Next Generation Sequencing, Human Genetics, and COVID-19

MISHIMA, Hiroyuki, D.D.S., Ph. D.
Nagasaki University, Nagasaki, Japan

JICA in Tunisia Online Seminar for
the Project to Strengthen the Detecting and Analyzing Capacity in the Fight against COVID-19

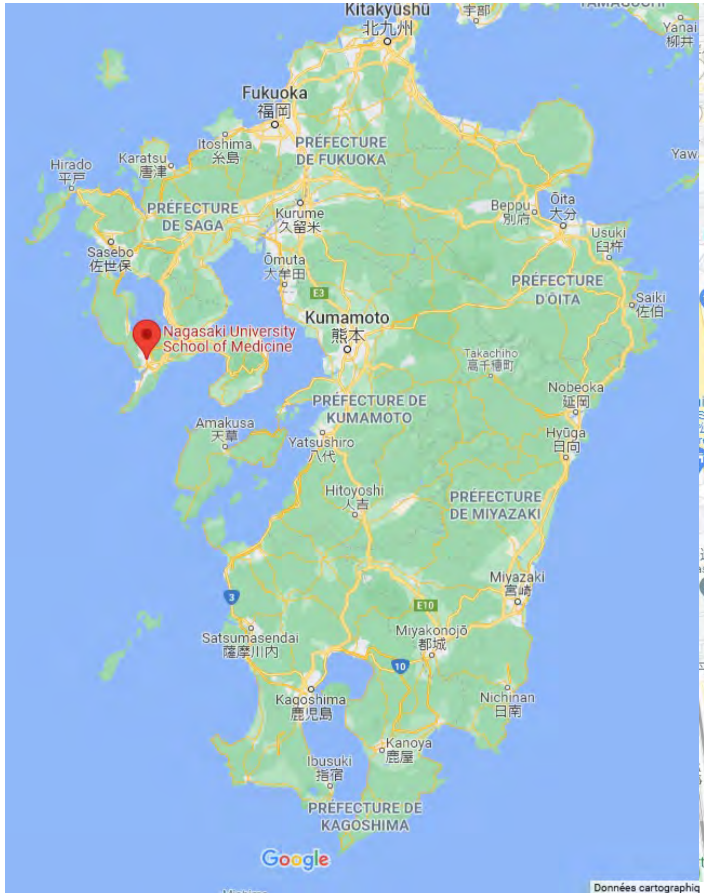
JICA チュニジア「新型コロナウィルス対策検査能力向上プロジェクト」遠隔研修 Feb 17, 2022

about 10,000 km

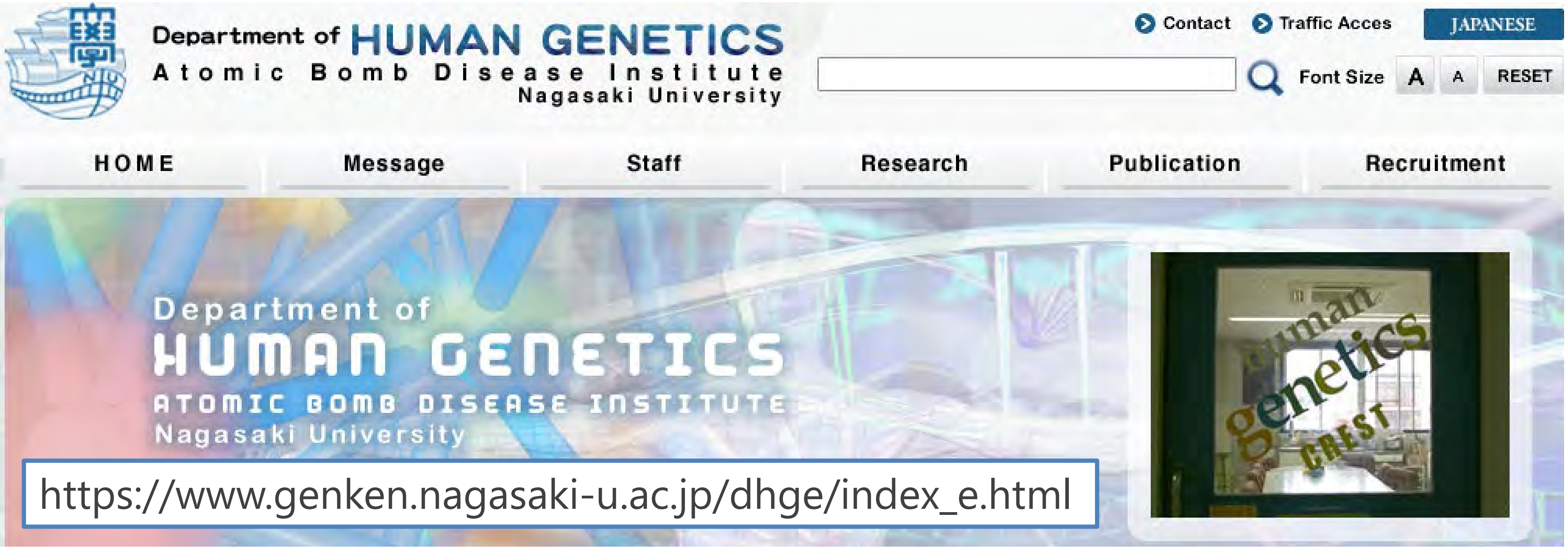
Tunis

Nagasaki

Hongrie
Mongolie
Bulgarie
Géorgie
Ouzbékistan
Kirghizistan
Istanbul
Ankara
Azerbaïdjan
Turquie
Turkménistan
Tadjikistan
Pékin
北京
Corée du Nord
Mer du Japon
Corée du Sud
Japon
Mer Tyrrhénienne
Tunisie
Mer Méditerranée
Syrie
Liban
Irak
Afghanistan
Chine
Nagasaki University School of Medicine
Shanghai
上海
Israël
Iran
Mer de Chine orientale
Le Caire
Pakistan
New Delhi
Népal
Bhoutan
Libye
Égypte
Golfe Persique
Émirats arabes unis
Arabie
Bangladesh
Taipei
台北
Taïwan



Atomic Bomb Disease Institute, Nagasaki University



https://www.genken.nagasaki-u.ac.jp/dhge/index_e.html

- Chair: Professor Dr. Koh-ichiro Yoshiura
- Human Genetics of …
  - rare genetic disorders
  - disorders with “missing inheritability”
  - epigenetic disorders

Next Generation Sequencers in Nagasaki U

Illumina
HiSeq2500

Illumina
MiSeq

Oxford Nanopore
PromethION
long-read sequencer

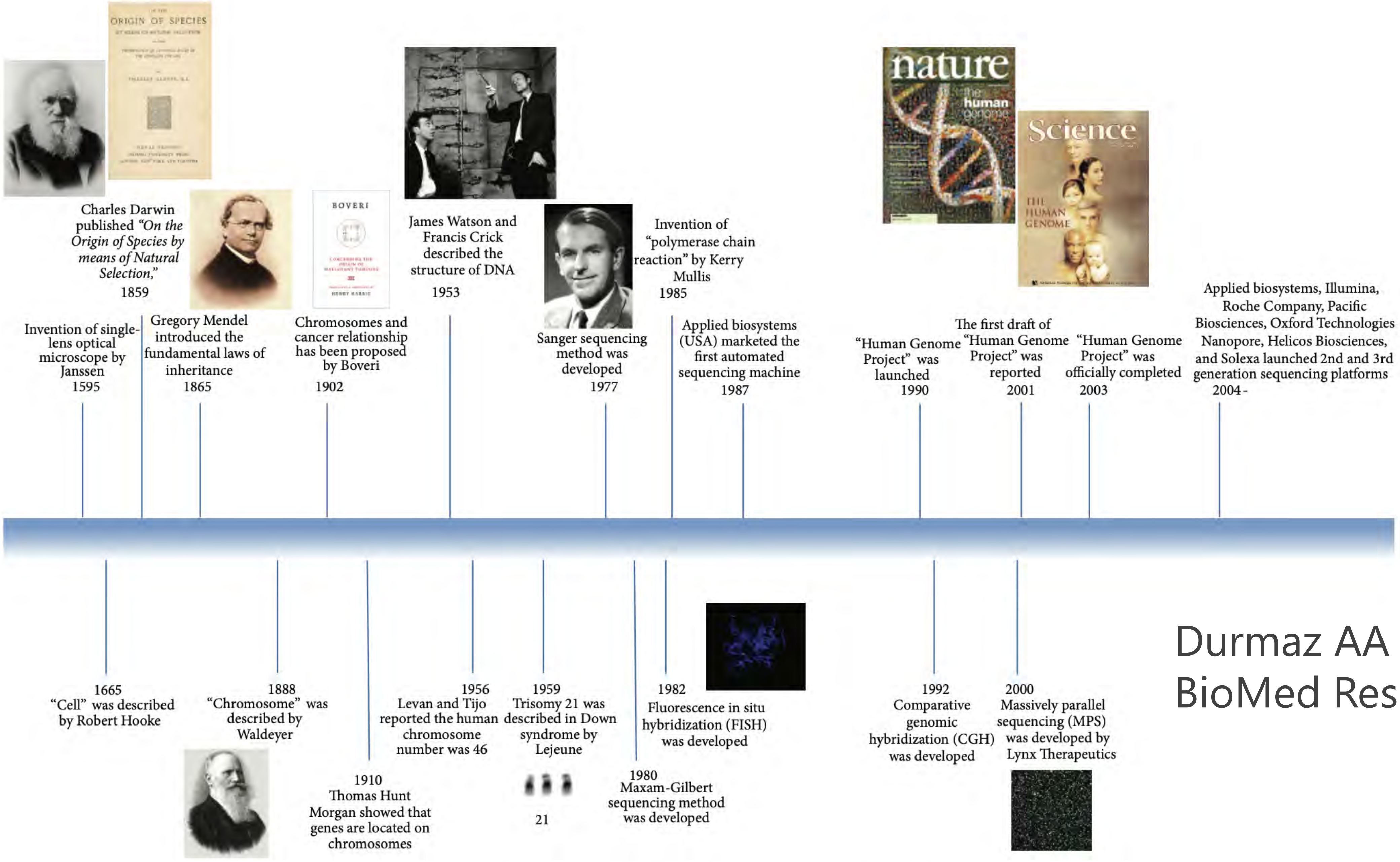
PART 1:

Japan’s experience and lessons learned with the next generation genomic sequencing system for the COVID-19 response

PART 2:

Human genome sequencing and analysis

Next Generation Sequencing (NGS)
a.k.a.
massively parallel sequencing



Invention of single-lens optical microscope by Janssen
1595

1665
"Cell" was described by Robert Hooke

Charles Darwin published *"On the Origin of Species by means of Natural Selection,"*
1859

Gregory Mendel introduced the fundamental laws of inheritance
1865

1888
"Chromosome" was described by Waldeyer

Chromosomes and cancer relationship has been proposed by Boveri
1902

1910
Thomas Hunt Morgan showed that genes are located on chromosomes

James Watson and Francis Crick described the structure of DNA
1953

1956
Levan and Tijo reported the human chromosome number was 46

1959
Trisomy 21 was described in Down syndrome by Lejeune

Sanger sequencing method was developed
1977

1980
Maxam-Gilbert sequencing method was developed

1982
Fluorescence in situ hybridization (FISH) was developed

Invention of "polymerase chain reaction" by Kerry Mullis
1985

Applied biosystems (USA) marketed the first automated sequencing machine
1987

"Human Genome Project" was launched
1990

1992
Comparative genomic hybridization (CGH) was developed

2000
Massively parallel sequencing (MPS) was developed by Lynx Therapeutics

The first draft of "Human Genome Project" was reported
2001

"Human Genome Project" was officially completed
2003

Applied biosystems, Illumina, Roche Company, Pacific Biosciences, Oxford Technologies Nanopore, Helicos Biosciences, and Solexa launched 2nd and 3rd generation sequencing platforms
2004 -

FIGURE 1: Landmarks in the history of genetics.

Durmaz AA et al. BioMed Res Int. 2015

Sanger sequencing method was developed
1977

Invention of "polymerase chain reaction" by Kerry Mullis
1985

Applied biosystems (USA) marketed the first automated sequencing machine
1987

"Human Genome Project" was launched
1990

The first draft of "Human Genome Project" was reported
2001

"Human Genome Project" was officially completed
2003

Applied biosystems, Illumina, Roche Company, Pacific Biosciences, Oxford Technologies Nanopore, Helicos Biosciences, and Solexa launched 2nd and 3rd generation sequencing platforms
2004 -

Durmaz AA et al.
BioMed Res Int. 2015

Next Generation Sequencers in Nagasaki U

Illumina
HiSeq2500

Illumina
MiSeq

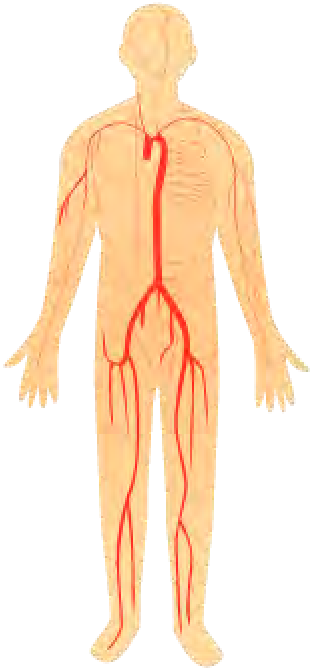
Oxford Nanopore
PromethION
long-read sequencer

Jan 20, 2015

THE PRECISION MEDICINE INITIATIVE

www.whitehouse.gov/precision-dedicine

Precision medicine for rare and undiagnosed diseases

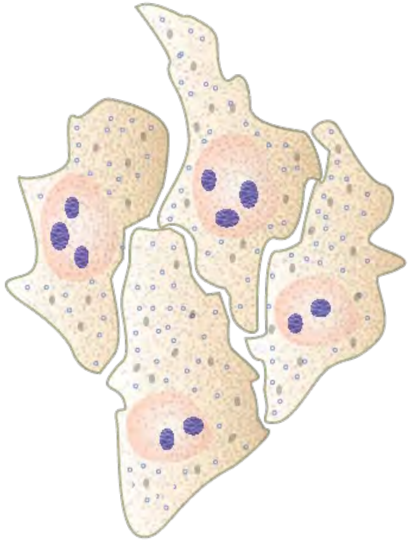


Genome analysis of patients and their families
↓
diagnosis

selection and development of treatment

Togo picture gallery by DBCLS is licensed under a Creative Commons Attribution 2.1 Japan license (c)

