UNIVERSITY OF CAMBRIDGE
Department of Medicine

Analysis of Human Genome Variation (HGV)

Stefan Gräf <sg550@cam.ac.uk>

Computational Genomic and Medicine

Division of Respiratory Medicine

TrainMalta, 15th / 16th September 2016

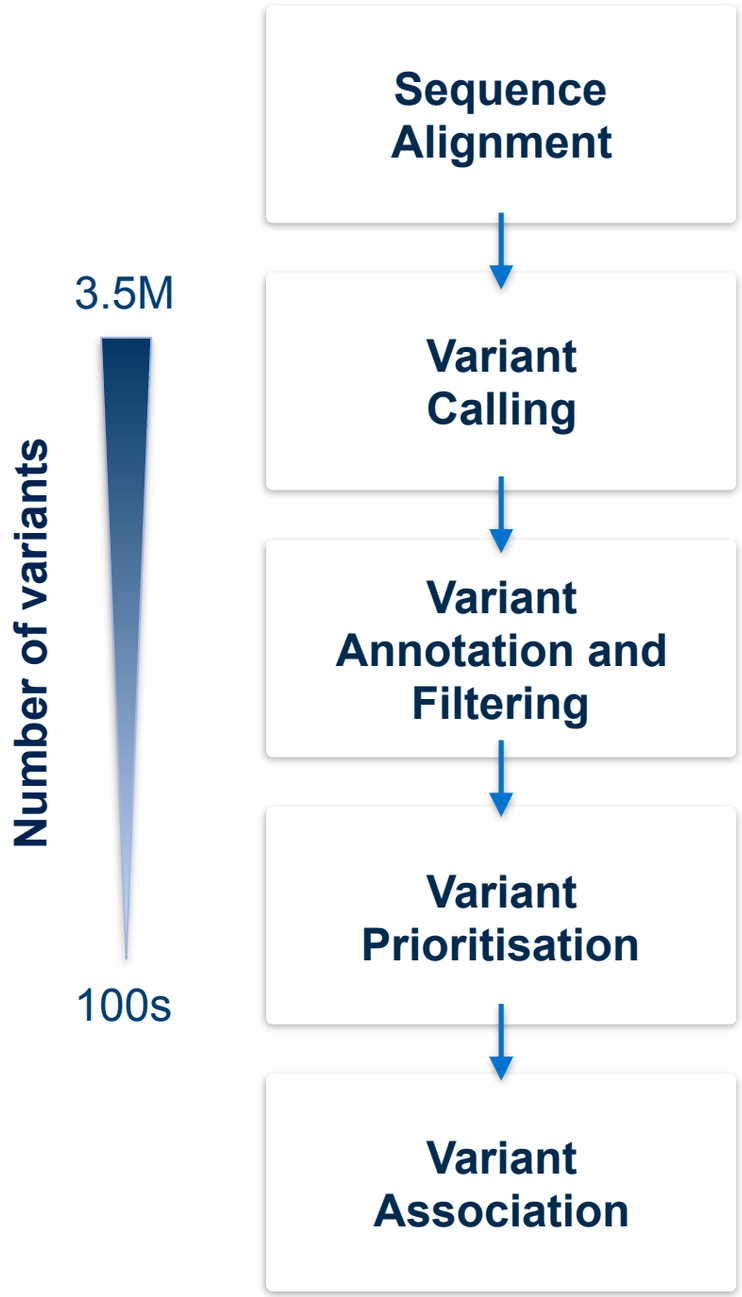
Course Tutors / Acknowledgements

Marta Bleda
Computational Biologist
PostDoc

Stefan Gräf
Computational Biologist
Senior Research Associate

Matthias Haimel
Software Engineer
PhD student

Analysis of Human Genome Variation (HGV)



Outline

Analysis of DNA Sequence Variation

- Introduction
  - Next-generation sequencing
  - Human genetics
- Identification of genetic variation
- Experimental Design
- Analysis pipeline overview

Next-Generation Sequencing

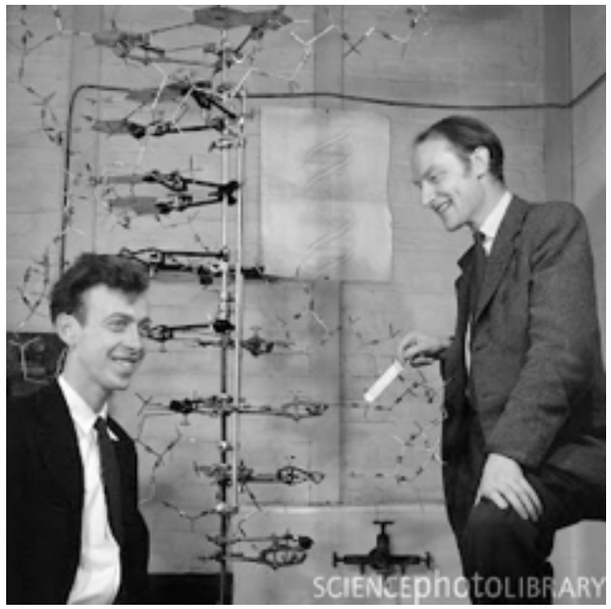
Definition

Non-Sanger-based high-throughput DNA sequencing technologies.

Millions or billions of DNA strands can be sequenced in parallel, yielding substantially more throughput and minimising the need for the fragment-cloning methods that are often used in Sanger sequencing of genomes.

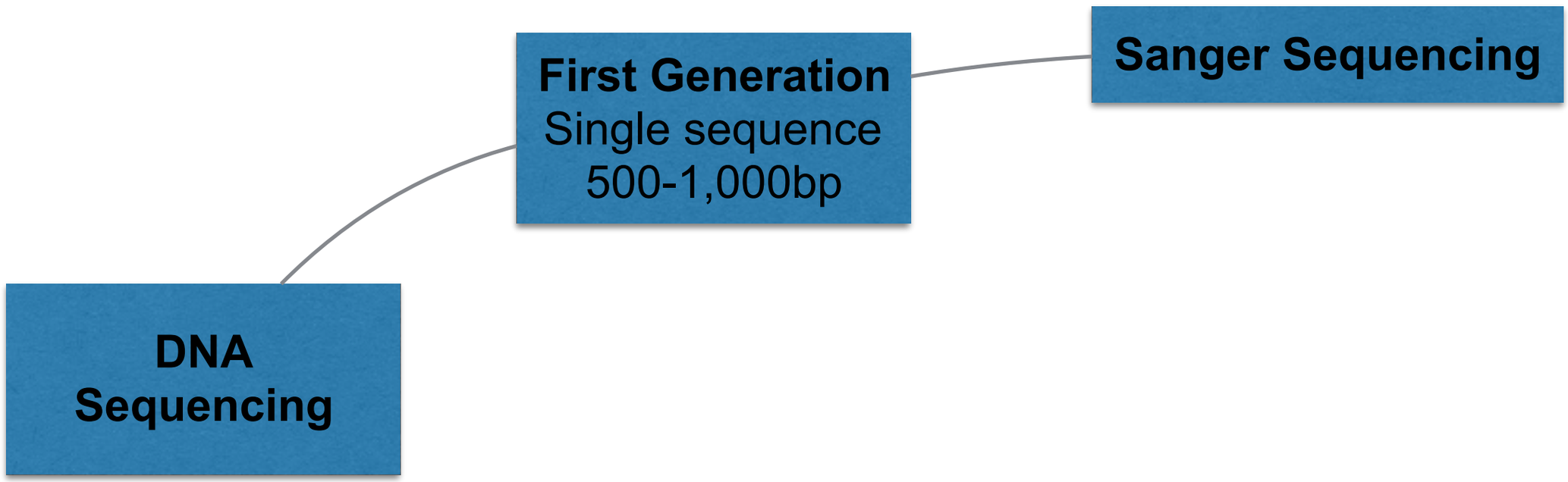
More than 60 Years of Genome Research

1953

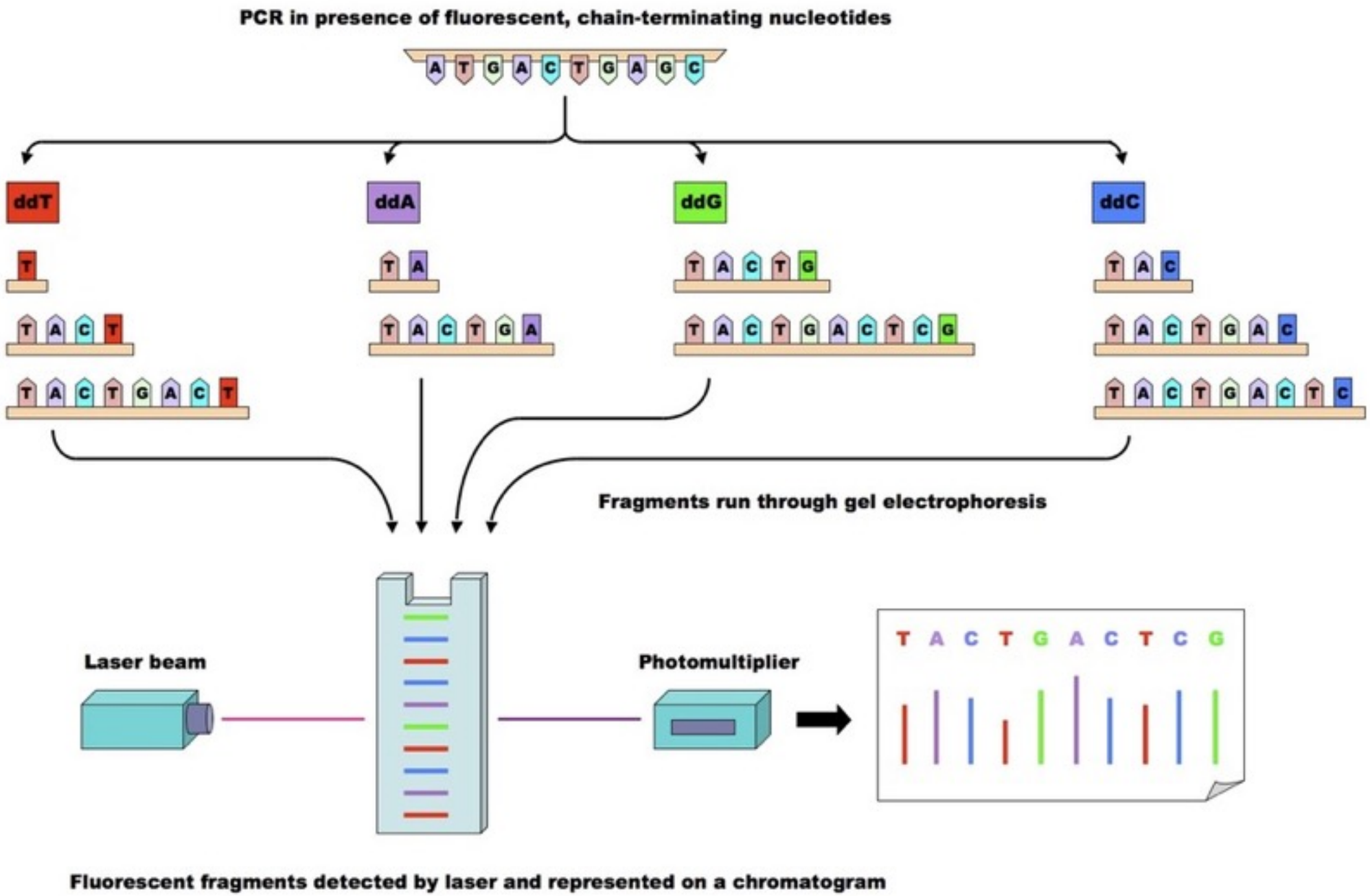


Structure of
the DNA

Evolution of Sequencing



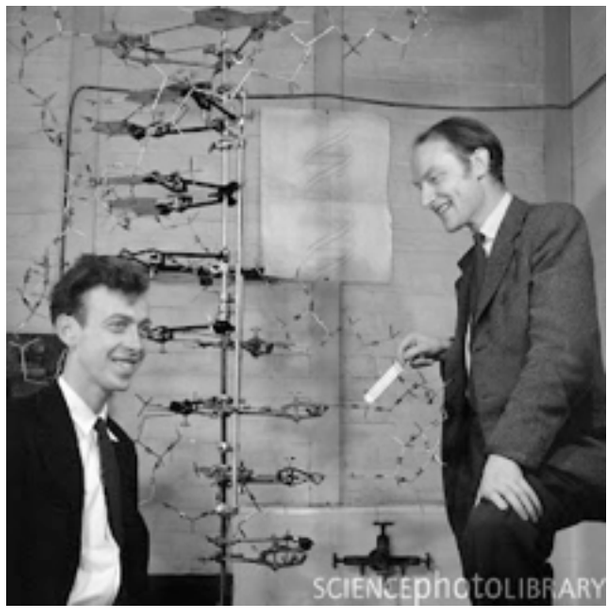
Sanger Sequencing



Fluorescent fragments detected by laser and represented on a chromatogram

More than 60 Years of Genome Research

1953



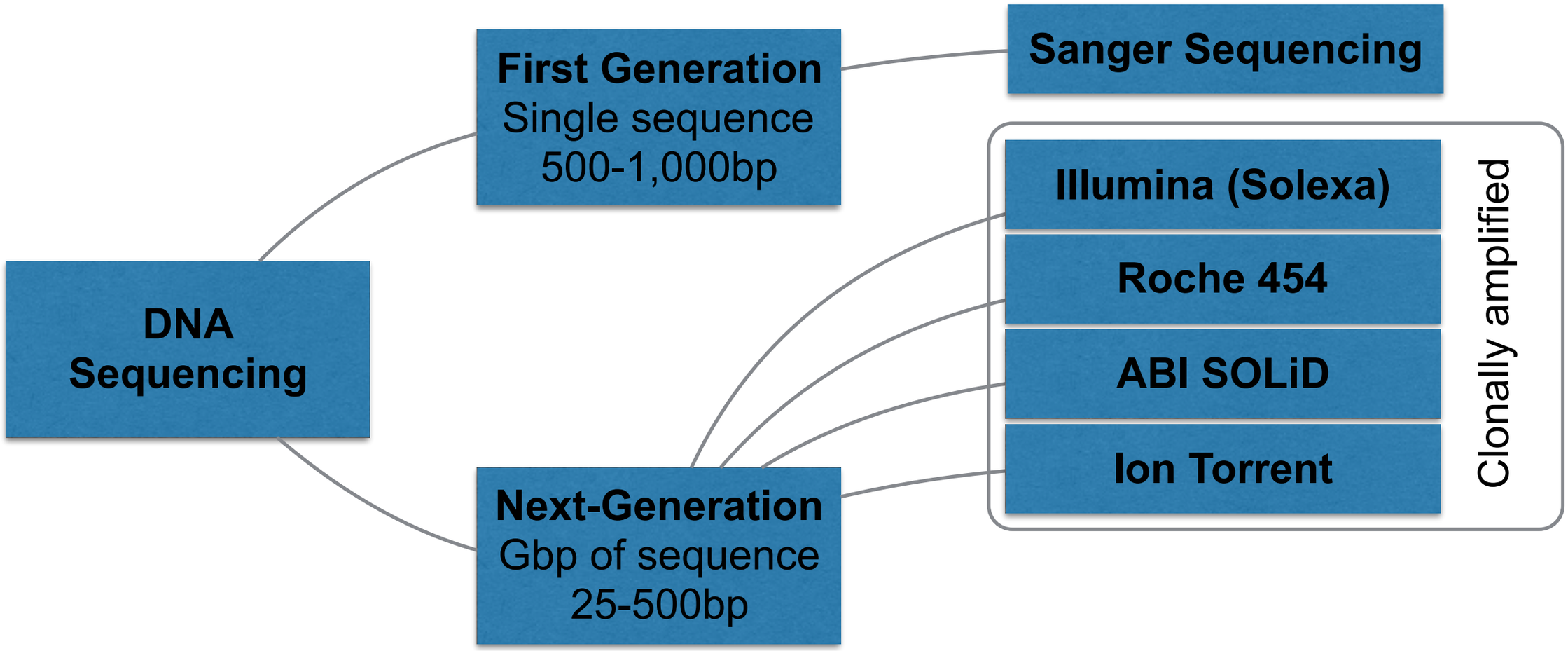
2001

Structure of the DNA

Reference genome

Sanger Sequencing

Evolution of Sequencing

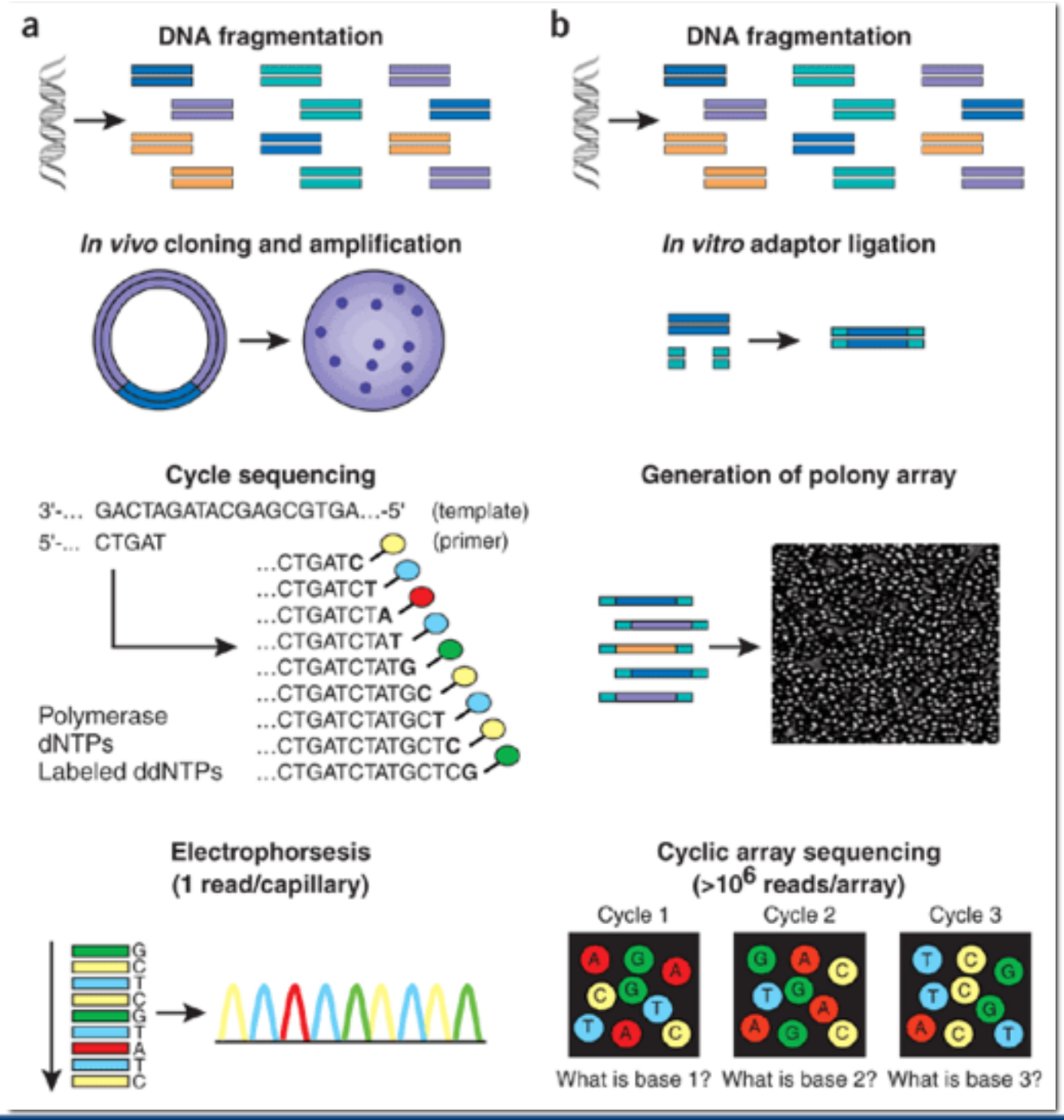


Illumina (Solexa)

Illumina Nextera library preparation, paired-end sequencing and analysis

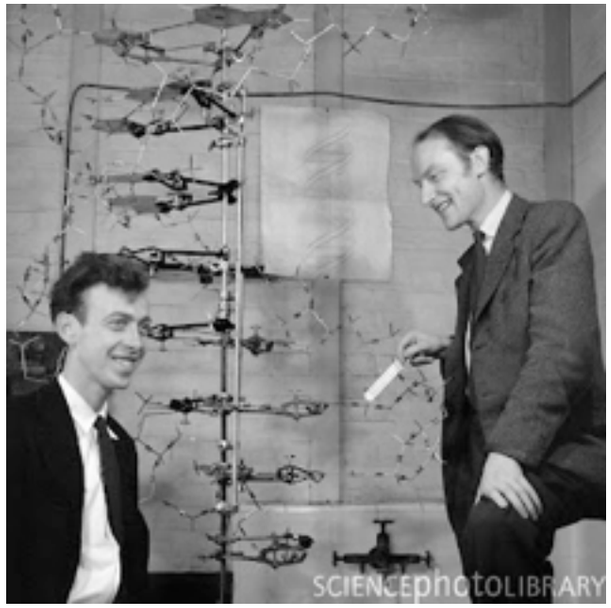
https://youtu.be/womKfikWlxM

Sanger vs. Next-Generation Sequencing



More than 60 Years of Genome Research

1953



2001

2007

Structure of the DNA

Reference genome

Personal genomes

Sanger Sequencing

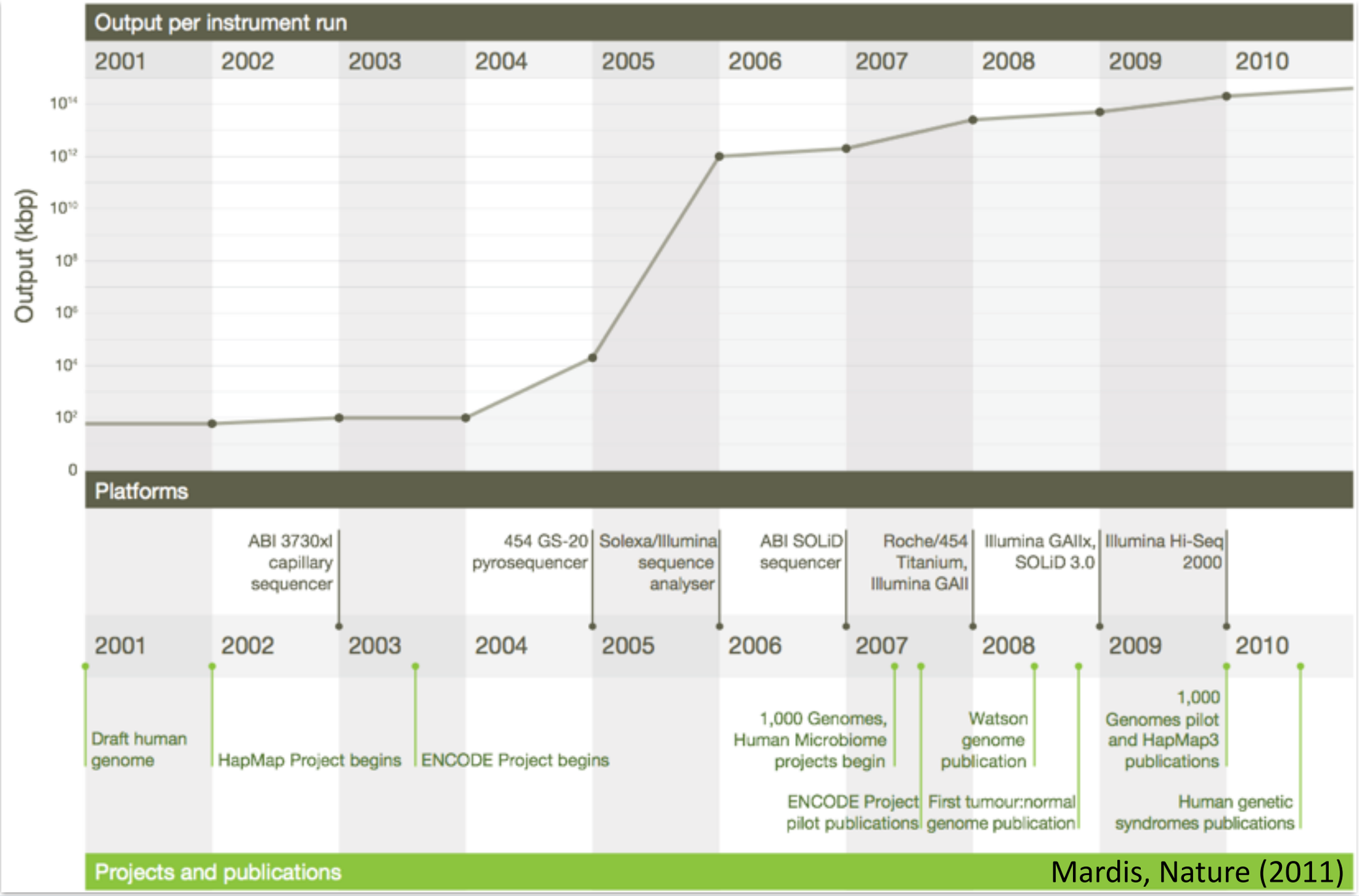
High-throughput Next-generation Sequencing

