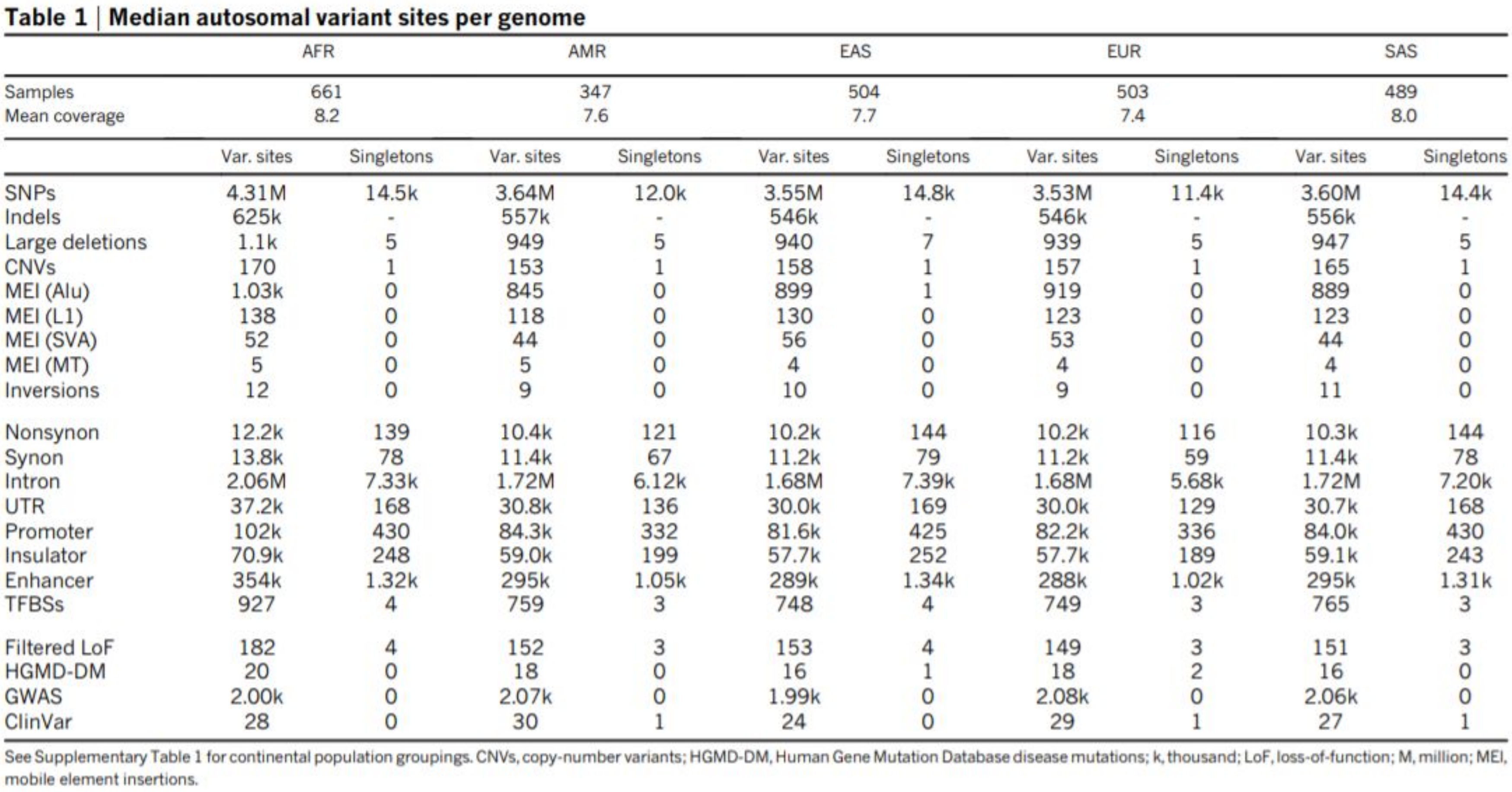
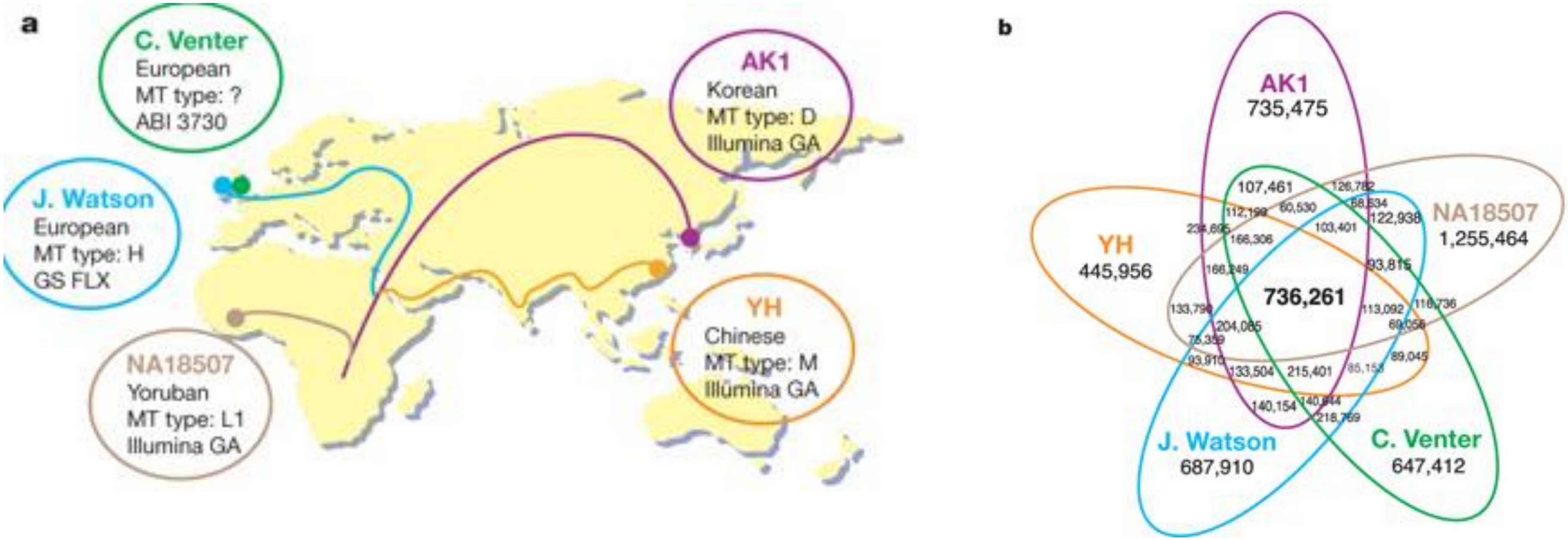


Table 1 | Median autosomal variant sites per genome

	AFR		AMR		EAS		EUR		SAS	
Samples	661		347		504		503		489	
Mean coverage	8.2		7.6		7.7		7.4		8.0	
	Var. sites	Singletons	Var. sites	Singletons	Var. sites	Singletons	Var. sites	Singletons	Var. sites	Singletons
SNPs	4.31M	14.5k	3.64M	12.0k	3.55M	14.8k	3.53M	11.4k	3.60M	14.4k
Indels	625k	-	557k	-	546k	-	546k	-	556k	-
Large deletions	1.1k	5	949	5	940	7	939	5	947	5
CNVs	170	1	153	1	158	1	157	1	165	1
MEI (Alu)	1.03k	0	845	0	899	1	919	0	889	0
MEI (L1)	138	0	118	0	130	0	123	0	123	0
MEI (SVA)	52	0	44	0	56	0	53	0	44	0
MEI (MT)	5	0	5	0	4	0	4	0	4	0
Inversions	12	0	9	0	10	0	9	0	11	0
Nonsynon	12.2k	139	10.4k	121	10.2k	144	10.2k	116	10.3k	144
Synon	13.8k	78	11.4k	67	11.2k	79	11.2k	59	11.4k	78
Intron	2.06M	7.33k	1.72M	6.12k	1.68M	7.39k	1.68M	5.68k	1.72M	7.20k
UTR	37.2k	168	30.8k	136	30.0k	169	30.0k	129	30.7k	168
Promoter	102k	430	84.3k	332	81.6k	425	82.2k	336	84.0k	430
Insulator	70.9k	248	59.0k	199	57.7k	252	57.7k	189	59.1k	243
Enhancer	354k	1.32k	295k	1.05k	289k	1.34k	288k	1.02k	295k	1.31k
TFBSs	927	4	759	3	748	4	749	3	765	3
Filtered LoF	182	4	152	3	153	4	149	3	151	3
HGMD-DM	20	0	18	0	16	1	18	2	16	0
GWAS	2.00k	0	2.07k	0	1.99k	0	2.08k	0	2.06k	0
ClinVar	28	0	30	1	24	0	29	1	27	1

See Supplementary Table 1 for continental population groupings. CNVs, copy-number variants; HGMD-DM, Human Gene Mutation Database disease mutations; k, thousand; LoF, loss-of-function; M, million; MEI, mobile element insertions.