Precision medicine for cancer

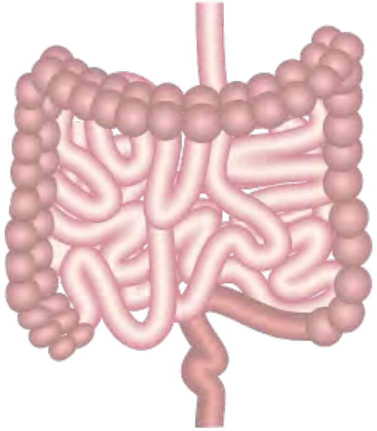


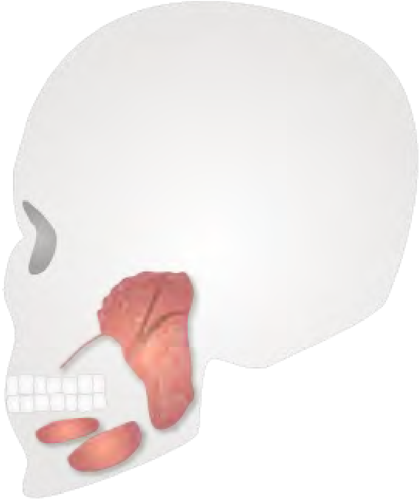
analysis of genome vary among lesions and stages

development and selection of treatment

Precision medicine based on metagenomics

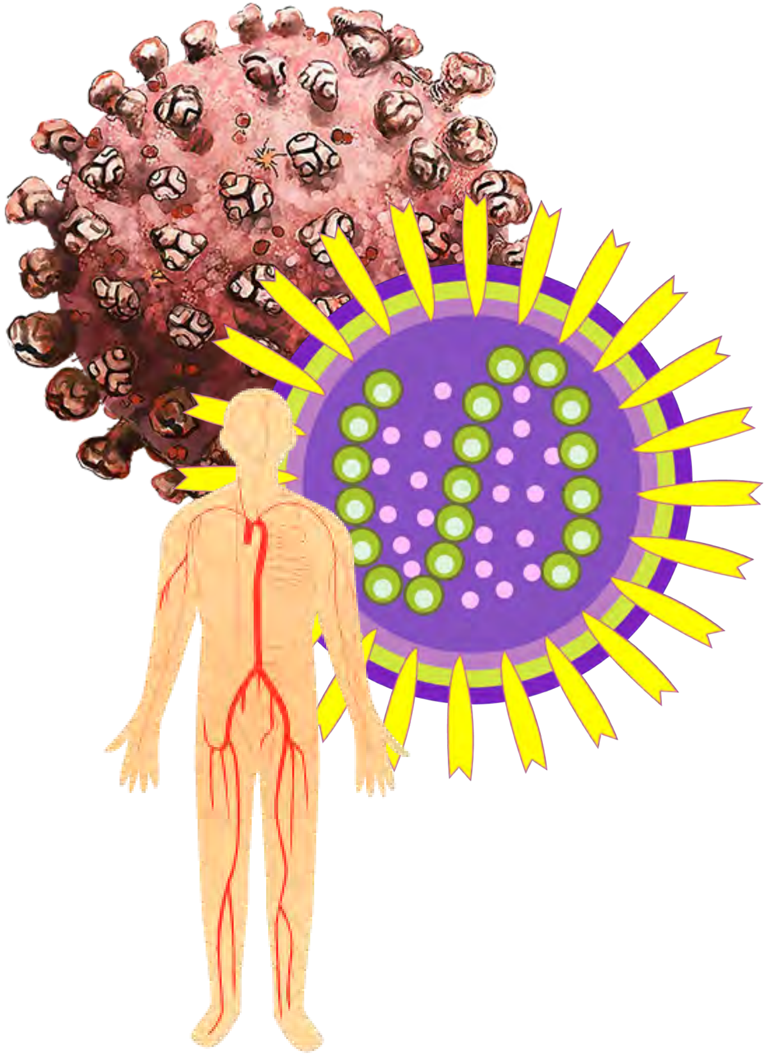
Metagenomic analysis of oral and gut bacterial flora





knowing body condition and disease stages objectively

Precision medicine of infectious diseases

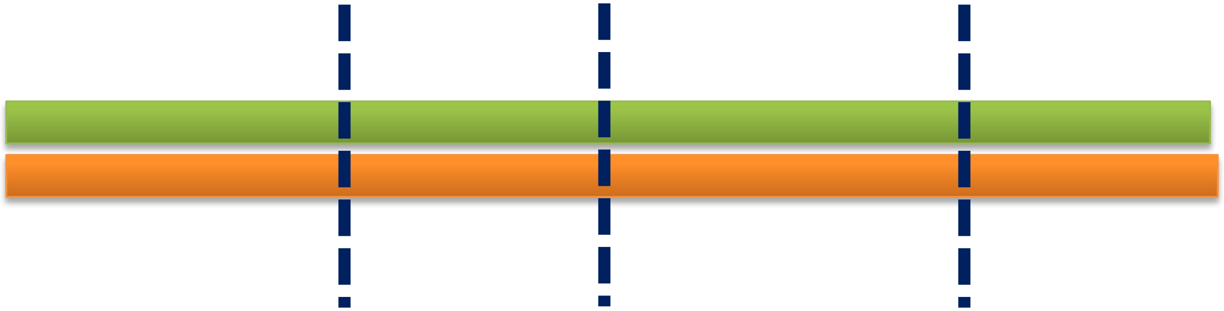


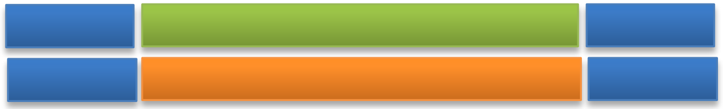
Rapid and accurate typing of bacteria and viruses
Finding hosts' genomic factors affecting disease severity.

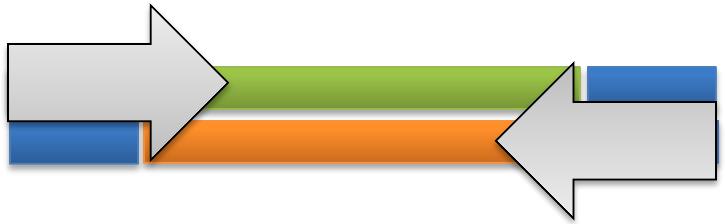
development and selection of treatment

Togo picture gallery by DBCLS is licensed under a Creative Commons Attribution 2.1 Japan license (c)

NGS simplified

1. 
DNA shearing

2. 
adaptor ligation to build a library

3. 
sequencing molecular termini

performing above in a massively parallel manner performing above in a massively parallel manner

- expensive instruments and delicate benchwork
- huge-size data and complicated analysis workflows

A good team both for “wet” and “dry” experiments are essential for NGS researches

”Wet” experiments

- Sample collection and transportation
- Sample storage and management
- DNA/RNA extraction from samples
- Reagent management
- Library Preparation
  - DNA shearing
  - target enrichment (PCR, hybridization, etc)
- NGS Sequencing
- NGS instrument maintenance

”Dry” experiments or Bioinformatics

- Raw data storage and routine backups
- Computer maintenance
- Data storage maintenance
- Building analysis workflows
- Updating system software
- Updating analysis software packages and workflows
- Analysis and interpret results
- Summarize results for noncomputer people
- Integrate and interpret multiple experiment results
- Keep data and workflows reproducible

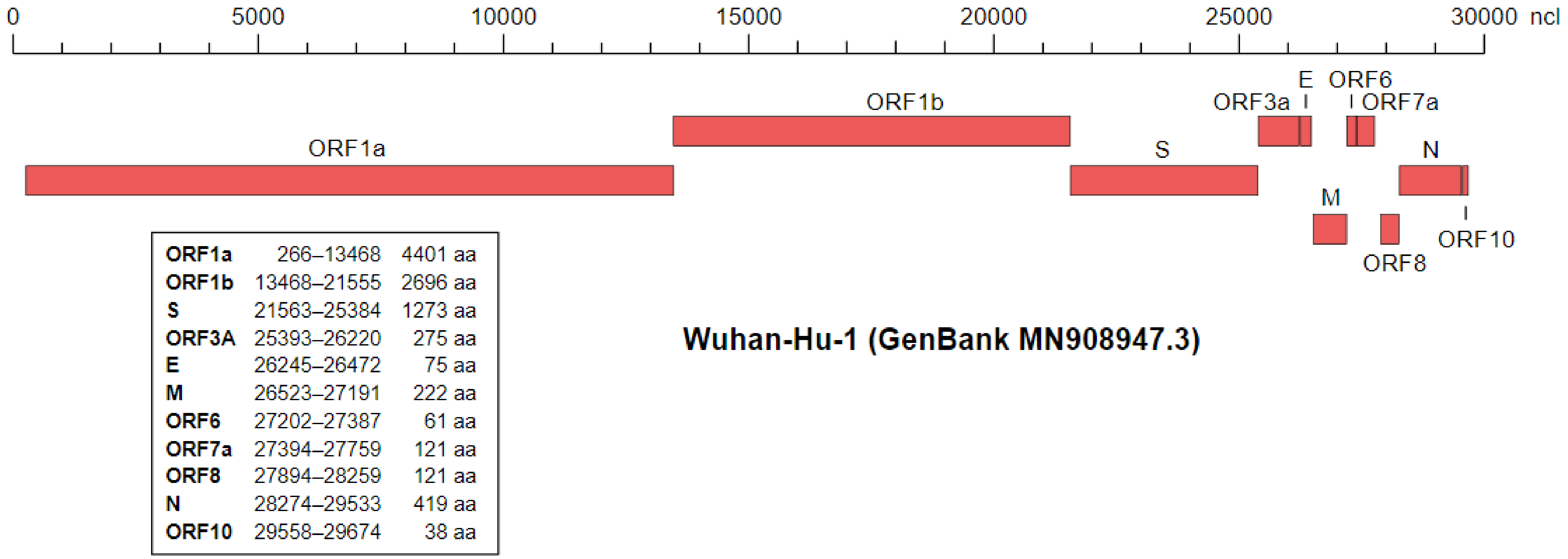
How can we fight against COVID-19 with NGSs?

1. revealing the genome of the virus, SARS-CoV-2
2. revealing the genome of the host, humans

DISCLAMER:

- The lecturer does not participate in COVID-19 projects personally.
- The lecturer intends to introduce current NGS-related COVID-19 research projects in Japan.

Original SARS-CoV-2 genome that isolated in Wuhan, China in December, 2019



ORF1a	266–13468	4401 aa
ORF1b	13468–21555	2696 aa
S	21563–25384	1273 aa
ORF3A	25393–26220	275 aa
E	26245–26472	75 aa
M	26523–27191	222 aa
ORF6	27202–27387	61 aa
ORF7a	27394–27759	121 aa
ORF8	27894–28259	121 aa
N	28274–29533	419 aa
ORF10	29558–29674	38 aa

Wuhan-Hu-1 (GenBank MN908947.3)

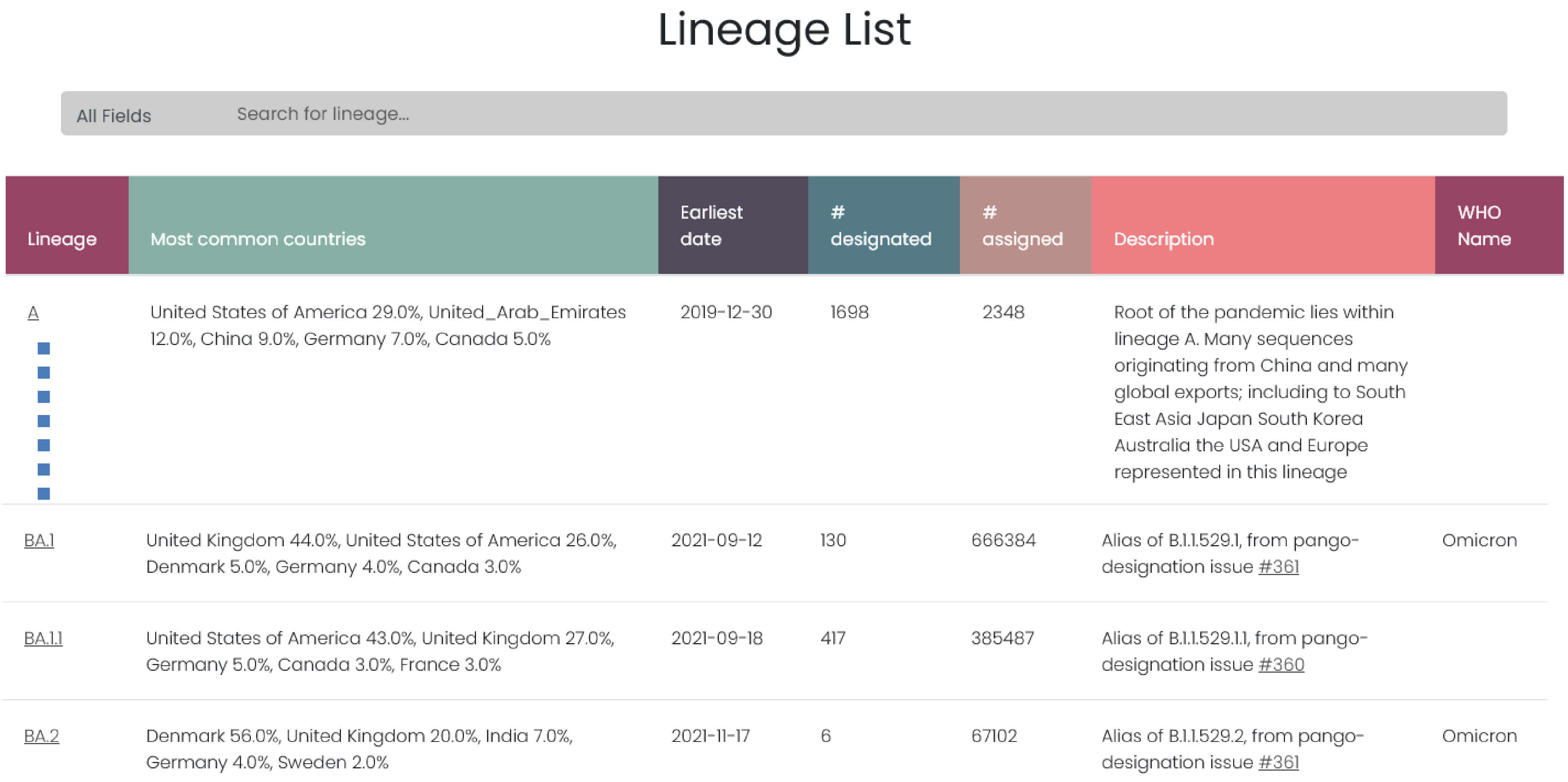
Wu, F et al., https://commons.wikimedia.org/wiki/File:SARS-CoV-2_genome.svg

Lineage of SARS-CoV-2

- A lineage is a group of a closely related viruses (= variation).
- Linages have (quite) different characters in such as pathogenicity, transmissibility and disease progression.