Illumina Flow Cell

- A flow cell contains **8 lanes**
- Each lane is subdivided into **100 image tiles**
- Per cycle **4 images** (A, G, T, C) are taken
- 8 lanes x 100 tiles x 4 bases x 50 cyles => **160,000 images**
- An image has a size of 7.3 MB => **1.2 TB per run**
- The imaging takes up most of the time
- **HiSeq X**: dual flow cell, 2x150bp reads, 5.3-6 billion reads pass QC, run takes less than 3 days, >75% of bases are above Q30
- Never write images to disc …

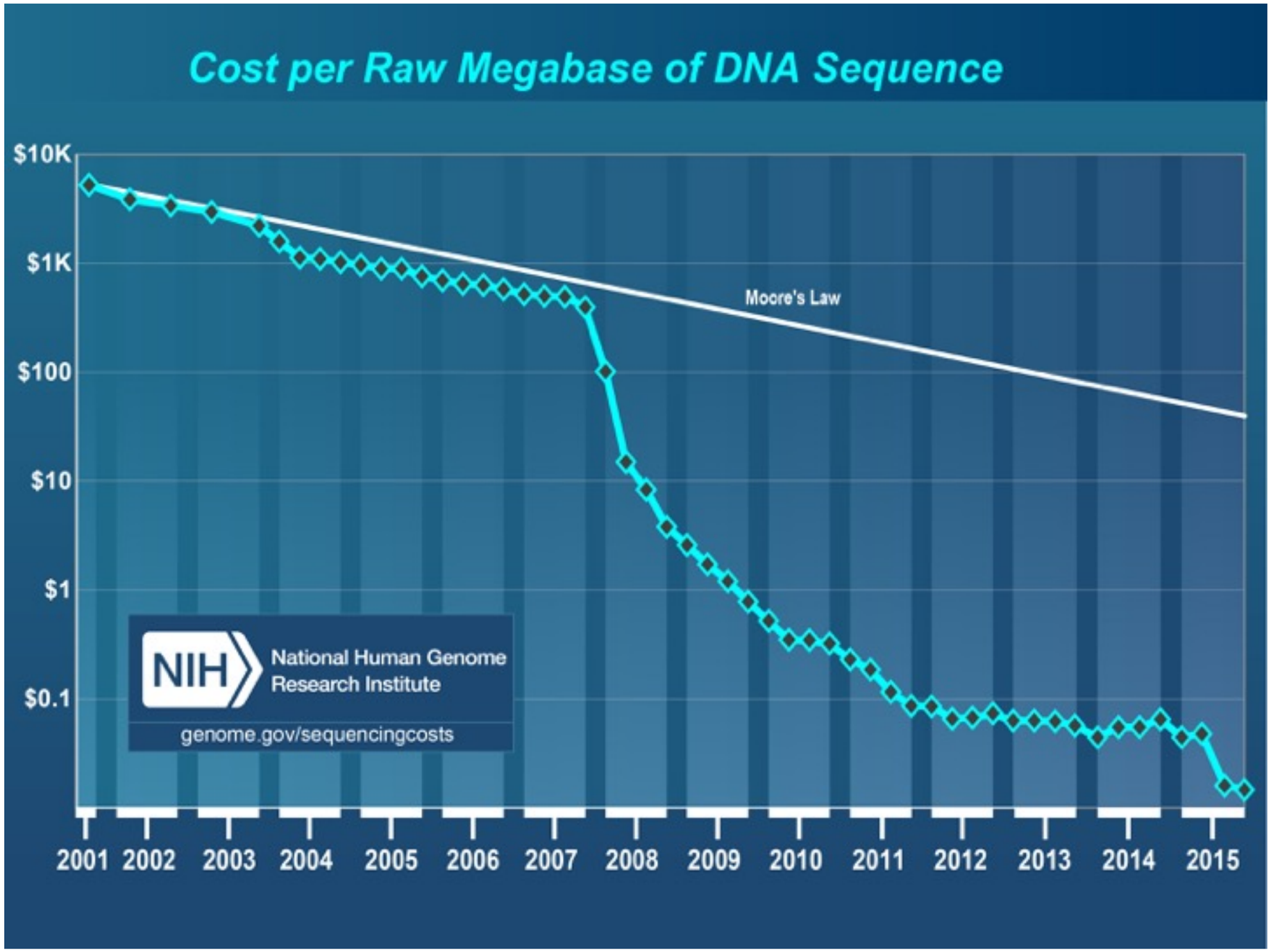
HISEQ X	DUAL FLOW CELL	SINGLE FLOW CELL
Output per Run	1.6-1.8 Tb	800-900 Gb
Reads Passing Filter	5.3-6 billion	2.6-3 billion
Supported Read Length	2 × 150 bp	
Run Time	< 3 days	
Quality Scores	≥ 75% of bases above Q30 at 2 x 150 bp	
Supported Library Preparation	• TruSeq DNA PCR-Free Library Prep Kit • TruSeq Nano DNA Library Prep Kit	

Throughput Growth Over 10 Years

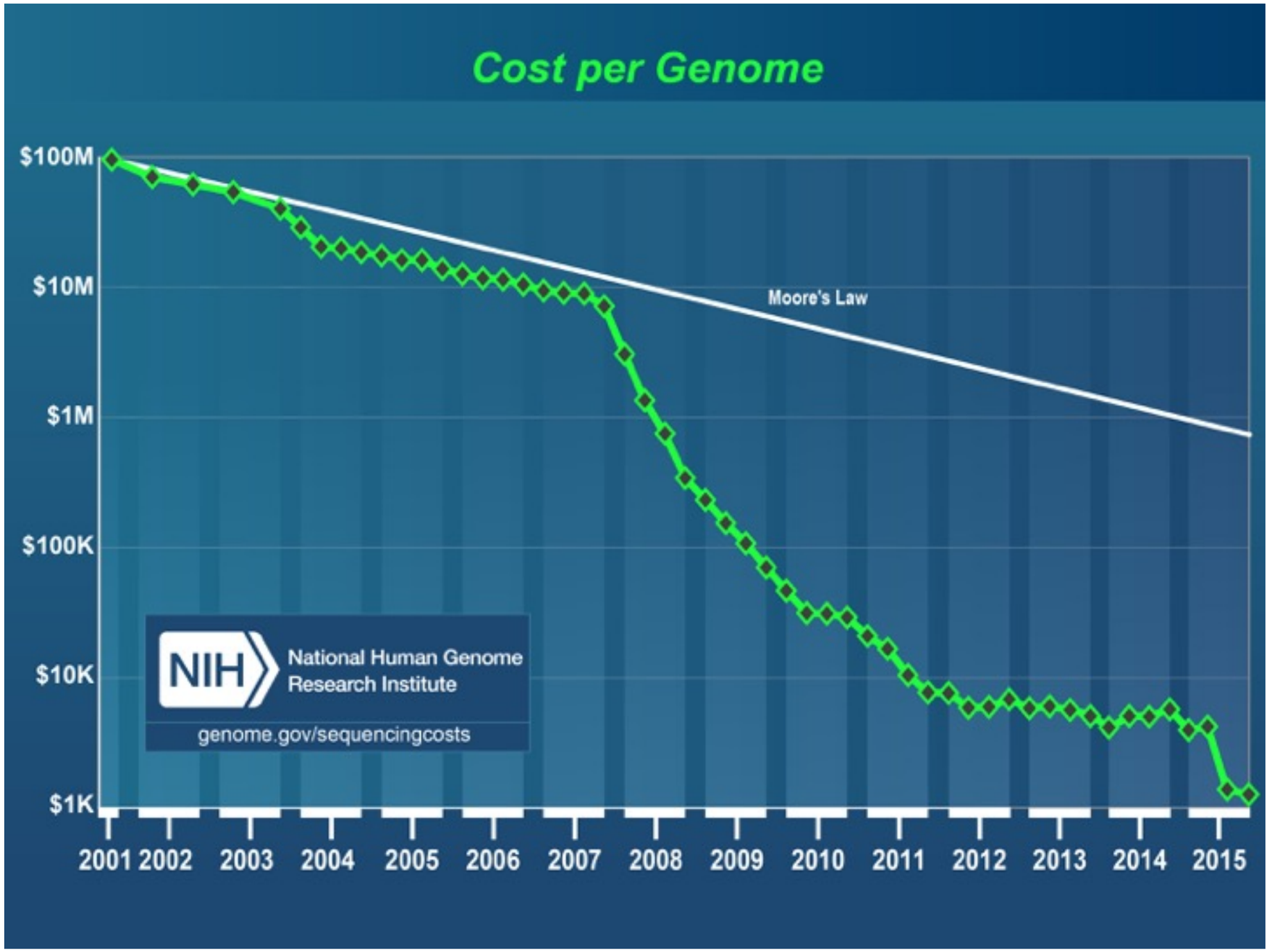


Mardis, Nature (2011)

Cost of Sequencing



Cost of Sequencing



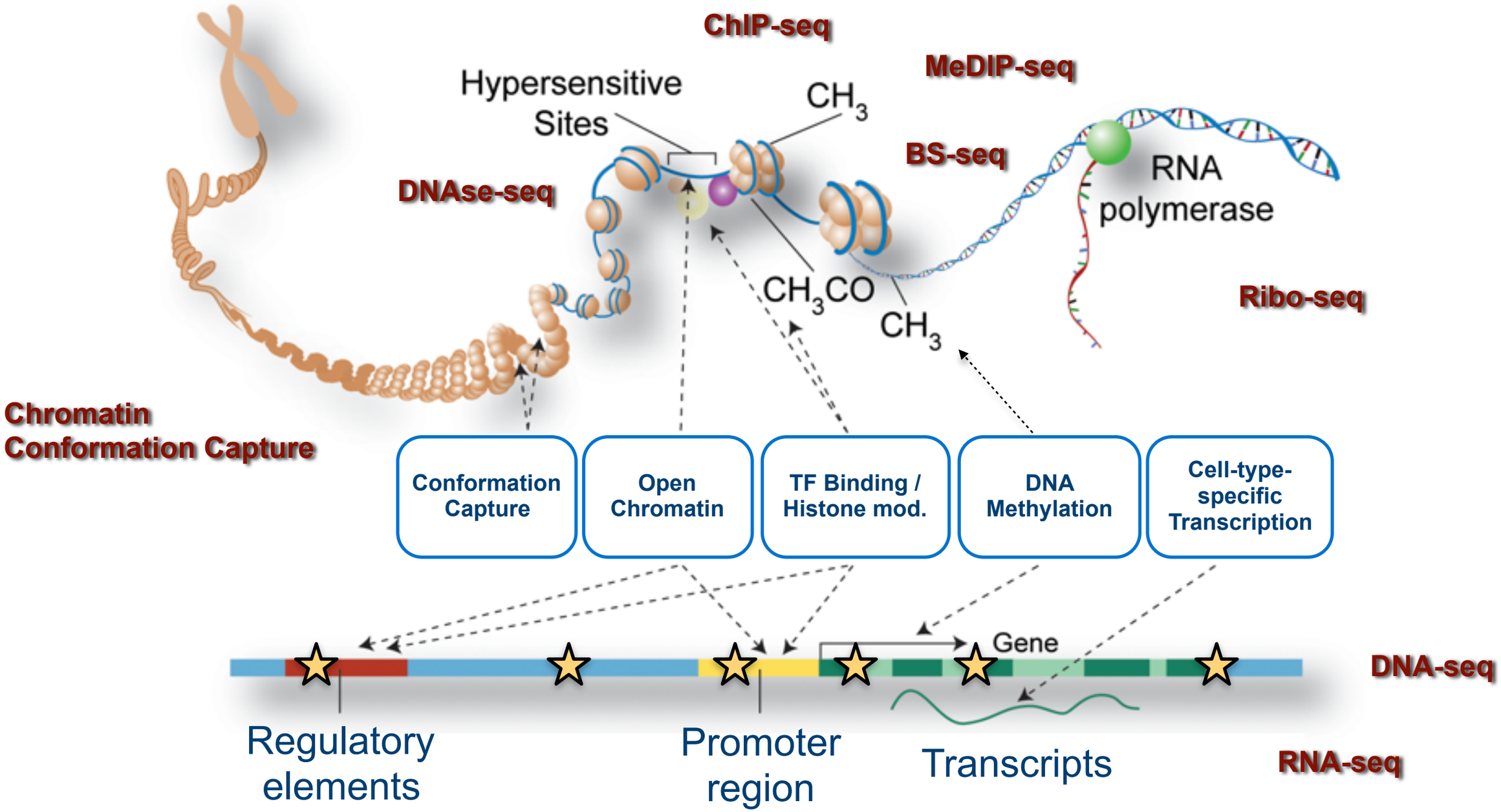
Personal Genomes

Personal Genome	Platform	Genomic template libraries	No. of reads (millions)	Read length (bases)	Base coverage (fold)	Assembly	Genome coverage (%)*	SNVs in millions (alignment tool)	No. of runs	Estimated cost (US$)
J. Craig Venter	Automated Sanger	MP from BACs, fosmids & plasmids	31.9	800	7.5	*De novo*	N/A	3.21	>340,000	70,000,000
James D. Watson	Roche/454	Frag: 500 bp	93.2‡	250§	7.4	Aligned*	95‖	3.32 (BLAT)	234	1,000,000¶
Yoruban male (NA18507)	Illumina/ Solexa	93% MP: 200 bp	3,410‡	35	40.6	Aligned*	99.9	3.83 (MAQ)	40	250,000¶
		7% MP: 1.8 kb	271	35				4.14 (ELAND)		
Han Chinese male	Illumina/ Solexa	66% Frag: 150–250 bp	1,921‡	35	36	Aligned*	99.9	3.07 (SOAP)	35	500,000¶
		34% MP: 135 bp & 440 bp	1,029	35						
Korean male (AK1)	Illumina/ Solexa	21% Frag: 130 bp & 440 bp	393‡	36	27.8	Aligned*	99.8	3.45 (GSNAP)	30	200,000¶
		79% MP: 130 bp, 390 bp & 2.7 kb	1,156	36, 88, 106						
Korean male (SJK)	Illumina/ Solexa	MP: 100 bp, 200 bp & 300 bp	1,647‡	35, 74	29.0	Aligned*	99.9	3.44 (MAQ)	15	250,000¶,#
Yoruban male (NA18507)	Life/APG	9% Frag: 100–500 bp	211‡	50	17.9	Aligned*	98.6	3.87 (Corona-lite)	9.5	60,000¶,**
		91% MP: 600–3,500 bp	2,075‡	25, 50						
Stephen R. Quake	Helicos BioSciences	Frag: 100–500 bp	2,725‡	32§	28	Aligned*	90	2.81 (IndexDP)	4	48,000¶
AML female	Illumina/ Solexa	Frag: 150–200 bp‡‡	2,730‡,‡‡	32	32.7	Aligned*	91	3.81‡‡ (MAQ)	98	1,600,000‖‖
		Frag: 150–200 bp§§	1,081‡,§§	35	13.9		83	2.92§§ (MAQ)	34	
AML male	Illumina/ Solexa	MP: 200–250 bp‡‡	1,620‡,‡‡	35	23.3	Aligned*	98.5	3.46‡‡ (MAQ)	16.5	500,000‖‖
		MP: 200–250 bp§§	1,351‡,§§	50	21.3		97.4	3.45§§ (MAQ)	13.1	
James R. Lupski CMT male	Life/APG	16% Frag: 100–500 bp	238‡	35	29.6	Aligned*	99.8	3.42 (Corona-lite)	3	75,000¶,¶¶
		84% MP: 600–3,500 bp	1,211‡	25, 50						

*A minimum of one read aligning to the National Center for Biotechnology Information build 36 reference genome. ‡Mappable reads for aligned assemblies. §Average read-length. ‖D. Wheeler, personal communication. ¶Reagent cost only. #S.-M. Ahn, personal communication. **K. McKernan, personal communication. ‡‡Tumour sample. §§Normal sample. ‖‖Tumour & normal samples: reagent, instrument, labour, bioinformatics and data storage cost, E. Mardis, personal communication. ¶¶R. Gibbs, personal communication. AML, acute myeloid leukaemia; BAC, bacterial artificial chromosome; CMT, Charcot–Marie–Tooth disease; Frag, fragment; MP, mate-pair; N/A, not available; SNV, single-nucleotide variant.

Metzker, Nat Rev Genet (2010)

Next-Generation Sequencing Helps Interrogating Many Omic Features of a Cell

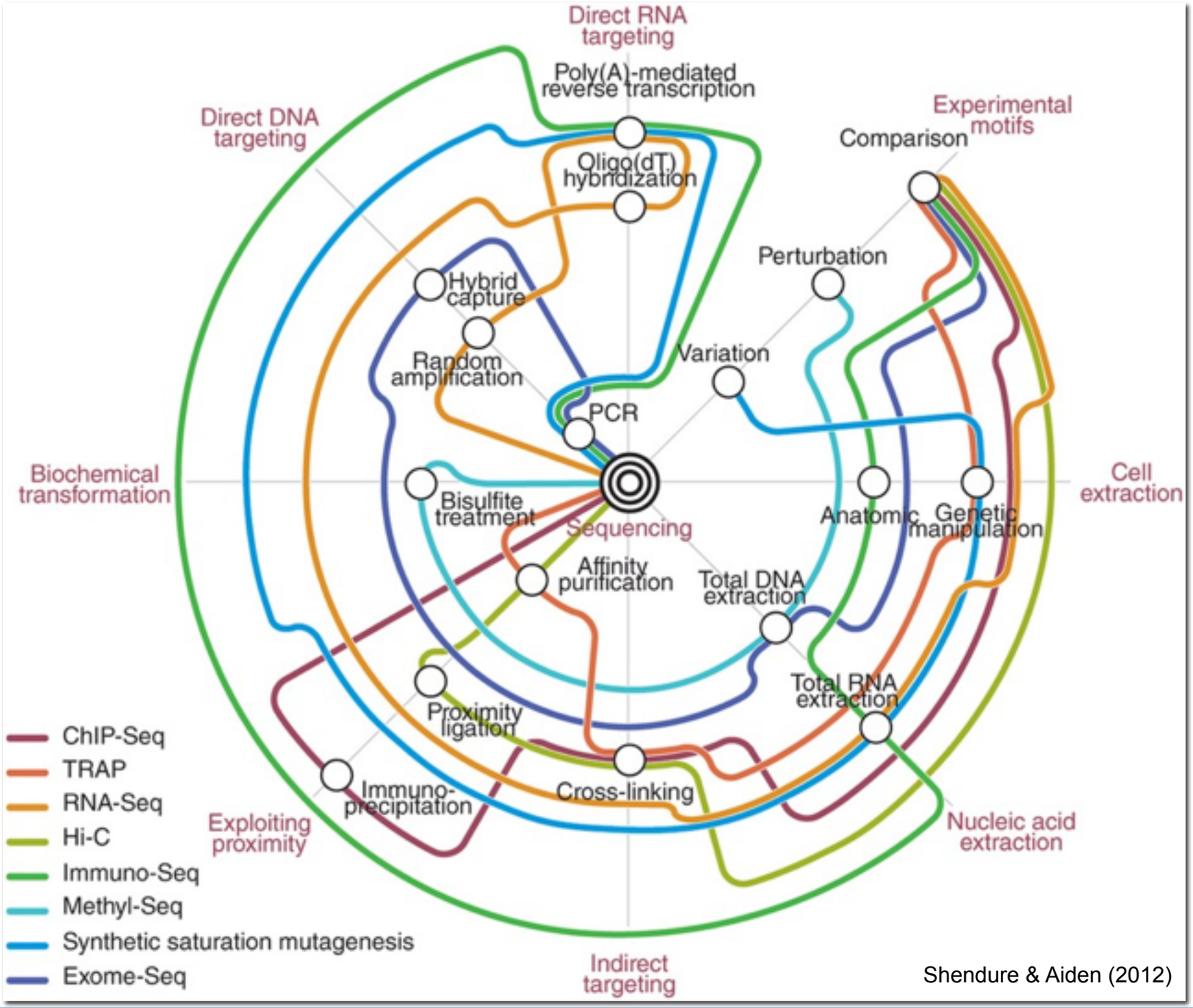


The ENCODE Project Consortium (adapted)

Method	Sequencing to determine:	Subway' route as defined in next figure
DNA-Seq	A genome sequence	Comparison, 'anatomic' (isolation by anatomic site), flow cytometery, DNA extraction, mechanical shearing, adaptor ligation, PCR and sequencing
Targeted DNA-Seq	A subset of a genome (for example, an exome)	Comparison, cell culture, DNA extraction, mechanical shearing, adaptor ligation, PCR, hybridization capture, PCR and sequencing
Methyl-Seq	Sites of DNA methylation, genome-wide	Perturbation, genetic manipulation, cell culture, DNA extraction, mechanical shearing, adaptor ligation, bisulfite conversion, PCR and sequencing
Targeted methyl-Seq	DNA methylation in a subset of the genome	Comparison, cell culture, DNA extraction, bisulfite conversion, molecular inversion probe capture, circularization, PCR and sequencing
DNase-Seq, Sono-Seq and FAIRE-Seq	Active regulatory chromatin (that is, nucleosome-depleted)	Perturbation, cell culture, nucleus extraction, DNase I digestion, DNA extraction, adaptor ligation, PCR and sequencing
MAINE-Seq	Histone-bound DNA (nucleosome	Comparison, cell culture, MNase I digestion, DNA extraction, adaptor ligation, PCR and sequencing
ChIP-Seq	Protein-DNA interactions (using chromatin immunoprecipitation)	Comparison, 'anatomic', cell culture, cross-linking, mechanical shearing, immunoprecipitation, DNA extraction, adaptor ligation, PCR and sequencing
RIP-Seq, CLIP-Seq, HITS-CLIP	Protein-RNA interactions	Variation, cross-linking, 'anatomic', RNase digestion, immunoprecipitation, RNA extraction, adaptor ligation, reverse transcription, PCR and sequencing
RNA-Seq	RNA (that is, the transcriptome)	Comparison, 'anatomic', RNA extraction, poly(A) selection, chemical fragmentation, reverse transcription, second-strand synthesis, adaptor ligation, PCR and sequencing
FRT-Seq	Amplification-free, strand-specific transcriptome sequencing	Comparison, 'anatomic', RNA extraction, poly(A) selection, chemical fragmentation, adaptor ligation, reverse transcription and sequencing
NET-Seq	Nascent transcription	Perturbation, genetic manipulation, cell culture, immunoprecipitation, RNA extraction, adaptor ligation, reverse transcription, circularization, PCR and sequencing
Hi-C	Three-dimensional genome structure	Comparison, cell culture, cross-linking, proximity ligation, mechanical shearing, affinity purification, adaptor
Chia-PET	Long-range interactions mediated by a protein	Perturbation, cell culture, cross-linking, mechanical shearing, immunoprecipitation, proximity ligation, affinity purification, adaptor ligation, PCR and sequencing
Ribo-Seq	Ribosome-protected mRNA fragments (that is, active translation)	Comparison, cell culture, RNase digestion, ribosome purification, RNA extraction, adaptor ligation, reverse transcription, rRNA depletion, circularization, PCR and sequencing
TRAP	Genetically targeted purification of polysomal mRNAs	Comparison, genetic manipulation, 'anatomic', cross-linking, affinity purification, RNA extraction, poly(A) selection, reverse transcription, second-strand synthesis, adaptor ligation, PCR and sequencing
PARS	Parallel analysis of RNA structure	Comparison, cell culture, RNA extraction, poly(A) selection, RNase digestion, chemical fragmentation, adaptor ligation, reverse transcription, PCR and sequencing
Synthetic saturation mutagenesis	Functional consequences of genetic variation	Variation, genetic manipulation, barcoding, RNA extraction, reverse transcription, PCR and sequencing
Immuno-Seq	The B-cell and T-cell repertoires	Perturbation, 'anatomic', DNA extraction, PCR and sequencing
Deep protein mutagenesis	Protein binding activity of synthetic peptide libraries or variants	Variation, genetic manipulation, phage display, *in vitro* competitive binding, DNA extraction, PCR and sequencing
PhIT-Seq	Relative fitness of cells containing disruptive insertions in diverse genes	Variation, genetic manipulation, cell culture, competitive growth, linear amplification, adaptor ligation, PCR and sequencing