Genetic variation are “shared”



Kim *et al.* Nature, 2009

PROJECTS FOR GENOMIC VARIATION DISCOVERY

International HapMap Project

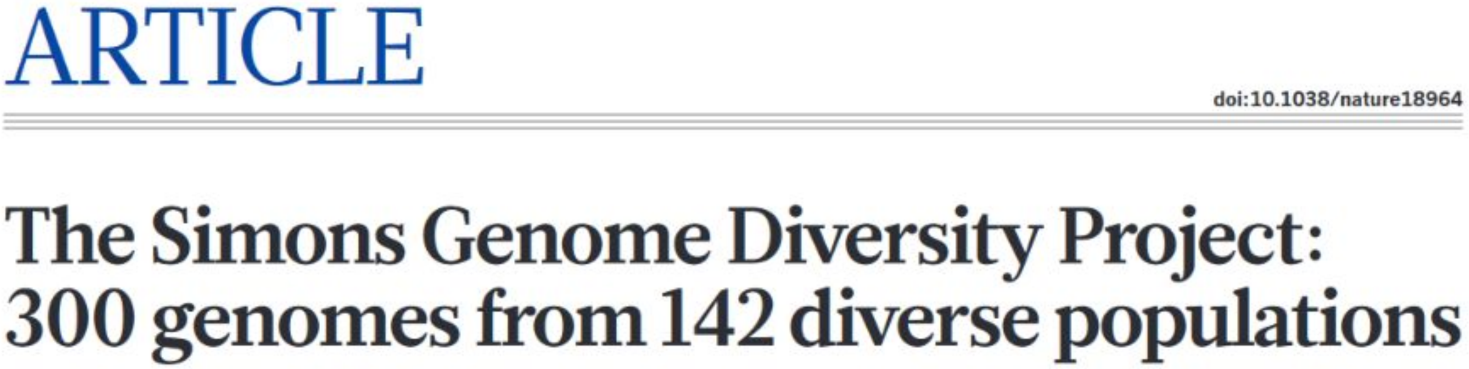
- Determine genotypes & haplotypes of 270 human individuals from 3 diverse populations:
  - Northern Americans (Utah / Mormons)
  - Africans (Yoruba from Nigeria)
  - Asians (Han Chinese and Japanese)
- 90 individuals from each population group, organized into parent-child **trios**.
- Each individual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

http://www.hapmap.org

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.



ARTICLE

doi:10.1038/nature18964

The Simons Genome Diversity Project: 300 genomes from 142 diverse populations

Mallick et al., 2016

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

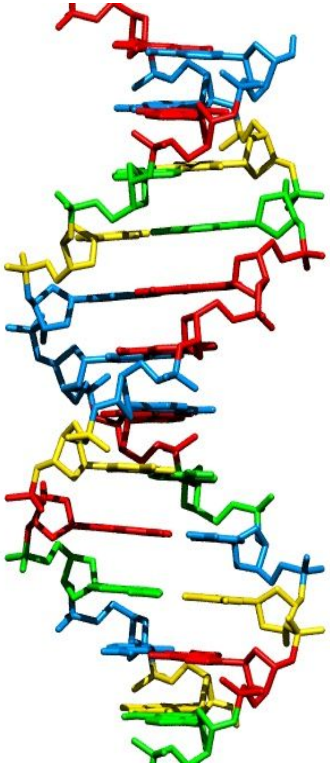
- 2007: “Sanger”-based capillary sequencing; one human genome (WGS): ~ $10 million (Levy et al., 2007)
- 2008: First “next-generation” sequencer 454 Life Sciences; genome of James Watson: ~$2 million (Wheeler et al., 2008)
- 2008: The Illumina platform; genome of an African (Bentley et al, 2008) and an Asian (Wang et al., 2008): ~$200K each
- 2009: The SOLiD platform: ~$200K
- Today with the Illumina platform: ~$1K/ genome
- Others: Oxford Nanopore, Pacific Biosciences SMRT

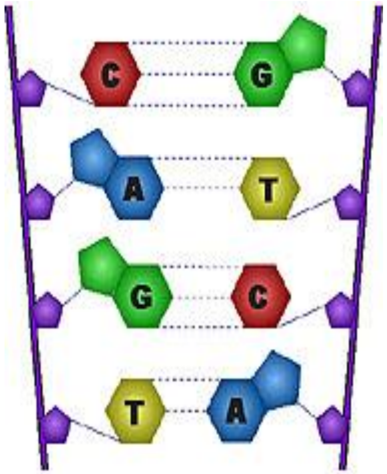
Sequencing-based projects

- The 1000 Genomes Project Consortium (www.1000genomes.org)
  - Large consortium: groups from USA, UK, China, Germany, Canada
  - 2.504 humans from 29 populations
- Independent
  - South African (Schuster et al., 2010), Korean, Japanese, UK (UK100K project), Ireland, Netherlands (GoNL project), France, US All of Us (> 1 million), UK Biobank (> 500K) …
- Cancer:
  - TCGA: >500 cases of 20 tumor types; 1.2 PB as of 2016
  - ICGC: > 20K samples (different types); 1.7 PB
- Ancient DNA: Neandertal (Green et al., 2010); Denisova (Reich et al., 2010); Çatalhöyük (METU)

DNA sequencing

How we obtain the sequence of nucleotides of a species



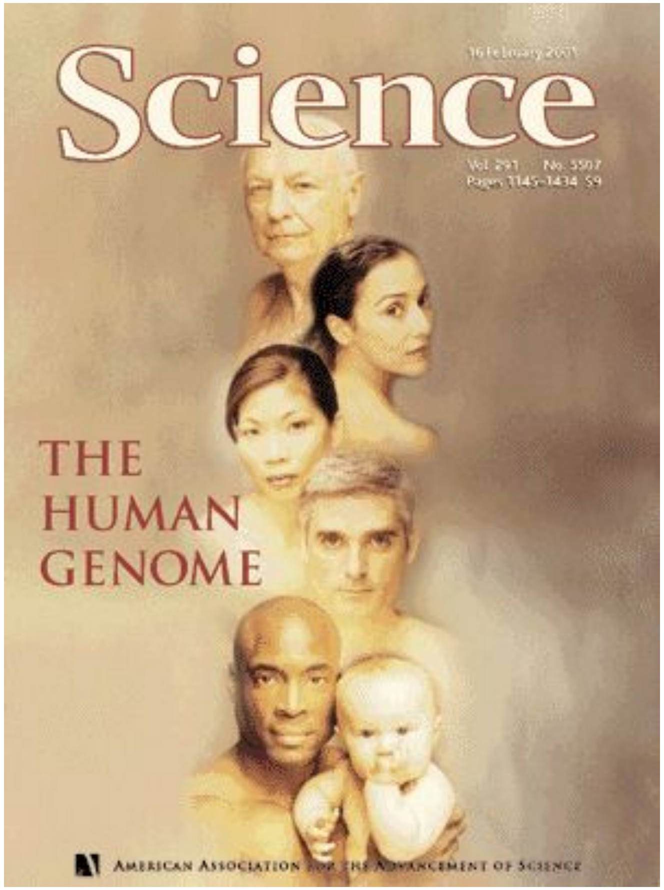


…ACGTGACTGAGGACCGTG
CGACTGAGACTGACTGGGT
CTAGCTAGACTACGTTTTA
TATATATATACGTCGTCGT
ACTGATGACTAGATTACAG
ACTGATTTAGATACCTGAC
TGATTTTAAAAAAATATT…