PANGO nomenclature at https://cov-lineages.org/

the Phylogenetic Assignment of Named Global Outbreak Lineages (PANGLIN) software package
https://github.com/cov-lineages/pangolin



Lineage List

Lineage	Most common countries	Earliest date	# designated	# assigned	Description	WHO Name
A	United States of America 29.0%, United_Arab_Emirates 12.0%, China 9.0%, Germany 7.0%, Canada 5.0%	2019-12-30	1698	2348	Root of the pandemic lies within lineage A. Many sequences originating from China and many global exports; including to South East Asia Japan South Korea Australia the USA and Europe represented in this lineage	
BA.1	United Kingdom 44.0%, United States of America 26.0%, Denmark 5.0%, Germany 4.0%, Canada 3.0%	2021-09-12	130	666384	Alias of B.1.1.529.1, from pango-designation issue #361	Omicron
BA.1.1	United States of America 43.0%, United Kingdom 27.0%, Germany 5.0%, Canada 3.0%, France 3.0%	2021-09-18	417	385487	Alias of B.1.1.529.1.1, from pango-designation issue #360	
BA.2	Denmark 56.0%, United Kingdom 20.0%, India 7.0%, Germany 4.0%, Sweden 2.0%	2021-11-17	6	67102	Alias of B.1.1.529.2, from pango-designation issue #361	Omicron

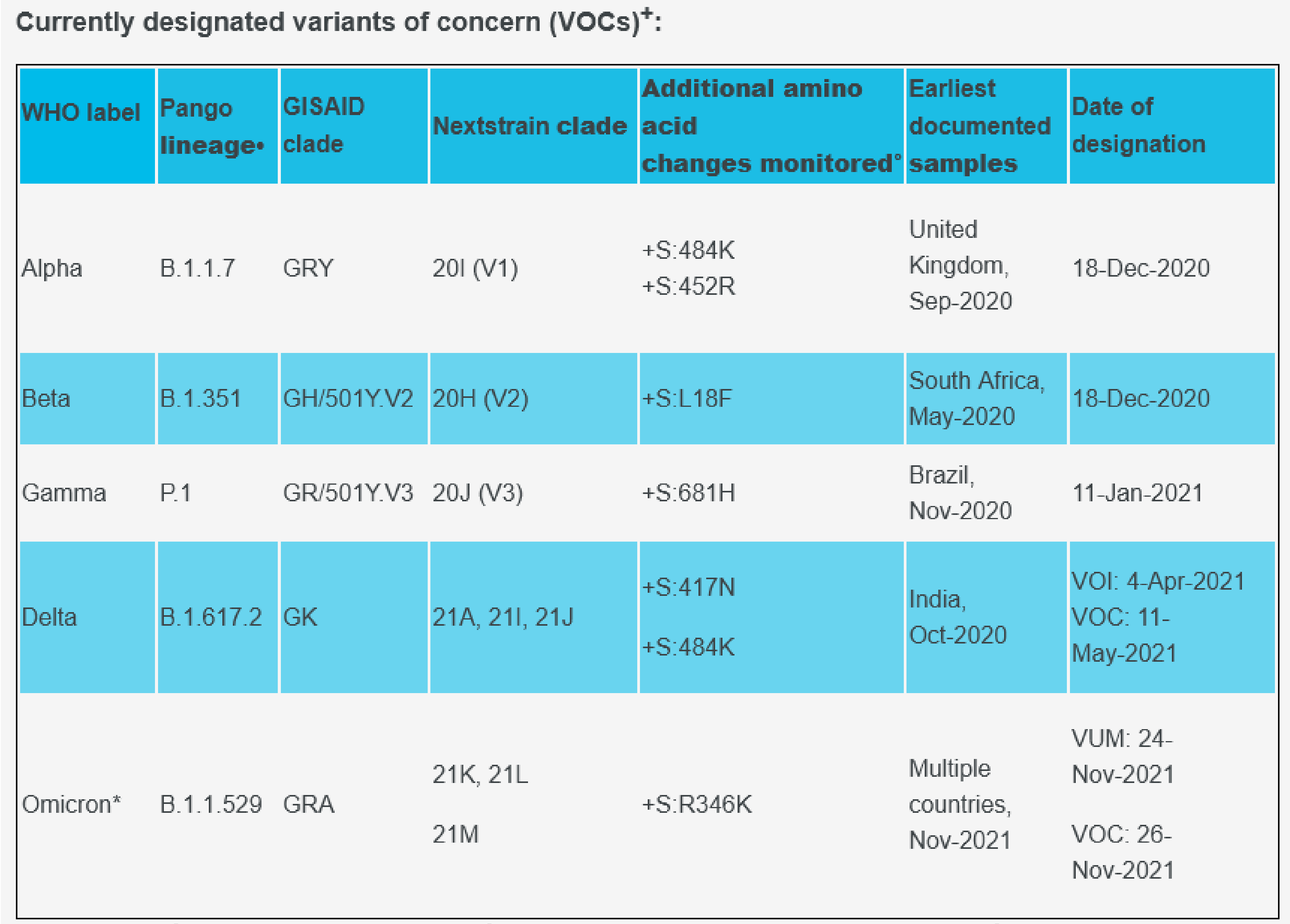
WHO labeling of SARS-CoV-2 variants

Variants of concern (VOC)

- Increase in transmissibility or detrimental change in COVID-19 epidemiology; OR
- Increase in virulence or change in clinical disease presentation; OR
- Decrease in effectiveness of public health and social measures or available diagnostics, vaccines, therapeutics.

https://www.who.int/en/activities/tracking-SARS-CoV-2-variants/

WHO labeling of SARS-CoV-2 variants



Currently designated variants of concern (VOCs)+:

WHO label	Pango lineage•	GISAID clade	Nextstrain clade	Additional amino acid changes monitored°	Earliest documented samples	Date of designation
Alpha	B.1.1.7	GRY	20I (V1)	+S:484K +S:452R	United Kingdom, Sep-2020	18-Dec-2020
Beta	B.1.351	GH/501Y.V2	20H (V2)	+S:L18F	South Africa, May-2020	18-Dec-2020
Gamma	P.1	GR/501Y.V3	20J (V3)	+S:681H	Brazil, Nov-2020	11-Jan-2021
Delta	B.1.617.2	GK	21A, 21I, 21J	+S:417N +S:484K	India, Oct-2020	VOI: 4-Apr-2021 VOC: 11-May-2021
Omicron*	B.1.1.529	GRA	21K, 21L 21M	+S:R346K	Multiple countries, Nov-2021	VUM: 24-Nov-2021 VOC: 26-Nov-2021

https://www.who.int/en/activities/tracking-SARS-CoV-2-variants/

Lineage detection/assignment

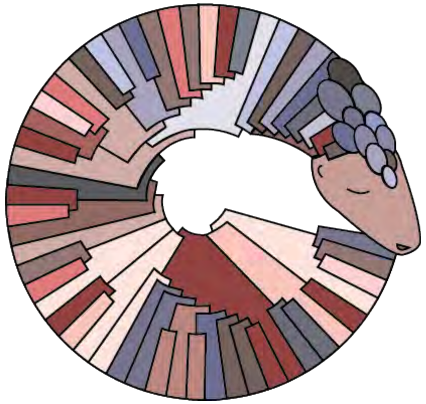
- Lineage can be assigned by determinizing lineage-specific genomic sequences or mutation (amino acid change).
- Determinate point mutations
  - PCR-based techniques
  - Protein (antigen)-based techniques
  - Cannot assign novel lineage
  - Cannot assign “stealth” lineage leading false negatives (e.g. Omicron variant)
- Whole genome sequencing (WGS) is necessary for accurate and robust lineage assignment

SARS-CoV-2 Whole Genome Sequencing Projects in Japan

- Center for Medical Genetics, Keio University
  https://cmg.med.keio.ac.jp/
- National Institute of Genetics (NIG)
  https://www.nig.ac.jp/nig/about-nig/covid19bcp
- National Institute of Infectious Diseases (NIID)
  https://www.niid.go.jp/niid/en/2019-ncov-e.html

Typical workflow of NGS-based SARS-Cov-2 genome analysis

1. RNA sample extracted clinical materials
2. cDNA synthesis and library construction
3. Performing NGS
4. data analysis
    1. mapping and variation detection
    2. lineage assignment (e.g. PANGOLIN)
5. Deposit sequence to public databases
    1. DDBJ/INSD (DDBJ+NCBI+ENA/EBI int’l alliance)
    2. GISAID (DB for pandemic genomes)



Statistics of SARS-CoV-2 genome sequencing at NIID

- number of samples: 114,502
- Detected PANGO lineages and WHO labels

B.1.351 (Beta):	117 cases	
P.1 (Gamma):	137 cases	B.1.1.28.1 = P.1
B.1.617.2 (Delta):	98,131 cases	B.1.617.2.28 = AY.29

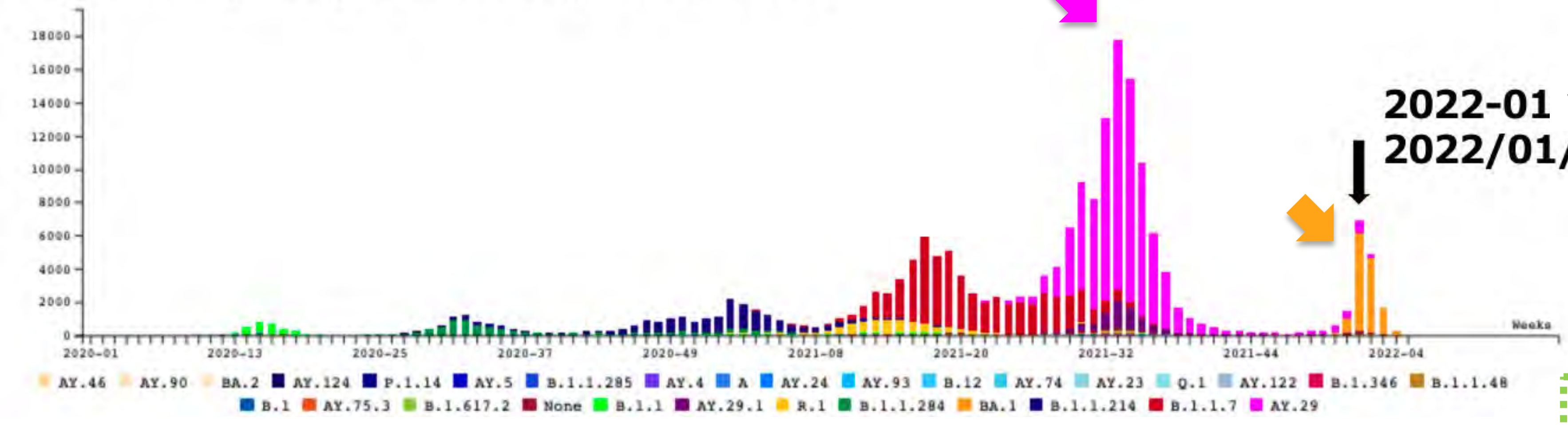
B1.1.529 (Omicron):	52,314 cases	B.1.1.519.1 = BA.1
		B.1.1.519.2 = BA.2

Ministry of Health, Labour and Welfare (as of Jan 31, 2022)

SARS-CoV-2 genome PANGO lineage transition by NIID

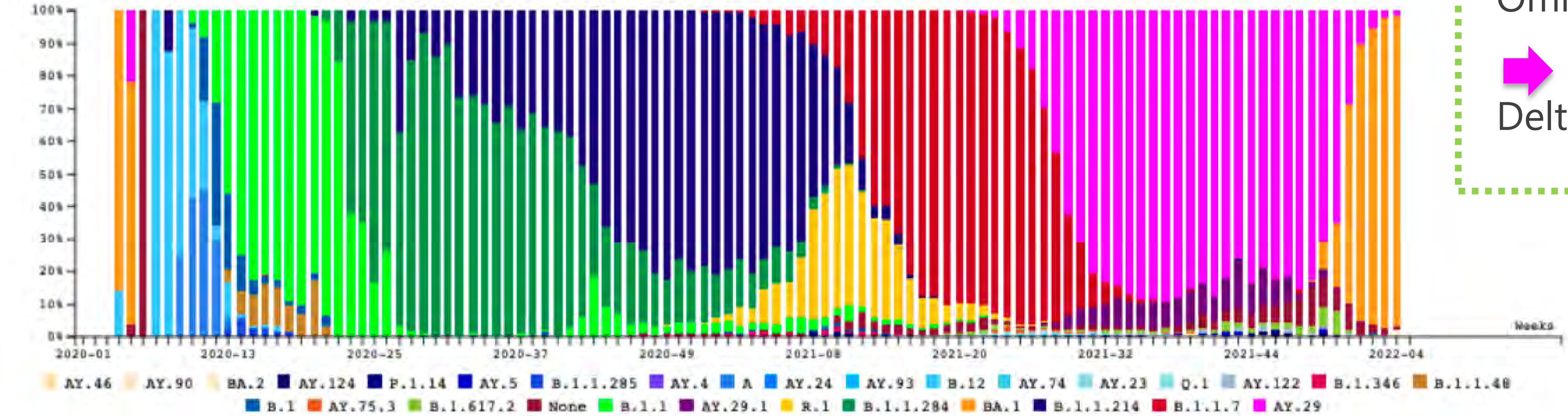
(as of Feb. 4, 2022)

[Only Domestic] Weekly Top 30 Graph (count each week)



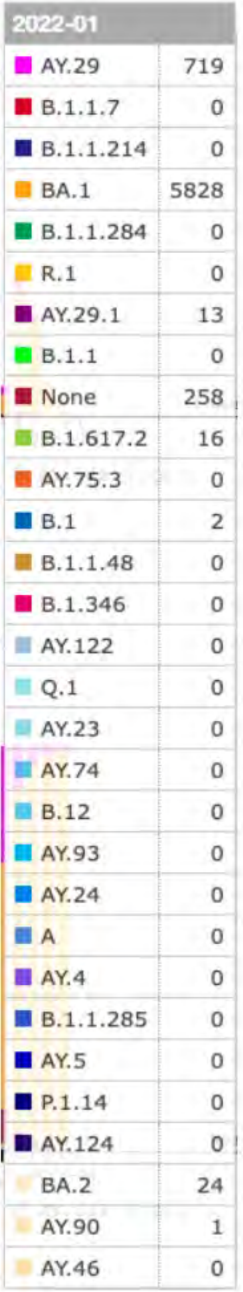
2022-01 week
2022/01/03 - 01/09

[Only Domestic] Weekly Top 30 Stacked Graph (count each week)



Omicron

Delta



2022-01	
AY.29	719
B.1.1.7	0
B.1.1.214	0
BA.1	5828
B.1.1.284	0
R.1	0
AY.29.1	13
B.1.1	0
None	258
B.1.617.2	16
AY.75.3	0
B.1	2
B.1.1.48	0
B.1.346	0
AY.122	0
Q.1	0
AY.23	0
AY.74	0
B.12	0
AY.93	0
AY.24	0
A	0
AY.4	0
B.1.1.285	0
AY.5	0
P.1.14	0
AY.124	0
BA.2	24
AY.90	1
AY.46	0

Japanese Ministry of Health, Labour and Welfare
https://www.mhlw.go.jp/stf/seisakunitsuite/bunya/0000121431_00333.html

How can we fight against COVID-19 with NGSs?

1. revealing the genome of the virus, SARS-CoV-2

2. revealing the genome of the host, humans

The COVID-19 Host Genetics Initiative

https://www.covid19hg.org/

Iceland
GenOMICC Genetics Of Mortality in Critical Care)
Canada
Ireland
Genes & Health
Ukraine
COVID-19 Host Response Pathogenesis (CHIRP) Study
Columbia University COVID19 Biobank
GEN-COVID Project
Genome-wide assessment of the gene variants associated with severe COVID-19 phenotype in Iran
Corea (Ge of COVID- Manifesta
Greece
MexGen-COVID Initiative
Morocco
Tunisia
Qatar Genome Program
SIGMA
Cuba
Libya
Egypt
Oman
India
Myanmar (Burma)
Mauritania
Niger
Yemen
Cambodia
Ethiopia
Cameroon
Kenya
Maldives
Malaysia
Colombia
Indonesia
mapbox
Genetic determinants
© Mapbox © OpenStreetMap Improve this map
+
−

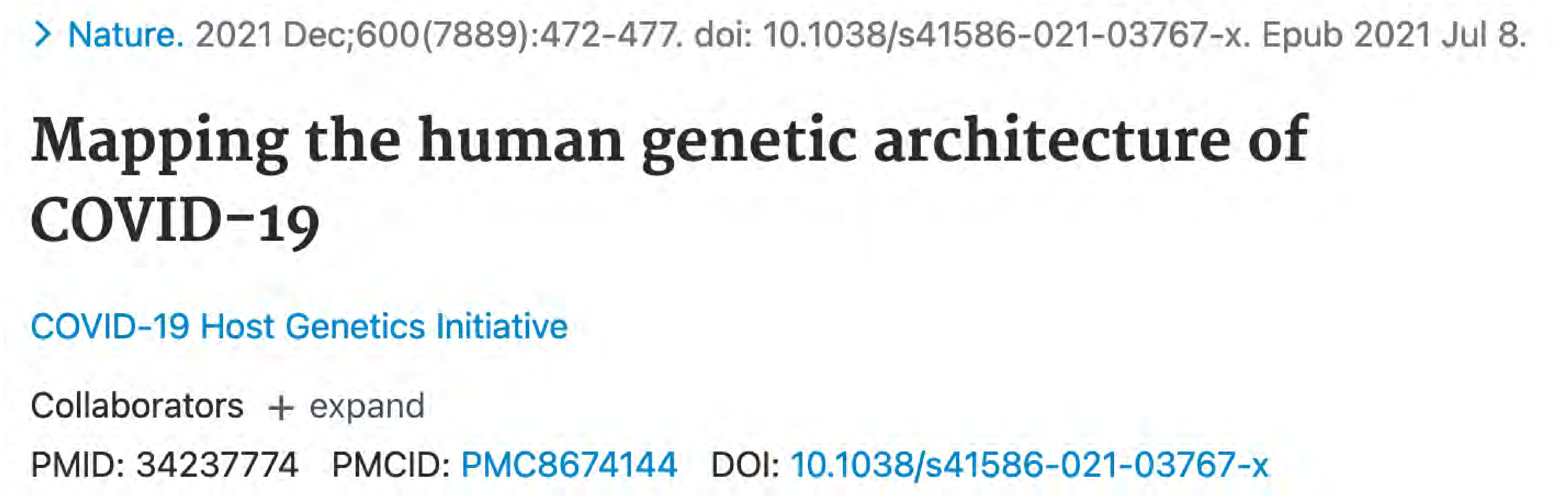
- 115 registered studies
- about 50,000 COVID-19 patients
- about 2,000,000 controls

Joint Research Coronavirus Task Force

https://www.covid19-taskforce.jp/

- Funded by the Japan Agency for Medical Research and Development (AMED)
- Started with 40 medical institutes in Japan
  - Currently over 100 institutes are participating.
- Biggest Asian participant of COVID-19HG

- Applied the Genome-Wide Association Study (GWAS) strategy
- GWAS focuses on association between single-nucleotide polymorphisms (SNPs) and traits such as disease susceptibility.