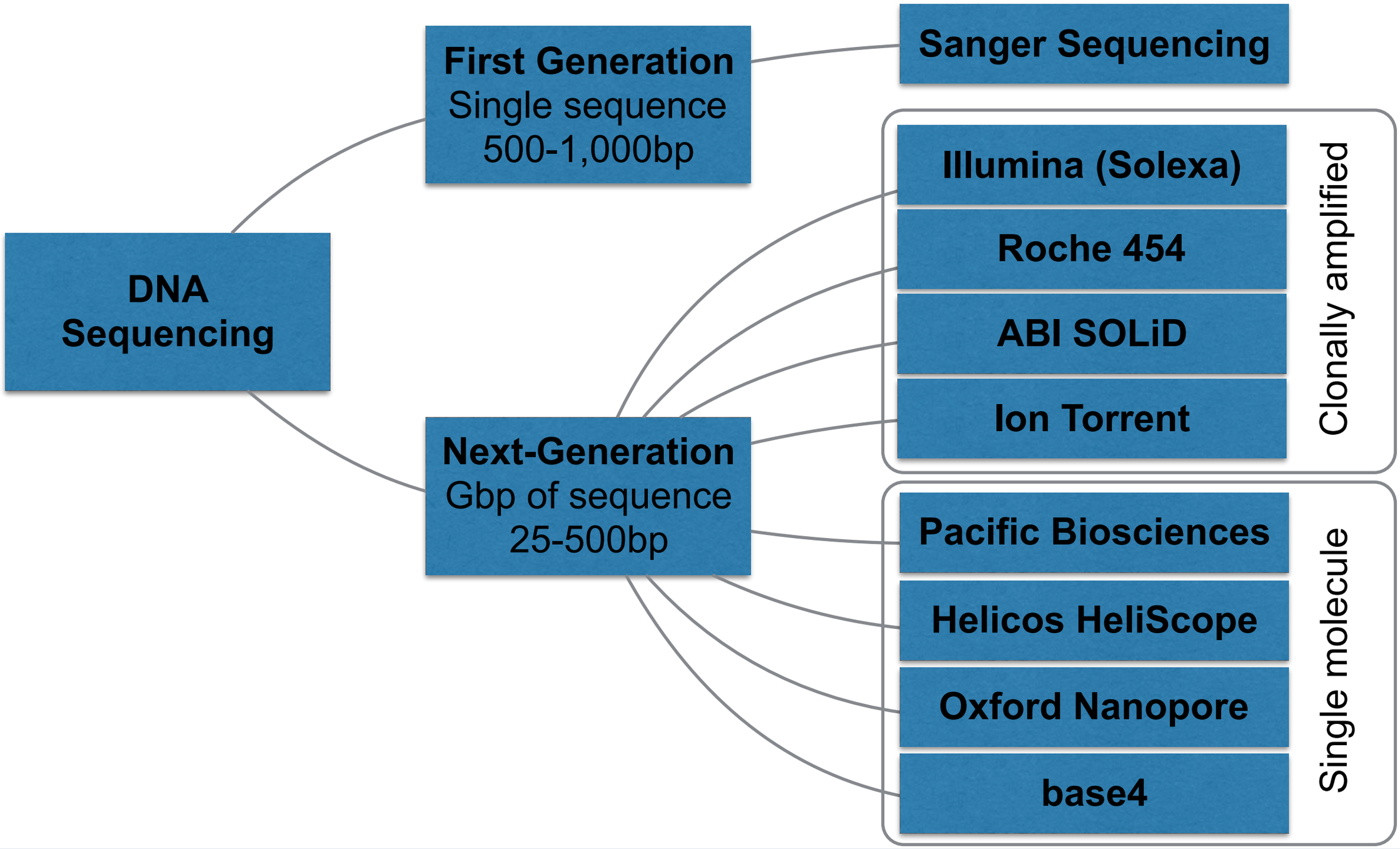
Shendure & Aiden (2012), adapted

Subway Map of Core Techniques



Shendure & Aiden (2012)

Evolution of Sequencing

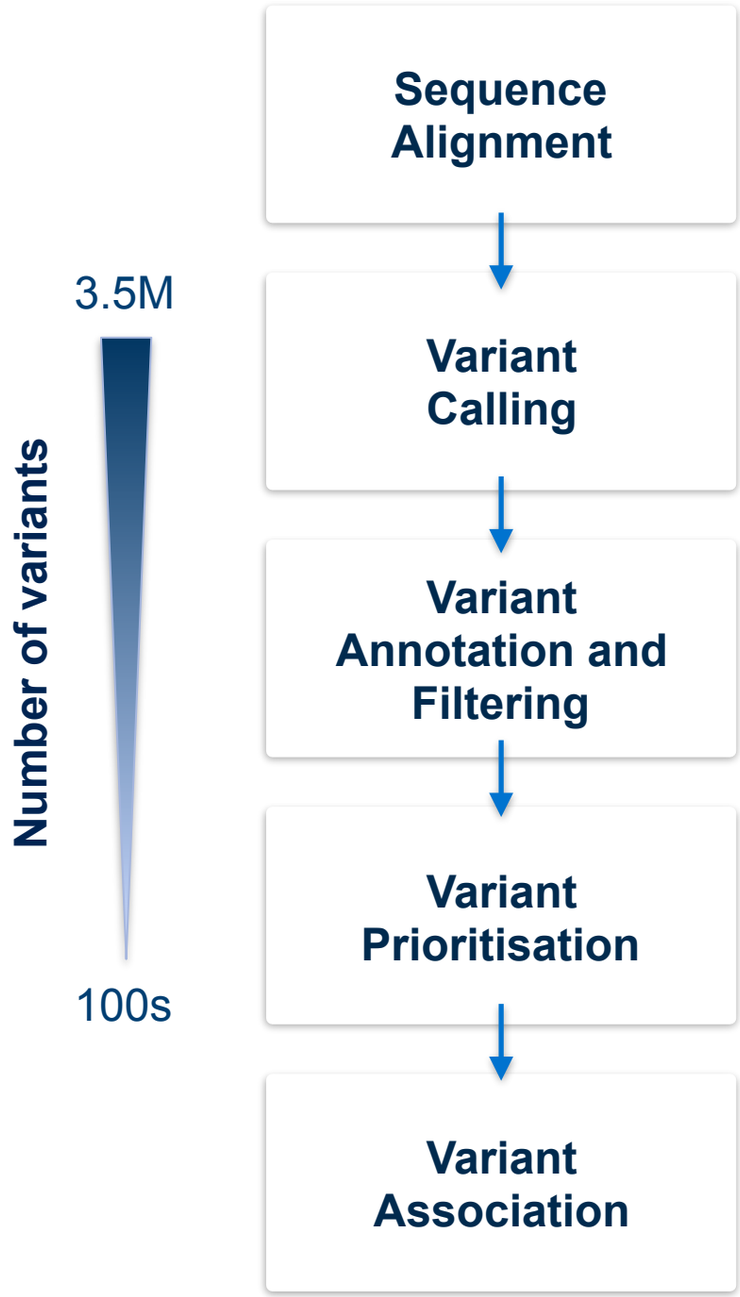


Oxford Nanopore Technology

Nanopore DNA sequencing:

https://vimeo.com/127689053?from=outro-embed

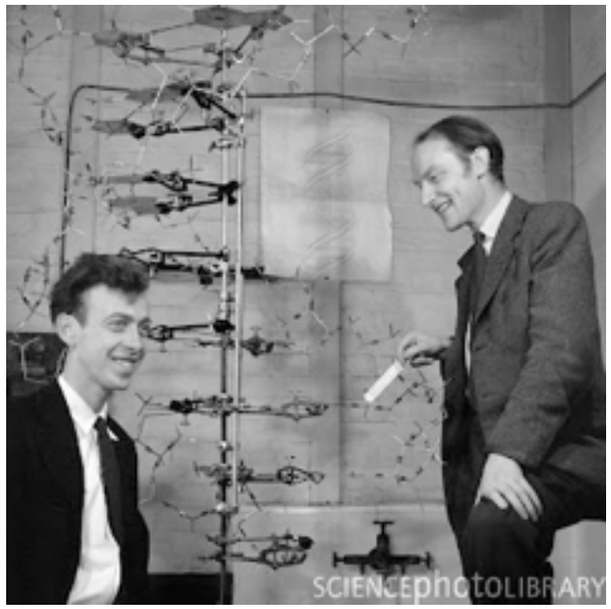
Analysis of Human Genome Variation (HGV)



Human Genome Variation

More than 60 Years of Genome Research

1953



Structure of the DNA

2001

Reference genome

Sanger Sequencing

2007

Personal genomes

High-throughput Next-generation Sequencing

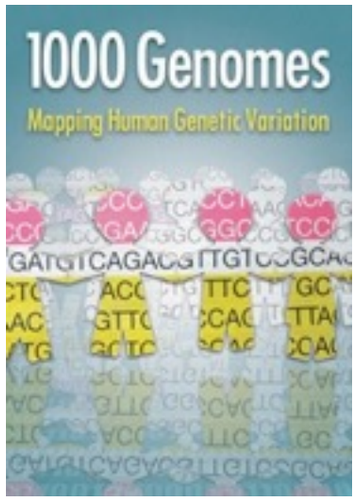
2010

Personal genomes x1000+

Human Genome Variation

1000 Genomes Project (Nature, 1 Nov 2012)

- Aims to understand the genetic contribution to disease
- 1092 individuals from 14 populations
- Low-coverage whole-exome and whole-genome sequencing
- Validated haplotype map of
  - 38 million single nucleotide polymorphisms
  - 1.4 million short insertions and deletions
  - more than 14,000 larger deletions



http://www.1000genomes.org

UK10K

Rare Genetic Variants in Health and Disease

- Better understand link between low-frequency and rare genetic changes and human disease caused by harmful changes to the proteins the body makes.
- Study the genetic code of 10,000 people in much finer detail than ever before.
  - 4,000 whole genomes of deeply phenotyped cohorts (i.e. TwinsUK and ASLPAC) at 6x depth
  - 6,000 whole exomes of extreme phenotypes of specific conditions
- Provide a sequence variation resource for future studies

http://www.uk10k.org

10,000 Whole Genomes …

UNIVERSITY OF CAMBRIDGE | Study at Cambridge | About the University | Research at Cambridge | Quick links | Search

/ Research / News / New initiative will sequence 10,000 whole genomes of people with rare genetic diseases

Research

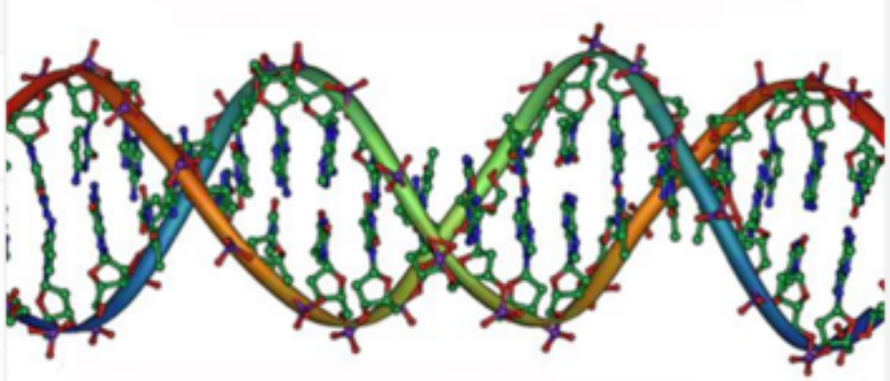
Home | News | Features | Discussion | Video and audio | Spotlight on… | Research at Cambridge | Innovation at Cambridge

New initiative will sequence 10,000 whole genomes of people with rare genetic diseases

Research

News

New initiative will sequence 10,000 whole genomes of people with rare genetic diseases



Published

21 Oct 2013

Image

An overview of the structure of DNA

Credit: Michael Ströck

Project will lay foundation for genomic medicine.

University of Cambridge, Genomics England and Illumina, Inc. today announced the start of a three-year project that will sequence 10,000 whole genomes of children and adults with rare genetic diseases. The project represents a pilot for Genomics England, which will provide 2,000 samples, and marks the beginning of the national endeavor to sequence 100,000 genomes in the UK National Health Service (NHS), announced recently by the Prime Minister, David Cameron.

"This project will bring enormous improvements to the care of patients with rare genetic diseases. It will shorten the gap between the first signs of ill-health in a person and providing a conclusive diagnosis by using the power of modern DNA sequencing methods," said Dr John Bradley,

> *"This project will bring enormous improvements to the care of patients with rare genetic diseases. It will shorten the gap between the first signs of ill-health in a person and providing a conclusive diagnosis by using the power of modern DNA sequencing methods"*
>
> *— Dr John Bradley, Director of the NIHR Cambridge Biomedical Research Centre*

Share

Email 10 | reddit 0 | Share 352 | Tweet 112 | ShareThis 609

Related articles

New technique will transform epigenetics research

Offensive manoeuvres in the war against HIV

A new dimension to DNA and personalised medicine of the future

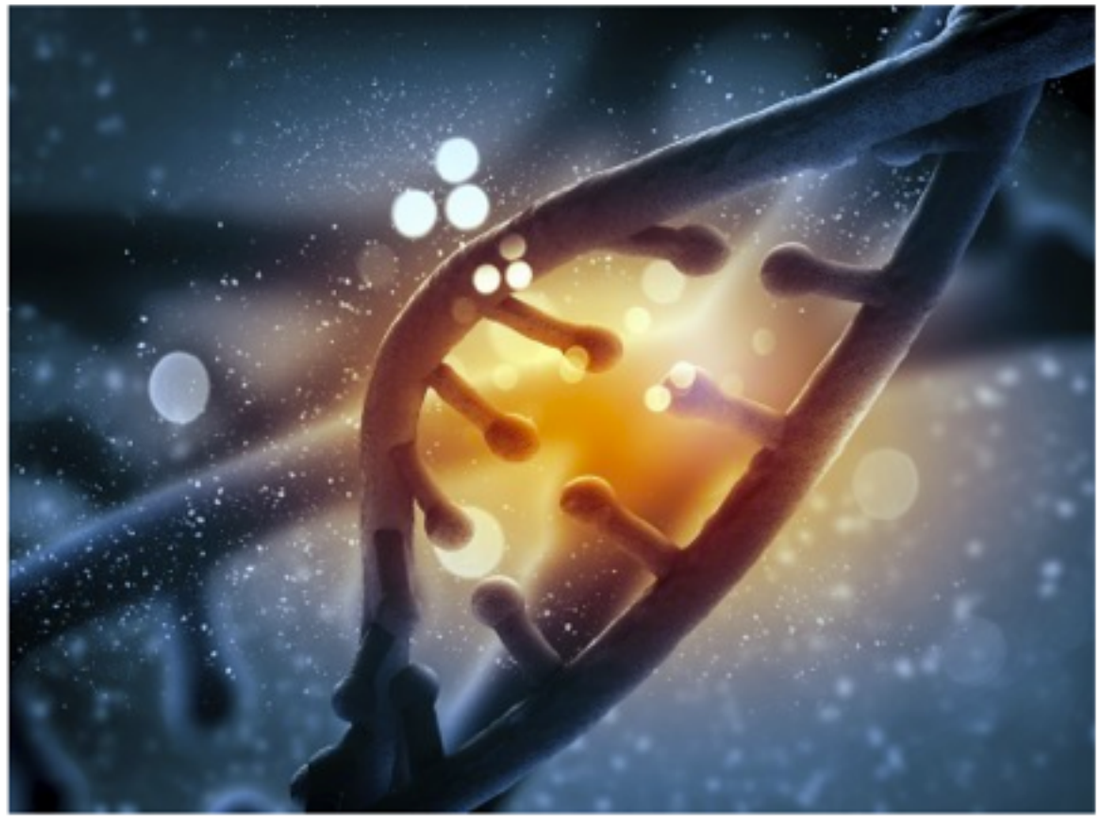
Scientists discover how antibiotic molecule found in bacteria stops breast cancer

Tracking MRSA in Real Time

Genomics England — 100,000 Genome Project

Genomics england

Home | About us | 100,000 Genomes Project | GeCIP | GENE Consortium | Library & resources | News | Contact us



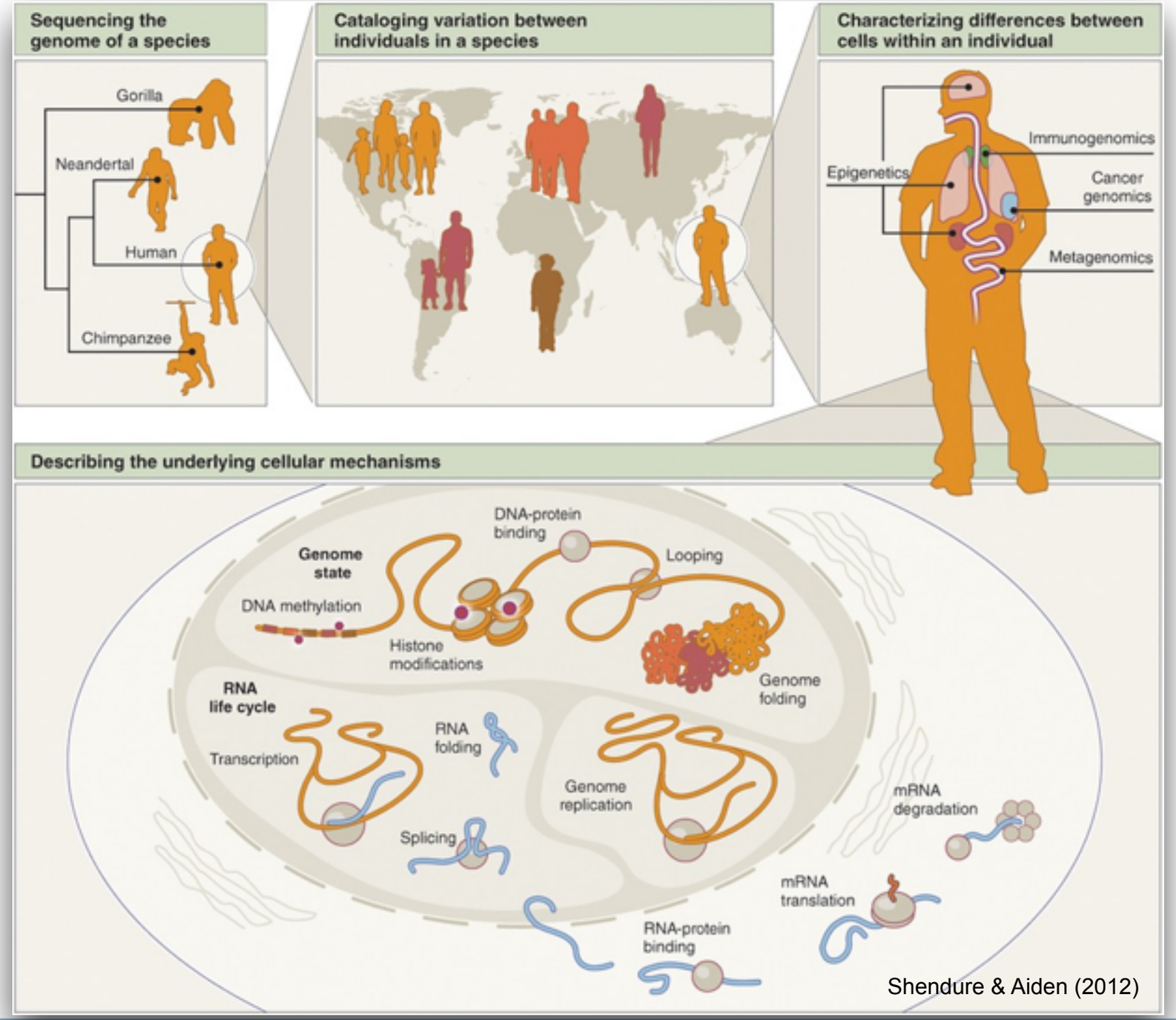
Genomics England, with the consent of participants and the support of the public, is creating a lasting legacy for patients, the NHS and the UK economy through the sequencing of 100,000 genomes: the 100,000 Genomes Project.

Genomics England was set up by the Department of Health to deliver the 100,000 Genomes Project. Initially the focus will be on rare disease, cancer and infectious disease.

Read more...

http://www.genomicsengland.co.uk/

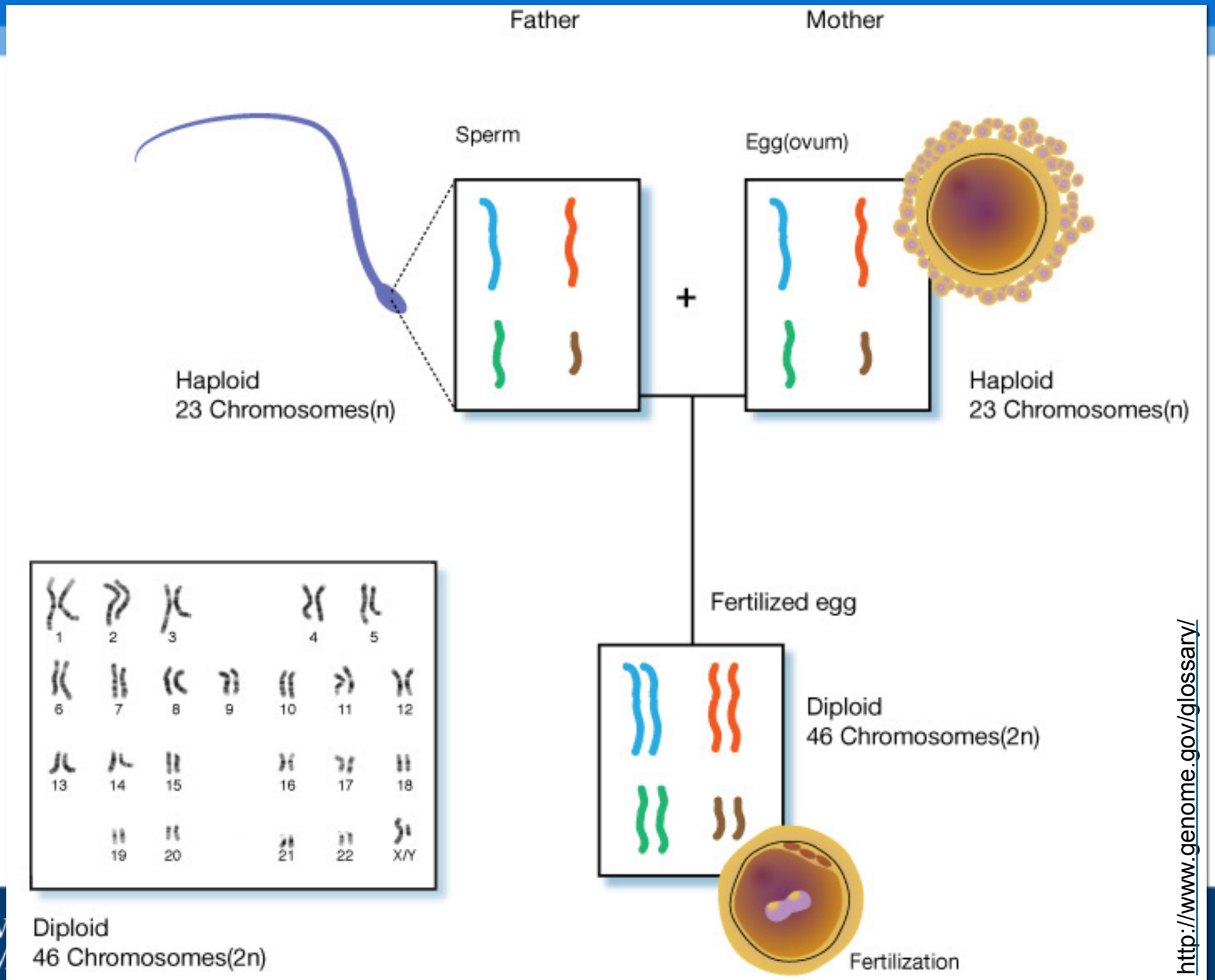
A Roadmap of Sequencing Science



Shendure & Aiden (2012)