HIGH THROUGHPUT SEQUENCING

Human genome reference

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



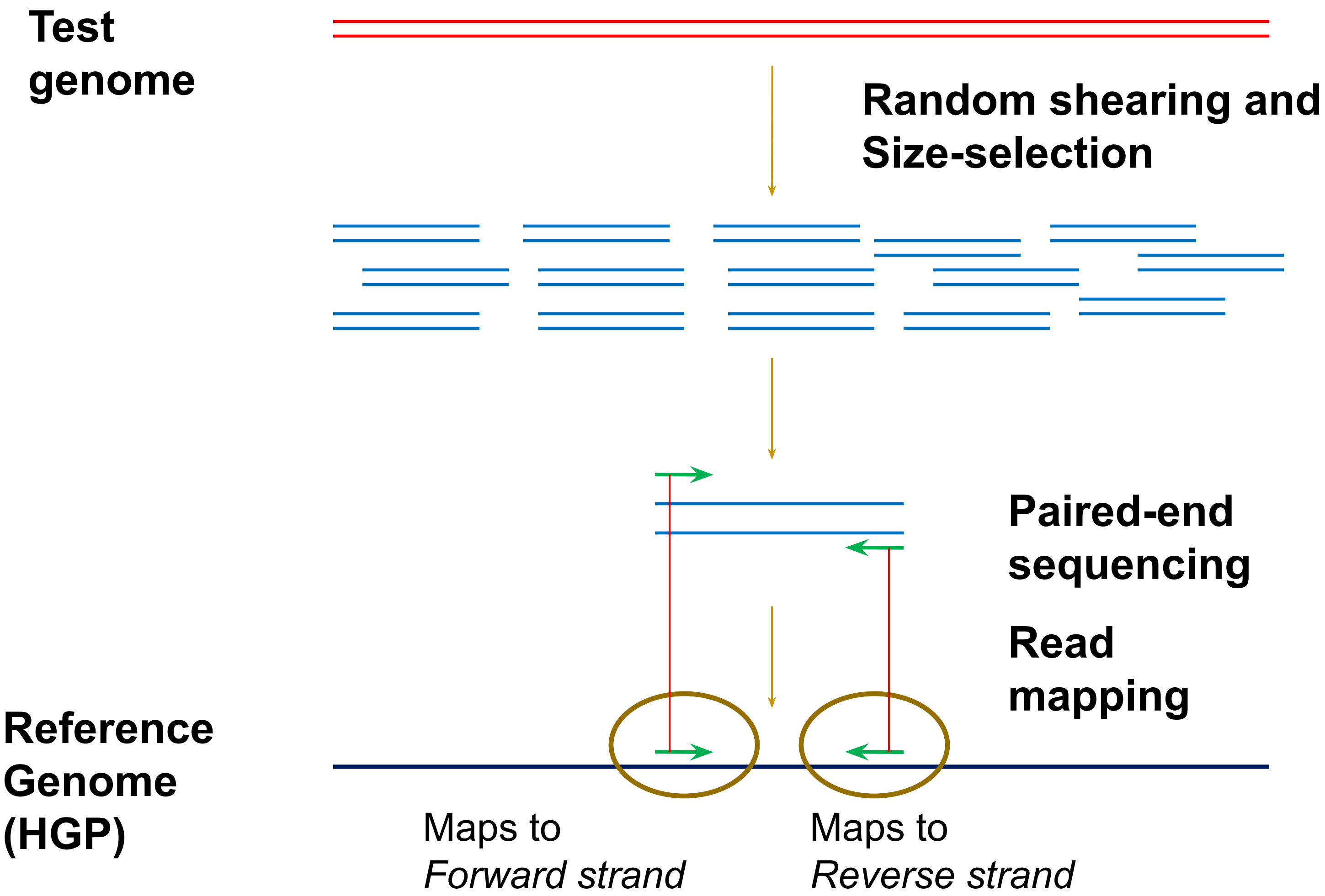
4 human samples, 14 years, 3-10 billion dollars

Current version: hg38

https://www.ncbi.nlm.nih.gov/grc

Chromosomes 1-22, X, Y, MT
Alternative haplotypes
HLA haplotypes

Whole Genome Shotgun sequencing



Test genome

Random shearing and Size-selection

Paired-end sequencing

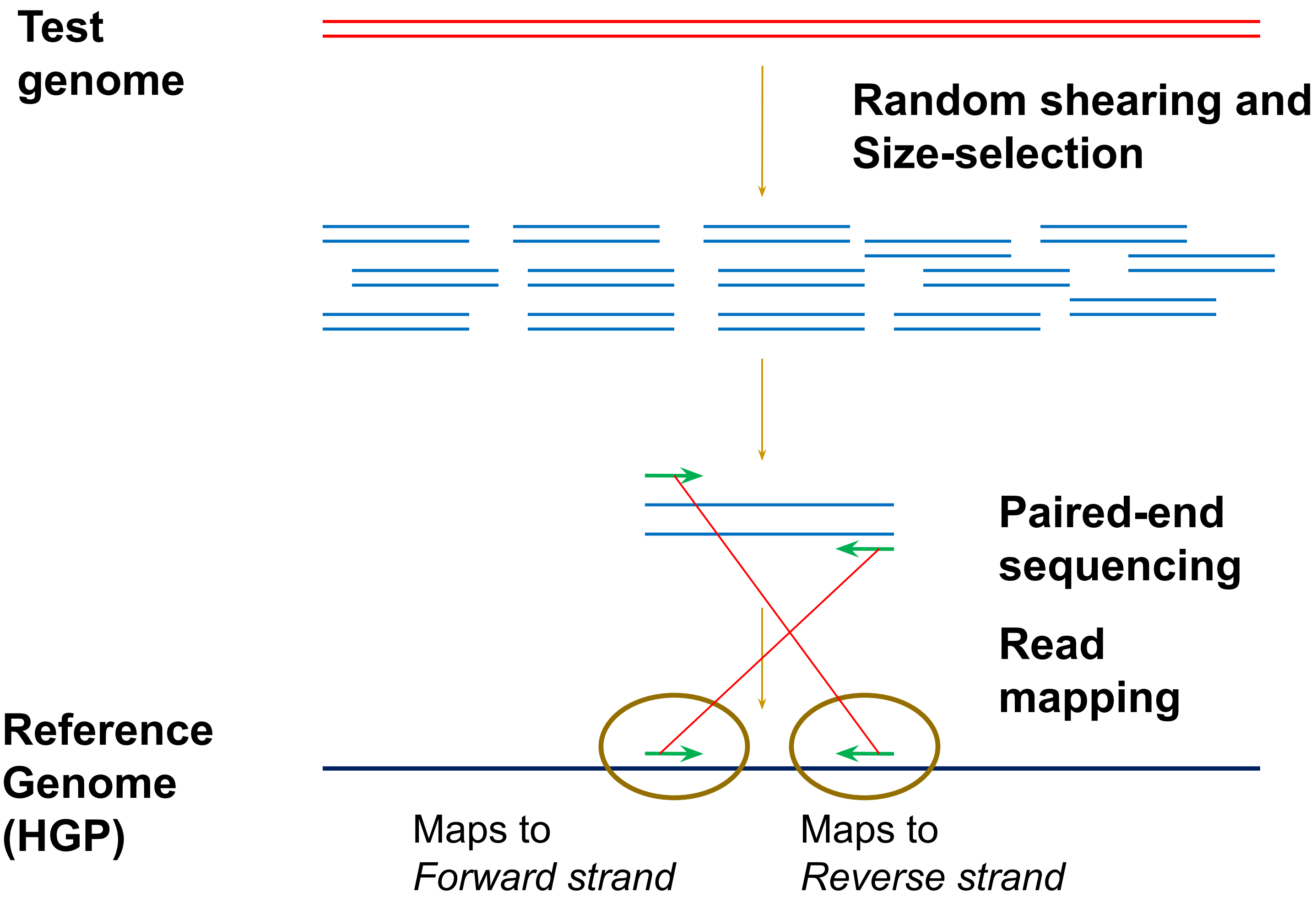
Read mapping

Reference Genome (HGP)