> Nature. 2021 Dec;600(7889):472-477. doi: 10.1038/s41586-021-03767-x. Epub 2021 Jul 8.

Mapping the human genetic architecture of COVID-19

COVID-19 Host Genetics Initiative

Collaborators + expand

PMID: 34237774 PMCID: PMC8674144 DOI: 10.1038/s41586-021-03767-x

The Japanese task force is a the biggest team in Asia, and only team that offered severe COVID-19 case data.

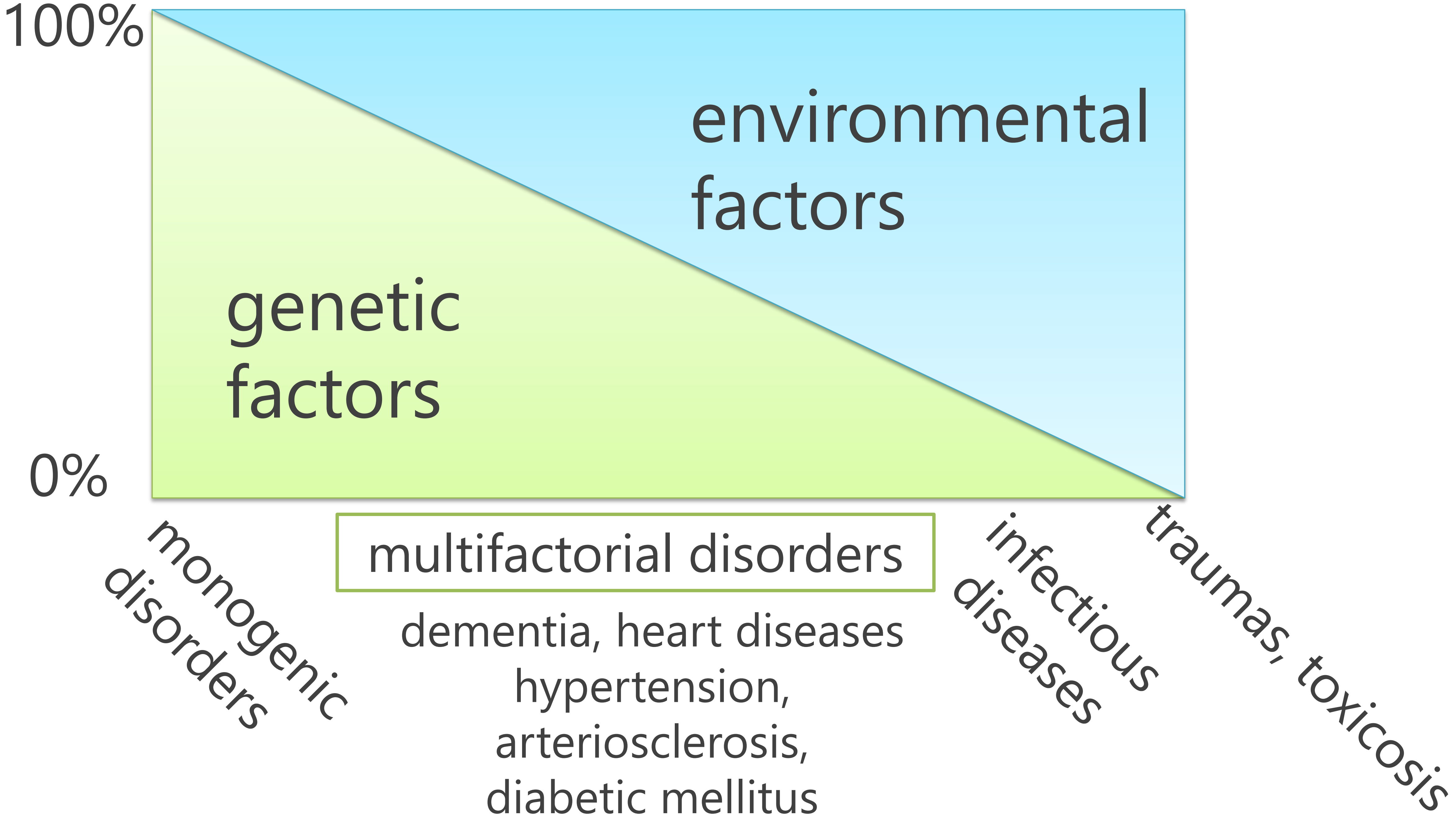
- Found 13 variants affect COVID-19 severity.
- Two variants showed high frequency in East and South Asians compared with Europeans.
  - *FOXP1* (lung cell proliferation)
  - *DPP9* (related to interstitial pneumonia)
  - *TYK2* (related to immunity)

- Previously found SNPs near *DOCK2* in Japanese as a disease progression factor was not reported in this paper.
- It may be explained by very low frequency of variations of *DOCK2* in Europeans.
- Population diversity is important for genetic researches.

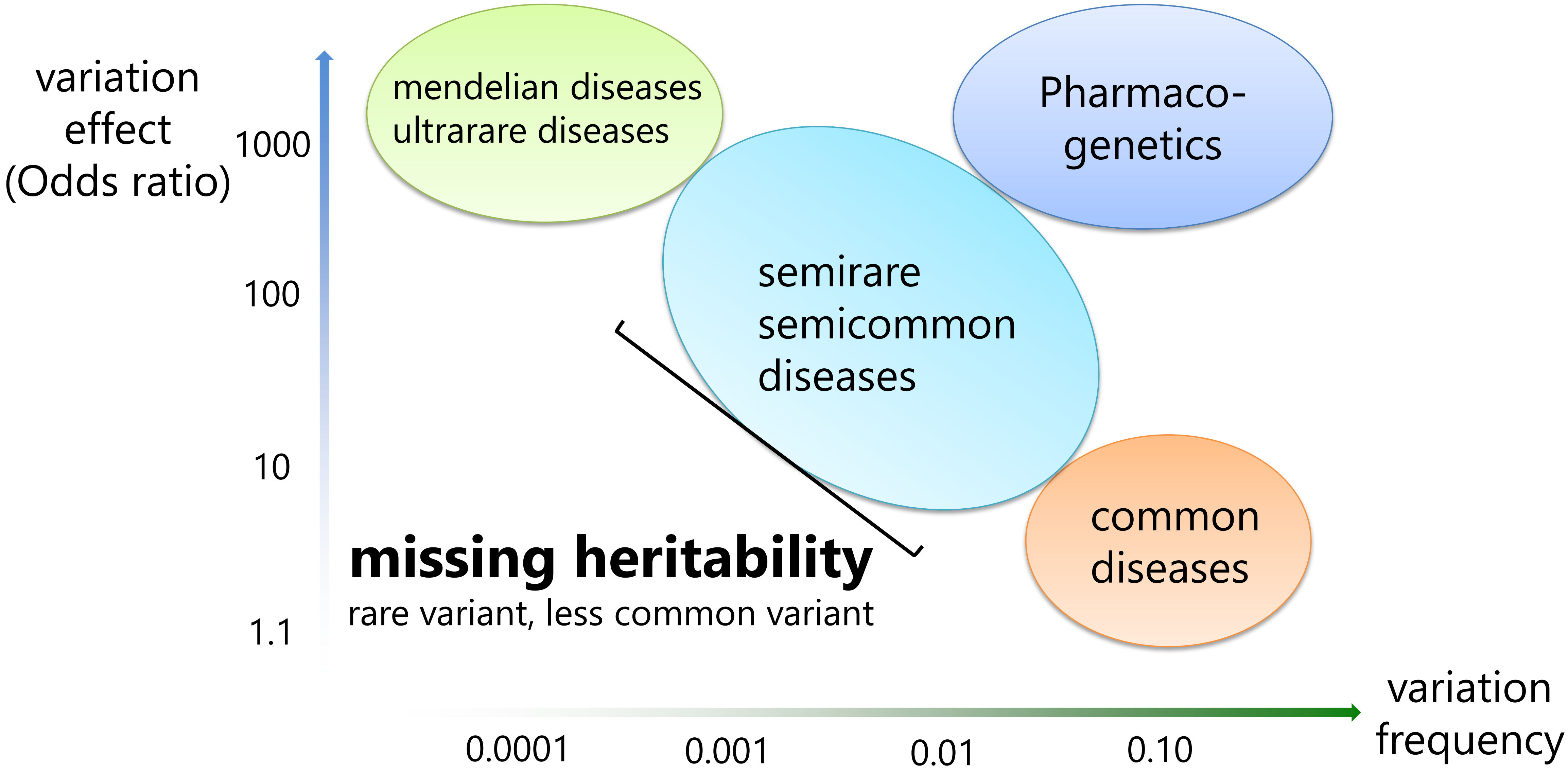
To be continued to the part 2….

PART 1:
Japan’s experience and lessons learned with the next generation genomic sequencing system for the COVID-19 response

PART 2:
Human genome sequencing and analysis



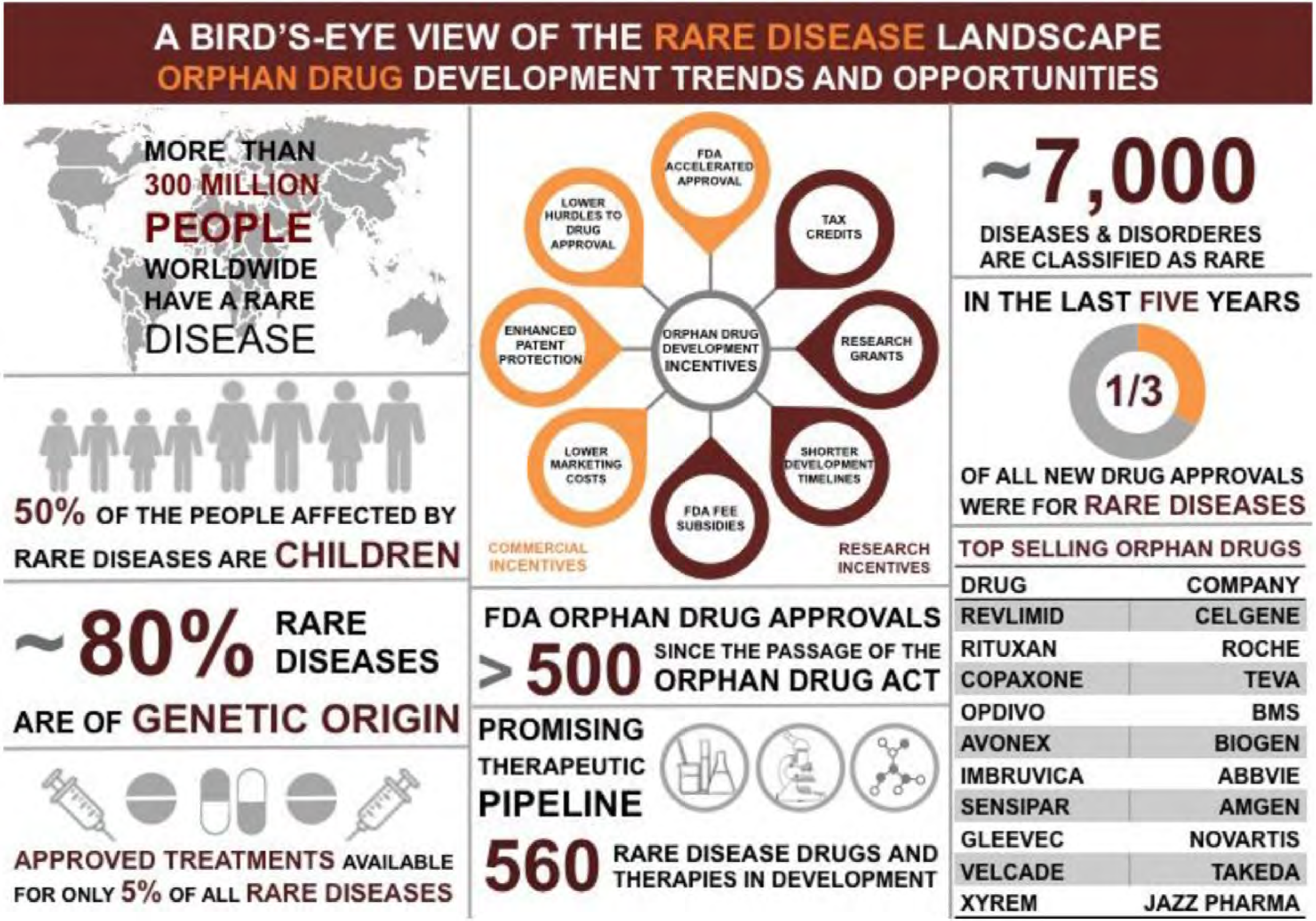
dementia, heart diseases
hypertension,
arteriosclerosis,
diabetic mellitus



genetic diseases clustered by variation frequency and effect odds ratio
（based on Manolio et al., Nature (2009) 461: 747-753）

Our mission is “Gene Hunting”

Revealing causative gene of inheritance disorders



A BIRD'S-EYE VIEW OF THE RARE DISEASE LANDSCAPE
ORPHAN DRUG DEVELOPMENT TRENDS AND OPPORTUNITIES

MORE THAN 300 MILLION PEOPLE WORLDWIDE HAVE A RARE DISEASE

50% OF THE PEOPLE AFFECTED BY RARE DISEASES ARE CHILDREN

~80% RARE DISEASES ARE OF GENETIC ORIGIN

APPROVED TREATMENTS AVAILABLE FOR ONLY 5% OF ALL RARE DISEASES

ORPHAN DRUG DEVELOPMENT INCENTIVES

- FDA ACCELERATED APPROVAL
- TAX CREDITS
- RESEARCH GRANTS
- SHORTER DEVELOPMENT TIMELINES
- FDA FEE SUBSIDIES
- LOWER MARKETING COSTS
- ENHANCED PATENT PROTECTION
- LOWER HURDLES TO DRUG APPROVAL

COMMERCIAL INCENTIVES

RESEARCH INCENTIVES

FDA ORPHAN DRUG APPROVALS

> 500 SINCE THE PASSAGE OF THE ORPHAN DRUG ACT

PROMISING THERAPEUTIC PIPELINE

560 RARE DISEASE DRUGS AND THERAPIES IN DEVELOPMENT

~7,000 DISEASES & DISORDERES ARE CLASSIFIED AS RARE

IN THE LAST FIVE YEARS

1/3

OF ALL NEW DRUG APPROVALS WERE FOR RARE DISEASES

TOP SELLING ORPHAN DRUGS

DRUG	COMPANY
REVLIMID	CELGENE
RITUXAN	ROCHE
COPAXONE	TEVA
OPDIVO	BMS
AVONEX	BIOGEN
IMBRUVICA	ABBVIE
SENSIPAR	AMGEN
GLEEVEC	NOVARTIS
VELCADE	TAKEDA
XYREM	JAZZ PHARMA

Bhattacharjee et al., *Cancers* 10 (9), 2018 https://doi.org/10.3390/cancers10090298.