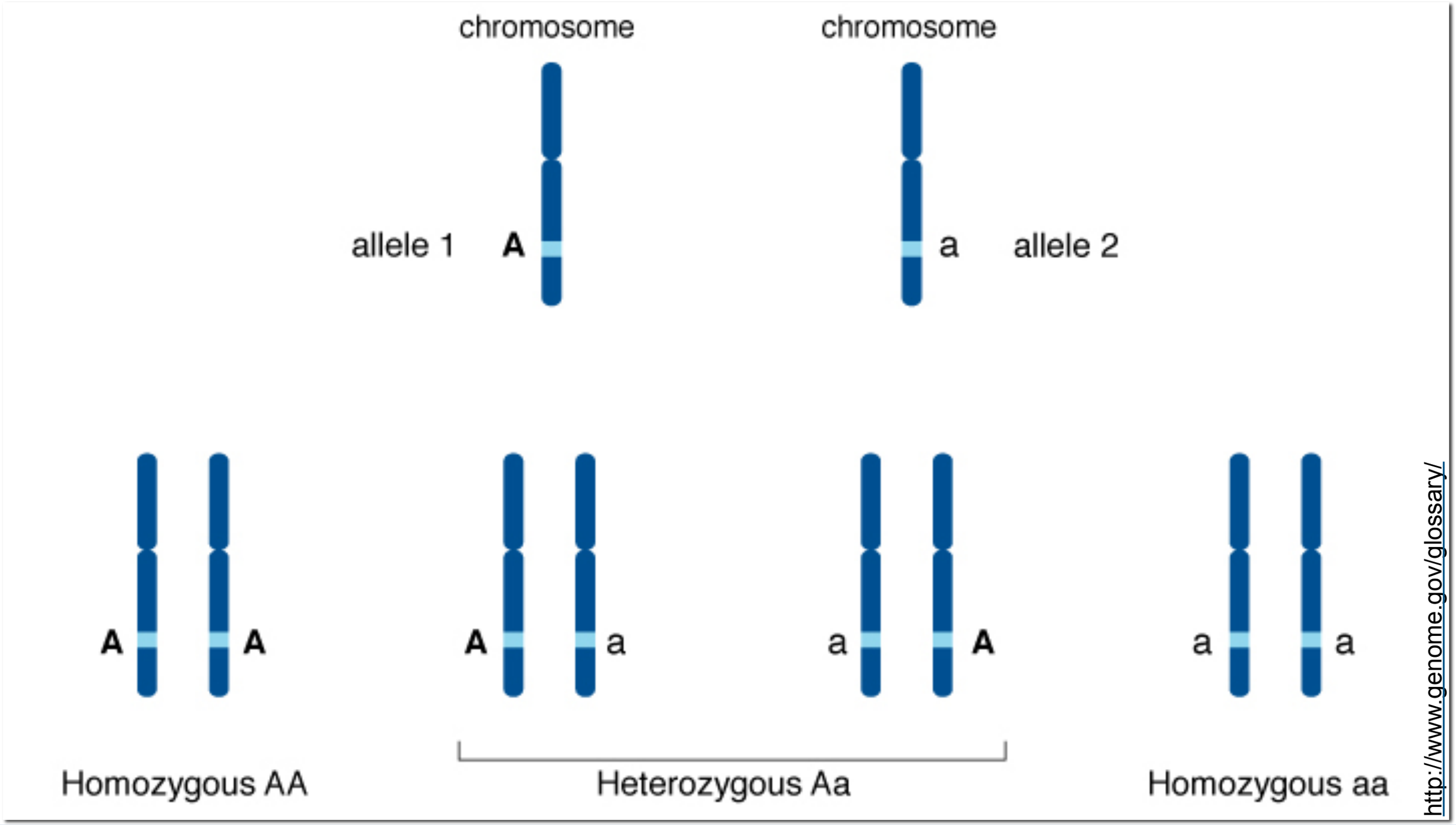
Human Genetics

Human Genetics



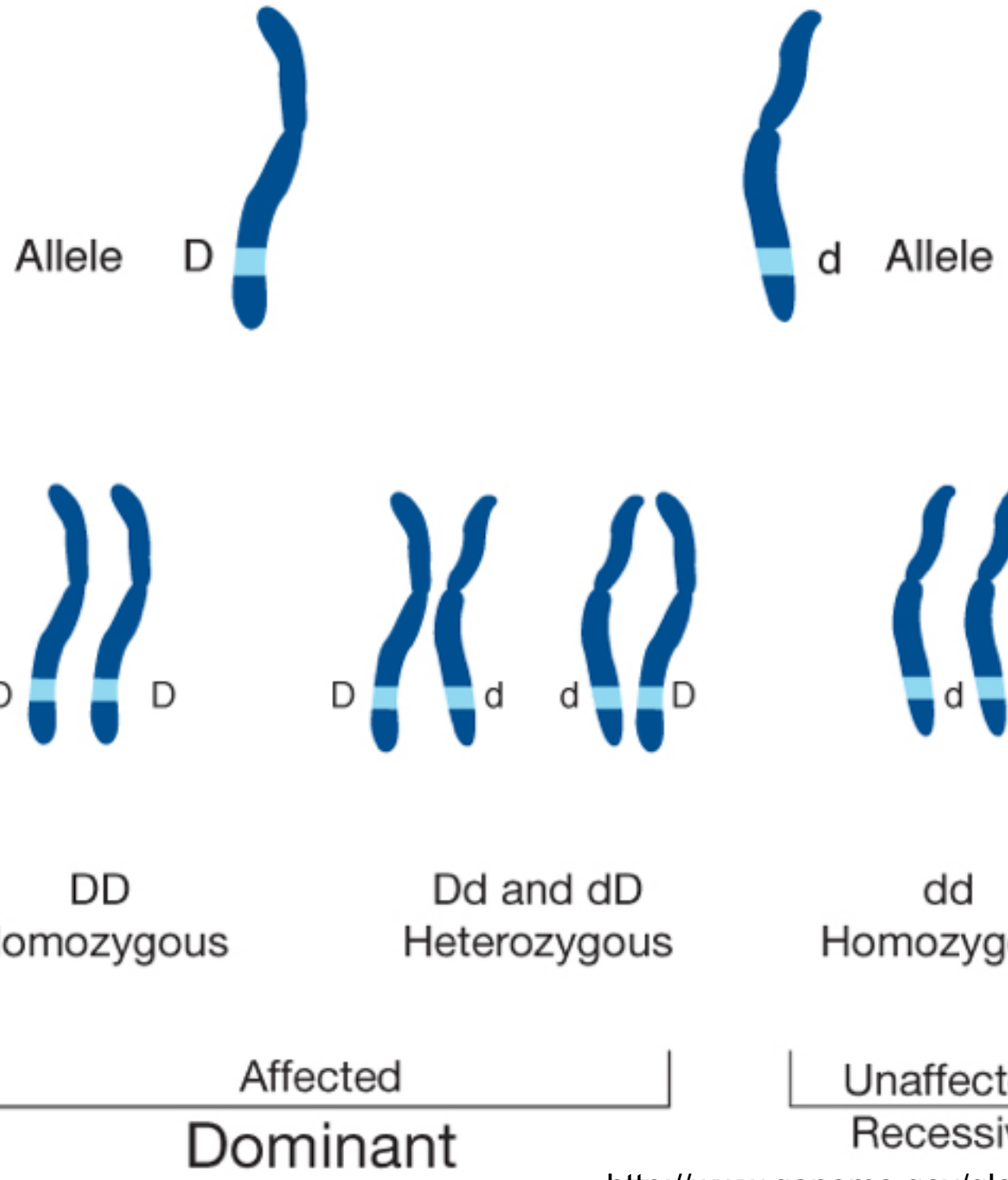
http://www.genome.gov/glossary/

Human Genetics



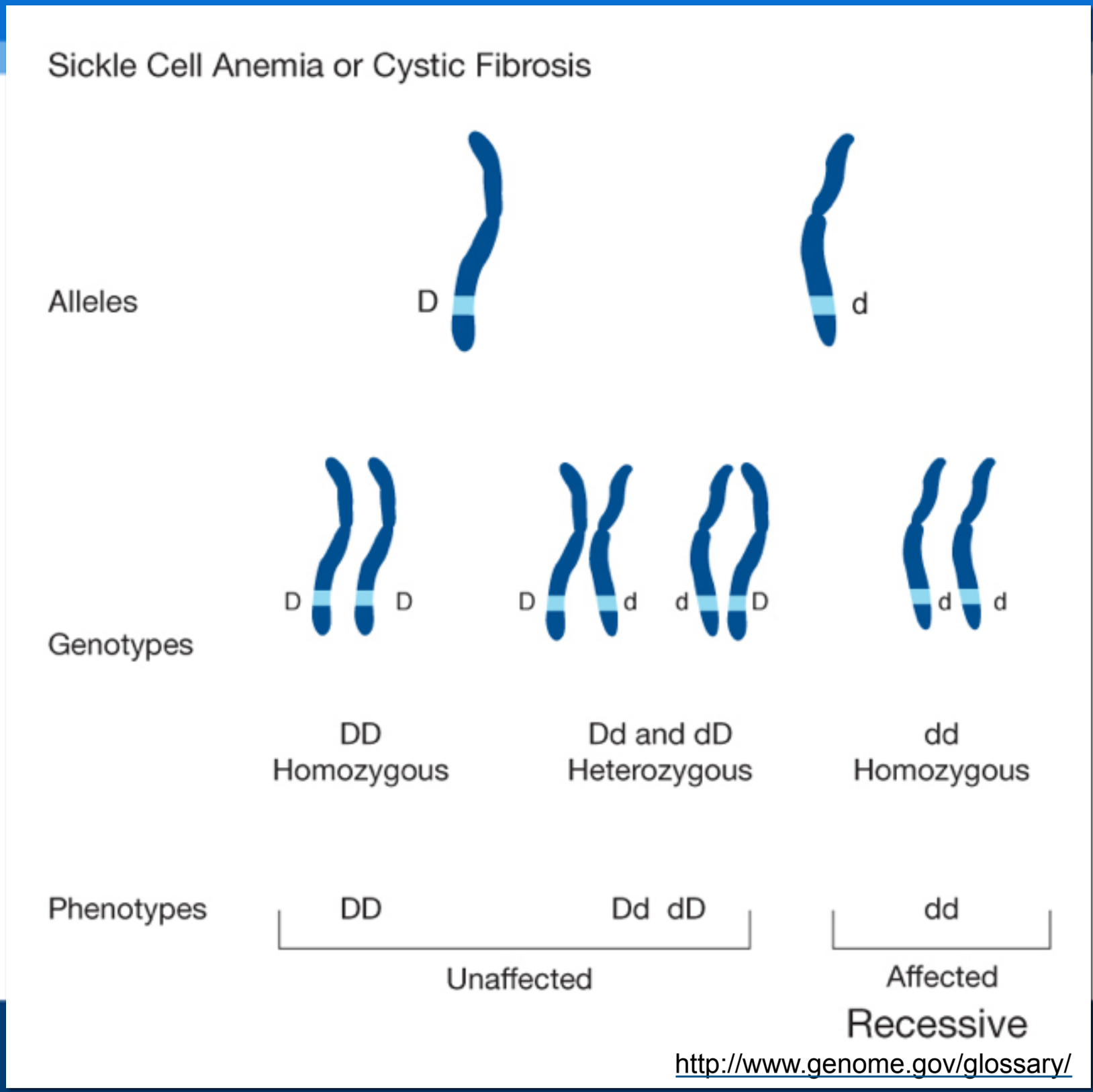
Human Genetics

Huntington's Disease



http://www.genome.gov/glossary/

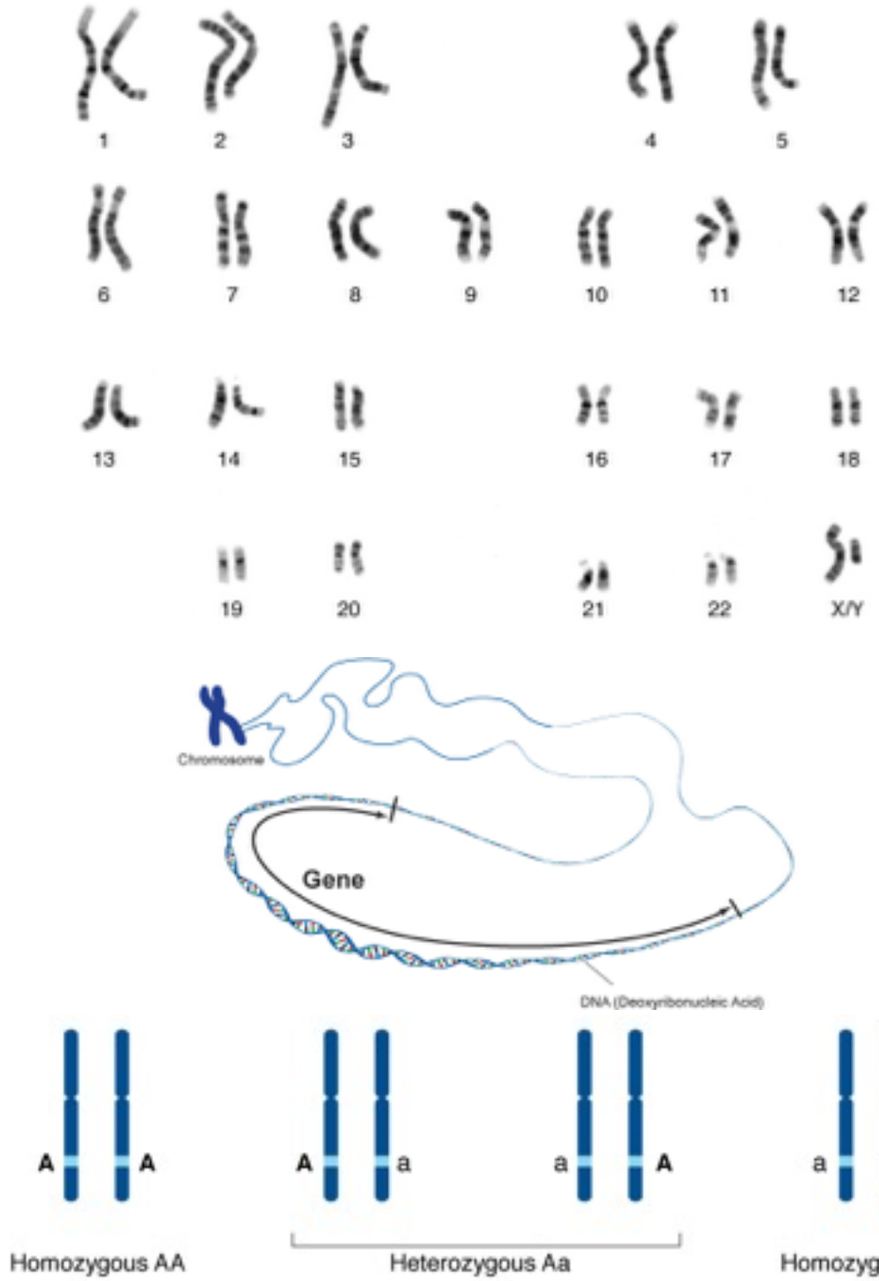
Human Genetics



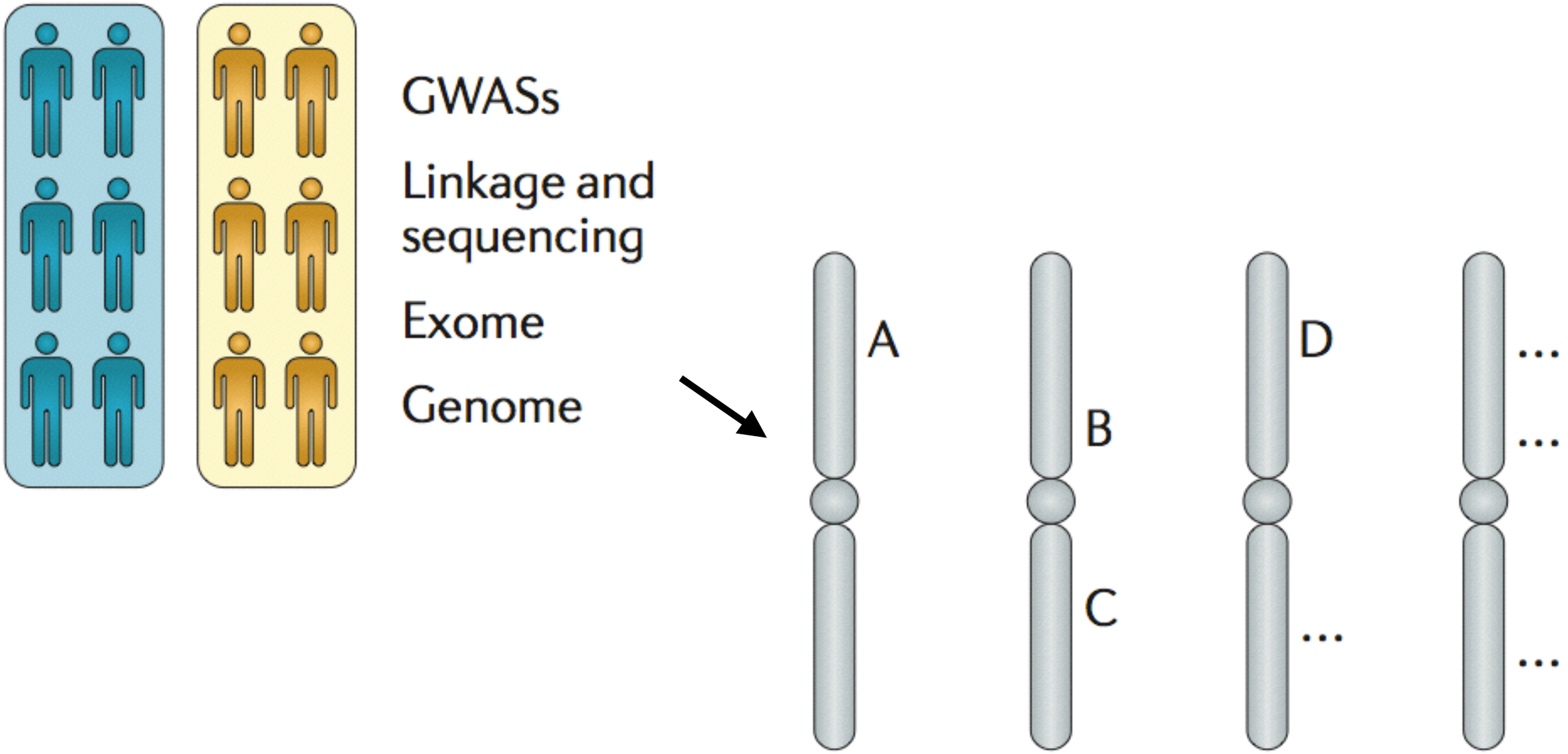
http://www.genome.gov/glossary/

Human Genetics

- Two sets of chromosomes (**diploid**), one from each parent
- Two alternative copies (**alleles**) of each gene
- Alleles can be
  - identical (**homozygous**) or
  - dissimilar (**heterozygous**)
- Only one allele (**dominant**), or both alleles (**recessive**) need to be mutated to be causative
- Genetic configuration (**genotype**) varies amongst individuals and populations
- Results in a observable trait (**phenotype**)

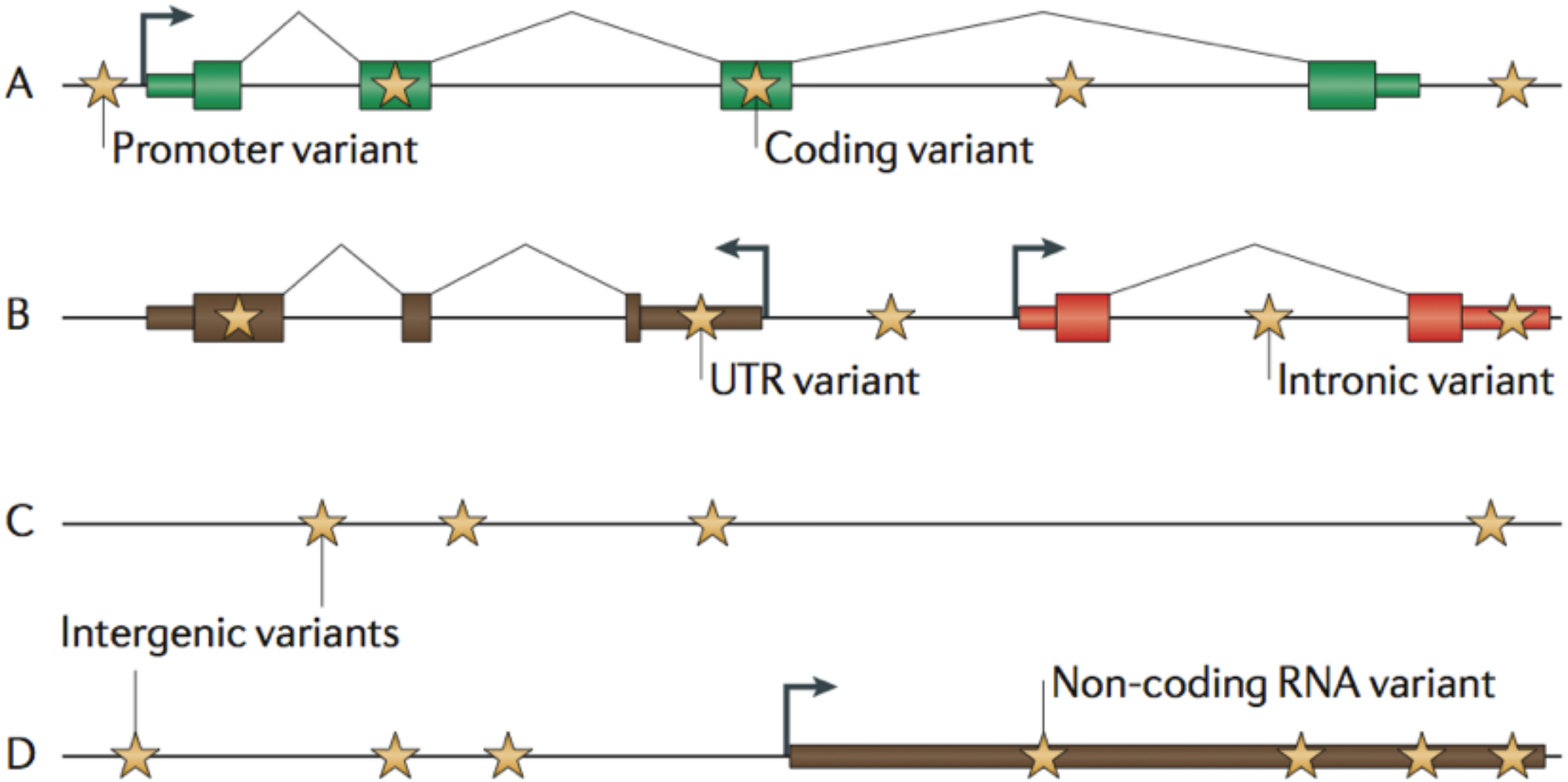


Identification of Genetic Variation



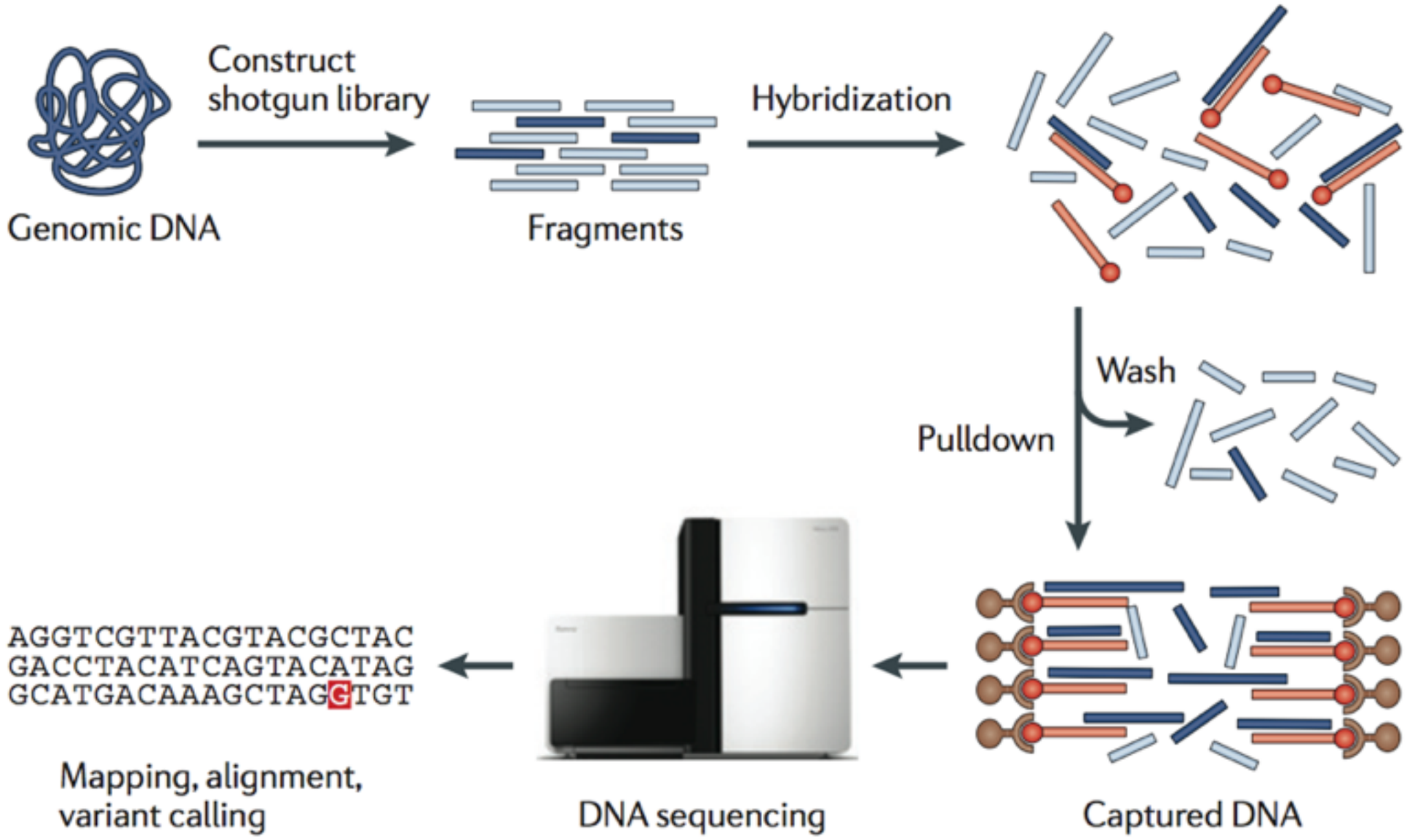
Cooper *et al.*, 2011

Genetic Variants have Different Functional Consequences



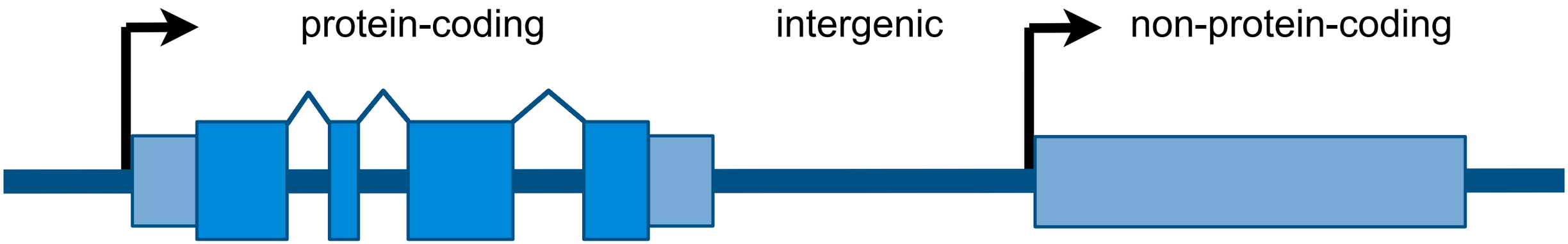
Cooper *et al.*, 2011

Exome-sequencing Interrogates the Protein-coding Portion of the Genome



Bamshad *et al.*, 2011

Whole Exome vs. Whole Genome



Regions covered by WES (64 MB, 2%)