Maps to *Forward strand*

Maps to *Reverse strand*

Whole Genome Shotgun sequencing



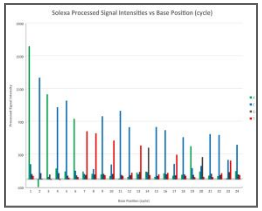
HTS Technologies

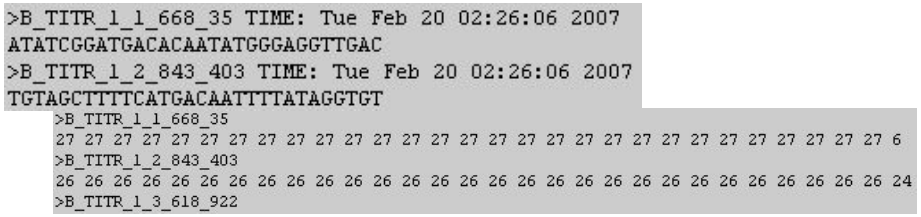
- Short read:
  - Illumina (Solexa): current market leader
    - *GAIIx, HiSeq2000, MiSeq, HiSeq2500, NovaSeq*
    - *Sequencing by synthesis*

- Long Read:
  - Pacific Biosciences Single Molecule Real Time
    - *RSII, Sequel*
  - Oxford Nanopore Technologies:
    - *MinION, Flongle, PromethION, GridION*

Fundamental informatics challenges

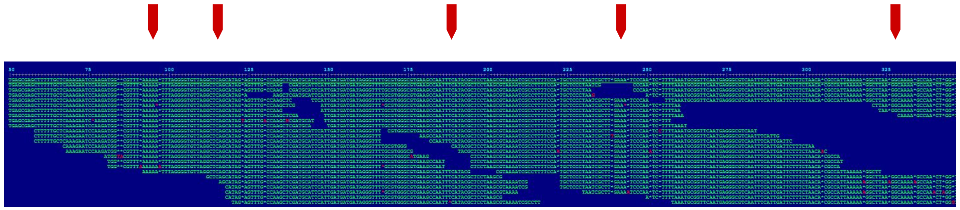
1. Interpreting machine readouts – base calling, base error estimation


Solexa Processed Signal Intensities vs Base Position (cycle)

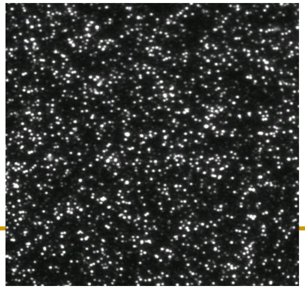


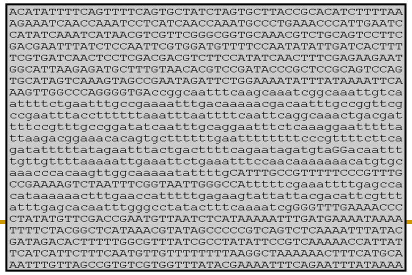
```
>B_TITR_1_1_668_35 TIME: Tue Feb 20 02:26:06 2007
ATATCGGATGACACAATATGGGAGGTTGAC
>B_TITR_1_2_843_403 TIME: Tue Feb 20 02:26:06 2007
TGTAGCTTTTCATGACAATTTTATAGGTGT
    >B_TITR_1_1_668_35
    27 27 27 27 27 27 27 27 27 27 27 27 27 27 27 27 27 27 27 27 27 27 27 27 27 27 27 27 27 6
    >B_TITR_1_2_843_403
    26 26 26 26 26 26 26 26 26 26 26 26 26 26 26 26 26 26 26 26 26 26 26 26 26 26 26 26 26 24
    >B_TITR_1_3_618_922
```

2. Data visualization



3. Data storage & management
Gzip compressed raw data for one human genome > 100 GB (Illumina)

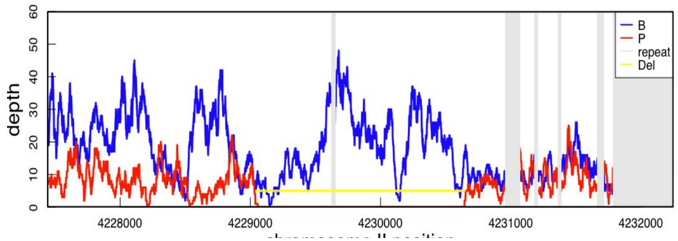


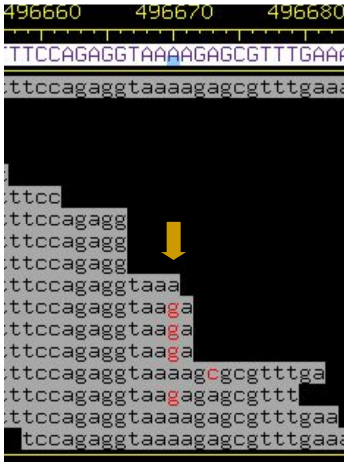


```
ACATATTTTCAGTTTTCAGTGCTATCTAGTGCTTACCGCACATCTTTTAA
AGAAATCAACCAAATCCTCATCAACCAAATGCCCTGAAACCCATTGAATC
CATATCAAATCATAACGTCGTTCGGGCGGTGCAAACGTCTGCAGTCCTTC
GACGAATTTATCTCCAATTCGTGGATGTTTTCCAATATATTGATCACTTT
TCGTGATCAACTCCTCGACGACGTCTTCCATATCAACTTTCGAGAAGAAT
GGCATTAAGAGATGCTTTGTAACACGTCCGATACCCGCTCCGCAGTCCAG
TGCATAGTCAAAGTAGCCGAATAGATTCTGGAAAATATTTATAAAATTCA
AAGTTGGCCCAGGGGTGAccggcaatttcaagcaaatcggcaaattgtca
attttctgaatttgccgaaaatttgacaaaaacgacaatttgccggttcg
ccgaatttacctttttttaaatttaattttcaattcaggcaaactgacgat
tttccgtttgccggatatcaatttgcaggaatttctcaaaggaattttta
ttaagacggaaacacagtgctttttttgaatttttttttcccgttttcttca
gatattttttatagaatttactgacttttcagaatagatgtaGGacaattt
tgttgttttaaaaattgaaattctgaaatttccaacaaaaaaacatgtgc
aaacccacaagttggcaaaaatattttgCATTTGCCGTTTTTCCCGTTTG
CCGAAAAGTCTAATTTCGGTAATTGGGCCAtttttcgaaattttgagcca
cataaaaaactttgaaccatttttgagaagtattattacgacattcgttt
atttgagcacaatttgggcctatactttcaaaatcgGGGTTTGAAAACCC
CTATATGTTCGACCGAATGTTAATCTCATAAAAATTTGATGAAAATAAAA
TTTTCTACGGCTCATAAACGTATAGCCCCCGTCAGTCTCAAAATTTATAC
GATAGACACTTTTTGGCGTTTATCGCCTATATTCCGTCAAAAACCATTAT
TCATCATTCTTTCAATGTTGTTTTTTTTAAGGCTAAAAAACTTTCATGCA
AATTTGTTAGCCGTGTCGTGGTTTATACGAAAATTTCAGAATTTATAAAA
```

Informatics challenges (cont'd)