IRUD-P

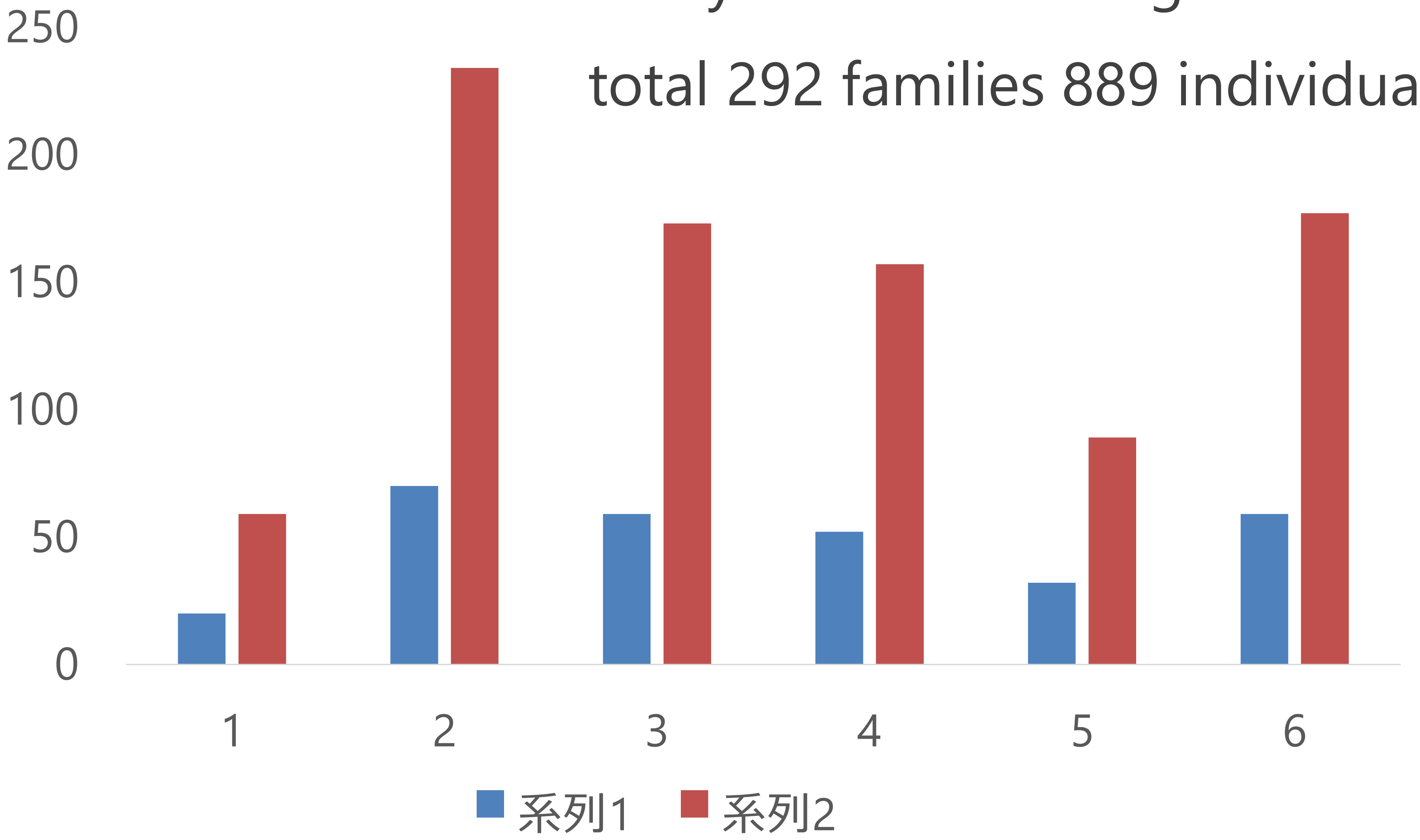
Initiative on Rare and Undiagnosed Diseases in Pediatrics

- Initiative on Rare and Undiagnosed Diseases in Pediatrics (IRUD-P)
- A project funded by the Japan Agency for Medical Research and Development (AMED).
- In the majority of participants, family trios are analyzed by whole exome sequencing (WES) with *de novo* inheritance model.

IRUD-P
Initiative on Rare and Undiagnosed Diseases in Pediatrics

Analyzed cases in Nagasaki U

total 292 families 889 individuals



NGS analysis is “big data science”

- a compressed 30x WGS FASTQ file ≈ 60 Gbyte
- Including intermediate files, about 3-5 times storage size is necessary
  - 300Gb / sample

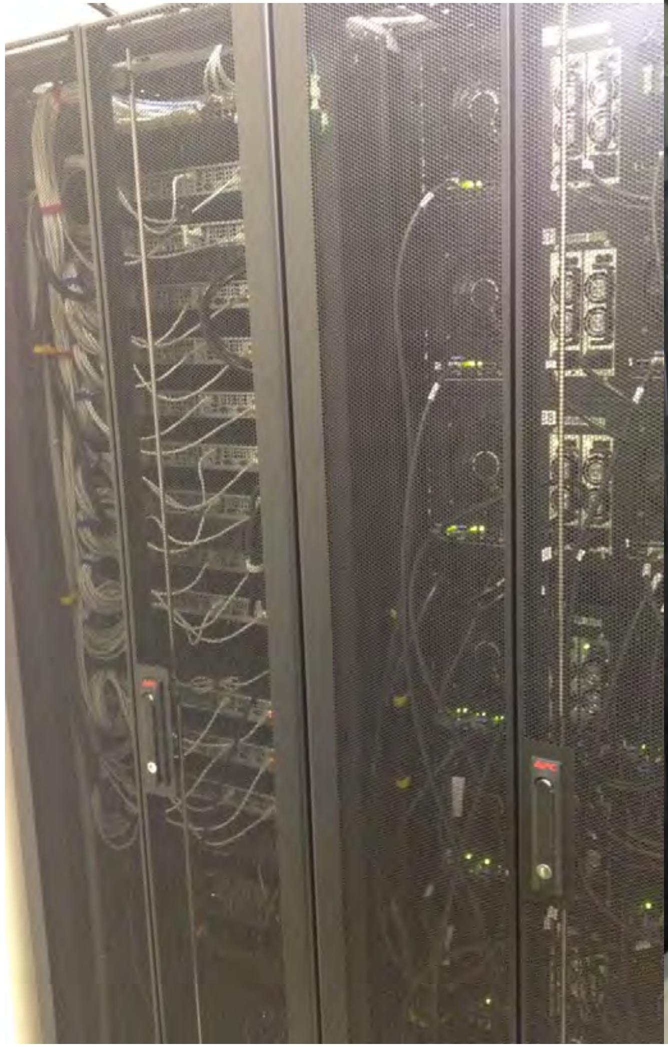
We built a hand-made large storage and HPC cluster

2014/04/08
ESD BAG
20 PACK
20 PACK
ESD BAG
ESD BAG
ESD BAG

ESD BAG
ESD BAG
20 PACK
20 PACK
2014/04/08

2014/04/08

3Tbyte
desktop SATA
x 384



549TB + 768 core

NGS researches require dry, wet and “sweat” experiments!

Graphic Processing Units (GPUs) in Bioinformatics

30x human WGS mapping

- HPC cluster: 12 samples in 3 days
- GPU server: 1 sample in40min. (108 sample/3days)

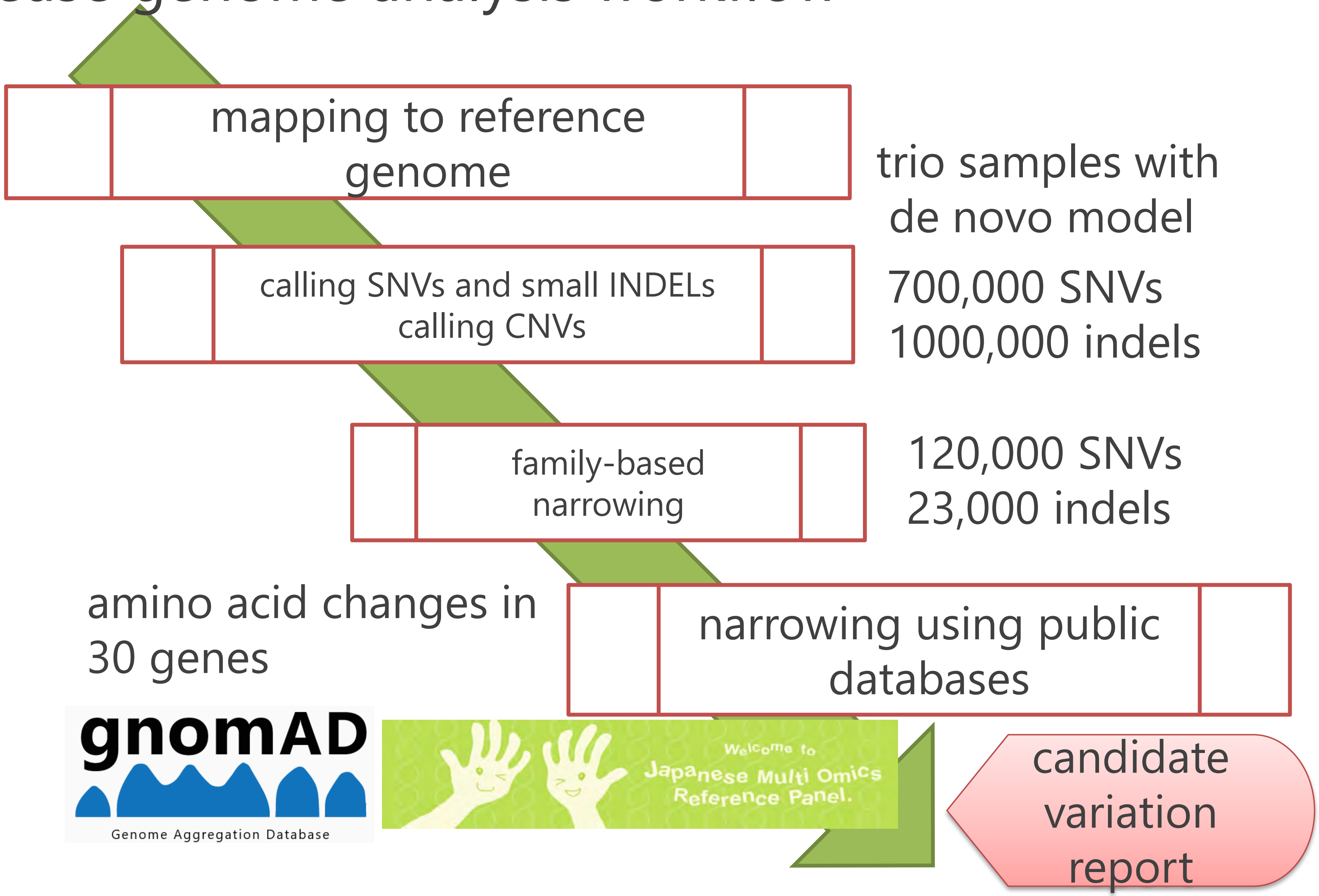
Bioinformatics and Unix-like systems

- Unix-like systems, such as Linux, are "mother-language" of Bioinformatics.
- MacOS's "terminal" and Windows' "Subsystem for Linux (WSL)" are also Unix-like systems
- Basically, not graphic user interface (GUI)-based but character user interface (CUI)-based system.
- Unix-like systems have advantages in handling multi-user, multi-process, large memory consuming, network distributed, and long continuous computing like bioinformatics analysis.

Bioinformatics and Unix-like systems

- Most of bioinformatics software packages require Unix-based systems
- Bioinformatics analyses requires workflow management combination of multiple packages written by different scientists including you.
- Yes, GUI systems are easy to use.
- CUI systems are good for workflow management.
- On CUI system, tiny scripts written by yourselves are self-documented, machine-readable, and reproducible workflow description. …. Use Linux.

Rare disease genome analysis workflow



mapping to reference genome

trio samples with de novo model

calling SNVs and small INDELs
calling CNVs

700,000 SNVs
1000,000 indels

family-based narrowing

120,000 SNVs
23,000 indels

amino acid changes in 30 genes

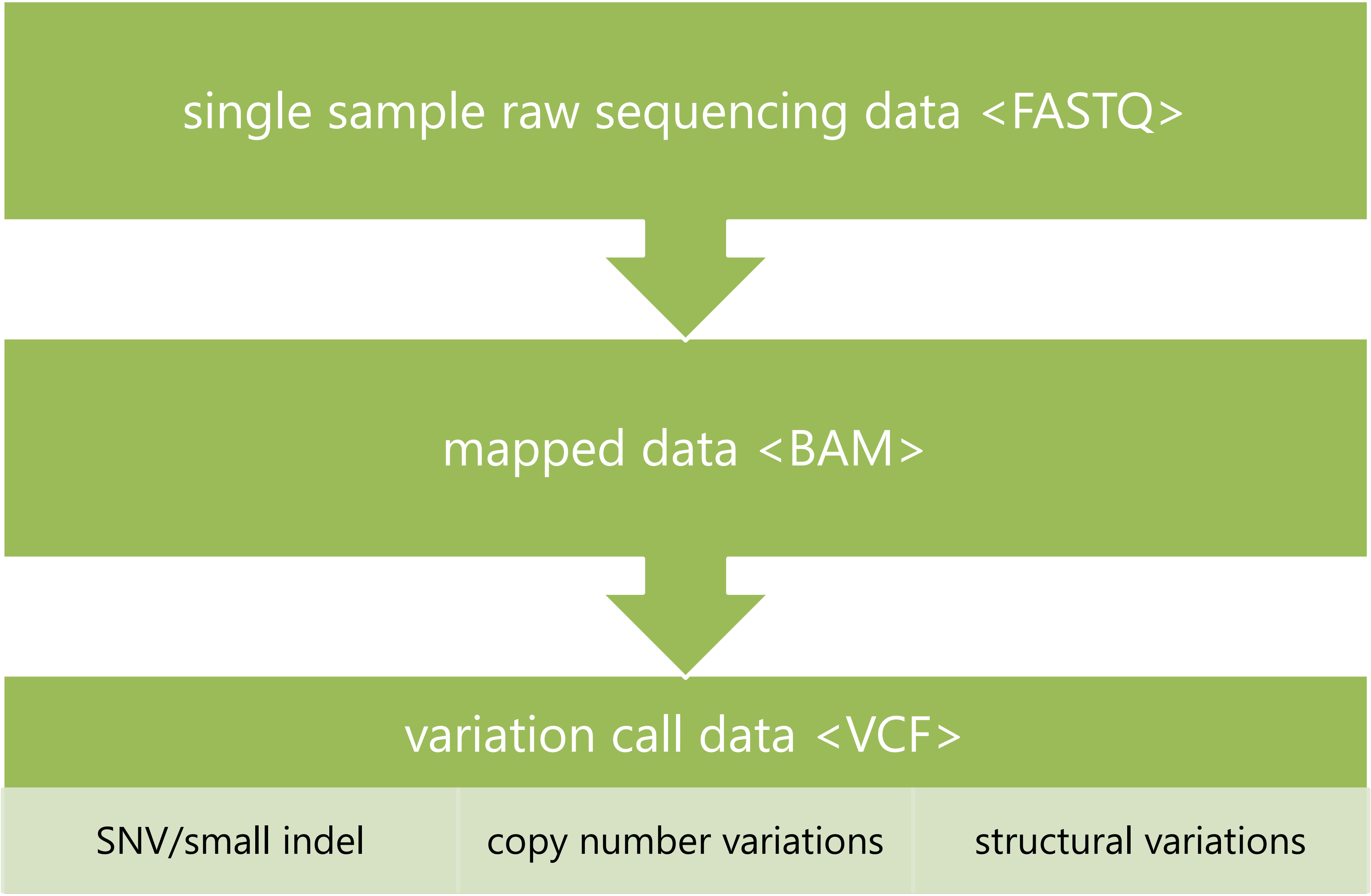
narrowing using public databases

gnomAD
Genome Aggregation Database

Welcome to
Japanese Multi Omics Reference Panel.