Regions covered by WGS (~3000 MB, 98%)

Consequences

Raised demands for resources

- Storage (talking peta (1015) bytes)
- Computation

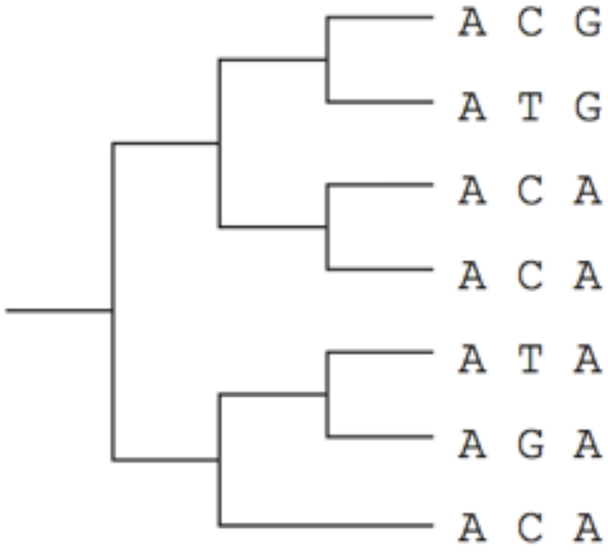
Data security

Sample requirements

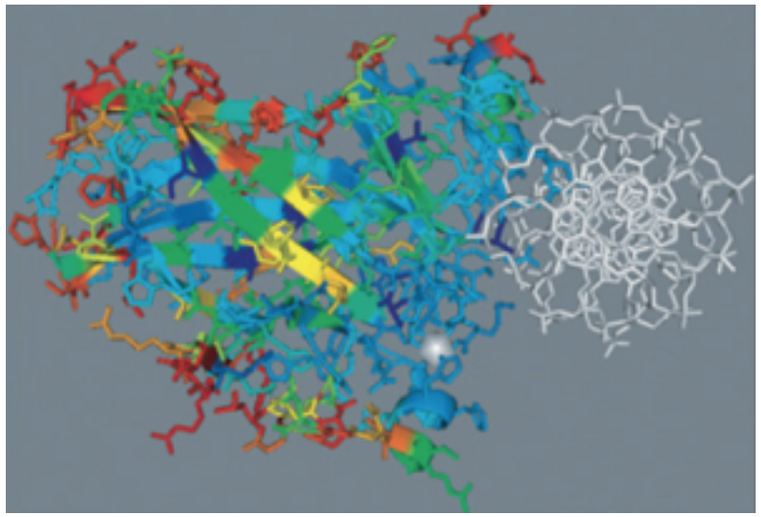

Darwin - University of Cambridge
High-Performance Computing (HPC)

Teasing out Disease-causing Variants

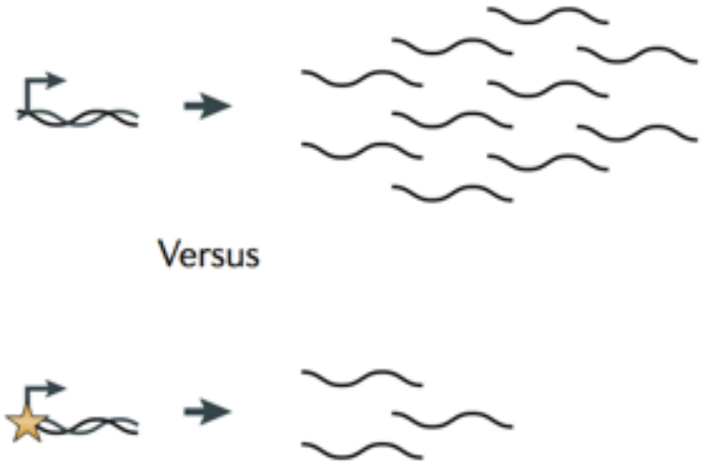
Long list of candidate variants



Comparative Genomics



Protein Structure / Biochemistry



Experimental Assay

Cooper *et al.*, 2011

Assessing Deleteriousness

Name	Category	Score used for analysis	Deleterious threshold	Information used
SIFT	Function prediction	1 − Score	>0.95	Protein sequence conservation among homologs
PolyPhen-2	Function prediction	Score	>0.5	Eight protein sequence features, three protein structure features
LRT	Function prediction	Score * 0.5 (if Omega ≥1) or 1 − Score * 0.5 (if Omega <1)	P	DNA sequence evolutionary model
MutationTaster	Function prediction	Score (if A or D) or 1 − Score (if N or P)	>0.5	DNA sequence conservation, splice site prediction, mRNA stability prediction and protein feature annotations
Mutation Assessor	Function prediction	(Score-Min)/(Max − Min)	>0.65	Sequence homology of protein families and sub-families within and between species
FATHMM	Function prediction	1 − (Score-Min)/(Max − Min)	≥0.45	Sequence homology
GERP++ RS	Conservation score	Score	>4.4	DNA sequence conservation
PhyloP	Conservation score	Score	>1.6	DNA sequence conservation
SiPhy	Conservation score	Score	>12.17	Inferred nucleotide substitution pattern per site
PON-P	Ensemble score	Score	P	Random forest methodology-based pipeline integrating five predictors
PANTHER	Function prediction	Score	P	Phylogenetic trees based on protein sequences
PhD-SNP	Function prediction	Score	P	SVM-based method using protein sequence and profile information
SNAP	Function prediction	Score	P	Neural network-based method using DNA sequence information as well as functional and structural annotations
SNPs&GO	Function prediction	Score	P	SVM-based method using information from protein sequence, protein sequence profile and protein function
MutPred	Function prediction	Score	>0.5	Protein sequence-based model using SIFT and a gain/loss of 14 different structural and functional properties
KGGSeq	Ensemble score	Score	P	Filtration and prioritization framework using information from three levels: genetic level, variant-gene level and knowledge level
CONDEL	Ensemble score	Score	>0.49	Weighted average of the normalized scores of five methods
CADD	Ensemble score	Score	>15	63 distinct variant annotation retrieved from Ensembl Variant Effect Predictor (VEP), data from the ENCODE project and information from UCSC genome browser tracks

Dong *et al.*, 2015

Experimental Design

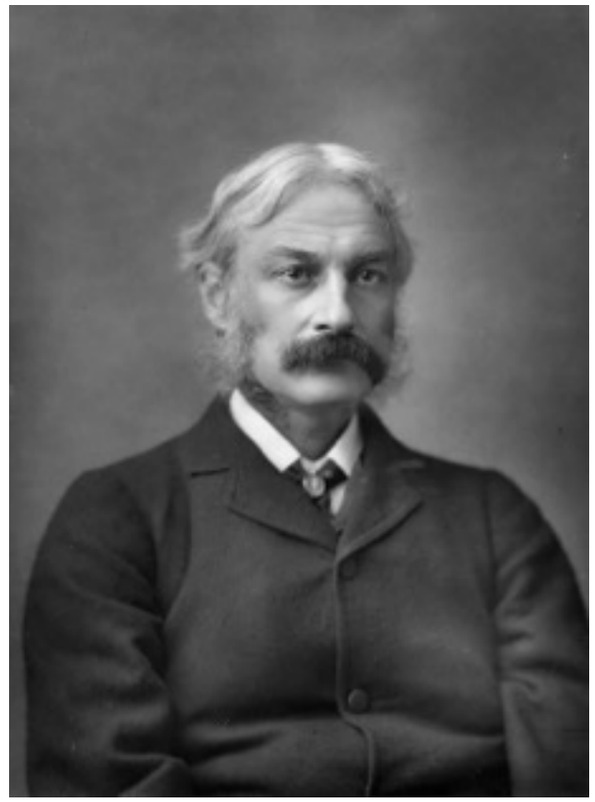
Sir Ronald A. Fisher

"To consult the statistician after an experiment is finished is often merely to ask him to conduct a post mortem examination. He can perhaps say what the experiment died of."

(1890 – 1962)
Evolutionary biologist, geneticist and statistician

Andrew Lang

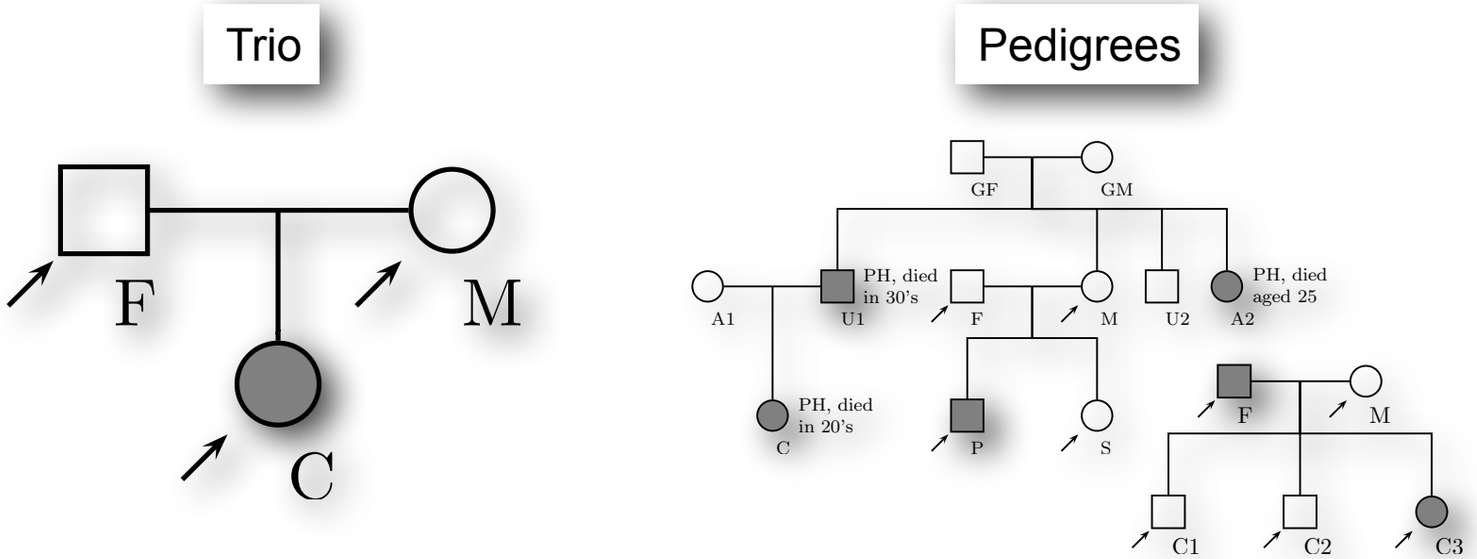
"An unsophisticated forecaster uses statistics as a drunken man uses lamp-posts - for support rather than for illumination."



(1844 — 1912)
Writer (poet, novelist),
literary critic and anthropologist

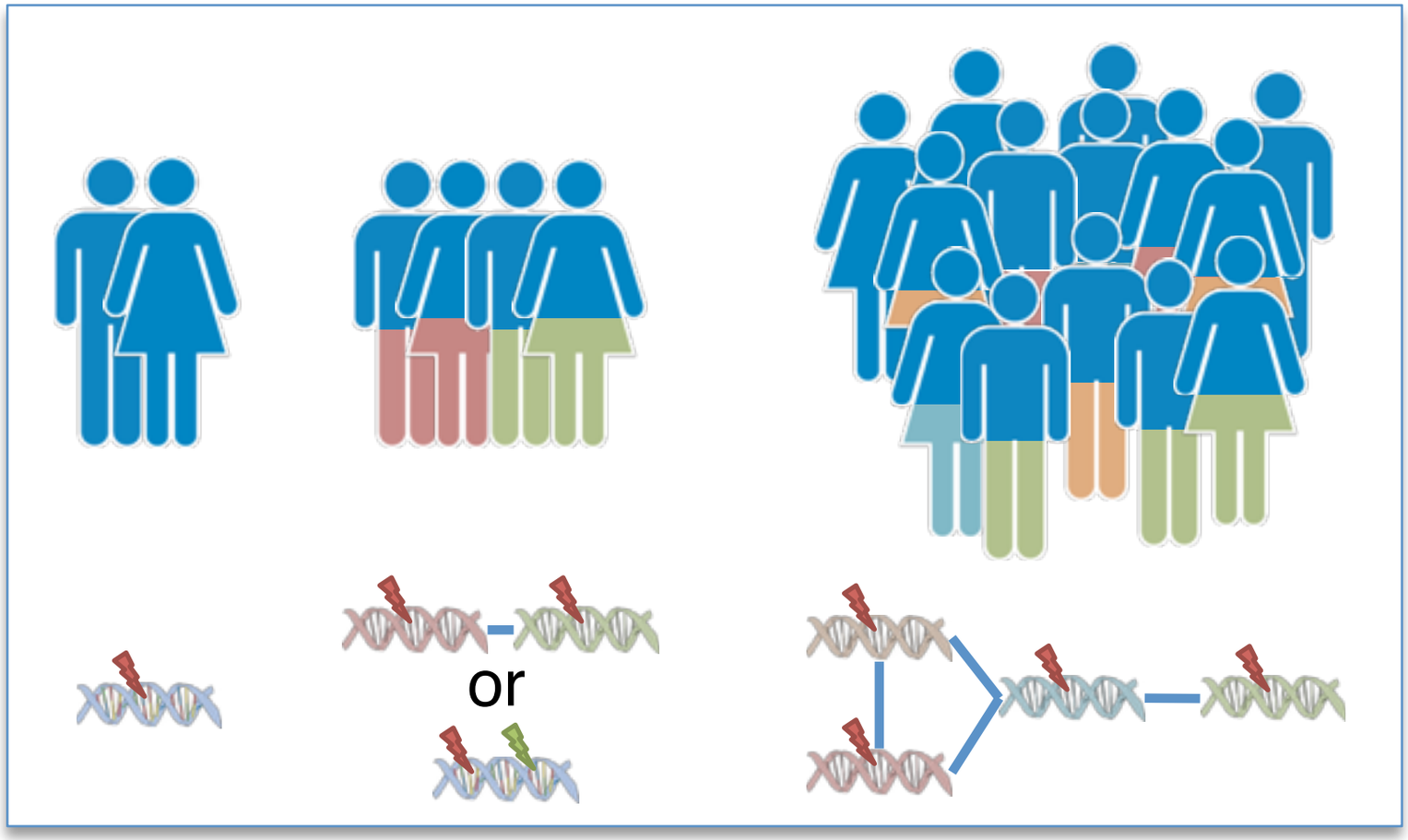
Variant Discovery Strategies and Sample Selection

- Select study design to achieve adequate statistical power (i.e. trios for de novo mutations, pedigree analysis, cohort of multiple unrelated patients)



- Focus on cases with extreme outcome
- Population stratification important for rare variant detection

Genetic and Phenotypic Heterogeneity Reduces Power



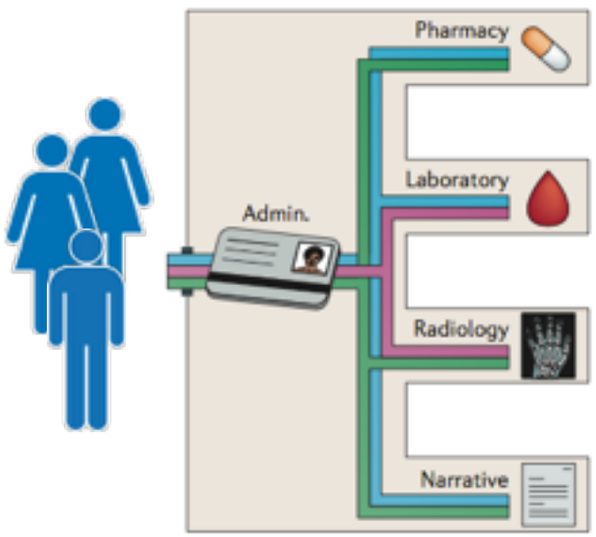
Number of samples to achieve 80% power

% Carriers	100	50	5
Recessive	4	9	170
Dominant	6	20	1100

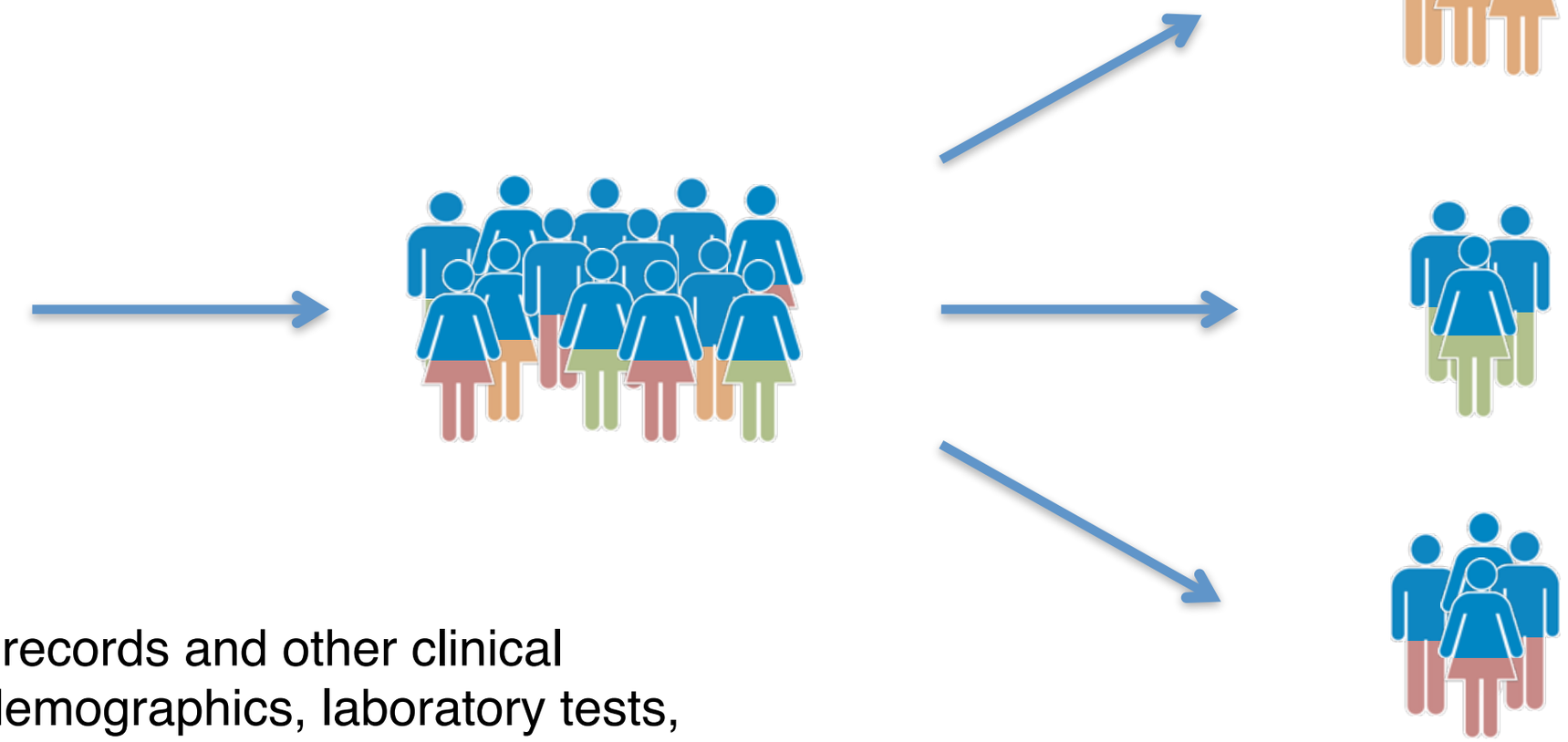
http://exomepower.ssg.uab.edu

Phenotype-based Clustering Can Restore Power

(Deep) Phenotyping



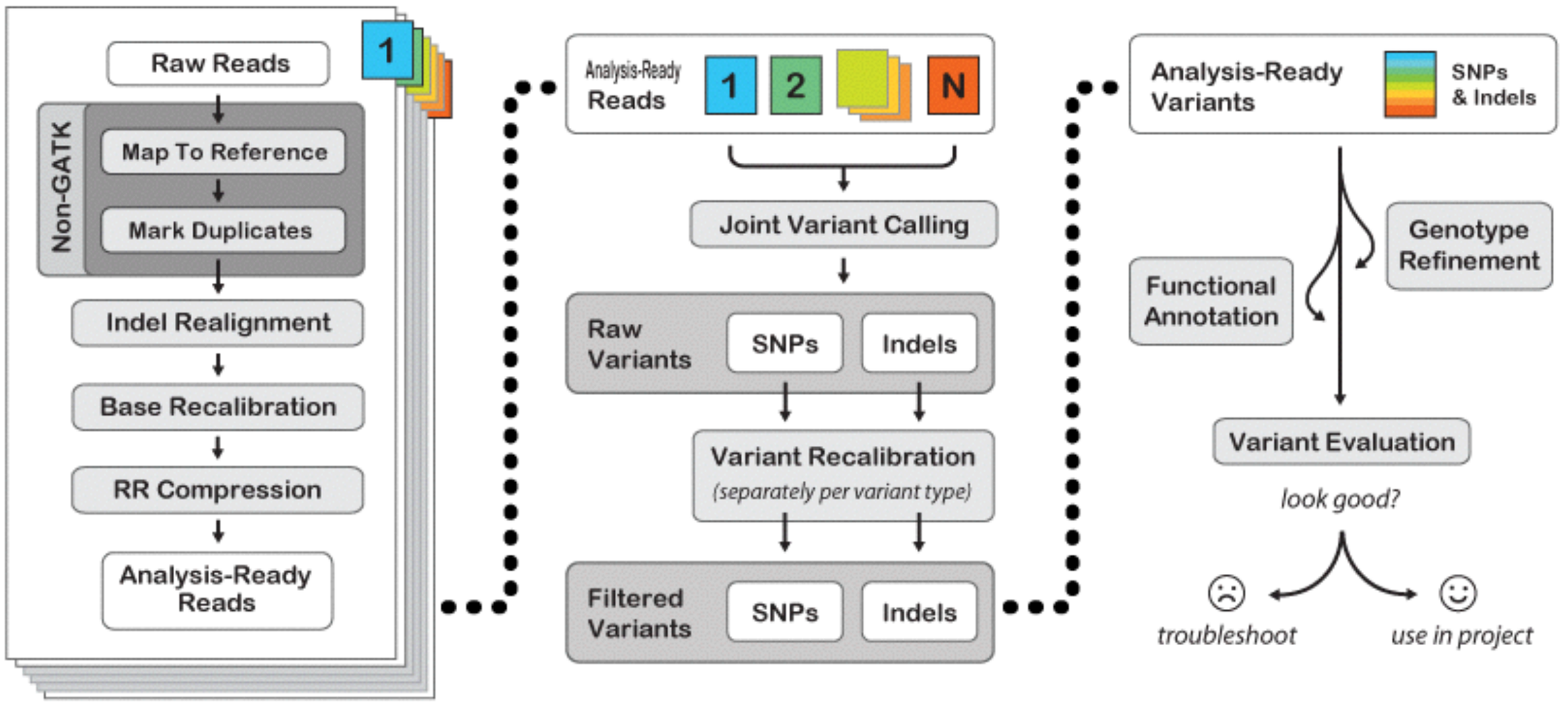
Clustering



Electronic health records and other clinical information (i.e. demographics, laboratory tests, human phenotype ontology (HPO), imaging, etc.)