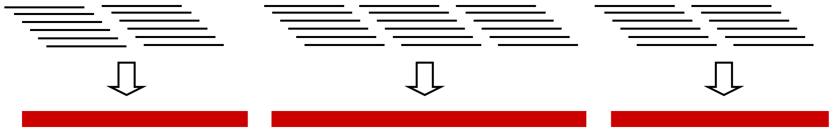
4. SNP, indel, and structural variation discovery





```
496660    496670    496680
TTCCAGAGGTAAAAGAGCGTTTGAAA
ttccagaggtaaaagagcgtttgaa
ttcc
ttccagagg
ttccagagg
ttccagagg
ttccagaggtaaa
ttccagaggtaaga
ttccagaggtaaga
ttccagaggtaaga
ttccagaggtaaaagcgcgtttga
ttccagaggtaagagagcgttt
ttccagaggtaaaagagcgtttgaa
 tccagaggtaaaagagcgtttgaa
```

5. ***De novo*** **Assembly**

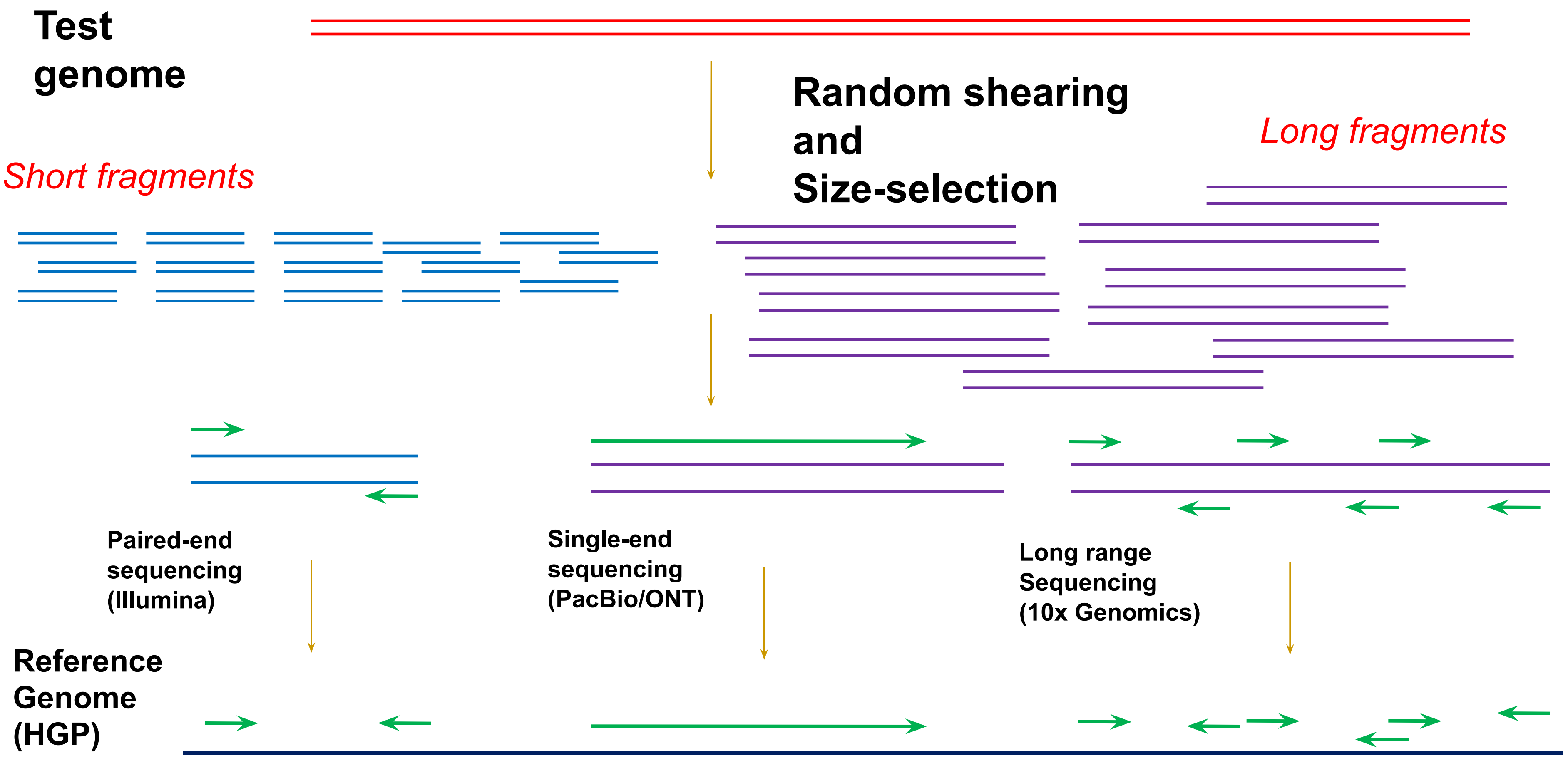


CURRENT PLATFORMS

Features of HTS data

- Short sequence reads
  - 150 - 300 bp Illumina
- Long, but error prone sequence reads
  - Average ~50 Kb PacBio - 12% error
  - Up to 1 Mb ONT – 20% error
- Huge amount of sequence per run
  - Up to terabases per run (3 Tbp for Illumina/NovaSeq 6000)
- Huge number of reads per run
  - Up to billions
- Higher error (compared with Sanger)
  - Illumina: mostly substitutions
  - PacBio / ONT: mostly indels

Whole Genome Sequencing


Test genome
Random shearing and Size-selection
Long fragments
Short fragments
Paired-end sequencing (Illumina)
Single-end sequencing (PacBio/ONT)
Long range Sequencing (10x Genomics)
Reference Genome (HGP)

Sequencing technologies

Short-Read

Illumina

- 100-200bp
- Paired-end
- Billions of reads
- < 0.1% error

Long Read

PacBio and Oxford Nanopore

- > 10 Kb, up to 1 Mb
- Single-end
- Hundreds of millions of reads
- 5-12% error – indel dominated
  - HiFi: 1% error

Long Range

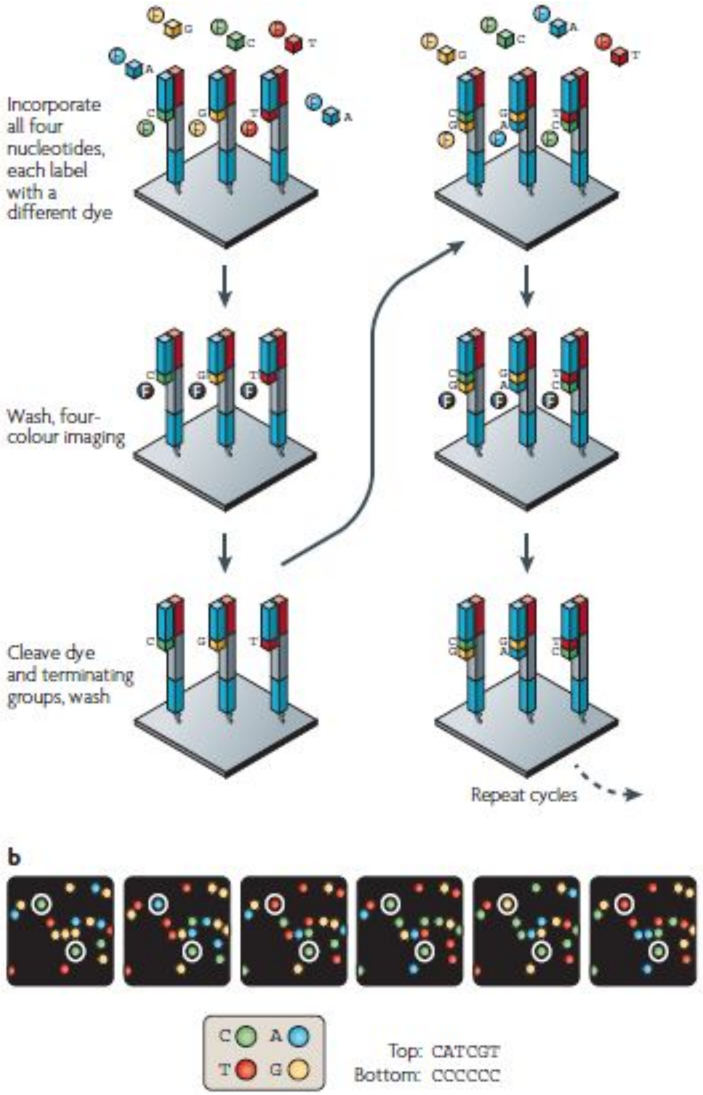
10X + Illumina

- 100-200bp
- Paired-end
- Billions of reads
- < 0.1% error
- Barcoded: 30-50 Kb molecule range