candidate variation report

NGS data processing outline



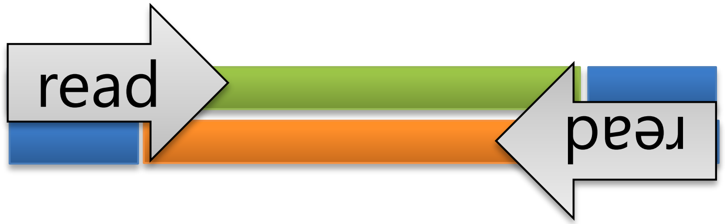
single sample raw sequencing data <FASTQ>

mapped data <BAM>

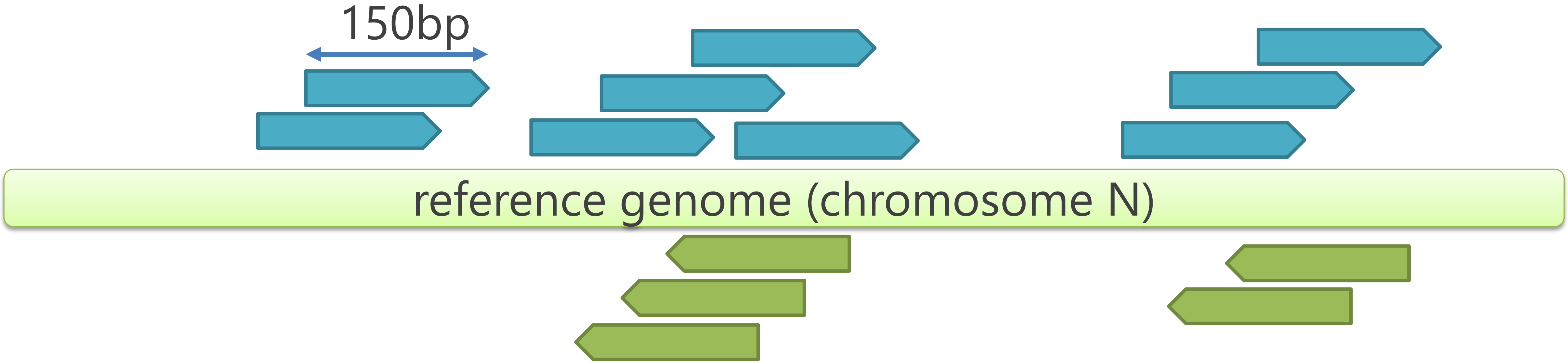
variation call data <VCF>

SNV/small indel	copy number variations	structural variations

mapping / alignment



- For each reads, finding the most similar, but not necessarily identical sequence in the reference genome.
- The BWA software package is de facto-standard.


150bp

reference genome (chromosome N)

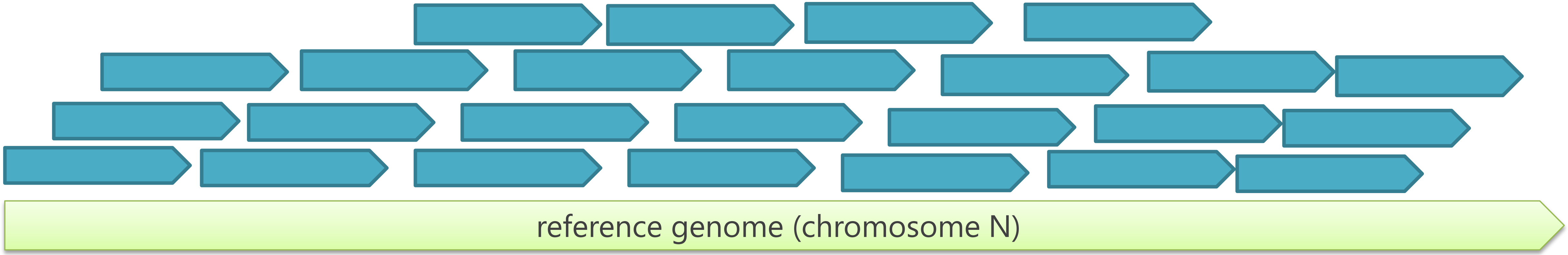
GRC Genome Reference Consortium

human reference genome assemblies

UCSC name	GRC name	release
hg18	NCBI 36	2006/03
hg19	GRCh37	2009/04
hg38	GRCh38	2013/12

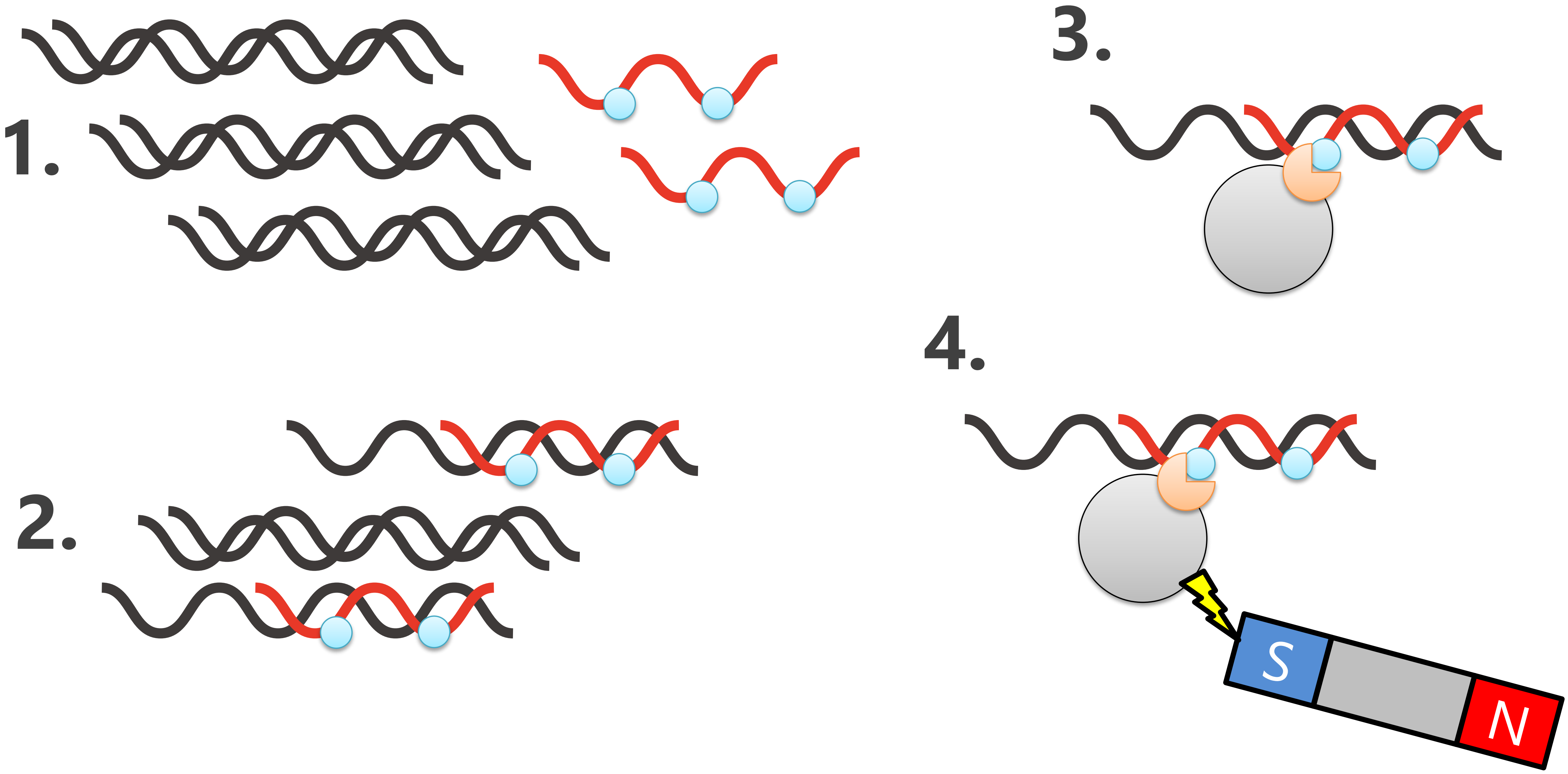
NGS human genomic data analysis

Whole Genome Sequencing (WGS)



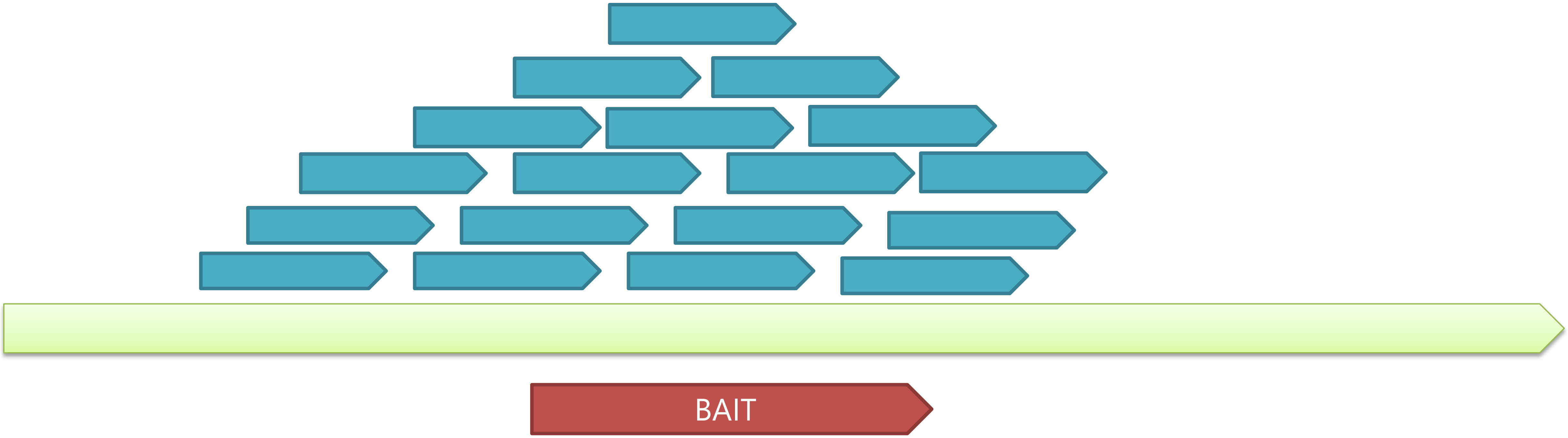
(indicating only reads mapped on the plus strand)

exome enrichment using RNA baits



NGS human genomic data analysis

Whole Exome Sequencing (WES)



Pros and cons in WGS and WES

	WGS	WES
Target	Whole genome (50x larger than WES)	Genes (about 2% of genome)
Effectivity to find gene mutations	Low	High
Exome capture bait reagents	No	Yes (not cheap)
Capturing benchwork	No	Yes
depth bias	Low	High
Outside of exons	Yes	No
SNVs and indel	Yes	Yes
CNVs	Yes	Yes but noisy

the FASTQ file format

- text-based format
- a nucleotide sequence (=FASTA) + quality scores

```
@READ_NAME
NUCLEATIDE_SEQUENCE
+
NUCLEOTIDE_QUALITY
```

```
@HWI-D00385:284:HKJCLBCX3:1:1101:1155:2118 1:N:0:ATTCCTTTTCTTTCCC
AGGTCAAGCAGAGTGCCACACAGGCCTGTGAGGCATCTGAGGTCCAACTAGCCAGTGTTGA
GTGTCCCAGCTGATCACTCACAGAATTTTCTAGTGATCCC
+
DDD<B<CDFHE@HHCHCHEHIIHHHHICD1FDCHIIIIHHIIGHHIIIIIGHHIFHHHH?GHGEG
HHHIIIIHHIIIIIIIIH?FHEHGHHII@HECHE@F
```

the SAM/BAM file format

- sequence alignment/mapping format
- text-based (SAM) and binary (BAM) formats
- FASTQ-based information + mapping information
- unsorted, read name-wise sorted, coordination sorted.
- Header:
  - sort status, reference genome, sequencer, sample information
- Body:
  - chromosome, position, FASTQ, quality scores, mapping details (CIGAR)

the VCF file format

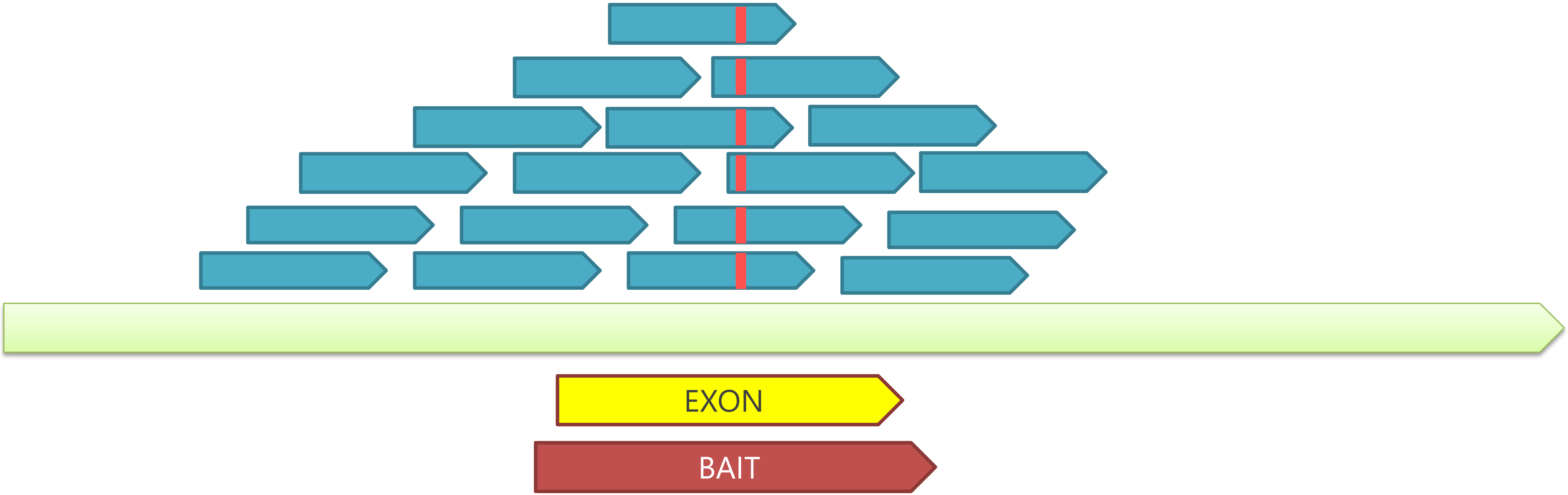
- variant call format
- text-based (vcf) and binary (bcf) formats
- variant information and their quality scores
- can include multiple sample information