Analysis Pipeline Overview

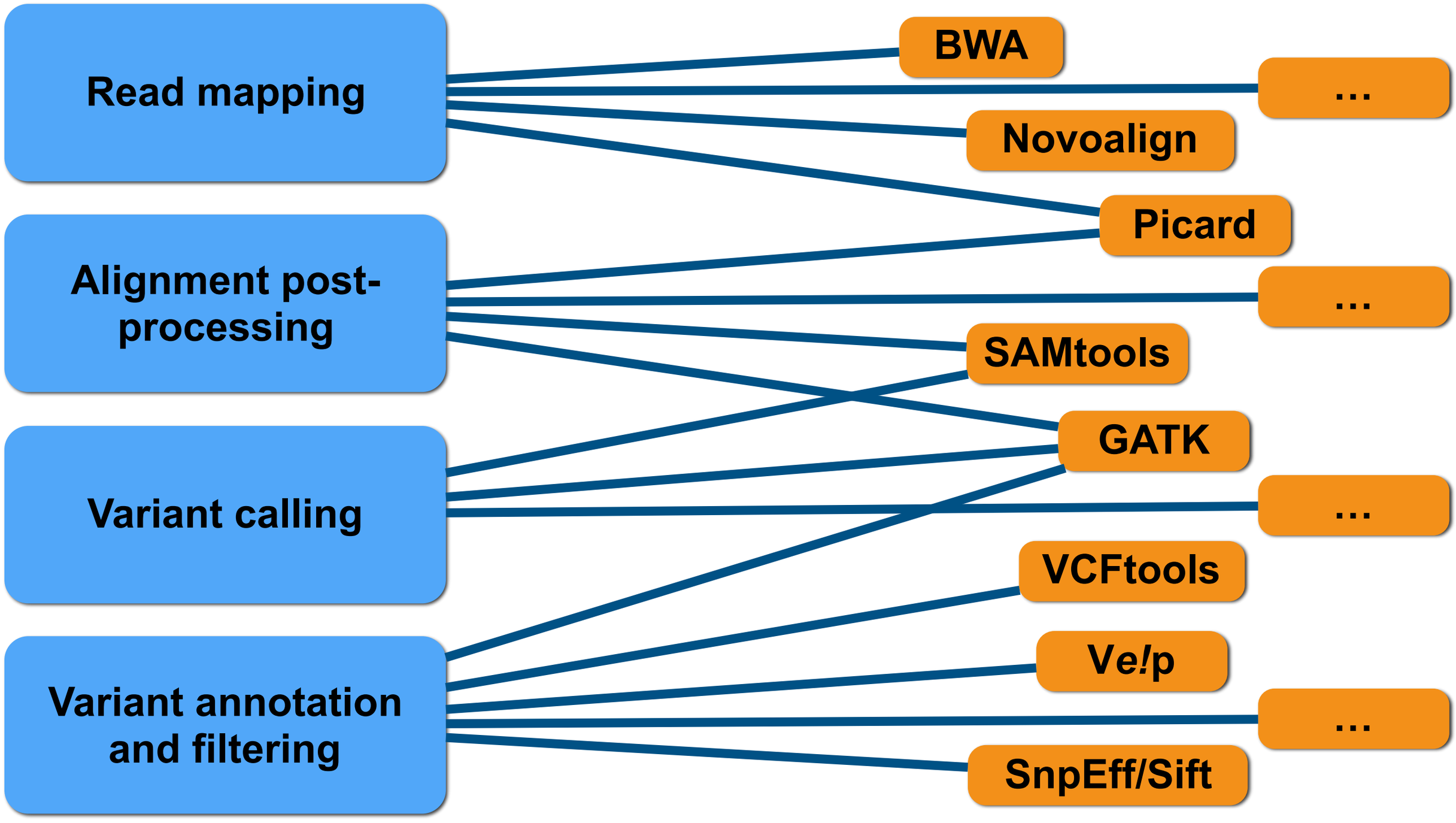
GATK's Best Practises



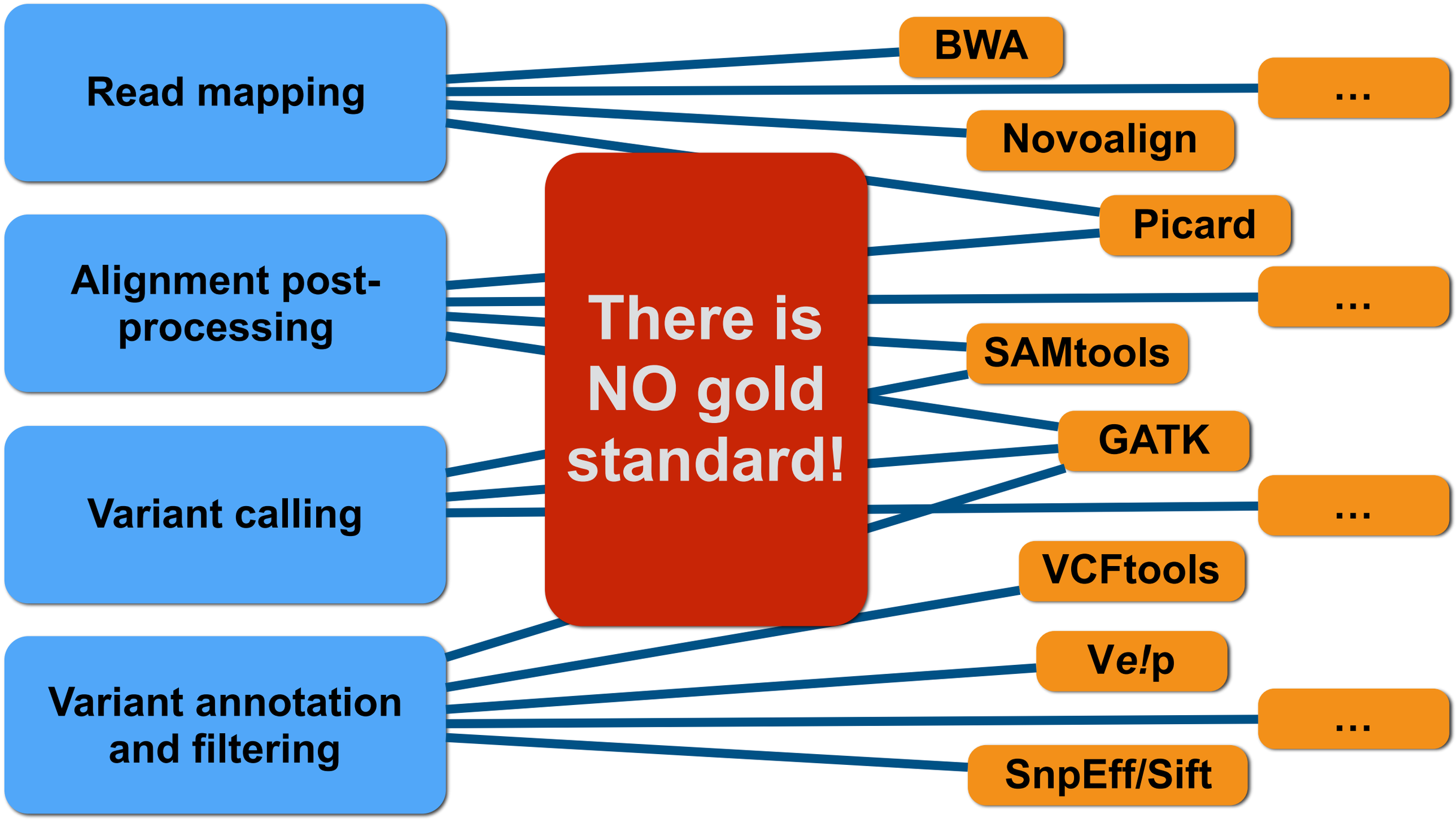
gatk

http://www.broadinstitute.org/gatk/

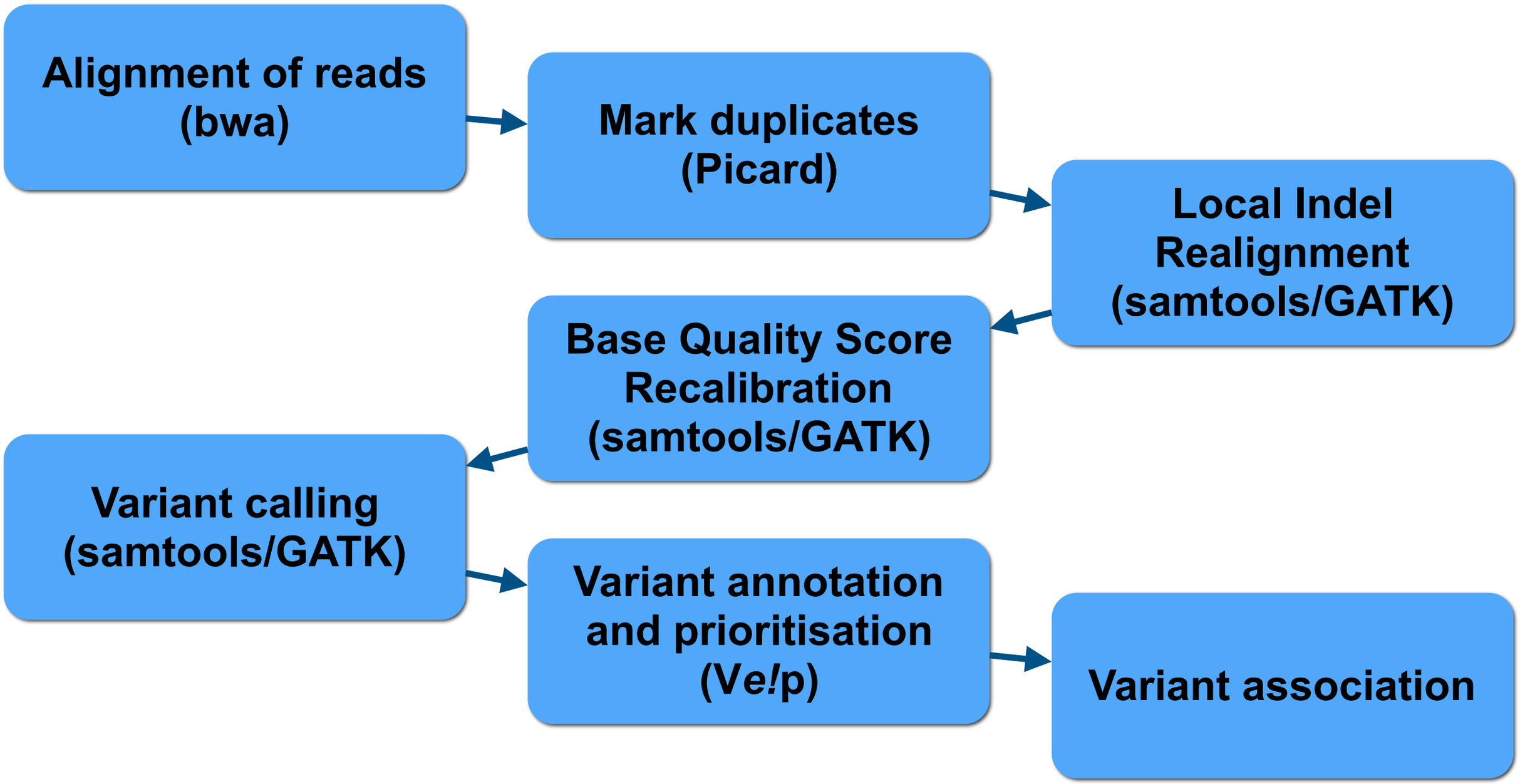
Analysis Pipeline Tasks and Tools



Analysis Pipeline Tasks and Tools



NGS-Course Analysis Pipeline



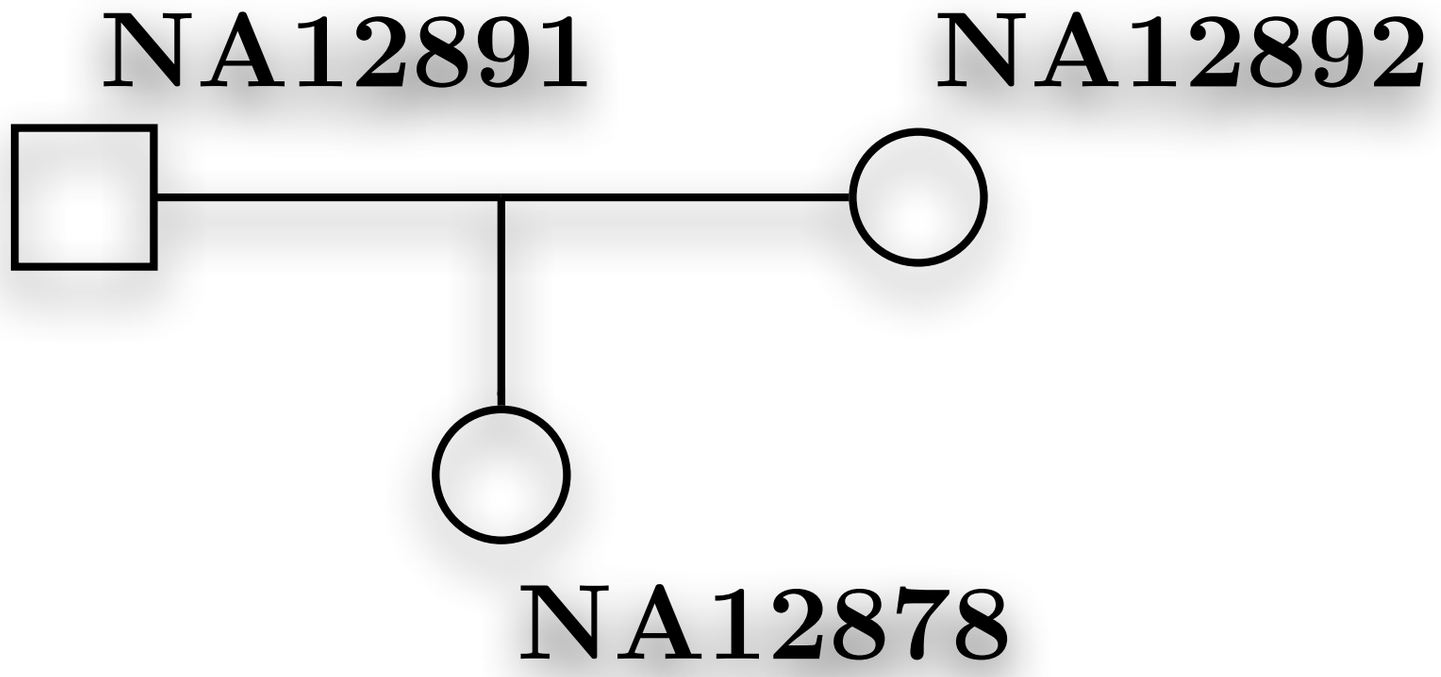
Data Formats

File extensions:

- `.fa` — reference sequence (fasta), i.e. GRCh37_chr19.fa
- `.fastq` — raw sequencing reads, i.e. NA12878_1.fq.gz
- `.sam` — aligned sequencing reads, i.e. NA12878.sam
- `.bam` — aligned reads (binary), i.e. NA12891.bam
- `.vcf` — called variants, i.e. trio_mpileup.vcf
- `.tbi` — files indexed with tabix
- `.gz` — compressed files

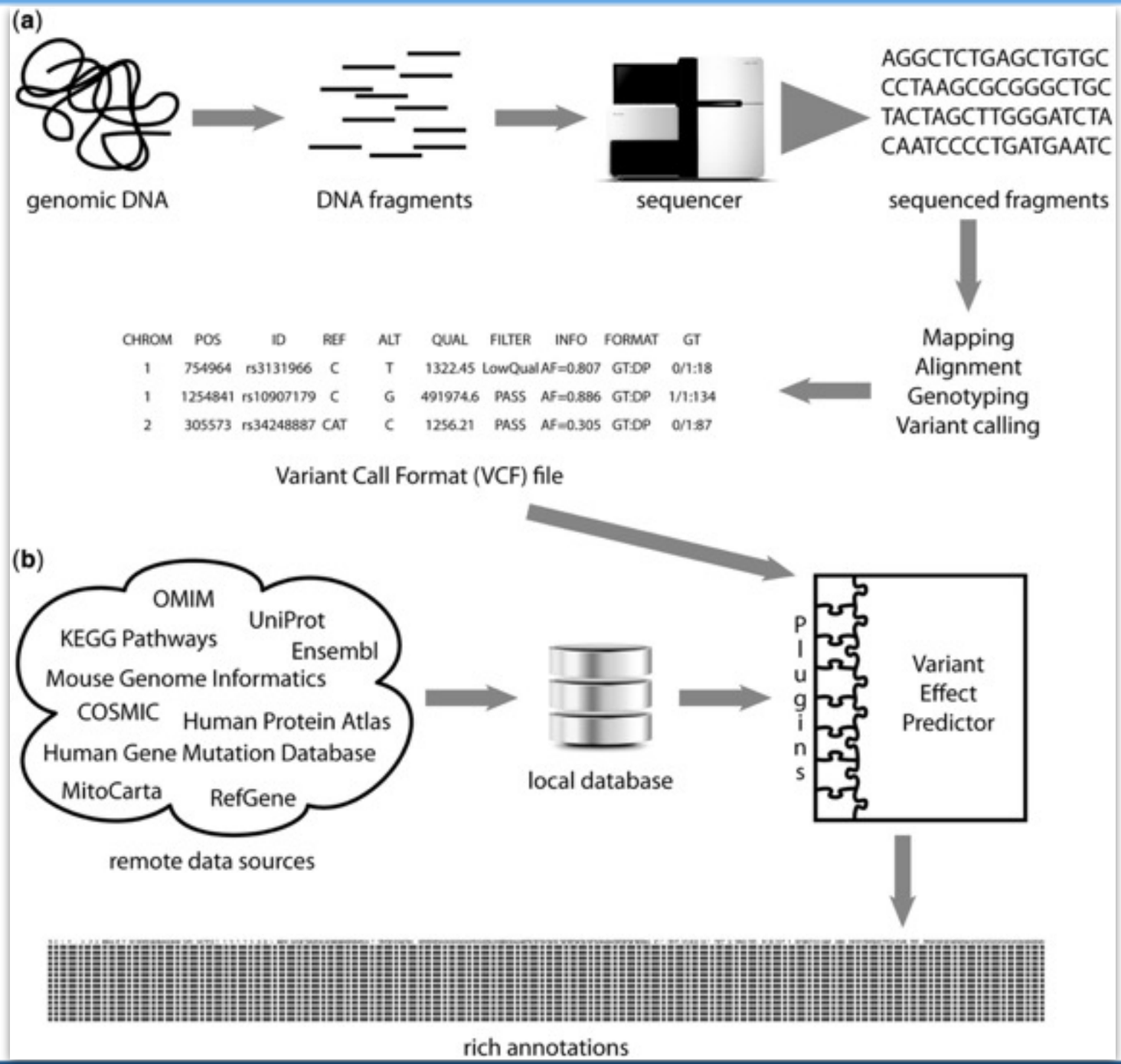
NGS-Course Data

Offspring trio of central european ancestry



Variant Annotation, Filtering and Prioritisation

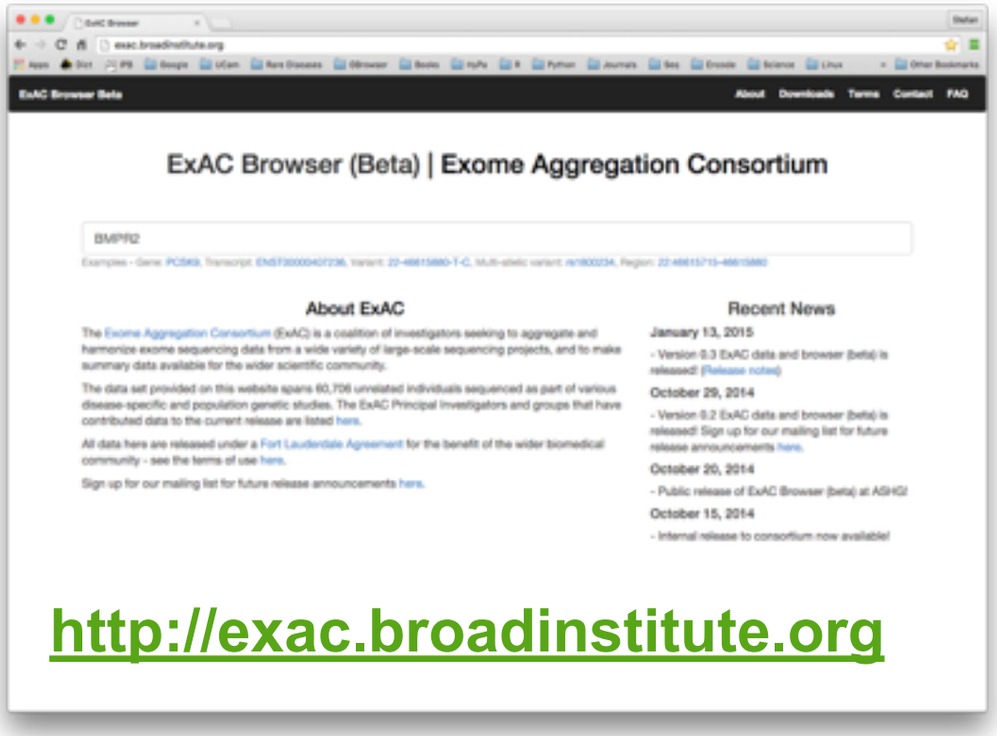
Variant Annotation and Effect Prediction



Yourshaw *et al.*, Brief Bioinform (2015)

Exome Aggregation Consortium (ExAC)

- Aggregation of high-quality exome (protein-coding region) sequence data for **60,706 individuals** of diverse ethnicities
- Resolution of **one variant every eight bases** of coding sequence
- Allows calculation of objective **metrics of pathogenicity** for sequence variants
- Can be used for **efficient filtering** of candidate disease-causing variants



http://exac.broadinstitute.org

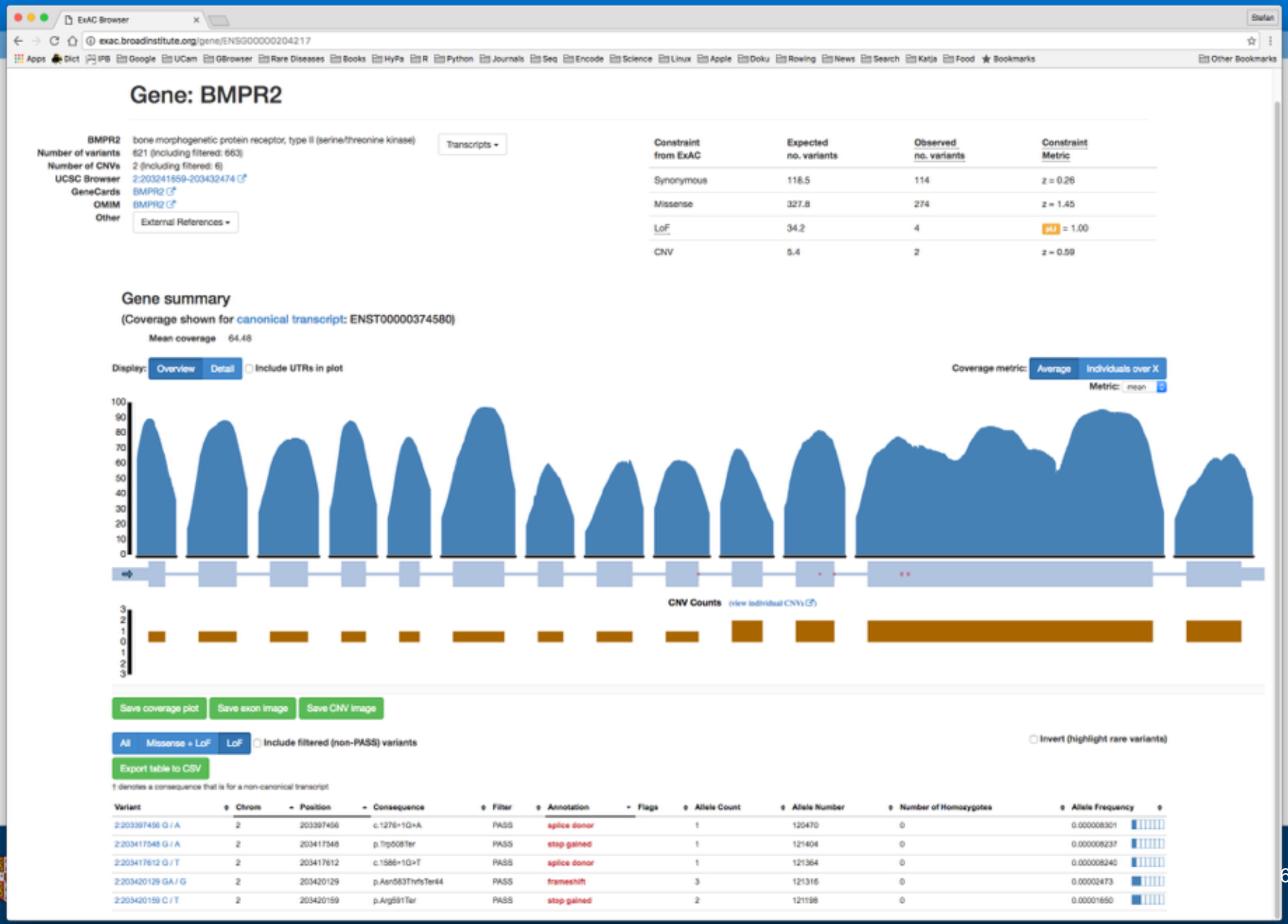
Contributing projects

- 1000 Genomes
- Bulgarian Trios
- Finland-United States Investigation of NIDDM Genetics (FUSION)
- GoT2D
- Inflammatory Bowel Disease
- METabolic Syndrome In Men (METSIM)
- Jackson Heart Study
- Myocardial Infarction Genetics Consortium:
  - Italian Atherosclerosis, Thrombosis, and Vascular Biology Working Group
  - Ottawa Genomics Heart Study
  - Pakistan Risk of Myocardial Infarction Study (PROMIS)
  - Precocious Coronary Artery Disease Study (PROCARDIS)
  - Registre Gironi del COR (REGICOR)
- NHLBI-GO Exome Sequencing Project (ESP), *incl. 96 PAH cases*
- National Institute of Mental Health (NIMH) Controls
- SIGMA-T2D
- Sequencing in Suomi (SISu)
- Swedish Schizophrenia & Bipolar Studies
- T2D-GENES
- Schizophrenia Trios from Taiwan
- The Cancer Genome Atlas (TCGA)
- Tourette Syndrome Association International Consortium for Genomics (TSAICG)

http://biorxiv.org/content/early/2015/10/30/030338

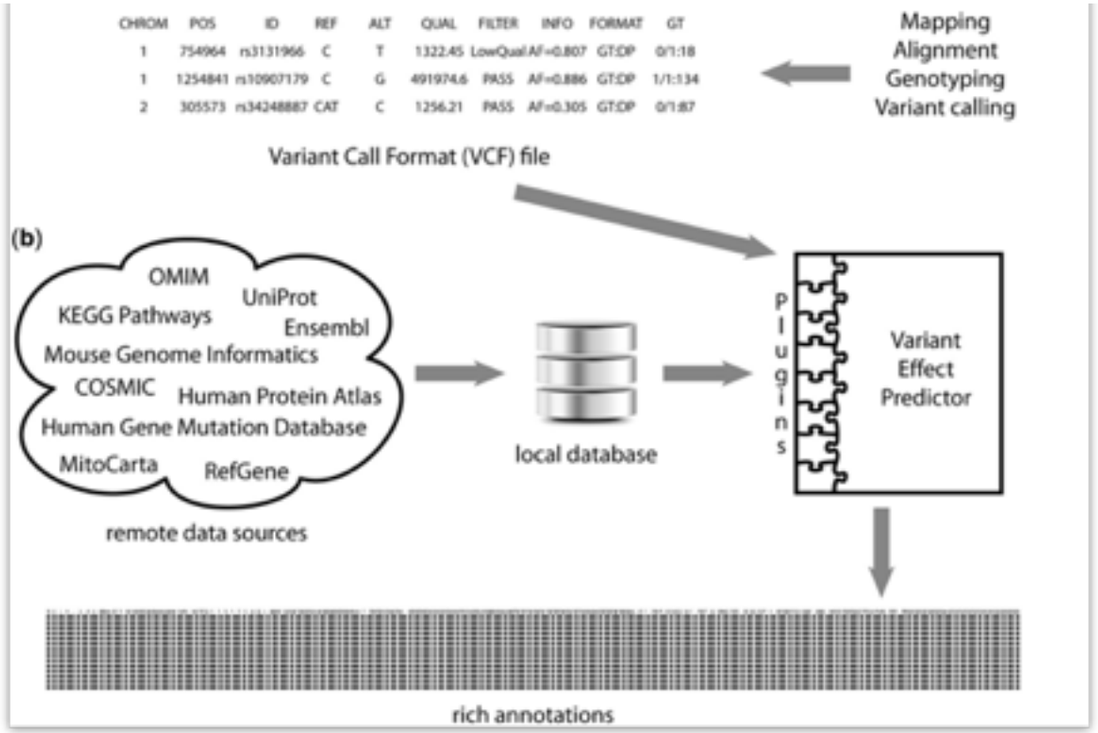
Exome Aggregation Consortium (ExAC)