Illumina

- Current market leader
- Based on *sequencing by synthesis*
- Current read length 150-300bp
- Paired-end sequencing
- Error ~0.1%
  - Substitution errors dominate
- Throughput: Up to 3 Tbp in one run (2 days)
- Cheapest sequencing technology
  - Cost: ~ $1,000 per human

Illumina


Incorporate all four nucleotides, each label with a different dye
Wash, four-colour imaging
Cleave dye and terminating groups, wash
Repeat cycles
b
C
A
T
G
Top: CATCGT
Bottom: CCCCCC

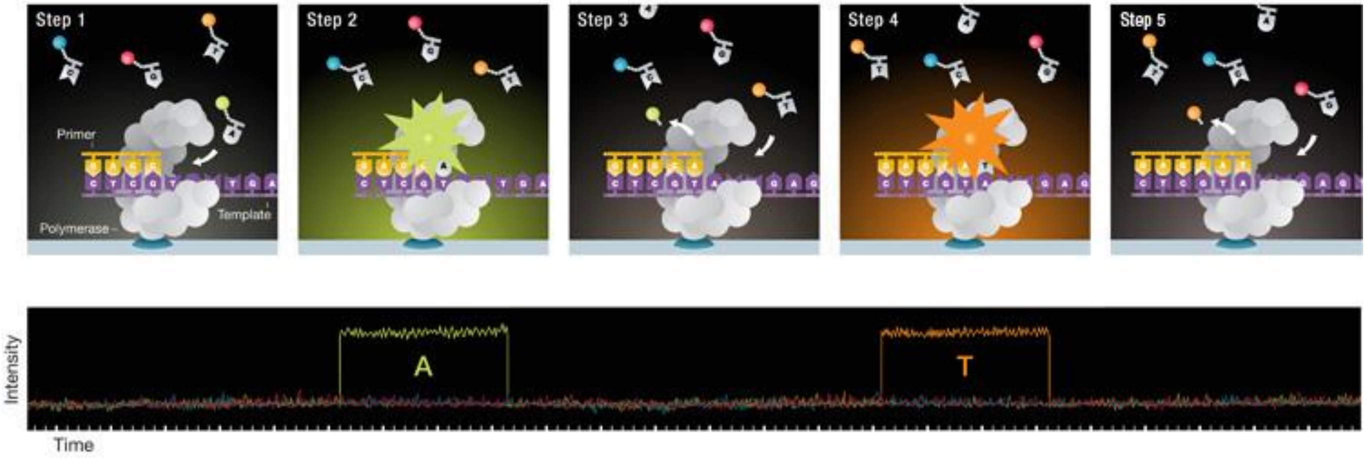
NovaSeq

MiSeq

HiSeq 2000/2500

Pacific Biosciences

- “Third generation”; single molecule real time sequencing (SMRT)
- Phosphates are labeled. Watches DNA polymerase in real-time while it copies single DNA molecules.
- Premise: long sequence reads in short time (median 60 Kbp)
- Errors: ~12%; indel dominated
  - HiFi: shorter reads with 1% error
- ~$ 3,000 / human



Pacific Biosciences

- For any DNA polymerase you can read a total of ~60 kb (median) sequence
- Two sequencing protocols:
  - CLR: single read
  - HiFi: Make a circle, re-read the same molecule 5-6 times
    - Multiple sequence alignment to correct errors
    - Median length = 60000 / 6= 10 Kbp
    - > 99% accuracy

Nanopore sequencing

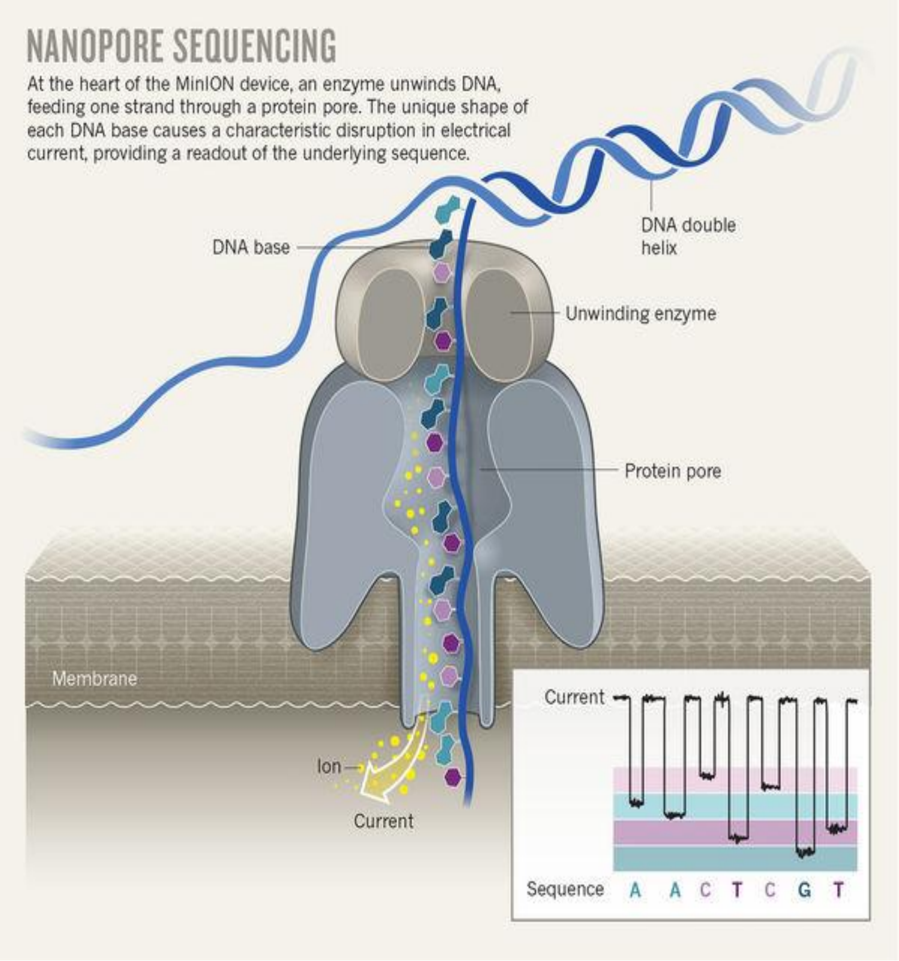
- Up to 2 Mbp reads
  - 5-20% error, indel dominated
- Real-time analysis supported
- RNN-based basecallers

Nanopore sequencing technology and tools for genome assembly: computational analysis of the current state, bottlenecks and future directions FREE