#CHROM	POS	ID	REF	ALT	QUAL	FILTER	INFO	FORMAT	MY_SAMPLE1
chr20	14370	rs6054257	G	A	29	PASS	NS=1;DP=14;AF=0.5	GT:GQ:DP:HQ	0|1:48:8:51,51
chr20	17330	.	T	A	3	q10	NS=1;DP=11;AF=0.017	GT:GQ:DP:HQ	0/1:3:5:65,3

Single Nucleotide Variation (SNV) calling

A homozygous SNV in WES

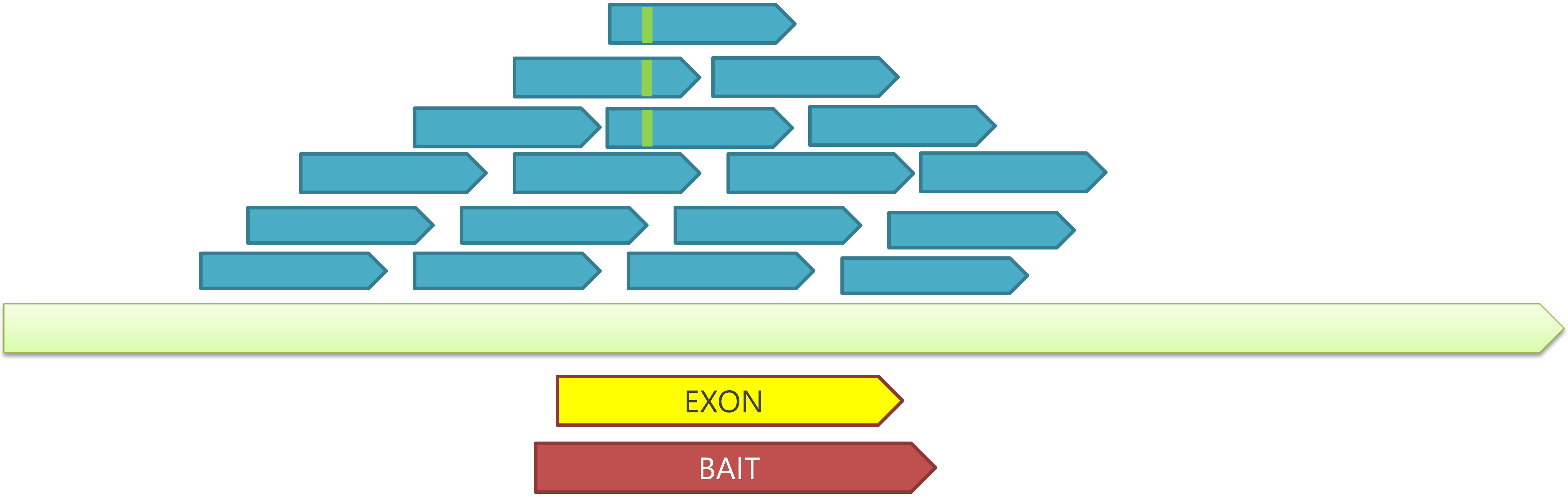
🟥 = alternative nucleotide



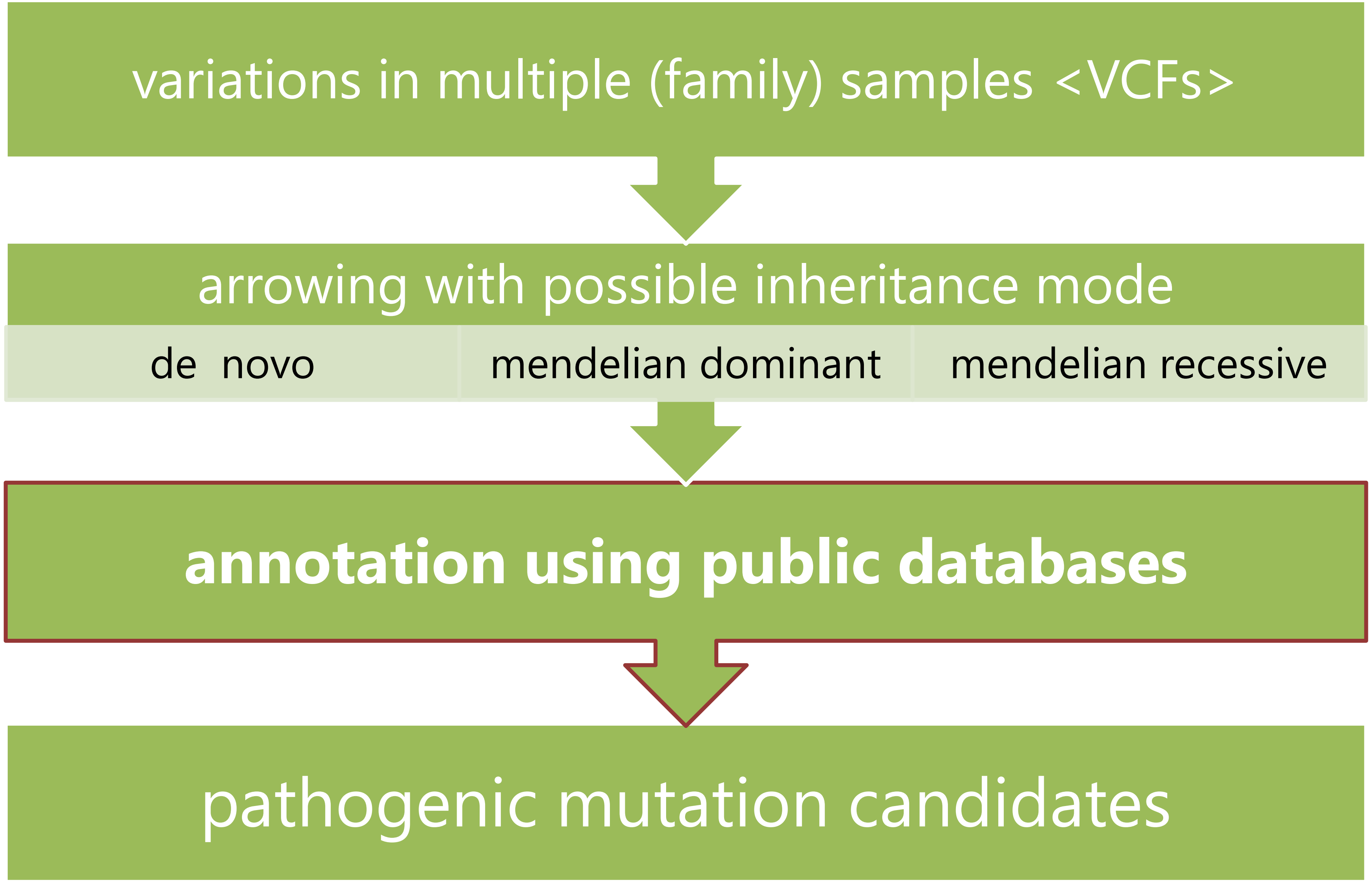
Single Nucleotide Variation (SNV) calling

A heterozygous SNV in WES

🟩 = alternative nucleotide



Interpretation of pathogenic variations



variations in multiple (family) samples <VCFs>

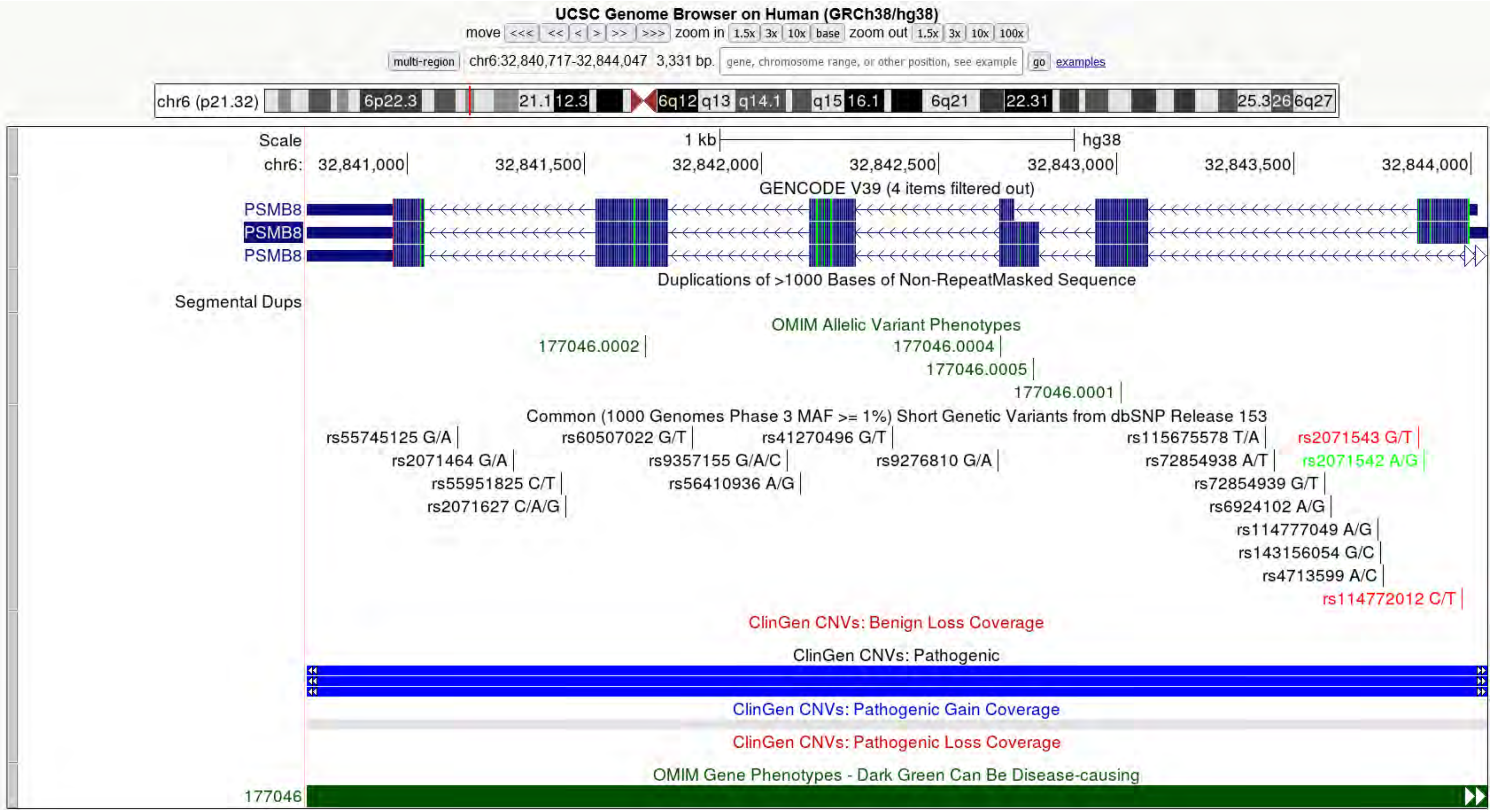
arrowing with possible inheritance mode

de novo	mendelian dominant	mendelian recessive

annotation using public databases

pathogenic mutation candidates

https://genome.ucsc.edu/



UCSC Genome Browser on Human (GRCh38/hg38)

chr6:32,840,717-32,844,047 3,331 bp.

Important Disease-related Public Databases

50 YEARS OMIM Human Genetics Knowledge for the World

OMIM®

Online Mendelian Inheritance in Man®

An Online Catalog of Human Genes and Genetic Disorders

https://www.omim.org/

ACTGATGGTATGGGGCCAAGAGATATATCT
CAGGTACGGCTGTCATCACTTAGACCTCAC
CAGGGCTGGGCATAAAAGTCAGGGCAGAGC
CCATGGTGCATCTGACTCCTGAGGAGAAGT
GCAGGTTGGTATCAAGGTTACAAGACAGGT
GGCACTGACTCTCTCTGCCTATTGGTCTAT

ClinVar

ClinVar aggregates information about genomic variation and its relationship to human health.

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

COSMIC

Catalogue Of Somatic Mutations In Cancer

https://cancer.sanger.ac.uk/cosmic

GWAS Catalog

The NHGRI-EBI Catalog of human genome-wide association studies

Search the catalog

Examples: breast carcinoma, rs7329174, Yao, 2q37.1, HBS1L, 6:16000000-25000000

https://www.ebi.ac.uk/gwas/

TOGOVAR A comprehensive Japanese genetic variation database

https://togovar.biosciencedbc.jp/

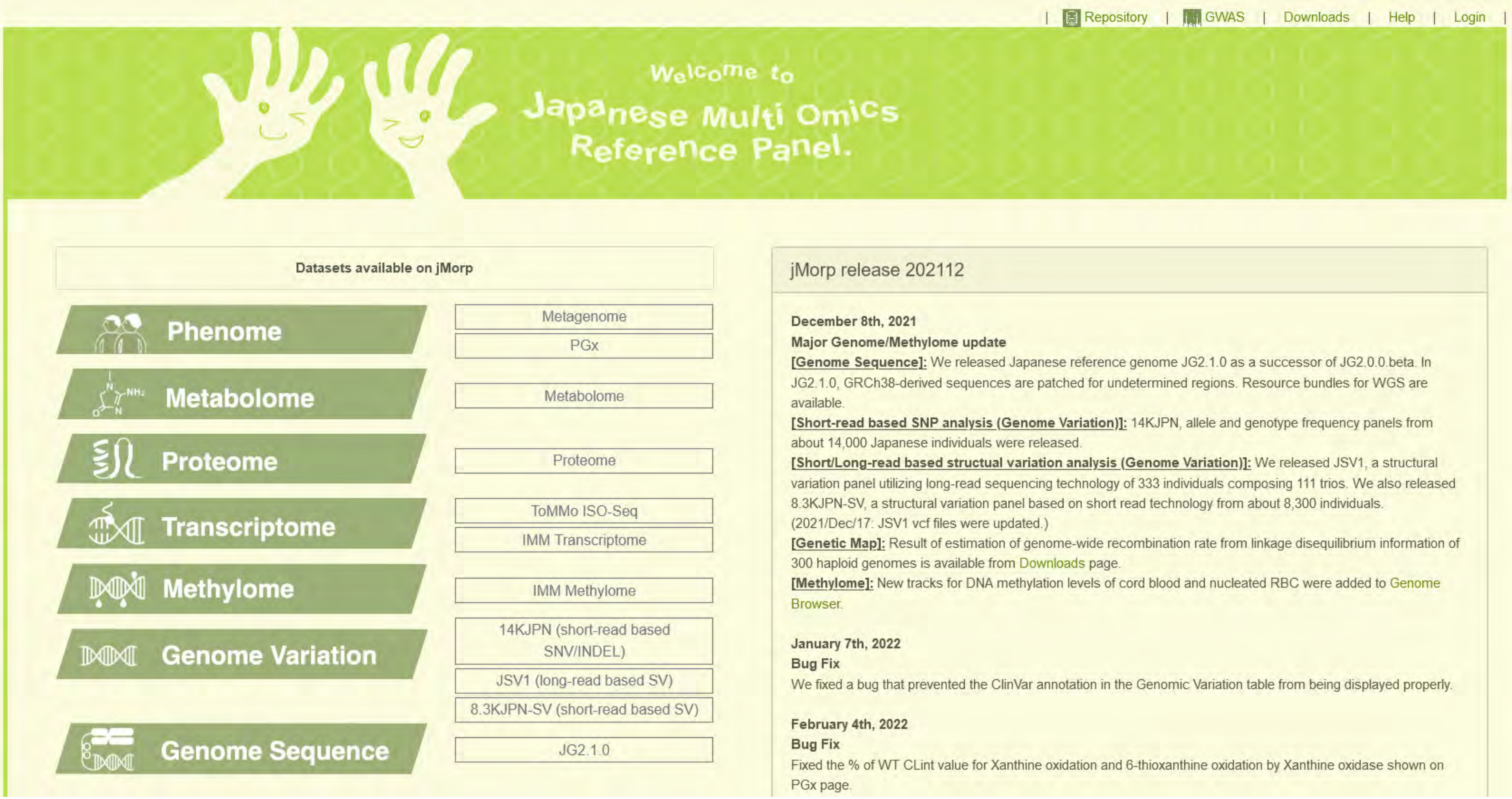
The Genome Aggregation Database (gnomAD)

https://gnomad.broadinstitute.org/

- The world biggest human genome variation DB
- WES 125,748 and WGS 76,156 samples
- Common variations in gnomAD can be omitted from candidates.
- Including European, African, Asian and Latin American populations.
- Sample size of Japanese (76) and Middle east (158) population is still small.

jMorp of Tohoku Medical Megabank (ToMMo)

https://jmorp.megabank.tohoku.ac.jp/ including “14KJPN”



| Repository | GWAS | Downloads | Help | Login |

Welcome to Japanese Multi Omics Reference Panel.