Exome Aggregation Consortium (ExAC)



Variant Annotation and Consequence Prediction

- **Deleteriousness scores**
  - **SIFT:** functional prediction, protein sequence conservation among homologs; score: 1 (tolerated) - 0 (deleterious)
  - **PolyPhen:** functional prediction, protein sequence and structure features; score: 0 (benign) - 1 (damaging)
  - **CADD:** ensemble score, combines 63 distinct variant annotation features retrieved from Ensembl VEP, Encode, UCSC genome browser; Phred score (i.e. 30 = 99.9% accurate or 1 in 1000 is incorrect)



Yourshaw *et al.,* Brief Bioinform (2015), adapted

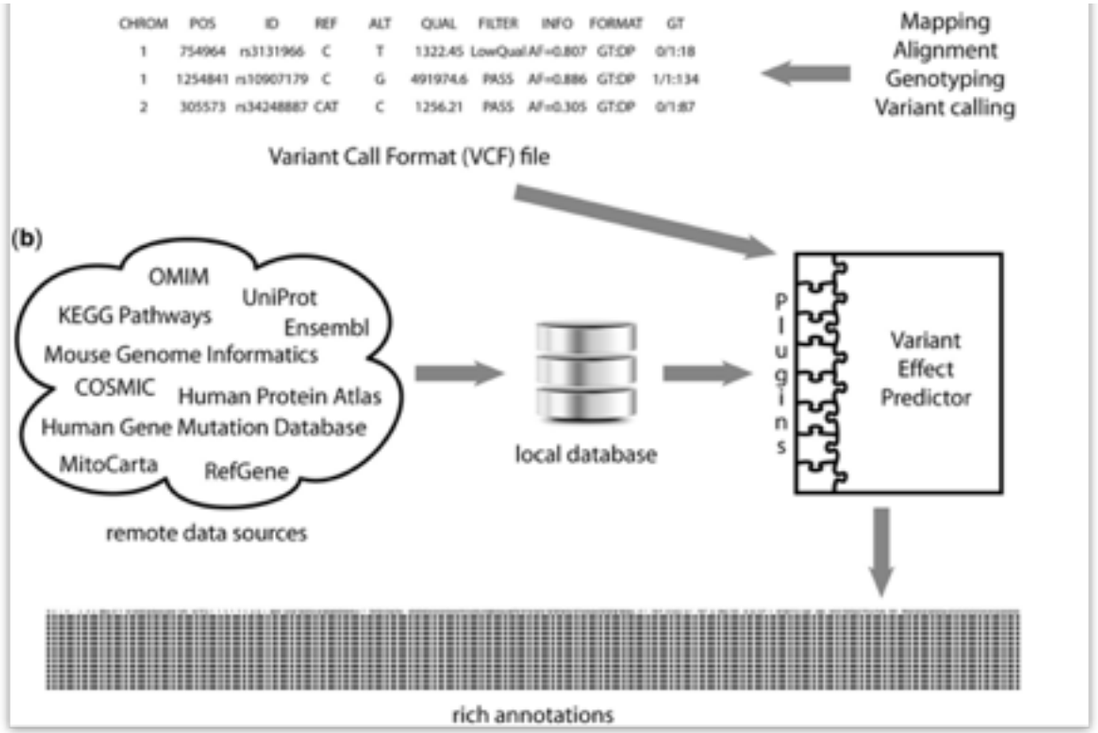
- **DNA sequence conservation scores**
  - **GERP:** maximum likelihood evolutionary rate estimation, predicts sites under evolutionary constraints
  - **PhyloP:** base-wise conservation score derived from Multiz alignment of 100 vertebrate species
  - **PhastCons:** evolutionary conserved elements derived from Multiz alignment of 100 vertebrate species (phylogenetic hidden Markov model)

Variant Annotation and Consequence Prediction

- **Deleteriousness scores**
  - **SIFT:** functional prediction, protein sequence conservation among homologs; score: 1 (tolerat
  - **PolyPhen:** func sequence and s score: 0 (benign) - 1 (damaging)
  - **CADD:** ensemble score, combines 63 distinct variant a Ensembl Phred sc is incorrect)

protein sequence and structure based prediction

score base on various informative genome-wide annotations



Yourshaw *et al.,* Brief Bioinform (2015), adapted

- **DNA sequence conservation scores**
  - **GERP:** maximum likelihood evolutionary rate estimation, predicts sites under evolutionary constraints
  - **PhyloP:** base-wise z alignment of 100 vertebrate species
  - **PhastCons:** evolutic m Multiz alignment of 100 vertebrate species (phylogenetic hidden Markov model)

measures DNA sequence conservation

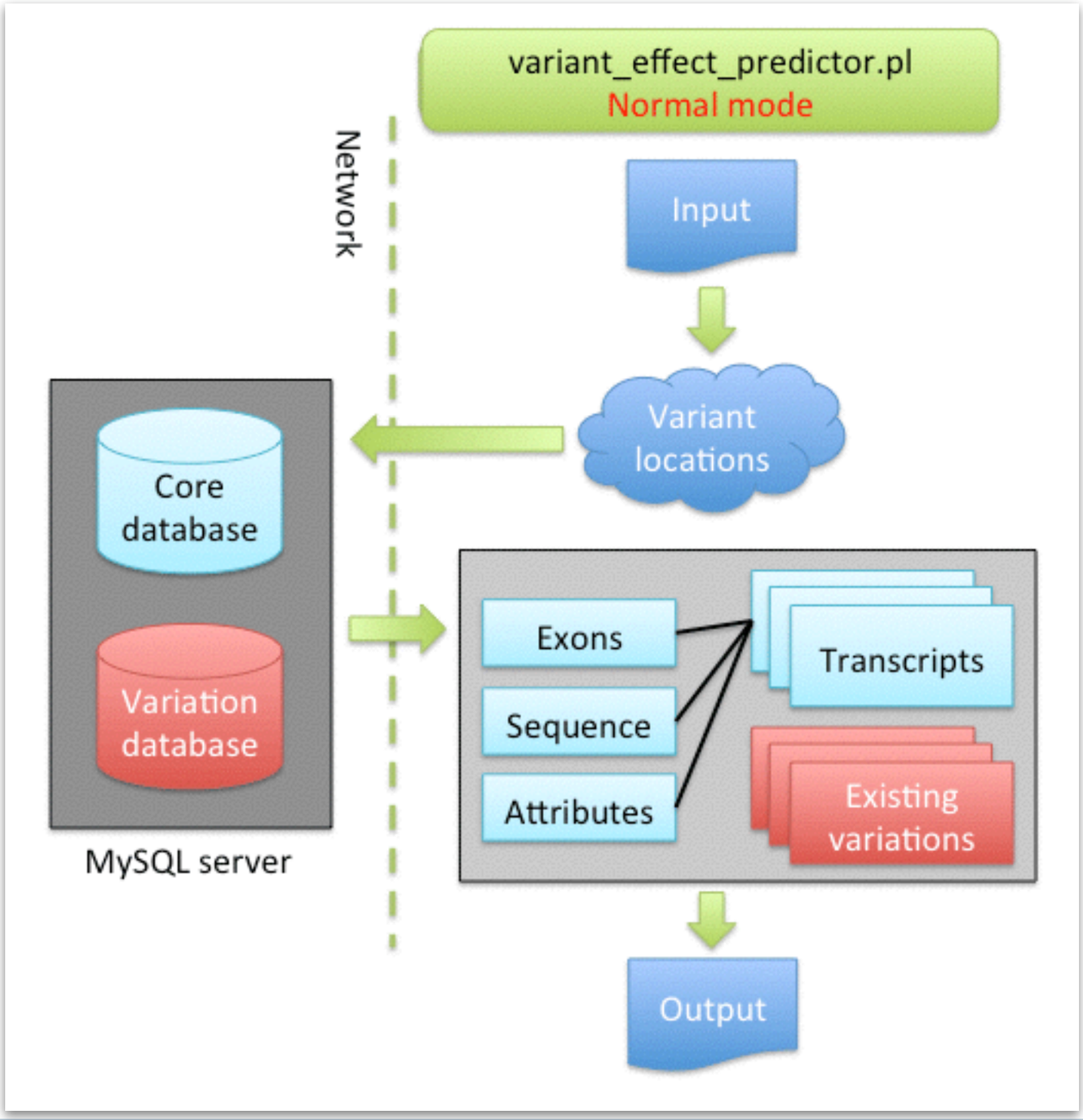
Variant Annotation Tools

- Ensembl Variant Effect Predictor (VEP)
  — http://www.ensembl.org/info/docs/tools/vep/index.html
- SnpEff / SnpSift
  — http://snpeff.sourceforge.net/
- AnnoVar
  — http://annovar.openbioinformatics.org/en/latest/

- Rich annotation of DNA sequencing variants by leveraging the Ensembl Variant Effect Predictor with plugins (Yourshaw *et al.*, 2015)
- The State of Variant Annotation: A Comparison of AnnoVar, snpEff and VEP (http://blog.goldenhelix.com/ajesaitis/the-sate-of-variant-annotation-a-comparison-of-annovar-snpeff-and-vep/)
- Choice of transcripts and software has a large effect on variant annotation (McCarthy et al., 2014)

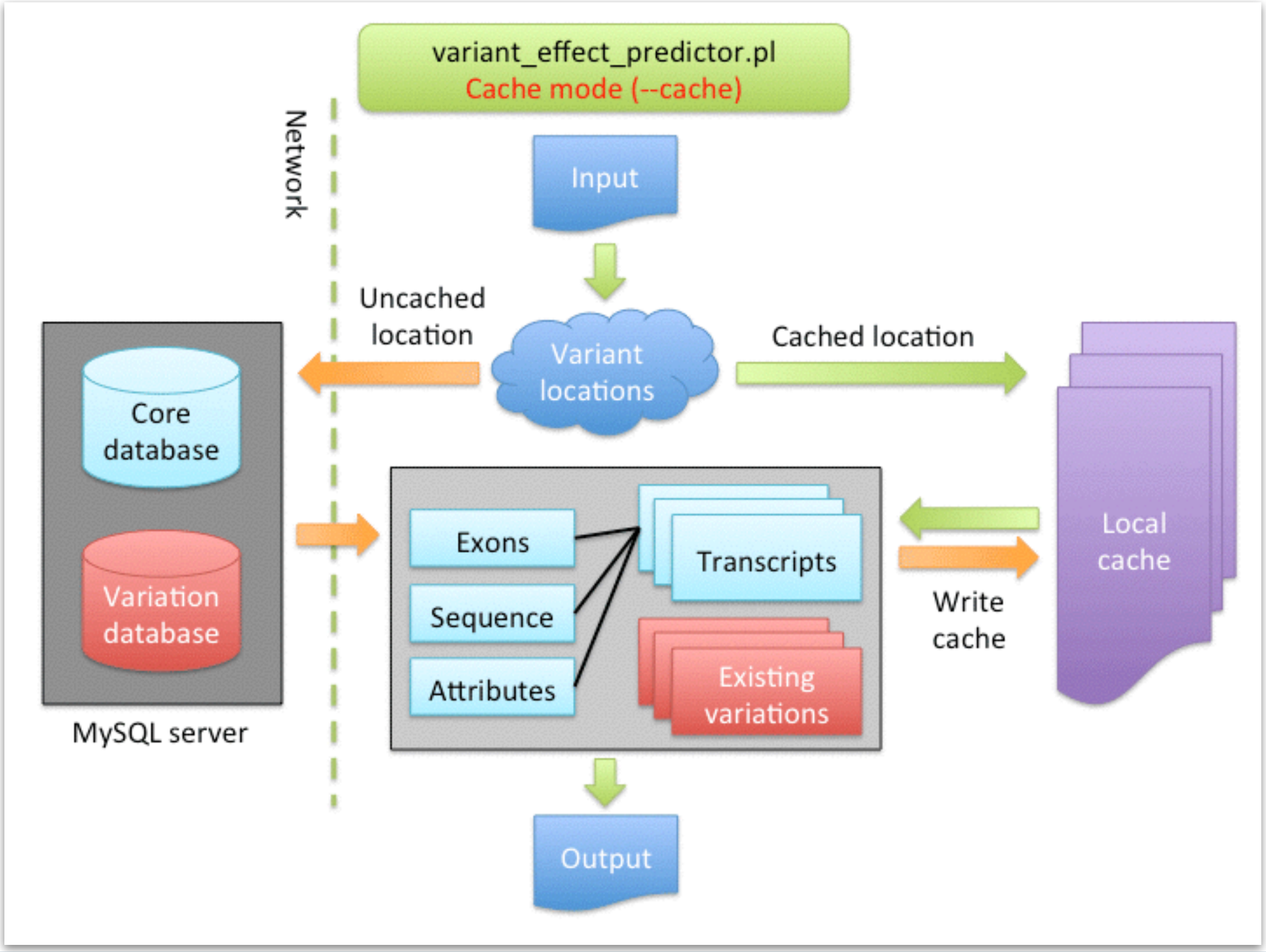
Ensembl Variant Effect Predictor (VEP)

Online



Ensembl Variant Effect Predictor (VEP)

Cache



Including External Resources

- Custom annotation
  — http://www.ensembl.org/info/docs/tools/vep/script/vep_custom.html
- VEP plugins
  — https://github.com/ensembl-variation/VEP_plugins
- Examples
  — http://www.ensembl.org/info/docs/tools/vep/script/vep_example.html

External Resources

- 1000 Genome Project
  — http://www.1000genomes.org/
- Exome Aggregation Consortium (ExAC) Database
  — http://exac.broadinstitute.org/
- dbNSFP
  — https://sites.google.com/site/jpopgen/dbNSFP

Assessing Deleteriousness

Name	Category	Score used for analysis	Deleterious threshold	Information used
SIFT	Function prediction	1 − Score	>0.95	Protein sequence conservation among homologs
PolyPhen-2	Function prediction	Score	>0.5	Eight protein sequence features, three protein structure features
LRT	Function prediction	Score * 0.5 (if Omega ≥1) or 1 − Score * 0.5 (if Omega <1)	P	DNA sequence evolutionary model
MutationTaster	Function prediction	Score (if A or D) or 1 − Score (if N or P)	>0.5	DNA sequence conservation, splice site prediction, mRNA stability prediction and protein feature annotations
Mutation Assessor	Function prediction	(Score-Min)/(Max − Min)	>0.65	Sequence homology of protein families and sub-families within and between species
FATHMM	Function prediction	1 − (Score-Min)/(Max − Min)	≥0.45	Sequence homology
GERP++ RS	Conservation score	Score	>4.4	DNA sequence conservation
PhyloP	Conservation score	Score	>1.6	DNA sequence conservation
SiPhy	Conservation score	Score	>12.17	Inferred nucleotide substitution pattern per site
PON-P	Ensemble score	Score	P	Random forest methodology-based pipeline integrating five predictors
PANTHER	Function prediction	Score	P	Phylogenetic trees based on protein sequences
PhD-SNP	Function prediction	Score	P	SVM-based method using protein sequence and profile information
SNAP	Function prediction	Score	P	Neural network-based method using DNA sequence information as well as functional and structural annotations
SNPs&GO	Function prediction	Score	P	SVM-based method using information from protein sequence, protein sequence profile and protein function
MutPred	Function prediction	Score	>0.5	Protein sequence-based model using SIFT and a gain/loss of 14 different structural and functional properties
KGGSeq	Ensemble score	Score	P	Filtration and prioritization framework using information from three levels: genetic level, variant-gene level and knowledge level
CONDEL	Ensemble score	Score	>0.49	Weighted average of the normalized scores of five methods
CADD	Ensemble score	Score	>15	63 distinct variant annotation retrieved from Ensembl Variant Effect Predictor (VEP), data from the ENCODE project and information from UCSC genome browser tracks

Dong *et al.*, 2015

Phred Quality Scores

- Assess/measure accuracy of base calling
- Defined as a property related to the **base calling error probabilities (P)**:

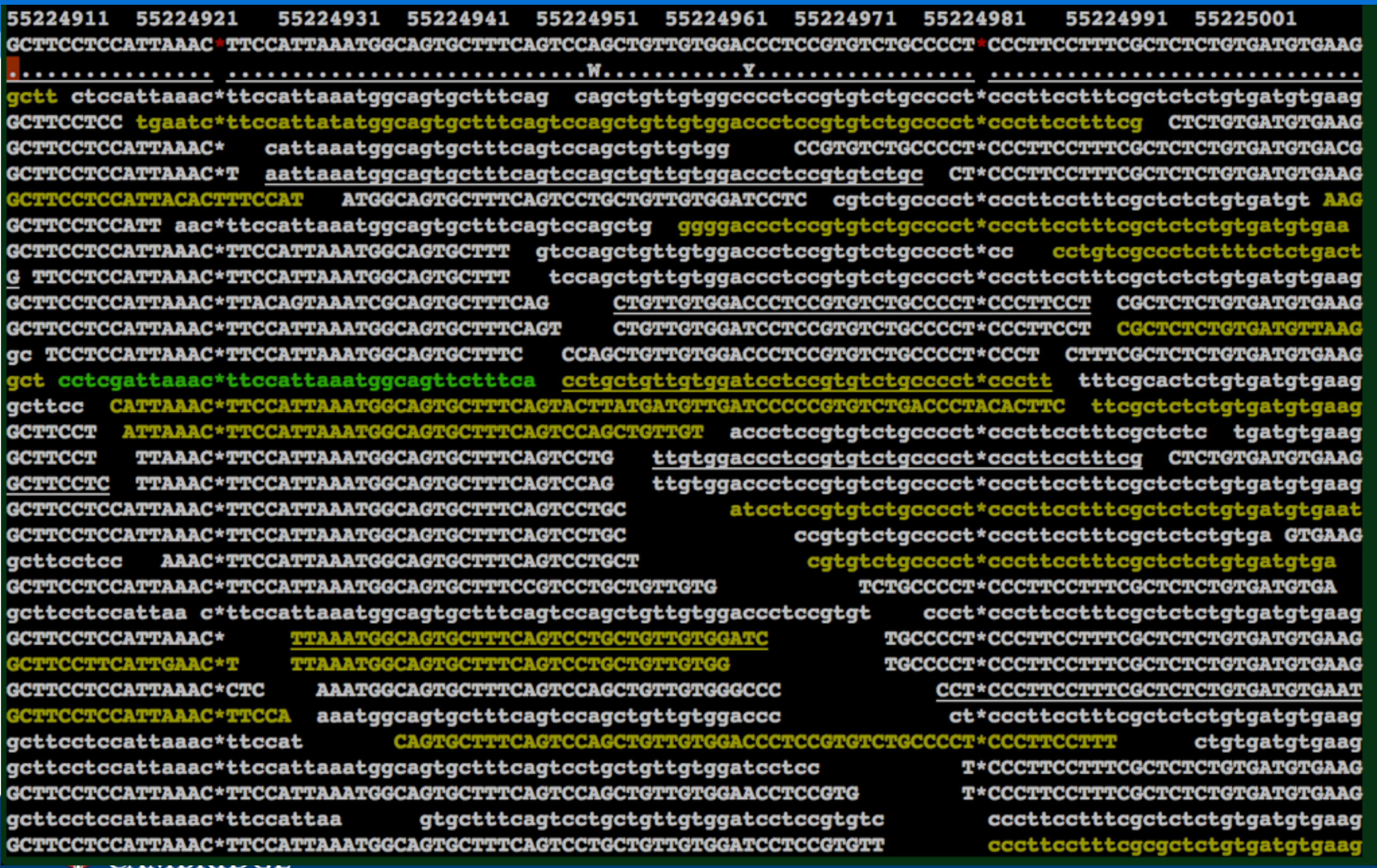
$$Q = -10 \log_{10}(P)$$

- Reaching Q30, virtually all bases in a read are called correctly:

Phred Quality Score	Probability of Incorrect Base Call	Base Call Accuracy
10	1 in 10	90%
20	1 in 100	99%
30	1 in 1,000	99.9%
40	1 in 10,000	99.99%
50	1 in 100,000	99.999%

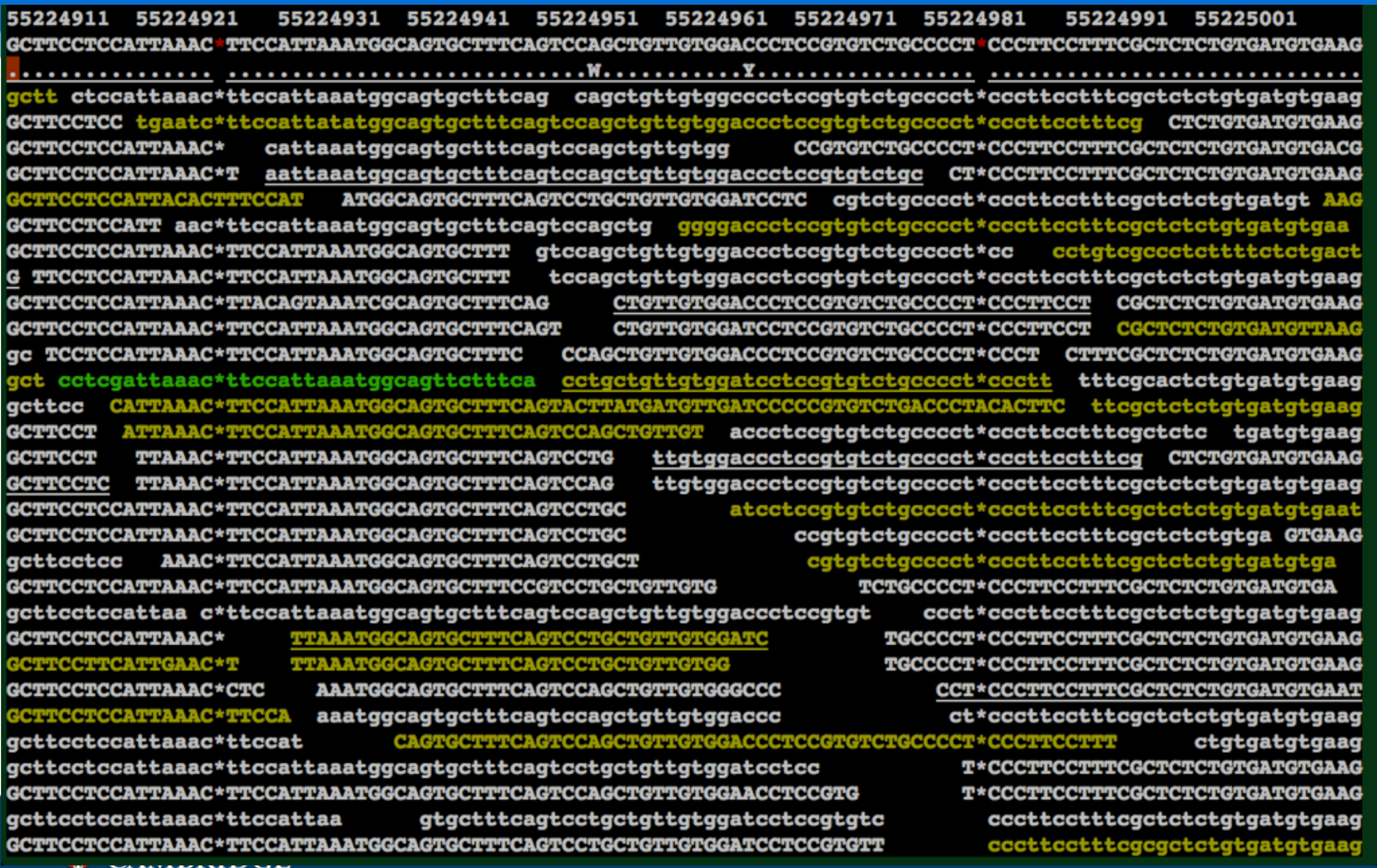
```
+-------------------------------------+
|          -=-    Help    -=-         |
|                                     |
|  ?           This window            |
|  Arrows      Small scroll movement  |
|  h,j,k,l     Small scroll movement  |
|  H,J,K,L     Large scroll movement  |
|  ctrl-H      Scroll 1k left         |
|  ctrl-L      Scroll 1k right        |
|  space       Scroll one screen      |
|  backspace   Scroll back one screen |
|  g           Go to specific location|
|  m           Color for mapping qual |
|  n           Color for nucleotide   |
|  b           Color for base quality |
|  c           Color for cs color     |
|  z           Color for cs qual      |
|  .           Toggle on/off dot view |
|  s           Toggle on/off ref skip |
|  r           Toggle on/off rd name  |
|  N           Turn on nt view        |
|  C           Turn on cs view        |
|  i           Toggle on/off ins      |
|  q           Exit                   |
|                                     |
|  Underline:      Secondary or orphan|
|  Blue:    0-9    Green: 10-19       |
|  Yellow: 20-29   White: >=30        |
+-------------------------------------+
```