Damla Senol Cali ✉, Jeremie S Kim, Saugata Ghose, Can Alkan, Onur Mutlu

Briefings in Bioinformatics, bby017, https://doi.org/10.1093/bib/bby017

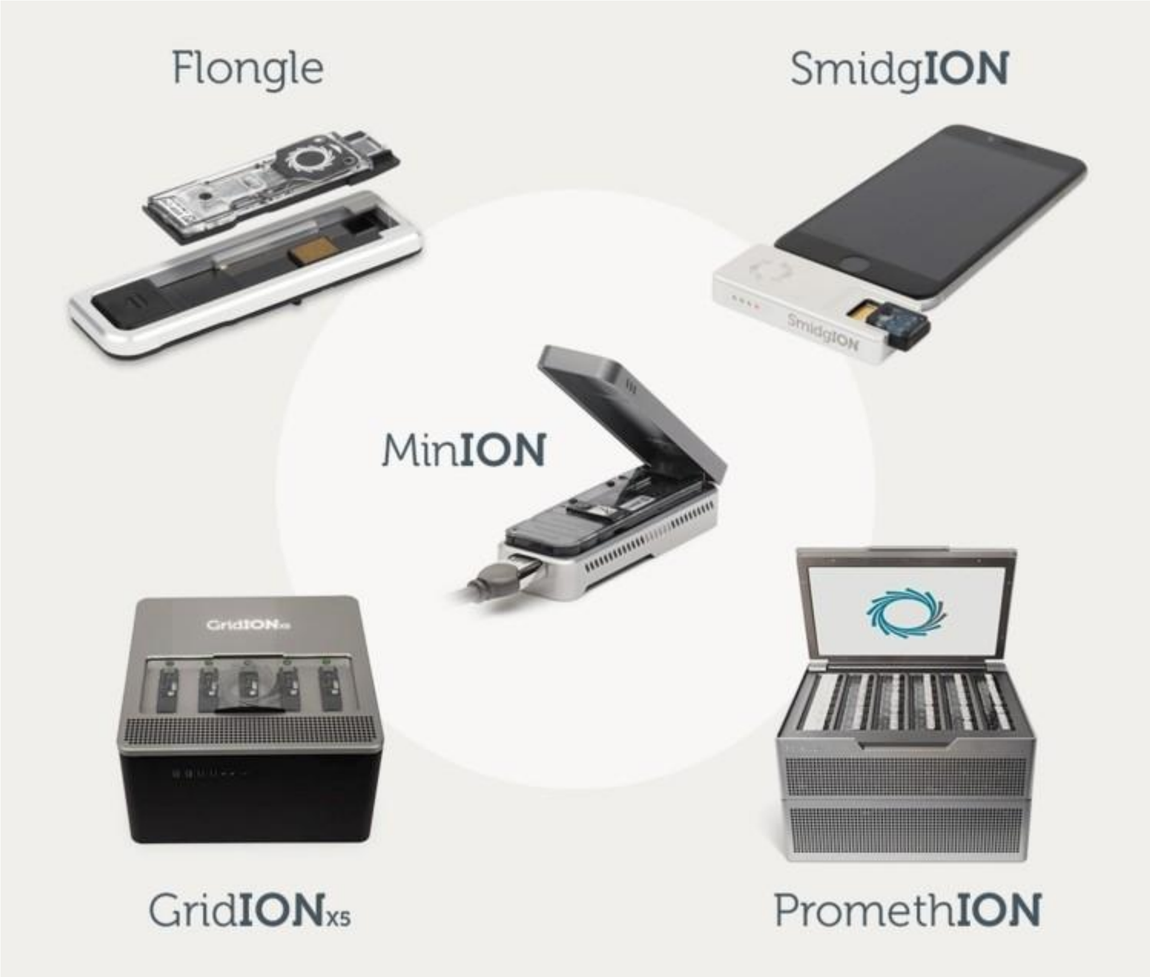
Published: 02 April 2018 **Article history** ▾



NANOPORE SEQUENCING

At the heart of the MinION device, an enzyme unwinds DNA, feeding one strand through a protein pore. The unique shape of each DNA base causes a characteristic disruption in electrical current, providing a readout of the underlying sequence.

Nanopore sequencing



This was used to sequence the first 2019-nCoV genome

HTS: Computational Challenges

- Data management
  - Files are very large; compression algorithms needed
- Read mapping
  - Finding the location on the reference genome
  - All platforms have different data types and error models
  - Repeats!!!!
- Variation discovery
  - Depends on mapping
  - Again, all platforms has strengths and weaknesses
- *De novo* assembly
  - It's very difficult to assemble short sequences and/or long sequences with high errors

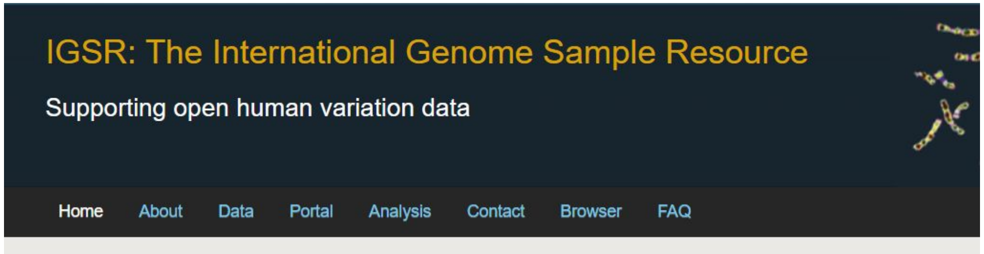
Data science pipeline

1. Identify problem
2. Locate data sources
3. Collect data
4. Prepare data (integrate, transform, clean, filter, aggregate)
5. Build model
6. Evaluate model
7. Communicate results

1) Identify problem

Are there mobile element insertion mutations that cause breast cancer?

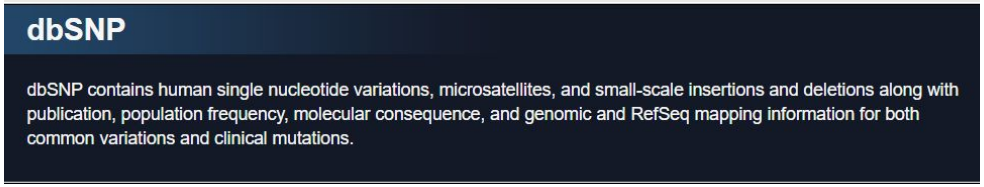
2) Locate data sources / normal



IGSR: The International Genome Sample Resource

Supporting open human variation data

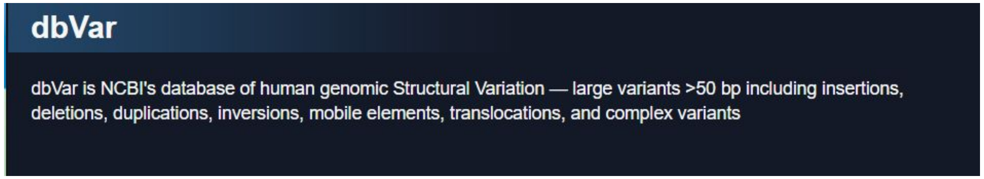
www.internationalgenome.org



dbSNP

dbSNP contains human single nucleotide variations, microsatellites, and small-scale insertions and deletions along with publication, population frequency, molecular consequence, and genomic and RefSeq mapping information for both common variations and clinical mutations.

https://www.ncbi.nlm.nih.gov/snp/



dbVar

dbVar is NCBI's database of human genomic Structural Variation — large variants >50 bp including insertions, deletions, duplications, inversions, mobile elements, translocations, and complex variants

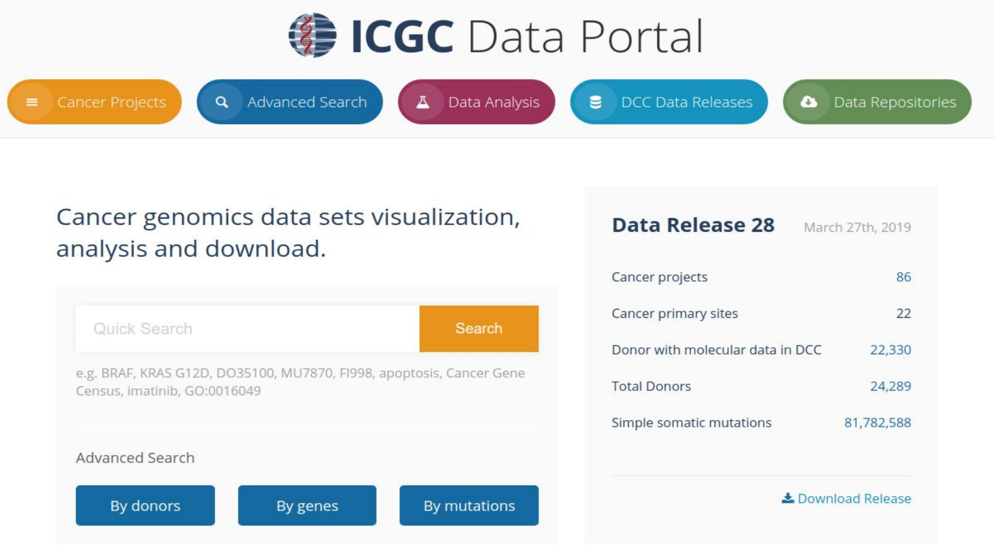
https://www.ncbi.nlm.nih.gov/dbvar/

gnomAD

genome aggregation database

https://gnomad.broadinstitute.org/

2) Locate data sources (II /cancer)



Cancer genomics data sets visualization, analysis and download.