Datasets available on jMorp

Phenome	Metagenome PGx
Metabolome	Metabolome
Proteome	Proteome
Transcriptome	ToMMo ISO-Seq IMM Transcriptome
Methylome	IMM Methylome
Genome Variation	14KJPN (short-read based SNV/INDEL) JSV1 (long-read based SV) 8.3KJPN-SV (short-read based SV)
Genome Sequence	JG2.1.0

jMorp release 202112

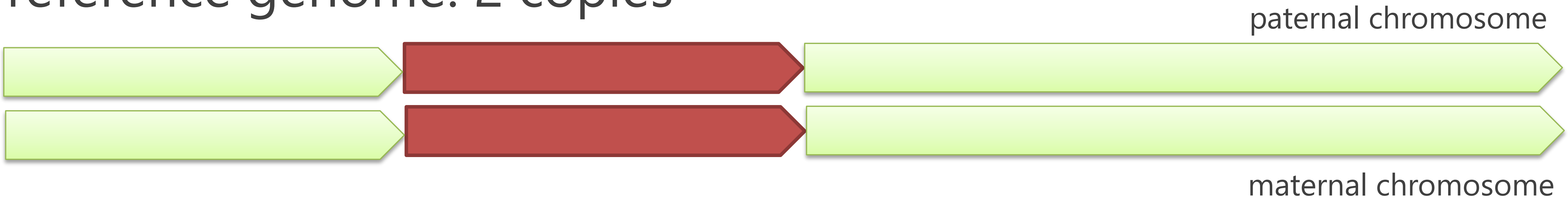
December 8th, 2021
Major Genome/Methylome update
[Genome Sequence]: We released Japanese reference genome JG2.1.0 as a successor of JG2.0.0.beta. In JG2.1.0, GRCh38-derived sequences are patched for undetermined regions. Resource bundles for WGS are available.
[Short-read based SNP analysis (Genome Variation)]: 14KJPN, allele and genotype frequency panels from about 14,000 Japanese individuals were released.
[Short/Long-read based structual variation analysis (Genome Variation)]: We released JSV1, a structural variation panel utilizing long-read sequencing technology of 333 individuals composing 111 trios. We also released 8.3KJPN-SV, a structural variation panel based on short read technology from about 8,300 individuals. (2021/Dec/17: JSV1 vcf files were updated.)
[Genetic Map]: Result of estimation of genome-wide recombination rate from linkage disequilibrium information of 300 haploid genomes is available from Downloads page.
[Methylome]: New tracks for DNA methylation levels of cord blood and nucleated RBC were added to Genome Browser.

January 7th, 2022
Bug Fix
We fixed a bug that prevented the ClinVar annotation in the Genomic Variation table from being displayed properly.

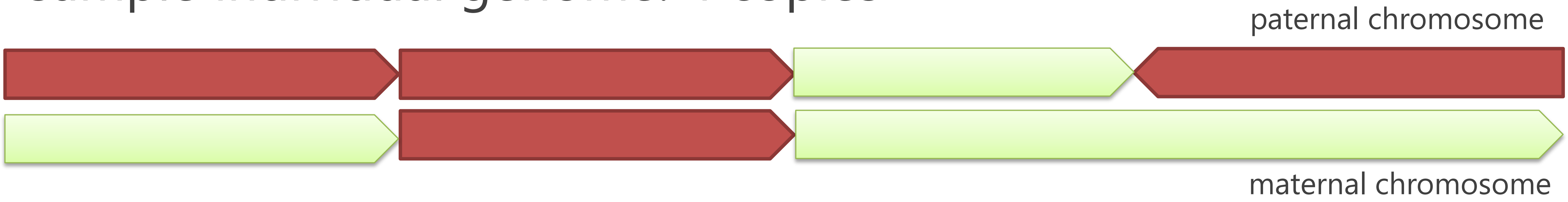
February 4th, 2022
Bug Fix
Fixed the % of WT CLint value for Xanthine oxidation and 6-thioxanthine oxidation by Xanthine oxidase shown on PGx page.

copy number variations (CNVs)

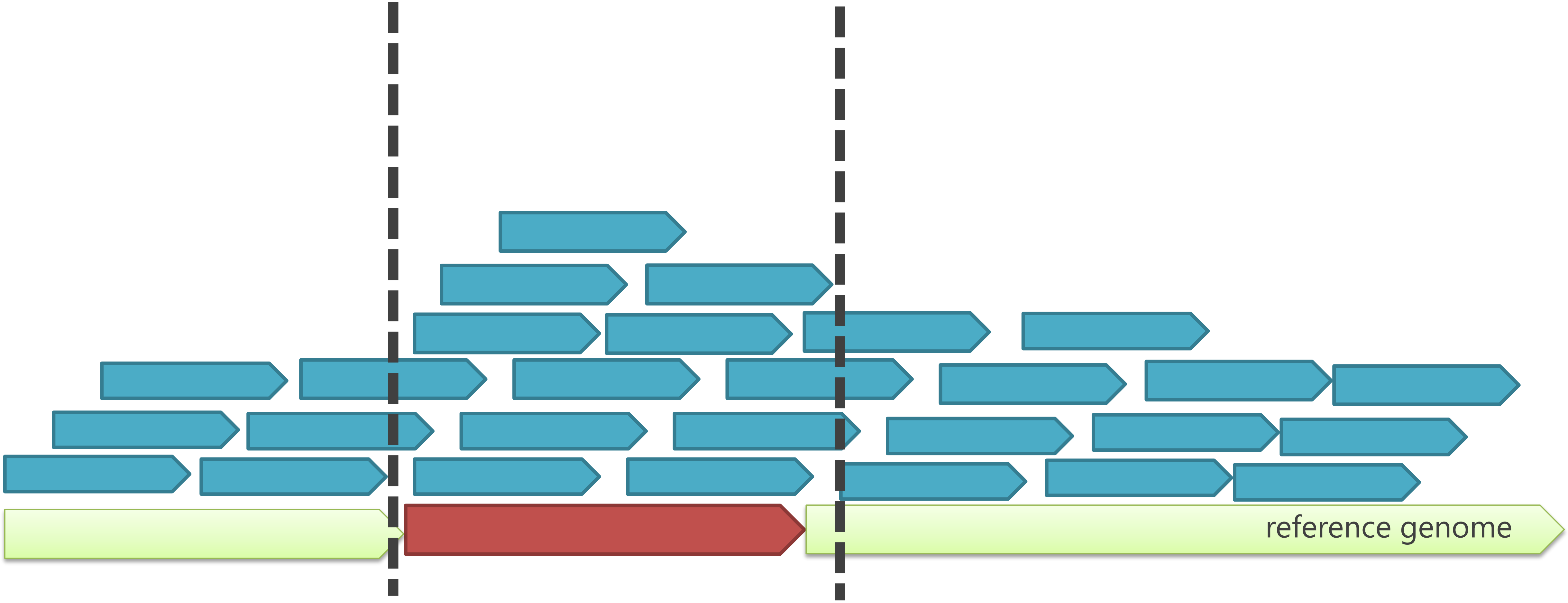
reference genome: 2 copies



sample individual genome: 4 copies



CNV calling in WGS

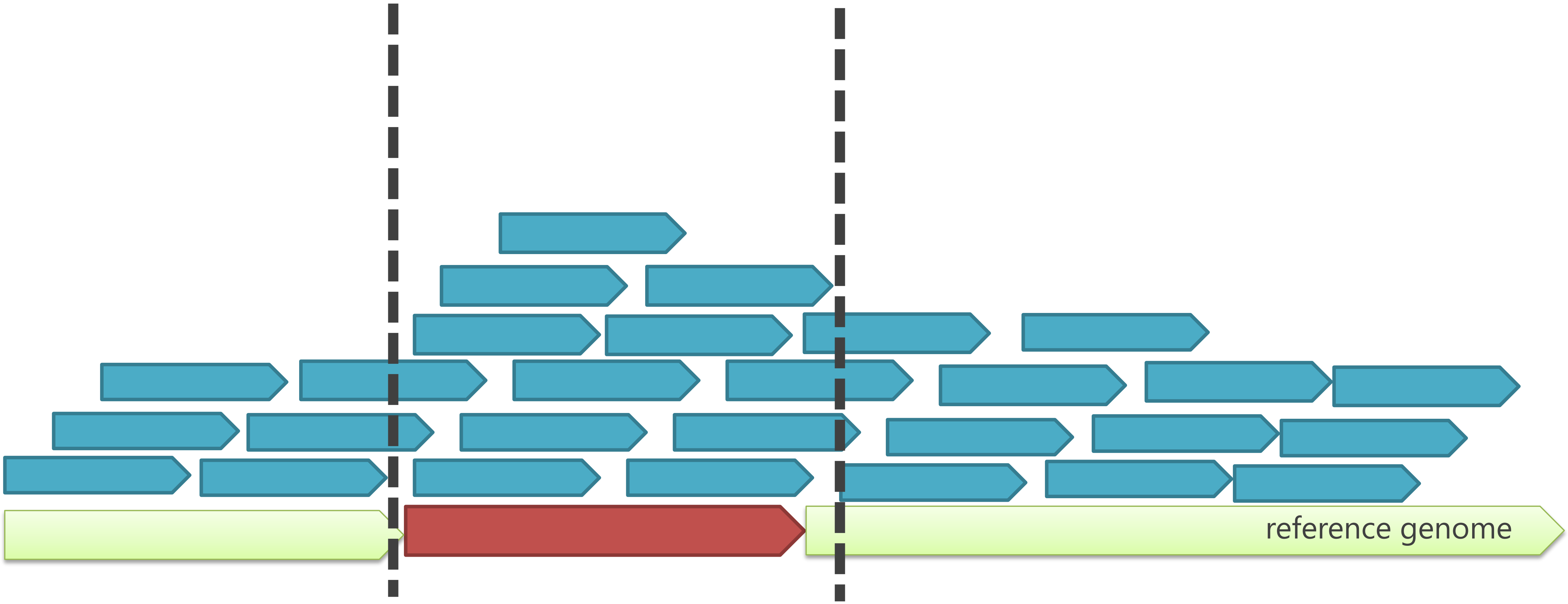


mean depth = 3.0

mean depth = 4.5.

mean depth = 3.1

CNV calling in WGS



2 copies estimated

3 copies estimated

2 copies estimated

normalization of WGS/WES depths

- WGS – lower noises
  - genome-wide average
  - reference GC%
  - reference complexity

CNVnatore
https://github.com/abyzovlab/CNVnator
EXCAVATOR2
https://sourceforge.net/projects/excavator2tool/
cn.MOPS
http://www.bioinf.jku.at/software/cnmops/cnmops.html

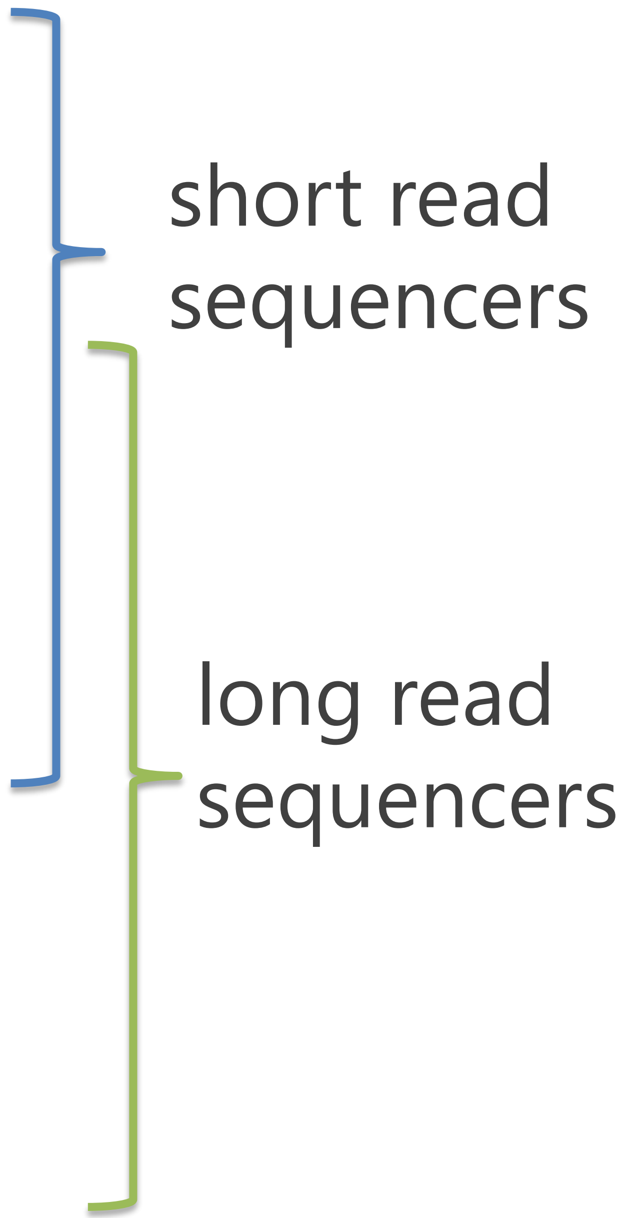
- WES – higher noises
  - bait bias / bait design
  - inter-experimental noise
  - batch effect

XHMM
https://statgen.bitbucket.io/xhmm/

CNVkit
https://cnvkit.readthedocs.io/

Genomic variations

- Single Nucleotide Variations (SNVs)
- small insertions and deletions (indels)
- copy-number variation (CNVs)
- genomic structural variations (SVs)
  - (large) insertion
  - (large) deletion
  - inversion
  - duplication
- Repeated sequence
  - simple repeats
  - interspersed elements (LINE/SINE)
  - heterochromatin / telomeres



short read sequencers

long read sequencers

Today’s take-home messages

In NGS research projects of human health…

- Both “Wet” and “dry” experiments are essential for successful analyses.
- A target population diversity, including Tunisians and Japanese, is essential.
- Building public databases for Tunisians and Japanese is a challenge for the future.