Exploring the Raw Data (m: Mapping Quality)

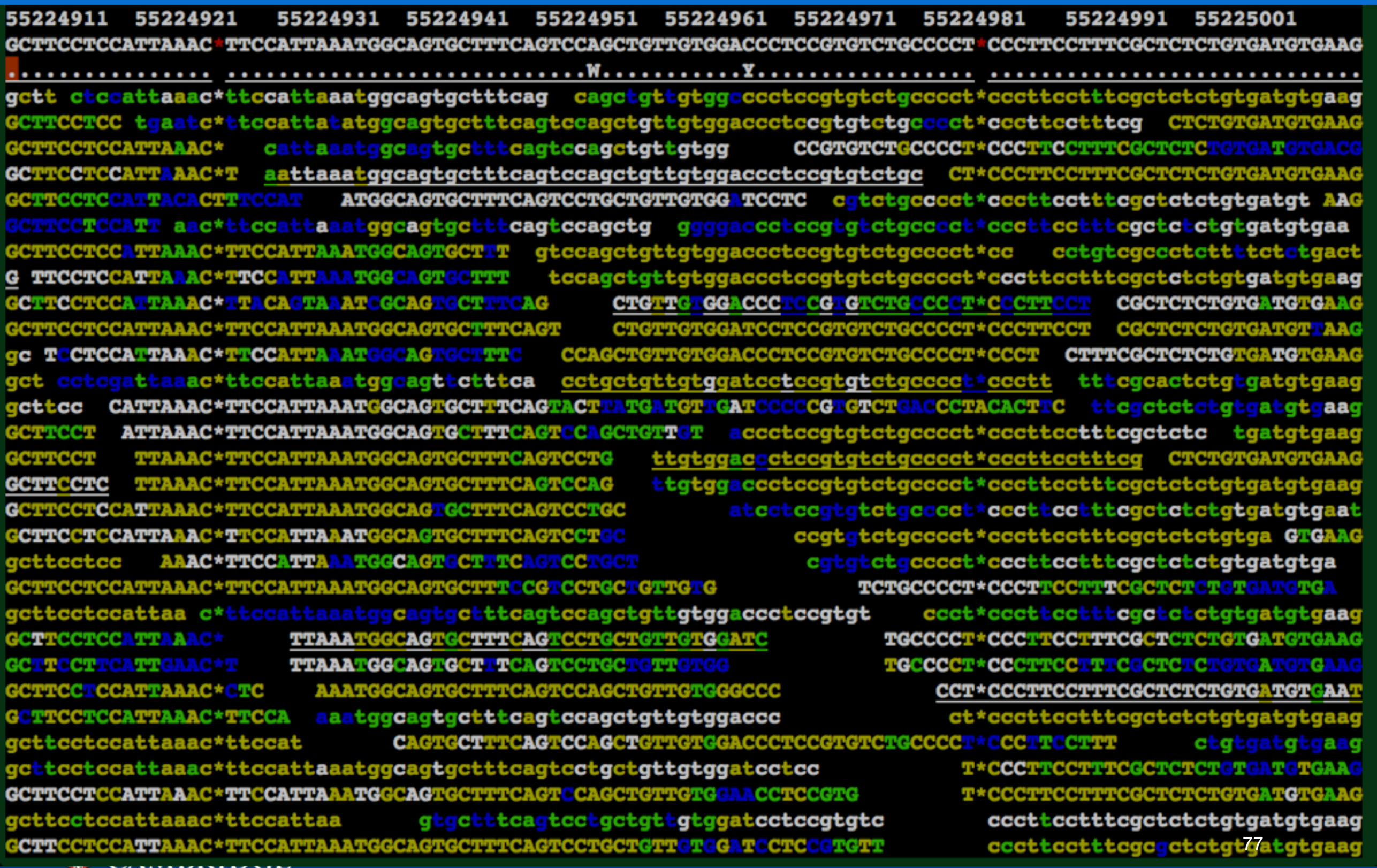


Exploring the Raw Data (r: read names)

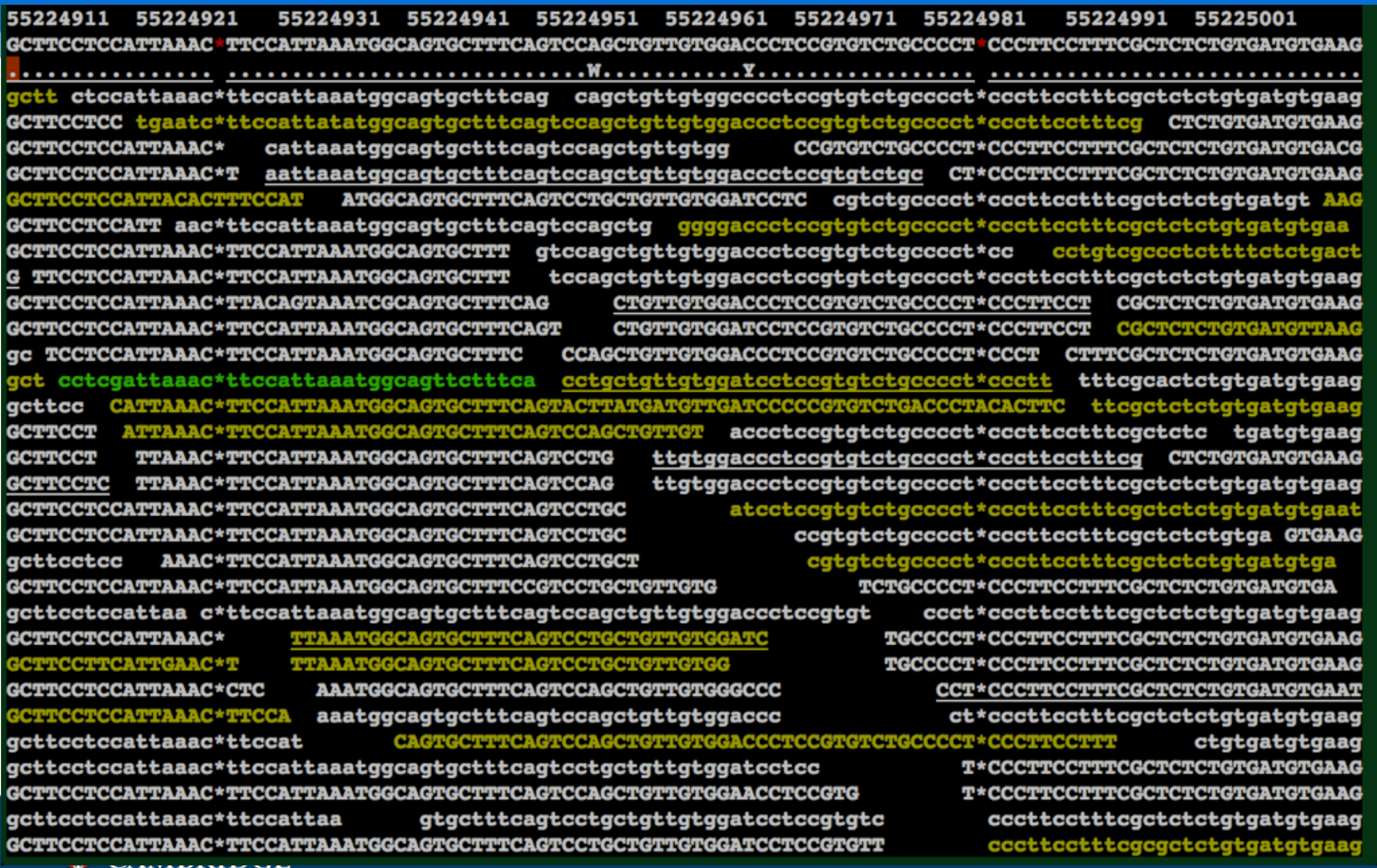
Exploring the Raw Data (m: Mapping Quality)



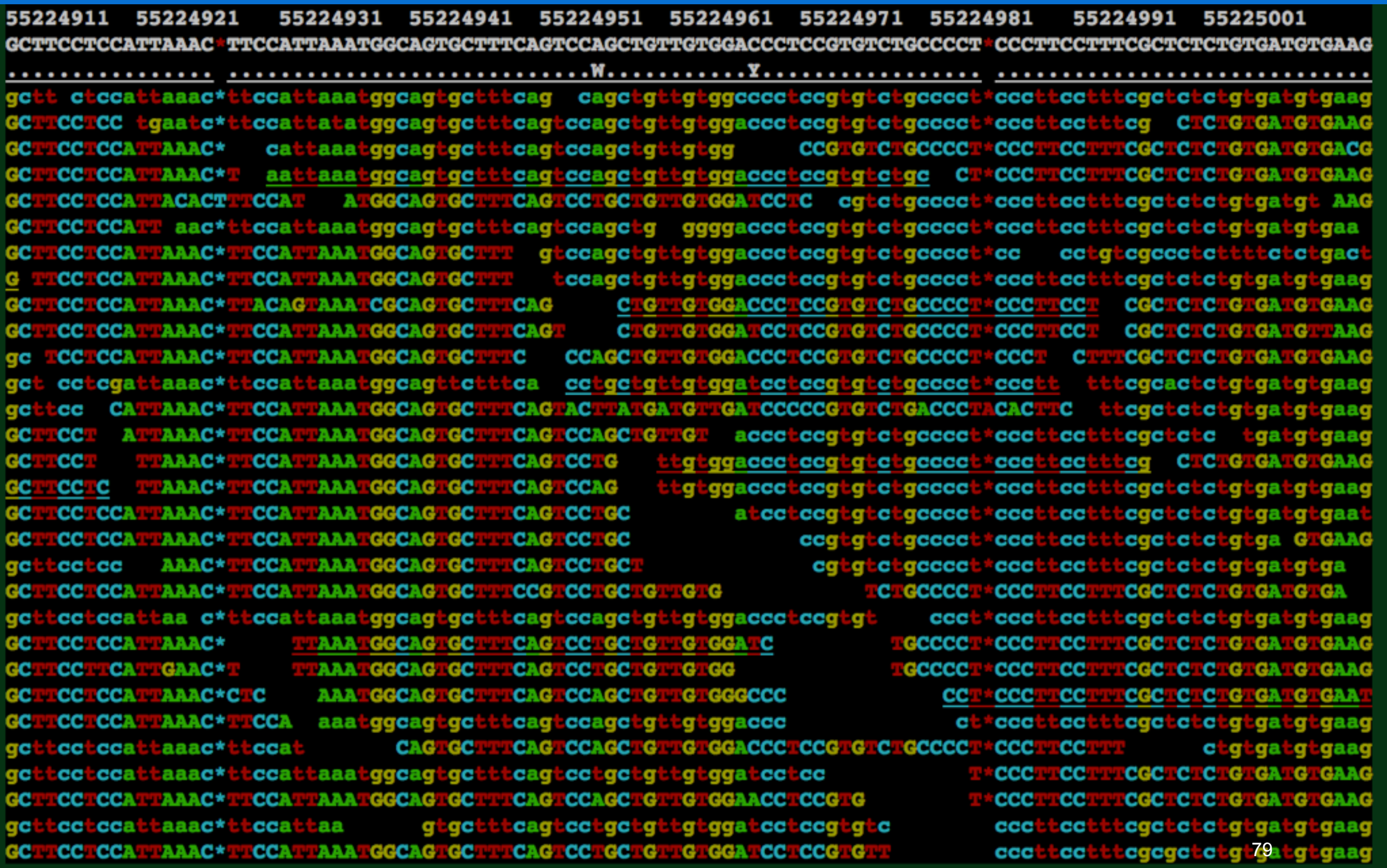
Exploring the Raw Data (b: Base Quality)



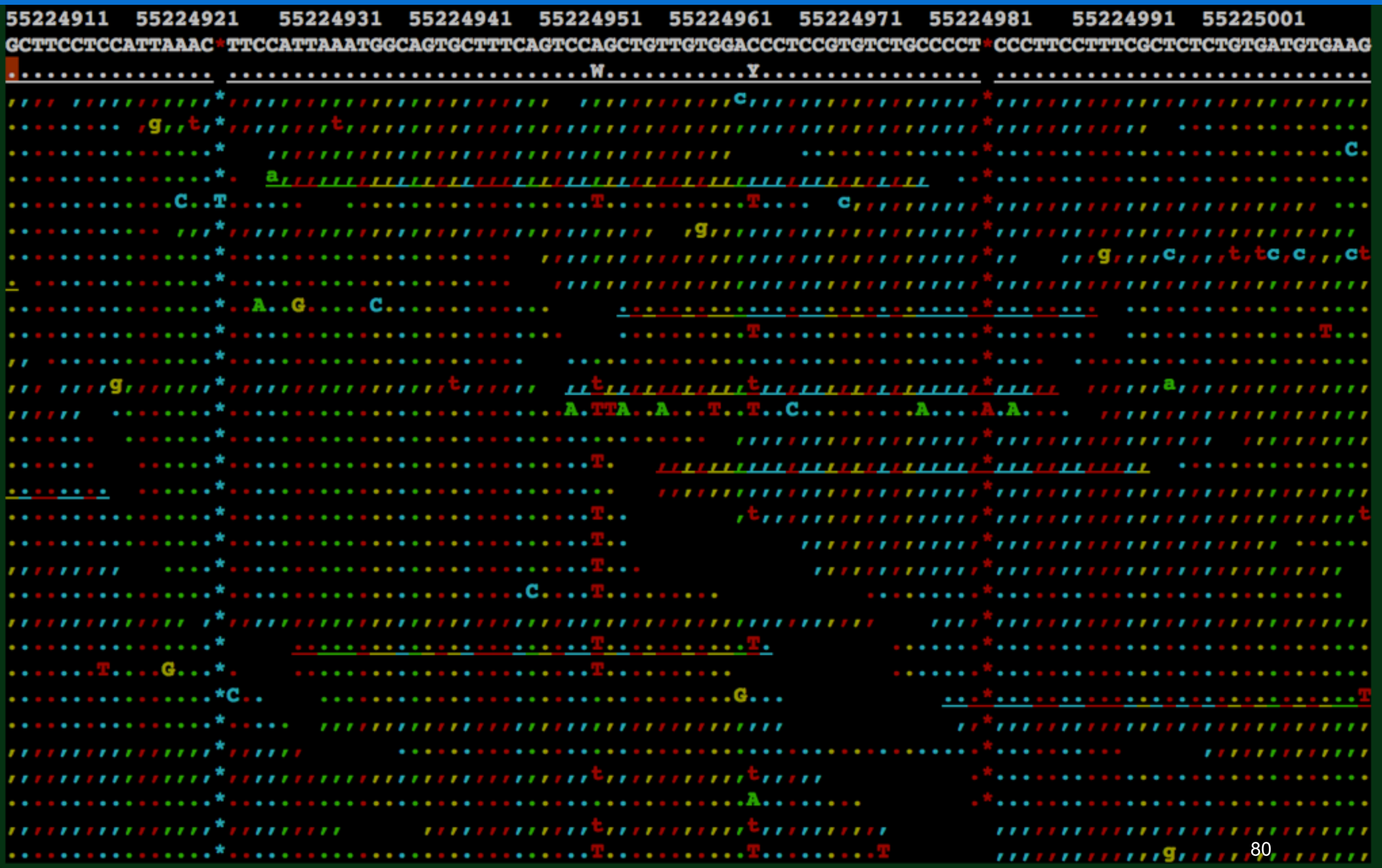
Exploring the Raw Data (m: Mapping Quality)



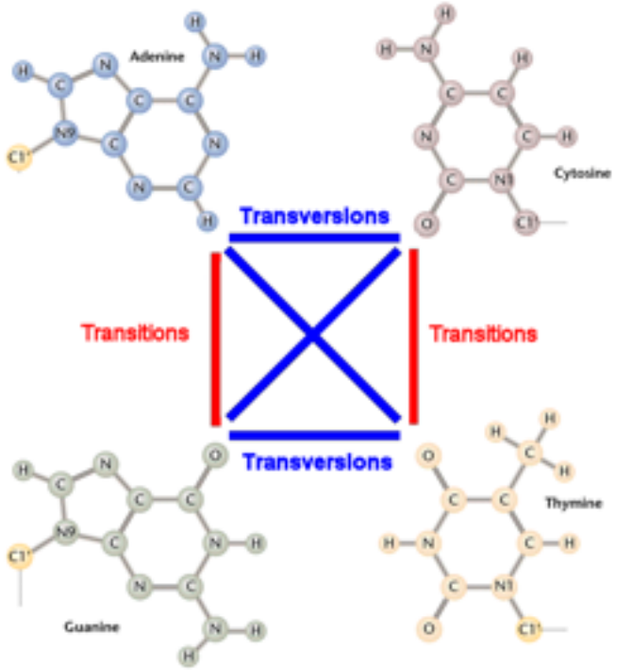
Exploring the Raw Data (n: Nucleotides Coloured)

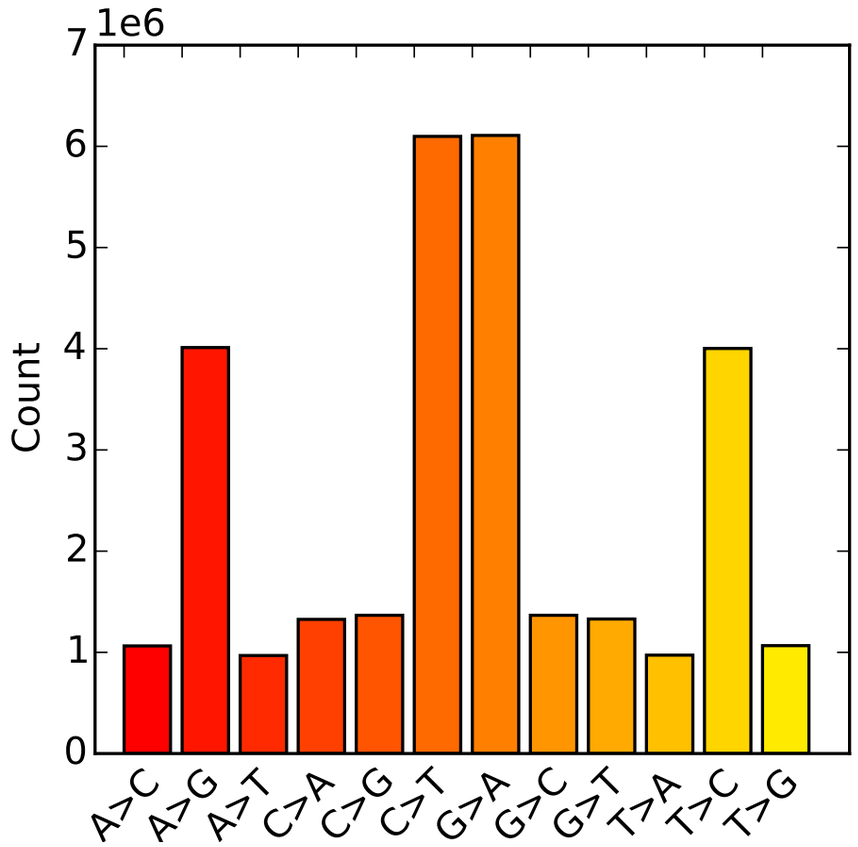


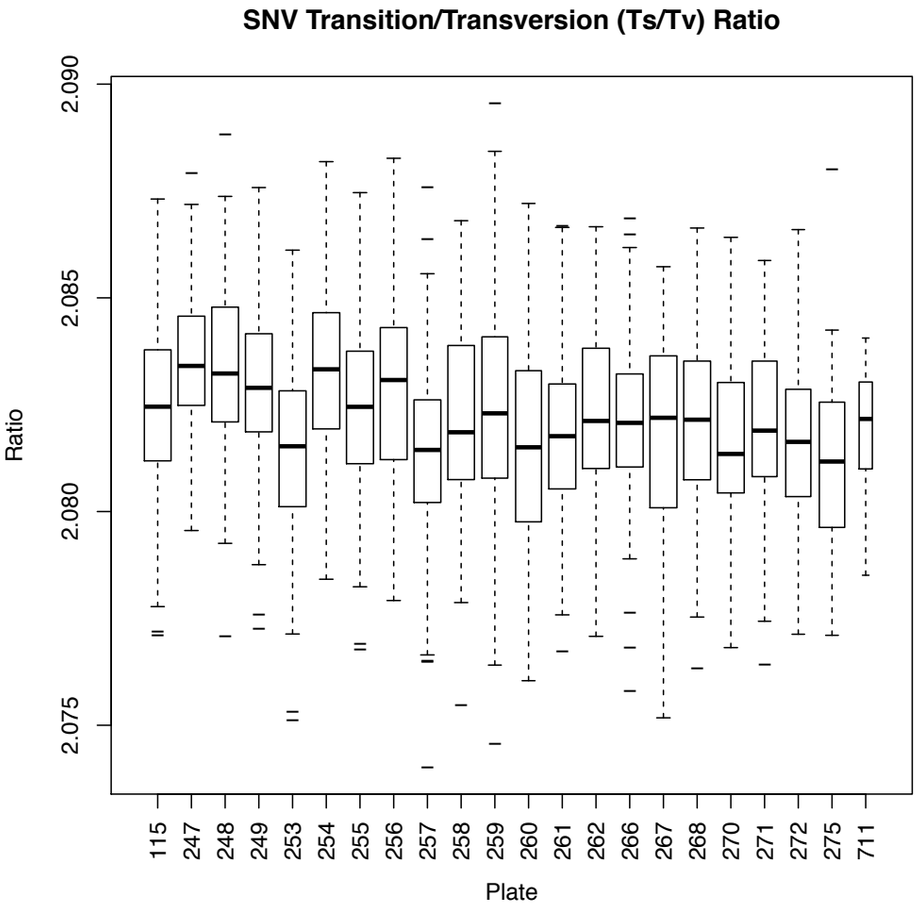
Exploring the Raw Data (.: Dot View)



Transition / Transversion







Variant Call Format (VCF)

```
##fileformat=VCFv4.2
##fileDate=20090805
##source=myImputationProgramV3.1
##reference=file:///seq/references/1000GenomesPilot-NCBI36.fasta
##contig=<ID=20,length=62435964,assembly=B36,md5=f126cdf8a6e0c7f379d618ff66beb2da,species="Homo sapiens",taxonomy=x>
##phasing=partial
##INFO=<ID=NS,Number=1,Type=Integer,Description="Number of Samples With Data">
##INFO=<ID=DP,Number=1,Type=Integer,Description="Total Depth">
##INFO=<ID=AF,Number=A,Type=Float,Description="Allele Frequency">
##INFO=<ID=AA,Number=1,Type=String,Description="Ancestral Allele">
##INFO=<ID=DB,Number=0,Type=Flag,Description="dbSNP membership, build 129">
##INFO=<ID=H2,Number=0,Type=Flag,Description="HapMap2 membership">
##FILTER=<ID=q10,Description="Quality below 10">
##FILTER=<ID=s50,Description="Less than 50% of samples have data">
##FORMAT=<ID=GT,Number=1,Type=String,Description="Genotype">
##FORMAT=<ID=GQ,Number=1,Type=Integer,Description="Genotype Quality">
##FORMAT=<ID=DP,Number=1,Type=Integer,Description="Read Depth">
##FORMAT=<ID=HQ,Number=2,Type=Integer,Description="Haplotype Quality">
#CHROM POS     ID        REF    ALT     QUAL FILTER INFO                              FORMAT      NA00001        NA00002        NA00003
20     14370   rs6054257 G      A       29   PASS   NS=3;DP=14;AF=0.5;DB;H2           GT:GQ:DP:HQ 0|0:48:1:51,51 1|0:48:8:51,51 1/1:43:5:.,.
20     17330   .         T      A       3    q10    NS=3;DP=11;AF=0.017               GT:GQ:DP:HQ 0|0:49:3:58,50 0|1:3:5:65,3   0/0:41:3
20     1110696 rs6040355 A      G,T     67   PASS   NS=2;DP=10;AF=0.333,0.667;AA=T;DB GT:GQ:DP:HQ 1|2:21:6:23,27 2|1:2:0:18,2   2/2:35:4
20     1230237 .         T      .       47   PASS   NS=3;DP=13;AA=T                   GT:GQ:DP:HQ 0|0:54:7:56,60 0|0:48:4:51,51 0/0:61:2
20     1234567 microsat1 GTC    G,GTCT  50   PASS   NS=3;DP=9;AA=G                    GT:GQ:DP    0/1:35:4       0/2:17:2       1/1:40:3
```

https://samtools.github.io/hts-specs/VCFv4.2.pdf