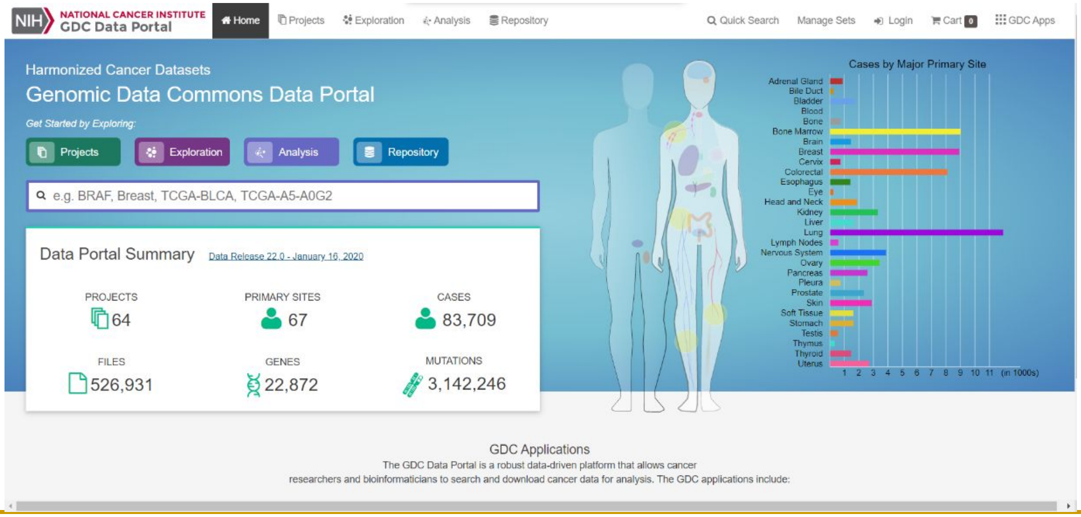
e.g. BRAF, KRAS G12D, DO35100, MU7870, FI998, apoptosis, Cancer Gene Census, imatinib, GO:0016049

Advanced Search

Data Release 28 March 27th, 2019

Cancer projects	86
Cancer primary sites	22
Donor with molecular data in DCC	22,330
Total Donors	24,289
Simple somatic mutations	81,782,588

ICGC



Harmonized Cancer Datasets

Genomic Data Commons Data Portal

Get Started by Exploring:

Data Portal Summary Data Release 22.0 - January 16, 2020

PROJECTS	PRIMARY SITES	CASES
64	67	83,709
FILES	**GENES**	**MUTATIONS**
526,931	22,872	3,142,246

GDC Applications

The GDC Data Portal is a robust data-driven platform that allows cancer researchers and bioinformaticians to search and download cancer data for analysis. The GDC applications include:

TCGA

Tumors and tumor/normal pairs

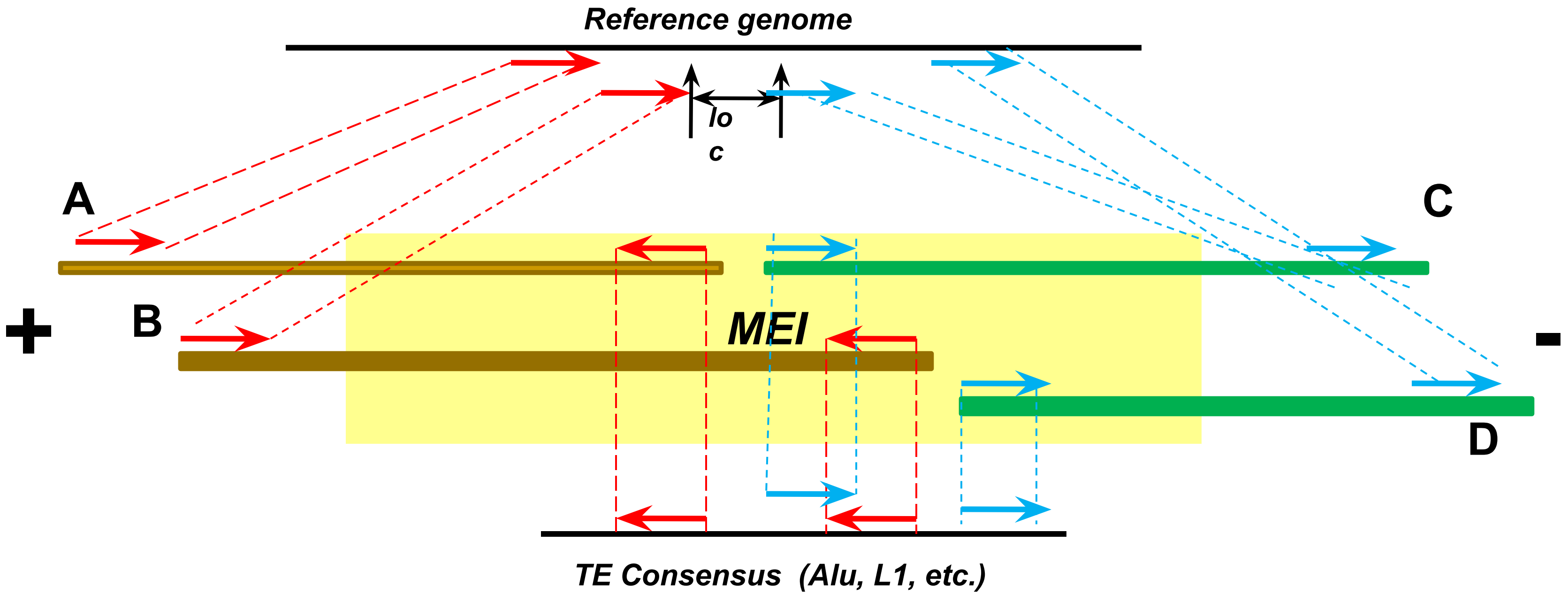
3) Collect data

- Some datasets (i.e., TCGA/ICGC) require access permissions
  - Legal documents, ethical review boards
- Most data available on AWS and GCP
- For local access, download:
  - FTP (will take a long time)
  - Aspera Connect (200-300 Mbit/sec)
    - NCBI and ENA have Aspera servers

4) Prepare data

- Depends on input data type:
  - Raw reads (FASTQ):
    - Map to human reference genome using BWA-MEM
    - Convert to BAM, sort, remove duplicates with SAMTools, sambamba, Picard
  - Aligned reads (BAM/CRAM)
    - Check if the human reference genome version is correct. If not, extract FASTQ, repeat the step above
  - Variation calls (VCF)
    - No preparation necessary, compare across samples
    - Very likely that variation types you are interested in are not listed

5) Build model



- Strand rules: MEI-mapping “+” reads and MEI mapping “-” reads should be in different orientations:
  - +/- and -/+ clusters; or +/+ and -/- clusters (inverted MEI)
- Span rules: A=(A1, A2); B=(B1, B2); C=(C1, C2); D=(D1, D2)
  - |A1-B1| ~ |A2-B2| and |C1-D1| ~ |C2-D2| (simplified; we have 8 rules)
- Location and 2-breakpoint rule:

$$\exists loc, \forall PE : RightMost(+) < loc < LeftMost(-)$$

Illumina paired-end data only

Hormozdiari et al., Bioinformatics 2010

6) Evaluate model

- Implement your new algorithm, or use:
  - TARDIS, MELT, Tangram, Mobster
- Run on normal genomes and tumors
- Filter MEI predictions in tumors that are also found in normal genomes
- Calculate variant allele frequency
  - Require high VAF
- Check if any hit oncogenes, or other functionally important regions of the genome

7) Communicate results

- Create VCF file
- Calculate all necessary statistics
  - Minor Allele Frequency
  - Variant Allele Frequency
  - Genes or functionally relevant regions
  - Pathway analyses
- Generate plots (genome.ucsc.edu)
- Release data